

# The Book of Statistical Proofs

DOI: 10.5281/zenodo.4305949

<https://statproofbook.github.io/>

StatProofBook@gmail.com

2025-12-30, 13:25

# Contents

|          |  |          |
|----------|--|----------|
| <b>I</b> | <b>General Theorems</b>                                  | <b>1</b> |
| 1        | Probability theory . . . . .                             | 2        |
| 1.1      | Random experiments . . . . .                             | 2        |
| 1.1.1    | <i>Random experiment</i> . . . . .                       | 2        |
| 1.1.2    | <i>Sample space</i> . . . . .                            | 2        |
| 1.1.3    | <i>Event space</i> . . . . .                             | 2        |
| 1.1.4    | <i>Probability space</i> . . . . .                       | 2        |
| 1.1.5    | <i>Measured data</i> . . . . .                           | 3        |
| 1.1.6    | <i>Statistical sample</i> . . . . .                      | 3        |
| 1.1.7    | <i>Sample size</i> . . . . .                             | 3        |
| 1.1.8    | <i>Sample statistic</i> . . . . .                        | 3        |
| 1.1.9    | <i>Descriptive vs. inferential</i> . . . . .             | 4        |
| 1.2      | Random variables . . . . .                               | 4        |
| 1.2.1    | <i>Random event</i> . . . . .                            | 4        |
| 1.2.2    | <i>Random variable</i> . . . . .                         | 4        |
| 1.2.3    | <i>Random vector</i> . . . . .                           | 5        |
| 1.2.4    | <i>Random matrix</i> . . . . .                           | 5        |
| 1.2.5    | <i>Constant</i> . . . . .                                | 5        |
| 1.2.6    | <i>Discrete vs. continuous</i> . . . . .                 | 5        |
| 1.2.7    | <i>Univariate vs. multivariate</i> . . . . .             | 6        |
| 1.2.8    | <i>independent and identically distributed</i> . . . . . | 6        |
| 1.3      | Probability . . . . .                                    | 6        |
| 1.3.1    | <i>Probability</i> . . . . .                             | 6        |
| 1.3.2    | <i>Joint probability</i> . . . . .                       | 7        |
| 1.3.3    | <i>Marginal probability</i> . . . . .                    | 7        |
| 1.3.4    | <i>Conditional probability</i> . . . . .                 | 7        |
| 1.3.5    | <i>Exceedance probability</i> . . . . .                  | 8        |
| 1.3.6    | <i>Statistical independence</i> . . . . .                | 8        |
| 1.3.7    | <i>Conditional independence</i> . . . . .                | 9        |
| 1.3.8    | <b>Self-independence</b> . . . . .                       | 10       |
| 1.3.9    | <b>Probability under independence</b> . . . . .          | 10       |
| 1.3.10   | <i>Mutual exclusivity</i> . . . . .                      | 11       |
| 1.3.11   | <b>Probability under exclusivity</b> . . . . .           | 11       |
| 1.4      | Probability axioms . . . . .                             | 12       |
| 1.4.1    | <i>Axioms of probability</i> . . . . .                   | 12       |
| 1.4.2    | <b>Monotonicity of probability</b> . . . . .             | 12       |
| 1.4.3    | <b>Monotonicity of probability</b> . . . . .             | 13       |
| 1.4.4    | <b>Probability of the empty set</b> . . . . .            | 14       |

|        |  |    |
|--------|--|----|
| 1.4.5  | <b>Probability of the empty set</b>  | 14 |
| 1.4.6  | <b>Probability of the complement</b>   | 15 |
| 1.4.7  | <b>Range of probability</b>  | 15 |
| 1.4.8  | <b>Addition law of probability</b>   | 16 |
| 1.4.9  | <b>Bonferroni's inequality</b>   | 17 |
| 1.4.10 | <b>Boole's inequality</b>  | 17 |
| 1.4.11 | <b>Law of total probability</b>  | 19 |
| 1.4.12 | <b>Probability of exhaustive events</b>  | 19 |
| 1.4.13 | <b>Probability of exhaustive events</b>  | 20 |
| 1.5    | <b>Probability distributions</b>   | 21 |
| 1.5.1  | <i>Probability distribution</i>  | 21 |
| 1.5.2  | <i>Joint distribution</i>  | 21 |
| 1.5.3  | <i>Marginal distribution</i>   | 22 |
| 1.5.4  | <i>Conditional distribution</i>  | 22 |
| 1.5.5  | <i>Unimodal vs. multimodal distribution</i>                                      | 22 |
| 1.5.6  | <i>Sampling distribution</i>   | 22 |
| 1.5.7  | <i>Statistical parameter</i>   | 23 |
| 1.5.8  | <i>Location parameter</i>  | 23 |
| 1.5.9  | <i>Scale parameter</i>   | 23 |
| 1.5.10 | <i>Rate parameter</i>  | 24 |
| 1.5.11 | <i>Shape parameter</i>   | 24 |
| 1.5.12 | <i>Degrees of freedom</i>  | 25 |
| 1.6    | <b>Probability mass function</b>   | 25 |
| 1.6.1  | <i>Definition</i>  | 25 |
| 1.6.2  | <b>Probability mass function of sum of independents</b>                          | 25 |
| 1.6.3  | <b>Probability mass function of strictly increasing function</b>                 | 26 |
| 1.6.4  | <b>Probability mass function of strictly decreasing function</b>                 | 27 |
| 1.6.5  | <b>Probability mass function of invertible function</b>                          | 28 |
| 1.7    | <b>Probability density function</b>  | 28 |
| 1.7.1  | <i>Definition</i>  | 28 |
| 1.7.2  | <b>Probability density function of sum of independents</b>                       | 29 |
| 1.7.3  | <b>Probability density function of strictly increasing function</b>              | 29 |
| 1.7.4  | <b>Probability density function of strictly decreasing function</b>              | 30 |
| 1.7.5  | <b>Probability density function of invertible function</b>                       | 32 |
| 1.7.6  | <b>Probability density function of linear transformation</b>                     | 34 |
| 1.7.7  | <b>Probability density function in terms of cumulative distribution function</b> | 34 |
| 1.7.8  | <i>Joint probability density function</i>  | 35 |
| 1.8    | <b>Cumulative distribution function</b>  | 36 |
| 1.8.1  | <i>Definition</i>  | 36 |
| 1.8.2  | <b>Cumulative distribution function of sum of independents</b>                   | 36 |
| 1.8.3  | <b>Cumulative distribution function of strictly increasing function</b>          | 37 |
| 1.8.4  | <b>Cumulative distribution function of strictly decreasing function</b>          | 38 |
| 1.8.5  | <b>Cumulative distribution function of discrete random variable</b>              | 39 |
| 1.8.6  | <b>Cumulative distribution function of continuous random variable</b>            | 39 |

|         |   |    |
|---------|---|----|
| 1.8.7   | <b>Exceedance probability based on cumulative distribution function</b> . . . . . | 40 |
| 1.8.8   | <b>Probability integral transform</b> . . . . .                                   | 41 |
| 1.8.9   | <b>Inverse transformation method</b> . . . . .                                    | 41 |
| 1.8.10  | <b>Distributional transformation</b> . . . . .                                    | 42 |
| 1.8.11  | <i>Joint cumulative distribution function</i> . . . . .                           | 42 |
| 1.9     | <b>Moment-generating function</b> . . . . .                                       | 43 |
| 1.9.1   | <i>Definition</i> . . . . .   | 43 |
| 1.9.2   | <b>Moment-generating function of sum of independents</b> . . . . .                | 43 |
| 1.9.3   | <b>Moment-generating function of arbitrary function</b> . . . . .                 | 44 |
| 1.9.4   | <b>Moment-generating function of linear transformation</b> . . . . .              | 44 |
| 1.9.5   | <b>Moment-generating function of linear combination</b> . . . . .                 | 45 |
| 1.10    | <b>Probability-generating function</b> . . . . .                                  | 46 |
| 1.10.1  | <i>Definition</i> . . . . .   | 46 |
| 1.10.2  | <b>Probability-generating function in terms of expected value</b> . . . . .       | 46 |
| 1.10.3  | <b>Probability-generating function of zero</b> . . . . .                          | 47 |
| 1.10.4  | <b>Probability-generating function of one</b> . . . . .                           | 47 |
| 1.11    | <b>Other probability functions</b> . . . . .                                      | 48 |
| 1.11.1  | <i>Quantile function</i> . . . . .  | 48 |
| 1.11.2  | <b>Quantile function in terms of cumulative distribution function</b> . . . . .   | 48 |
| 1.11.3  | <i>Characteristic function</i> . . . . .  | 49 |
| 1.11.4  | <b>Characteristic function of arbitrary function</b> . . . . .                    | 49 |
| 1.11.5  | <i>Cumulant-generating function</i> . . . . .                                     | 50 |
| 1.12    | <b>Expected value</b> . . . . .   | 50 |
| 1.12.1  | <i>Definition</i> . . . . .   | 50 |
| 1.12.2  | <i>Sample mean</i> . . . . .  | 51 |
| 1.12.3  | <b>Non-negative random variable</b> . . . . .                                     | 51 |
| 1.12.4  | <b>Non-negativity</b> . . . . .   | 52 |
| 1.12.5  | <b>Linearity</b> . . . . .  | 53 |
| 1.12.6  | <b>Monotonicity</b> . . . . .   | 54 |
| 1.12.7  | <b>(Non-)Multiplicativity</b> . . . . .   | 55 |
| 1.12.8  | <b>Law of total expectation</b> . . . . .   | 57 |
| 1.12.9  | <b>Law of the unconscious statistician</b> . . . . .                              | 58 |
| 1.12.10 | <b>Squared expectation of a product</b> . . . . .                                 | 59 |
| 1.12.11 | <b>Jensen's inequality</b> . . . . .  | 61 |
| 1.12.12 | <b>Markov's inequality</b> . . . . .  | 62 |
| 1.12.13 | <b>Chebyshev's inequality</b> . . . . .   | 62 |
| 1.12.14 | <b>Weak law of large numbers</b> . . . . .  | 63 |
| 1.12.15 | <b>Expected value minimizes squared error</b> . . . . .                           | 65 |
| 1.12.16 | <i>Expected value of a random vector</i> . . . . .                                | 66 |
| 1.12.17 | <b>Expectation of a quadratic form</b> . . . . .                                  | 66 |
| 1.12.18 | <b>Expectation of a bilinear form</b> . . . . .                                   | 67 |
| 1.12.19 | <i>Expected value of a random matrix</i> . . . . .                                | 68 |
| 1.12.20 | <b>Expectation of a trace</b> . . . . .   | 68 |
| 1.13    | <b>Variance</b> . . . . .   | 69 |
| 1.13.1  | <i>Definition</i> . . . . .   | 69 |
| 1.13.2  | <i>Sample variance</i> . . . . .  | 69 |
| 1.13.3  | <i>Pooled sample variance</i> . . . . .   | 70 |

|         |   |    |
|---------|---|----|
| 1.13.4  | <b>Partition into expected values</b>           | 70 |
| 1.13.5  | <b>Non-negativity</b>                           | 71 |
| 1.13.6  | <b>Variance of a constant</b>                   | 71 |
| 1.13.7  | <b>Invariance under addition</b>                | 72 |
| 1.13.8  | <b>Scaling upon multiplication</b>              | 73 |
| 1.13.9  | <b>Variance of a sum</b>                        | 73 |
| 1.13.10 | <b>Variance of linear combination</b>           | 74 |
| 1.13.11 | <b>Additivity under independence</b>            | 75 |
| 1.13.12 | <b>Law of total variance</b>                    | 75 |
| 1.13.13 | <i>Precision</i>                                | 76 |
| 1.14    | <b>Skewness</b>                                 | 76 |
| 1.14.1  | <i>Definition</i>                               | 76 |
| 1.14.2  | <i>Sample skewness</i>                          | 77 |
| 1.14.3  | <b>Partition into expected values</b>           | 77 |
| 1.15    | <b>Covariance</b>                               | 78 |
| 1.15.1  | <i>Definition</i>                               | 78 |
| 1.15.2  | <i>Sample covariance</i>                        | 78 |
| 1.15.3  | <b>Partition into expected values</b>           | 78 |
| 1.15.4  | <b>Symmetry</b>                                 | 79 |
| 1.15.5  | <b>Self-covariance</b>                          | 79 |
| 1.15.6  | <b>Covariance under independence</b>            | 80 |
| 1.15.7  | <b>Law of total covariance</b>                  | 80 |
| 1.15.8  | <b>Relationship to correlation</b>              | 81 |
| 1.16    | <b>Covariance matrix</b>                        | 82 |
| 1.16.1  | <i>Definition</i>                               | 82 |
| 1.16.2  | <i>Sample covariance matrix</i>                 | 82 |
| 1.16.3  | <b>Partition into expected values</b>           | 82 |
| 1.16.4  | <b>Symmetry</b>                                 | 83 |
| 1.16.5  | <b>Positive semi-definiteness</b>               | 84 |
| 1.16.6  | <b>Invariance under addition of vector</b>      | 85 |
| 1.16.7  | <b>Scaling upon multiplication with matrix</b>  | 85 |
| 1.16.8  | <b>Covariance matrix and correlation matrix</b> | 86 |
| 1.16.9  | <i>Cross-covariance matrix</i>                  | 87 |
| 1.16.10 | <b>Cross-covariance and expected values</b>     | 88 |
| 1.16.11 | <b>Covariance matrix of a sum</b>               | 88 |
| 1.16.12 | <i>Precision matrix</i>                         | 89 |
| 1.16.13 | <b>Precision matrix and correlation matrix</b>  | 89 |
| 1.17    | <b>Correlation</b>                              | 90 |
| 1.17.1  | <i>Definition</i>                               | 90 |
| 1.17.2  | <b>Range</b>                                    | 91 |
| 1.17.3  | <b>Correlation under independence</b>           | 92 |
| 1.17.4  | <i>Sample correlation coefficient</i>           | 92 |
| 1.17.5  | <b>Relationship to standard scores</b>          | 92 |
| 1.17.6  | <i>Correlation matrix</i>                       | 93 |
| 1.17.7  | <i>Sample correlation matrix</i>                | 93 |
| 1.18    | <b>Measures of central tendency</b>             | 94 |
| 1.18.1  | <i>Median</i>                                   | 94 |
| 1.18.2  | <b>Median minimizes mean absolute error</b>     | 95 |

|      |        |  |     |
|------|--------|--|-----|
|      | 1.18.3 | <i>Mode</i> . . . . .  | 96  |
| 1.19 |        | Measures of statistical dispersion . . . . .                                     | 96  |
|      | 1.19.1 | <i>Standard deviation</i> . . . . .  | 96  |
|      | 1.19.2 | <i>Sample standard deviation</i> . . . . .                                       | 96  |
|      | 1.19.3 | <i>Pooled sample standard deviation</i> . . . . .                                | 97  |
|      | 1.19.4 | <i>Full width at half maximum</i> . . . . .                                      | 97  |
| 1.20 |        | Further summary statistics . . . . .   | 98  |
|      | 1.20.1 | <i>Minimum</i> . . . . .   | 98  |
|      | 1.20.2 | <i>Maximum</i> . . . . .   | 98  |
| 1.21 |        | Further moments . . . . .  | 99  |
|      | 1.21.1 | <i>Moment</i> . . . . .  | 99  |
|      | 1.21.2 | <b>Moment in terms of moment-generating function</b> . . . . .                   | 99  |
|      | 1.21.3 | <i>Raw moment</i> . . . . .  | 100 |
|      | 1.21.4 | <b>First raw moment is mean</b> . . . . .  | 101 |
|      | 1.21.5 | <b>Second raw moment and variance</b> . . . . .                                  | 101 |
|      | 1.21.6 | <i>Central moment</i> . . . . .  | 102 |
|      | 1.21.7 | <b>First central moment is zero</b> . . . . .                                    | 102 |
|      | 1.21.8 | <b>Second central moment is variance</b> . . . . .                               | 102 |
|      | 1.21.9 | <i>Standardized moment</i> . . . . .   | 103 |
| 2    |        | Information theory . . . . .   | 104 |
|      | 2.1    | Shannon entropy . . . . .  | 104 |
|      |        | 2.1.1 <i>Definition</i> . . . . .  | 104 |
|      |        | 2.1.2 <b>Non-negativity</b> . . . . .  | 104 |
|      |        | 2.1.3 <b>Concavity</b> . . . . .   | 105 |
|      |        | 2.1.4 <i>Conditional entropy</i> . . . . .                                       | 106 |
|      |        | 2.1.5 <i>Joint entropy</i> . . . . .   | 106 |
|      |        | 2.1.6 <i>Cross-entropy</i> . . . . .   | 107 |
|      |        | 2.1.7 <b>Convexity of cross-entropy</b> . . . . .                                | 107 |
|      |        | 2.1.8 <b>Gibbs' inequality</b> . . . . .   | 108 |
|      |        | 2.1.9 <b>Log sum inequality</b> . . . . .  | 109 |
|      | 2.2    | Differential entropy . . . . .   | 110 |
|      |        | 2.2.1 <i>Definition</i> . . . . .  | 110 |
|      |        | 2.2.2 <b>Negativity</b> . . . . .  | 110 |
|      |        | 2.2.3 <b>Invariance under addition</b> . . . . .                                 | 111 |
|      |        | 2.2.4 <b>Addition upon multiplication</b> . . . . .                              | 112 |
|      |        | 2.2.5 <b>Addition upon matrix multiplication</b> . . . . .                       | 113 |
|      |        | 2.2.6 <b>Non-invariance and transformation</b> . . . . .                         | 114 |
|      |        | 2.2.7 <i>Conditional differential entropy</i> . . . . .                          | 116 |
|      |        | 2.2.8 <i>Joint differential entropy</i> . . . . .                                | 116 |
|      |        | 2.2.9 <i>Differential cross-entropy</i> . . . . .                                | 116 |
|      | 2.3    | Discrete mutual information . . . . .  | 117 |
|      |        | 2.3.1 <i>Definition</i> . . . . .  | 117 |
|      |        | 2.3.2 <b>Relation to marginal and conditional entropy</b> . . . . .              | 117 |
|      |        | 2.3.3 <b>Relation to marginal and joint entropy</b> . . . . .                    | 118 |
|      |        | 2.3.4 <b>Relation to joint and conditional entropy</b> . . . . .                 | 119 |
|      | 2.4    | Continuous mutual information . . . . .  | 120 |
|      |        | 2.4.1 <i>Definition</i> . . . . .  | 120 |
|      |        | 2.4.2 <b>Relation to marginal and conditional differential entropy</b> . . . . . | 121 |

|     |        |   |     |
|-----|--------|---|-----|
|     | 2.4.3  | Relation to marginal and joint differential entropy . . . . .       | 122 |
|     | 2.4.4  | Relation to joint and conditional differential entropy . . . . .    | 123 |
| 2.5 |        | Kullback-Leibler divergence . . . . .                               | 124 |
|     | 2.5.1  | <i>Definition</i> . . . . .   | 124 |
|     | 2.5.2  | <b>Non-negativity</b> . . . . .                                     | 124 |
|     | 2.5.3  | <b>Non-negativity</b> . . . . .                                     | 125 |
|     | 2.5.4  | <b>Non-negativity</b> . . . . .                                     | 126 |
|     | 2.5.5  | <b>Non-symmetry</b> . . . . .                                       | 127 |
|     | 2.5.6  | <b>Convexity</b> . . . . .  | 128 |
|     | 2.5.7  | <b>Additivity for independent distributions</b> . . . . .           | 129 |
|     | 2.5.8  | <b>Invariance under parameter transformation</b> . . . . .          | 130 |
|     | 2.5.9  | <b>Relation to discrete entropy</b> . . . . .                       | 131 |
|     | 2.5.10 | <b>Relation to differential entropy</b> . . . . .                   | 132 |
| 3   |        | Estimation theory . . . . .   | 133 |
|     | 3.1    | Basic concepts of estimation . . . . .                              | 133 |
|     | 3.1.1  | <i>Estimator</i> . . . . .  | 133 |
|     | 3.1.2  | <i>Biased vs. unbiased</i> . . . . .                                | 133 |
|     | 3.2    | Point estimates . . . . .   | 133 |
|     | 3.2.1  | <i>Mean squared error</i> . . . . .                                 | 133 |
|     | 3.2.2  | <b>Partition of the mean squared error into bias and variance</b> . | 134 |
|     | 3.3    | Interval estimates . . . . .  | 135 |
|     | 3.3.1  | <i>Confidence interval</i> . . . . .                                | 135 |
|     | 3.3.2  | <b>Construction of confidence intervals using Wilks' theorem</b> .  | 135 |
| 4   |        | Frequentist statistics . . . . .                                    | 137 |
|     | 4.1    | Likelihood theory . . . . .   | 137 |
|     | 4.1.1  | <i>Likelihood function</i> . . . . .                                | 137 |
|     | 4.1.2  | <i>Log-likelihood function</i> . . . . .                            | 137 |
|     | 4.1.3  | <i>Maximum likelihood estimation</i> . . . . .                      | 137 |
|     | 4.1.4  | <i>Maximum log-likelihood</i> . . . . .                             | 137 |
|     | 4.1.5  | <b>MLE can be biased</b> . . . . .                                  | 137 |
|     | 4.1.6  | <i>Likelihood ratio</i> . . . . .                                   | 138 |
|     | 4.1.7  | <i>Log-likelihood ratio</i> . . . . .                               | 138 |
|     | 4.1.8  | <i>Method of moments</i> . . . . .                                  | 138 |
|     | 4.2    | Statistical hypotheses . . . . .                                    | 139 |
|     | 4.2.1  | <i>Statistical hypothesis</i> . . . . .                             | 139 |
|     | 4.2.2  | <i>Simple vs. composite</i> . . . . .                               | 139 |
|     | 4.2.3  | <i>Point/exact vs. set/inexact</i> . . . . .                        | 140 |
|     | 4.2.4  | <i>One-tailed vs. two-tailed</i> . . . . .                          | 140 |
|     | 4.3    | Hypothesis testing . . . . .  | 141 |
|     | 4.3.1  | <i>Statistical test</i> . . . . .                                   | 141 |
|     | 4.3.2  | <i>Null hypothesis</i> . . . . .                                    | 141 |
|     | 4.3.3  | <i>Alternative hypothesis</i> . . . . .                             | 142 |
|     | 4.3.4  | <i>One-tailed vs. two-tailed</i> . . . . .                          | 142 |
|     | 4.3.5  | <i>Test statistic</i> . . . . .                                     | 142 |
|     | 4.3.6  | <i>Size of a test</i> . . . . .                                     | 143 |
|     | 4.3.7  | <i>Power of a test</i> . . . . .                                    | 143 |
|     | 4.3.8  | <i>Significance level</i> . . . . .                                 | 143 |
|     | 4.3.9  | <i>Critical value</i> . . . . .                                     | 144 |

|   |                     |  |     |
|---|---------------------|--|-----|
|   | 4.3.10              | <i>p-value</i>   | 144 |
|   | 4.3.11              | <b>Distribution of p-value under null hypothesis</b>                     | 144 |
|   | 4.3.12              | <i>Minimum detectable effect</i>   | 145 |
|   | 4.3.13              | <i>Minimum required sample size</i>                                      | 146 |
| 5 | Bayesian statistics |  | 147 |
|   | 5.1                 | Probabilistic modeling   | 147 |
|   | 5.1.1               | <i>Generative model</i>  | 147 |
|   | 5.1.2               | <i>Likelihood function</i>   | 147 |
|   | 5.1.3               | <i>Prior distribution</i>  | 147 |
|   | 5.1.4               | <i>Prior predictive distribution</i>                                     | 147 |
|   | 5.1.5               | <b>Prior predictive distribution is marginal of joint likelihood</b>     | 148 |
|   | 5.1.6               | <i>Full probability model</i>  | 148 |
|   | 5.1.7               | <i>Joint likelihood</i>  | 148 |
|   | 5.1.8               | <b>Joint likelihood is product of likelihood and prior</b>               | 149 |
|   | 5.1.9               | <i>Posterior distribution</i>  | 149 |
|   | 5.1.10              | <b>Posterior density is proportional to joint likelihood</b>             | 149 |
|   | 5.1.11              | <b>Combined posterior distribution from independent data</b>             | 150 |
|   | 5.1.12              | <i>Posterior predictive distribution</i>                                 | 151 |
|   | 5.1.13              | <b>Posterior predictive distribution is marginal of joint likelihood</b> | 151 |
|   | 5.1.14              | <i>Maximum-a-posteriori estimation</i>                                   | 152 |
|   | 5.1.15              | <i>Marginal likelihood</i>   | 152 |
|   | 5.1.16              | <b>Marginal likelihood is integral of joint likelihood</b>               | 152 |
|   | 5.2                 | Prior distributions  | 153 |
|   | 5.2.1               | <i>Flat vs. hard vs. soft</i>  | 153 |
|   | 5.2.2               | <i>Uniform vs. non-uniform</i>   | 153 |
|   | 5.2.3               | <i>Informative vs. non-informative</i>                                   | 153 |
|   | 5.2.4               | <i>Empirical vs. non-empirical</i>                                       | 154 |
|   | 5.2.5               | <i>Conjugate vs. non-conjugate</i>                                       | 154 |
|   | 5.2.6               | <i>Maximum entropy priors</i>  | 154 |
|   | 5.2.7               | <i>Empirical Bayes priors</i>  | 155 |
|   | 5.2.8               | <i>Reference priors</i>  | 155 |
|   | 5.3                 | Bayesian inference   | 155 |
|   | 5.3.1               | <i>Odds ratios</i>   | 155 |
|   | 5.3.2               | <b>Bayes' theorem</b>  | 156 |
|   | 5.3.3               | <b>Bayes' rule</b>   | 157 |
|   | 5.3.4               | <i>Empirical Bayes</i>   | 157 |
|   | 5.3.5               | <i>Variational Bayes</i>   | 158 |
|   | 5.3.6               | <b>Decomposition of the free energy</b>                                  | 159 |
|   | 5.3.7               | <b>Free energy is lower bound on log model evidence</b>                  | 161 |
| 6 | Machine learning    |  | 162 |
|   | 6.1                 | Scoring rules  | 162 |
|   | 6.1.1               | <i>Scoring rule</i>  | 162 |
|   | 6.1.2               | <i>Proper scoring rule</i>   | 162 |
|   | 6.1.3               | <i>Strictly proper scoring rule</i>                                      | 162 |
|   | 6.1.4               | <i>Log probability scoring rule</i>                                      | 163 |
|   | 6.1.5               | <b>Log probability is strictly proper scoring rule</b>                   | 163 |
|   | 6.1.6               | <i>Brier scoring rule</i>  | 166 |
|   | 6.1.7               | <b>Brier scoring rule is strictly proper scoring rule</b>                | 167 |



|           |   |            |
|-----------|---|------------|
| <b>II</b> | <b>Probability Distributions</b>                                  | <b>171</b> |
| 1         | Univariate discrete distributions . . . . .                       | 172        |
| 1.0.1     | <i>Definition</i> . . . . .                                       | 172        |
| 1.0.2     | <b>Probability mass function</b> . . . . .                        | 172        |
| 1.0.3     | <b>Cumulative distribution function</b> . . . . .                 | 172        |
| 1.0.4     | <b>Quantile function</b> . . . . .                                | 173        |
| 1.0.5     | <b>Shannon entropy</b> . . . . .                                  | 174        |
| 1.0.6     | <b>Kullback-Leibler divergence</b> . . . . .                      | 175        |
| 1.0.7     | <b>Maximum entropy distribution</b> . . . . .                     | 176        |
| 1.1       | Bernoulli distribution . . . . .                                  | 177        |
| 1.1.1     | <i>Definition</i> . . . . .                                       | 177        |
| 1.1.2     | <b>Special case of categorical distribution</b> . . . . .         | 178        |
| 1.1.3     | <b>Probability mass function</b> . . . . .                        | 178        |
| 1.1.4     | <b>Mean</b> . . . . .   | 179        |
| 1.1.5     | <b>Variance</b> . . . . .   | 179        |
| 1.1.6     | <b>Range of variance</b> . . . . .                                | 180        |
| 1.1.7     | <b>Shannon entropy</b> . . . . .                                  | 181        |
| 1.1.8     | <b>Kullback-Leibler divergence</b> . . . . .                      | 181        |
| 1.2       | Binomial distribution . . . . .                                   | 182        |
| 1.2.1     | <i>Definition</i> . . . . .                                       | 182        |
| 1.2.2     | <b>Special case of multinomial distribution</b> . . . . .         | 183        |
| 1.2.3     | <b>Probability mass function</b> . . . . .                        | 183        |
| 1.2.4     | <b>Cumulative distribution function</b> . . . . .                 | 184        |
| 1.2.5     | <b>Probability-generating function</b> . . . . .                  | 185        |
| 1.2.6     | <b>Mean</b> . . . . .   | 186        |
| 1.2.7     | <b>Variance</b> . . . . .   | 186        |
| 1.2.8     | <b>Range of variance</b> . . . . .                                | 187        |
| 1.2.9     | <b>Shannon entropy</b> . . . . .                                  | 188        |
| 1.2.10    | <b>Kullback-Leibler divergence</b> . . . . .                      | 189        |
| 1.2.11    | <b>Conditional binomial</b> . . . . .                             | 190        |
| 1.3       | Beta-binomial distribution . . . . .                              | 192        |
| 1.3.1     | <i>Definition</i> . . . . .                                       | 192        |
| 1.3.2     | <b>Probability mass function</b> . . . . .                        | 192        |
| 1.3.3     | <b>Probability mass function in terms of gamma function</b> . . . | 194        |
| 1.3.4     | <b>Cumulative distribution function</b> . . . . .                 | 194        |
| 1.4       | Poisson distribution . . . . .                                    | 196        |
| 1.4.1     | <i>Definition</i> . . . . .                                       | 196        |
| 1.4.2     | <b>Probability mass function</b> . . . . .                        | 196        |
| 1.4.3     | <b>Mean</b> . . . . .   | 196        |
| 1.4.4     | <b>Variance</b> . . . . .   | 197        |
| 1.4.5     | <b>Shannon entropy</b> . . . . .                                  | 199        |
| 2         | Multivariate discrete distributions . . . . .                     | 201        |
| 2.1       | Categorical distribution . . . . .                                | 201        |
| 2.1.1     | <i>Definition</i> . . . . .                                       | 201        |
| 2.1.2     | <b>Probability mass function</b> . . . . .                        | 201        |
| 2.1.3     | <b>Mean</b> . . . . .   | 201        |
| 2.1.4     | <b>Covariance</b> . . . . .                                       | 202        |
| 2.1.5     | <b>Shannon entropy</b> . . . . .                                  | 202        |

|        |  |     |
|--------|--|-----|
| 2.2    | Multinomial distribution . . . . .   | 203 |
| 2.2.1  | <i>Definition</i> . . . . .  | 203 |
| 2.2.2  | <b>Probability mass function</b> . . . . .                                 | 203 |
| 2.2.3  | <b>Cumulative distribution function</b> . . . . .                          | 204 |
| 2.2.4  | <b>Mean</b> . . . . .  | 205 |
| 2.2.5  | <b>Covariance</b> . . . . .  | 206 |
| 2.2.6  | <b>Shannon entropy</b> . . . . .   | 207 |
| 2.2.7  | <b>Marginal distributions</b> . . . . .                                    | 209 |
| 3      | Univariate continuous distributions . . . . .                              | 210 |
| 3.1    | Continuous uniform distribution . . . . .                                  | 210 |
| 3.1.1  | <i>Definition</i> . . . . .  | 210 |
| 3.1.2  | <i>Standard uniform distribution</i> . . . . .                             | 210 |
| 3.1.3  | <b>Probability density function</b> . . . . .                              | 210 |
| 3.1.4  | <b>Cumulative distribution function</b> . . . . .                          | 211 |
| 3.1.5  | <b>Quantile function</b> . . . . .   | 212 |
| 3.1.6  | <b>Mean</b> . . . . .  | 213 |
| 3.1.7  | <b>Median</b> . . . . .  | 214 |
| 3.1.8  | <b>Mode</b> . . . . .  | 214 |
| 3.1.9  | <b>Variance</b> . . . . .  | 215 |
| 3.1.10 | <b>Differential entropy</b> . . . . .                                      | 216 |
| 3.1.11 | <b>Kullback-Leibler divergence</b> . . . . .                               | 217 |
| 3.1.12 | <b>Maximum entropy distribution</b> . . . . .                              | 218 |
| 3.2    | Normal distribution . . . . .  | 219 |
| 3.2.1  | <i>Definition</i> . . . . .  | 219 |
| 3.2.2  | <b>Special case of multivariate normal distribution</b> . . . . .          | 220 |
| 3.2.3  | <i>Standard normal distribution</i> . . . . .                              | 220 |
| 3.2.4  | <b>Relationship to standard normal distribution</b> . . . . .              | 220 |
| 3.2.5  | <b>Relationship to standard normal distribution</b> . . . . .              | 222 |
| 3.2.6  | <b>Relationship to standard normal distribution</b> . . . . .              | 222 |
| 3.2.7  | <b>Relationship to chi-squared distribution</b> . . . . .                  | 223 |
| 3.2.8  | <b>Relationship to t-distribution</b> . . . . .                            | 225 |
| 3.2.9  | <b>Gaussian integral</b> . . . . .   | 227 |
| 3.2.10 | <b>Probability density function</b> . . . . .                              | 228 |
| 3.2.11 | <b>Moment-generating function</b> . . . . .                                | 229 |
| 3.2.12 | <b>Cumulative distribution function</b> . . . . .                          | 230 |
| 3.2.13 | <b>Cumulative distribution function without error function</b> . . . . .   | 232 |
| 3.2.14 | <b>Probability of being within standard deviations from mean</b> . . . . . | 234 |
| 3.2.15 | <b>Quantile function</b> . . . . .   | 235 |
| 3.2.16 | <b>Mean</b> . . . . .  | 235 |
| 3.2.17 | <b>Median</b> . . . . .  | 237 |
| 3.2.18 | <b>Mode</b> . . . . .  | 237 |
| 3.2.19 | <b>Variance</b> . . . . .  | 238 |
| 3.2.20 | <b>Full width at half maximum</b> . . . . .                                | 240 |
| 3.2.21 | <b>Extreme points</b> . . . . .  | 241 |
| 3.2.22 | <b>Inflection points</b> . . . . .   | 242 |
| 3.2.23 | <b>Differential entropy</b> . . . . .                                      | 243 |
| 3.2.24 | <b>Kullback-Leibler divergence</b> . . . . .                               | 244 |
| 3.2.25 | <b>Maximum entropy distribution</b> . . . . .                              | 246 |

|     |                                    |  |     |
|-----|------------------------------------|--|-----|
|     | 3.2.26                             | Linear combination of independent normals . . . . .        | 247 |
|     | 3.2.27                             | Normal and uncorrelated does not imply independent . . . . | 249 |
|     | 3.2.28                             | Marginally normal does not imply jointly normal . . . . .  | 251 |
| 3.3 | t-distribution . . . . .           |  | 252 |
|     | 3.3.1                              | <i>Definition</i> . . . . .                                | 252 |
|     | 3.3.2                              | Special case of multivariate t-distribution . . . . .      | 252 |
|     | 3.3.3                              | <i>Non-standardized t-distribution</i> . . . . .           | 253 |
|     | 3.3.4                              | Relationship to non-standardized t-distribution . . . . .  | 253 |
|     | 3.3.5                              | Probability density function . . . . .                     | 254 |
| 3.4 | Gamma distribution . . . . .       |  | 256 |
|     | 3.4.1                              | <i>Definition</i> . . . . .                                | 256 |
|     | 3.4.2                              | Special case of Wishart distribution . . . . .             | 257 |
|     | 3.4.3                              | <i>Standard gamma distribution</i> . . . . .               | 257 |
|     | 3.4.4                              | Relationship to standard gamma distribution . . . . .      | 258 |
|     | 3.4.5                              | Relationship to standard gamma distribution . . . . .      | 259 |
|     | 3.4.6                              | Scaling of a gamma random variable . . . . .               | 259 |
|     | 3.4.7                              | Probability density function . . . . .                     | 260 |
|     | 3.4.8                              | Moment-generating function . . . . .                       | 261 |
|     | 3.4.9                              | Cumulative distribution function . . . . .                 | 262 |
|     | 3.4.10                             | Quantile function . . . . .                                | 263 |
|     | 3.4.11                             | Mean . . . . .   | 263 |
|     | 3.4.12                             | Median . . . . .   | 264 |
|     | 3.4.13                             | Mode . . . . .   | 265 |
|     | 3.4.14                             | Variance . . . . .   | 266 |
|     | 3.4.15                             | Logarithmic expectation . . . . .                          | 268 |
|     | 3.4.16                             | Expectation of $x \ln x$ . . . . .                         | 270 |
|     | 3.4.17                             | Differential entropy . . . . .                             | 271 |
|     | 3.4.18                             | Kullback-Leibler divergence . . . . .                      | 272 |
| 3.5 | Exponential distribution . . . . . |  | 273 |
|     | 3.5.1                              | <i>Definition</i> . . . . .                                | 273 |
|     | 3.5.2                              | Special case of gamma distribution . . . . .               | 273 |
|     | 3.5.3                              | Probability density function . . . . .                     | 274 |
|     | 3.5.4                              | Moment-generating function . . . . .                       | 274 |
|     | 3.5.5                              | Cumulative distribution function . . . . .                 | 275 |
|     | 3.5.6                              | Quantile function . . . . .                                | 276 |
|     | 3.5.7                              | Mean . . . . .   | 277 |
|     | 3.5.8                              | Median . . . . .   | 278 |
|     | 3.5.9                              | Mode . . . . .   | 278 |
|     | 3.5.10                             | Variance . . . . .   | 279 |
|     | 3.5.11                             | Skewness . . . . .   | 280 |
| 3.6 | Log-normal distribution . . . . .  |  | 282 |
|     | 3.6.1                              | <i>Definition</i> . . . . .                                | 282 |
|     | 3.6.2                              | Probability density function . . . . .                     | 283 |
|     | 3.6.3                              | Cumulative distribution function . . . . .                 | 284 |
|     | 3.6.4                              | Quantile function . . . . .                                | 285 |
|     | 3.6.5                              | Mean . . . . .   | 286 |
|     | 3.6.6                              | Median . . . . .   | 288 |
|     | 3.6.7                              | Mode . . . . .   | 288 |

|      |        |   |     |
|------|--------|---|-----|
|      | 3.6.8  | <b>Variance</b> . . . . .                                   | 290 |
|      | 3.6.9  | <b>Product of independent log-normals</b> . . . . .         | 292 |
|      | 3.6.10 | <b>Geometric mean of independent log-normals</b> . . . . .  | 293 |
| 3.7  |        | Chi-squared distribution . . . . .                          | 294 |
|      | 3.7.1  | <i>Definition</i> . . . . .                                 | 294 |
|      | 3.7.2  | <b>Special case of gamma distribution</b> . . . . .         | 294 |
|      | 3.7.3  | <b>Probability density function</b> . . . . .               | 295 |
|      | 3.7.4  | <b>Raw moments</b> . . . . .                                | 297 |
| 3.8  |        | F-distribution . . . . .                                    | 298 |
|      | 3.8.1  | <i>Definition</i> . . . . .                                 | 298 |
|      | 3.8.2  | <b>Probability density function</b> . . . . .               | 298 |
| 3.9  |        | Beta distribution . . . . .                                 | 300 |
|      | 3.9.1  | <i>Definition</i> . . . . .                                 | 300 |
|      | 3.9.2  | <b>Special case of Dirichlet distribution</b> . . . . .     | 300 |
|      | 3.9.3  | <b>Relationship to chi-squared distribution</b> . . . . .   | 301 |
|      | 3.9.4  | <b>Relationship to F-distribution</b> . . . . .             | 303 |
|      | 3.9.5  | <b>Probability density function</b> . . . . .               | 305 |
|      | 3.9.6  | <b>Moment-generating function</b> . . . . .                 | 306 |
|      | 3.9.7  | <b>Cumulative distribution function</b> . . . . .           | 307 |
|      | 3.9.8  | <b>Mean</b> . . . . .                                       | 308 |
|      | 3.9.9  | <b>Median</b> . . . . .                                     | 309 |
|      | 3.9.10 | <b>Mode</b> . . . . .                                       | 310 |
|      | 3.9.11 | <b>Variance</b> . . . . .                                   | 313 |
| 3.10 |        | Wald distribution . . . . .                                 | 314 |
|      | 3.10.1 | <i>Definition</i> . . . . .                                 | 314 |
|      | 3.10.2 | <b>Probability density function</b> . . . . .               | 315 |
|      | 3.10.3 | <b>Moment-generating function</b> . . . . .                 | 315 |
|      | 3.10.4 | <b>Mean</b> . . . . .                                       | 317 |
|      | 3.10.5 | <b>Variance</b> . . . . .                                   | 318 |
|      | 3.10.6 | <b>Skewness</b> . . . . .                                   | 319 |
|      | 3.10.7 | <b>Method of moments</b> . . . . .                          | 322 |
| 3.11 |        | ex-Gaussian distribution . . . . .                          | 324 |
|      | 3.11.1 | <i>Definition</i> . . . . .                                 | 324 |
|      | 3.11.2 | <b>Probability density function</b> . . . . .               | 324 |
|      | 3.11.3 | <b>Moment-generating function</b> . . . . .                 | 326 |
|      | 3.11.4 | <b>Mean</b> . . . . .                                       | 327 |
|      | 3.11.5 | <b>Variance</b> . . . . .                                   | 328 |
|      | 3.11.6 | <b>Skewness</b> . . . . .                                   | 329 |
|      | 3.11.7 | <b>Method of moments</b> . . . . .                          | 331 |
| 4    |        | Multivariate continuous distributions . . . . .             | 334 |
| 4.1  |        | Multivariate normal distribution . . . . .                  | 334 |
|      | 4.1.1  | <i>Definition</i> . . . . .                                 | 334 |
|      | 4.1.2  | <b>Special case of matrix-normal distribution</b> . . . . . | 334 |
|      | 4.1.3  | <b>Relationship to chi-squared distribution</b> . . . . .   | 335 |
|      | 4.1.4  | <b>Probability density function</b> . . . . .               | 336 |
|      | 4.1.5  | <b>Moment-generating function</b> . . . . .                 | 336 |
|      | 4.1.6  | <b>Mean</b> . . . . .                                       | 337 |
|      | 4.1.7  | <b>Mode</b> . . . . .                                       | 338 |

|        |  |     |
|--------|--|-----|
| 4.1.8  | Covariance . . . . .   | 340 |
| 4.1.9  | Expectation of a quadratic form . . . . .                                  | 341 |
| 4.1.10 | Differential entropy . . . . .   | 342 |
| 4.1.11 | Mutual information . . . . .   | 343 |
| 4.1.12 | Kullback-Leibler divergence . . . . .                                      | 344 |
| 4.1.13 | Maximum entropy distribution . . . . .                                     | 346 |
| 4.1.14 | Linear transformation . . . . .  | 347 |
| 4.1.15 | Marginal distributions . . . . .   | 348 |
| 4.1.16 | Conditional distributions . . . . .  | 349 |
| 4.1.17 | Conditions for independence . . . . .                                      | 352 |
| 4.1.18 | Independence of products . . . . .   | 354 |
| 4.1.19 | Drawing samples . . . . .  | 355 |
| 4.2    | Bivariate normal distribution . . . . .                                    | 355 |
| 4.2.1  | <i>Definition</i> . . . . .  | 355 |
| 4.2.2  | Probability density function . . . . .                                     | 356 |
| 4.2.3  | Probability density function in terms of correlation coefficient . . . . . | 357 |
| 4.2.4  | Construction from standard normal distributions . . . . .                  | 359 |
| 4.2.5  | Mutual information . . . . .   | 361 |
| 4.2.6  | Linear combination . . . . .   | 363 |
| 4.3    | Multivariate t-distribution . . . . .                                      | 364 |
| 4.3.1  | <i>Definition</i> . . . . .  | 364 |
| 4.3.2  | Probability density function . . . . .                                     | 365 |
| 4.3.3  | Relationship to F-distribution . . . . .                                   | 365 |
| 4.4    | Normal-gamma distribution . . . . .  | 366 |
| 4.4.1  | <i>Definition</i> . . . . .  | 366 |
| 4.4.2  | Special case of normal-Wishart distribution . . . . .                      | 367 |
| 4.4.3  | Probability density function . . . . .                                     | 368 |
| 4.4.4  | Mean . . . . .   | 369 |
| 4.4.5  | Covariance . . . . .   | 371 |
| 4.4.6  | Differential entropy . . . . .   | 372 |
| 4.4.7  | Kullback-Leibler divergence . . . . .                                      | 373 |
| 4.4.8  | Marginal distributions . . . . .   | 375 |
| 4.4.9  | Conditional distributions . . . . .  | 377 |
| 4.4.10 | Drawing samples . . . . .  | 379 |
| 4.5    | Dirichlet distribution . . . . .   | 380 |
| 4.5.1  | <i>Definition</i> . . . . .  | 380 |
| 4.5.2  | Probability density function . . . . .                                     | 380 |
| 4.5.3  | Kullback-Leibler divergence . . . . .                                      | 381 |
| 4.5.4  | Exceedance probabilities . . . . .   | 382 |
| 5      | Matrix-variate continuous distributions . . . . .                          | 386 |
| 5.1    | Matrix-normal distribution . . . . .                                       | 386 |
| 5.1.1  | <i>Definition</i> . . . . .  | 386 |
| 5.1.2  | Equivalence to multivariate normal distribution . . . . .                  | 386 |
| 5.1.3  | Probability density function . . . . .                                     | 387 |
| 5.1.4  | Mean . . . . .   | 388 |
| 5.1.5  | Covariance . . . . .   | 388 |
| 5.1.6  | Differential entropy . . . . .   | 389 |
| 5.1.7  | Kullback-Leibler divergence . . . . .                                      | 390 |

|                               |   |            |
|-------------------------------|---|------------|
| 5.1.8                         | Transposition . . . . .                                   | 391        |
| 5.1.9                         | Linear transformation . . . . .                           | 393        |
| 5.1.10                        | Marginal distributions . . . . .                          | 394        |
| 5.1.11                        | Redundancy of parameters . . . . .                        | 396        |
| 5.1.12                        | Redundancy of parameters . . . . .                        | 397        |
| 5.1.13                        | Drawing samples . . . . .                                 | 397        |
| 5.2                           | Wishart distribution . . . . .                            | 398        |
| 5.2.1                         | <i>Definition</i> . . . . .                               | 398        |
| 5.2.2                         | Kullback-Leibler divergence . . . . .                     | 399        |
| 5.3                           | Normal-Wishart distribution . . . . .                     | 401        |
| 5.3.1                         | <i>Definition</i> . . . . .                               | 401        |
| 5.3.2                         | Probability density function . . . . .                    | 401        |
| 5.3.3                         | Mean . . . . .  | 402        |
| <b>III Statistical Models</b> |   | <b>405</b> |
| 1                             | Univariate normal data . . . . .                          | 406        |
| 1.1                           | Univariate Gaussian . . . . .                             | 406        |
| 1.1.1                         | <i>Definition</i> . . . . .                               | 406        |
| 1.1.2                         | Maximum likelihood estimation . . . . .                   | 406        |
| 1.1.3                         | One-sample t-test . . . . .                               | 408        |
| 1.1.4                         | Two-sample t-test . . . . .                               | 409        |
| 1.1.5                         | Paired t-test . . . . .                                   | 411        |
| 1.1.6                         | F-test for equality of variances . . . . .                | 412        |
| 1.1.7                         | Power analysis for one-sample t-test . . . . .            | 413        |
| 1.1.8                         | Conjugate prior distribution . . . . .                    | 416        |
| 1.1.9                         | Posterior distribution . . . . .                          | 418        |
| 1.1.10                        | Log model evidence . . . . .                              | 421        |
| 1.1.11                        | Accuracy and complexity . . . . .                         | 423        |
| 1.1.12                        | Cross-validated log model evidence . . . . .              | 424        |
| 1.1.13                        | Cross-validated log Bayes factor . . . . .                | 429        |
| 1.2                           | Univariate Gaussian with known variance . . . . .         | 430        |
| 1.2.1                         | <i>Definition</i> . . . . .                               | 430        |
| 1.2.2                         | Maximum likelihood estimation . . . . .                   | 430        |
| 1.2.3                         | One-sample z-test . . . . .                               | 432        |
| 1.2.4                         | Two-sample z-test . . . . .                               | 433        |
| 1.2.5                         | Paired z-test . . . . .                                   | 434        |
| 1.2.6                         | Conjugate prior distribution . . . . .                    | 435        |
| 1.2.7                         | Posterior distribution . . . . .                          | 437        |
| 1.2.8                         | Log model evidence . . . . .                              | 439        |
| 1.2.9                         | Accuracy and complexity . . . . .                         | 441        |
| 1.2.10                        | Log Bayes factor . . . . .                                | 442        |
| 1.2.11                        | Expectation of log Bayes factor . . . . .                 | 443        |
| 1.2.12                        | Cross-validated log model evidence . . . . .              | 445        |
| 1.2.13                        | Cross-validated log Bayes factor . . . . .                | 447        |
| 1.2.14                        | Expectation of cross-validated log Bayes factor . . . . . | 448        |
| 1.3                           | Analysis of variance . . . . .                            | 450        |
| 1.3.1                         | One-way ANOVA . . . . .                                   | 450        |
| 1.3.2                         | Treatment sum of squares . . . . .                        | 450        |

|        |  |     |
|--------|--|-----|
| 1.3.3  | Ordinary least squares for one-way ANOVA . . . . .               | 451 |
| 1.3.4  | Sums of squares in one-way ANOVA . . . . .                       | 452 |
| 1.3.5  | F-test for main effect in one-way ANOVA . . . . .                | 453 |
| 1.3.6  | F-statistic in terms of OLS estimates . . . . .                  | 457 |
| 1.3.7  | Reparametrization of one-way ANOVA . . . . .                     | 458 |
| 1.3.8  | <i>Two-way ANOVA</i> . . . . .                                   | 461 |
| 1.3.9  | <i>Interaction sum of squares</i> . . . . .                      | 462 |
| 1.3.10 | Ordinary least squares for two-way ANOVA . . . . .               | 462 |
| 1.3.11 | Sums of squares in two-way ANOVA . . . . .                       | 466 |
| 1.3.12 | Cochran's theorem for two-way ANOVA . . . . .                    | 468 |
| 1.3.13 | F-test for main effect in two-way ANOVA . . . . .                | 474 |
| 1.3.14 | F-test for interaction in two-way ANOVA . . . . .                | 477 |
| 1.3.15 | F-test for grand mean in two-way ANOVA . . . . .                 | 478 |
| 1.3.16 | F-statistics in terms of OLS estimates . . . . .                 | 480 |
| 1.4    | Simple linear regression . . . . .                               | 481 |
| 1.4.1  | <i>Definition</i> . . . . .                                      | 481 |
| 1.4.2  | Special case of multiple linear regression . . . . .             | 482 |
| 1.4.3  | Ordinary least squares . . . . .                                 | 483 |
| 1.4.4  | Ordinary least squares . . . . .                                 | 485 |
| 1.4.5  | Expectation of estimates . . . . .                               | 487 |
| 1.4.6  | Variance of estimates . . . . .                                  | 489 |
| 1.4.7  | Distribution of estimates . . . . .                              | 491 |
| 1.4.8  | Correlation of estimates . . . . .                               | 494 |
| 1.4.9  | Effects of mean-centering . . . . .                              | 494 |
| 1.4.10 | <i>Regression line</i> . . . . .                                 | 496 |
| 1.4.11 | Regression line includes center of mass . . . . .                | 496 |
| 1.4.12 | Projection of data point to regression line . . . . .            | 497 |
| 1.4.13 | Sums of squares . . . . .  | 498 |
| 1.4.14 | Partition of sums of squares . . . . .                           | 500 |
| 1.4.15 | Transformation matrices . . . . .                                | 503 |
| 1.4.16 | Weighted least squares . . . . .                                 | 505 |
| 1.4.17 | Weighted least squares . . . . .                                 | 507 |
| 1.4.18 | Maximum likelihood estimation . . . . .                          | 508 |
| 1.4.19 | Maximum likelihood estimation . . . . .                          | 510 |
| 1.4.20 | t-test for intercept parameter . . . . .                         | 511 |
| 1.4.21 | t-test for slope parameter . . . . .                             | 513 |
| 1.4.22 | F-test for model comparison . . . . .                            | 515 |
| 1.4.23 | Sum of residuals is zero . . . . .                               | 517 |
| 1.4.24 | Correlation with covariate is zero . . . . .                     | 518 |
| 1.4.25 | Residual variance in terms of sample variance . . . . .          | 519 |
| 1.4.26 | Correlation coefficient in terms of slope estimate . . . . .     | 521 |
| 1.4.27 | Coefficient of determination in terms of correlation coefficient | 521 |
| 1.5    | Multiple linear regression . . . . .                             | 523 |
| 1.5.1  | <i>Definition</i> . . . . .                                      | 523 |
| 1.5.2  | Special case of general linear model . . . . .                   | 524 |
| 1.5.3  | Ordinary least squares . . . . .                                 | 524 |
| 1.5.4  | Ordinary least squares . . . . .                                 | 525 |
| 1.5.5  | Ordinary least squares . . . . .                                 | 526 |

|        |   |     |
|--------|---|-----|
| 1.5.6  | Ordinary least squares for two regressors . . . . .                         | 527 |
| 1.5.7  | <i>Total sum of squares</i> . . . . .                                       | 529 |
| 1.5.8  | <i>Explained sum of squares</i> . . . . .                                   | 529 |
| 1.5.9  | <i>Residual sum of squares</i> . . . . .                                    | 530 |
| 1.5.10 | <b>Total, explained and residual sum of squares</b> . . . . .               | 530 |
| 1.5.11 | <i>Estimation matrix</i> . . . . .  | 532 |
| 1.5.12 | <i>Projection matrix</i> . . . . .  | 532 |
| 1.5.13 | <i>Residual-forming matrix</i> . . . . .                                    | 532 |
| 1.5.14 | <b>Estimation, projection and residual-forming matrix</b> . . . . .         | 532 |
| 1.5.15 | <b>Symmetry of projection and residual-forming matrix</b> . . . . .         | 534 |
| 1.5.16 | <b>Idempotence of projection and residual-forming matrix</b> . . . . .      | 535 |
| 1.5.17 | <b>Independence of estimated parameters and residuals</b> . . . . .         | 535 |
| 1.5.18 | <b>Distribution of OLS estimates, signal and residuals</b> . . . . .        | 537 |
| 1.5.19 | <b>Distribution of WLS estimates, signal and residuals</b> . . . . .        | 538 |
| 1.5.20 | <b>Distribution of residual sum of squares</b> . . . . .                    | 540 |
| 1.5.21 | <b>Weighted least squares</b> . . . . .                                     | 542 |
| 1.5.22 | <b>Weighted least squares</b> . . . . .                                     | 543 |
| 1.5.23 | <b>Maximum likelihood estimation</b> . . . . .                              | 544 |
| 1.5.24 | <b>Maximum log-likelihood</b> . . . . .                                     | 545 |
| 1.5.25 | <b>Log-likelihood ratio</b> . . . . .                                       | 547 |
| 1.5.26 | <i>t-contrast</i> . . . . .   | 548 |
| 1.5.27 | <i>F-contrast</i> . . . . .   | 549 |
| 1.5.28 | <b>Contrast-based t-test</b> . . . . .                                      | 549 |
| 1.5.29 | <b>Contrast-based F-test</b> . . . . .                                      | 552 |
| 1.5.30 | <b>t-test for single regressor</b> . . . . .                                | 554 |
| 1.5.31 | <b>F-test for multiple regressors</b> . . . . .                             | 556 |
| 1.5.32 | <b>Deviance function</b> . . . . .  | 561 |
| 1.5.33 | <b>Akaike information criterion</b> . . . . .                               | 563 |
| 1.5.34 | <b>Bayesian information criterion</b> . . . . .                             | 563 |
| 1.5.35 | <b>Corrected Akaike information criterion</b> . . . . .                     | 564 |
| 1.6    | Bayesian linear regression . . . . .  | 565 |
| 1.6.1  | <b>Conjugate prior distribution</b> . . . . .                               | 565 |
| 1.6.2  | <b>Posterior distribution</b> . . . . .                                     | 567 |
| 1.6.3  | <b>Log model evidence</b> . . . . .   | 569 |
| 1.6.4  | <b>Accuracy and complexity</b> . . . . .                                    | 571 |
| 1.6.5  | <b>Deviance information criterion</b> . . . . .                             | 575 |
| 1.6.6  | <b>Maximum-a-posteriori estimation</b> . . . . .                            | 577 |
| 1.6.7  | <b>Expression of posterior parameters using error terms</b> . . . . .       | 579 |
| 1.6.8  | <b>Posterior probability of alternative hypothesis</b> . . . . .            | 580 |
| 1.6.9  | <b>Posterior credibility region excluding null hypothesis</b> . . . . .     | 582 |
| 1.6.10 | <b>Combined posterior distribution from independent data sets</b> . . . . . | 583 |
| 1.6.11 | <b>Log Bayes factor for comparison of two regression models</b> . . . . .   | 587 |
| 1.7    | Bayesian linear regression with known covariance . . . . .                  | 589 |
| 1.7.1  | <b>Conjugate prior distribution</b> . . . . .                               | 589 |
| 1.7.2  | <b>Posterior distribution</b> . . . . .                                     | 590 |
| 1.7.3  | <b>Log model evidence</b> . . . . .   | 592 |
| 1.7.4  | <b>Accuracy and complexity</b> . . . . .                                    | 595 |
| 2      | Multivariate normal data . . . . .  | 598 |



|        |  |     |
|--------|--|-----|
| 2.1    | Multivariate Gaussian . . . . .  | 598 |
| 2.1.1  | <i>Bivariate normally distributed data</i> . . . . .                   | 598 |
| 2.1.2  | <i>Multivariate normally distributed data</i> . . . . .                | 598 |
| 2.1.3  | <b>Maximum likelihood estimation (<math>p = 2</math>)</b> . . . . .    | 598 |
| 2.1.4  | <b>Maximum likelihood estimation (<math>p &gt; 2</math>)</b> . . . . . | 600 |
| 2.2    | General linear model . . . . .   | 603 |
| 2.2.1  | <i>Definition</i> . . . . .  | 603 |
| 2.2.2  | <b>Ordinary least squares</b> . . . . .                                | 604 |
| 2.2.3  | <b>Weighted least squares</b> . . . . .                                | 604 |
| 2.2.4  | <b>Maximum likelihood estimation</b> . . . . .                         | 605 |
| 2.2.5  | <b>Maximum log-likelihood</b> . . . . .                                | 607 |
| 2.2.6  | <b>Log-likelihood ratio</b> . . . . .                                  | 608 |
| 2.2.7  | <b>Mutual information</b> . . . . .                                    | 610 |
| 2.2.8  | <b>Log-likelihood ratio and estimated mutual information</b> . . .     | 612 |
| 2.3    | Transformed general linear model . . . . .                             | 613 |
| 2.3.1  | <i>Definition</i> . . . . .  | 613 |
| 2.3.2  | <b>Derivation of the distribution</b> . . . . .                        | 613 |
| 2.3.3  | <b>Equivalence of parameter estimates</b> . . . . .                    | 614 |
| 2.4    | Inverse general linear model . . . . .                                 | 615 |
| 2.4.1  | <i>Definition</i> . . . . .  | 615 |
| 2.4.2  | <b>Derivation of the distribution</b> . . . . .                        | 616 |
| 2.4.3  | <b>Best linear unbiased estimator</b> . . . . .                        | 616 |
| 2.4.4  | <b>Equivalence of log-likelihood ratios</b> . . . . .                  | 618 |
| 2.4.5  | <i>Corresponding forward model</i> . . . . .                           | 620 |
| 2.4.6  | <b>Derivation of parameters</b> . . . . .                              | 621 |
| 2.4.7  | <b>Proof of existence</b> . . . . .                                    | 622 |
| 2.5    | Multivariate Bayesian linear regression . . . . .                      | 623 |
| 2.5.1  | <b>Conjugate prior distribution</b> . . . . .                          | 623 |
| 2.5.2  | <b>Posterior distribution</b> . . . . .                                | 624 |
| 2.5.3  | <b>Log model evidence</b> . . . . .                                    | 627 |
| 3      | Count data . . . . .   | 630 |
| 3.1    | Binomial observations . . . . .  | 630 |
| 3.1.1  | <i>Definition</i> . . . . .  | 630 |
| 3.1.2  | <b>Binomial test</b> . . . . .   | 630 |
| 3.1.3  | <b>Maximum likelihood estimation</b> . . . . .                         | 631 |
| 3.1.4  | <b>Maximum log-likelihood</b> . . . . .                                | 632 |
| 3.1.5  | <b>Maximum-a-posteriori estimation</b> . . . . .                       | 633 |
| 3.1.6  | <b>Conjugate prior distribution</b> . . . . .                          | 634 |
| 3.1.7  | <b>Posterior distribution</b> . . . . .                                | 635 |
| 3.1.8  | <b>Log model evidence</b> . . . . .                                    | 636 |
| 3.1.9  | <b>Log Bayes factor</b> . . . . .                                      | 637 |
| 3.1.10 | <b>Posterior probability</b> . . . . .                                 | 639 |
| 3.1.11 | <b>Cross-validated log model evidence</b> . . . . .                    | 640 |
| 3.2    | Multinomial observations . . . . .                                     | 642 |
| 3.2.1  | <i>Definition</i> . . . . .  | 642 |
| 3.2.2  | <b>Multinomial test</b> . . . . .                                      | 642 |
| 3.2.3  | <b>Maximum likelihood estimation</b> . . . . .                         | 644 |
| 3.2.4  | <b>Maximum log-likelihood</b> . . . . .                                | 644 |

|                           |  |            |
|---------------------------|--|------------|
| 3.2.5                     | Maximum-a-posteriori estimation . . . . .                    | 646        |
| 3.2.6                     | Conjugate prior distribution . . . . .                       | 647        |
| 3.2.7                     | Posterior distribution . . . . .                             | 648        |
| 3.2.8                     | Log model evidence . . . . .                                 | 649        |
| 3.2.9                     | Log Bayes factor . . . . .                                   | 651        |
| 3.2.10                    | Posterior probability . . . . .                              | 652        |
| 3.2.11                    | Cross-validated log model evidence . . . . .                 | 654        |
| 3.3                       | Poisson-distributed data . . . . .                           | 657        |
| 3.3.1                     | <i>Definition</i> . . . . .                                  | 657        |
| 3.3.2                     | Maximum likelihood estimation . . . . .                      | 657        |
| 3.3.3                     | Conjugate prior distribution . . . . .                       | 659        |
| 3.3.4                     | Posterior distribution . . . . .                             | 660        |
| 3.3.5                     | Log model evidence . . . . .                                 | 661        |
| 3.4                       | Poisson distribution with exposure values . . . . .          | 663        |
| 3.4.1                     | <i>Definition</i> . . . . .                                  | 663        |
| 3.4.2                     | Maximum likelihood estimation . . . . .                      | 663        |
| 3.4.3                     | Conjugate prior distribution . . . . .                       | 665        |
| 3.4.4                     | Posterior distribution . . . . .                             | 666        |
| 3.4.5                     | Log model evidence . . . . .                                 | 668        |
| 4                         | Frequency data . . . . .                                     | 671        |
| 4.1                       | Beta-distributed data . . . . .                              | 671        |
| 4.1.1                     | <i>Definition</i> . . . . .                                  | 671        |
| 4.1.2                     | Method of moments . . . . .                                  | 671        |
| 4.2                       | Dirichlet-distributed data . . . . .                         | 673        |
| 4.2.1                     | <i>Definition</i> . . . . .                                  | 673        |
| 4.2.2                     | Maximum likelihood estimation . . . . .                      | 673        |
| 4.3                       | Beta-binomial data . . . . .                                 | 676        |
| 4.3.1                     | <i>Definition</i> . . . . .                                  | 676        |
| 4.3.2                     | Method of moments . . . . .                                  | 676        |
| 5                         | Categorical data . . . . .                                   | 679        |
| 5.1                       | Logistic regression . . . . .                                | 679        |
| 5.1.1                     | <i>Definition</i> . . . . .                                  | 679        |
| 5.1.2                     | Probability and log-odds . . . . .                           | 679        |
| 5.1.3                     | Log-odds and probability . . . . .                           | 680        |
| <b>IV Model Selection</b> |  | <b>683</b> |
| 1                         | Goodness-of-fit measures . . . . .                           | 684        |
| 1.0.1                     | <i>Definition</i> . . . . .                                  | 684        |
| 1.0.2                     | Maximum likelihood estimator is biased ( $p = 1$ ) . . . . . | 684        |
| 1.0.3                     | Maximum likelihood estimator is biased ( $p > 1$ ) . . . . . | 686        |
| 1.0.4                     | Construction of unbiased estimator ( $p = 1$ ) . . . . .     | 688        |
| 1.0.5                     | Construction of unbiased estimator ( $p > 1$ ) . . . . .     | 689        |
| 1.1                       | R-squared . . . . .  | 690        |
| 1.1.1                     | <i>Definition</i> . . . . .                                  | 690        |
| 1.1.2                     | Derivation of $R^2$ and adjusted $R^2$ . . . . .             | 690        |
| 1.1.3                     | Relationship to residual variance . . . . .                  | 691        |
| 1.1.4                     | Relationship to maximum log-likelihood . . . . .             | 692        |
| 1.1.5                     | Statistical significance test for $R^2$ . . . . .            | 694        |

|     |        |  |     |
|-----|--------|--|-----|
|     | 1.1.6  | Distribution under null hypothesis . . . . .             | 696 |
|     | 1.1.7  | Mean/mode/median under null hypothesis . . . . .         | 697 |
|     | 1.1.8  | Variance under null hypothesis . . . . .                 | 698 |
| 1.2 |        | F-statistic . . . . .                                    | 699 |
|     | 1.2.1  | <i>Definition</i> . . . . .                              | 699 |
|     | 1.2.2  | Relationship to coefficient of determination . . . . .   | 699 |
|     | 1.2.3  | Relationship to maximum log-likelihood . . . . .         | 701 |
| 1.3 |        | Signal-to-noise ratio . . . . .                          | 703 |
|     | 1.3.1  | <i>Definition</i> . . . . .                              | 703 |
|     | 1.3.2  | Relationship to coefficient of determination . . . . .   | 704 |
|     | 1.3.3  | Relationship to maximum log-likelihood . . . . .         | 705 |
| 2   |        | Classical information criteria . . . . .                 | 706 |
| 2.1 |        | Akaike information criterion . . . . .                   | 706 |
|     | 2.1.1  | <i>Definition</i> . . . . .                              | 706 |
|     | 2.1.2  | <i>Corrected AIC</i> . . . . .                           | 706 |
|     | 2.1.3  | Corrected AIC and uncorrected AIC . . . . .              | 706 |
|     | 2.1.4  | Corrected AIC and maximum log-likelihood . . . . .       | 707 |
| 2.2 |        | Bayesian information criterion . . . . .                 | 708 |
|     | 2.2.1  | <i>Definition</i> . . . . .                              | 708 |
|     | 2.2.2  | <b>Derivation</b> . . . . .                              | 708 |
| 2.3 |        | Deviance information criterion . . . . .                 | 709 |
|     | 2.3.1  | <i>Definition</i> . . . . .                              | 709 |
|     | 2.3.2  | <i>Deviance</i> . . . . .                                | 710 |
| 3   |        | Bayesian model selection . . . . .                       | 711 |
| 3.1 |        | Model evidence . . . . .                                 | 711 |
|     | 3.1.1  | <i>Definition</i> . . . . .                              | 711 |
|     | 3.1.2  | <b>Derivation</b> . . . . .                              | 711 |
|     | 3.1.3  | <i>Log model evidence</i> . . . . .                      | 711 |
|     | 3.1.4  | Derivation of the log model evidence . . . . .           | 712 |
|     | 3.1.5  | Expression using prior and posterior . . . . .           | 712 |
|     | 3.1.6  | Partition into accuracy and complexity . . . . .         | 713 |
|     | 3.1.7  | Subtraction of mean from LMEs . . . . .                  | 714 |
|     | 3.1.8  | <i>Uniform-prior log model evidence</i> . . . . .        | 715 |
|     | 3.1.9  | <i>Cross-validated log model evidence</i> . . . . .      | 715 |
|     | 3.1.10 | <i>Empirical Bayesian log model evidence</i> . . . . .   | 716 |
|     | 3.1.11 | <i>Variational Bayesian log model evidence</i> . . . . . | 717 |
| 3.2 |        | Family evidence . . . . .                                | 717 |
|     | 3.2.1  | <i>Definition</i> . . . . .                              | 717 |
|     | 3.2.2  | <b>Derivation</b> . . . . .                              | 718 |
|     | 3.2.3  | <i>Log family evidence</i> . . . . .                     | 718 |
|     | 3.2.4  | Derivation of the log family evidence . . . . .          | 719 |
|     | 3.2.5  | Calculation from log model evidences . . . . .           | 720 |
|     | 3.2.6  | Approximation of log family evidences . . . . .          | 721 |
| 3.3 |        | Bayes factor . . . . .                                   | 722 |
|     | 3.3.1  | <i>Definition</i> . . . . .                              | 722 |
|     | 3.3.2  | <b>Transitivity</b> . . . . .                            | 723 |
|     | 3.3.3  | Computation using Savage-Dickey density ratio . . . . .  | 724 |
|     | 3.3.4  | Computation using encompassing prior method . . . . .    | 725 |

|          |   |            |
|----------|---|------------|
| 3.3.5    | <i>Encompassing model</i> . . . . .                   | 726        |
| 3.3.6    | <i>Log Bayes factor</i> . . . . .                     | 727        |
| 3.3.7    | <b>Derivation of the log Bayes factor</b> . . . . .   | 727        |
| 3.3.8    | <b>Calculation from log model evidences</b> . . . . . | 728        |
| 3.4      | Posterior model probability . . . . .                 | 728        |
| 3.4.1    | <i>Definition</i> . . . . .                           | 728        |
| 3.4.2    | <b>Derivation</b> . . . . .                           | 729        |
| 3.4.3    | <b>Calculation from Bayes factors</b> . . . . .       | 729        |
| 3.4.4    | <b>Calculation from log Bayes factor</b> . . . . .    | 730        |
| 3.4.5    | <b>Calculation from log model evidences</b> . . . . . | 731        |
| 3.5      | Bayesian model averaging . . . . .                    | 732        |
| 3.5.1    | <i>Definition</i> . . . . .                           | 732        |
| 3.5.2    | <b>Derivation</b> . . . . .                           | 732        |
| 3.5.3    | <b>Calculation from log model evidences</b> . . . . . | 733        |
| <b>V</b> | <b>Appendix</b>                                       | <b>735</b> |
| 1        | Proof by Number . . . . .                             | 736        |
| 2        | Definition by Number . . . . .                        | 766        |
| 3        | Proof by Topic . . . . .                              | 775        |
| 4        | Definition by Topic . . . . .                         | 788        |

# Chapter I

## General Theorems

# 1 Probability theory

## 1.1 Random experiments

### 1.1.1 Random experiment

**Definition:** A random experiment is any repeatable procedure that results in one ( $\rightarrow$  I/1.2.2) out of a well-defined set of possible outcomes.

- The set of possible outcomes is called sample space ( $\rightarrow$  I/1.1.2).
- A set of zero or more outcomes is called a random event ( $\rightarrow$  I/1.2.1).
- A function that maps from events to probabilities is called a probability function ( $\rightarrow$  I/1.5.1).

Together, sample space ( $\rightarrow$  I/1.1.2), event space ( $\rightarrow$  I/1.1.3) and probability function ( $\rightarrow$  I/1.1.4) characterize a random experiment.

**Sources:**

- Wikipedia (2020): “Experiment (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Experiment\\_\(probability\\_theory\)](https://en.wikipedia.org/wiki/Experiment_(probability_theory)).

### 1.1.2 Sample space

**Definition:** Given a random experiment ( $\rightarrow$  I/1.1.1), the set of all possible outcomes from this experiment is called the sample space of the experiment. A sample space is usually denoted as  $\Omega$  and specified using set notation.

**Sources:**

- Wikipedia (2021): “Sample space”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Sample\\_space](https://en.wikipedia.org/wiki/Sample_space).

### 1.1.3 Event space

**Definition:** Given a random experiment ( $\rightarrow$  I/1.1.1), an event space  $\mathcal{E}$  is any set of events, where an event ( $\rightarrow$  I/1.2.1) is any set of zero or more elements from the sample space ( $\rightarrow$  I/1.1.2)  $\Omega$  of this experiment.

**Sources:**

- Wikipedia (2021): “Event (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Event\\_\(probability\\_theory\)](https://en.wikipedia.org/wiki/Event_(probability_theory)).

### 1.1.4 Probability space

**Definition:** Given a random experiment ( $\rightarrow$  I/1.1.1), a probability space  $(\Omega, \mathcal{E}, P)$  is a triple consisting of

- the sample space ( $\rightarrow$  I/1.1.2)  $\Omega$ , i.e. the set of all possible outcomes from this experiment;
- an event space ( $\rightarrow$  I/1.1.3)  $\mathcal{E} \subseteq 2^\Omega$ , i.e. a set of subsets from the sample space, called events ( $\rightarrow$  I/1.2.1);
- a probability measure  $P : \mathcal{E} \rightarrow [0, 1]$ , i.e. a function mapping from the event space ( $\rightarrow$  I/1.1.3) to the real numbers, observing the axioms of probability ( $\rightarrow$  I/1.4.1).

**Sources:**

- Wikipedia (2021): “Probability space”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Probability\\_space#Definition](https://en.wikipedia.org/wiki/Probability_space#Definition).

**1.1.5 Measured data**

**Definition:** Data, also “measured data”, are any set of one or more random variables ( $\rightarrow$  I/1.2.2) – discrete or continuous ( $\rightarrow$  I/1.2.6), univariate or multivariate ( $\rightarrow$  I/1.2.7) – that have been observed, measured or collected in the course of a random experiment ( $\rightarrow$  I/1.1.1).

The random variables that are part of data have probability distributions ( $\rightarrow$  I/1.5.1). Assumptions for these probability distributions are made with generative models ( $\rightarrow$  I/5.1.1). Generative models can form the basis for statistical tests ( $\rightarrow$  I/4.3.1) checking statistical hypotheses ( $\rightarrow$  I/4.2.1).

**1.1.6 Statistical sample**

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) or a set of random variables ( $\rightarrow$  I/1.2.2). Then, a sample, also “statistical sample”, is a set of realizations of this random variable, obtained via observation and measurement ( $\rightarrow$  I/1.1.5) within random experiments ( $\rightarrow$  I/1.1.1):

$$x = \{x_1, \dots, x_n\} . \quad (1)$$

**Sources:**

- Elliott, Graham (2024): “Sampling”; in: *Understanding Statistics*, retrieved on 2024-11-15; URL: <https://econweb.ucsd.edu/~gelliott/Chapter5.html>.

**1.1.7 Sample size**

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a statistical sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2) or a set of random variables ( $\rightarrow$  I/1.2.2)  $X$ . Then, the number of elements  $n$  in this sample is called the “sample size”.

**Sources:**

- Wikipedia (2024): “Sampling (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-11-22; URL: [https://en.wikipedia.org/wiki/Sampling\\_\(statistics\)](https://en.wikipedia.org/wiki/Sampling_(statistics)).

**1.1.8 Sample statistic**

**Definition:** A statistic, also “sample statistic”, is any quantity calculated from the data points ( $\rightarrow$  I/1.1.5) in a sample ( $\rightarrow$  I/1.1.6) that is used for statistical purposes.

Examples include statistics used to estimate the parameters ( $\rightarrow$  I/1.5.7) of a probability distribution ( $\rightarrow$  I/1.5.1) and test statistics ( $\rightarrow$  I/4.3.5) to evaluate statistical hypotheses ( $\rightarrow$  I/4.2.1).

**Sources:**

- Wikipedia (2024): “Statistic”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-04; URL: <https://en.wikipedia.org/wiki/Statistic>.

### 1.1.9 Descriptive vs. inferential

**Definition:** A statistic ( $\rightarrow$  I/1.1.8) may be called

- a descriptive statistic, if it quantitatively describes or summarizes a statistical sample ( $\rightarrow$  I/1.1.6) or estimates ( $\rightarrow$  I/3.1.1) aspects of the probability distribution ( $\rightarrow$  I/1.5.1) from which the sample is drawn, e.g. sample mean ( $\rightarrow$  I/1.12.2);
- an inferential statistic, if its sampling distribution ( $\rightarrow$  I/1.5.6) under the null hypothesis ( $\rightarrow$  I/4.3.2) can be used for testing hypotheses ( $\rightarrow$  I/4.3.1) about parameters ( $\rightarrow$  I/1.5.7) of the distribution ( $\rightarrow$  I/1.5.1) from which the sample is drawn, e.g. F-statistic ( $\rightarrow$  IV/1.2.1).

**Sources:**

- Wikipedia (2025): “Descriptive statistics”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-04-04; URL: [https://en.wikipedia.org/wiki/Descriptive\\_statistics](https://en.wikipedia.org/wiki/Descriptive_statistics).
- Wikipedia (2025): “Statistical inference”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-04-04; URL: [https://en.wikipedia.org/wiki/Statistical\\_inference](https://en.wikipedia.org/wiki/Statistical_inference).

## 1.2 Random variables

### 1.2.1 Random event

**Definition:** A random event  $E$  is the outcome of a random experiment ( $\rightarrow$  I/1.1.1) which can be described by a statement that is either true or false.

- If the statement is true, the event is said to take place, denoted as  $E$ .
- If the statement is false, the complement of  $E$  occurs, denoted as  $\overline{E}$ .

In other words, a random event is a random variable ( $\rightarrow$  I/1.2.2) with two possible values (true and false, or 1 and 0). A random experiment ( $\rightarrow$  I/1.1.1) with two possible outcomes is called a Bernoulli trial ( $\rightarrow$  II/1.1.1).

**Sources:**

- Wikipedia (2020): “Event (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Event\\_\(probability\\_theory\)](https://en.wikipedia.org/wiki/Event_(probability_theory)).

### 1.2.2 Random variable

**Definition:** A random variable may be understood

- informally, as a real number  $X \in \mathbb{R}$  whose value is the outcome of a random experiment ( $\rightarrow$  I/1.1.1);
- formally, as a measurable function  $X$  defined on a probability space ( $\rightarrow$  I/1.1.4)  $(\Omega, \mathcal{E}, P)$  that maps from a sample space ( $\rightarrow$  I/1.1.2)  $\Omega$  to the real numbers  $\mathbb{R}$  using an event space ( $\rightarrow$  I/1.1.3)  $\mathcal{E}$  and a probability function ( $\rightarrow$  I/1.5.1)  $P$ ;
- more broadly, as any random quantity  $X$  such as a random event ( $\rightarrow$  I/1.2.1), a random scalar ( $\rightarrow$  I/1.2.2), a random vector ( $\rightarrow$  I/1.2.3) or a random matrix ( $\rightarrow$  I/1.2.4).

**Sources:**

- Wikipedia (2020): “Random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Random\\_variable#Definition](https://en.wikipedia.org/wiki/Random_variable#Definition).



### 1.2.3 Random vector

**Definition:** A random vector, also called “multivariate random variable”, is an  $n$ -dimensional column vector  $X \in \mathbb{R}^{n \times 1}$  whose entries are random variables ( $\rightarrow$  I/1.2.2).

**Sources:**

- Wikipedia (2020): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Multivariate\\_random\\_variable](https://en.wikipedia.org/wiki/Multivariate_random_variable).

### 1.2.4 Random matrix

**Definition:** A random matrix, also called “matrix-valued random variable”, is an  $n \times p$  matrix  $X \in \mathbb{R}^{n \times p}$  whose entries are random variables ( $\rightarrow$  I/1.2.2). Equivalently, a random matrix is an  $n \times p$  matrix whose columns are  $n$ -dimensional random vectors ( $\rightarrow$  I/1.2.3).

**Sources:**

- Wikipedia (2020): “Random matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Random\\_matrix](https://en.wikipedia.org/wiki/Random_matrix).

### 1.2.5 Constant

**Definition:** A constant is a quantity which does not change and thus always has the same value. From a statistical perspective, a constant is a random variable ( $\rightarrow$  I/1.2.2) which is equal to its expected value ( $\rightarrow$  I/1.12.1)

$$X = E(X) \quad (1)$$

or equivalently, whose variance is zero ( $\rightarrow$  I/1.13.6)

$$\text{Var}(X) = 0. \quad (2)$$

**Sources:**

- ProofWiki (2020): “Definition: Constant”; in: *ProofWiki*, retrieved on 2020-09-09; URL: <https://proofwiki.org/wiki/Definition:Constant#Definition>.

### 1.2.6 Discrete vs. continuous

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$ . Then,

- $X$  is called a discrete random variable, if  $\mathcal{X}$  is either a finite set or a countably infinite set; in this case,  $X$  can be described by a probability mass function ( $\rightarrow$  I/1.6.1);
- $X$  is called a continuous random variable, if  $\mathcal{X}$  is an uncountably infinite set; if it is absolutely continuous,  $X$  can be described by a probability density function ( $\rightarrow$  I/1.7.1).

**Sources:**

- Wikipedia (2020): “Random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-29; URL: [https://en.wikipedia.org/wiki/Random\\_variable#Standard\\_case](https://en.wikipedia.org/wiki/Random_variable#Standard_case).

### 1.2.7 Univariate vs. multivariate

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$ . Then,

- $X$  is called a two-valued random variable or random event ( $\rightarrow$  I/1.2.1), if  $\mathcal{X}$  has exactly two elements, e.g.  $\mathcal{X} = \{E, \bar{E}\}$  or  $\mathcal{X} = \{\text{true}, \text{false}\}$  or  $\mathcal{X} = \{1, 0\}$ ;
- $X$  is called a univariate random variable or random scalar ( $\rightarrow$  I/1.2.2), if  $\mathcal{X}$  is one-dimensional, i.e. (a subset of) the real numbers  $\mathbb{R}$ ;
- $X$  is called a multivariate random variable or random vector ( $\rightarrow$  I/1.2.3), if  $\mathcal{X}$  is multi-dimensional, e.g. (a subset of) the  $n$ -dimensional Euclidean space  $\mathbb{R}^n$ ;
- $X$  is called a matrix-valued random variable or random matrix ( $\rightarrow$  I/1.2.4), if  $\mathcal{X}$  is (a subset of) the set of  $n \times p$  real matrices  $\mathbb{R}^{n \times p}$ .

**Sources:**

- Wikipedia (2020): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-06; URL: [https://en.wikipedia.org/wiki/Multivariate\\_random\\_variable](https://en.wikipedia.org/wiki/Multivariate_random_variable).

### 1.2.8 independent and identically distributed

**Definition:** Let  $X_i$  for  $i = 1, \dots, n$  be random variables ( $\rightarrow$  I/1.2.2). Then,  $X_1, \dots, X_n$  are called independent and identically distributed (i.i.d.), if (i) they are statistically independent ( $\rightarrow$  I/1.3.6) and (ii) they follow the same probability distribution ( $\rightarrow$  I/1.5.1)  $\mathcal{D}$  with the same parameters  $\theta$ :

$$X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}(\theta), \quad i = 1, \dots, n. \quad (1)$$

Often, especially in linear regression models, error terms ( $\rightarrow$  III/1.5.1) are independent and identically distributed according to a normal distribution ( $\rightarrow$  II/3.2.1) with mean zero and unknown variance:

$$\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n. \quad (2)$$

**Sources:**

- Wikipedia (2024): “Independent and identically distributed random variables”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-08-08; URL: [https://en.wikipedia.org/wiki/Independent\\_and\\_identically\\_distributed\\_random\\_variables#Introduction](https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables#Introduction).

## 1.3 Probability

### 1.3.1 Probability

**Definition:** Let  $E$  be a statement about an arbitrary event such as the outcome of a random experiment ( $\rightarrow$  I/1.1.1). Then,  $p(E)$  is called the probability of  $E$  and may be interpreted as

- (objectivist interpretation of probability:) some physical state of affairs, e.g. the relative frequency of occurrence of  $E$ , when repeating the experiment (“Frequentist probability”); or
- (subjectivist interpretation of probability:) a degree of belief in  $E$ , e.g. the price at which someone would buy or sell a bet that pays 1 unit of utility if  $E$  and 0 if not  $E$  (“Bayesian probability”).

**Sources:**

- Wikipedia (2020): “Probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: <https://en.wikipedia.org/wiki/Probability#Interpretations>.

### 1.3.2 Joint probability

**Definition:** Let  $A$  and  $B$  be two arbitrary statements about random variables ( $\rightarrow$  I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then,  $p(A, B)$  is called the joint probability of  $A$  and  $B$  and is defined as the probability ( $\rightarrow$  I/1.3.1) that  $A$  and  $B$  are both true.

**Sources:**

- Wikipedia (2020): “Joint probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: [https://en.wikipedia.org/wiki/Joint\\_probability\\_distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution).
- Jason Browlee (2019): “A Gentle Introduction to Joint, Marginal, and Conditional Probability”; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: <https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/>.

### 1.3.3 Marginal probability

**Definition:** (law of marginal probability, also called “sum rule”) Let  $A$  and  $X$  be two arbitrary statements about random variables ( $\rightarrow$  I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability ( $\rightarrow$  I/1.3.2) distribution  $p(A, X)$ . Then,  $p(A)$  is called the marginal probability of  $A$  and, 1) if  $X$  is a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with domain  $\mathcal{X}$ , is given by

$$p(A) = \sum_{x \in \mathcal{X}} p(A, x) ; \quad (1)$$

2) if  $X$  is a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with domain  $\mathcal{X}$ , is given by

$$p(A) = \int_{\mathcal{X}} p(A, x) dx . \quad (2)$$

The law of marginal probability can be motivated from the law of total probability ( $\rightarrow$  I/1.4.11) which follows from the Kolmogorov axioms of probability ( $\rightarrow$  I/1.4.1).

**Sources:**

- Wikipedia (2020): “Marginal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: [https://en.wikipedia.org/wiki/Marginal\\_distribution#Definition](https://en.wikipedia.org/wiki/Marginal_distribution#Definition).
- Jason Browlee (2019): “A Gentle Introduction to Joint, Marginal, and Conditional Probability”; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: <https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/>.

### 1.3.4 Conditional probability

**Definition:** (law of conditional probability, also called “product rule”) Let  $A$  and  $B$  be two arbitrary statements about random variables ( $\rightarrow$  I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability ( $\rightarrow$  I/1.3.2) distribution  $p(A, B)$ . Then,  $p(A|B)$  is called the conditional probability that  $A$  is true, given that  $B$  is true, and is given by

$$p(A|B) = \frac{p(A, B)}{p(B)} \quad (1)$$

where  $p(B)$  is the marginal probability ( $\rightarrow$  I/1.3.3) of  $B$ .

**Sources:**

- Wikipedia (2020): “Conditional probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: [https://en.wikipedia.org/wiki/Conditional\\_probability#Definition](https://en.wikipedia.org/wiki/Conditional_probability#Definition).
- Jason Browlee (2019): “A Gentle Introduction to Joint, Marginal, and Conditional Probability”; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: <https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/>.

### 1.3.5 Exceedance probability

**Definition:** Let  $X = \{X_1, \dots, X_n\}$  be a set of  $n$  random variables ( $\rightarrow$  I/1.2.2) which the joint probability distribution ( $\rightarrow$  I/1.5.2)  $p(X) = p(X_1, \dots, X_n)$ . Then, the exceedance probability for random variable  $X_i$  is the probability ( $\rightarrow$  I/1.3.1) that  $X_i$  is larger than all other random variables  $X_j$ ,  $j \neq i$ :

$$\begin{aligned} \varphi(X_i) &= \Pr(\forall j \in \{1, \dots, n | j \neq i\} : X_i > X_j) \\ &= \Pr\left(\bigwedge_{j \neq i} X_i > X_j\right) \\ &= \Pr(X_i = \max(\{X_1, \dots, X_n\})) \\ &= \int_{X_i = \max(X)} p(X) dX . \end{aligned} \tag{1}$$

**Sources:**

- Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009): “Bayesian model selection for group studies”; in: *NeuroImage*, vol. 46, pp. 1004–1017, eq. 16; URL: <https://www.sciencedirect.com/science/article/abs/pii/S1053811909002638>; DOI: 10.1016/j.neuroimage.2009.03.025.
- Soch J, Allefeld C (2016): “Exceedance Probabilities for the Dirichlet Distribution”; in: *arXiv stat.AP*, 1611.01439; URL: <https://arxiv.org/abs/1611.01439>.

### 1.3.6 Statistical independence

**Definition:** Generally speaking, random variables ( $\rightarrow$  I/1.2.2) are statistically independent, if their joint probability ( $\rightarrow$  I/1.3.2) can be expressed in terms of their marginal probabilities ( $\rightarrow$  I/1.3.3).

1) A set of discrete random variables ( $\rightarrow$  I/1.2.2)  $X_1, \dots, X_n$  with possible values  $\mathcal{X}_1, \dots, \mathcal{X}_n$  is called statistically independent, if

$$p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p(X_i = x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \tag{1}$$

where  $p(x_1, \dots, x_n)$  are the joint probabilities ( $\rightarrow$  I/1.3.2) of  $X_1, \dots, X_n$  and  $p(x_i)$  are the marginal probabilities ( $\rightarrow$  I/1.3.3) of  $X_i$ .

2) A set of continuous random variables ( $\rightarrow$  I/1.2.2)  $X_1, \dots, X_n$  defined on the domains  $\mathcal{X}_1, \dots, \mathcal{X}_n$  is called statistically independent, if

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \quad (2)$$

or equivalently, if the probability densities ( $\rightarrow$  I/1.7.1) exist, if

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \quad (3)$$

where  $F$  are the joint ( $\rightarrow$  I/1.5.2) or marginal ( $\rightarrow$  I/1.5.3) cumulative distribution functions ( $\rightarrow$  I/1.8.1) and  $f$  are the respective probability density functions ( $\rightarrow$  I/1.7.1).

#### Sources:

- Wikipedia (2020): “Independence (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Independence\\_\(probability\\_theory\)#Definition](https://en.wikipedia.org/wiki/Independence_(probability_theory)#Definition).

### 1.3.7 Conditional independence

**Definition:** Generally speaking, random variables ( $\rightarrow$  I/1.2.2) are conditionally independent given another random variable, if they are statistically independent ( $\rightarrow$  I/1.3.6) in their conditional probability distributions ( $\rightarrow$  I/1.5.4) given this random variable.

1) A set of discrete random variables ( $\rightarrow$  I/1.2.6)  $X_1, \dots, X_n$  with possible values  $\mathcal{X}_1, \dots, \mathcal{X}_n$  is called conditionally independent given the random variable  $Y$  with possible values  $\mathcal{Y}$ , if

$$p(X_1 = x_1, \dots, X_n = x_n | Y = y) = \prod_{i=1}^n p(X_i = x_i | Y = y) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \quad (1)$$

where  $p(x_1, \dots, x_n | y)$  are the joint (conditional) probabilities ( $\rightarrow$  I/1.3.2) of  $X_1, \dots, X_n$  given  $Y$  and  $p(x_i | y)$  are the marginal (conditional) probabilities ( $\rightarrow$  I/1.3.3) of  $X_i$  given  $Y$ .

2) A set of continuous random variables ( $\rightarrow$  I/1.2.6)  $X_1, \dots, X_n$  with possible values  $\mathcal{X}_1, \dots, \mathcal{X}_n$  is called conditionally independent given the random variable  $Y$  with possible values  $\mathcal{Y}$ , if

$$F_{X_1, \dots, X_n | Y=y}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i | Y=y}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \quad (2)$$

or equivalently, if the probability densities ( $\rightarrow$  I/1.7.1) exist, if

$$f_{X_1, \dots, X_n | Y=y}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i | Y=y}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \quad (3)$$

where  $F$  are the joint (conditional) ( $\rightarrow$  I/1.5.2) or marginal (conditional) ( $\rightarrow$  I/1.5.3) cumulative distribution functions ( $\rightarrow$  I/1.8.1) and  $f$  are the respective probability density functions ( $\rightarrow$  I/1.7.1).

#### Sources:

- Wikipedia (2020): “Conditional independence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Conditional\\_independence#Conditional\\_independence\\_of\\_random\\_variables](https://en.wikipedia.org/wiki/Conditional_independence#Conditional_independence_of_random_variables).

### 1.3.8 Self-independence

**Theorem:** Let  $E$  be a random event ( $\rightarrow$  I/1.2.1). Then,  $E$  is independent of itself ( $\rightarrow$  I/1.3.6), if and only if its probability ( $\rightarrow$  I/1.3.1) is zero or one:

$$E \text{ self-independent} \Leftrightarrow p(E) = 0 \quad \text{or} \quad p(E) = 1 . \quad (1)$$

**Proof:** According to the definition of statistical independence ( $\rightarrow$  I/1.3.6), it must hold that:

$$\begin{aligned} p(E, E) &= p(E) \cdot p(E) \\ p(E) &= (p(E))^2 . \end{aligned} \quad (2)$$

For  $0 \leq p(E) \leq 1$ , this is only fulfilled, if

$$p(E) = 0 \quad \text{or} \quad p(E) = 1 . \quad (3)$$

Both is possible, since the lower bound of probability is zero ( $\rightarrow$  I/1.4.1) and the upper bound of probability is one ( $\rightarrow$  I/1.4.7). ■

#### Sources:

- Wikipedia (2024): “Independence (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-20; URL: [https://en.wikipedia.org/wiki/Independence\\_\(probability\\_theory\)#Self-independence](https://en.wikipedia.org/wiki/Independence_(probability_theory)#Self-independence).
- Soch, Joram (2023): “Suppose A is an event. Can A be independent of itself?”; in: *X*, Aug 7, 2023, 03:59 PM; URL: <https://x.com/JoramSoch/status/1688550557034651648>.

### 1.3.9 Probability under independence

**Theorem:** Let  $A$  and  $B$  be two statements about random variables ( $\rightarrow$  I/1.2.2). Then, if  $A$  and  $B$  are independent ( $\rightarrow$  I/1.3.6), marginal ( $\rightarrow$  I/1.3.3) and conditional ( $\rightarrow$  I/1.3.4) probabilities are equal:

$$\begin{aligned} p(A) &= p(A|B) \\ p(B) &= p(B|A) . \end{aligned} \quad (1)$$

**Proof:** If  $A$  and  $B$  are independent ( $\rightarrow$  I/1.3.6), then the joint probability ( $\rightarrow$  I/1.3.2) is equal to the product of the marginal probabilities ( $\rightarrow$  I/1.3.3):

$$p(A, B) = p(A) \cdot p(B) . \quad (2)$$

The law of conditional probability ( $\rightarrow$  I/1.3.4) states that

$$p(A|B) = \frac{p(A, B)}{p(B)} . \quad (3)$$

Combining (2) and (3), we have:

$$p(A|B) = \frac{p(A) \cdot p(B)}{p(B)} = p(A) . \quad (4)$$

Equivalently, we can write:

$$p(B|A) \stackrel{(3)}{=} \frac{p(A, B)}{p(A)} \stackrel{(2)}{=} \frac{p(A) \cdot p(B)}{p(A)} = p(B) . \quad (5)$$

■

#### Sources:

- Wikipedia (2021): “Independence (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: [https://en.wikipedia.org/wiki/Independence\\_\(probability\\_theory\)#Definition](https://en.wikipedia.org/wiki/Independence_(probability_theory)#Definition).

### 1.3.10 Mutual exclusivity

**Definition:** Generally speaking, random events ( $\rightarrow$  I/1.2.1) are mutually exclusive, if they cannot occur together, such that their intersection is equal to the empty set ( $\rightarrow$  I/1.4.4).

More precisely, a set of statements  $A_1, \dots, A_n$  is called mutually exclusive, if

$$p(A_1, \dots, A_n) = 0 \quad (1)$$

where  $p(A_1, \dots, A_n)$  is the joint probability ( $\rightarrow$  I/1.3.2) of the statements  $A_1, \dots, A_n$ .

#### Sources:

- Wikipedia (2021): “Mutual exclusivity”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: [https://en.wikipedia.org/wiki/Mutual\\_exclusivity#Probability](https://en.wikipedia.org/wiki/Mutual_exclusivity#Probability).

### 1.3.11 Probability under exclusivity

**Theorem:** Let  $A$  and  $B$  be two statements about random variables ( $\rightarrow$  I/1.2.2). Then, if  $A$  and  $B$  are mutually exclusive ( $\rightarrow$  I/1.3.10), the probability ( $\rightarrow$  I/1.3.1) of their disjunction is equal to the sum of the marginal probabilities ( $\rightarrow$  I/1.3.3):

$$p(A \vee B) = p(A) + p(B) . \quad (1)$$

**Proof:** If  $A$  and  $B$  are mutually exclusive ( $\rightarrow$  I/1.3.10), then their joint probability ( $\rightarrow$  I/1.3.2) is zero:

$$p(A, B) = 0 . \quad (2)$$

The addition law of probability states that

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \quad (3)$$

which, in logical rather than set-theoretic expression, becomes

$$p(A \vee B) = p(A) + p(B) - p(A, B) . \quad (4)$$

Because the union of mutually exclusive events is the empty set ( $\rightarrow$  I/1.3.10) and the probability of the empty set is zero ( $\rightarrow$  I/1.4.4), the joint probability ( $\rightarrow$  I/1.3.2) term cancels out:

$$p(A \vee B) = p(A) + p(B) - p(A, B) \stackrel{(2)}{=} p(A) + p(B) . \quad (5)$$

■

#### Sources:

- Wikipedia (2021): “Mutual exclusivity”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: [https://en.wikipedia.org/wiki/Mutual\\_exclusivity#Probability](https://en.wikipedia.org/wiki/Mutual_exclusivity#Probability).

## 1.4 Probability axioms

### 1.4.1 Axioms of probability

**Definition:** Let there be a sample space ( $\rightarrow$  I/1.1.2)  $\Omega$ , an event space ( $\rightarrow$  I/1.1.3)  $\mathcal{E}$  and a probability measure  $P$ , such that  $p(E)$  is the probability ( $\rightarrow$  I/1.3.1) of some event ( $\rightarrow$  I/1.2.1)  $E \in \mathcal{E}$ . Then, we introduce three axioms of probability:

- First axiom: The probability of an event is a non-negative real number:

$$p(E) \in \mathbb{R}, p(E) \geq 0, \text{ for all } E \in \mathcal{E} . \quad (1)$$

- Second axiom: The probability that at least one elementary event in the sample space will occur is one:

$$p(\Omega) = 1 . \quad (2)$$

- Third axiom: The probability of any countable sequence of disjoint (i.e. mutually exclusive ( $\rightarrow$  I/1.3.10)) events  $E_1, E_2, E_3, \dots$  is equal to the sum of the probabilities of the individual events:

$$p\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} p(E_i) . \quad (3)$$

#### Sources:

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 2; URL: <https://archive.org/details/foundationsofthe00kolm/page/2/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eqs. 8.2-8.4; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549>.
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#Axioms](https://en.wikipedia.org/wiki/Probability_axioms#Axioms).

### 1.4.2 Monotonicity of probability

**Theorem:** Probability ( $\rightarrow$  I/1.3.1) is monotonic, i.e. if  $A$  is a subset of or equal to  $B$ , then the probability of  $A$  is smaller than or equal to  $B$ :

$$A \subseteq B \quad \Rightarrow \quad p(A) \leq p(B) . \quad (1)$$



**Proof:** Set  $E_1 = A$ ,  $E_2 = B \setminus A$  and  $E_i = \emptyset$  for  $i \geq 3$ . Then, the sets  $E_i$  are pairwise disjoint and  $E_1 \cup E_2 \cup \dots = B$ , because  $A \subseteq B$ . Thus, from the third axiom of probability ( $\rightarrow$  I/1.4.1), we have:

$$p(B) = p(A) + p(B \setminus A) + \sum_{i=3}^{\infty} p(E_i) . \quad (2)$$

Since, by the first axiom of probability ( $\rightarrow$  I/1.4.1), the right-hand side is a series of non-negative numbers converging to  $p(B)$  on the left-hand side, it follows that

$$p(A) \leq p(B) . \quad (3)$$

#### Sources:

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 6; URL: <https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9>
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#Monotonicity](https://en.wikipedia.org/wiki/Probability_axioms#Monotonicity).

### 1.4.3 Monotonicity of probability

**Theorem:** Probability ( $\rightarrow$  I/1.3.1) is monotonic, i.e. if  $A$  is a subset of or equal to  $B$ , then the probability of  $A$  is smaller than or equal to  $B$ :

$$A \subseteq B \quad \Rightarrow \quad p(A) \leq p(B) . \quad (1)$$

**Proof:** When  $A \subseteq B$ , then  $B$  is equal to the union of  $A$  and the intersection of  $B$  with the complement ( $\rightarrow$  I/1.4.6) of  $A$ :

$$B = A \cup (B \cap \overline{A}) . \quad (2)$$

Moreover, the intersection of  $A$  and the intersection of  $B$  with the complement of  $A$  is equal to the empty set:

$$A \cap (B \cap \overline{A}) = \emptyset . \quad (3)$$

Thus, the third axiom of probability ( $\rightarrow$  I/1.4.1) implies:

$$\begin{aligned} p(B) &= p(A) + p(B \cap \overline{A}) \\ p(A) &= p(B) - p(B \cap \overline{A}) . \end{aligned} \quad (4)$$

Since  $p(B \cap \overline{A}) \geq 0$  by the first axiom of probability ( $\rightarrow$  I/1.4.1), it must hold that  $p(A) \leq p(B)$ .

#### Sources:

- Ostwald, Dirk (2023): “Elementare Wahrscheinlichkeiten”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (2), Folien 8-10; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/2\\_Elementare\\_Wahrscheinlichkeiten.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/2_Elementare_Wahrscheinlichkeiten.pdf).

#### 1.4.4 Probability of the empty set

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of the empty set is zero:

$$p(\emptyset) = 0 . \quad (1)$$

**Proof:** Let  $A$  and  $B$  be two events fulfilling  $A \subseteq B$ . Set  $E_1 = A$ ,  $E_2 = B \setminus A$  and  $E_i = \emptyset$  for  $i \geq 3$ . Then, the sets  $E_i$  are pairwise disjoint and  $E_1 \cup E_2 \cup \dots = B$ . Thus, from the third axiom of probability ( $\rightarrow$  I/1.4.1), we have:

$$p(B) = p(A) + p(B \setminus A) + \sum_{i=3}^{\infty} p(E_i) . \quad (2)$$

Assume that the probability of the empty set is not zero, i.e.  $p(\emptyset) > 0$ . Then, the right-hand side of (2) would be infinite. However, by the first axiom of probability ( $\rightarrow$  I/1.4.1), the left-hand side must be finite. This is a contradiction. Therefore,  $p(\emptyset) = 0$ . ■

#### Sources:

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 6, eq. 3; URL: <https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (b); URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549>.
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#The\\_probability\\_of\\_the\\_empty\\_set](https://en.wikipedia.org/wiki/Probability_axioms#The_probability_of_the_empty_set).

#### 1.4.5 Probability of the empty set

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of the empty set is zero:

$$p(\emptyset) = 0 . \quad (1)$$

**Proof:** Let  $E_i = \emptyset$  for  $i = 1, 2, \dots$ . Then,  $E_i \cap E_j = \emptyset \cap \emptyset = \emptyset$  for  $i, j \geq 1$  and  $\bigcup_{i=1}^{\infty} E_i = \bigcup_{i=1}^{\infty} \emptyset = \emptyset$ . Thus,  $E_1, E_2, \dots$  is a countable sequence of disjoint events so that, with the third axiom of probability ( $\rightarrow$  I/1.4.1), it holds that:

$$\begin{aligned} p\left(\bigcup_{i=1}^{\infty} E_i\right) &= \sum_{i=1}^{\infty} p(E_i) \\ p(\emptyset) &= \sum_{i=1}^{\infty} p(\emptyset) . \end{aligned} \quad (2)$$

Since, by the first axiom of probability ( $\rightarrow$  I/1.4.1), probabilities are non-negative, i.e.  $p(\emptyset) \geq 0$ , we are searching for a non-negative number which, when added to itself infinitely, is equal to itself. The only such number is zero, i.e.  $p(\emptyset) = 0$ .

**Sources:**

- Ostwald, Dirk (2023): “Wahrscheinlichkeitsräume”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (1), Folie 19; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/1\\_Wahrscheinlichkeitsraeume.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/1_Wahrscheinlichkeitsraeume.pdf).

**1.4.6 Probability of the complement**

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of a complement of a set is one minus the probability of this set:

$$p(\bar{A}) = 1 - p(A) \quad (1)$$

where  $\bar{A} = \Omega \setminus A$  and  $\Omega$  is the sample space ( $\rightarrow$  I/1.1.2).

**Proof:** Since  $A$  and  $\bar{A}$  are mutually exclusive ( $\rightarrow$  I/1.3.10) and  $A \cup \bar{A} = \Omega$ , the third axiom of probability ( $\rightarrow$  I/1.4.1) implies:

$$\begin{aligned} p(A \cup \bar{A}) &= p(A) + p(\bar{A}) \\ p(\Omega) &= p(A) + p(\bar{A}) \\ p(\bar{A}) &= p(\Omega) - p(A) . \end{aligned} \quad (2)$$

The second axiom of probability ( $\rightarrow$  I/1.4.1) states that  $p(\Omega) = 1$ , such that we obtain:

$$p(\bar{A}) = 1 - p(A) . \quad (3)$$

**Sources:**

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 6, eq. 2; URL: <https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (c); URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549>.
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#The\\_complement\\_rule](https://en.wikipedia.org/wiki/Probability_axioms#The_complement_rule).

**1.4.7 Range of probability**

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of an event is bounded between 0 and 1:

$$0 \leq p(E) \leq 1 . \quad (1)$$

**Proof:** From the first axiom of probability ( $\rightarrow$  I/1.4.1), we have:

$$p(E) \geq 0 . \quad (2)$$

By combining the first axiom of probability ( $\rightarrow$  I/1.4.1) and the probability of the complement ( $\rightarrow$  I/1.4.6), we obtain:

$$\begin{aligned} 1 - p(E) = p(\overline{E}) &\geq 0 \\ 1 - p(E) &\geq 0 \\ p(E) &\leq 1 . \end{aligned} \tag{3}$$

Together, (2) and (3) imply that

$$0 \leq p(E) \leq 1 . \tag{4}$$

■

#### Sources:

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 6; URL: <https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9>
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#The\\_numeric\\_bound](https://en.wikipedia.org/wiki/Probability_axioms#The_numeric_bound).

#### 1.4.8 Addition law of probability

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of the union of  $A$  and  $B$  is the sum of the probabilities of  $A$  and  $B$  minus the probability of the intersection of  $A$  and  $B$ :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) . \tag{1}$$

**Proof:** Let  $E_1 = A$  and  $E_2 = B \setminus A$ , such that  $E_1 \cup E_2 = A \cup B$ . Then, by the third axiom of probability ( $\rightarrow$  I/1.4.1), we have:

$$\begin{aligned} p(A \cup B) &= p(A) + p(B \setminus A) \\ p(A \cup B) &= p(A) + p(B \setminus [A \cap B]) . \end{aligned} \tag{2}$$

Then, let  $E_1 = B \setminus [A \cap B]$  and  $E_2 = A \cap B$ , such that  $E_1 \cup E_2 = B$ . Again, from the third axiom of probability ( $\rightarrow$  I/1.4.1), we obtain:

$$\begin{aligned} p(B) &= p(B \setminus [A \cap B]) + p(A \cap B) \\ p(B \setminus [A \cap B]) &= p(B) - p(A \cap B) . \end{aligned} \tag{3}$$

Plugging (3) into (2), we finally get:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) . \tag{4}$$

■

#### Sources:

- A.N. Kolmogorov (1950): “Elementary Theory of Probability”; in: *Foundations of the Theory of Probability*, p. 2; URL: <https://archive.org/details/foundationsofthe00kolm/page/2/mode/2up>.
- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (a); URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549>.
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#Further\\_consequences](https://en.wikipedia.org/wiki/Probability_axioms#Further_consequences).

### 1.4.9 Bonferroni’s inequality

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of the intersection of  $A$  and  $B$  is larger than or equal to the sum of the probabilities of  $A$  and  $B$  minus one:

$$p(A \cap B) \geq p(A) + p(B) - 1 . \quad (1)$$

**Proof:** The addition law of probability ( $\rightarrow$  I/1.4.8) states that, for two events ( $\rightarrow$  I/1.2.1)  $A$  and  $B$ , it holds true that

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) . \quad (2)$$

Rearranging for  $p(A \cap B)$ , we have:

$$p(A \cap B) = p(A) + p(B) - p(A \cup B) . \quad (3)$$

The range of probability ( $\rightarrow$  I/1.4.7) is

$$0 \leq p(E) \leq 1 . \quad (4)$$

Thus,  $p(A \cup B)$  is at most one. (This is the case, if  $A$  and  $B$  are collectively exhaustive ( $\rightarrow$  I/1.4.12) and their union is thus equal to the sample space ( $\rightarrow$  I/1.4.1)  $\Omega$ .) With this, we are able to derive a lower bound for  $p(A \cap B)$ , as given by the theorem:

$$p(A \cap B) \geq p(A) + p(B) - 1 . \quad (5)$$

■

#### Sources:

- Probability Fact (2021): “Bonferroni’s inequality”; in:  $X$ , retrieved on 2025-01-10; URL: <https://x.com/ProbFact/status/1462828945968189443>.
- Wikipedia (2025): “Boole’s inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Boole%27s\\_inequality#Bonferroni\\_inequalities](https://en.wikipedia.org/wiki/Boole%27s_inequality#Bonferroni_inequalities).

### 1.4.10 Boole’s inequality

**Theorem:** The probability ( $\rightarrow$  I/1.3.1) of any countable sequence of events ( $\rightarrow$  I/1.2.1)  $A_1, A_2, A_3, \dots$  is smaller than or equal to the sum of the probabilities of the individual events ( $\rightarrow$  I/1.4.1):

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} p(A_i) . \quad (1)$$

**Proof:** The addition law of probability ( $\rightarrow$  I/1.4.8) states that, for two events ( $\rightarrow$  I/1.2.1)  $A$  and  $B$ , we have:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) . \quad (2)$$

We will prove the statement by induction, i.e. observe that it holds for  $n = 1$  event and show that, if it holds for  $n$  events, it also holds for  $n + 1$  events. For  $n = 1$ , it is obviously true that:

$$p(A_1) \leq p(A_1) . \quad (3)$$

We shall suppose that statement (1) holds for  $A_1, \dots, A_n$ :

$$p\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n p(A_i) . \quad (4)$$

Using (2), for  $A_1, \dots, A_n, A_{n+1}$ , we then have:

$$p\left(\bigcup_{i=1}^{n+1} A_i\right) = p\left(\bigcup_{i=1}^n A_i \cup A_{n+1}\right) = p\left(\bigcup_{i=1}^n A_i\right) + p(A_{n+1}) - p\left(\bigcup_{i=1}^n A_i \cap A_{n+1}\right) . \quad (5)$$

Since, by the first axiom of probability ( $\rightarrow$  I/1.4.1), any probability is positive

$$p\left(\bigcup_{i=1}^n A_i \cap A_{n+1}\right) \geq 0 , \quad (6)$$

it follows from (5) that

$$p\left(\bigcup_{i=1}^{n+1} A_i\right) \leq p\left(\bigcup_{i=1}^n A_i\right) + p(A_{n+1}) \quad (7)$$

and using (4), we get

$$p\left(\bigcup_{i=1}^{n+1} A_i\right) \leq \sum_{i=1}^n p(A_i) + p(A_{n+1}) , \quad (8)$$

such that

$$p\left(\bigcup_{i=1}^{n+1} A_i\right) \leq \sum_{i=1}^{n+1} p(A_i) . \quad (9)$$

Together, (3) and the transition from (4) to (9), prove the theorem. ■

#### Sources:

- Wikipedia (2025): “Boole’s inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Boole%27s\\_inequality#Proof\\_using\\_induction](https://en.wikipedia.org/wiki/Boole%27s_inequality#Proof_using_induction).

### 1.4.11 Law of total probability

**Theorem:** Let  $A$  be a subset of sample space ( $\rightarrow$  I/1.1.2)  $\Omega$  and let  $B_1, \dots, B_n$  be finite or countably infinite partition of  $\Omega$ , such that  $B_i \cap B_j = \emptyset$  for all  $i \neq j$  and  $\cup_i B_i = \Omega$ . Then, the probability ( $\rightarrow$  I/1.3.1) of the event  $A$  is

$$p(A) = \sum_i p(A \cap B_i) . \quad (1)$$

**Proof:** Because all  $B_i$  are disjoint, sets  $(A \cap B_i)$  are also disjoint:

$$B_i \cap B_j = \emptyset \quad \Rightarrow \quad (A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset . \quad (2)$$

Because the  $B_i$  are exhaustive, the sets  $(A \cap B_i)$  are also exhaustive:

$$\cup_i B_i = \Omega \quad \Rightarrow \quad \cup_i (A \cap B_i) = A \cap (\cup_i B_i) = A \cap \Omega = A . \quad (3)$$

Thus, the third axiom of probability ( $\rightarrow$  I/1.4.1) implies that

$$p(A) = \sum_i p(A \cap B_i) . \quad (4)$$

■

#### Sources:

- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, p. 288, eq. (d); p. 289, eq. 8.7; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549>.
- Wikipedia (2021): “Law of total probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-08; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_total\\_probability#Statement](https://en.wikipedia.org/wiki/Law_of_total_probability#Statement).

### 1.4.12 Probability of exhaustive events

**Theorem:** Let  $B_1, \dots, B_n$  be mutually exclusive ( $\rightarrow$  I/1.3.10) and collectively exhaustive subsets of a sample space ( $\rightarrow$  I/1.1.2)  $\Omega$ . Then, their total probability ( $\rightarrow$  I/1.4.11) is one:

$$\sum_i p(B_i) = 1 . \quad (1)$$

**Proof:** Because all  $B_i$  are mutually exclusive, we have:

$$B_i \cap B_j = \emptyset \quad \text{for all } i \neq j . \quad (2)$$

Because the  $B_i$  are collectively exhaustive, we have:

$$\cup_i B_i = \Omega . \quad (3)$$

Thus, the third axiom of probability ( $\rightarrow$  I/1.4.1) implies that

$$\sum_i p(B_i) = p(\Omega) . \quad (4)$$

and the second axiom of probability ( $\rightarrow$  I/1.4.1) implies that

$$\sum_i p(B_i) = 1 . \quad (5)$$

■

#### Sources:

- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9>
- Wikipedia (2021): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-08; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#Axioms](https://en.wikipedia.org/wiki/Probability_axioms#Axioms).

#### 1.4.13 Probability of exhaustive events

**Theorem:** Let  $B_1, \dots, B_n$  be mutually exclusive ( $\rightarrow$  I/1.3.10) and collectively exhaustive subsets of a sample space ( $\rightarrow$  I/1.1.2)  $\Omega$ . Then, their total probability ( $\rightarrow$  I/1.4.11) is one:

$$\sum_i p(B_i) = 1 . \quad (1)$$

**Proof:** The addition law of probability ( $\rightarrow$  I/1.4.8) states that for two events ( $\rightarrow$  I/1.2.1)  $A$  and  $B$ , the probability ( $\rightarrow$  I/1.3.1) of at least one of them occurring is:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) . \quad (2)$$

Recursively applying this law to the events  $B_1, \dots, B_n$ , we have:

$$\begin{aligned} p(B_1 \cup \dots \cup B_n) &= p(B_1) + p(B_2 \cup \dots \cup B_n) - p(B_1 \cap [B_2 \cup \dots \cup B_n]) \\ &= p(B_1) + p(B_2) + p(B_3 \cup \dots \cup B_n) - p(B_2 \cap [B_3 \cup \dots \cup B_n]) - p(B_1 \cap [B_2 \cup \dots \cup B_n]) \\ &\vdots \\ &= p(B_1) + \dots + p(B_n) - p(B_1 \cap [B_2 \cup \dots \cup B_n]) - \dots - p(B_{n-1} \cap B_n) \\ p(\cup_i^n B_i) &= \sum_i^n p(B_i) - \sum_i^{n-1} p(B_i \cap [\cup_{j=i+1}^n B_j]) \\ &= \sum_i^n p(B_i) - \sum_i^{n-1} p(\cup_{j=i+1}^n [B_i \cap B_j]) . \end{aligned} \quad (3)$$

Because all  $B_i$  are mutually exclusive, we have:

$$B_i \cap B_j = \emptyset \quad \text{for all } i \neq j . \quad (4)$$

Since the probability of the empty set is zero ( $\rightarrow$  I/1.4.4), this means that the second sum on the right-hand side of (3) disappears:

$$p(\cup_i^n B_i) = \sum_i^n p(B_i) . \quad (5)$$



Because the  $B_i$  are collectively exhaustive, we have:

$$\cup_i B_i = \Omega . \quad (6)$$

Since the probability of the sample space is one ( $\rightarrow$  I/1.4.1), this means that the left-hand side of (5) becomes equal to one:

$$1 = \sum_i^n p(B_i) . \quad (7)$$

This proves the statement in (1). ■

#### Sources:

- Alan Stuart & J. Keith Ord (1994): “Probability and Statistical Inference”; in: *Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: <https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9>
- Wikipedia (2022): “Probability axioms”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: [https://en.wikipedia.org/wiki/Probability\\_axioms#Consequences](https://en.wikipedia.org/wiki/Probability_axioms#Consequences).

## 1.5 Probability distributions

### 1.5.1 Probability distribution

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with the set of possible outcomes  $\mathcal{X}$ . Then, a probability distribution of  $X$  is a mathematical function that gives the probabilities ( $\rightarrow$  I/1.3.1) of occurrence of all possible outcomes  $x \in \mathcal{X}$  of this random variable.

#### Sources:

- Wikipedia (2020): “Probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: [https://en.wikipedia.org/wiki/Probability\\_distribution](https://en.wikipedia.org/wiki/Probability_distribution).

### 1.5.2 Joint distribution

**Definition:** Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2) with sets of possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$ . Then, a joint distribution of  $X$  and  $Y$  is a probability distribution ( $\rightarrow$  I/1.5.1) that specifies the probability of the event that  $X = x$  and  $Y = y$  for each possible combination of  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ .

- The joint distribution of two scalar random variables ( $\rightarrow$  I/1.2.2) is called a bivariate distribution.
- The joint distribution of a random vector ( $\rightarrow$  I/1.2.3) is called a multivariate distribution.
- The joint distribution of a random matrix ( $\rightarrow$  I/1.2.4) is called a matrix-variate distribution.

#### Sources:

- Wikipedia (2020): “Joint probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: [https://en.wikipedia.org/wiki/Joint\\_probability\\_distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution).

### 1.5.3 Marginal distribution

**Definition:** Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2) with sets of possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$ . Then, the marginal distribution of  $X$  is a probability distribution ( $\rightarrow$  I/1.5.1) that specifies the probability of the event that  $X = x$  irrespective of the value of  $Y$  for each possible value  $x \in \mathcal{X}$ . The marginal distribution can be obtained from the joint distribution ( $\rightarrow$  I/1.5.2) of  $X$  and  $Y$  using the law of marginal probability ( $\rightarrow$  I/1.3.3).

**Sources:**

- Wikipedia (2020): “Marginal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: [https://en.wikipedia.org/wiki/Marginal\\_distribution](https://en.wikipedia.org/wiki/Marginal_distribution).

### 1.5.4 Conditional distribution

**Definition:** Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2) with sets of possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$ . Then, the conditional distribution of  $X$  given that  $Y$  is a probability distribution ( $\rightarrow$  I/1.5.1) that specifies the probability of the event that  $X = x$  given that  $Y = y$  for each possible combination of  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . The conditional distribution of  $X$  can be obtained from the joint distribution ( $\rightarrow$  I/1.5.2) of  $X$  and  $Y$  and the marginal distribution ( $\rightarrow$  I/1.5.3) of  $Y$  using the law of conditional probability ( $\rightarrow$  I/1.3.4).

**Sources:**

- Wikipedia (2020): “Conditional probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: [https://en.wikipedia.org/wiki/Conditional\\_probability\\_distribution](https://en.wikipedia.org/wiki/Conditional_probability_distribution).

### 1.5.5 Unimodal vs. multimodal distribution

**Definition:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with some probability distribution ( $\rightarrow$  I/1.5.1)  $P$  characterized by probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ . Then,

- $P$  is called a unimodal probability distribution, if  $f_X(x)$  has exactly one local maximum;
- $P$  is called a bimodal probability distribution, if  $f_X(x)$  has exactly two local maxima;
- $P$  is called a trimodal probability distribution, if  $f_X(x)$  has exactly three local maxima;
- $P$  is called a multimodal probability distribution, if  $f_X(x)$  has more than one local maximum.

Note that this definition of multimodality differs from the strict definition of the mode ( $\rightarrow$  I/1.18.3) in which only the global maximum of  $f_X(x)$  would be considered the single mode ( $\rightarrow$  I/1.18.3).

**Sources:**

- Weisstein, Eric W. (2024): “Mode”; in: *Wolfram MathWorld*, retrieved on 2024-10-25; URL: <https://mathworld.wolfram.com/Mode.html>.
- Wikipedia (2024): “Unimodality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-25; URL: [https://en.wikipedia.org/wiki/Unimodality#Unimodal\\_probability\\_distribution](https://en.wikipedia.org/wiki/Unimodality#Unimodal_probability_distribution).

### 1.5.6 Sampling distribution

**Definition:** Let there be a random sample ( $\rightarrow$  I/1.1.6) with finite sample size ( $\rightarrow$  I/1.1.7). Then, the probability distribution ( $\rightarrow$  I/1.5.1) of a given statistic ( $\rightarrow$  I/1.1.8) computed from this sample, e.g. a test statistic ( $\rightarrow$  I/4.3.5), is called a sampling distribution.

**Sources:**

- Wikipedia (2021): “Sampling distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: [https://en.wikipedia.org/wiki/Sampling\\_distribution](https://en.wikipedia.org/wiki/Sampling_distribution).

**1.5.7 Statistical parameter**

**Definition:** A parameter, also “statistical parameter”, is any fixed quantity, i.e. constant ( $\rightarrow$  I/1.2.5) scalar, vector or matrix, that describes a parametrized probability distribution ( $\rightarrow$  I/1.5.1) by influencing its probability mass function ( $\rightarrow$  I/1.6.1) or probability density function ( $\rightarrow$  I/1.7.1).

Examples of parameters include the mean and variance parameters of a normal distribution ( $\rightarrow$  II/3.2.1), covariance parameters in a multivariate ( $\rightarrow$  II/4.1.1) or matrix ( $\rightarrow$  II/5.1.1)-normal distribution, shape and rate parameters of the gamma distribution ( $\rightarrow$  II/3.4.1) or the vector of category probabilities in a multinomial distribution ( $\rightarrow$  II/2.2.1).

**Sources:**

- Wikipedia (2024): “Statistical parameter”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-27; URL: [https://en.wikipedia.org/wiki/Statistical\\_parameter#Parameterised\\_distributions](https://en.wikipedia.org/wiki/Statistical_parameter#Parameterised_distributions).

**1.5.8 Location parameter**

**Definition:** Let  $\mathcal{D}(\mu, \lambda)$  denote a family of probability distributions ( $\rightarrow$  I/1.5.1) with statistical parameters ( $\rightarrow$  I/1.5.7)  $\mu$  and  $\lambda$ . Then, the parameter  $\mu$  is referred to as a “location parameter”, if and only if the following holds: When a random variable ( $\rightarrow$  I/1.2.2)  $X$  is following the probability distribution  $\mathcal{D}(\mu, \lambda)$ , then the random variable  $Y = X - \mu$  is following the probability distribution  $\mathcal{D}(0, \lambda)$ .

$$X \sim \mathcal{D}(\mu^*, \lambda^*) \quad \Rightarrow \quad Y = X - \mu^* \sim \mathcal{D}(0, \lambda^*) . \quad (1)$$

This implies an identity of the cumulative distribution functions ( $\rightarrow$  I/1.8.1):

$$\begin{aligned} F_Y(y) &= F_X(y + \mu^*) \\ F_{\mathcal{D}}(y; 0, \lambda^*) &= F_{\mathcal{D}}(y + \mu^*; \mu^*, \lambda^*) . \end{aligned} \quad (2)$$

The parameter ( $\rightarrow$  I/1.5.7)  $\mu$  can be scalar-, vector- or matrix-valued for random variables ( $\rightarrow$  I/1.2.2), vectors ( $\rightarrow$  I/1.2.3) or matrices ( $\rightarrow$  I/1.2.4)  $X$ .

**Sources:**

- Wikipedia (2025): “Location–scale family”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Location%E2%80%93scale\\_family](https://en.wikipedia.org/wiki/Location%E2%80%93scale_family).
- Wikipedia (2025): “Location parameter”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Location\\_parameter#Definition](https://en.wikipedia.org/wiki/Location_parameter#Definition).

**1.5.9 Scale parameter**

**Definition:** Let  $\mathcal{D}(\lambda, \sigma)$  denote a family of probability distributions ( $\rightarrow$  I/1.5.1) with statistical parameters ( $\rightarrow$  I/1.5.7)  $\lambda$  and  $\sigma$ . Then, the parameter  $\sigma$  is referred to as a “scale parameter”, if and only if the following holds: When a random variable ( $\rightarrow$  I/1.2.2)  $X$  is following the probability

distribution  $\mathcal{D}(\lambda, \sigma)$ , then the random variable  $Y = X/\sigma$  is following the probability distribution  $\mathcal{D}(\lambda, 1)$ .

$$X \sim \mathcal{D}(\lambda^*, \sigma^*) \quad \Rightarrow \quad Y = \frac{X}{\sigma^*} \sim \mathcal{D}(\lambda^*, 1) . \quad (1)$$

This implies an identity of the cumulative distribution functions ( $\rightarrow$  I/1.8.1):

$$\begin{aligned} F_Y(y) &= F_X(\sigma^* y) \\ F_{\mathcal{D}}(y; \lambda^*, 1) &= F_{\mathcal{D}}(\sigma^* y; \lambda^*, \sigma^*) . \end{aligned} \quad (2)$$

The parameter ( $\rightarrow$  I/1.5.7)  $\sigma$  can be scalar for a random variable ( $\rightarrow$  I/1.2.2)  $X$  or matrix-valued for a random matrix ( $\rightarrow$  I/1.2.4)  $X$ .

#### Sources:

- Wikipedia (2025): “Location–scale family”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Location%E2%80%93scale\\_family](https://en.wikipedia.org/wiki/Location%E2%80%93scale_family).
- Wikipedia (2025): “Scale parameter”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Scale\\_parameter#Definition](https://en.wikipedia.org/wiki/Scale_parameter#Definition).

#### 1.5.10 Rate parameter

**Definition:** Let  $\mathcal{D}(\lambda, \rho)$  denote a family of probability distributions ( $\rightarrow$  I/1.5.1) with statistical parameters ( $\rightarrow$  I/1.5.7)  $\lambda$  and  $\rho$ . Then, the parameter  $\rho$  is referred to as a “rate parameter” or “inverse scale parameter”, if it is the reciprocal inverse of a scale parameter ( $\rightarrow$  I/1.5.9) with respect to this family of distributions.

$$\rho \text{ rate} \quad \Leftrightarrow \quad \sigma = 1/\rho \text{ scale} . \quad (1)$$

#### Sources:

- Wikipedia (2025): “Rate parameter”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Scale\\_parameter#Rate\\_parameter](https://en.wikipedia.org/wiki/Scale_parameter#Rate_parameter).

#### 1.5.11 Shape parameter

**Definition:** Let  $\mathcal{D}(\lambda, \chi)$  denote a family of probability distributions ( $\rightarrow$  I/1.5.1) with statistical parameters ( $\rightarrow$  I/1.5.7)  $\lambda$  and  $\chi$ . Then, the parameter  $\chi$  is referred to as a “shape parameter”, if it is neither a location parameter ( $\rightarrow$  I/1.5.8) nor a scale parameter ( $\rightarrow$  I/1.5.9) nor a function of these (such as a rate parameter ( $\rightarrow$  I/1.5.10)).

#### Sources:

- Wikipedia (2025): “Shape parameter”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-03-14; URL: [https://en.wikipedia.org/wiki/Shape\\_parameter](https://en.wikipedia.org/wiki/Shape_parameter).

### 1.5.12 Degrees of freedom

**Definition:** Degrees of freedom can refer to

- a parameter ( $\rightarrow$  I/1.5.7) or parameters ( $\rightarrow$  I/1.5.7) of certain probability distributions ( $\rightarrow$  I/1.5.1), such as the chi-squared distribution ( $\rightarrow$  II/3.7.1), the t-distribution ( $\rightarrow$  II/3.3.1) or the F-distribution ( $\rightarrow$  II/3.8.1);
- the number of values in the final calculation of a statistic ( $\rightarrow$  I/1.1.8) that are free to vary, such as  $n - 1$  for a sample variance ( $\rightarrow$  I/1.13.2) calculated from  $n$  data points ( $\rightarrow$  I/1.1.5).

Some statistics ( $\rightarrow$  I/1.1.8) have sampling distributions ( $\rightarrow$  I/1.5.6) whose parameters ( $\rightarrow$  I/1.5.7) (in the former sense) are equal to their degrees of freedom (in the latter sense). For example, a function of the sample variance follows a chi-squared distribution with  $n - 1$  degrees of freedom ( $\rightarrow$  II/3.2.7).

**Sources:**

- Wikipedia (2024): “Degrees of freedom (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-11; URL: [https://en.wikipedia.org/wiki/Degrees\\_of\\_freedom\\_\(statistics\)](https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)).
- Animated Software (2024): “Degrees of Freedom”; in: *Glossary of Statistical Terms*, retrieved on 2024-10-11; URL: <https://www.animatedsoftware.com/statglos/sgdegree.htm>.

## 1.6 Probability mass function

### 1.6.1 Definition

**Definition:** Let  $X$  be a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$ . Then,  $f_X(x) : \mathbb{R} \rightarrow [0, 1]$  is the probability mass function (PMF) of  $X$ , if

$$f_X(x) = 0 \tag{1}$$

for all  $x \notin \mathcal{X}$ ,

$$\Pr(X = x) = f_X(x) \tag{2}$$

for all  $x \in \mathcal{X}$  and

$$\sum_{x \in \mathcal{X}} f_X(x) = 1. \tag{3}$$

**Sources:**

- Wikipedia (2020): “Probability mass function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: [https://en.wikipedia.org/wiki/Probability\\_mass\\_function](https://en.wikipedia.org/wiki/Probability_mass_function).

### 1.6.2 Probability mass function of sum of independents

**Theorem:** Let  $X$  and  $Y$  be two independent ( $\rightarrow$  I/1.3.6) discrete ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$  and  $\mathcal{Y}$  and let  $Z = X + Y$ . Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Z$  is given by

$$\begin{aligned}
f_Z(z) &= \sum_{y \in \mathcal{Y}} f_X(z - y) f_Y(y) \\
\text{or } f_Z(z) &= \sum_{x \in \mathcal{X}} f_Y(z - x) f_X(x)
\end{aligned} \tag{1}$$

where  $f_X(x)$ ,  $f_Y(y)$  and  $f_Z(z)$  are the probability mass functions ( $\rightarrow$  I/1.6.1) of  $X$ ,  $Y$  and  $Z$ .

**Proof:** Using the definition of the probability mass function ( $\rightarrow$  I/1.6.1) and the expected value ( $\rightarrow$  I/1.12.1), the first equation can be derived as follows:

$$\begin{aligned}
f_Z(z) &= \Pr(Z = z) \\
&= \Pr(X + Y = z) \\
&= \Pr(X = z - Y) \\
&= \mathbb{E}[\Pr(X = z - Y | Y = y)] .
\end{aligned} \tag{2}$$

By construction,  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), such that conditional probabilities are equal to marginal probabilities ( $\rightarrow$  I/1.3.9), i.e.  $\Pr(X = z - Y | Y = y) = \Pr(X = z - Y)$  and we have:

$$\begin{aligned}
f_Z(z) &= \mathbb{E}[\Pr(X = z - Y)] \\
&= \mathbb{E}[f_X(z - Y)] \\
&= \sum_{y \in \mathcal{Y}} f_X(z - y) f_Y(y) .
\end{aligned} \tag{3}$$

The second equation can be derived by switching  $X$  and  $Y$ . ■

#### Sources:

- Taboga, Marco (2017): “Sums of independent random variables”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/sums-of-independent-random-variables>.

### 1.6.3 Probability mass function of strictly increasing function

**Theorem:** Let  $X$  be a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly increasing function on the support of  $X$ . Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \tag{2}$$

**Proof:** Because a strictly increasing function is invertible, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y$  can be derived as follows:

$$\begin{aligned} f_Y(y) &= \Pr(Y = y) \\ &= \Pr(g(X) = y) \\ &= \Pr(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) . \end{aligned} \tag{3}$$

■

**Sources:**

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid3>.

#### 1.6.4 Probability mass function of strictly decreasing function

**Theorem:** Let  $X$  be a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly decreasing function on the support of  $X$ . Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \tag{2}$$

**Proof:** Because a strictly decreasing function is invertible, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y$  can be derived as follows:

$$\begin{aligned} f_Y(y) &= \Pr(Y = y) \\ &= \Pr(g(X) = y) \\ &= \Pr(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) . \end{aligned} \tag{3}$$

■

**Sources:**

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid6>.

### 1.6.5 Probability mass function of invertible function

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) of discrete random variables ( $\rightarrow$  I/1.2.6) with possible outcomes  $\mathcal{X}$  and let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be an invertible function on the support of  $X$ . Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (1)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** Because an invertible function is a one-to-one mapping, the probability mass function ( $\rightarrow$  I/1.6.1) of  $Y$  can be derived as follows:

$$\begin{aligned} f_Y(y) &= \Pr(Y = y) \\ &= \Pr(g(X) = y) \\ &= \Pr(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) . \end{aligned} \quad (3)$$

■

#### Sources:

- Taboga, Marco (2017): “Functions of random vectors and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors>.

## 1.7 Probability density function

### 1.7.1 Definition

**Definition:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$ . Then,  $f_X(x) : \mathbb{R} \rightarrow \mathbb{R}$  is the probability density function (PDF) of  $X$ , if

$$f_X(x) \geq 0 \quad (1)$$

for all  $x \in \mathbb{R}$ ,

$$\Pr(X \in A) = \int_A f_X(x) dx \quad (2)$$

for any  $A \subset \mathcal{X}$  and

$$\int_{\mathcal{X}} f_X(x) dx = 1 . \quad (3)$$

#### Sources:

- Wikipedia (2020): “Probability density function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: [https://en.wikipedia.org/wiki/Probability\\_density\\_function](https://en.wikipedia.org/wiki/Probability_density_function).



### 1.7.2 Probability density function of sum of independents

**Theorem:** Let  $X$  and  $Y$  be two independent ( $\rightarrow$  I/1.3.6) continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$  and  $\mathcal{Y}$  and let  $Z = X + Y$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Z$  is given by

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y) \, dy \\ \text{or } f_Z(z) &= \int_{-\infty}^{+\infty} f_Y(z-x)f_X(x) \, dx \end{aligned} \quad (1)$$

where  $f_X(x)$ ,  $f_Y(y)$  and  $f_Z(z)$  are the probability density functions ( $\rightarrow$  I/1.7.1) of  $X$ ,  $Y$  and  $Z$ .

**Proof:** The cumulative distribution function of a sum of independent random variables ( $\rightarrow$  I/1.8.2) is

$$F_Z(z) = E[F_X(z-Y)] . \quad (2)$$

The probability density function is the first derivative of the cumulative distribution function ( $\rightarrow$  I/1.7.7), such that

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{d}{dz} E[F_X(z-Y)] \\ &= E \left[ \frac{d}{dz} F_X(z-Y) \right] \\ &= E[f_X(z-Y)] \\ &= \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y) \, dy . \end{aligned} \quad (3)$$

The second equation can be derived by switching  $X$  and  $Y$ . ■

#### Sources:

- Taboga, Marco (2017): “Sums of independent random variables”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/sums-of-independent-random-variables>.

### 1.7.3 Probability density function of strictly increasing function

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly increasing function on the support of  $X$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (1)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** The cumulative distribution function of a strictly increasing function ( $\rightarrow$  I/1.8.3) is

$$F_Y(y) = \begin{cases} 0 , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 1 , & \text{if } y > \max(\mathcal{Y}) \end{cases} \quad (3)$$

Because the probability density function is the first derivative of the cumulative distribution function ( $\rightarrow$  I/1.7.7)

$$f_X(x) = \frac{dF_X(x)}{dx} , \quad (4)$$

the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$  can be derived as follows:

1) If  $y$  does not belong to the support of  $Y$ ,  $F_Y(y)$  is constant, such that

$$f_Y(y) = 0, \quad \text{if } y \notin \mathcal{Y} . \quad (5)$$

2) If  $y$  belongs to the support of  $Y$ , then  $f_Y(y)$  can be derived using the chain rule:

$$\begin{aligned} f_Y(y) &\stackrel{(4)}{=} \frac{d}{dy} F_Y(y) \\ &\stackrel{(3)}{=} \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} . \end{aligned} \quad (6)$$

Taking together (5) and (6), eventually proves (1). ■

#### Sources:

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid4>.

#### 1.7.4 Probability density function of strictly decreasing function

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly decreasing function on the support of  $X$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (1)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** The cumulative distribution function of a strictly decreasing function ( $\rightarrow$  I/1.8.4) is

$$F_Y(y) = \begin{cases} 1, & \text{if } y > \max(\mathcal{Y}) \\ 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)), & \text{if } y \in \mathcal{Y} \\ 0, & \text{if } y < \min(\mathcal{Y}) \end{cases} \quad (3)$$

Note that for continuous random variables, the probability ( $\rightarrow$  I/1.7.1) of point events is

$$\Pr(X = a) = \int_a^a f_X(x) dx = 0 . \quad (4)$$

Because the probability density function is the first derivative of the cumulative distribution function ( $\rightarrow$  I/1.7.7)

$$f_X(x) = \frac{dF_X(x)}{dx} , \quad (5)$$

the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$  can be derived as follows:

1) If  $y$  does not belong to the support of  $Y$ ,  $F_Y(y)$  is constant, such that

$$f_Y(y) = 0, \quad \text{if } y \notin \mathcal{Y} . \quad (6)$$

2) If  $y$  belongs to the support of  $Y$ , then  $f_Y(y)$  can be derived using the chain rule:

$$\begin{aligned} f_Y(y) &\stackrel{(5)}{=} \frac{d}{dy} F_Y(y) \\ &\stackrel{(3)}{=} \frac{d}{dy} [1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y))] \\ &\stackrel{(4)}{=} \frac{d}{dy} [1 - F_X(g^{-1}(y))] \\ &= -\frac{d}{dy} F_X(g^{-1}(y)) \\ &= -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} . \end{aligned} \quad (7)$$

Taking together (6) and (7), eventually proves (1). ■

#### Sources:

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid7>.

### 1.7.5 Probability density function of invertible function

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) of continuous random variables ( $\rightarrow$  I/1.2.6) with possible outcomes  $\mathcal{X} \subseteq \mathbb{R}^n$  and let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be an invertible and differentiable function on the support of  $X$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Y = g(X)$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |J_{g^{-1}}(y)|, & \text{if } y \in \mathcal{Y} \\ 0, & \text{if } y \notin \mathcal{Y} \end{cases}, \quad (1)$$

if the Jacobian determinant satisfies

$$|J_{g^{-1}}(y)| \neq 0 \quad \text{for all } y \in \mathcal{Y} \quad (2)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$ ,  $J_{g^{-1}}(y)$  is the Jacobian matrix of  $g^{-1}(y)$

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ \vdots & \ddots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix}, \quad (3)$$

$|J|$  is the determinant of  $J$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}. \quad (4)$$

**Proof:**

1) First, we obtain the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y = g(X)$ . The joint CDF ( $\rightarrow$  I/1.8.11) is given by

$$\begin{aligned} F_Y(y) &= \Pr(Y_1 \leq y_1, \dots, Y_n \leq y_n) \\ &= \Pr(g_1(X) \leq y_1, \dots, g_n(X) \leq y_n) \\ &= \int_{A(y)} f_X(x) dx \end{aligned} \quad (5)$$

where  $A(y)$  is the following subset of the  $n$ -dimensional Euclidean space:

$$A(y) = \{x \in \mathbb{R}^n : g_j(x) \leq y_j \text{ for all } j = 1, \dots, n\} \quad (6)$$

and  $g_j(X)$  is the function which returns the  $j$ -th element of  $Y$ , given a vector  $X$ .

2) Next, we substitute  $x = g^{-1}(y)$  into the integral which gives us

$$\begin{aligned} F_Y(z) &= \int_{B(z)} f_X(g^{-1}(y)) dg^{-1}(y) \\ &= \int_{-\infty}^{z_n} \cdots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) dg^{-1}(y). \end{aligned} \quad (7)$$

where we have modified the integration regime  $B(z)$  which reads

$$B(z) = \{y \in \mathbb{R}^n : y \leq z_j \text{ for all } j = 1, \dots, n\} . \quad (8)$$

3) The formula for change of variables in multivariable calculus states that

$$y = f(x) \quad \Rightarrow \quad dy = |J_f(x)| dx . \quad (9)$$

Applied to equation (7), this yields

$$\begin{aligned} F_Y(z) &= \int_{-\infty}^{z_n} \dots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| dy \\ &= \int_{-\infty}^{z_n} \dots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| dy_1 \dots dy_n . \end{aligned} \quad (10)$$

4) Finally, we obtain the probability density function ( $\rightarrow$  I/1.7.1) of  $Y = g(X)$ . Because the PDF is the derivative of the CDF ( $\rightarrow$  I/1.7.7), we can differentiate the joint CDF to get

$$\begin{aligned} f_Y(z) &= \frac{d^n}{dz_1 \dots dz_n} F_Y(z) \\ &= \frac{d^n}{dz_1 \dots dz_n} \int_{-\infty}^{z_n} \dots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| dy_1 \dots dy_n \\ &= f_X(g^{-1}(z)) |J_{g^{-1}}(z)| \end{aligned} \quad (11)$$

which can also be written as

$$f_Y(y) = f_X(g^{-1}(y)) |J_{g^{-1}}(y)| . \quad (12)$$

■

#### Sources:

- Taboga, Marco (2017): “Functions of random vectors and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors>.
- Lebanon, Guy (2017): “Functions of a Random Vector”; in: *Probability: The Analysis of Data, Vol. 1*, retrieved on 2021-08-30; URL: [http://theanalysisofdata.com/probability/4\\_4.html](http://theanalysisofdata.com/probability/4_4.html).
- Poirier, Dale J. (1995): “Distributions of Functions of Random Variables”; in: *Intermediate Statistics and Econometrics: A Comparative Approach*, ch. 4, pp. 149ff.; URL: [https://books.google.de/books?id=K52\\_YvD1YNwC&hl=de&source=gbp\\_navlinks\\_s](https://books.google.de/books?id=K52_YvD1YNwC&hl=de&source=gbp_navlinks_s).
- Devore, Jay L.; Berk, Kenneth N. (2011): “Conditional Distributions”; in: *Modern Mathematical Statistics with Applications*, ch. 5.2, pp. 253ff.; URL: [https://books.google.de/books?id=5PRLUho-YYgC&hl=de&source=gbp\\_navlinks\\_s](https://books.google.de/books?id=5PRLUho-YYgC&hl=de&source=gbp_navlinks_s).
- peek-a-boo (2019): “How to come up with the Jacobian in the change of variables formula”; in: *StackExchange Mathematics*, retrieved on 2021-08-30; URL: <https://math.stackexchange.com/a/3239222>.
- Bazett, Trefor (2019): “Change of Variables & The Jacobian | Multi-variable Integration”; in: *YouTube*, retrieved on 2021-08-30; URL: <https://www.youtube.com/watch?v=wUF-lyyWpUc>.

### 1.7.6 Probability density function of linear transformation

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) of continuous random variables ( $\rightarrow$  I/1.2.6) with possible outcomes  $\mathcal{X} \subseteq \mathbb{R}^n$  and let  $Y = \Sigma X + \mu$  be a linear transformation of this random variable with constant ( $\rightarrow$  I/1.2.5)  $n$ -dimensional vector  $\mu$  and constant ( $\rightarrow$  I/1.2.5)  $n \times n$  matrix  $\Sigma$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$  is

$$f_Y(y) = \begin{cases} \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y - \mu)) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (1)$$

where  $|\Sigma|$  is the determinant of  $\Sigma$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = \Sigma x + \mu : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** Because the linear function  $g(X) = \Sigma X + \mu$  is invertible and differentiable, we can determine the probability density function of an invertible function of a continuous random vector ( $\rightarrow$  I/1.7.5) using the relation

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} . \quad (3)$$

The inverse function is

$$X = g^{-1}(Y) = \Sigma^{-1}(Y - \mu) = \Sigma^{-1}Y - \Sigma^{-1}\mu \quad (4)$$

and the Jacobian matrix is

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ \vdots & \ddots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix} = \Sigma^{-1} . \quad (5)$$

Plugging (4) and (5) into (3) and applying the determinant property  $|A^{-1}| = |A|^{-1}$ , we obtain

$$f_Y(y) = \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y - \mu)) . \quad (6)$$

■

#### Sources:

- Taboga, Marco (2017): “Functions of random vectors and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors>.

### 1.7.7 Probability density function in terms of cumulative distribution function

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2). Then, the probability distribution function ( $\rightarrow$  I/1.7.1) of  $X$  is the first derivative of the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$ :

$$f_X(x) = \frac{dF_X(x)}{dx} . \quad (1)$$

**Proof:** The cumulative distribution function in terms of the probability density function of a continuous random variable ( $\rightarrow$  I/1.8.6) is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad x \in \mathbb{R} . \quad (2)$$

Taking the derivative with respect to  $x$ , we have:

$$\frac{dF_X(x)}{dx} = \frac{d}{dx} \int_{-\infty}^x f_X(t) dt . \quad (3)$$

The fundamental theorem of calculus states that, if  $f(x)$  is a continuous real-valued function defined on the interval  $[a, b]$ , then it holds that

$$F(x) = \int_a^x f(t) dt \quad \Rightarrow \quad F'(x) = f(x) \quad \text{for all } x \in (a, b) . \quad (4)$$

Applying (4) to (2), it follows that

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad \Rightarrow \quad \frac{dF_X(x)}{dx} = f_X(x) \quad \text{for all } x \in \mathbb{R} . \quad (5)$$

■

#### Sources:

- Wikipedia (2020): “Fundamental theorem of calculus”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: [https://en.wikipedia.org/wiki/Fundamental\\_theorem\\_of\\_calculus#Formal\\_statements](https://en.wikipedia.org/wiki/Fundamental_theorem_of_calculus#Formal_statements).

### 1.7.8 Joint probability density function

**Definition:** Let  $X \in \mathbb{R}^n$  be a continuous ( $\rightarrow$  I/1.2.6) random vector ( $\rightarrow$  I/1.2.3). Then, a function  $f_X(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  is the joint ( $\rightarrow$  I/1.5.2) probability density function ( $\rightarrow$  I/1.7.1) of  $X$ , if

$$f_X(x) \geq 0 \quad (1)$$

for all  $x \in \mathbb{R}^n$ ,

$$\Pr(X \in A) = \int_A f_X(x) dx \quad (2)$$

for any  $A \subset \mathbb{R}^n$  and

$$\int_{\mathbb{R}^n} f_X(x) dx = 1 . \quad (3)$$

#### Sources:

- Wikipedia (2025): “Probability density function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-02-06; URL: [https://en.wikipedia.org/wiki/Probability\\_density\\_function#Densities\\_associated\\_with\\_multiple\\_variables](https://en.wikipedia.org/wiki/Probability_density_function#Densities_associated_with_multiple_variables).

## 1.8 Cumulative distribution function

### 1.8.1 Definition

**Definition:** The cumulative distribution function (CDF) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  at a given value  $x$  is defined as the probability ( $\rightarrow$  I/1.3.1) that  $X$  is smaller than  $x$ :

$$F_X(x) = \Pr(X \leq x) . \quad (1)$$

1) If  $X$  is a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and the probability mass function ( $\rightarrow$  I/1.6.1)  $f_X(x)$ , then the cumulative distribution function is the function ( $\rightarrow$  I/1.8.5)  $F_X(x) : \mathbb{R} \rightarrow [0, 1]$  with

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) . \quad (2)$$

2) If  $X$  is a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and the probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ , then the cumulative distribution function is the function ( $\rightarrow$  I/1.8.6)  $F_X(x) : \mathbb{R} \rightarrow [0, 1]$  with

$$F_X(x) = \int_{-\infty}^x f_X(t) dt . \quad (3)$$

#### Sources:

- Wikipedia (2020): “Cumulative distribution function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: [https://en.wikipedia.org/wiki/Cumulative\\_distribution\\_function#Definition](https://en.wikipedia.org/wiki/Cumulative_distribution_function#Definition).

### 1.8.2 Cumulative distribution function of sum of independents

**Theorem:** Let  $X$  and  $Y$  be two independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) and let  $Z = X + Y$ . Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Z$  is given by

$$\begin{aligned} F_Z(z) &= \mathbb{E}[F_X(z - Y)] \\ \text{or } F_Z(z) &= \mathbb{E}[F_Y(z - X)] \end{aligned} \quad (1)$$

where  $F_X(x)$ ,  $F_Y(y)$  and  $F_Z(z)$  are the cumulative distribution functions ( $\rightarrow$  I/1.8.1) of  $X$ ,  $Y$  and  $Z$  and  $\mathbb{E}[\cdot]$  denotes the expected value ( $\rightarrow$  I/1.12.1).

**Proof:** Using the definition of the cumulative distribution function ( $\rightarrow$  I/1.8.1), the first equation can be derived as follows:

$$\begin{aligned} F_Z(z) &= \Pr(Z \leq z) \\ &= \Pr(X + Y \leq z) \\ &= \Pr(X \leq z - Y) \\ &= \mathbb{E}[\Pr(X \leq z - Y | Y = y)] \\ &= \mathbb{E}[\Pr(X \leq z - Y)] \\ &= \mathbb{E}[F_X(z - Y)] . \end{aligned} \quad (2)$$



Note that the second-last transition is justified by the fact that  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), such that conditional probabilities are equal to marginal probabilities ( $\rightarrow$  I/1.3.9). The second equation can be derived by switching  $X$  and  $Y$ . ■

#### Sources:

- Taboga, Marco (2017): “Sums of independent random variables”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: <https://www.statlect.com/fundamentals-of-probability/sums-of-independent-random-variables>.

### 1.8.3 Cumulative distribution function of strictly increasing function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly increasing function on the support of  $X$ . Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y = g(X)$  is given by

$$F_Y(y) = \begin{cases} 0, & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)), & \text{if } y \in \mathcal{Y} \\ 1, & \text{if } y > \max(\mathcal{Y}) \end{cases} \quad (1)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** The support of  $Y$  is determined by  $g(x)$  and by the set of possible outcomes of  $X$ . Moreover, if  $g(x)$  is strictly increasing, then  $g^{-1}(y)$  is also strictly increasing. Therefore, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y$  can be derived as follows:

1) If  $y$  is lower than the lowest value ( $\rightarrow$  I/1.20.1)  $Y$  can take, then  $\Pr(Y \leq y) = 0$ , so

$$F_Y(y) = 0, \quad \text{if } y < \min(\mathcal{Y}) . \quad (3)$$

2) If  $y$  belongs to the support of  $Y$ , then  $F_Y(y)$  can be derived as follows:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(g(X) \leq y) \\ &= \Pr(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) . \end{aligned} \quad (4)$$

3) If  $y$  is higher than the highest value ( $\rightarrow$  I/1.20.2)  $Y$  can take, then  $\Pr(Y \leq y) = 1$ , so

$$F_Y(y) = 1, \quad \text{if } y > \max(\mathcal{Y}) . \quad (5)$$

Taking together (3), (4), (5), eventually proves (1). ■

#### Sources:

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid2>.

#### 1.8.4 Cumulative distribution function of strictly decreasing function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $g(x)$  be a strictly decreasing function on the support of  $X$ . Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y = g(X)$  is given by

$$F_Y(y) = \begin{cases} 1, & \text{if } y > \max(\mathcal{Y}) \\ 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)), & \text{if } y \in \mathcal{Y} \\ 0, & \text{if } y < \min(\mathcal{Y}) \end{cases} \quad (1)$$

where  $g^{-1}(y)$  is the inverse function of  $g(x)$  and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (2)$$

**Proof:** The support of  $Y$  is determined by  $g(x)$  and by the set of possible outcomes of  $X$ . Moreover, if  $g(x)$  is strictly decreasing, then  $g^{-1}(y)$  is also strictly decreasing. Therefore, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y$  can be derived as follows:

1) If  $y$  is higher than the highest value ( $\rightarrow$  I/1.20.2)  $Y$  can take, then  $\Pr(Y \leq y) = 1$ , so

$$F_Y(y) = 1, \quad \text{if } y > \max(\mathcal{Y}) . \quad (3)$$

2) If  $y$  belongs to the support of  $Y$ , then  $F_Y(y)$  can be derived as follows:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= 1 - \Pr(Y > y) \\ &= 1 - \Pr(g(X) > y) \\ &= 1 - \Pr(X < g^{-1}(y)) \\ &= 1 - \Pr(X < g^{-1}(y)) - \Pr(X = g^{-1}(y)) + \Pr(X = g^{-1}(y)) \\ &= 1 - [\Pr(X < g^{-1}(y)) + \Pr(X = g^{-1}(y))] + \Pr(X = g^{-1}(y)) \\ &= 1 - \Pr(X \leq g^{-1}(y)) + \Pr(X = g^{-1}(y)) \\ &= 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)) . \end{aligned} \quad (4)$$

3) If  $y$  is lower than the lowest value ( $\rightarrow$  I/1.20.1)  $Y$  can take, then  $\Pr(Y \leq y) = 0$ , so

$$F_Y(y) = 0, \quad \text{if } y < \min(\mathcal{Y}) . \quad (5)$$

Taking together (3), (4), (5), eventually proves (1). ■

#### Sources:

- Taboga, Marco (2017): “Functions of random variables and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid5>.

### 1.8.5 Cumulative distribution function of discrete random variable

**Theorem:** Let  $X$  be a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$  and probability mass function ( $\rightarrow$  I/1.6.1)  $f_X(x)$ . Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) . \quad (1)$$

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as the probability that  $X$  is smaller than  $x$ :

$$F_X(x) = \Pr(X \leq x) . \quad (2)$$

The probability mass function ( $\rightarrow$  I/1.6.1) of a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2)  $X$  returns the probability that  $X$  takes a particular value  $x$ :

$$f_X(x) = \Pr(X = x) . \quad (3)$$

Taking these two definitions together, we have:

$$\begin{aligned} F_X(x) &\stackrel{(2)}{=} \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} \Pr(X = t) \\ &\stackrel{(3)}{=} \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) . \end{aligned} \quad (4)$$

■

### 1.8.6 Cumulative distribution function of continuous random variable

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$  and probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ . Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \int_{-\infty}^x f_X(t) dt . \quad (1)$$

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as the probability that  $X$  is smaller than  $x$ :

$$F_X(x) = \Pr(X \leq x) . \quad (2)$$

The probability density function ( $\rightarrow$  I/1.7.1) of a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2)  $X$  can be used to calculate the probability that  $X$  falls into a particular interval  $A$ :

$$\Pr(X \in A) = \int_A f_X(x) dx . \quad (3)$$

Taking these two definitions together, we have:

$$\begin{aligned}
F_X(x) &\stackrel{(2)}{=} \Pr(X \in (-\infty, x]) \\
&\stackrel{(3)}{=} \int_{-\infty}^x f_X(t) dt .
\end{aligned} \tag{4}$$

■

### 1.8.7 Exceedance probability based on cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$  and cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$ . Then, the exceedance probability ( $\rightarrow$  I/1.3.11) that  $X$  is larger than some value  $x$  is

$$\Pr(X > x) = 1 - F_X(x) . \tag{1}$$

**Proof:** Note that  $\{X \mid X > x\}$  and  $\{X \mid X \leq x\}$  are disjoint sets

$$\{X \mid X > x\} \cap \{X \mid X \leq x\} = \emptyset \tag{2}$$

and that they comprise the set of all outcomes, i.e. the sample space ( $\rightarrow$  I/1.1.2):

$$\{X \mid X > x\} \cup \{X \mid X \leq x\} = \mathcal{X} = \Omega . \tag{3}$$

Using the second axiom of probability ( $\rightarrow$  I/1.4.1), we have:

$$\begin{aligned}
p(\Omega) &= 1 \\
p(\{X \mid X > x\} \cup \{X \mid X \leq x\}) &= 1 .
\end{aligned} \tag{4}$$

Using the third axiom of probability ( $\rightarrow$  I/1.4.1), we get:

$$\begin{aligned}
p(\{X \mid X > x\}) + p(\{X \mid X \leq x\}) &= 1 \\
p(\{X \mid X > x\}) &= 1 - p(\{X \mid X \leq x\}) \\
\Pr(X > x) &= 1 - \Pr(X \leq x) .
\end{aligned} \tag{5}$$

Using the definition of the cumulative distribution function ( $\rightarrow$  I/1.8.1), we finally have:

$$\Pr(X > x) = 1 - F_X(x) . \tag{6}$$

■

#### Sources:

- Ostwald, Dirk (2023): “Zufallsvariablen”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (3), Folie 34; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Wintersemester2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/3\\_Zufallsvariablen.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Wintersemester2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/3_Zufallsvariablen.pdf).

### 1.8.8 Probability integral transform

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with invertible ( $\rightarrow$  I/1.11.1) cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$ . Then, the random variable ( $\rightarrow$  I/1.2.2)

$$Y = F_X(X) \quad (1)$$

has a standard uniform distribution ( $\rightarrow$  II/3.1.2).

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y = F_X(X)$  can be derived as

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(F_X(X) \leq y) \\ &= \Pr(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned} \quad (2)$$

which is the cumulative distribution function of a continuous uniform distribution ( $\rightarrow$  II/3.1.4) with  $a = 0$  and  $b = 1$ , i.e. the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the standard uniform distribution ( $\rightarrow$  II/3.1.2)  $\mathcal{U}(0, 1)$ . ■

#### Sources:

- Wikipedia (2021): “Probability integral transform”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: [https://en.wikipedia.org/wiki/Probability\\_integral\\_transform#Proof](https://en.wikipedia.org/wiki/Probability_integral_transform#Proof).

### 1.8.9 Inverse transformation method

**Theorem:** Let  $U$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) having a standard uniform distribution ( $\rightarrow$  II/3.1.2). Then, the random variable ( $\rightarrow$  I/1.2.2)

$$X = F_X^{-1}(U) \quad (1)$$

has a probability distribution ( $\rightarrow$  I/1.5.1) characterized by the invertible ( $\rightarrow$  I/1.11.1) cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$ .

**Proof:** With the cumulative distribution function of the continuous uniform distribution ( $\rightarrow$  II/3.1.4), the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the standard uniform distribution ( $\rightarrow$  II/3.1.2) is

$$U \sim \mathcal{U}(0, 1) \quad \Rightarrow \quad F_U(u) = \Pr(U \leq u) = u. \quad (2)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the transformation  $X = F_X^{-1}(U)$  can be derived as

$$\begin{aligned} \Pr(X \leq x) &= \Pr(F_X^{-1}(U) \leq x) \\ &= \Pr(U \leq F_X(x)) \\ &= F_U(F_X(x)) \\ &= F_X(x). \end{aligned} \quad (3)$$

**Sources:**

- Wikipedia (2021): “Inverse transform sampling”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: [https://en.wikipedia.org/wiki/Inverse\\_transform\\_sampling#Proof\\_of\\_correctness](https://en.wikipedia.org/wiki/Inverse_transform_sampling#Proof_of_correctness).

**1.8.10 Distributional transformation**

**Theorem:** Let  $X$  and  $Y$  be two continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$  and invertible cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_Y(y)$ . Then, the random variable ( $\rightarrow$  I/1.2.2)

$$\tilde{X} = F_Y^{-1}(F_X(X)) \quad (1)$$

has the same probability distribution ( $\rightarrow$  I/1.5.1) as  $Y$ .

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) of the transformation  $\tilde{X} = F_Y^{-1}(F_X(X))$  can be derived as

$$\begin{aligned} F_{\tilde{X}}(y) &= \Pr(\tilde{X} \leq y) \\ &= \Pr(F_Y^{-1}(F_X(X)) \leq y) \\ &= \Pr(F_X(X) \leq F_Y(y)) \\ &= \Pr(X \leq F_X^{-1}(F_Y(y))) \\ &= F_X(F_X^{-1}(F_Y(y))) \\ &= F_Y(y) \end{aligned} \quad (2)$$

which shows that  $\tilde{X}$  and  $Y$  have the same cumulative distribution function ( $\rightarrow$  I/1.8.1) and are thus identically distributed ( $\rightarrow$  I/1.5.1).

**Sources:**

- Soch, Joram (2020): “Distributional Transformation Improves Decoding Accuracy When Predicting Chronological Age From Structural MRI”; in: *Frontiers in Psychiatry*, vol. 11, art. 604268; URL: <https://www.frontiersin.org/articles/10.3389/fpsy.2020.604268/full>; DOI: 10.3389/fpsy.2020.604268

**1.8.11 Joint cumulative distribution function**

**Definition:** Let  $X \in \mathbb{R}^n$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Then, the joint ( $\rightarrow$  I/1.5.2) cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is defined as the probability ( $\rightarrow$  I/1.3.1) that each entry  $X_i$  is smaller than a specific value  $x_i$  for  $i = 1, \dots, n$ :

$$F_X(x) = \Pr(X_1 \leq x_1, \dots, X_n \leq x_n) . \quad (1)$$

**Sources:**

- Wikipedia (2021): “Cumulative distribution function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: [https://en.wikipedia.org/wiki/Cumulative\\_distribution\\_function#Definition\\_for\\_more\\_than\\_two\\_random\\_variables](https://en.wikipedia.org/wiki/Cumulative_distribution_function#Definition_for_more_than_two_random_variables).

## 1.9 Moment-generating function

### 1.9.1 Definition

**Definition:**

1) The moment-generating function of a random variable ( $\rightarrow$  I/1.2.2)  $X \in \mathbb{R}$  is

$$M_X(t) = \mathbb{E} [e^{tX}], \quad t \in \mathbb{R}. \quad (1)$$

2) The moment-generating function of a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^n$  is

$$M_X(t) = \mathbb{E} [e^{t^T X}], \quad t \in \mathbb{R}^n. \quad (2)$$

**Sources:**

- Wikipedia (2020): “Moment-generating function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-22; URL: [https://en.wikipedia.org/wiki/Moment-generating\\_function#Definition](https://en.wikipedia.org/wiki/Moment-generating_function#Definition).
- Taboga, Marco (2017): “Joint moment generating function”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-10-07; URL: <https://www.statlect.com/fundamentals-of-probability/joint-moment-generating-function>.

### 1.9.2 Moment-generating function of sum of independents

**Theorem:** Let  $X$  and  $Y$  be two independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) and let  $Z = X + Y$ . Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $Z$  is given by

$$M_Z(t) = M_X(t) \cdot M_Y(t) \quad (1)$$

where  $M_X(t)$ ,  $M_Y(t)$  and  $M_Z(t)$  are the moment-generating functions ( $\rightarrow$  I/1.9.1) of  $X$ ,  $Y$  and  $Z$ .

**Proof:** The moment-generating function of a random variable ( $\rightarrow$  I/1.9.1)  $X$  is

$$M_X(t) = \mathbb{E} (\exp [tX]) \quad (2)$$

and therefore the moment-generating function of the sum  $Z$  is given by

$$\begin{aligned} M_Z(t) &= \mathbb{E} (\exp [tZ]) \\ &= \mathbb{E} (\exp [t(X + Y)]) \\ &= \mathbb{E} (\exp [tX] \cdot \exp [tY]) . \end{aligned} \quad (3)$$

Because the expected value is multiplicative for independent random variables ( $\rightarrow$  I/1.12.7), we have

$$\begin{aligned} M_Z(t) &= \mathbb{E} (\exp [tX]) \cdot \mathbb{E} (\exp [tY]) \\ &= M_X(t) \cdot M_Y(t) . \end{aligned} \quad (4)$$

■

**Sources:**

- Probability Fact (2021): “If  $X$  and  $Y$  are independent, the moment generating function (MGF)”; in:  $X$ , retrieved on 2024-11-08; URL: <https://x.com/ProbFact/status/1468264616706859016>.

### 1.9.3 Moment-generating function of arbitrary function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with the expected value ( $\rightarrow$  I/1.12.1) function  $E_X[\cdot]$ . Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $Y = g(X)$  is equal to

$$M_Y(t) = E_X [\exp(t g(X))] . \quad (1)$$

**Proof:** The moment-generating function ( $\rightarrow$  I/1.9.1) is defined as

$$M_Y(t) = E [\exp(t Y)] . \quad (2)$$

Due of the law of the unconscious statistician ( $\rightarrow$  I/1.12.9)

$$\begin{aligned} E[g(X)] &= \sum_{x \in \mathcal{X}} g(x) f_X(x) \\ E[g(X)] &= \int_{\mathcal{X}} g(x) f_X(x) dx , \end{aligned} \quad (3)$$

$Y = g(X)$  can simply be substituted into (2) to give

$$M_Y(t) = E_X [\exp(t g(X))] . \quad (4)$$

■

#### Sources:

- Taboga, Marco (2017): “Functions of random vectors and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors>.

### 1.9.4 Moment-generating function of linear transformation

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) with the moment-generating function ( $\rightarrow$  I/1.9.1)  $M_X(t)$ . Then, the moment-generating function of the linear transformation  $Y = AX + b$  is given by

$$M_Y(t) = \exp [t^T b] \cdot M_X(At) \quad (1)$$

where  $A$  is an  $m \times n$  matrix and  $b$  is an  $m \times 1$  vector.

**Proof:** The moment-generating function of a random vector ( $\rightarrow$  I/1.9.1)  $X$  is

$$M_X(t) = E (\exp [t^T X]) \quad (2)$$

and therefore the moment-generating function of the random vector ( $\rightarrow$  I/1.2.3)  $Y$  is given by

$$\begin{aligned} M_Y(t) &= E (\exp [t^T (AX + b)]) \\ &= E (\exp [t^T AX] \cdot \exp [t^T b]) \\ &= \exp [t^T b] \cdot E (\exp [(At)^T X]) \\ &= \exp [t^T b] \cdot M_X(At) . \end{aligned} \quad (3)$$



**Sources:**

- ProofWiki (2020): “Moment Generating Function of Linear Transformation of Random Variable”; in: *ProofWiki*, retrieved on 2020-08-19; URL: [https://proofwiki.org/wiki/Moment\\_Generating\\_Function\\_of\\_Linear\\_Transformation\\_of\\_Random\\_Variable](https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Transformation_of_Random_Variable).

**1.9.5 Moment-generating function of linear combination**

**Theorem:** Let  $X_1, \dots, X_n$  be  $n$  independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) with moment-generating functions ( $\rightarrow$  I/1.9.1)  $M_{X_i}(t)$ . Then, the moment-generating function of the linear combination  $X = \sum_{i=1}^n a_i X_i$  is given by

$$M_X(t) = \prod_{i=1}^n M_{X_i}(a_i t) \quad (1)$$

where  $a_1, \dots, a_n$  are  $n$  real numbers.

**Proof:** The moment-generating function of a random variable ( $\rightarrow$  I/1.9.1)  $X_i$  is

$$M_{X_i}(t) = E(\exp[tX_i]) \quad (2)$$

and therefore the moment-generating function of the linear combination  $X$  is given by

$$\begin{aligned} M_X(t) &= E(\exp[tX]) \\ &= E\left(\exp\left[t \sum_{i=1}^n a_i X_i\right]\right) \\ &= E\left(\prod_{i=1}^n \exp[t a_i X_i]\right). \end{aligned} \quad (3)$$

Because the expected value is multiplicative for independent random variables ( $\rightarrow$  I/1.12.7), we have

$$\begin{aligned} M_X(t) &= \prod_{i=1}^n E(\exp[(a_i t)X_i]) \\ &= \prod_{i=1}^n M_{X_i}(a_i t). \end{aligned} \quad (4)$$

**Sources:**

- ProofWiki (2020): “Moment Generating Function of Linear Combination of Independent Random Variables”; in: *ProofWiki*, retrieved on 2020-08-19; URL: [https://proofwiki.org/wiki/Moment\\_Generating\\_Function\\_of\\_Linear\\_Combination\\_of\\_Independent\\_Random\\_Variables](https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Combination_of_Independent_Random_Variables).

## 1.10 Probability-generating function

### 1.10.1 Definition

**Definition:**

1) If  $X$  is a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) taking values in the non-negative integers  $\{0, 1, \dots\}$ , then the probability-generating function of  $X$  is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x) z^x \quad (1)$$

where  $z \in \mathbb{C}$  and  $f_X(x)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ .

2) If  $X$  is a discrete ( $\rightarrow$  I/1.2.6) random vector ( $\rightarrow$  I/1.2.3) taking values in the  $n$ -dimensional integer lattice  $x \in \{0, 1, \dots\}^n$ , then the probability-generating function of  $X$  is defined as

$$G_X(z) = \sum_{x_1=0}^{\infty} \cdots \sum_{x_n=0}^{\infty} f_{X_1, \dots, X_n}(x_1, \dots, x_n) z_1^{x_1} \cdots z_n^{x_n} \quad (2)$$

where  $z \in \mathbb{C}^n$  and  $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ .

**Sources:**

- Wikipedia (2020): “Probability-generating function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: [https://en.wikipedia.org/wiki/Probability-generating\\_function#Definition](https://en.wikipedia.org/wiki/Probability-generating_function#Definition).

### 1.10.2 Probability-generating function in terms of expected value

**Theorem:** Let  $X$  be a discrete ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) whose set of possible values  $\mathcal{X}$  is a subset of the natural numbers  $\mathbb{N}$ . Then, the probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$  can be expressed in terms of an expected value ( $\rightarrow$  I/1.12.1) of a function of  $X$

$$G_X(z) = \mathbb{E} [z^X] \quad (1)$$

where  $z \in \mathbb{C}$ .

**Proof:** The law of the unconscious statistician ( $\rightarrow$  I/1.12.9) states that

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \quad (2)$$

where  $f_X(x)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ . Here, we have  $g(X) = z^X$ , such that

$$\mathbb{E} [z^X] = \sum_{x \in \mathcal{X}} z^x f_X(x) . \quad (3)$$

By the definition of  $X$ , this is equal to

$$\mathbb{E} [z^X] = \sum_{x=0}^{\infty} f_X(x) z^x . \quad (4)$$

The right-hand side is equal to the probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$ :

$$\mathbb{E} [z^X] = G_X(z) . \quad (5)$$

**Sources:**

- ProofWiki (2022): “Probability Generating Function as Expectation”; in: *ProofWiki*, retrieved on 2022-10-11; URL: [https://proofwiki.org/wiki/Probability\\_Generating\\_Function\\_as\\_Expectation](https://proofwiki.org/wiki/Probability_Generating_Function_as_Expectation).

**1.10.3 Probability-generating function of zero**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with probability-generating function ( $\rightarrow$  I/1.10.1)  $G_X(z)$  and probability mass function ( $\rightarrow$  I/1.6.1)  $f_X(x)$ . Then, the value of the probability-generating function at zero is equal to the value of the probability mass function at zero:

$$G_X(0) = f_X(0) . \quad (1)$$

**Proof:** The probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$  is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x) z^x \quad (2)$$

where  $f_X(x)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ . Setting  $z = 0$ , we obtain:

$$\begin{aligned} G_X(0) &= \sum_{x=0}^{\infty} f_X(x) \cdot 0^x \\ &= f_X(0) + 0^1 \cdot f_X(1) + 0^2 \cdot f_X(2) + \dots \\ &= f_X(0) + 0 + 0 + \dots \\ &= f_X(0) . \end{aligned} \quad (3)$$

**Sources:**

- ProofWiki (2022): “Probability Generating Function of Zero”; in: *ProofWiki*, retrieved on 2022-10-11; URL: [https://proofwiki.org/wiki/Probability\\_Generating\\_Function\\_of\\_Zero](https://proofwiki.org/wiki/Probability_Generating_Function_of_Zero).

**1.10.4 Probability-generating function of one**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with probability-generating function ( $\rightarrow$  I/1.10.1)  $G_X(z)$  and set of possible values  $\mathcal{X}$ . Then, the value of the probability-generating function at one is equal to one:

$$G_X(1) = 1 . \quad (1)$$

**Proof:** The probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$  is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x) z^x \quad (2)$$

where  $f_X(x)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ . Setting  $z = 1$ , we obtain:

$$\begin{aligned}
G_X(1) &= \sum_{x=0}^{\infty} f_X(x) \cdot 1^x \\
&= \sum_{x=0}^{\infty} f_X(x) \cdot 1 \\
&= \sum_{x=0}^{\infty} f_X(x) .
\end{aligned} \tag{3}$$

Because the probability mass function ( $\rightarrow$  I/1.6.1) sums up to one, this becomes:

$$\begin{aligned}
G_X(1) &= \sum_{x \in \mathcal{X}} f_X(x) \\
&= 1 .
\end{aligned} \tag{4}$$

■

#### Sources:

- ProofWiki (2022): “Probability Generating Function of One”; in: *ProofWiki*, retrieved on 2022-10-11; URL: [https://proofwiki.org/wiki/Probability\\_Generating\\_Function\\_of\\_One](https://proofwiki.org/wiki/Probability_Generating_Function_of_One).

## 1.11 Other probability functions

### 1.11.1 Quantile function

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with the cumulative distribution function ( $\rightarrow$  I/1.8.1) (CDF)  $F_X(x)$ . Then, the function  $Q_X(p) : [0, 1] \rightarrow \mathbb{R}$  which is the inverse CDF is the quantile function (QF) of  $X$ . More precisely, the QF is the function that, for a given quantile  $p \in [0, 1]$ , returns the smallest  $x$  for which  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \tag{1}$$

#### Sources:

- Wikipedia (2020): “Probability density function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: [https://en.wikipedia.org/wiki/Quantile\\_function#Definition](https://en.wikipedia.org/wiki/Quantile_function#Definition).

### 1.11.2 Quantile function in terms of cumulative distribution function

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2) with the cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$ . If the cumulative distribution function is strictly monotonically increasing, then the quantile function ( $\rightarrow$  I/1.11.1) is identical to the inverse of  $F_X(x)$ :

$$Q_X(p) = F_X^{-1}(x) . \tag{1}$$

**Proof:** The quantile function ( $\rightarrow$  I/1.11.1)  $Q_X(p)$  is defined as the function that, for a given quantile  $p \in [0, 1]$ , returns the smallest  $x$  for which  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \quad (2)$$

If  $F_X(x)$  is continuous and strictly monotonically increasing, then there is exactly one  $x$  for which  $F_X(x) = p$  and  $F_X(x)$  is an invertible function, such that

$$Q_X(p) = F_X^{-1}(p) . \quad (3)$$

■

#### Sources:

- Wikipedia (2020): “Quantile function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: [https://en.wikipedia.org/wiki/Quantile\\_function#Definition](https://en.wikipedia.org/wiki/Quantile_function#Definition).

### 1.11.3 Characteristic function

#### Definition:

1) The characteristic function of a random variable ( $\rightarrow$  I/1.2.2)  $X \in \mathbb{R}$  is

$$\varphi_X(t) = \mathbb{E} [e^{itX}] , \quad t \in \mathbb{R} . \quad (1)$$

2) The characteristic function of a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^n$  is

$$\varphi_X(t) = \mathbb{E} [e^{it^T X}] , \quad t \in \mathbb{R}^n . \quad (2)$$

3) The characteristic function of a random matrix ( $\rightarrow$  I/1.2.4)  $X \in \mathbb{R}^{n \times p}$  is

$$\varphi_X(t) = \mathbb{E} [e^{i \operatorname{tr}(t^T X)}] , \quad t \in \mathbb{R}^{n \times p} . \quad (3)$$

#### Sources:

- Wikipedia (2021): “Characteristic function (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-09-22; URL: [https://en.wikipedia.org/wiki/Characteristic\\_function\\_\(probability\\_theory\)#Definition](https://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)#Definition).
- Taboga, Marco (2017): “Joint characteristic function”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-10-07; URL: <https://www.statlect.com/fundamentals-of-probability/joint-characteristic-function>.

### 1.11.4 Characteristic function of arbitrary function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with the expected value ( $\rightarrow$  I/1.12.1) function  $\mathbb{E}_X[\cdot]$ . Then, the characteristic function ( $\rightarrow$  I/1.11.3) of  $Y = g(X)$  is equal to

$$\varphi_Y(t) = \mathbb{E}_X [\exp(it g(X))] . \quad (1)$$

**Proof:** The characteristic function ( $\rightarrow$  I/1.11.3) is defined as

$$\varphi_Y(t) = \mathbb{E} [\exp(it Y)] . \quad (2)$$

Due to the law of the unconscious statistician ( $\rightarrow$  I/1.12.9)

$$\begin{aligned} \mathbb{E}[g(X)] &= \sum_{x \in \mathcal{X}} g(x) f_X(x) \\ \mathbb{E}[g(X)] &= \int_{\mathcal{X}} g(x) f_X(x) \, dx, \end{aligned} \tag{3}$$

$Y = g(X)$  can simply be substituted into (2) to give

$$\varphi_Y(t) = \mathbb{E}_X [\exp(it g(X))] . \tag{4}$$

■

#### Sources:

- Taboga, Marco (2017): “Functions of random vectors and their distribution”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors>.

### 1.11.5 Cumulant-generating function

#### Definition:

1) The cumulant-generating function of a random variable ( $\rightarrow$  I/1.2.2)  $X \in \mathbb{R}$  is

$$K_X(t) = \log \mathbb{E} [e^{tX}] , \quad t \in \mathbb{R} . \tag{1}$$

2) The cumulant-generating function of a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^n$  is

$$K_X(t) = \log \mathbb{E} [e^{t^T X}] , \quad t \in \mathbb{R}^n . \tag{2}$$

#### Sources:

- Wikipedia (2020): “Cumulant”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: <https://en.wikipedia.org/wiki/Cumulant#Definition>.

## 1.12 Expected value

### 1.12.1 Definition

#### Definition:

1) The expected value (or, mean) of a discrete random variable ( $\rightarrow$  I/1.2.2)  $X$  with domain  $\mathcal{X}$  is

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{1}$$

where  $f_X(x)$  is the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$ .

2) The expected value (or, mean) of a continuous random variable ( $\rightarrow$  I/1.2.2)  $X$  with domain  $\mathcal{X}$  is

$$\mathbb{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx \tag{2}$$

where  $f_X(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ .

#### Sources:

- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: [https://en.wikipedia.org/wiki/Expected\\_value#Definition](https://en.wikipedia.org/wiki/Expected_value#Definition).

### 1.12.2 Sample mean

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the sample mean of  $x$  is denoted as  $\bar{x}$  and is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i . \quad (1)$$

**Sources:**

- Wikipedia (2021): “Sample mean and covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-16; URL: [https://en.wikipedia.org/wiki/Sample\\_mean\\_and\\_covariance#Definition\\_of\\_the\\_sample\\_mean](https://en.wikipedia.org/wiki/Sample_mean_and_covariance#Definition_of_the_sample_mean).

### 1.12.3 Non-negative random variable

**Theorem:** Let  $X$  be a non-negative random variable ( $\rightarrow$  I/1.2.2). Then, the expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \int_0^{\infty} (1 - F_X(x)) dx \quad (1)$$

where  $F_X(x)$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$ .

**Proof:** Because the cumulative distribution function gives the probability of a random variable being smaller than a given value ( $\rightarrow$  I/1.8.1),

$$F_X(x) = \Pr(X \leq x) , \quad (2)$$

we have ( $\rightarrow$  I/1.8.7)

$$1 - F_X(x) = \Pr(X > x) , \quad (3)$$

such that

$$\int_0^{\infty} (1 - F_X(x)) dx = \int_0^{\infty} \Pr(X > x) dx \quad (4)$$

which, using the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ , can be rewritten as

$$\begin{aligned}
\int_0^\infty (1 - F_X(x)) \, dx &= \int_0^\infty \int_x^\infty f_X(z) \, dz \, dx \\
&= \int_0^\infty \int_0^z f_X(z) \, dx \, dz \\
&= \int_0^\infty f_X(z) \int_0^z 1 \, dx \, dz \\
&= \int_0^\infty [x]_0^z \cdot f_X(z) \, dz \\
&= \int_0^\infty z \cdot f_X(z) \, dz
\end{aligned} \tag{5}$$

and by applying the definition of the expected value ( $\rightarrow$  I/1.12.1), we see that

$$\int_0^\infty (1 - F_X(x)) \, dx = \int_0^\infty z \cdot f_X(z) \, dz = E(X) \tag{6}$$

which proves the identity given above. ■

#### Sources:

- Kemp, Graham (2014): “Expected value of a non-negative random variable”; in: *StackExchange Mathematics*, retrieved on 2020-05-18; URL: <https://math.stackexchange.com/questions/958472/expected-value-of-a-non-negative-random-variable>.

#### 1.12.4 Non-negativity

**Theorem:** If a random variable ( $\rightarrow$  I/1.2.2) is strictly non-negative, its expected value ( $\rightarrow$  I/1.12.1) is also non-negative, i.e.

$$E(X) \geq 0, \quad \text{if } X \geq 0. \tag{1}$$

#### Proof:

1) If  $X \geq 0$  is a discrete random variable, then, because the probability mass function ( $\rightarrow$  I/1.6.1) is always non-negative, all the addends in

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{2}$$

are non-negative, thus the entire sum must be non-negative.

2) If  $X \geq 0$  is a continuous random variable, then, because the probability density function ( $\rightarrow$  I/1.7.1) is always non-negative, the integrand in

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx \tag{3}$$

is strictly non-negative, thus the term on the right-hand side is a Lebesgue integral, so that the result on the left-hand side must be non-negative.



**Sources:**

- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: [https://en.wikipedia.org/wiki/Expected\\_value#Basic\\_properties](https://en.wikipedia.org/wiki/Expected_value#Basic_properties).

**1.12.5 Linearity**

**Theorem:** The expected value ( $\rightarrow$  I/1.12.1) is a linear operator, i.e.

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) \\ E(a X) &= a E(X) \end{aligned} \quad (1)$$

for random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  and a constant ( $\rightarrow$  I/1.2.5)  $a$ .

**Proof:**

1) If  $X$  and  $Y$  are discrete random variables ( $\rightarrow$  I/1.2.6), the expected value ( $\rightarrow$  I/1.12.1) is

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \quad (2)$$

and the law of marginal probability ( $\rightarrow$  I/1.3.3) states that

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y) . \quad (3)$$

Applying this, we have

$$\begin{aligned} E(X + Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x + y) \cdot f_{X,Y}(x, y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot f_{X,Y}(x, y) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y \cdot f_{X,Y}(x, y) \\ &= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} f_{X,Y}(x, y) + \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} f_{X,Y}(x, y) \\ &\stackrel{(3)}{=} \sum_{x \in \mathcal{X}} x \cdot f_X(x) + \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\ &\stackrel{(2)}{=} E(X) + E(Y) \end{aligned} \quad (4)$$

as well as

$$\begin{aligned} E(a X) &= \sum_{x \in \mathcal{X}} a x \cdot f_X(x) \\ &= a \sum_{x \in \mathcal{X}} x \cdot f_X(x) \\ &\stackrel{(2)}{=} a E(X) . \end{aligned} \quad (5)$$

2) If  $X$  and  $Y$  are continuous random variables ( $\rightarrow$  I/1.2.6), the expected value ( $\rightarrow$  I/1.12.1) is

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx \quad (6)$$

and the law of marginal probability ( $\rightarrow$  I/1.3.3) states that

$$p(x) = \int_{\mathcal{Y}} p(x, y) dy. \quad (7)$$

Applying this, we have

$$\begin{aligned} E(X + Y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x + y) \cdot f_{X,Y}(x, y) dy dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} x \cdot f_{X,Y}(x, y) dy dx + \int_{\mathcal{X}} \int_{\mathcal{Y}} y \cdot f_{X,Y}(x, y) dy dx \\ &= \int_{\mathcal{X}} x \int_{\mathcal{Y}} f_{X,Y}(x, y) dy dx + \int_{\mathcal{Y}} y \int_{\mathcal{X}} f_{X,Y}(x, y) dx dy \\ &\stackrel{(7)}{=} \int_{\mathcal{X}} x \cdot f_X(x) dx + \int_{\mathcal{Y}} y \cdot f_Y(y) dy \\ &\stackrel{(6)}{=} E(X) + E(Y) \end{aligned} \quad (8)$$

as well as

$$\begin{aligned} E(aX) &= \int_{\mathcal{X}} ax \cdot f_X(x) dx \\ &= a \int_{\mathcal{X}} x \cdot f_X(x) dx \\ &\stackrel{(6)}{=} aE(X). \end{aligned} \quad (9)$$

Collectively, this shows that both requirements for linearity are fulfilled for the expected value ( $\rightarrow$  I/1.12.1), for discrete ( $\rightarrow$  I/1.2.6) as well as for continuous ( $\rightarrow$  I/1.2.6) random variables. The present derivation also holds for the expected value of random vectors ( $\rightarrow$  I/1.12.16) as well as for the expected value of random matrices ( $\rightarrow$  I/1.12.19). ■

#### Sources:

- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: [https://en.wikipedia.org/wiki/Expected\\_value#Basic\\_properties](https://en.wikipedia.org/wiki/Expected_value#Basic_properties).
- Michael B, Kuldeep Guha Mazumder, Geoff Pilling et al. (2020): “Linearity of Expectation”; in: *brilliant.org*, retrieved on 2020-02-13; URL: <https://brilliant.org/wiki/linearity-of-expectation/>.

#### 1.12.6 Monotonicity

**Theorem:** The expected value ( $\rightarrow$  I/1.12.1) is monotonic, i.e.

$$E(X) \leq E(Y), \quad \text{if } X \leq Y. \quad (1)$$

**Proof:** Let  $Z = Y - X$ . Due to the linearity of the expected value ( $\rightarrow$  I/1.12.5), we have

$$E(Z) = E(Y - X) = E(Y) - E(X) . \quad (2)$$

With the non-negativity property of the expected value ( $\rightarrow$  I/1.12.4), it also holds that

$$Z \geq 0 \quad \Rightarrow \quad E(Z) \geq 0 . \quad (3)$$

Together with (2), this yields

$$E(Y) - E(X) \geq 0 \quad (4)$$

and rearranging, we have

$$E(X) \leq E(Y) . \quad (5)$$

■

#### Sources:

- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: [https://en.wikipedia.org/wiki/Expected\\_value#Basic\\_properties](https://en.wikipedia.org/wiki/Expected_value#Basic_properties).

### 1.12.7 (Non-)Multiplicativity

#### Theorem:

1) If two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), the expected value ( $\rightarrow$  I/1.12.1) is multiplicative, i.e.

$$E(XY) = E(X)E(Y) . \quad (1)$$

2) If two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  are dependent ( $\rightarrow$  I/1.3.6), the expected value ( $\rightarrow$  I/1.12.1) is not necessarily multiplicative, i.e. there exist  $X$  and  $Y$  such that

$$E(XY) \neq E(X)E(Y) . \quad (2)$$

#### Proof:

1) If  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), it holds that

$$p(x, y) = p(x)p(y) \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y} . \quad (3)$$

Applying this to the expected value for discrete random variables ( $\rightarrow$  I/1.12.1), we have

$$\begin{aligned} E(XY) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x, y) \\ &\stackrel{(3)}{=} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \\ &= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\ &= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \cdot E(Y) \\ &= E(X)E(Y) . \end{aligned} \quad (4)$$

And applying it to the expected value for continuous random variables ( $\rightarrow$  I/1.12.1), we have

$$\begin{aligned}
 E(XY) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x, y) \, dy \, dx \\
 &\stackrel{(3)}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \, dy \, dx \\
 &= \int_{\mathcal{X}} x \cdot f_X(x) \int_{\mathcal{Y}} y \cdot f_Y(y) \, dy \, dx \\
 &= \int_{\mathcal{X}} x \cdot f_X(x) \cdot E(Y) \, dx \\
 &= E(X) E(Y) .
 \end{aligned} \tag{5}$$

2) Let  $X$  and  $Y$  be Bernoulli random variables ( $\rightarrow$  II/1.1.1) with the following joint probability ( $\rightarrow$  I/1.3.2) mass function ( $\rightarrow$  I/1.6.1)

$$\begin{aligned}
 p(X = 0, Y = 0) &= 1/2 \\
 p(X = 0, Y = 1) &= 0 \\
 p(X = 1, Y = 0) &= 0 \\
 p(X = 1, Y = 1) &= 1/2
 \end{aligned} \tag{6}$$

and thus, the following marginal probabilities ( $\rightarrow$  I/1.3.3):

$$\begin{aligned}
 p(X = 0) &= p(X = 1) = 1/2 \\
 p(Y = 0) &= p(Y = 1) = 1/2 .
 \end{aligned} \tag{7}$$

Then,  $X$  and  $Y$  are dependent, because

$$p(X = 0, Y = 1) \stackrel{(6)}{=} 0 \neq \frac{1}{2} \cdot \frac{1}{2} \stackrel{(7)}{=} p(X = 0) p(Y = 1) , \tag{8}$$

and the expected value of their product is

$$\begin{aligned}
 E(XY) &= \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} (x \cdot y) \cdot p(x, y) \\
 &= (1 \cdot 1) \cdot p(X = 1, Y = 1) \\
 &\stackrel{(6)}{=} \frac{1}{2}
 \end{aligned} \tag{9}$$

while the product of their expected values is

$$\begin{aligned}
 E(X) E(Y) &= \left( \sum_{x \in \{0,1\}} x \cdot p(x) \right) \cdot \left( \sum_{y \in \{0,1\}} y \cdot p(y) \right) \\
 &= (1 \cdot p(X = 1)) \cdot (1 \cdot p(Y = 1)) \\
 &\stackrel{(7)}{=} \frac{1}{4}
 \end{aligned} \tag{10}$$

and thus,

$$E(XY) \neq E(X)E(Y) . \quad (11)$$

■

**Sources:**

- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: [https://en.wikipedia.org/wiki/Expected\\_value#Basic\\_properties](https://en.wikipedia.org/wiki/Expected_value#Basic_properties).

### 1.12.8 Law of total expectation

**Theorem:** (law of total expectation, also called “law of iterated expectations”) Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $E(X)$  and let  $Y$  be any random variable ( $\rightarrow$  I/1.13.1) defined on the same probability space ( $\rightarrow$  I/1.1.4). Then, the expected value ( $\rightarrow$  I/1.12.1) of the conditional expectation of  $X$  given  $Y$  is the same as the expected value ( $\rightarrow$  I/1.12.1) of  $X$ :

$$E(X) = E[E(X|Y)] . \quad (1)$$

**Proof:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.6) with sets of possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$ . Then, the expectation of the conditional expectation can be rewritten as:

$$\begin{aligned} E[E(X|Y)] &= E \left[ \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x|Y) \right] \\ &= \sum_{y \in \mathcal{Y}} \left[ \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x|Y = y) \right] \cdot \Pr(Y = y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot \Pr(X = x|Y = y) \cdot \Pr(Y = y) . \end{aligned} \quad (2)$$

Using the law of conditional probability ( $\rightarrow$  I/1.3.4), this becomes:

$$\begin{aligned} E[E(X|Y)] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot \Pr(X = x, Y = y) \\ &= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y) . \end{aligned} \quad (3)$$

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), this becomes:

$$\begin{aligned} E[E(X|Y)] &= \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x) \\ &= E(X) . \end{aligned} \quad (4)$$

■

**Sources:**

- Wikipedia (2021): “Law of total expectation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_total\\_expectation#Proof\\_in\\_the\\_finite\\_and\\_countable\\_cases](https://en.wikipedia.org/wiki/Law_of_total_expectation#Proof_in_the_finite_and_countable_cases).

### 1.12.9 Law of the unconscious statistician

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) and let  $Y = g(X)$  be a function of this random variable.

1) If  $X$  is a discrete random variable with possible outcomes  $\mathcal{X}$  and probability mass function ( $\rightarrow$  I/1.6.1)  $f_X(x)$ , the expected value ( $\rightarrow$  I/1.12.1) of  $g(X)$  is

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) . \quad (1)$$

2) If  $X$  is a continuous random variable with possible outcomes  $\mathcal{X}$  and probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ , the expected value ( $\rightarrow$  I/1.12.1) of  $g(X)$  is

$$E[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx . \quad (2)$$

**Proof:** Suppose that  $g$  is differentiable and that its inverse  $g^{-1}$  is monotonic.

1) The expected value ( $\rightarrow$  I/1.12.1) of  $Y = g(X)$  is defined as

$$E[Y] = \sum_{y \in \mathcal{Y}} y f_Y(y) . \quad (3)$$

Writing the probability mass function  $f_Y(y)$  in terms of  $y = g(x)$ , we have:

$$\begin{aligned} E[g(X)] &= \sum_{y \in \mathcal{Y}} y \Pr(g(x) = y) \\ &= \sum_{y \in \mathcal{Y}} y \Pr(x = g^{-1}(y)) \\ &= \sum_{y \in \mathcal{Y}} y \sum_{x=g^{-1}(y)} f_X(x) \\ &= \sum_{y \in \mathcal{Y}} \sum_{x=g^{-1}(y)} y f_X(x) \\ &= \sum_{y \in \mathcal{Y}} \sum_{x=g^{-1}(y)} g(x) f_X(x) . \end{aligned} \quad (4)$$

Finally, noting that “for all  $y$ , then for all  $x = g^{-1}(y)$ ” is equivalent to “for all  $x$ ”, if  $g^{-1}$  is a monotonic function, we can conclude that

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) . \quad (5)$$

2) Let  $y = g(x)$ . The derivative of an inverse function is

$$\frac{d}{dy}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))} \quad (6)$$

Because  $x = g^{-1}(y)$ , this can be rearranged into

$$dx = \frac{1}{g'(g^{-1}(y))} dy \quad (7)$$

and substituting (7) into (2), we get

$$\int_{\mathcal{X}} g(x) f_X(x) dx = \int_{\mathcal{Y}} y f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} dy . \quad (8)$$

Considering the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y$ , one can deduce:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(g(X) \leq y) \\ &= \Pr(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) . \end{aligned} \quad (9)$$

Differentiating to get the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$ , the result is:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &\stackrel{(9)}{=} \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)) \\ &\stackrel{(6)}{=} f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} . \end{aligned} \quad (10)$$

Finally, substituting (10) into (8), we have:

$$\int_{\mathcal{X}} g(x) f_X(x) dx = \int_{\mathcal{Y}} y f_Y(y) dy = E[Y] = E[g(X)] . \quad (11)$$

■

#### Sources:

- Wikipedia (2020): “Law of the unconscious statistician”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_the\\_unconscious\\_statistician#Proof](https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician#Proof).
- Taboga, Marco (2017): “Transformation theorem”; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: <https://www.statlect.com/glossary/transformation-theorem>.

#### 1.12.10 Squared expectation of a product

**Theorem:** Let  $X$  and  $Y$  be two random variables ( $\rightarrow$  I/1.2.2) with expected values ( $\rightarrow$  I/1.12.1)  $E(X)$  and  $E(Y)$  and let  $E(XY)$  exist and be finite. Then, the square of the expectation of the product of  $X$  and  $Y$  is less than or equal to the product of the expectation of the squares of  $X$  and  $Y$ :

$$[E(XY)]^2 \leq E(X^2) E(Y^2) . \quad (1)$$

**Proof:** Note that  $Y^2$  is a non-negative random variable ( $\rightarrow$  I/1.2.2) whose expected value is also non-negative ( $\rightarrow$  I/1.12.4):

$$E(Y^2) \geq 0 . \quad (2)$$

1) First, consider the case that  $E(Y^2) > 0$ . Define a new random variable  $Z$  as

$$Z = X - Y \frac{E(XY)}{E(Y^2)} . \quad (3)$$

Once again, because  $Z^2$  is always non-negative, we have the expected value:

$$E(Z^2) \geq 0 . \quad (4)$$

Thus, using the linearity of the expected value ( $\rightarrow$  I/1.12.5), we have

$$\begin{aligned} 0 &\leq E(Z^2) \\ &\leq E\left(\left(X - Y \frac{E(XY)}{E(Y^2)}\right)^2\right) \\ &\leq E\left(X^2 - 2XY \frac{E(XY)}{E(Y^2)} + Y^2 \frac{[E(XY)]^2}{[E(Y^2)]^2}\right) \\ &\leq E(X^2) - 2E(XY) \frac{E(XY)}{E(Y^2)} + E(Y^2) \frac{[E(XY)]^2}{[E(Y^2)]^2} \\ &\leq E(X^2) - 2 \frac{[E(XY)]^2}{E(Y^2)} + \frac{[E(XY)]^2}{E(Y^2)} \\ &\leq E(X^2) - \frac{[E(XY)]^2}{E(Y^2)} , \end{aligned} \quad (5)$$

giving

$$[E(XY)]^2 \leq E(X^2) E(Y^2) \quad (6)$$

as required.

2) Next, consider the case that  $E(Y^2) = 0$ . In this case,  $Y$  must be a constant ( $\rightarrow$  I/1.2.5) with mean ( $\rightarrow$  I/1.12.1)  $E(Y) = 0$  and variance ( $\rightarrow$  I/1.13.1)  $\text{Var}(Y) = 0$ , thus we have

$$\Pr(Y = 0) = 1 . \quad (7)$$

This implies

$$\Pr(XY = 0) = 1 , \quad (8)$$

such that

$$E(XY) = 0 . \quad (9)$$

Thus, we can write

$$0 = [E(XY)]^2 = E(X^2) E(Y^2) = 0 , \quad (10)$$

giving

$$[E(XY)]^2 \leq E(X^2) E(Y^2) \quad (11)$$

as required.



**Sources:**

- ProofWiki (2022): “Square of Expectation of Product is Less Than or Equal to Product of Expectation of Squares”; in: *ProofWiki*, retrieved on 2022-10-11; URL: [https://proofwiki.org/wiki/Square\\_of\\_Expectation\\_of\\_Product\\_is\\_Less\\_Than\\_or\\_Equal\\_to\\_Product\\_of\\_Expectation\\_of\\_Squares](https://proofwiki.org/wiki/Square_of_Expectation_of_Product_is_Less_Than_or_Equal_to_Product_of_Expectation_of_Squares).

**1.12.11 Jensen’s inequality**

**Theorem:** Let  $X$  be a univariate ( $\rightarrow$  I/1.2.7) random variable ( $\rightarrow$  I/1.2.2) with probability mass function ( $\rightarrow$  I/1.6.1) (if discrete ( $\rightarrow$  I/1.2.6)) or probability density function ( $\rightarrow$  I/1.7.1) (if continuous ( $\rightarrow$  I/1.2.6))  $f_X(x)$  and let  $g(x)$  be a convex function. Then, the function value, evaluated at the expected value ( $\rightarrow$  I/1.12.1) of  $X$ , is smaller than or equal to the expected value ( $\rightarrow$  I/1.12.1) of the function value, evaluated at  $X$ :

$$g(E[X]) \leq E[g(X)] . \quad (1)$$

**Proof:** A function  $g(x)$  is said to be convex on the interval  $[x_1, x_2]$ , if every straight line connecting two points of the function’s graph lies above the function’s graph:

$$g(qx_1 + (1 - q)x_2) \leq qg(x_1) + (1 - q)g(x_2) \quad \text{for any } q \in [0, 1] . \quad (2)$$

This property can be extended to an arbitrary number of points  $x_i$ ,  $i = 1, \dots, n$ , with  $q_i \geq 0$ ,  $i = 1, \dots, n$  and  $\sum_{i=1}^n q_i = 1$ , such that

$$g\left(\sum_{i=1}^n q_i x_i\right) \leq \sum_{i=1}^n q_i g(x_i) . \quad (3)$$

If  $\{x_1, \dots, x_n\}$  is the set of possible values for a discrete random variable ( $\rightarrow$  I/1.2.6)  $X$ , then the  $q_i$  fulfill the definition of a probability mass function ( $\rightarrow$  I/1.6.1)  $f_X(x_i) = q_i$ . With the law of the unconscious statistician ( $\rightarrow$  I/1.12.9), it then follows that

$$g(E[X]) \leq E[g(X)] . \quad (4)$$

The above property can be further extended to a continuum of points  $x \in \mathcal{X}$ , with  $q(x) \geq 0$  and  $\int_{\mathcal{X}} q(x) dx = 1$ , such that

$$g\left(\int_{\mathcal{X}} q(x)x dx\right) \leq \int_{\mathcal{X}} q(x)g(x) dx . \quad (5)$$

If  $\mathcal{X}$  is the set of possible values for a continuous random variable ( $\rightarrow$  I/1.2.6)  $X$ , then  $q(x)$  fulfill the definition of a probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x) = q(x)$ . With the law of the unconscious statistician ( $\rightarrow$  I/1.12.9), it then follows that

$$g(E[X]) \leq E[g(X)] . \quad (6)$$

This completes the proof.

**Sources:**

- Ostwald et al. (2014): “A tutorial on variational Bayes for latent linear stochastic time-series models”; in: *Journal of Mathematical Psychology*, vol. 60, pp. 1-19, App. A; URL: <https://www.sciencedirect.com/science/article/pii/S0022249614000352>; DOI: 10.1016/j.jmp.2014.04.0.

### 1.12.12 Markov’s inequality

**Theorem:** Let  $X$  be a non-negative ( $\rightarrow$  I/1.12.3) random variable ( $\rightarrow$  I/1.2.2) and let  $a > 0$  be a positive real number. Then, the probability ( $\rightarrow$  I/1.3.1) that  $X$  is larger than or equal to  $a$  is smaller than or equal to the expected value ( $\rightarrow$  I/1.12.1) of  $X$ , divided by  $a$ :

$$\Pr(X \geq a) \leq \frac{E(X)}{a} . \quad (1)$$

**Proof:** Let us define an indicator variable that is one whenever  $X$  is larger than or equal to  $a$ :

$$I_{X \geq a} = \begin{cases} 0 , & \text{if } X < a \\ 1 , & \text{if } X \geq a . \end{cases} \quad (2)$$

It is clear that the following inequality holds:

$$a I_{X \geq a} \leq X . \quad (3)$$

Taking the expectation ( $\rightarrow$  I/1.12.1) on both sides, we have:

$$\begin{aligned} E(a I_{X \geq a}) &\leq E(X) \\ a E(I_{X \geq a}) &\leq E(X) \\ a [0 \cdot \Pr(I_{X \geq a} = 0) + 1 \cdot \Pr(I_{X \geq a} = 1)] &\leq E(X) \\ a \Pr(X \geq a) &\leq E(X) \\ \Pr(X \geq a) &\leq \frac{E(X)}{a} . \end{aligned} \quad (4)$$

■

#### Sources:

- Probability and Statistics (2023): “On Markov’s Inequality”; in:  $X$ , retrieved on 2025-01-10; URL: <https://x.com/probnstat/status/1713998688530370794>.
- Wikipedia (2025): “Markov’s inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Markov%27s\\_inequality#Probability-theoretic\\_proof](https://en.wikipedia.org/wiki/Markov%27s_inequality#Probability-theoretic_proof).

### 1.12.13 Chebyshev’s inequality

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with finite expected value ( $\rightarrow$  I/1.12.1)  $\mu$  and finite variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ . Then, for any positive real number  $k > 0$ , the following inequality holds:

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} . \quad (1)$$

**Proof:** Markov's inequality ( $\rightarrow$  I/1.12.12) states that, for any positive real number  $a > 0$ , the probability ( $\rightarrow$  I/1.3.1) that  $Y$  is larger than or equal to  $a$  is smaller than or equal to the expected value ( $\rightarrow$  I/1.12.1) of  $Y$ , divided by  $a$ :

$$\Pr(Y \geq a) \leq \frac{E(Y)}{a} . \quad (2)$$

The variance of a random variable ( $\rightarrow$  I/1.13.1) is defined as:

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] \quad \text{where} \quad \mu = E(X) . \quad (3)$$

Let  $Y = (X - \mu)^2$  and  $a = (k\sigma)^2$ . Using Markov's inequality, we then have:

$$\begin{aligned} \Pr(|X - \mu| \geq k\sigma) &= \Pr((X - \mu)^2 \geq (k\sigma)^2) \\ &\stackrel{(2)}{\leq} \frac{E[(X - \mu)^2]}{(k\sigma)^2} \\ &\stackrel{(3)}{=} \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2} . \end{aligned} \quad (4)$$

■

#### Sources:

- Wikipedia (2025): “Chebyshev’s inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Chebyshev%27s\\_inequality#Proof](https://en.wikipedia.org/wiki/Chebyshev%27s_inequality#Proof).

#### 1.12.14 Weak law of large numbers

**Theorem:** Let  $X_1, \dots, X_n$  be independent and identically distributed ( $\rightarrow$  I/1.2.8) random variables ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $E(X_i) = \mu$  and finite variance ( $\rightarrow$  I/1.13.1)  $\text{Var}(X_i) < \infty$  for  $i = 1, \dots, n$ . The sample mean ( $\rightarrow$  I/1.12.2) is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i . \quad (1)$$

Then, for any positive number  $\epsilon > 0$ , the probability that the absolute difference of the sample mean ( $\rightarrow$  I/1.12.2) from the expected value ( $\rightarrow$  I/1.12.1)  $\mu$  is smaller than  $\epsilon$  will approach one, as sample size ( $\rightarrow$  I/1.1.7) goes to infinity:

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X} - \mu| < \epsilon) = 1 . \quad (2)$$

**Proof:** Since  $X_1, \dots, X_n$  are independent and identically distributed ( $\rightarrow$  I/1.2.8), they have the same mean, denoted as  $\mu$ , and the same variance, denoted as  $\sigma^2$ . Using the linearity of the expected value ( $\rightarrow$  I/1.12.5), the expected value of the sample mean becomes:

$$\begin{aligned}
E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\
&= \frac{1}{n} \sum_{i=1}^n E(X_i) \\
&= \frac{1}{n} n E(X_i) \\
&= \mu .
\end{aligned} \tag{3}$$

Moreover, with the scaling of the variance upon multiplication ( $\rightarrow$  I/1.13.8) and the additivity of the variance under independence ( $\rightarrow$  I/1.13.11), the variance of the sample mean becomes:

$$\begin{aligned}
\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\
&= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \\
&= \frac{1}{n^2} n \sigma^2 \\
&= \frac{\sigma^2}{n} .
\end{aligned} \tag{4}$$

Chebyshev's inequality ( $\rightarrow$  I/1.12.13) makes a statement about a random variable ( $\rightarrow$  I/1.2.2)  $X$  in relation to its mean  $\mu$  and variance  $\sigma^2$  for any positive number  $k > 0$ :

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} . \tag{5}$$

Applying this inequality to the random variable  $\bar{X}$ , we have:

$$\begin{aligned}
\Pr(|\bar{X} - E(\bar{X})| \geq x) &\leq \frac{\text{Var}(\bar{X})}{x^2} \\
\Pr(|\bar{X} - \mu| \geq \epsilon) &\leq \frac{\sigma^2}{n\epsilon^2} .
\end{aligned} \tag{6}$$

Since the cumulative distribution function can be used to relate probabilities of inverse events ( $\rightarrow$  I/1.8.7), i.e.  $\Pr(X \geq x) = 1 - \Pr(X < x)$ , we have:

$$\begin{aligned}
1 - \Pr(|\bar{X} - \mu| < \epsilon) &\leq \frac{\sigma^2}{n\epsilon^2} \\
\Pr(|\bar{X} - \mu| < \epsilon) &\geq 1 - \frac{\sigma^2}{n\epsilon^2} .
\end{aligned} \tag{7}$$

Now taking the limit for  $n \rightarrow \infty$  on both sides, while considering that  $\epsilon$  and  $\sigma^2$  are finite, gives:

$$\begin{aligned}
\lim_{n \rightarrow \infty} \Pr(|\bar{X} - \mu| < \epsilon) &\geq \lim_{n \rightarrow \infty} \left(1 - \frac{\sigma^2}{n\epsilon^2}\right) \\
&\geq 1 - \lim_{n \rightarrow \infty} \frac{\sigma^2/\epsilon^2}{n} \\
&\geq 1.
\end{aligned} \tag{8}$$

Since the upper limit of probability is one ( $\rightarrow$  I/1.4.7), this implies that:

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X} - \mu| < \epsilon) = 1. \tag{9}$$

■

#### Sources:

- Ostwald, Dirk (2023): “Ungleichungen und Grenzwerte”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (6), Folie 19-20; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/6\\_Ungleichungen\\_und\\_Grenzwerte.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/6_Ungleichungen_und_Grenzwerte.pdf).
- Wikipedia (2024): “Law of large numbers”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-13; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_large\\_numbers#Proof\\_of\\_the\\_weak\\_law](https://en.wikipedia.org/wiki/Law_of_large_numbers#Proof_of_the_weak_law).

#### 1.12.15 Expected value minimizes squared error

**Theorem:** Let  $X_1, \dots, X_n$  be a collection of random variables ( $\rightarrow$  I/1.2.2) with common mean ( $\rightarrow$  I/1.12.1)  $E(X_i) = \mu$ ,  $i = 1, \dots, n$ . Then,  $\mu$  minimizes the mean squared error:

$$\mu = \arg \min_{a \in \mathbb{R}} E[(X_i - a)^2]. \tag{1}$$

**Proof:** Using the linearity of expectation ( $\rightarrow$  I/1.12.5), we can simplify the objective function:

$$E[(X_i - a)^2] = E[X_i^2 - 2aX_i + a^2] = a^2 - 2a\mu + E(X_i^2). \tag{2}$$

Setting the first derivative

$$\frac{d}{da} [a^2 - 2a\mu + E(X_i^2)] = 2a - 2\mu \tag{3}$$

to zero to perform a derivative test, we obtain:

$$2a - 2\mu = 0 \quad \Leftrightarrow \quad a = \mu. \tag{4}$$

The second derivative is equal to 2, which is greater than 0. Since  $a = \mu$  is the sole critical point, we can conclude that this value is the unique global minimum. This completes the proof that the expected value minimizes the mean squared error.

■

#### Sources:

- Wikipedia (2024): “Derivative test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-13; URL: [https://en.wikipedia.org/wiki/Derivative\\_test](https://en.wikipedia.org/wiki/Derivative_test).

### 1.12.16 Expected value of a random vector

**Definition:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Then, the expected value ( $\rightarrow$  I/1.12.1) of  $X$  is an  $n$ -dimensional vector whose entries correspond to the expected values of the entries of the random vector:

$$E(X) = E \left( \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \right) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix} . \quad (1)$$

**Sources:**

- Taboga, Marco (2017): “Expected value”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2021-07-08; URL: <https://www.statlect.com/fundamentals-of-probability/expected-value#hid12>.
- Wikipedia (2021): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-08; URL: [https://en.wikipedia.org/wiki/Multivariate\\_random\\_variable#Expected\\_value](https://en.wikipedia.org/wiki/Multivariate_random_variable#Expected_value).

### 1.12.17 Expectation of a quadratic form

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) with mean ( $\rightarrow$  I/1.12.1)  $\mu$  and covariance ( $\rightarrow$  I/1.16.1)  $\Sigma$  and let  $A$  be a symmetric  $n \times n$  matrix. Then, the expectation of the quadratic form  $X^T A X$  is

$$E [X^T A X] = \mu^T A \mu + \text{tr}(A \Sigma) . \quad (1)$$

**Proof:** Note that  $X^T A X$  is a  $1 \times 1$  matrix. We can therefore write

$$E [X^T A X] = E [\text{tr} (X^T A X)] . \quad (2)$$

Using the trace property  $\text{tr}(ABC) = \text{tr}(BCA)$ , this becomes

$$E [X^T A X] = E [\text{tr} (A X X^T)] . \quad (3)$$

Because mean and trace are linear operators ( $\rightarrow$  I/1.12.20), we have

$$E [X^T A X] = \text{tr} (A E [X X^T]) . \quad (4)$$

Note that the covariance matrix can be partitioned into expected values ( $\rightarrow$  I/1.16.3)

$$\text{Cov}(X, X) = E(X X^T) - E(X)E(X)^T , \quad (5)$$

such that the expected value of the quadratic form becomes

$$E [X^T A X] = \text{tr} (A [\text{Cov}(X, X) + E(X)E(X)^T]) . \quad (6)$$

Finally, applying mean ( $\rightarrow$  I/1.12.1) and covariance ( $\rightarrow$  I/1.16.1) of  $X$ , we have

$$\begin{aligned}
E[X^T A X] &= \text{tr}(A[\Sigma + \mu\mu^T]) \\
&= \text{tr}(A\Sigma + A\mu\mu^T) \\
&= \text{tr}(A\Sigma) + \text{tr}(A\mu\mu^T) \\
&= \text{tr}(A\Sigma) + \text{tr}(\mu^T A \mu) \\
&= \mu^T A \mu + \text{tr}(A\Sigma) .
\end{aligned} \tag{7}$$

■

**Sources:**

- Kendrick, David (1981): “Expectation of a quadratic form”; in: *Stochastic Control for Economic Models*, pp. 170-171.
- Wikipedia (2020): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-13; URL: [https://en.wikipedia.org/wiki/Multivariate\\_random\\_variable#Expectation\\_of\\_a\\_quadratic\\_form](https://en.wikipedia.org/wiki/Multivariate_random_variable#Expectation_of_a_quadratic_form).
- Halvorsen, Kjetil B. (2012): “Expected value and variance of trace function”; in: *StackExchange Cross Validated*, retrieved on 2020-07-13; URL: <https://stats.stackexchange.com/questions/34477/expected-value-and-variance-of-trace-function>.
- Sarwate, Dilip (2013): “Expected Value of Quadratic Form”; in: *StackExchange Cross Validated*, retrieved on 2020-07-13; URL: <https://stats.stackexchange.com/questions/48066/expected-value-of-quadratic-form>.

**1.12.18 Expectation of a bilinear form**

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) with mean ( $\rightarrow$  I/1.12.1)  $\mu_X$  and let  $Y$  be an  $m$ -dimensional random vector ( $\rightarrow$  I/1.2.3) with mean ( $\rightarrow$  I/1.12.1)  $\mu_Y$  where  $X$  and  $Y$  can or cannot be of equal of size ( $n = m$  or  $n \neq m$ ). Further, let  $A$  be an  $n \times m$  matrix and let  $\Sigma_{XY}$  be the cross-covariance matrix ( $\rightarrow$  I/1.16.9) of  $X$  and  $Y$ . Then, the expectation of the bilinear form  $X^T A Y$  is

$$E[X^T A Y] = \mu_X^T A \mu_Y + \text{tr}(A^T \Sigma_{XY}) . \tag{1}$$

**Proof:** Note that  $X^T A Y = Y^T A^T X$  is a  $1 \times 1$  matrix. We can therefore write

$$E[X^T A Y] = E[\text{tr}(Y^T A^T X)] . \tag{2}$$

Using the trace property  $\text{tr}(ABC) = \text{tr}(BCA)$ , this becomes

$$E[X^T A Y] = E[\text{tr}(A^T X Y^T)] . \tag{3}$$

Because mean and trace are linear operators ( $\rightarrow$  I/1.12.20), we have

$$E[X^T A Y] = \text{tr}(A^T E[XY^T]) . \tag{4}$$

Note that the cross-covariance matrix can be partitioned into expected values ( $\rightarrow$  I/1.16.10)

$$\text{Cov}(X, Y) = E(XY^T) - E(X)E(Y)^T , \tag{5}$$

such that the expected value of the quadratic form becomes

$$E[X^T A Y] = \text{tr}(A[\text{Cov}(X, Y) + E(X)E(Y)^T]) . \tag{6}$$

Finally, applying the means ( $\rightarrow$  I/1.12.1) and cross-covariance ( $\rightarrow$  I/1.16.9) of  $X$  and  $Y$ , we have

$$\begin{aligned}
 E[X^T AY] &= \text{tr}(A^T [\Sigma_{XY} + \mu_X \mu_Y^T]) \\
 &= \text{tr}(A^T \Sigma_{XY} + A^T \mu_X \mu_Y^T) \\
 &= \text{tr}(A^T \Sigma_{XY}) + \text{tr}(A^T \mu_X \mu_Y^T) \\
 &= \text{tr}(A^T \Sigma_{XY}) + \text{tr}(\mu_Y^T A^T \mu_X) \\
 &= \text{tr}(A^T \Sigma_{XY}) + \text{tr}(\mu_X^T A \mu_Y) \\
 &= \mu_X^T A \mu_Y + \text{tr}(A^T \Sigma_{XY}) .
 \end{aligned} \tag{7}$$

■

### 1.12.19 Expected value of a random matrix

**Definition:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4). Then, the expected value ( $\rightarrow$  I/1.12.1) of  $X$  is an  $n \times p$  matrix whose entries correspond to the expected values of the entries of the random matrix:

$$E(X) = E \left( \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix} \right) = \begin{bmatrix} E(X_{11}) & \dots & E(X_{1p}) \\ \vdots & \ddots & \vdots \\ E(X_{n1}) & \dots & E(X_{np}) \end{bmatrix} . \tag{1}$$

#### Sources:

- Taboga, Marco (2017): “Expected value”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2021-07-08; URL: <https://www.statlect.com/fundamentals-of-probability/expected-value#hid13>.

### 1.12.20 Expectation of a trace

**Theorem:** Let  $A$  be an  $n \times n$  random matrix ( $\rightarrow$  I/1.2.4). Then, the expectation ( $\rightarrow$  I/1.12.1) of the trace of  $A$  is equal to the trace of the expectation ( $\rightarrow$  I/1.12.1) of  $A$ :

$$E[\text{tr}(A)] = \text{tr}(E[A]) . \tag{1}$$

**Proof:** The trace of an  $n \times n$  matrix  $A$  is defined as:

$$\text{tr}(A) = \sum_{i=1}^n a_{ii} . \tag{2}$$

Using this definition of the trace, the linearity of the expected value ( $\rightarrow$  I/1.12.5) and the expected value of a random matrix ( $\rightarrow$  I/1.12.19), we have:



$$\begin{aligned}
\mathbb{E}[\text{tr}(A)] &= \mathbb{E} \left[ \sum_{i=1}^n a_{ii} \right] \\
&= \sum_{i=1}^n \mathbb{E}[a_{ii}] \\
&= \text{tr} \left( \begin{bmatrix} \mathbb{E}[a_{11}] & \dots & \mathbb{E}[a_{1n}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[a_{n1}] & \dots & \mathbb{E}[a_{nn}] \end{bmatrix} \right) \\
&= \text{tr}(\mathbb{E}[A]) .
\end{aligned} \tag{3}$$

■

**Sources:**

- drerD (2018): “‘Trace trick’ for expectations of quadratic forms”; in: *StackExchange Mathematics*, retrieved on 2021-12-07; URL: <https://math.stackexchange.com/a/3004034/480910>.

**1.13 Variance****1.13.1 Definition**

**Definition:** The variance of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as the expected value ( $\rightarrow$  I/1.12.1) of the squared deviation from its expected value ( $\rightarrow$  I/1.12.1):

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] . \tag{1}$$

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

**1.13.2 Sample variance**

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the sample variance of  $x$  is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \tag{1}$$

and the unbiased sample variance of  $x$  is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \tag{2}$$

where  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Sources:**

- Wikipedia (2021): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-16; URL: [https://en.wikipedia.org/wiki/Variance#Sample\\_variance](https://en.wikipedia.org/wiki/Variance#Sample_variance).

### 1.13.3 Pooled sample variance

**Definition:** Let  $x_i = \{x_{i1}, \dots, x_{in_i}\}$  for  $i = 1, \dots, k$  be samples ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$  whose expected value ( $\rightarrow$  I/1.12.1), but not variance ( $\rightarrow$  I/1.13.1) potentially depends on sample index  $i$ . Then, the pooled sample variance ( $\rightarrow$  I/1.13.2) of  $x = \{x_1, \dots, x_k\}$  is given by

$$s_{1\dots k}^2 = \frac{1}{n - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad (1)$$

where  $\bar{x}_i$  is the sample mean ( $\rightarrow$  I/1.12.2) of the  $i$ -th sample

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad (2)$$

and  $n$  is the total sample size ( $\rightarrow$  I/1.1.7):

$$n = \sum_{i=1}^k n_i. \quad (3)$$

#### Sources:

- Soch, Joram; Ostwald, Dirk (2024): “Einfaktorielle Varianzanalyse”; in: *Allgemeines Lineares Modell*, Einheit (10), Folien 27-29; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2024/Allgemeines+Lineares+Modell/10\\_Einfaktorielle\\_Varianzanalyse.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2024/Allgemeines+Lineares+Modell/10_Einfaktorielle_Varianzanalyse.pdf).
- Wikipedia (2025): “Pooled variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Pooled\\_variance#Definition\\_and\\_computation](https://en.wikipedia.org/wiki/Pooled_variance#Definition_and_computation).

### 1.13.4 Partition into expected values

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is equal to the mean ( $\rightarrow$  I/1.12.1) of the square of  $X$  minus the square of the mean ( $\rightarrow$  I/1.12.1) of  $X$ :

$$\text{Var}(X) = E(X^2) - E(X)^2. \quad (1)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) of  $X$  is defined as

$$\text{Var}(X) = E[(X - E[X])^2] \quad (2)$$

which, due to the linearity of the expected value ( $\rightarrow$  I/1.12.5), can be rewritten as

$$\begin{aligned} \text{Var}(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2X E(X) + E(X)^2] \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned} \quad (3)$$

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-19; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

**1.13.5 Non-negativity**

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) is always non-negative, i.e.

$$\text{Var}(X) \geq 0 . \quad (1)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) of a random variable ( $\rightarrow$  I/1.2.2) is defined as

$$\text{Var}(X) = \text{E} [(X - \text{E}(X))^2] . \quad (2)$$

1) If  $X$  is a discrete random variable ( $\rightarrow$  I/1.2.2), then, because squares and probabilities are strictly non-negative, all the addends in

$$\text{Var}(X) = \sum_{x \in \mathcal{X}} (x - \text{E}(X))^2 \cdot f_X(x) \quad (3)$$

are also non-negative, thus the entire sum must be non-negative.

2) If  $X$  is a continuous random variable ( $\rightarrow$  I/1.2.2), then, because squares and probability densities are strictly non-negative, the integrand in

$$\text{Var}(X) = \int_{\mathcal{X}} (x - \text{E}(X))^2 \cdot f_X(x) \, dx \quad (4)$$

is always non-negative, thus the term on the right-hand side is a Lebesgue integral, so that the result on the left-hand side must be non-negative.

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

**1.13.6 Variance of a constant**

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) of a constant ( $\rightarrow$  I/1.2.5) is zero

$$a = \text{const.} \quad \Rightarrow \quad \text{Var}(a) = 0 \quad (1)$$

and if the variance ( $\rightarrow$  I/1.13.1) of  $X$  is zero, then  $X$  is a constant ( $\rightarrow$  I/1.2.5)

$$\text{Var}(X) = 0 \quad \Rightarrow \quad X = \text{const.} \quad (2)$$

**Proof:**

1) A constant ( $\rightarrow$  I/1.2.5) is defined as a quantity that always has the same value. Thus, if understood as a random variable ( $\rightarrow$  I/1.2.2), the expected value ( $\rightarrow$  I/1.12.1) of a constant is equal to itself:

$$E(a) = a . \quad (3)$$

Plugged into the formula of the variance ( $\rightarrow$  I/1.13.1), we have

$$\begin{aligned} \text{Var}(a) &= E [(a - E(a))^2] \\ &= E [(a - a)^2] \\ &= E(0) . \end{aligned} \quad (4)$$

Applied to the formula of the expected value ( $\rightarrow$  I/1.12.1), this gives

$$E(0) = \sum_{x=0} x \cdot f_X(x) = 0 \cdot 1 = 0 . \quad (5)$$

Together, (4) and (5) imply (1).

2) The variance ( $\rightarrow$  I/1.13.1) is defined as

$$\text{Var}(X) = E [(X - E(X))^2] . \quad (6)$$

Because  $(X - E(X))^2$  is strictly non-negative ( $\rightarrow$  I/1.12.4), the only way for the variance to become zero is, if the squared deviation is always zero:

$$(X - E(X))^2 = 0 . \quad (7)$$

This, in turn, requires that  $X$  is equal to its expected value ( $\rightarrow$  I/1.12.1)

$$X = E(X) \quad (8)$$

which can only be the case, if  $X$  always has the same value ( $\rightarrow$  I/1.2.5):

$$X = \text{const.} \quad (9)$$

■

#### Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-27; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

### 1.13.7 Invariance under addition

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) is invariant under addition of a constant ( $\rightarrow$  I/1.2.5):

$$\text{Var}(X + a) = \text{Var}(X) \quad (1)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is defined in terms of the expected value ( $\rightarrow$  I/1.12.1) as

$$\text{Var}(X) = E [(X - E(X))^2] . \quad (2)$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5), we can derive (1) as follows:

$$\begin{aligned}
\text{Var}(X + a) &\stackrel{(2)}{=} \text{E} [((X + a) - \text{E}(X + a))^2] \\
&= \text{E} [(X + a - \text{E}(X) - a)^2] \\
&= \text{E} [(X - \text{E}(X))^2] \\
&\stackrel{(2)}{=} \text{Var}(X) .
\end{aligned} \tag{3}$$

■

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

**1.13.8 Scaling upon multiplication**

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) scales upon multiplication with a constant ( $\rightarrow$  I/1.2.5):

$$\text{Var}(aX) = a^2 \text{Var}(X) \tag{1}$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is defined in terms of the expected value ( $\rightarrow$  I/1.12.1) as

$$\text{Var}(X) = \text{E} [(X - \text{E}(X))^2] . \tag{2}$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5), we can derive (1) as follows:

$$\begin{aligned}
\text{Var}(aX) &\stackrel{(2)}{=} \text{E} [((aX) - \text{E}(aX))^2] \\
&= \text{E} [(aX - a\text{E}(X))^2] \\
&= \text{E} [(a[X - \text{E}(X)])^2] \\
&= \text{E} [a^2(X - \text{E}(X))^2] \\
&= a^2 \text{E} [(X - \text{E}(X))^2] \\
&\stackrel{(2)}{=} a^2 \text{Var}(X) .
\end{aligned} \tag{3}$$

■

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

**1.13.9 Variance of a sum**

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) of the sum of two random variables ( $\rightarrow$  I/1.2.2) equals the sum of the variances of those random variables, plus two times their covariance ( $\rightarrow$  I/1.15.1):

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \tag{1}$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is defined in terms of the expected value ( $\rightarrow$  I/1.12.1) as

$$\text{Var}(X) = E[(X - E(X))^2] . \quad (2)$$

Using the linearity of the expected value ( $\rightarrow$  I/1.12.5) and the definition of the covariance ( $\rightarrow$  I/1.15.1), we can derive (1) as follows:

$$\begin{aligned} \text{Var}(X + Y) &\stackrel{(2)}{=} E[((X + Y) - E(X + Y))^2] \\ &= E[(X - E(X) + Y - E(Y))^2] \\ &= E[(X - E(X))^2 + (Y - E(Y))^2 + 2(X - E(X))(Y - E(Y))] \\ &= E[(X - E(X))^2] + E[(Y - E(Y))^2] + E[2(X - E(X))(Y - E(Y))] \\ &\stackrel{(2)}{=} \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \end{aligned} \quad (3)$$

■

#### Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

#### 1.13.10 Variance of linear combination

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) of the linear combination of two random variables ( $\rightarrow$  I/1.2.2) is a function of the variances as well as the covariance ( $\rightarrow$  I/1.15.1) of those random variables:

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) . \quad (1)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is defined in terms of the expected value ( $\rightarrow$  I/1.12.1) as

$$\text{Var}(X) = E[(X - E(X))^2] . \quad (2)$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5), we can derive (1) as follows:

$$\begin{aligned} \text{Var}(aX + bY) &\stackrel{(2)}{=} E[((aX + bY) - E(aX + bY))^2] \\ &= E[(a(X - E(X)) + b(Y - E(Y)))^2] \\ &= E[a^2(X - E(X))^2 + b^2(Y - E(Y))^2 + 2ab(X - E(X))(Y - E(Y))] \\ &= E[a^2(X - E(X))^2] + E[b^2(Y - E(Y))^2] + E[2ab(X - E(X))(Y - E(Y))] \\ &\stackrel{(2)}{=} a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) . \end{aligned} \quad (3)$$

■

#### Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

**1.13.11 Additivity under independence**

**Theorem:** The variance ( $\rightarrow$  I/1.13.1) is additive for independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2):

$$p(X, Y) = p(X)p(Y) \quad \Rightarrow \quad \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) . \quad (1)$$

**Proof:** The variance of the sum of two random variables ( $\rightarrow$  I/1.13.9) is given by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \quad (2)$$

The covariance of independent random variables ( $\rightarrow$  I/1.15.6) is zero:

$$p(X, Y) = p(X)p(Y) \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 . \quad (3)$$

Combining (2) and (3), we have:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) . \quad (4)$$

■

**Sources:**

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: [https://en.wikipedia.org/wiki/Variance#Basic\\_properties](https://en.wikipedia.org/wiki/Variance#Basic_properties).

**1.13.12 Law of total variance**

**Theorem:** (law of total variance, also called “conditional variance formula”) Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2) defined on the same probability space ( $\rightarrow$  I/1.1.4) and assume that the variance ( $\rightarrow$  I/1.13.1) of  $Y$  is finite. Then, the sum of the expectation ( $\rightarrow$  I/1.12.1) of the conditional variance and the variance ( $\rightarrow$  I/1.13.1) of the conditional expectation of  $Y$  given  $X$  is equal to the variance ( $\rightarrow$  I/1.13.1) of  $Y$ :

$$\text{Var}(Y) = \text{E}[\text{Var}(Y|X)] + \text{Var}[\text{E}(Y|X)] . \quad (1)$$

**Proof:** The variance can be decomposed into expected values ( $\rightarrow$  I/1.13.4) as follows:

$$\text{Var}(Y) = \text{E}(Y^2) - \text{E}(Y)^2 . \quad (2)$$

This can be rearranged into:

$$\text{E}(Y^2) = \text{Var}(Y) + \text{E}(Y)^2 . \quad (3)$$

Applying the law of total expectation ( $\rightarrow$  I/1.12.8), we have:

$$\text{E}(Y^2) = \text{E} [\text{Var}(Y|X) + \text{E}(Y|X)^2] . \quad (4)$$

Now subtract the second term from (2):

$$\text{E}(Y^2) - \text{E}(Y)^2 = \text{E} [\text{Var}(Y|X) + \text{E}(Y|X)^2] - \text{E}(Y)^2 . \quad (5)$$

Again applying the law of total expectation ( $\rightarrow$  I/1.12.8), we have:

$$E(Y^2) - E(Y)^2 = E [\text{Var}(Y|X) + E(Y|X)^2] - E [E(Y|X)]^2 . \quad (6)$$

With the linearity of the expected value ( $\rightarrow$  I/1.12.5), the terms can be regrouped to give:

$$E(Y^2) - E(Y)^2 = E [\text{Var}(Y|X)] + (E [E(Y|X)^2] - E [E(Y|X)]^2) . \quad (7)$$

Using the decomposition of variance into expected values ( $\rightarrow$  I/1.13.4), we finally have:

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)] . \quad (8)$$

■

#### Sources:

- Wikipedia (2021): “Law of total variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_total\\_variance#Proof](https://en.wikipedia.org/wiki/Law_of_total_variance#Proof).

### 1.13.13 Precision

**Definition:** The precision of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as the inverse of the variance ( $\rightarrow$  I/1.13.1), i.e. one divided by the expected value ( $\rightarrow$  I/1.12.1) of the squared deviation from its expected value ( $\rightarrow$  I/1.12.1):

$$\text{Prec}(X) = \text{Var}(X)^{-1} = \frac{1}{E [(X - E(X))^2]} . \quad (1)$$

#### Sources:

- Wikipedia (2020): “Precision (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-21; URL: [https://en.wikipedia.org/wiki/Precision\\_\(statistics\)](https://en.wikipedia.org/wiki/Precision_(statistics)).

## 1.14 Skewness

### 1.14.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $\mu$  and standard deviation ( $\rightarrow$  I/1.19.1)  $\sigma$ . Then, the skewness of  $X$  is defined as the third standardized moment ( $\rightarrow$  I/1.21.9) of  $X$ :

$$\text{Skew}(X) = \frac{E[(X - \mu)^3]}{\sigma^3} . \quad (1)$$

#### Sources:

- Wikipedia (2023): “Skewness”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-04-20; URL: <https://en.wikipedia.org/wiki/Skewness>.



### 1.14.2 Sample skewness

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the sample skewness of  $x$  is given by

$$\hat{s} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}, \quad (1)$$

where  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Sources:**

- Joanes, D. N. and Gill, C. A. (1998): “Comparing measures of sample skewness and kurtosis”; in: *The Statistician*, vol. 47, part 1, pp. 183-189; URL: <https://www.jstor.org/stable/2988433>.

### 1.14.3 Partition into expected values

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $\mu$  and standard deviation ( $\rightarrow$  I/1.19.1)  $\sigma$ . Then, the skewness ( $\rightarrow$  I/1.14.1) of  $X$  can be computed as:

$$\text{Skew}(X) = \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \quad (1)$$

**Proof:** The skewness ( $\rightarrow$  I/1.14.1) of  $X$  is defined as

$$\text{Skew}(X) = \frac{E[(X - \mu)^3]}{\sigma^3}. \quad (2)$$

Because the expected value is a linear operator ( $\rightarrow$  I/1.12.5), we can rewrite (2) as

$$\begin{aligned} \text{Skew}(X) &= \frac{E[(X - \mu)^3]}{\sigma^3} \\ &= \frac{E[X^3 - 3X^2\mu + 3X\mu^2 - \mu^3]}{\sigma^3} \\ &= \frac{E(X^3) - 3E(X^2)\mu + 3E(X)\mu^2 - \mu^3}{\sigma^3} \\ &= \frac{E(X^3) - 3\mu[E(X^2) - E(X)\mu] - \mu^3}{\sigma^3} \\ &= \frac{E(X^3) - 3\mu[E(X^2) - E(X)^2] - \mu^3}{\sigma^3}. \end{aligned} \quad (3)$$

Because the variance can be written in terms of expected values ( $\rightarrow$  I/1.13.4) as

$$\sigma^2 = E(X^2) - E(X)^2, \quad (4)$$

we can rewrite (3) as

$$\text{Skew}(X) = \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \quad (5)$$

This finishes the proof of (1). ■

**Sources:**

- Wikipedia (2023): “Skewness”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-04-20; URL: <https://en.wikipedia.org/wiki/Skewness>.

**1.15 Covariance****1.15.1 Definition**

**Definition:** The covariance of two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as the expected value ( $\rightarrow$  I/1.12.1) of the product of their deviations from their individual expected values ( $\rightarrow$  I/1.12.1):

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

**1.15.2 Sample covariance**

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  and  $y = \{y_1, \dots, y_n\}$  be samples ( $\rightarrow$  I/1.1.6) from random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ . Then, the sample covariance of  $x$  and  $y$  is given by

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (1)$$

and the unbiased sample covariance of  $x$  and  $y$  is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2).

**Sources:**

- Wikipedia (2021): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: [https://en.wikipedia.org/wiki/Covariance#Calculating\\_the\\_sample\\_covariance](https://en.wikipedia.org/wiki/Covariance#Calculating_the_sample_covariance).

**1.15.3 Partition into expected values**

**Theorem:** Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2). Then, the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  is equal to the mean ( $\rightarrow$  I/1.12.1) of the product of  $X$  and  $Y$  minus the product of the means ( $\rightarrow$  I/1.12.1) of  $X$  and  $Y$ :

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) . \quad (1)$$

**Proof:** The covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  is defined as

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (2)$$

which, due to the linearity of the expected value ( $\rightarrow$  I/1.12.5), can be rewritten as

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - X E(Y) - E(X) Y + E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) . \end{aligned} \quad (3)$$

■

#### Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-02; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

#### 1.15.4 Symmetry

**Theorem:** The covariance ( $\rightarrow$  I/1.15.1) of two random variables ( $\rightarrow$  I/1.2.2) is a symmetric function:

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) . \quad (1)$$

**Proof:** The covariance ( $\rightarrow$  I/1.15.1) of random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (2)$$

Switching  $X$  and  $Y$  in (2), we can easily see:

$$\begin{aligned} \text{Cov}(Y, X) &\stackrel{(2)}{=} E[(Y - E[Y])(X - E[X])] \\ &= E[(X - E[X])(Y - E[Y])] \\ &= \text{Cov}(X, Y) . \end{aligned} \quad (3)$$

■

#### Sources:

- Wikipedia (2022): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Covariance#Covariance\\_of\\_linear\\_combinations](https://en.wikipedia.org/wiki/Covariance#Covariance_of_linear_combinations).

#### 1.15.5 Self-covariance

**Theorem:** The covariance ( $\rightarrow$  I/1.15.1) of a random variable ( $\rightarrow$  I/1.2.2) with itself is equal to the variance ( $\rightarrow$  I/1.13.1):

$$\text{Cov}(X, X) = \text{Var}(X) . \quad (1)$$

**Proof:** The covariance ( $\rightarrow$  I/1.15.1) of random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (2)$$

Inserting  $X$  for  $Y$  in (2), the result is the variance ( $\rightarrow$  I/1.13.1) of  $X$ :

$$\begin{aligned}\text{Cov}(X, X) &\stackrel{(2)}{=} \text{E}[(X - \text{E}[X])(X - \text{E}[X])] \\ &= \text{E}[(X - \text{E}[X])^2] \\ &= \text{Var}(X) .\end{aligned}\tag{3}$$

■

**Sources:**

- Wikipedia (2022): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Covariance#Covariance\\_with\\_itself](https://en.wikipedia.org/wiki/Covariance#Covariance_with_itself).

### 1.15.6 Covariance under independence

**Theorem:** Let  $X$  and  $Y$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2). Then, the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  is zero:

$$X, Y \text{ independent} \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 .\tag{1}$$

**Proof:** The covariance can be expressed in terms of expected values ( $\rightarrow$  I/1.15.3) as

$$\text{Cov}(X, Y) = \text{E}(X Y) - \text{E}(X) \text{E}(Y) .\tag{2}$$

For independent random variables, the expected value of the product is equal to the product of the expected values ( $\rightarrow$  I/1.12.7):

$$\text{E}(X Y) = \text{E}(X) \text{E}(Y) .\tag{3}$$

Taking (2) and (3) together, we have

$$\begin{aligned}\text{Cov}(X, Y) &\stackrel{(2)}{=} \text{E}(X Y) - \text{E}(X) \text{E}(Y) \\ &\stackrel{(3)}{=} \text{E}(X) \text{E}(Y) - \text{E}(X) \text{E}(Y) \\ &= 0 .\end{aligned}\tag{4}$$

■

**Sources:**

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: [https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness\\_and\\_independence](https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence).

### 1.15.7 Law of total covariance

**Theorem:** (law of total covariance, also called “conditional covariance formula”) Let  $X$ ,  $Y$  and  $Z$  be random variables ( $\rightarrow$  I/1.2.2) defined on the same probability space ( $\rightarrow$  I/1.1.4) and assume that the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  is finite. Then, the sum of the expectation ( $\rightarrow$  I/1.12.1) of the conditional covariance and the covariance ( $\rightarrow$  I/1.15.1) of the conditional expectations of  $X$  and  $Y$  given  $Z$  is equal to the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$ :

$$\text{Cov}(X, Y) = \text{E}[\text{Cov}(X, Y|Z)] + \text{Cov}[\text{E}(X|Z), \text{E}(Y|Z)] . \quad (1)$$

**Proof:** The covariance can be decomposed into expected values ( $\rightarrow$  I/1.15.3) as follows:

$$\text{Cov}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) . \quad (2)$$

Then, conditioning on  $Z$  and applying the law of total expectation ( $\rightarrow$  I/1.12.8), we have:

$$\text{Cov}(X, Y) = \text{E}[\text{E}(XY|Z)] - \text{E}[\text{E}(X|Z)] \text{E}[\text{E}(Y|Z)] . \quad (3)$$

Applying the decomposition of covariance into expected values ( $\rightarrow$  I/1.15.3) to the first term gives:

$$\text{Cov}(X, Y) = \text{E}[\text{Cov}(X, Y|Z) + \text{E}(X|Z)\text{E}(Y|Z)] - \text{E}[\text{E}(X|Z)] \text{E}[\text{E}(Y|Z)] . \quad (4)$$

With the linearity of the expected value ( $\rightarrow$  I/1.12.5), the terms can be regrouped to give:

$$\text{Cov}(X, Y) = \text{E}[\text{Cov}(X, Y|Z)] + (\text{E}[\text{E}(X|Z)\text{E}(Y|Z)] - \text{E}[\text{E}(X|Z)] \text{E}[\text{E}(Y|Z)]) . \quad (5)$$

Once more using the decomposition of covariance into expected values ( $\rightarrow$  I/1.15.3), we finally have:

$$\text{Cov}(X, Y) = \text{E}[\text{Cov}(X, Y|Z)] + \text{Cov}[\text{E}(X|Z), \text{E}(Y|Z)] . \quad (6)$$

■

#### Sources:

- Wikipedia (2021): “Law of total covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: [https://en.wikipedia.org/wiki/Law\\_of\\_total\\_covariance#Proof](https://en.wikipedia.org/wiki/Law_of_total_covariance#Proof).

### 1.15.8 Relationship to correlation

**Theorem:** Let  $X$  and  $Y$  be random variables ( $\rightarrow$  I/1.2.2). Then, the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  is equal to the product of their correlation ( $\rightarrow$  I/1.17.1) and the standard deviations ( $\rightarrow$  I/1.19.1) of  $X$  and  $Y$ :

$$\text{Cov}(X, Y) = \text{Corr}(X, Y) \sigma_X \sigma_Y . \quad (1)$$

**Proof:** The correlation ( $\rightarrow$  I/1.17.1) of  $X$  and  $Y$  is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} . \quad (2)$$

which can be rearranged for the covariance ( $\rightarrow$  I/1.15.1) to give

$$\text{Cov}(X, Y) = \text{Corr}(X, Y) \sigma_X \sigma_Y . \quad (3)$$

■

## 1.16 Covariance matrix

### 1.16.1 Definition

**Definition:** Let  $X = [X_1, \dots, X_n]^T$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the covariance matrix of  $X$  is defined as the  $n \times n$  matrix in which the entry  $(i, j)$  is the covariance ( $\rightarrow$  I/1.15.1) of  $X_i$  and  $X_j$ :

$$\Sigma_{XX} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix} \quad (1)$$

**Sources:**

- Wikipedia (2020): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Definition](https://en.wikipedia.org/wiki/Covariance_matrix#Definition).

### 1.16.2 Sample covariance matrix

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^{p \times 1}$ . Then, the sample covariance matrix of  $x$  is given by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad (1)$$

and the unbiased sample covariance matrix of  $x$  is given by

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad (2)$$

where  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Sources:**

- Wikipedia (2021): “Sample mean and covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: [https://en.wikipedia.org/wiki/Sample\\_mean\\_and\\_covariance#Definition\\_of\\_sample\\_covariance](https://en.wikipedia.org/wiki/Sample_mean_and_covariance#Definition_of_sample_covariance).

### 1.16.3 Partition into expected values

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  is equal to the mean ( $\rightarrow$  I/1.12.1) of the outer product of  $X$  with itself minus the outer product of the mean ( $\rightarrow$  I/1.12.1) of  $X$  with itself:

$$\Sigma_{XX} = E(XX^T) - E(X)E(X)^T. \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  is defined as

$$\Sigma_{XX} = \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix} \quad (2)$$

which can also be expressed using matrix multiplication as

$$\Sigma_{XX} = E[(X - E[X])(X - E[X])^T] . \quad (3)$$

Due to the linearity of the expected value ( $\rightarrow$  I/1.12.5), this can be rewritten as

$$\begin{aligned} \Sigma_{XX} &= E[(X - E[X])(X - E[X])^T] \\ &= E[XX^T - XE(X)^T - E(X)X^T + E(X)E(X)^T] \\ &= E(XX^T) - E(X)E(X)^T - E(X)E(X)^T + E(X)E(X)^T \\ &= E(XX^T) - E(X)E(X)^T . \end{aligned} \quad (4)$$

■

#### Sources:

- Taboga, Marco (2010): “Covariance matrix”; in: *Lectures on probability and statistics*, retrieved on 2020-06-06; URL: <https://www.statlect.com/fundamentals-of-probability/covariance-matrix>.

#### 1.16.4 Symmetry

**Theorem:** Each covariance matrix ( $\rightarrow$  I/1.16.1) is symmetric:

$$\Sigma_{XX}^T = \Sigma_{XX} . \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of a random vector ( $\rightarrow$  I/1.2.3)  $X$  is defined as

$$\Sigma_{XX} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} . \quad (2)$$

A symmetric matrix is a matrix whose transpose is equal to itself. The transpose of  $\Sigma_{XX}$  is

$$\Sigma_{XX}^T = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_n, X_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_n) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} . \quad (3)$$

Because the covariance is a symmetric function ( $\rightarrow$  I/1.15.4), i.e.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ , this matrix is equal to

$$\Sigma_{XX}^T = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} \quad (4)$$

which is equivalent to our original definition in (2). ■

**Sources:**

- Wikipedia (2022): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Basic\\_properties](https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties).

### 1.16.5 Positive semi-definiteness

**Theorem:** Each covariance matrix ( $\rightarrow$  I/1.16.1) is positive semi-definite:

$$a^T \Sigma_{XX} a \geq 0 \quad \text{for all } a \in \mathbb{R}^n. \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  can be expressed ( $\rightarrow$  I/1.16.3) in terms of expected values ( $\rightarrow$  I/1.12.1) as follows

$$\Sigma_{XX} = \Sigma(X) = E[(X - E[X])(X - E[X])^T] \quad (2)$$

A positive semi-definite matrix is a matrix whose eigenvalues are all non-negative or, equivalently,

$$M \text{ pos. semi-def.} \Leftrightarrow x^T M x \geq 0 \quad \text{for all } x \in \mathbb{R}^n. \quad (3)$$

Here, for an arbitrary real column vector  $a \in \mathbb{R}^n$ , we have:

$$a^T \Sigma_{XX} a \stackrel{(2)}{=} a^T E[(X - E[X])(X - E[X])^T] a. \quad (4)$$

Because the expected value is a linear operator ( $\rightarrow$  I/1.12.5), we can write:

$$a^T \Sigma_{XX} a = E[a^T (X - E[X])(X - E[X])^T a]. \quad (5)$$

Now define the scalar random variable ( $\rightarrow$  I/1.2.2)

$$Y = a^T (X - \mu_X). \quad (6)$$

where  $\mu_X = E[X]$  and note that

$$a^T (X - \mu_X) = (X - \mu_X)^T a. \quad (7)$$

Thus, combining (5) with (6), we have:

$$a^T \Sigma_{XX} a = E[Y^2]. \quad (8)$$

Because  $Y^2$  is a random variable that cannot become negative and the expected value of a strictly non-negative random variable is also non-negative ( $\rightarrow$  I/1.12.4), we finally have

$$a^T \Sigma_{XX} a \geq 0 \quad (9)$$

for any  $a \in \mathbb{R}^n$ . ■

**Sources:**



- hkBattousai (2013): “What is the proof that covariance matrices are always semi-definite?”; in: *StackExchange Mathematics*, retrieved on 2022-09-26; URL: <https://math.stackexchange.com/a/327872>.
- Wikipedia (2022): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Basic\\_properties](https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties).

### 1.16.6 Invariance under addition of vector

**Theorem:** The covariance matrix ( $\rightarrow$  I/1.16.1)  $\Sigma_{XX}$  of a random vector ( $\rightarrow$  I/1.2.3)  $X$  is invariant under addition of a constant vector ( $\rightarrow$  I/1.2.5)  $a$ :

$$\Sigma(X + a) = \Sigma(X) . \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  can be expressed ( $\rightarrow$  I/1.16.3) in terms of expected values ( $\rightarrow$  I/1.12.1) as follows:

$$\Sigma_{XX} = \Sigma(X) = E [(X - E[X])(X - E[X])^T] . \quad (2)$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5), we can derive (1) as follows:

$$\begin{aligned} \Sigma(X + a) &\stackrel{(2)}{=} E [([X + a] - E[X + a])([X + a] - E[X + a])^T] \\ &= E [(X + a - E[X] - a)(X + a - E[X] - a)^T] \\ &= E [(X - E[X])(X - E[X])^T] \\ &\stackrel{(2)}{=} \Sigma(X) . \end{aligned} \quad (3)$$

■

**Sources:**

- Wikipedia (2022): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Basic\\_properties](https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties).

### 1.16.7 Scaling upon multiplication with matrix

**Theorem:** The covariance matrix ( $\rightarrow$  I/1.16.1)  $\Sigma_{XX}$  of a random vector ( $\rightarrow$  I/1.2.3)  $X$  scales upon multiplication with a constant matrix ( $\rightarrow$  I/1.2.5)  $A$ :

$$\Sigma(AX) = A \Sigma(X) A^T . \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  can be expressed ( $\rightarrow$  I/1.16.3) in terms of expected values ( $\rightarrow$  I/1.12.1) as follows:

$$\Sigma_{XX} = \Sigma(X) = E [(X - E[X])(X - E[X])^T] . \quad (2)$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5), we can derive (1) as follows:

$$\begin{aligned}
\Sigma(AX) &\stackrel{(2)}{=} E [(AX) - E[AX]](AX) - E[AX]]^T] \\
&= E [(A[X - E[X]]) (A[X - E[X]])^T] \\
&= E [A(X - E[X])(X - E[X])^T A^T] \\
&= A E [(X - E[X])(X - E[X])^T] A^T \\
&\stackrel{(2)}{=} A \Sigma(X) A^T .
\end{aligned} \tag{3}$$

■

**Sources:**

- Wikipedia (2022): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Basic\\_properties](https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties).

**1.16.8 Covariance matrix and correlation matrix**

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the covariance matrix ( $\rightarrow$  I/1.16.1)  $\Sigma_{XX}$  of  $X$  can be expressed in terms of its correlation matrix ( $\rightarrow$  I/1.17.6)  $C_{XX}$  as follows

$$\Sigma_{XX} = D_X \cdot C_{XX} \cdot D_X , \tag{1}$$

where  $D_X$  is a diagonal matrix with the standard deviations ( $\rightarrow$  I/1.19.1) of  $X_1, \dots, X_n$  as entries on the diagonal:

$$D_X = \text{diag}(\sigma_{X_1}, \dots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} . \tag{2}$$

**Proof:** Reiterating (1) and applying (2), we have:

$$\Sigma_{XX} = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot C_{XX} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} . \tag{3}$$

Together with the definition of the correlation matrix ( $\rightarrow$  I/1.17.6), this gives

$$\begin{aligned}
\Sigma_{XX} &= \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \\
&= \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix}
\end{aligned} \tag{4}$$

which is nothing else than the definition of the covariance matrix ( $\rightarrow$  I/1.16.1).

■

#### Sources:

- Penny, William (2006): “The correlation matrix”; in: *Mathematics for Brain Imaging*, ch. 1.4.5, p. 28, eq. 1.60; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).

#### 1.16.9 Cross-covariance matrix

**Definition:** Let  $X = [X_1, \dots, X_n]^T$  and  $Y = [Y_1, \dots, Y_m]^T$  be two random vectors ( $\rightarrow$  I/1.2.3) that can or cannot be of equal size. Then, the cross-covariance matrix of  $X$  and  $Y$  is defined as the  $n \times m$  matrix in which the entry  $(i, j)$  is the covariance ( $\rightarrow$  I/1.15.1) of  $X_i$  and  $Y_j$ :

$$\Sigma_{XY} = \begin{bmatrix} \text{Cov}(X_1, Y_1) & \dots & \text{Cov}(X_1, Y_m) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y_1) & \dots & \text{Cov}(X_n, Y_m) \end{bmatrix} = \begin{bmatrix} E[(X_1 - E[X_1])(Y_1 - E[Y_1])] & \dots & E[(X_1 - E[X_1])(Y_m - E[Y_m])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(Y_1 - E[Y_1])] & \dots & E[(X_n - E[X_n])(Y_m - E[Y_m])] \end{bmatrix} \tag{1}$$

#### Sources:

- Wikipedia (2022): “Cross-covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Cross-covariance\\_matrix#Definition](https://en.wikipedia.org/wiki/Cross-covariance_matrix#Definition).

### 1.16.10 Cross-covariance and expected values

**Theorem:** Let  $X$  and  $Y$  be a random vectors ( $\rightarrow$  I/1.2.3). Then, the cross-covariance matrix ( $\rightarrow$  I/1.16.9) of  $X$  and  $Y$  is equal to the mean ( $\rightarrow$  I/1.12.1) of the outer product of  $X$  with  $Y$  minus the outer product of the means ( $\rightarrow$  I/1.12.1) of  $X$  and  $Y$ :

$$\Sigma_{XY} = E(XY^T) - E(X)E(Y)^T. \quad (1)$$

**Proof:** The cross-covariance matrix ( $\rightarrow$  I/1.16.9) of  $X$  and  $Y$  is defined as

$$\Sigma_{XY} = \begin{bmatrix} E[(X_1 - E[X_1])(Y_1 - E[Y_1])] & \dots & E[(X_1 - E[X_1])(Y_m - E[Y_m])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(Y_1 - E[Y_1])] & \dots & E[(X_n - E[X_n])(Y_m - E[Y_m])] \end{bmatrix} \quad (2)$$

which can also be expressed using matrix multiplication as

$$\Sigma_{XY} = E[(X - E[X])(Y - E[Y])^T]. \quad (3)$$

Due to the linearity of the expected value ( $\rightarrow$  I/1.12.5), this can be rewritten as

$$\begin{aligned} \Sigma_{XY} &= E[(X - E[X])(Y - E[Y])^T] \\ &= E[XY^T - XE(Y)^T - E(X)Y^T + E(X)E(Y)^T] \\ &= E(XY^T) - E(X)E(Y)^T - E(X)E(Y)^T + E(X)E(Y)^T \\ &= E(XY^T) - E(X)E(Y)^T. \end{aligned} \quad (4)$$

■

### 1.16.11 Covariance matrix of a sum

**Theorem:** The covariance matrix ( $\rightarrow$  I/1.16.1) of the sum of two random vectors ( $\rightarrow$  I/1.2.3) of the same dimension equals the sum of the covariances of those random vectors, plus the sum of their cross-covariances ( $\rightarrow$  I/1.16.9):

$$\Sigma(X + Y) = \Sigma_{XX} + \Sigma_{YY} + \Sigma_{XY} + \Sigma_{YX}. \quad (1)$$

**Proof:** The covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  can be expressed ( $\rightarrow$  I/1.16.3) in terms of expected values ( $\rightarrow$  I/1.12.1) as follows

$$\Sigma_{XX} = \Sigma(X) = E[(X - E[X])(X - E[X])^T] \quad (2)$$

and the cross-covariance matrix ( $\rightarrow$  I/1.16.9) of  $X$  and  $Y$  can similarly be written as

$$\Sigma_{XY} = \Sigma(X, Y) = E[(X - E[X])(Y - E[Y])^T] \quad (3)$$

Using this and the linearity of the expected value ( $\rightarrow$  I/1.12.5) as well as the definitions of covariance matrix ( $\rightarrow$  I/1.16.1) and cross-covariance matrix ( $\rightarrow$  I/1.16.9), we can derive (1) as follows:

$$\begin{aligned}
\Sigma(X+Y) &\stackrel{(2)}{=} \mathbb{E} \left[ ([X+Y] - \mathbb{E}[X+Y])([X+Y] - \mathbb{E}[X+Y])^T \right] \\
&= \mathbb{E} \left[ ([X - \mathbb{E}(X)] + [Y - \mathbb{E}(Y)])([X - \mathbb{E}(X)] + [Y - \mathbb{E}(Y)])^T \right] \\
&= \mathbb{E} \left[ (X - \mathbb{E}[X])(X - \mathbb{E}[X])^T + (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T + (Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T + (Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T \right] \\
&= \mathbb{E} \left[ (X - \mathbb{E}[X])(X - \mathbb{E}[X])^T \right] + \mathbb{E} \left[ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T \right] + \mathbb{E} \left[ (Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T \right] + \mathbb{E} \left[ (Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T \right] \\
&\stackrel{(2)}{=} \Sigma_{XX} + \Sigma_{YY} + \mathbb{E} \left[ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T \right] + \mathbb{E} \left[ (Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T \right] \\
&\stackrel{(3)}{=} \Sigma_{XX} + \Sigma_{YY} + \Sigma_{XY} + \Sigma_{YX} .
\end{aligned} \tag{4}$$

**Sources:**

- Wikipedia (2022): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: [https://en.wikipedia.org/wiki/Covariance\\_matrix#Basic\\_properties](https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties).

**1.16.12 Precision matrix**

**Definition:** Let  $X = [X_1, \dots, X_n]^T$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the precision matrix of  $X$  is defined as the inverse of the covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$ :

$$\Lambda_{XX} = \Sigma_{XX}^{-1} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix}^{-1} . \tag{1}$$

**Sources:**

- Wikipedia (2020): “Precision (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Precision\\_\(statistics\)](https://en.wikipedia.org/wiki/Precision_(statistics)).

**1.16.13 Precision matrix and correlation matrix**

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the precision matrix ( $\rightarrow$  I/1.16.12)  $\Lambda_{XX}$  of  $X$  can be expressed in terms of its correlation matrix ( $\rightarrow$  I/1.17.6)  $C_{XX}$  as follows

$$\Lambda_{XX} = D_X^{-1} \cdot C_{XX}^{-1} \cdot D_X^{-1} , \tag{1}$$

where  $D_X^{-1}$  is a diagonal matrix with the inverse standard deviations ( $\rightarrow$  I/1.19.1) of  $X_1, \dots, X_n$  as entries on the diagonal:

$$D_X^{-1} = \text{diag}(1/\sigma_{X_1}, \dots, 1/\sigma_{X_n}) = \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} . \tag{2}$$

**Proof:** The precision matrix ( $\rightarrow$  I/1.16.12) is defined as the inverse of the covariance matrix ( $\rightarrow$  I/1.16.1)

$$\Lambda_{XX} = \Sigma_{XX}^{-1} \quad (3)$$

and the relation between covariance matrix and correlation matrix ( $\rightarrow$  I/1.16.8) is given by

$$\Sigma_{XX} = D_X \cdot C_{XX} \cdot D_X \quad (4)$$

where

$$D_X = \text{diag}(\sigma_{X_1}, \dots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (5)$$

Using the matrix product property

$$(A \cdot B \cdot C)^{-1} = C^{-1} \cdot B^{-1} \cdot A^{-1} \quad (6)$$

and the diagonal matrix property

$$\text{diag}(a_1, \dots, a_n)^{-1} = \begin{bmatrix} a_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{a_n} \end{bmatrix} = \text{diag}(1/a_1, \dots, 1/a_n), \quad (7)$$

we obtain

$$\begin{aligned} \Lambda_{XX} &\stackrel{(3)}{=} \Sigma_{XX}^{-1} \\ &\stackrel{(4)}{=} (D_X \cdot C_{XX} \cdot D_X)^{-1} \\ &\stackrel{(6)}{=} D_X^{-1} \cdot C_{XX}^{-1} \cdot D_X^{-1} \\ &\stackrel{(7)}{=} \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \cdot C_{XX}^{-1} \cdot \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \end{aligned} \quad (8)$$

which conforms to equation (1). ■

## 1.17 Correlation

### 1.17.1 Definition

**Definition:** The correlation of two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ , also called Pearson product-moment correlation coefficient (PPMCC), is defined as the ratio of the covariance ( $\rightarrow$  I/1.15.1) of  $X$  and  $Y$  relative to the product of their standard deviations ( $\rightarrow$  I/1.19.1):

$$\text{Corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\text{E}[(X - \text{E}[X])(Y - \text{E}[Y])]}{\sqrt{\text{E}[(X - \text{E}[X])^2]} \sqrt{\text{E}[(Y - \text{E}[Y])^2]}}. \quad (1)$$

Two random variables are called uncorrelated, if  $\text{Corr}(X, Y) = 0$ .

**Sources:**

- Wikipedia (2020): “Correlation and dependence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: [https://en.wikipedia.org/wiki/Correlation\\_and\\_dependence#Pearson's\\_product-moment\\_coefficient](https://en.wikipedia.org/wiki/Correlation_and_dependence#Pearson's_product-moment_coefficient).

### 1.17.2 Range

**Theorem:** Let  $X$  and  $Y$  be two random variables ( $\rightarrow$  I/1.2.2). Then, the correlation of  $X$  and  $Y$  is between and including  $-1$  and  $+1$ :

$$-1 \leq \text{Corr}(X, Y) \leq +1 . \quad (1)$$

**Proof:** Consider the variance ( $\rightarrow$  I/1.13.1) of  $X$  plus or minus  $Y$ , each divided by their standard deviations ( $\rightarrow$  I/1.19.1):

$$\text{Var} \left( \frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) . \quad (2)$$

Because the variance is non-negative ( $\rightarrow$  I/1.13.5), this term is larger than or equal to zero:

$$0 \leq \text{Var} \left( \frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) . \quad (3)$$

Using the variance of a linear combination ( $\rightarrow$  I/1.13.10), it can also be written as:

$$\begin{aligned} \text{Var} \left( \frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) &= \text{Var} \left( \frac{X}{\sigma_X} \right) + \text{Var} \left( \frac{Y}{\sigma_Y} \right) \pm 2 \text{Cov} \left( \frac{X}{\sigma_X}, \frac{Y}{\sigma_Y} \right) \\ &= \frac{1}{\sigma_X^2} \text{Var}(X) + \frac{1}{\sigma_Y^2} \text{Var}(Y) \pm 2 \frac{1}{\sigma_X \sigma_Y} \text{Cov}(X, Y) \\ &= \frac{1}{\sigma_X^2} \sigma_X^2 + \frac{1}{\sigma_Y^2} \sigma_Y^2 \pm 2 \frac{1}{\sigma_X \sigma_Y} \sigma_{XY} . \end{aligned} \quad (4)$$

Using the relationship between covariance and correlation ( $\rightarrow$  I/1.15.8), we have:

$$\text{Var} \left( \frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) = 1 + 1 \pm 2 \text{Corr}(X, Y) . \quad (5)$$

Thus, the combination of (3) with (5) yields

$$0 \leq 2 \pm 2 \text{Corr}(X, Y) \quad (6)$$

which is equivalent to

$$-1 \leq \text{Corr}(X, Y) \leq +1 . \quad (7)$$

**Sources:**

- Dor Leventer (2021): “How can I simply prove that the pearson correlation coefficient is between -1 and 1?”; in: *StackExchange Mathematics*, retrieved on 2021-12-14; URL: <https://math.stackexchange.com/a/4260655/480910>.

### 1.17.3 Correlation under independence

**Theorem:** Independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) are uncorrelated.

**Proof:** The correlation ( $\rightarrow$  I/1.17.1) of two random variables is defined as:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}. \quad (1)$$

The covariance of independent random variables is zero ( $\rightarrow$  I/1.15.6):

$$X, Y \text{ independent} \Rightarrow \text{Cov}(X, Y) = 0. \quad (2)$$

Thus, the correlation of independent random variables is also zero:

$$X, Y \text{ independent} \Rightarrow \text{Corr}(X, Y) = 0. \quad (3)$$

■

**Sources:**

- StatProofBook (2022): “Uncorrelated random variables are not necessarily independent.”; in: *X*, Nov 22, 2022, 06:34 AM; URL: <https://x.com/StatProofBook/status/1594927275514134528>.

### 1.17.4 Sample correlation coefficient

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  and  $y = \{y_1, \dots, y_n\}$  be samples ( $\rightarrow$  I/1.1.6) from random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ . Then, the sample correlation coefficient of  $x$  and  $y$  is given by

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2).

**Sources:**

- Wikipedia (2021): “Pearson correlation coefficient”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-14; URL: [https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient#For\\_a\\_sample](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#For_a_sample).

### 1.17.5 Relationship to standard scores

**Theorem:** Let  $x = \{x_1, \dots, x_n\}$  and  $y = \{y_1, \dots, y_n\}$  be samples ( $\rightarrow$  I/1.1.6) from random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ . Then, the sample correlation coefficient ( $\rightarrow$  I/1.17.4)  $r_{xy}$  can be expressed in terms of the standard scores of  $x$  and  $y$ :

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n z_i^{(x)} \cdot z_i^{(y)} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \quad (1)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2) and  $s_x$  and  $s_y$  are the sample variances ( $\rightarrow$  I/1.13.2).

**Proof:** The sample correlation coefficient ( $\rightarrow$  I/1.17.4) is defined as



$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} . \quad (2)$$

Using the sample variances ( $\rightarrow$  I/1.13.2) of  $x$  and  $y$ , we can write:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(n-1)s_x^2} \sqrt{(n-1)s_y^2}} . \quad (3)$$

Rearranging the terms, we arrive at:

$$r_{xy} = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) . \quad (4)$$

Further simplifying, the result is:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) . \quad (5)$$

■

#### Sources:

- Wikipedia (2021): “Pearson correlation coefficient”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-14; URL: [https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient#For\\_a\\_sample](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#For_a_sample).

### 1.17.6 Correlation matrix

**Definition:** Let  $X = [X_1, \dots, X_n]^T$  be a random vector ( $\rightarrow$  I/1.2.3). Then, the correlation matrix of  $X$  is defined as the  $n \times n$  matrix in which the entry  $(i, j)$  is the correlation ( $\rightarrow$  I/1.17.1) of  $X_i$  and  $X_j$ :

$$C_{XX} = \begin{bmatrix} \text{Corr}(X_1, X_1) & \dots & \text{Corr}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Corr}(X_n, X_1) & \dots & \text{Corr}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \frac{\mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_n - \mathbb{E}[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\mathbb{E}[(X_n - \mathbb{E}[X_n])(X_1 - \mathbb{E}[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\mathbb{E}[(X_n - \mathbb{E}[X_n])(X_n - \mathbb{E}[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} . \quad (1)$$

#### Sources:

- Wikipedia (2020): “Correlation and dependence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Correlation\\_and\\_dependence#Correlation\\_matrices](https://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_matrices).

### 1.17.7 Sample correlation matrix

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^{p \times 1}$ . Then, the sample correlation matrix of  $x$  is the matrix whose entries are the sample correlation coefficients ( $\rightarrow$  I/1.17.4) between pairs of entries of  $x_1, \dots, x_n$ :

$$R_{xx} = \begin{bmatrix} r_{x^{(1)},x^{(1)}} & \dots & r_{x^{(1)},x^{(n)}} \\ \vdots & \ddots & \vdots \\ r_{x^{(n)},x^{(1)}} & \dots & r_{x^{(n)},x^{(n)}} \end{bmatrix} \quad (1)$$

where the  $r_{x^{(j)},x^{(k)}}$  is the sample correlation ( $\rightarrow$  I/1.17.4) between the  $j$ -th and the  $k$ -th entry of  $X$  given by

$$r_{x^{(j)},x^{(k)}} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}^{(j)})(x_{ik} - \bar{x}^{(k)})}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}^{(j)})^2} \sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}^{(k)})^2}} \quad (2)$$

in which  $\bar{x}^{(j)}$  and  $\bar{x}^{(k)}$  are the sample means ( $\rightarrow$  I/1.12.2)

$$\begin{aligned} \bar{x}^{(j)} &= \frac{1}{n} \sum_{i=1}^n x_{ij} \\ \bar{x}^{(k)} &= \frac{1}{n} \sum_{i=1}^n x_{ik} . \end{aligned} \quad (3)$$

## 1.18 Measures of central tendency

### 1.18.1 Median

**Definition:** The median of a sample or random variable is the value separating the higher half from the lower half of its values.

1) Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the median of  $x$  is

$$\text{median}(x) = \begin{cases} x_{(n+1)/2} , & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) , & \text{if } n \text{ is even} , \end{cases} \quad (1)$$

i.e. the median is the “middle” number when all numbers are sorted from smallest to largest.

2) Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2) with cumulative distribution function ( $\rightarrow$  I/1.8.1)  $F_X(x)$ . Then, the median of  $X$  is

$$\text{median}(X) = x, \quad \text{s.t.} \quad F_X(x) = \frac{1}{2} , \quad (2)$$

i.e. the median is the value at which the CDF is 1/2.

#### Sources:

- Wikipedia (2020): “Median”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-15; URL: <https://en.wikipedia.org/wiki/Median>.

### 1.18.2 Median minimizes mean absolute error

**Theorem:** Let  $X_1, \dots, X_n$  be a collection of continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) drawn from a probability distribution ( $\rightarrow$  I/1.5.1) with the probability density function ( $\rightarrow$  I/1.7.1)  $f(x)$  supported on  $(-\infty, \infty)$  with common median ( $\rightarrow$  I/1.18.1)  $m$ . Then,  $m$  minimizes the mean absolute error:

$$m = \arg \min_{a \in \mathbb{R}} E[|X_i - a|] . \quad (1)$$

**Proof:** We can find the optimum by performing a derivative test. First, since an absolute value function is not differentiable at 0, we simplify the objective function by splitting it into two separate integrals:

$$E(|X_i - a|) = \int_{-\infty}^a (a - x)f(x) dx + \int_a^{\infty} (x - a)f(x) dx . \quad (2)$$

Now note that  $|\frac{\partial}{\partial a}(a - x)f(x)| = |\frac{\partial}{\partial a}(x - a)f(x)| = f(x)$ . Consequently,  $\int_{-\infty}^a f(x) dx = \Pr(X_i < a)$  and  $\int_a^{\infty} f(x) dx = \Pr(X_i > a)$ , both of which must be finite by the axioms of probability ( $\rightarrow$  I/1.4.1). Therefore, these integrals meet the conditions for application of Leibniz's rule.

Applying Leibniz's integral rule, we can differentiate the objective function as follows:

$$\begin{aligned} & \frac{\partial}{\partial a} \left( \int_{-\infty}^a (a - x)f(x) dx + \int_a^{\infty} (x - a)f(x) dx \right) \\ &= (a - x)f(x) \Big|_{x=a} + \int_{-\infty}^a f(x) dx - (x - a)f(x) \Big|_{x=a} - \int_a^{\infty} f(x) dx . \end{aligned} \quad (3)$$

Canceling non-integral terms and setting this derivative to 0, it must be true that

$$\int_{-\infty}^a f(x) dx - \int_a^{\infty} f(x) dx = 0 \quad \Rightarrow \quad \Pr(X_i < a) = \Pr(X_i > a) . \quad (4)$$

Together with the probability of the complement ( $\rightarrow$  I/1.4.6), this yields the implication

$$\Pr(X_i < a) = \Pr(X_i > a) \quad \Rightarrow \quad \Pr(X_i < a) = 1 - \Pr(X_i < a) \quad \Rightarrow \quad \Pr(X_i < a) = 0.5 \quad (5)$$

As a result,  $a$  satisfies the definition of a median ( $\rightarrow$  I/1.18.1) at the critical point of the objective function.

Finally, the absolute value is a convex function. Thus, by Jensen's inequality, its expected value is also convex. This implies, since the median is the sole critical point, that it must be a global minimum. Therefore, the median must minimize the mean absolute error, completing the proof. ■

#### Sources:

- Wikipedia (2024): "Derivative test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-23; URL: [https://en.wikipedia.org/wiki/Derivative\\_test](https://en.wikipedia.org/wiki/Derivative_test).
- Wikipedia (2024): "Leibniz integral rule"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-23; URL: [https://en.wikipedia.org/wiki/Leibniz\\_integral\\_rule](https://en.wikipedia.org/wiki/Leibniz_integral_rule).
- Wikipedia (2024): "Jensen's inequality"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-23; URL: [https://en.wikipedia.org/wiki/Jensen%27s\\_inequality](https://en.wikipedia.org/wiki/Jensen%27s_inequality).
- Wikipedia (2024): "Convex function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-09-23; URL: [https://en.wikipedia.org/wiki/Convex\\_function](https://en.wikipedia.org/wiki/Convex_function).

### 1.18.3 Mode

**Definition:** The mode of a sample or random variable is the value which occurs most often or with largest probability among all its values.

1) Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the mode of  $x$  is the value which occurs most often in the list  $x_1, \dots, x_n$ .

2) Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with probability mass function ( $\rightarrow$  I/1.6.1) or probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ . Then, the mode of  $X$  is the value which maximizes the PMF or PDF:

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Mode (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-15; URL: [https://en.wikipedia.org/wiki/Mode\\_\(statistics\)](https://en.wikipedia.org/wiki/Mode_(statistics)).

## 1.19 Measures of statistical dispersion

### 1.19.1 Standard deviation

**Definition:** The standard deviation  $\sigma$  of a random variable ( $\rightarrow$  I/1.2.2)  $X$  with expected value ( $\rightarrow$  I/1.12.1)  $\mu$  is defined as the square root of the variance ( $\rightarrow$  I/1.13.1), i.e.

$$\sigma(X) = \sqrt{\text{E}[(X - \mu)^2]} . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Standard deviation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: [https://en.wikipedia.org/wiki/Standard\\_deviation#Definition\\_of\\_population\\_values](https://en.wikipedia.org/wiki/Standard_deviation#Definition_of_population_values).

### 1.19.2 Sample standard deviation

**Definition:** Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the sample standard deviation of  $x$  is given by

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

and the unbiased sample standard deviation of  $x$  is given by

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

where  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Sources:**

- Ostwald, Dirk (2023): “Erwartungswerte”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (5), Folie 32; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/5\\_Erwartungswerte.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Wintersemester+2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/5_Erwartungswerte.pdf).
- Wikipedia (2025): “Standard deviation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Standard\\_deviation#Estimation](https://en.wikipedia.org/wiki/Standard_deviation#Estimation).

### 1.19.3 Pooled sample standard deviation

**Definition:** Let  $x_i = \{x_{i1}, \dots, x_{in_i}\}$  for  $i = 1, \dots, k$  be samples ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$  whose expected value ( $\rightarrow$  I/1.12.1), but not variance ( $\rightarrow$  I/1.13.1) potentially depends on sample index  $i$ . Then, the pooled sample standard deviation ( $\rightarrow$  I/1.19.2) is defined as the square root of the pooled sample variance ( $\rightarrow$  I/1.13.3), i.e.

$$s_{1\dots k} = \sqrt{s_{1\dots k}^2} = \sqrt{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \quad (1)$$

where  $\bar{x}_i$  is the sample mean ( $\rightarrow$  I/1.12.2) of the  $i$ -th sample

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad (2)$$

and  $n$  is the total sample size ( $\rightarrow$  I/1.1.7):

$$n = \sum_{i=1}^k n_i . \quad (3)$$

#### Sources:

- Ostwald, Dirk (2023): “T-Tests”; in: *Allgemeines Lineares Modell*, Einheit (9), Folien 66/73; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/9\\_T\\_Tests.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/9_T_Tests.pdf).
- Wikipedia (2025): “Pooled variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-01-10; URL: [https://en.wikipedia.org/wiki/Pooled\\_variance#Definition\\_and\\_computation](https://en.wikipedia.org/wiki/Pooled_variance#Definition_and_computation).

### 1.19.4 Full width at half maximum

**Definition:** Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2) with a unimodal probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$  and mode ( $\rightarrow$  I/1.18.3)  $x_M$ . Then, the full width at half maximum of  $X$  is defined as

$$\text{FWHM}(X) = \Delta x = x_2 - x_1 \quad (1)$$

where  $x_1$  and  $x_2$  are specified, such that

$$f_X(x_1) = f_X(x_2) = \frac{1}{2} f_X(x_M) \quad \text{and} \quad x_1 < x_M < x_2 . \quad (2)$$

#### Sources:

- Wikipedia (2020): “Full width at half maximum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: [https://en.wikipedia.org/wiki/Full\\_width\\_at\\_half\\_maximum](https://en.wikipedia.org/wiki/Full_width_at_half_maximum).

## 1.20 Further summary statistics

### 1.20.1 Minimum

**Definition:** The minimum of a sample or random variable is its lowest observed or possible value.

1) Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the minimum of  $x$  is

$$\min(x) = x_j, \quad \text{such that } x_j \leq x_i \quad \text{for all } i = 1, \dots, n, i \neq j, \quad (1)$$

i.e. the minimum is the value which is smaller than or equal to all other observed values.

2) Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$ . Then, the minimum of  $X$  is

$$\min(X) = \tilde{x}, \quad \text{such that } \tilde{x} < x \quad \text{for all } x \in \mathcal{X} \setminus \{\tilde{x}\}, \quad (2)$$

i.e. the minimum is the value which is smaller than all other possible values.

**Sources:**

- Wikipedia (2020): “Sample maximum and minimum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: [https://en.wikipedia.org/wiki/Sample\\_maximum\\_and\\_minimum](https://en.wikipedia.org/wiki/Sample_maximum_and_minimum).

### 1.20.2 Maximum

**Definition:** The maximum of a sample or random variable is its highest observed or possible value.

1) Let  $x = \{x_1, \dots, x_n\}$  be a sample ( $\rightarrow$  I/1.1.6) from a random variable ( $\rightarrow$  I/1.2.2)  $X$ . Then, the maximum of  $x$  is

$$\max(x) = x_j, \quad \text{such that } x_j \geq x_i \quad \text{for all } i = 1, \dots, n, i \neq j, \quad (1)$$

i.e. the maximum is the value which is larger than or equal to all other observed values.

2) Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible values  $\mathcal{X}$ . Then, the maximum of  $X$  is

$$\max(X) = \tilde{x}, \quad \text{such that } \tilde{x} > x \quad \text{for all } x \in \mathcal{X} \setminus \{\tilde{x}\}, \quad (2)$$

i.e. the maximum is the value which is larger than all other possible values.

**Sources:**

- Wikipedia (2020): “Sample maximum and minimum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: [https://en.wikipedia.org/wiki/Sample\\_maximum\\_and\\_minimum](https://en.wikipedia.org/wiki/Sample_maximum_and_minimum).

## 1.21 Further moments

### 1.21.1 Moment

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2), let  $c$  be a constant ( $\rightarrow$  I/1.2.5) and let  $n$  be a positive integer. Then, the  $n$ -th moment of  $X$  about  $c$  is defined as the expected value ( $\rightarrow$  I/1.12.1) of the  $n$ -th power of  $X$  minus  $c$ :

$$\mu_n(c) = E[(X - c)^n] . \quad (1)$$

The “ $n$ -th moment of  $X$ ” may also refer to:

- the  $n$ -th raw moment ( $\rightarrow$  I/1.21.3)  $\mu'_n = \mu_n(0)$ ;
- the  $n$ -th central moment ( $\rightarrow$  I/1.21.6)  $\mu_n = \mu_n(\mu)$ ;
- the  $n$ -th standardized moment ( $\rightarrow$  I/1.21.9)  $\mu_n^* = \mu_n/\sigma^n$ .

**Sources:**

- Wikipedia (2020): “Moment (mathematics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: [https://en.wikipedia.org/wiki/Moment\\_\(mathematics\)#Significance\\_of\\_the\\_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments).

### 1.21.2 Moment in terms of moment-generating function

**Theorem:** Let  $X$  be a scalar random variable ( $\rightarrow$  I/1.2.2) with the moment-generating function ( $\rightarrow$  I/1.9.1)  $M_X(t)$ . Then, the  $n$ -th raw moment ( $\rightarrow$  I/1.21.3) of  $X$  can be calculated from the moment-generating function via

$$E(X^n) = M_X^{(n)}(0) \quad (1)$$

where  $n$  is a positive integer and  $M_X^{(n)}(t)$  is the  $n$ -th derivative of  $M_X(t)$ .

**Proof:** Using the definition of the moment-generating function ( $\rightarrow$  I/1.9.1), we can write:

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} E(e^{tX}) . \quad (2)$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} , \quad (3)$$

equation (2) becomes

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} E\left(\sum_{m=0}^{\infty} \frac{t^m X^m}{m!}\right) . \quad (4)$$

Because the expected value is a linear operator ( $\rightarrow$  I/1.12.5), we have:

$$\begin{aligned} M_X^{(n)}(t) &= \frac{d^n}{dt^n} \sum_{m=0}^{\infty} E\left(\frac{t^m X^m}{m!}\right) \\ &= \sum_{m=0}^{\infty} \frac{d^n}{dt^n} \frac{t^m}{m!} E(X^m) . \end{aligned} \quad (5)$$

Using the  $n$ -th derivative of the  $m$ -th power

$$\frac{d^n}{dx^n} x^m = \begin{cases} m^n x^{m-n}, & \text{if } n \leq m \\ 0, & \text{if } n > m. \end{cases} \quad (6)$$

with the falling factorial

$$m^n = \prod_{i=0}^{n-1} (m - i) = \frac{m!}{(m - n)!}, \quad (7)$$

equation (5) becomes

$$\begin{aligned} M_X^{(n)}(t) &= \sum_{m=n}^{\infty} \frac{m^n t^{m-n}}{m!} E(X^m) \\ &\stackrel{(7)}{=} \sum_{m=n}^{\infty} \frac{m! t^{m-n}}{(m - n)! m!} E(X^m) \\ &= \sum_{m=n}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= \frac{t^{n-n}}{(n - n)!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= \frac{t^0}{0!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m). \end{aligned} \quad (8)$$

Setting  $t = 0$  in (8) yields

$$\begin{aligned} M_X^{(n)}(0) &= E(X^n) + \sum_{m=n+1}^{\infty} \frac{0^{m-n}}{(m - n)!} E(X^m) \\ &= E(X^n) \end{aligned} \quad (9)$$

which conforms to equation (1). ■

#### Sources:

- ProofWiki (2020): “Moment in terms of Moment Generating Function”; in: *ProofWiki*, retrieved on 2020-08-19; URL: [https://proofwiki.org/wiki/Moment\\_in\\_terms\\_of\\_Moment\\_Generating\\_Function](https://proofwiki.org/wiki/Moment_in_terms_of_Moment_Generating_Function).

### 1.21.3 Raw moment

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) and let  $n$  be a positive integer. Then, the  $n$ -th raw moment of  $X$ , also called ( $n$ -th) “crude moment”, is defined as the  $n$ -th moment ( $\rightarrow$  I/1.21.1) of  $X$  about the value 0:



$$\mu'_n = \mu_n(0) = E[(X - 0)^n] = E[X^n] . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Moment (mathematics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: [https://en.wikipedia.org/wiki/Moment\\_\(mathematics\)#Significance\\_of\\_the\\_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments).

**1.21.4 First raw moment is mean**

**Theorem:** The first raw moment ( $\rightarrow$  I/1.21.3) equals the mean ( $\rightarrow$  I/1.12.1), i.e.

$$\mu'_1 = \mu . \quad (1)$$

**Proof:** The first raw moment ( $\rightarrow$  I/1.21.3) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as

$$\mu'_1 = E[(X - 0)^1] \quad (2)$$

which is equal to the expected value ( $\rightarrow$  I/1.12.1) of  $X$ :

$$\mu'_1 = E[X] = \mu . \quad (3)$$

■

**1.21.5 Second raw moment and variance**

**Theorem:** The second raw moment ( $\rightarrow$  I/1.21.3) can be expressed as

$$\mu'_2 = \text{Var}(X) + E(X)^2 \quad (1)$$

where  $\text{Var}(X)$  is the variance ( $\rightarrow$  I/1.13.1) of  $X$  and  $E(X)$  is the expected value ( $\rightarrow$  I/1.12.1) of  $X$ .

**Proof:** The second raw moment ( $\rightarrow$  I/1.21.3) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  is defined as

$$\mu'_2 = E[(X - 0)^2] . \quad (2)$$

Using the partition of variance into expected values ( $\rightarrow$  I/1.13.4)

$$\text{Var}(X) = E(X^2) - E(X)^2 , \quad (3)$$

the second raw moment can be rearranged into:

$$\mu'_2 \stackrel{(2)}{=} E(X^2) \stackrel{(3)}{=} \text{Var}(X) + E(X)^2 . \quad (4)$$

■

### 1.21.6 Central moment

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $\mu$  and let  $n$  be a positive integer. Then, the  $n$ -th central moment of  $X$  is defined as the  $n$ -th moment ( $\rightarrow$  I/1.21.1) of  $X$  about the value  $\mu$ :

$$\mu_n = E[(X - \mu)^n] . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Moment (mathematics)” ; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: [https://en.wikipedia.org/wiki/Moment\\_\(mathematics\)#Significance\\_of\\_the\\_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments).

### 1.21.7 First central moment is zero

**Theorem:** The first central moment ( $\rightarrow$  I/1.21.6) is zero, i.e.

$$\mu_1 = 0 . \quad (1)$$

**Proof:** The first central moment ( $\rightarrow$  I/1.21.6) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  with mean ( $\rightarrow$  I/1.12.1)  $\mu$  is defined as

$$\mu_1 = E[(X - \mu)^1] . \quad (2)$$

Due to the linearity of the expected value ( $\rightarrow$  I/1.12.5) and by plugging in  $\mu = E(X)$ , we have

$$\begin{aligned} \mu_1 &= E[X - \mu] \\ &= E(X) - \mu \\ &= E(X) - E(X) \\ &= 0 . \end{aligned} \quad (3)$$

■

**Sources:**

- ProofWiki (2020): “First Central Moment is Zero”; in: *ProofWiki*, retrieved on 2020-09-09; URL: [https://proofwiki.org/wiki/First\\_Central\\_Moment\\_is\\_Zero](https://proofwiki.org/wiki/First_Central_Moment_is_Zero).

### 1.21.8 Second central moment is variance

**Theorem:** The second central moment ( $\rightarrow$  I/1.21.6) equals the variance ( $\rightarrow$  I/1.13.1), i.e.

$$\mu_2 = \text{Var}(X) . \quad (1)$$

**Proof:** The second central moment ( $\rightarrow$  I/1.21.6) of a random variable ( $\rightarrow$  I/1.2.2)  $X$  with mean ( $\rightarrow$  I/1.12.1)  $\mu$  is defined as

$$\mu_2 = E[(X - \mu)^2] \quad (2)$$

which is equivalent to the definition of the variance ( $\rightarrow$  I/1.13.1):

$$\mu_2 = E[(X - E(X))^2] = \text{Var}(X) . \quad (3)$$

■

**Sources:**

- Wikipedia (2020): “Moment (mathematics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: [https://en.wikipedia.org/wiki/Moment\\_\(mathematics\)#Significance\\_of\\_the\\_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments).

### 1.21.9 Standardized moment

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with expected value ( $\rightarrow$  I/1.12.1)  $\mu$  and standard deviation ( $\rightarrow$  I/1.19.1)  $\sigma$  and let  $n$  be a positive integer. Then, the  $n$ -th standardized moment of  $X$  is defined as the  $n$ -th moment ( $\rightarrow$  I/1.21.1) of  $X$  about the value  $\mu$ , divided by the  $n$ -th power of  $\sigma$ :

$$\mu_n^* = \frac{\mu_n}{\sigma^n} = \frac{E[(X - \mu)^n]}{\sigma^n} . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Moment (mathematics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: [https://en.wikipedia.org/wiki/Moment\\_\(mathematics\)#Standardized\\_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Standardized_moments).

## 2 Information theory

### 2.1 Shannon entropy

#### 2.1.1 Definition

**Definition:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and the (observed or assumed) probability mass function ( $\rightarrow$  I/1.6.1)  $p(x) = f_X(x)$ . Then, the entropy (also referred to as “Shannon entropy”) of  $X$  is defined as

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the entropy is determined. By convention,  $0 \cdot \log 0$  is taken to be zero when calculating the entropy of  $X$ .

**Sources:**

- Shannon CE (1948): “A Mathematical Theory of Communication”; in: *Bell System Technical Journal*, vol. 27, iss. 3, pp. 379-423; URL: <https://ieeexplore.ieee.org/document/6773024>; DOI: 10.1002/j.1538-7305.1948.tb01338.x.

#### 2.1.2 Non-negativity

**Theorem:** The entropy of a discrete random variable ( $\rightarrow$  I/1.2.2) is a non-negative number:

$$H(X) \geq 0 . \quad (1)$$

**Proof:** The entropy of a discrete random variable ( $\rightarrow$  I/2.1.1) is defined as

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \quad (2)$$

The minus sign can be moved into the sum:

$$H(X) = \sum_{x \in \mathcal{X}} [p(x) \cdot (-\log_b p(x))] \quad (3)$$

Because the co-domain of probability mass functions ( $\rightarrow$  I/1.6.1) is  $[0, 1]$ , we can deduce:

$$\begin{aligned} 0 &\leq p(x) \leq 1 \\ -\infty &\leq \log_b p(x) \leq 0 \\ 0 &\leq -\log_b p(x) \leq +\infty \\ 0 &\leq p(x) \cdot (-\log_b p(x)) \leq +\infty . \end{aligned} \quad (4)$$

By convention,  $0 \cdot \log_b(0)$  is taken to be 0 when calculating entropy, consistent with

$$\lim_{p \rightarrow 0} [p \log_b(p)] = 0 . \quad (5)$$

Taking this together, each addend in (3) is positive or zero and thus, the entire sum must also be non-negative.

**Sources:**

- Cover TM, Thomas JA (1991): “Elements of Information Theory”, p. 15; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

**2.1.3 Concavity**

**Theorem:** The entropy ( $\rightarrow$  I/2.1.1) is concave in the probability mass function ( $\rightarrow$  I/1.6.1)  $p$ , i.e.

$$H[\lambda p_1 + (1 - \lambda)p_2] \geq \lambda H[p_1] + (1 - \lambda)H[p_2] \quad (1)$$

where  $p_1$  and  $p_2$  are probability mass functions and  $0 \leq \lambda \leq 1$ .

**Proof:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $u(x)$  be the probability mass function ( $\rightarrow$  I/1.6.1) of a discrete uniform distribution ( $\rightarrow$  II/1.0.1) on  $X \in \mathcal{X}$ . Then, the entropy ( $\rightarrow$  I/2.1.1) of an arbitrary probability mass function ( $\rightarrow$  I/1.6.1)  $p(x)$  can be rewritten as

$$\begin{aligned} H[p] &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} u(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \cdot \log u(x) \\ &= -\text{KL}[p||u] - \log \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \\ &= \log |\mathcal{X}| - \text{KL}[p||u] \end{aligned} \quad (2)$$

$$\log |\mathcal{X}| - H[p] = \text{KL}[p||u]$$

where we have applied the definition of the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1), the probability mass function of the discrete uniform distribution ( $\rightarrow$  II/1.0.2) and the total sum over the probability mass function ( $\rightarrow$  I/1.6.1).

Note that the KL divergence is convex ( $\rightarrow$  I/2.5.6) in the pair of probability distributions ( $\rightarrow$  I/1.5.1)  $(p, q)$ :

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \quad (3)$$

A special case of this is given by

$$\begin{aligned} \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda u + (1 - \lambda)u] &\leq \lambda \text{KL}[p_1 || u] + (1 - \lambda) \text{KL}[p_2 || u] \\ \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || u] &\leq \lambda \text{KL}[p_1 || u] + (1 - \lambda) \text{KL}[p_2 || u] \end{aligned} \quad (4)$$

and applying equation (2), we have

$$\begin{aligned} \log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \lambda (\log |\mathcal{X}| - H[p_1]) + (1 - \lambda) (\log |\mathcal{X}| - H[p_2]) \\ \log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \log |\mathcal{X}| - \lambda H[p_1] - (1 - \lambda)H[p_2] \\ -H[\lambda p_1 + (1 - \lambda)p_2] &\leq -\lambda H[p_1] - (1 - \lambda)H[p_2] \\ H[\lambda p_1 + (1 - \lambda)p_2] &\geq \lambda H[p_1] + (1 - \lambda)H[p_2] \end{aligned} \quad (5)$$

which is equivalent to (1). ■

**Sources:**

- Wikipedia (2020): “Entropy (information theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: [https://en.wikipedia.org/wiki/Entropy\\_\(information\\_theory\)#Further\\_properties](https://en.wikipedia.org/wiki/Entropy_(information_theory)#Further_properties).
- Cover TM, Thomas JA (1991): “Elements of Information Theory”, p. 30; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.
- Xie, Yao (2012): “Chain Rules and Inequalities”; in: *ECE587: Information Theory*, Lecture 3, Slide 25; URL: <https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf>.
- Goh, Siong Thye (2016): “Understanding the proof of the concavity of entropy”; in: *StackExchange Mathematics*, retrieved on 2020-11-08; URL: <https://math.stackexchange.com/questions/2000194/understanding-the-proof-of-the-concavity-of-entropy>.

### 2.1.4 Conditional entropy

**Definition:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$  and probability mass functions ( $\rightarrow$  I/1.6.1)  $p(x)$  and  $p(y)$ . Then, the conditional entropy of  $Y$  given  $X$  or, entropy of  $Y$  conditioned on  $X$ , is defined as

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) \cdot H(Y|X = x) \quad (1)$$

where  $H(Y|X = x)$  is the (marginal) entropy ( $\rightarrow$  I/2.1.1) of  $Y$ , evaluated at  $x$ .

**Sources:**

- Cover TM, Thomas JA (1991): “Joint Entropy and Conditional Entropy”; in: *Elements of Information Theory*, ch. 2.2, p. 15; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

### 2.1.5 Joint entropy

**Definition:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$  and joint probability ( $\rightarrow$  I/1.3.2) mass function ( $\rightarrow$  I/1.6.1)  $p(x, y)$ . Then, the joint entropy of  $X$  and  $Y$  is defined as

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the entropy is determined.

**Sources:**

- Cover TM, Thomas JA (1991): “Joint Entropy and Conditional Entropy”; in: *Elements of Information Theory*, ch. 2.2, p. 16; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

### 2.1.6 Cross-entropy

**Definition:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $P$  and  $Q$  be two probability distributions ( $\rightarrow$  I/1.5.1) on  $X$  with the probability mass functions ( $\rightarrow$  I/1.6.1)  $p(x)$  and  $q(x)$ . Then, the cross-entropy of  $Q$  relative to  $P$  is defined as

$$H(P, Q) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b q(x) \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the cross-entropy is determined.

**Sources:**

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: [https://en.wikipedia.org/wiki/Cross\\_entropy#Definition](https://en.wikipedia.org/wiki/Cross_entropy#Definition).

### 2.1.7 Convexity of cross-entropy

**Theorem:** The cross-entropy ( $\rightarrow$  I/2.1.6) is convex in the second probability distribution ( $\rightarrow$  I/1.5.1), i.e.

$$H[p, \lambda q_1 + (1 - \lambda)q_2] \leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] \quad (1)$$

where  $p$  is a fixed and  $q_1$  and  $q_2$  are any two probability mass functions ( $\rightarrow$  I/1.6.1) and  $0 \leq \lambda \leq 1$ .

**Proof:** The relationship between Kullback-Leibler divergence, entropy and cross-entropy ( $\rightarrow$  I/2.5.9) is:

$$\text{KL}[P||Q] = H(P, Q) - H(P) . \quad (2)$$

Note that the KL divergence is convex ( $\rightarrow$  I/2.5.6) in the pair of probability distributions ( $\rightarrow$  I/1.5.1)  $(p, q)$ :

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda)\text{KL}[p_2 || q_2] . \quad (3)$$

A special case of this is given by

$$\begin{aligned} \text{KL}[\lambda p + (1 - \lambda)p || \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda \text{KL}[p || q_1] + (1 - \lambda)\text{KL}[p || q_2] \\ \text{KL}[p || \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda \text{KL}[p || q_1] + (1 - \lambda)\text{KL}[p || q_2] \end{aligned} \quad (4)$$

and applying equation (2), we have

$$\begin{aligned} H[p, \lambda q_1 + (1 - \lambda)q_2] - H[p] &\leq \lambda (H[p, q_1] - H[p]) + (1 - \lambda) (H[p, q_2] - H[p]) \\ H[p, \lambda q_1 + (1 - \lambda)q_2] - H[p] &\leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] - H[p] \\ H[p, \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] \end{aligned} \quad (5)$$

which is equivalent to (1). ■

**Sources:**

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: [https://en.wikipedia.org/wiki/Cross\\_entropy#Definition](https://en.wikipedia.org/wiki/Cross_entropy#Definition).
- gunes (2019): “Convexity of cross entropy”; in: *StackExchange CrossValidated*, retrieved on 2020-11-08; URL: <https://stats.stackexchange.com/questions/394463/convexity-of-cross-entropy>.

### 2.1.8 Gibbs’ inequality

**Theorem:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2) and consider two probability distributions ( $\rightarrow$  I/1.5.1) with probability mass functions ( $\rightarrow$  I/1.6.1)  $p(x)$  and  $q(x)$ . Then, Gibbs’ inequality states that the entropy ( $\rightarrow$  I/2.1.1) of  $X$  according to  $P$  is smaller than or equal to the cross-entropy ( $\rightarrow$  I/2.1.6) of  $P$  and  $Q$ :

$$-\sum_{x \in \mathcal{X}} p(x) \log_b p(x) \leq -\sum_{x \in \mathcal{X}} p(x) \log_b q(x) . \quad (1)$$

**Proof:** Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_b a = \frac{\ln a}{\ln b} . \quad (2)$$

Let  $I$  be the set of all  $x$  for which  $p(x)$  is non-zero. Then, proving (1) requires to show that

$$\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} \geq 0 . \quad (3)$$

For all  $x > 0$ , it holds that  $\ln x \leq x - 1$ , with equality only if  $x = 1$ . Multiplying this with  $-1$ , we have  $\ln \frac{1}{x} \geq 1 - x$ . Applying this to (3), we can say about the left-hand side that

$$\begin{aligned} \sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) \left( 1 - \frac{q(x)}{p(x)} \right) \\ &= \sum_{x \in I} p(x) - \sum_{x \in I} q(x) . \end{aligned} \quad (4)$$

Finally, since  $p(x)$  and  $q(x)$  are probability mass functions ( $\rightarrow$  I/1.6.1), we have

$$\begin{aligned} 0 \leq p(x) \leq 1, \quad \sum_{x \in I} p(x) &= 1 \quad \text{and} \\ 0 \leq q(x) \leq 1, \quad \sum_{x \in I} q(x) &\leq 1 , \end{aligned} \quad (5)$$

such that it follows from (4) that

$$\begin{aligned} \sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \\ &= 1 - \sum_{x \in I} q(x) \geq 0 . \end{aligned} \quad (6)$$



**Sources:**

- Wikipedia (2020): “Gibbs’ inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: [https://en.wikipedia.org/wiki/Gibbs%27\\_inequality#Proof](https://en.wikipedia.org/wiki/Gibbs%27_inequality#Proof).

**2.1.9 Log sum inequality**

**Theorem:** Let  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  be non-negative real numbers and define  $a = \sum_{i=1}^n a_i$  and  $b = \sum_{i=1}^n b_i$ . Then, the log sum inequality states that

$$\sum_{i=1}^n a_i \log_c \frac{a_i}{b_i} \geq a \log_c \frac{a}{b} . \quad (1)$$

**Proof:** Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_c a = \frac{\ln a}{\ln c} . \quad (2)$$

Let  $f(x) = x \ln x$ . Then, the left-hand side of (1) can be rewritten as

$$\begin{aligned} \sum_{i=1}^n a_i \ln \frac{a_i}{b_i} &= \sum_{i=1}^n b_i f\left(\frac{a_i}{b_i}\right) \\ &= b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) . \end{aligned} \quad (3)$$

Because  $f(x)$  is a convex function and

$$\begin{aligned} \frac{b_i}{b} &\geq 0 \\ \sum_{i=1}^n \frac{b_i}{b} &= 1 , \end{aligned} \quad (4)$$

applying Jensen’s inequality ( $\rightarrow$  I/1.12.11) yields

$$\begin{aligned} b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) &\geq b f\left(\sum_{i=1}^n \frac{b_i}{b} \frac{a_i}{b_i}\right) \\ &= b f\left(\frac{1}{b} \sum_{i=1}^n a_i\right) \\ &= b f\left(\frac{a}{b}\right) \\ &= a \ln \frac{a}{b} . \end{aligned} \quad (5)$$

Finally, combining (3) and (5), this demonstrates (1).

**Sources:**

- Wikipedia (2020): “Log sum inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: [https://en.wikipedia.org/wiki/Log\\_sum\\_inequality#Proof](https://en.wikipedia.org/wiki/Log_sum_inequality#Proof).

**2.2 Differential entropy****2.2.1 Definition**

**Definition:** Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and the (estimated or assumed) probability density function ( $\rightarrow$  I/1.7.1)  $p(x) = f_X(x)$ . Then, the differential entropy (also referred to as “continuous entropy”) of  $X$  is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the entropy is determined.

**Sources:**

- Cover TM, Thomas JA (1991): “Differential Entropy”; in: *Elements of Information Theory*, ch. 8.1, p. 243; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

**2.2.2 Negativity**

**Theorem:** Unlike its discrete analogue ( $\rightarrow$  I/2.1.2), the differential entropy ( $\rightarrow$  I/2.2.1) can become negative.

**Proof:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1) with minimum 0 and maximum  $1/2$ :

$$X \sim \mathcal{U}(0, 1/2) . \quad (1)$$

Then, its probability density function ( $\rightarrow$  II/3.1.3) is:

$$f_X(x) = 2 \quad \text{for} \quad 0 \leq x \leq \frac{1}{2} . \quad (2)$$

Thus, the differential entropy ( $\rightarrow$  I/2.2.1) follows as

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} f_X(x) \log_b f_X(x) dx \\ &= - \int_0^{\frac{1}{2}} 2 \log_b(2) dx \\ &= - \log_b(2) \int_0^{\frac{1}{2}} 2 dx \\ &= - \log_b(2) [2x]_0^{\frac{1}{2}} \\ &= - \log_b(2) \end{aligned} \quad (3)$$

which is negative for any base  $b > 1$ .

**Sources:**

- Wikipedia (2020): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-02; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Definition](https://en.wikipedia.org/wiki/Differential_entropy#Definition).

**2.2.3 Invariance under addition**

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2). Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  remains constant under addition of a constant:

$$h(X + c) = h(X) . \quad (1)$$

**Proof:** By definition, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx \quad (2)$$

where  $p(x) = f_X(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ . Define the mappings between  $X$  and  $Y = X + c$  as

$$Y = g(X) = X + c \quad \Leftrightarrow \quad X = g^{-1}(Y) = Y - c . \quad (3)$$

Note that  $g(X)$  is a strictly increasing function, such that the probability density function ( $\rightarrow$  I/1.7.3) of  $Y$  is

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} \stackrel{(3)}{=} f_X(y - c) . \quad (4)$$

Writing down the differential entropy for  $Y$ , we have:

$$\begin{aligned} h(Y) &= - \int_{\mathcal{Y}} f_Y(y) \log f_Y(y) dy \\ &\stackrel{(4)}{=} - \int_{\mathcal{Y}} f_X(y - c) \log f_X(y - c) dy \end{aligned} \quad (5)$$

Substituting  $x = y - c$ , such that  $y = x + c$ , this yields:

$$\begin{aligned} h(Y) &= - \int_{\{y-c \mid y \in \mathcal{Y}\}} f_X(x + c - c) \log f_X(x + c - c) d(x + c) \\ &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx \\ &\stackrel{(2)}{=} h(X) . \end{aligned} \quad (6)$$

**Sources:**

- Wikipedia (2020): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-12; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Properties\\_of\\_differential\\_entropy](https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy).

### 2.2.4 Addition upon multiplication

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random variable ( $\rightarrow$  I/1.2.2). Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  increases additively upon multiplication with a constant:

$$h(aX) = h(X) + \log |a| . \quad (1)$$

**Proof:** By definition, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx \quad (2)$$

where  $p(x) = f_X(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ .

Define the mappings between  $X$  and  $Y = aX$  as

$$Y = g(X) = aX \quad \Leftrightarrow \quad X = g^{-1}(Y) = \frac{Y}{a} . \quad (3)$$

If  $a > 0$ , then  $g(X)$  is a strictly increasing function, such that the probability density function ( $\rightarrow$  I/1.7.3) of  $Y$  is

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} \stackrel{(3)}{=} \frac{1}{a} f_X\left(\frac{y}{a}\right) ; \quad (4)$$

if  $a < 0$ , then  $g(X)$  is a strictly decreasing function, such that the probability density function ( $\rightarrow$  I/1.7.4) of  $Y$  is

$$f_Y(y) = -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} \stackrel{(3)}{=} -\frac{1}{a} f_X\left(\frac{y}{a}\right) ; \quad (5)$$

thus, we can write

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right) . \quad (6)$$

Writing down the differential entropy for  $Y$ , we have:

$$\begin{aligned} h(Y) &= - \int_{\mathcal{Y}} f_Y(y) \log f_Y(y) dy \\ &\stackrel{(6)}{=} - \int_{\mathcal{Y}} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log \left[ \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \right] dy \end{aligned} \quad (7)$$

Substituting  $x = y/a$ , such that  $y = ax$ , this yields:

$$\begin{aligned} h(Y) &= - \int_{\{y/a \mid y \in \mathcal{Y}\}} \frac{1}{|a|} f_X\left(\frac{ax}{a}\right) \log \left[ \frac{1}{|a|} f_X\left(\frac{ax}{a}\right) \right] d(ax) \\ &= - \int_{\mathcal{X}} f_X(x) \log \left[ \frac{1}{|a|} f_X(x) \right] dx \\ &= - \int_{\mathcal{X}} f_X(x) [\log f_X(x) - \log |a|] dx \\ &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx + \log |a| \int_{\mathcal{X}} f_X(x) dx \\ &\stackrel{(2)}{=} h(X) + \log |a| . \end{aligned} \quad (8)$$

**Sources:**

- Wikipedia (2020): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-12; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Properties\\_of\\_differential\\_entropy](https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy).

**2.2.5 Addition upon matrix multiplication**

**Theorem:** Let  $X$  be a continuous ( $\rightarrow$  I/1.2.6) random vector ( $\rightarrow$  I/1.2.3). Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  increases additively when multiplied with an invertible matrix  $A$ :

$$h(AX) = h(X) + \log |A| . \quad (1)$$

**Proof:** By definition, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx \quad (2)$$

where  $f_X(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  and  $\mathcal{X}$  is the set of possible values of  $X$ .

The probability density function of a linear function of a continuous random vector ( $\rightarrow$  I/1.7.6)  $Y = g(X) = \Sigma X + \mu$  is

$$f_Y(y) = \begin{cases} \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y - \mu)) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (3)$$

where  $\mathcal{Y} = \{y = \Sigma x + \mu : x \in \mathcal{X}\}$  is the set of possible outcomes of  $Y$ .

Therefore, with  $Y = g(X) = AX$ , i.e.  $\Sigma = A$  and  $\mu = 0_n$ , the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$  is given by

$$f_Y(y) = \begin{cases} \frac{1}{|A|} f_X(A^{-1}y) , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (4)$$

where  $\mathcal{Y} = \{y = Ax : x \in \mathcal{X}\}$ .

Thus, the differential entropy ( $\rightarrow$  I/2.2.1) of  $Y$  is

$$\begin{aligned} h(Y) &\stackrel{(2)}{=} - \int_{\mathcal{Y}} f_Y(y) \log f_Y(y) dy \\ &\stackrel{(4)}{=} - \int_{\mathcal{Y}} \left[ \frac{1}{|A|} f_X(A^{-1}y) \right] \log \left[ \frac{1}{|A|} f_X(A^{-1}y) \right] dy . \end{aligned} \quad (5)$$

Substituting  $y = Ax$  into the integral, we obtain

$$\begin{aligned} h(Y) &= - \int_{\mathcal{X}} \left[ \frac{1}{|A|} f_X(A^{-1}Ax) \right] \log \left[ \frac{1}{|A|} f_X(A^{-1}Ax) \right] d(Ax) \\ &= - \frac{1}{|A|} \int_{\mathcal{X}} f_X(x) \log \left[ \frac{1}{|A|} f_X(x) \right] d(Ax) . \end{aligned} \quad (6)$$

Using the differential  $d(Ax) = |A|dx$ , this becomes

$$\begin{aligned} h(Y) &= -\frac{|A|}{|A|} \int_{\mathcal{X}} f_X(x) \log \left[ \frac{1}{|A|} f_X(x) \right] dx \\ &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx - \int_{\mathcal{X}} f_X(x) \log \frac{1}{|A|} dx . \end{aligned} \quad (7)$$

Finally, employing the fact ( $\rightarrow$  I/1.7.1) that  $\int_{\mathcal{X}} f_X(x) dx = 1$ , we can derive the differential entropy ( $\rightarrow$  I/2.2.1) of  $Y$  as

$$\begin{aligned} h(Y) &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx + \log |A| \int_{\mathcal{X}} f_X(x) dx \\ &\stackrel{(2)}{=} h(X) + \log |A| . \end{aligned} \quad (8)$$

■

#### Sources:

- Cover, Thomas M. & Thomas, Joy A. (1991): “Properties of Differential Entropy, Relative Entropy, and Mutual Information”; in: *Elements of Information Theory*, sect. 8.6, p. 253; URL: [https://www.google.de/books/edition/Elements\\_of\\_Information\\_Theory/j0DBDwAAQBAJ](https://www.google.de/books/edition/Elements_of_Information_Theory/j0DBDwAAQBAJ).
- Wikipedia (2021): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Properties\\_of\\_differential\\_entropy](https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy).

### 2.2.6 Non-invariance and transformation

**Theorem:** The differential entropy ( $\rightarrow$  I/2.2.1) is not invariant under change of variables, i.e. there exist random variables  $X$  and  $Y = g(X)$ , such that

$$h(Y) \neq h(X) . \quad (1)$$

In particular, for an invertible transformation  $g : \mathcal{X} \rightarrow \mathcal{Y}$  from a random vector  $X$  to another random vector of the same dimension  $Y$ , it holds that

$$h(Y) = h(X) + \int_{\mathcal{X}} f_X(x) \log |J_g(x)| dx . \quad (2)$$

where  $J_g(x)$  is the Jacobian matrix of the vector-valued function  $g$  and  $\mathcal{X}$  and  $\mathcal{Y}$  are the sets of possible values of  $X$  of  $Y$ , respectively.

**Proof:** By definition, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx \quad (3)$$

where  $f_X(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ .

The probability density function of an invertible function of a continuous random vector ( $\rightarrow$  I/1.7.5)  $Y = g(X)$  is

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (4)$$

where  $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$  is the set of possible outcomes of  $Y$  and  $J_{g^{-1}}(y)$  is the Jacobian matrix of  $g^{-1}(y)$

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ \vdots & \ddots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix}. \quad (5)$$

Thus, the differential entropy ( $\rightarrow$  I/2.2.1) of  $Y$  is

$$\begin{aligned} h(Y) &\stackrel{(3)}{=} - \int_{\mathcal{Y}} f_Y(y) \log f_Y(y) dy \\ &\stackrel{(4)}{=} - \int_{\mathcal{Y}} [f_X(g^{-1}(y)) |J_{g^{-1}}(y)|] \log [f_X(g^{-1}(y)) |J_{g^{-1}}(y)|] dy. \end{aligned} \quad (6)$$

Substituting  $y = g(x)$  into the integral and applying  $J_{f^{-1}}(y) = J_f^{-1}(x)$ , we obtain

$$\begin{aligned} h(Y) &= - \int_{\mathcal{X}} [f_X(g^{-1}(g(x))) |J_{g^{-1}}(y)|] \log [f_X(g^{-1}(g(x))) |J_{g^{-1}}(y)|] d[g(x)] \\ &= - \int_{\mathcal{X}} [f_X(x) |J_g^{-1}(x)|] \log [f_X(x) |J_g^{-1}(x)|] d[g(x)]. \end{aligned} \quad (7)$$

Using the relations  $y = f(x) \Rightarrow dy = |J_f(x)| dx$  and  $|A||B| = |AB|$ , this becomes

$$\begin{aligned} h(Y) &= - \int_{\mathcal{X}} [f_X(x) |J_g^{-1}(x)| |J_g(x)|] \log [f_X(x) |J_g^{-1}(x)|] dx \\ &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx - \int_{\mathcal{X}} f_X(x) \log |J_g^{-1}(x)| dx. \end{aligned} \quad (8)$$

Finally, employing the fact ( $\rightarrow$  I/1.7.8) that  $\int_{\mathcal{X}} f_X(x) dx = 1$  and the determinant property  $|A^{-1}| = 1/|A|$ , we can derive the differential entropy ( $\rightarrow$  I/2.2.1) of  $Y$  as

$$\begin{aligned} h(Y) &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) dx - \int_{\mathcal{X}} f_X(x) \log \frac{1}{|J_g(x)|} dx \\ &\stackrel{(3)}{=} h(X) + \int_{\mathcal{X}} f_X(x) \log |J_g(x)| dx. \end{aligned} \quad (9)$$

Because there exist  $X$  and  $Y$ , such that the integral term in (9) is non-zero, this also demonstrates that there exist  $X$  and  $Y$ , such that (1) is fulfilled. ■

### Sources:

- Wikipedia (2021): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Properties\\_of\\_differential\\_entropy](https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy).
- Bernhard (2016): “proof of upper bound on differential entropy of  $f(X)$ ”; in: *StackExchange Mathematics*, retrieved on 2021-10-07; URL: <https://math.stackexchange.com/a/1759531>.

- peek-a-boo (2019): “How to come up with the Jacobian in the change of variables formula”; in: *StackExchange Mathematics*, retrieved on 2021-08-30; URL: <https://math.stackexchange.com/a/3239222>.
- Wikipedia (2021): “Jacobian matrix and determinant”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: [https://en.wikipedia.org/wiki/Jacobian\\_matrix\\_and\\_determinant#Inverse](https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant#Inverse).
- Wikipedia (2021): “Inverse function theorem”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: [https://en.wikipedia.org/wiki/Inverse\\_function\\_theorem#Statement](https://en.wikipedia.org/wiki/Inverse_function_theorem#Statement).
- Wikipedia (2021): “Determinant”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: [https://en.wikipedia.org/wiki/Determinant#Properties\\_of\\_the\\_determinant](https://en.wikipedia.org/wiki/Determinant#Properties_of_the_determinant).

### 2.2.7 Conditional differential entropy

**Definition:** Let  $X$  and  $Y$  be continuous random variables ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$  and probability density functions ( $\rightarrow$  I/1.7.1)  $p(x)$  and  $p(y)$ . Then, the conditional differential entropy of  $Y$  given  $X$  or, differential entropy of  $Y$  conditioned on  $X$ , is defined as

$$h(Y|X) = \int_{x \in \mathcal{X}} p(x) \cdot h(Y|X = x) dx \quad (1)$$

where  $h(Y|X = x)$  is the (marginal) differential entropy ( $\rightarrow$  I/2.2.1) of  $Y$ , evaluated at  $x$ .

### 2.2.8 Joint differential entropy

**Definition:** Let  $X$  and  $Y$  be continuous random variables ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and  $\mathcal{Y}$  and joint probability ( $\rightarrow$  I/1.3.2) density function ( $\rightarrow$  I/1.7.1)  $p(x, y)$ . Then, the joint differential entropy of  $X$  and  $Y$  is defined as

$$h(X, Y) = - \int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) dy dx \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the differential entropy is determined.

### 2.2.9 Differential cross-entropy

**Definition:** Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $P$  and  $Q$  be two probability distributions ( $\rightarrow$  I/1.5.1) on  $X$  with the probability density functions ( $\rightarrow$  I/1.7.1)  $p(x)$  and  $q(x)$ . Then, the differential cross-entropy of  $Q$  relative to  $P$  is defined as

$$h(P, Q) = - \int_{\mathcal{X}} p(x) \log_b q(x) dx \quad (1)$$

where  $b$  is the base of the logarithm specifying in which unit the differential cross-entropy is determined.

#### Sources:

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: [https://en.wikipedia.org/wiki/Cross\\_entropy#Definition](https://en.wikipedia.org/wiki/Cross_entropy#Definition).



## 2.3 Discrete mutual information

### 2.3.1 Definition

**Definition:**

1) The mutual information of two discrete random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as

$$I(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} \quad (1)$$

where  $p(x)$  and  $p(y)$  are the probability mass functions ( $\rightarrow$  I/1.6.1) of  $X$  and  $Y$  and  $p(x, y)$  is the joint probability ( $\rightarrow$  I/1.3.2) mass function of  $X$  and  $Y$ .

2) The mutual information of two continuous random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as

$$I(X, Y) = - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} dy dx \quad (2)$$

where  $p(x)$  and  $p(y)$  are the probability density functions ( $\rightarrow$  I/1.6.1) of  $X$  and  $Y$  and  $p(x, y)$  is the joint probability ( $\rightarrow$  I/1.3.2) density function of  $X$  and  $Y$ .

**Sources:**

- Cover TM, Thomas JA (1991): “Relative Entropy and Mutual Information”; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

### 2.3.2 Relation to marginal and conditional entropy

**Theorem:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \quad (1)$$

where  $H(X)$  and  $H(Y)$  are the marginal entropies ( $\rightarrow$  I/2.1.1) of  $X$  and  $Y$  and  $H(X|Y)$  and  $H(Y|X)$  are the conditional entropies ( $\rightarrow$  I/2.1.4).

**Proof:** The mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is defined as

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x) p(y)}. \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(y)} - \sum_x \sum_y p(x, y) \log p(x). \quad (3)$$

Applying the law of conditional probability ( $\rightarrow$  I/1.3.4), i.e.  $p(x, y) = p(x|y) p(y)$ , we get:

$$I(X, Y) = \sum_x \sum_y p(x|y) p(y) \log p(x|y) - \sum_x \sum_y p(x, y) \log p(x). \quad (4)$$

Regrouping the variables, we have:

$$I(X, Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x \left( \sum_y p(x, y) \right) \log p(x) . \quad (5)$$

Applying the law of marginal probability ( $\rightarrow$  I/1.3.3), i.e.  $p(x) = \sum_y p(x, y)$ , we get:

$$I(X, Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x p(x) \log p(x) . \quad (6)$$

Now considering the definitions of marginal ( $\rightarrow$  I/2.1.1) and conditional ( $\rightarrow$  I/2.1.4) entropy

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ H(X|Y) &= \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) , \end{aligned} \quad (7)$$

we can finally show:

$$\begin{aligned} I(X, Y) &= -H(X|Y) + H(X) \\ &= H(X) - H(X|Y) . \end{aligned} \quad (8)$$

The conditioning of  $X$  on  $Y$  in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional entropy of  $Y$  given  $X$  is obtained by simply switching  $x$  and  $y$  in the derivation. ■

#### Sources:

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

### 2.3.3 Relation to marginal and joint entropy

**Theorem:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \quad (1)$$

where  $H(X)$  and  $H(Y)$  are the marginal entropies ( $\rightarrow$  I/2.1.1) of  $X$  and  $Y$  and  $H(X, Y)$  is the joint entropy ( $\rightarrow$  I/2.1.5).

**Proof:** The mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is defined as

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} . \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log p(x, y) - \sum_x \sum_y p(x, y) \log p(x) - \sum_x \sum_y p(x, y) \log p(y) . \quad (3)$$

Regrouping the variables, this reads:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log p(x, y) - \sum_x \left( \sum_y p(x, y) \right) \log p(x) - \sum_y \left( \sum_x p(x, y) \right) \log p(y) . \quad (4)$$

Applying the law of marginal probability ( $\rightarrow$  I/1.3.3), i.e.  $p(x) = \sum_y p(x, y)$ , we get:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log p(x, y) - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) . \quad (5)$$

Now considering the definitions of marginal ( $\rightarrow$  I/2.1.1) and joint ( $\rightarrow$  I/2.1.5) entropy

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) , \end{aligned} \quad (6)$$

we can finally show:

$$\begin{aligned} I(X, Y) &= -H(X, Y) + H(X) + H(Y) \\ &= H(X) + H(Y) - H(X, Y) . \end{aligned} \quad (7)$$

■

#### Sources:

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

### 2.3.4 Relation to joint and conditional entropy

**Theorem:** Let  $X$  and  $Y$  be discrete random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$I(X, Y) = H(X, Y) - H(X|Y) - H(Y|X) \quad (1)$$

where  $H(X, Y)$  is the joint entropy ( $\rightarrow$  I/2.1.5) of  $X$  and  $Y$  and  $H(X|Y)$  and  $H(Y|X)$  are the conditional entropies ( $\rightarrow$  I/2.1.4).

**Proof:** The existence of the joint probability mass function ( $\rightarrow$  I/1.6.1) ensures that the mutual information ( $\rightarrow$  I/2.4.1) is defined:

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} . \quad (2)$$

The relation of mutual information to conditional entropy ( $\rightarrow$  I/2.3.2) is:

$$I(X, Y) = H(X) - H(X|Y) \quad (3)$$

$$I(X, Y) = H(Y) - H(Y|X) \quad (4)$$

The relation of mutual information to joint entropy ( $\rightarrow$  I/2.3.3) is:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) . \quad (5)$$

It is true that

$$I(X, Y) = I(X, Y) + I(X, Y) - I(X, Y) . \quad (6)$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) + H(Y) - H(Y|X) - H(X) - H(Y) + H(X, Y) \\ &= H(X, Y) - H(X|Y) - H(Y|X) \end{aligned} \quad (7)$$

which proves the identity given above. ■

#### Sources:

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

## 2.4 Continuous mutual information

### 2.4.1 Definition

#### Definition:

1) The mutual information of two discrete random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as

$$I(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} \quad (1)$$

where  $p(x)$  and  $p(y)$  are the probability mass functions ( $\rightarrow$  I/1.6.1) of  $X$  and  $Y$  and  $p(x, y)$  is the joint probability ( $\rightarrow$  I/1.3.2) mass function of  $X$  and  $Y$ .

2) The mutual information of two continuous random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$  is defined as

$$I(X, Y) = - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} dy dx \quad (2)$$

where  $p(x)$  and  $p(y)$  are the probability density functions ( $\rightarrow$  I/1.6.1) of  $X$  and  $Y$  and  $p(x, y)$  is the joint probability ( $\rightarrow$  I/1.3.2) density function of  $X$  and  $Y$ .

#### Sources:

- Cover TM, Thomas JA (1991): “Relative Entropy and Mutual Information”; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.

### 2.4.2 Relation to marginal and conditional differential entropy

**Theorem:** Let  $X$  and  $Y$  be continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$\begin{aligned} I(X, Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned} \quad (1)$$

where  $h(X)$  and  $h(Y)$  are the marginal differential entropies ( $\rightarrow$  I/2.2.1) of  $X$  and  $Y$  and  $h(X|Y)$  and  $h(Y|X)$  are the conditional differential entropies ( $\rightarrow$  I/2.2.7).

**Proof:** The mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is defined as

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx . \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dx dy . \quad (3)$$

Applying the law of conditional probability ( $\rightarrow$  I/1.3.4), i.e.  $p(x, y) = p(x|y)p(y)$ , we get:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x|y)p(y) \log p(x|y) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dy dx . \quad (4)$$

Regrouping the variables, we have:

$$I(X, Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) dx dy - \int_{\mathcal{X}} \left( \int_{\mathcal{Y}} p(x, y) dy \right) \log p(x) dx . \quad (5)$$

Applying the law of marginal probability ( $\rightarrow$  I/1.3.3), i.e.  $p(x) = \int_{\mathcal{Y}} p(x, y) dy$ , we get:

$$I(X, Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) dx dy - \int_{\mathcal{X}} p(x) \log p(x) dx . \quad (6)$$

Now considering the definitions of marginal ( $\rightarrow$  I/2.2.1) and conditional ( $\rightarrow$  I/2.2.7) differential entropy

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(X|Y) &= \int_{\mathcal{Y}} p(y) h(X|Y = y) dy , \end{aligned} \quad (7)$$

we can finally show:

$$I(X, Y) = -h(X|Y) + h(X) = h(X) - h(X|Y) . \quad (8)$$

The conditioning of  $X$  on  $Y$  in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional differential entropy of  $Y$  given  $X$  is obtained by simply switching  $x$  and  $y$  in the derivation.



### Sources:

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

### 2.4.3 Relation to marginal and joint differential entropy

**Theorem:** Let  $X$  and  $Y$  be continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$I(X, Y) = h(X) + h(Y) - h(X, Y) \quad (1)$$

where  $h(X)$  and  $h(Y)$  are the marginal differential entropies ( $\rightarrow$  I/2.2.1) of  $X$  and  $Y$  and  $h(X, Y)$  is the joint differential entropy ( $\rightarrow$  I/2.2.8).

**Proof:** The mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is defined as

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx . \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(y) dy dx . \quad (3)$$

Regrouping the variables, this reads:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} \left( \int_{\mathcal{Y}} p(x, y) dy \right) \log p(x) dx - \int_{\mathcal{Y}} \left( \int_{\mathcal{X}} p(x, y) dx \right) \log p(y) dy . \quad (4)$$

Applying the law of marginal probability ( $\rightarrow$  I/1.3.3), i.e.  $p(x) = \int_{\mathcal{Y}} p(x, y) dy$ , we get:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} p(x) \log p(x) dx - \int_{\mathcal{Y}} p(y) \log p(y) dy . \quad (5)$$

Now considering the definitions of marginal ( $\rightarrow$  I/2.2.1) and joint ( $\rightarrow$  I/2.2.8) differential entropy

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(X, Y) &= - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx , \end{aligned} \quad (6)$$

we can finally show:

$$\begin{aligned} I(X, Y) &= -h(X, Y) + h(X) + h(Y) \\ &= h(X) + h(Y) - h(X, Y) . \end{aligned} \quad (7)$$

■

**Sources:**

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

**2.4.4 Relation to joint and conditional differential entropy**

**Theorem:** Let  $X$  and  $Y$  be continuous ( $\rightarrow$  I/1.2.6) random variables ( $\rightarrow$  I/1.2.2) with the joint probability ( $\rightarrow$  I/1.3.2)  $p(x, y)$  for  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  can be expressed as

$$I(X, Y) = h(X, Y) - h(X|Y) - h(Y|X) \quad (1)$$

where  $h(X, Y)$  is the joint differential entropy ( $\rightarrow$  I/2.2.8) of  $X$  and  $Y$  and  $h(X|Y)$  and  $h(Y|X)$  are the conditional differential entropies ( $\rightarrow$  I/2.2.7).

**Proof:** The existence of the joint probability density function ( $\rightarrow$  I/1.7.1) ensures that the mutual information ( $\rightarrow$  I/2.4.1) is defined:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx . \quad (2)$$

The relation of mutual information to conditional differential entropy ( $\rightarrow$  I/2.4.2) is:

$$I(X, Y) = h(X) - h(X|Y) \quad (3)$$

$$I(X, Y) = h(Y) - h(Y|X) \quad (4)$$

The relation of mutual information to joint differential entropy ( $\rightarrow$  I/2.4.3) is:

$$I(X, Y) = h(X) + h(Y) - h(X, Y) . \quad (5)$$

It is true that

$$I(X, Y) = I(X, Y) + I(X, Y) - I(X, Y) . \quad (6)$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{aligned} I(X, Y) &= h(X) - h(X|Y) + h(Y) - h(Y|X) - h(X) - h(Y) + h(X, Y) \\ &= h(X, Y) - h(X|Y) - h(Y|X) \end{aligned} \quad (7)$$

which proves the identity given above.

■

**Sources:**

- Wikipedia (2020): “Mutual information”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: [https://en.wikipedia.org/wiki/Mutual\\_information#Relation\\_to\\_conditional\\_and\\_joint\\_entropy](https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy).

## 2.5 Kullback-Leibler divergence

### 2.5.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $P$  and  $Q$  be two probability distributions ( $\rightarrow$  I/1.5.1) on  $X$ .

1) The Kullback-Leibler divergence of  $P$  from  $Q$  for a discrete random variable  $X$  is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (1)$$

where  $p(x)$  and  $q(x)$  are the probability mass functions ( $\rightarrow$  I/1.6.1) of  $P$  and  $Q$ .

2) The Kullback-Leibler divergence of  $P$  from  $Q$  for a continuous random variable  $X$  is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

where  $p(x)$  and  $q(x)$  are the probability density functions ( $\rightarrow$  I/1.7.1) of  $P$  and  $Q$ .

By convention ( $\rightarrow$  I/2.1.1),  $0 \cdot \log 0$  is taken to be zero when calculating the divergence between  $P$  and  $Q$ .

#### Sources:

- MacKay, David J.C. (2003): “Probability, Entropy, and Inference”; in: *Information Theory, Inference, and Learning Algorithms*, ch. 2.6, eq. 2.45, p. 34; URL: <https://www.inference.org.uk/itprnn/book.pdf>.

### 2.5.2 Non-negativity

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is always non-negative

$$\text{KL}[P||Q] \geq 0 \quad (1)$$

with  $\text{KL}[P||Q] = 0$ , if and only if  $P = Q$ .

**Proof:** The discrete Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (2)$$

which can be reformulated into

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) - \sum_{x \in \mathcal{X}} p(x) \cdot \log q(x) . \quad (3)$$

Gibbs’ inequality ( $\rightarrow$  I/2.1.8) states that the entropy ( $\rightarrow$  I/2.1.1) of a probability distribution is always less than or equal to the cross-entropy ( $\rightarrow$  I/2.1.6) with another probability distribution – with equality only if the distributions are identical –,

$$-\sum_{i=1}^n p(x_i) \log p(x_i) \leq -\sum_{i=1}^n p(x_i) \log q(x_i) \quad (4)$$

which can be reformulated into



$$\sum_{i=1}^n p(x_i) \log p(x_i) - \sum_{i=1}^n p(x_i) \log q(x_i) \geq 0 . \quad (5)$$

Applying (5) to (3), this proves equation (1). ■

**Sources:**

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Properties](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties).

### 2.5.3 Non-negativity

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is always non-negative

$$\text{KL}[P||Q] \geq 0 \quad (1)$$

with  $\text{KL}[P||Q] = 0$ , if and only if  $P = Q$ .

**Proof:** The discrete Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} . \quad (2)$$

The log sum inequality ( $\rightarrow$  I/2.1.9) states that

$$\sum_{i=1}^n a_i \log_c \frac{a_i}{b_i} \geq a \log_c \frac{a}{b} . \quad (3)$$

where  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  are non-negative real numbers as well as  $a = \sum_{i=1}^n a_i$  and  $b = \sum_{i=1}^n b_i$ . Because  $p(x)$  and  $q(x)$  are probability mass functions ( $\rightarrow$  I/1.6.1), such that

$$\begin{aligned} p(x) &\geq 0, & \sum_{x \in \mathcal{X}} p(x) &= 1 & \text{ and} \\ q(x) &\geq 0, & \sum_{x \in \mathcal{X}} q(x) &= 1 , \end{aligned} \quad (4)$$

theorem (1) is simply a special case of (3), i.e.

$$\text{KL}[P||Q] \stackrel{(2)}{=} \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \stackrel{(3)}{\geq} 1 \log \frac{1}{1} = 0 . \quad (5)$$

**Sources:**

- Wikipedia (2020): “Log sum inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: [https://en.wikipedia.org/wiki/Log\\_sum\\_inequality#Applications](https://en.wikipedia.org/wiki/Log_sum_inequality#Applications). ■

### 2.5.4 Non-negativity

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is always non-negative

$$\text{KL}[P||Q] \geq 0 \quad (1)$$

with  $\text{KL}[P||Q] = 0$ , if and only if  $P = Q$ .

**Proof:** The continuous Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx . \quad (2)$$

For a continuous random variable ( $\rightarrow$  I/1.2.6)  $X$  with possible values  $x \in \mathcal{X}$ , Jensen's inequality ( $\rightarrow$  I/1.12.11) implies that

$$\int_{\mathcal{X}} q(x)g(x) dx \geq g\left(\int_{\mathcal{X}} q(x)x dx\right) \quad (3)$$

where  $g(x)$  is a convex function and  $q(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of  $X$ . The negative KL divergence is equal to

$$\begin{aligned} -\text{KL}[P||Q] &= -\int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \\ &= \int_{\mathcal{X}} p(x) \cdot \log \frac{q(x)}{p(x)} dx . \end{aligned} \quad (4)$$

Applying (3) to (4), we have

$$\begin{aligned} -\text{KL}[P||Q] &\leq \log \int_{\mathcal{X}} p(x) \cdot \frac{q(x)}{p(x)} dx \\ &= \log \int_{\mathcal{X}} q(x) dx \\ &= \log 1 = 0 \end{aligned} \quad (5)$$

where the inequality sign has been reversed, because  $\log(x)$  is a concave function, and we have used the fact that a PDF integrates to one ( $\rightarrow$  I/1.7.1). From (5), we get

$$\begin{aligned} -\text{KL}[P||Q] &\leq 0 \\ \text{KL}[P||Q] &\geq 0 . \end{aligned} \quad (6)$$

■

#### Sources:

- Ostwald et al. (2014): “A tutorial on variational Bayes for latent linear stochastic time-series models”; in: *Journal of Mathematical Psychology*, vol. 60, pp. 1-19, App. A; URL: <https://www.sciencedirect.com/science/article/pii/S0022249614000352>; DOI: 10.1016/j.jmp.2014.04.0.

### 2.5.5 Non-symmetry

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is non-symmetric, i.e.

$$\text{KL}[P||Q] \neq \text{KL}[Q||P] \quad (1)$$

for some probability distributions ( $\rightarrow$  I/1.5.1)  $P$  and  $Q$ .

**Proof:** Let  $X \in \mathcal{X} = \{0, 1, 2\}$  be a discrete random variable ( $\rightarrow$  I/1.2.2) and consider the two probability distributions ( $\rightarrow$  I/1.5.1)

$$\begin{aligned} P : X &\sim \text{Bin}(2, 0.5) \\ Q : X &\sim \mathcal{U}(0, 2) \end{aligned} \quad (2)$$

where  $\text{Bin}(n, p)$  indicates a binomial distribution ( $\rightarrow$  II/1.2.1) and  $\mathcal{U}(a, b)$  indicates a discrete uniform distribution ( $\rightarrow$  II/1.0.1).

Then, the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3) entails that

$$p(x) = \begin{cases} 1/4, & \text{if } x = 0 \\ 1/2, & \text{if } x = 1 \\ 1/4, & \text{if } x = 2 \end{cases} \quad (3)$$

and the probability mass function of the discrete uniform distribution ( $\rightarrow$  II/1.0.2) entails that

$$q(x) = \frac{1}{3}, \quad (4)$$

such that the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is

$$\begin{aligned} \text{KL}[P||Q] &= \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \\ &= \frac{1}{4} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} + \frac{1}{4} \log \frac{3}{4} \\ &= \frac{1}{2} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} \\ &= \frac{1}{2} \left( \log \frac{3}{4} + \log \frac{3}{2} \right) \\ &= \frac{1}{2} \log \left( \frac{3}{4} \cdot \frac{3}{2} \right) \\ &= \frac{1}{2} \log \frac{9}{8} = 0.0589 \end{aligned} \quad (5)$$

and the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $Q$  from  $P$  is

$$\begin{aligned}
\text{KL}[Q||P] &= \sum_{x \in \mathcal{X}} q(x) \cdot \log \frac{q(x)}{p(x)} \\
&= \frac{1}{3} \log \frac{4}{3} + \frac{1}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{4}{3} \\
&= \frac{1}{3} \left( \log \frac{4}{3} + \log \frac{2}{3} + \log \frac{4}{3} \right) \\
&= \frac{1}{3} \log \left( \frac{4}{3} \cdot \frac{2}{3} \cdot \frac{4}{3} \right) \\
&= \frac{1}{3} \log \frac{32}{27} = 0.0566
\end{aligned} \tag{6}$$

which provides an example for

$$\text{KL}[P||Q] \neq \text{KL}[Q||P] \tag{7}$$

and thus proves the theorem. ■

#### Sources:

- Kullback, Solomon (1959): “Divergence”; in: *Information Theory and Statistics*, ch. 1.3, pp. 6ff.; URL: <http://index-of.co.uk/Information-Theory/Information%20theory%20and%20statistics%20-%20Solomon%20Kullback.pdf>.
- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Basic\\_example](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Basic_example).

### 2.5.6 Convexity

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is convex in pairs of probability distributions ( $\rightarrow$  I/1.5.1), i.e.

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \tag{1}$$

where  $(p_1, q_1)$  and  $(p_2, q_2)$  are two pairs of probability density functions ( $\rightarrow$  I/1.7.1) and  $0 \leq \lambda \leq 1$ .

**Proof:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

and the log sum inequality ( $\rightarrow$  I/2.1.9) states that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \tag{3}$$

where  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  are non-negative real numbers.

Thus, we can rewrite the KL divergence of the mixture distribution as

$$\begin{aligned}
& \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \\
& \stackrel{(2)}{=} \sum_{x \in \mathcal{X}} \left[ [\lambda p_1(x) + (1 - \lambda)p_2(x)] \cdot \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \right] \\
& \stackrel{(3)}{\leq} \sum_{x \in \mathcal{X}} \left[ \lambda p_1(x) \cdot \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \cdot \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)} \right] \\
& = \lambda \sum_{x \in \mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda) \sum_{x \in \mathcal{X}} p_2(x) \cdot \log \frac{p_2(x)}{q_2(x)} \\
& \stackrel{(2)}{=} \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2]
\end{aligned} \tag{4}$$

which is equivalent to (1). ■

### Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Properties](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties).
- Xie, Yao (2012): “Chain Rules and Inequalities”; in: *ECE587: Information Theory*, Lecture 3, Slides 22/24; URL: <https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf>.

### 2.5.7 Additivity for independent distributions

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is additive for independent distributions, i.e.

$$\text{KL}[P || Q] = \text{KL}[P_1 || Q_1] + \text{KL}[P_2 || Q_2] \tag{1}$$

where  $P_1$  and  $P_2$  are independent ( $\rightarrow$  I/1.3.6) distributions ( $\rightarrow$  I/1.5.1) with the joint distribution ( $\rightarrow$  I/1.5.2)  $P$ , such that  $p(x, y) = p_1(x) p_2(y)$ , and equivalently for  $Q_1$ ,  $Q_2$  and  $Q$ .

**Proof:** The continuous Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P || Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{2}$$

which, applied to the joint distributions  $P$  and  $Q$ , yields

$$\text{KL}[P || Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{q(x, y)} dy dx . \tag{3}$$

Applying  $p(x, y) = p_1(x) p_2(y)$  and  $q(x, y) = q_1(x) q_2(y)$ , we have

$$\text{KL}[P || Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x) p_2(y)}{q_1(x) q_2(y)} dy dx . \tag{4}$$

Now we can separate the logarithm and evaluate the integrals:

$$\begin{aligned}
\text{KL}[P||Q] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \left( \log \frac{p_1(x)}{q_1(x)} + \log \frac{p_2(y)}{q_2(y)} \right) dy dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x)}{q_1(x)} dy dx + \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy dx \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} \int_{\mathcal{Y}} p_2(y) dy dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \int_{\mathcal{X}} p_1(x) dx dy \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy \\
&\stackrel{(2)}{=} \text{KL}[P_1||Q_1] + \text{KL}[P_2||Q_2] .
\end{aligned} \tag{5}$$

■

**Sources:**

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Properties](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties).

**2.5.8 Invariance under parameter transformation**

**Theorem:** The Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is invariant under parameter transformation, i.e.

$$\text{KL}[p(x)||q(x)] = \text{KL}[p(y)||q(y)] \tag{1}$$

where  $y(x) = mx + n$  is an affine transformation of  $x$  and  $p(x)$  and  $q(x)$  are the probability density functions ( $\rightarrow$  I/1.7.1) of the probability distributions ( $\rightarrow$  I/1.5.1)  $P$  and  $Q$  on the continuous random variable ( $\rightarrow$  I/1.2.2)  $X$ .

**Proof:** The continuous Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) (KL divergence) is defined as

$$\text{KL}[p(x)||q(x)] = \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{2}$$

where  $a = \min(\mathcal{X})$  and  $b = \max(\mathcal{X})$  are the lower and upper bound of the possible outcomes  $\mathcal{X}$  of  $X$ . Since  $y(x) = mx + n$  is a strictly increasing ( $\rightarrow$  I/1.7.3) or strictly decreasing ( $\rightarrow$  I/1.7.4) function, we have the identity of the differentials

$$\begin{aligned}
p(x) dx &= p(y) dy \\
q(x) dx &= q(y) dy
\end{aligned} \tag{3}$$

which can be rearranged into

$$\begin{aligned}
p(x) &= p(y) \frac{dy}{dx} \\
q(x) &= q(y) \frac{dy}{dx} ,
\end{aligned} \tag{4}$$

the KL divergence can be evaluated as follows:

$$\begin{aligned}
\text{KL}[p(x)||q(x)] &= \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \\
&= \int_{y(a)}^{y(b)} p(y) \frac{dy}{dx} \cdot \log \left( \frac{p(y) \frac{dy}{dx}}{q(y) \frac{dy}{dx}} \right) dx \\
&= \int_{y(a)}^{y(b)} p(y) \cdot \log \frac{p(y)}{q(y)} dy \\
&= \text{KL}[p(y)||q(y)] .
\end{aligned} \tag{5}$$

■

**Sources:**

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Properties](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties).
- shimao (2018): “KL divergence invariant to affine transformation?”; in: *StackExchange CrossValidated*, retrieved on 2020-05-28; URL: <https://stats.stackexchange.com/questions/341922/kl-divergence-invariant-to-affine-transformation>.

**2.5.9 Relation to discrete entropy**

**Theorem:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $P$  and  $Q$  be two probability distributions ( $\rightarrow$  I/1.5.1) on  $X$ . Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  can be expressed as

$$\text{KL}[P||Q] = H(P, Q) - H(P) \tag{1}$$

where  $H(P, Q)$  is the cross-entropy ( $\rightarrow$  I/2.1.6) of  $P$  and  $Q$  and  $H(P)$  is the marginal entropy ( $\rightarrow$  I/2.1.1) of  $P$ .

**Proof:** The discrete Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

where  $p(x)$  and  $q(x)$  are the probability mass functions ( $\rightarrow$  I/1.6.1) of  $P$  and  $Q$ . Separating the logarithm, we have:

$$\text{KL}[P||Q] = - \sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) . \tag{3}$$

Now considering the definitions of marginal entropy ( $\rightarrow$  I/2.1.1) and cross-entropy ( $\rightarrow$  I/2.1.6)

$$\begin{aligned}
H(P) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
H(P, Q) &= - \sum_{x \in \mathcal{X}} p(x) \log q(x) ,
\end{aligned} \tag{4}$$

we can finally show:

$$\text{KL}[P||Q] = H(P, Q) - H(P) . \tag{5}$$

**Sources:**

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Motivation](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation).

**2.5.10 Relation to differential entropy**

**Theorem:** Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2) with possible outcomes  $\mathcal{X}$  and let  $P$  and  $Q$  be two probability distributions ( $\rightarrow$  I/1.5.1) on  $X$ . Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  can be expressed as

$$\text{KL}[P||Q] = h(P, Q) - h(P) \quad (1)$$

where  $h(P, Q)$  is the differential cross-entropy ( $\rightarrow$  I/2.2.9) of  $P$  and  $Q$  and  $h(P)$  is the marginal differential entropy ( $\rightarrow$  I/2.2.1) of  $P$ .

**Proof:** The continuous Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

where  $p(x)$  and  $q(x)$  are the probability density functions ( $\rightarrow$  I/1.7.1) of  $P$  and  $Q$ . Separating the logarithm, we have:

$$\text{KL}[P||Q] = - \int_{\mathcal{X}} p(x) \log q(x) dx + \int_{\mathcal{X}} p(x) \log p(x) dx . \quad (3)$$

Now considering the definitions of marginal differential entropy ( $\rightarrow$  I/2.2.1) and differential cross-entropy ( $\rightarrow$  I/2.2.9)

$$\begin{aligned} h(P) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(P, Q) &= - \int_{\mathcal{X}} p(x) \log q(x) dx , \end{aligned} \quad (4)$$

we can finally show:

$$\text{KL}[P||Q] = h(P, Q) - h(P) . \quad (5)$$

**Sources:**

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence#Motivation](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation).



## 3 Estimation theory

### 3.1 Basic concepts of estimation

#### 3.1.1 Estimator

**Definition:** Let  $y \in \mathcal{Y}$  be measured data ( $\rightarrow$  I/1.1.5), governed by a probability distribution ( $\rightarrow$  I/1.5.1) described by some statistical parameter ( $\rightarrow$  I/1.5.7)  $\theta \in \Theta$ . Then, a function  $\hat{\theta} : \mathcal{Y} \rightarrow \Theta$  exemplifying a rule for calculating an estimate of  $\theta$  from  $y$  is called an “estimator”. Estimation theory distinguishes:

- the quantify of interest  $\theta$  is called the “estimand”;
- the rule  $\hat{\theta}$  for estimating it is called “estimator”;
- the result of estimation  $\hat{\theta}(y)$  is called “estimate”.

**Sources:**

- Ostwald, Dirk (2023): “Punktschätzung”; in: *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*, Einheit (9), Folie 7; URL: [https://www.ipsy.ovgu.de/ipsy\\_\\_media/Methodenlehre+I/Wintersemester2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/9\\_Punktsch%C3%A4tzung.pdf](https://www.ipsy.ovgu.de/ipsy__media/Methodenlehre+I/Wintersemester2324/Wahrscheinlichkeitstheorie+und+Frequentistische+Inferenz/9_Punktsch%C3%A4tzung.pdf).
- Wikipedia (2024): “Estimator”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-11-01; URL: <https://en.wikipedia.org/wiki/Estimator#Definition>.

#### 3.1.2 Biased vs. unbiased

**Definition:** Let  $\hat{\theta} : \mathcal{Y} \rightarrow \Theta$  be an estimator ( $\rightarrow$  I/3.1.1) of a parameter ( $\rightarrow$  I/1.5.7)  $\theta \in \Theta$  from data ( $\rightarrow$  I/1.1.5)  $y \in \mathcal{Y}$ . Then,

- $\hat{\theta}$  is called an unbiased estimator when its expected value ( $\rightarrow$  I/1.12.1) is equal to the parameter that it is estimating:  $E_{\hat{\theta}}(\hat{\theta}) = \theta$ , where the expectation is calculated over all possible samples  $y$  leading to values of  $\hat{\theta}$ .
- $\hat{\theta}$  is called a biased estimator otherwise, i.e. when  $E_{\hat{\theta}}(\hat{\theta}) \neq \theta$ .

**Sources:**

- Wikipedia (2024): “Estimator”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-11-08; URL: <https://en.wikipedia.org/wiki/Estimator#Bias>.

## 3.2 Point estimates

### 3.2.1 Mean squared error

**Definition:** Let  $\hat{\theta}$  be an estimator ( $\rightarrow$  I/3.1.1) of an unknown parameter ( $\rightarrow$  I/1.5.7)  $\theta$  based on measured data ( $\rightarrow$  I/1.1.5)  $y$ . Then, the mean squared error is defined as the expected value ( $\rightarrow$  I/1.12.1) of the squared difference between the estimated value and the true value of the parameter:

$$\text{MSE} = E_{\hat{\theta}} \left[ \left( \hat{\theta} - \theta \right)^2 \right] . \quad (1)$$

where  $E_{\hat{\theta}} [\cdot]$  is expectation calculated over all possible samples ( $\rightarrow$  I/1.1.6)  $y$  leading to values of  $\hat{\theta}$ .

**Sources:**

- Wikipedia (2022): “Estimator”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: [https://en.wikipedia.org/wiki/Estimator#Mean\\_squared\\_error](https://en.wikipedia.org/wiki/Estimator#Mean_squared_error).

### 3.2.2 Partition of the mean squared error into bias and variance

**Theorem:** The mean squared error ( $\rightarrow$  I/3.2.1) can be partitioned into variance and squared bias

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2 \quad (1)$$

where the variance ( $\rightarrow$  I/1.13.1) is given by

$$\text{Var}(\hat{\theta}) = E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] \quad (2)$$

and the bias is given by

$$\text{Bias}(\hat{\theta}, \theta) = \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right). \quad (3)$$

**Proof:** The mean squared error ( $\rightarrow$  I/3.2.1) (MSE) is defined as the expected value ( $\rightarrow$  I/1.12.1) of the squared deviation of the estimated value  $\hat{\theta}$  from the true value  $\theta$  of a parameter, over all values  $\hat{\theta}$ :

$$\text{MSE}(\hat{\theta}) = E_{\hat{\theta}} \left[ \left( \hat{\theta} - \theta \right)^2 \right]. \quad (4)$$

This formula can be evaluated in the following way:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - \theta \right)^2 \right] \\ &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) + E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right] \\ &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 + 2 \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right) \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right) + \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right] \\ &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + E_{\hat{\theta}} \left[ 2 \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right) \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right) \right] + E_{\hat{\theta}} \left[ \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right]. \end{aligned} \quad (5)$$

Because  $E_{\hat{\theta}}(\hat{\theta}) - \theta$  is constant as a function of  $\hat{\theta}$ , we have:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + 2 \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right) E_{\hat{\theta}} \left[ \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right] + \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \\ &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + 2 \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right) \left( E_{\hat{\theta}}(\hat{\theta}) - E_{\hat{\theta}}(\hat{\theta}) \right) + \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \\ &= E_{\hat{\theta}} \left[ \left( \hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + \left( E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2. \end{aligned} \quad (6)$$

This proves the partition given by (1). ■

**Sources:**

- Wikipedia (2019): “Mean squared error”; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-11-27; URL: [https://en.wikipedia.org/wiki/Mean\\_squared\\_error#Proof\\_of\\_variance\\_and\\_bias\\_relationship](https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship).

**3.3 Interval estimates****3.3.1 Confidence interval**

**Definition:** Let  $y$  be a random sample ( $\rightarrow$  I/1.1.6) from a probability distributions ( $\rightarrow$  I/1.5.1) governed by a parameter ( $\rightarrow$  I/1.5.7) of interest  $\theta$  and quantities not of interest  $\varphi$ . A confidence interval for  $\theta$  is defined as an interval  $[u(y), v(y)]$  determined by the random variables ( $\rightarrow$  I/1.2.2)  $u(y)$  and  $v(y)$  with the property

$$\Pr(u(y) < \theta < v(y) \mid \theta, \varphi) = \gamma \quad \text{for all } (\theta, \varphi). \quad (1)$$

where  $\gamma = 1 - \alpha$  is called the confidence level.

**Sources:**

- Wikipedia (2022): “Confidence interval”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: [https://en.wikipedia.org/wiki/Confidence\\_interval#Definition](https://en.wikipedia.org/wiki/Confidence_interval#Definition).

**3.3.2 Construction of confidence intervals using Wilks’ theorem**

**Theorem:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) for measured data  $y$  with model parameters  $\theta \in \Theta$ , consisting of a parameter of interest  $\phi \in \Phi$  and nuisance parameters  $\lambda \in \Lambda$ :

$$m : p(y \mid \theta) = \mathcal{D}(y; \theta), \quad \theta = \{\phi, \lambda\}. \quad (1)$$

Further, let  $\hat{\theta}$  be an estimate of  $\theta$ , obtained using maximum-likelihood-estimation ( $\rightarrow$  I/4.1.3):

$$\hat{\theta} = \arg \max_{\theta} \log p(y \mid \theta), \quad \hat{\theta} = \{\hat{\phi}, \hat{\lambda}\}. \quad (2)$$

Then, an asymptotic confidence interval ( $\rightarrow$  I/3.3.1) for  $\theta$  is given by

$$\text{CI}_{1-\alpha}(\hat{\phi}) = \left\{ \phi \mid \log p(y \mid \phi, \hat{\lambda}) \geq \log p(y \mid \hat{\phi}, \hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2 \right\} \quad (3)$$

where  $1 - \alpha$  is the confidence level and  $\chi_{1,1-\alpha}^2$  is the  $(1 - \alpha)$ -quantile of the chi-squared distribution ( $\rightarrow$  II/3.7.1) with 1 degree of freedom ( $\rightarrow$  I/1.5.12).

**Proof:** The confidence interval ( $\rightarrow$  I/3.3.1) is defined as the interval that, under infinitely repeated random experiments ( $\rightarrow$  I/1.1.1), contains the true parameter value with a certain probability.

Let us define the likelihood ratio ( $\rightarrow$  I/4.1.6)

$$\Lambda(\phi) = \frac{p(y \mid \phi, \hat{\lambda})}{p(y \mid \hat{\phi}, \hat{\lambda})} \quad \text{for all } \phi \in \Phi \quad (4)$$

and compute the log-likelihood ratio ( $\rightarrow$  I/4.1.7)

$$\log \Lambda(\phi) = \log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) . \quad (5)$$

Wilks' theorem states that, when comparing two statistical models with parameter spaces  $\Theta_1$  and  $\Theta_0 \subset \Theta_1$ , as the sample size ( $\rightarrow$  I/1.1.7) approaches infinity, the quantity calculated as  $-2$  times the log-ratio of maximum likelihoods follows a chi-squared distribution ( $\rightarrow$  II/3.7.1), if the null hypothesis is true:

$$H_0 : \theta \in \Theta_0 \quad \Rightarrow \quad -2 \log \frac{\max_{\theta \in \Theta_0} p(y|\theta)}{\max_{\theta \in \Theta_1} p(y|\theta)} \sim \chi_{\Delta k}^2 \quad \text{as } n \rightarrow \infty \quad (6)$$

where  $\Delta k$  is the difference in dimensionality between  $\Theta_0$  and  $\Theta_1$ . Applied to our example in (5), we note that  $\Theta_1 = \{\phi, \hat{\phi}\}$  and  $\Theta_0 = \{\phi\}$ , such that  $\Delta k = 1$  and Wilks' theorem implies:

$$-2 \log \Lambda(\phi) \sim \chi_1^2 . \quad (7)$$

Using the quantile function ( $\rightarrow$  I/1.11.1)  $\chi_{k,p}^2$  of the chi-squared distribution ( $\rightarrow$  II/3.7.1), an  $(1 - \alpha)$ -confidence interval is therefore given by all values  $\phi$  that satisfy

$$-2 \log \Lambda(\phi) \leq \chi_{1,1-\alpha}^2 . \quad (8)$$

Applying (5) and rearranging, we can evaluate

$$\begin{aligned} -2 \left[ \log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) \right] &\leq \chi_{1,1-\alpha}^2 \\ \log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) &\geq -\frac{1}{2} \chi_{1,1-\alpha}^2 \\ \log p(y|\phi, \hat{\lambda}) &\geq \log p(y|\hat{\phi}, \hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2 \end{aligned} \quad (9)$$

which is equivalent to the confidence interval given by (3). ■

#### Sources:

- Wikipedia (2020): “Confidence interval”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: [https://en.wikipedia.org/wiki/Confidence\\_interval#Methods\\_of\\_derivation](https://en.wikipedia.org/wiki/Confidence_interval#Methods_of_derivation).
- Wikipedia (2020): “Likelihood-ratio test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: [https://en.wikipedia.org/wiki/Likelihood-ratio\\_test#Definition](https://en.wikipedia.org/wiki/Likelihood-ratio_test#Definition).
- Wikipedia (2020): “Wilks' theorem”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: [https://en.wikipedia.org/wiki/Wilks%27\\_theorem](https://en.wikipedia.org/wiki/Wilks%27_theorem).

## 4 Frequentist statistics

### 4.1 Likelihood theory

#### 4.1.1 Likelihood function

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of the distribution of  $y$  given  $\theta$  is called the likelihood function of  $m$ :

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) . \quad (1)$$

#### 4.1.2 Log-likelihood function

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the logarithm of the probability density function ( $\rightarrow$  I/1.7.1) of the distribution of  $y$  given  $\theta$  is called the log-likelihood function ( $\rightarrow$  I/5.1.2) of  $m$ :

$$\text{LL}_m(\theta) = \log p(y|\theta, m) = \log \mathcal{D}(y; \theta) . \quad (1)$$

#### 4.1.3 Maximum likelihood estimation

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the parameter values maximizing the likelihood function ( $\rightarrow$  I/5.1.2) or log-likelihood function ( $\rightarrow$  I/4.1.2) are called “maximum likelihood estimates” of  $\theta$ :

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}_m(\theta) = \arg \max_{\theta} \text{LL}_m(\theta) . \quad (1)$$

The process of calculating  $\hat{\theta}$  is called “maximum likelihood estimation” and the functional form leading from  $y$  to  $\hat{\theta}$  given  $m$  is called “maximum likelihood estimator”. Maximum likelihood estimation, estimator and estimates may all be abbreviated as “MLE”.

#### 4.1.4 Maximum log-likelihood

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the maximum log-likelihood (MLL) of  $m$  is the maximal value of the log-likelihood function ( $\rightarrow$  I/4.1.2) of this model:

$$\text{MLL}(m) = \max_{\theta} \text{LL}_m(\theta) . \quad (1)$$

The maximum log-likelihood can be obtained by plugging the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) into the log-likelihood function ( $\rightarrow$  I/4.1.2).

#### 4.1.5 MLE can be biased

**Theorem:** Maximum likelihood estimation ( $\rightarrow$  I/4.1.3) can result in biased estimates ( $\rightarrow$  I/3.1.2) of model parameters ( $\rightarrow$  I/1.5.7), i.e. estimates ( $\rightarrow$  I/3.1.1) whose long-term expected value ( $\rightarrow$  I/1.12.1) is unequal to the quantities they estimate:

$$\text{E} \left[ \hat{\theta}_{\text{MLE}} \right] = \text{E} \left[ \arg \max_{\theta} \text{LL}_m(\theta) \right] \neq \theta . \quad (1)$$

**Proof:** Consider a set of independent and identical normally distributed observations ( $\rightarrow$  III/1.1.1)  $x = \{x_1, \dots, x_n\}$  with unknown mean ( $\rightarrow$  I/1.12.1)  $\mu$  and variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ :

$$x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (2)$$

Then, we know that the maximum likelihood estimator for the variance ( $\rightarrow$  III/1.1.2)  $\sigma^2$  is underestimating the true variance of the data distribution ( $\rightarrow$  IV/1.0.2):

$$\mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] = \frac{n-1}{n} \sigma^2 \neq \sigma^2. \quad (3)$$

This proves the existence of cases such as those stated by the theorem. ■

#### 4.1.6 Likelihood ratio

**Definition:** Let  $m_0$  and  $m_1$  be two generative models ( $\rightarrow$  I/5.1.1) describing the same measured data  $y$  using different model parameters  $\theta_0 \in \Theta_0$  and  $\theta_1 \in \Theta_1$ . Then, the quotient of the maximized likelihood functions ( $\rightarrow$  I/4.1.3) of these two models is denoted as  $\Lambda_{01}$  and is called the likelihood ( $\rightarrow$  I/5.1.2) ratio of  $m_0$  relative to  $m_1$ :

$$\Lambda_{01} = \frac{\max_{\theta_0 \in \Theta_0} \mathcal{L}_{m_0}(\theta_0)}{\max_{\theta_1 \in \Theta_1} \mathcal{L}_{m_1}(\theta_1)} = \frac{p(y|\hat{\theta}_0, m_0)}{p(y|\hat{\theta}_1, m_1)}. \quad (1)$$

##### Sources:

- Wikipedia (2024): “Neyman-Pearson lemma”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-14; URL: [https://en.wikipedia.org/wiki/Neyman%E2%80%93Pearson\\_lemma#Example](https://en.wikipedia.org/wiki/Neyman%E2%80%93Pearson_lemma#Example).

#### 4.1.7 Log-likelihood ratio

**Definition:** Let  $m_0$  and  $m_1$  be two generative models ( $\rightarrow$  I/5.1.1) describing the same measured data  $y$  using different model parameters  $\theta_0 \in \Theta_0$  and  $\theta_1 \in \Theta_1$ . Then, the logarithmized quotient of the maximized likelihood functions ( $\rightarrow$  I/4.1.3) of these two models is denoted as  $\log \Lambda_{01}$  and is called the log-likelihood ( $\rightarrow$  I/4.1.2) ratio of  $m_0$  relative to  $m_1$ :

$$\log \Lambda_{01} = \log \frac{\max_{\theta_0 \in \Theta_0} \mathcal{L}_{m_0}(\theta_0)}{\max_{\theta_1 \in \Theta_1} \mathcal{L}_{m_1}(\theta_1)} = \log p(y|\hat{\theta}_0, m_0) - \log p(y|\hat{\theta}_1, m_1). \quad (1)$$

##### Sources:

- Wikipedia (2024): “Likelihood-ratio test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-14; URL: [https://en.wikipedia.org/wiki/Likelihood-ratio\\_test#Definition](https://en.wikipedia.org/wiki/Likelihood-ratio_test#Definition).

#### 4.1.8 Method of moments

**Definition:** Let measured data  $y$  follow a probability distribution ( $\rightarrow$  I/1.5.1) with probability mass ( $\rightarrow$  I/1.6.1) or probability density ( $\rightarrow$  I/1.7.1)  $p(y|\theta)$  governed by unknown parameters  $\theta_1, \dots, \theta_k$ . Then, method-of-moments estimation, also referred to as “method of moments” or “matching the moments”, consists in

1) expressing the first  $k$  moments ( $\rightarrow$  I/1.21.1) of  $y$  in terms of  $\theta$

$$\begin{aligned}\mu_1 &= f_1(\theta_1, \dots, \theta_k) \\ &\vdots \\ \mu_k &= f_k(\theta_1, \dots, \theta_k) ,\end{aligned}\tag{1}$$

2) calculating the first  $k$  sample moments ( $\rightarrow$  I/1.21.1) from  $y$

$$\hat{\mu}_1(y), \dots, \hat{\mu}_k(y)\tag{2}$$

3) and solving the system of  $k$  equations

$$\begin{aligned}\hat{\mu}_1(y) &= f_1(\hat{\theta}_1, \dots, \hat{\theta}_k) \\ &\vdots \\ \hat{\mu}_k(y) &= f_k(\hat{\theta}_1, \dots, \hat{\theta}_k)\end{aligned}\tag{3}$$

for  $\hat{\theta}_1, \dots, \hat{\theta}_k$ , which are subsequently referred to as “method-of-moments estimates”.

#### Sources:

- Wikipedia (2021): “Method of moments (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: [https://en.wikipedia.org/wiki/Method\\_of\\_moments\\_\(statistics\)#Method](https://en.wikipedia.org/wiki/Method_of_moments_(statistics)#Method).

## 4.2 Statistical hypotheses

### 4.2.1 Statistical hypothesis

**Definition:** A statistical hypothesis is a statement about the parameters of a distribution describing a population from which observations can be sampled as measured data ( $\rightarrow$  I/1.1.5).

More precisely, let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) describing measured data  $y$  in terms of a distribution  $\mathcal{D}(\theta)$  with model parameters  $\theta \in \Theta$ . Then, a statistical hypothesis is formally specified as

$$H : \theta \in \Theta^* \quad \text{where} \quad \Theta^* \subset \Theta .\tag{1}$$

#### Sources:

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#Definition\\_of\\_terms](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms).

### 4.2.2 Simple vs. composite

**Definition:** Let  $H$  be a statistical hypothesis ( $\rightarrow$  I/4.2.1). Then,

- $H$  is called a simple hypothesis, if it completely specifies the population distribution; in this case, the sampling distribution ( $\rightarrow$  I/1.5.6) of the test statistic ( $\rightarrow$  I/4.3.5) is a function of sample size ( $\rightarrow$  I/1.1.7) alone.
- $H$  is called a composite hypothesis, if it does not completely specify the population distribution; for example, the hypothesis may only specify one parameter of the distribution and leave others unspecified.

**Sources:**

- Wikipedia (2021): “Exclusion of the null hypothesis”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Exclusion\\_of\\_the\\_null\\_hypothesis#Terminology](https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Terminology).

### 4.2.3 Point/exact vs. set/inexact

**Definition:** Let  $H$  be a statistical hypothesis ( $\rightarrow$  I/4.2.1). Then,

- $H$  is called a point hypothesis or exact hypothesis, if it specifies an exact parameter value:

$$H : \theta = \theta^* ; \quad (1)$$

- $H$  is called a set hypothesis or inexact hypothesis, if it specifies a set of possible values with more than one element for the parameter value (e.g. a range or an interval):

$$H : \theta \in \Theta^* . \quad (2)$$

**Sources:**

- Wikipedia (2021): “Exclusion of the null hypothesis”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Exclusion\\_of\\_the\\_null\\_hypothesis#Terminology](https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Terminology).

### 4.2.4 One-tailed vs. two-tailed

**Definition:** Let  $H_0$  be a point ( $\rightarrow$  I/4.2.3) null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \theta = \theta_0 \quad (1)$$

and consider a set ( $\rightarrow$  I/4.2.3) alternative hypothesis ( $\rightarrow$  I/4.3.3)  $H_1$ . Then,

- $H_1$  is called a left-sided one-tailed hypothesis, if  $\theta$  is assumed to be smaller than  $\theta_0$ :

$$H_1 : \theta < \theta_0 ; \quad (2)$$

- $H_1$  is called a right-sided one-tailed hypothesis, if  $\theta$  is assumed to be larger than  $\theta_0$ :

$$H_1 : \theta > \theta_0 ; \quad (3)$$

- $H_1$  is called a two-tailed hypothesis, if  $\theta$  is assumed to be unequal to  $\theta_0$ :

$$H_1 : \theta \neq \theta_0 . \quad (4)$$

**Sources:**



- Wikipedia (2021): “One- and two-tailed tests”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: [https://en.wikipedia.org/wiki/One\\_and\\_two-tailed\\_tests](https://en.wikipedia.org/wiki/One_and_two-tailed_tests).

## 4.3 Hypothesis testing

### 4.3.1 Statistical test

**Definition:** Let  $y$  be a set of measured data ( $\rightarrow$  I/1.1.5). Then, a statistical hypothesis test consists of the following:

- an assumption about the distribution ( $\rightarrow$  I/1.5.1) of the data, often expressed in terms of a statistical model ( $\rightarrow$  I/5.1.1)  $m$ ;
- a null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$  and an alternative hypothesis ( $\rightarrow$  I/4.3.3)  $H_1$  which make specific statements about the distribution of the data;
- a test statistic ( $\rightarrow$  I/4.3.5)  $T(Y)$  which is a function of the data and whose distribution under the null hypothesis ( $\rightarrow$  I/4.3.2) is known;
- a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$  which imposes an upper bound on the probability ( $\rightarrow$  I/1.3.1) of rejecting  $H_0$ , given that  $H_0$  is true.

Procedurally, the statistical hypothesis test works as follows:

- Given the null hypothesis  $H_0$  and the significance level  $\alpha$ , a critical value ( $\rightarrow$  I/4.3.9)  $t_{\text{crit}}$  is determined which partitions the set of possible values of  $T(Y)$  into “acceptance region” and “rejection region”.
- Then, the observed test statistic ( $\rightarrow$  I/4.3.5)  $t_{\text{obs}} = T(y)$  is calculated from the actually measured data  $y$ . If it is in the rejection region,  $H_0$  is rejected in favor of  $H_1$ . Otherwise, the test fails to reject  $H_0$ .

**Sources:**

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#The\\_testing\\_process](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#The_testing_process).

### 4.3.2 Null hypothesis

**Definition:** The statement which is tested in a statistical hypothesis test ( $\rightarrow$  I/4.3.1) is called the “null hypothesis”, denoted as  $H_0$ . The test is designed to assess the strength of evidence against  $H_0$  and possibly reject it. The opposite of  $H_0$  is called the “alternative hypothesis ( $\rightarrow$  I/4.3.3)”. Usually,  $H_0$  is a statement that a particular parameter is zero, that there is no effect of a particular treatment or that there is no difference between particular conditions.

More precisely, let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) describing measured data  $y$  using model parameters  $\theta \in \Theta$ . Then, a null hypothesis is formally specified as

$$H_0 : \theta \in \Theta_0 \quad \text{where} \quad \Theta_0 \subset \Theta . \quad (1)$$

**Sources:**

- Wikipedia (2021): “Exclusion of the null hypothesis”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Exclusion\\_of\\_the\\_null\\_hypothesis#Basic\\_definitions](https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Basic_definitions).

### 4.3.3 Alternative hypothesis

**Definition:** Let  $H_0$  be a null hypothesis ( $\rightarrow$  I/4.3.2) of a statistical hypothesis test ( $\rightarrow$  I/4.3.1). Then, the corresponding alternative hypothesis, denoted as  $H_1$ , is either the negation of  $H_0$  or an interesting sub-case in the negation of  $H_0$ , depending on context. The test is designed to assess the strength of evidence against  $H_0$  and possibly reject it in favor of  $H_1$ . Usually,  $H_1$  is a statement that a particular parameter is non-zero, that there is an effect of a particular treatment or that there is a difference between particular conditions.

More precisely, let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) describing measured data  $y$  using model parameters  $\theta \in \Theta$ . Then, null and alternative hypothesis are formally specified as

$$\begin{aligned} H_0 : \theta \in \Theta_0 \quad \text{where} \quad \Theta_0 \subset \Theta \\ H_1 : \theta \in \Theta_1 \quad \text{where} \quad \Theta_1 = \Theta \setminus \Theta_0 . \end{aligned} \tag{1}$$

**Sources:**

- Wikipedia (2021): “Exclusion of the null hypothesis”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Exclusion\\_of\\_the\\_null\\_hypothesis#Basic\\_definitions](https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Basic_definitions).

### 4.3.4 One-tailed vs. two-tailed

**Definition:** Let there be a statistical test ( $\rightarrow$  I/4.3.1) of an alternative hypothesis ( $\rightarrow$  I/4.3.3)  $H_1$  against a null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$ . Then,

- the test is called a one-tailed test, if  $H_1$  is a one-tailed hypothesis ( $\rightarrow$  I/4.2.4);
- the test is called a two-tailed test, if  $H_1$  is a two-tailed hypothesis ( $\rightarrow$  I/4.2.4).

The fact whether a test ( $\rightarrow$  I/4.3.1) is one-tailed or two-tailed has consequences for the computation of critical value ( $\rightarrow$  I/4.3.9) and p-value ( $\rightarrow$  I/4.3.10).

**Sources:**

- Wikipedia (2021): “One- and two-tailed tests”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: [https://en.wikipedia.org/wiki/One-\\_and\\_two-tailed\\_tests](https://en.wikipedia.org/wiki/One-_and_two-tailed_tests).

### 4.3.5 Test statistic

**Definition:** In a statistical hypothesis test ( $\rightarrow$  I/4.3.1), the test statistic  $T(Y)$  is a scalar function of the measured data ( $\rightarrow$  I/1.1.5)  $y$  whose distribution under the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$  can be established. Together with a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ , this distribution implies a critical value ( $\rightarrow$  I/4.3.9)  $t_{\text{crit}}$  of the test statistic which determines whether the test rejects or fails to reject  $H_0$ .

**Sources:**

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#The\\_testing\\_process](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#The_testing_process).

#### 4.3.6 Size of a test

**Definition:** Let there be a statistical hypothesis test ( $\rightarrow$  I/4.3.1) with null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$ . Then, the size of the test is the probability of a false-positive result or making a type I error, i.e. the probability ( $\rightarrow$  I/1.3.1) of rejecting the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$ , given that  $H_0$  is actually true.

For a simple null hypothesis ( $\rightarrow$  I/4.2.2), the size is determined by the following conditional probability ( $\rightarrow$  I/1.3.4):

$$\Pr(\text{test rejects } H_0 | H_0) . \quad (1)$$

For a composite null hypothesis ( $\rightarrow$  I/4.2.2), the size is the supremum over all possible realizations of the null hypothesis ( $\rightarrow$  I/4.3.2):

$$\sup_{h \in H_0} \Pr(\text{test rejects } H_0 | h) . \quad (2)$$

**Sources:**

- Wikipedia (2021): “Size (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Size\\_\(statistics\)](https://en.wikipedia.org/wiki/Size_(statistics)).

#### 4.3.7 Power of a test

**Definition:** Let there be a statistical hypothesis test ( $\rightarrow$  I/4.3.1) with null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$  and alternative hypothesis ( $\rightarrow$  I/4.3.3)  $H_1$ . Then, the power of the test is the probability of a true-positive result or not making a type II error, i.e. the probability ( $\rightarrow$  I/1.3.1) of rejecting  $H_0$ , given that  $H_1$  is actually true.

For given null ( $\rightarrow$  I/4.3.2) and alternative ( $\rightarrow$  I/4.3.3) hypothesis ( $\rightarrow$  I/4.2.1), the size is determined by the following conditional probability ( $\rightarrow$  I/1.3.4):

$$\Pr(\text{test rejects } H_0 | H_1) . \quad (1)$$

**Sources:**

- Wikipedia (2021): “Power of a test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: [https://en.wikipedia.org/wiki/Power\\_of\\_a\\_test#Description](https://en.wikipedia.org/wiki/Power_of_a_test#Description).

#### 4.3.8 Significance level

**Definition:** Let the size ( $\rightarrow$  I/4.3.6) of a statistical hypothesis test ( $\rightarrow$  I/4.3.1) be the probability of a false-positive result or making a type I error, i.e. the probability ( $\rightarrow$  I/1.3.1) of rejecting the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$ , given that  $H_0$  is actually true:

$$\Pr(\text{test rejects } H_0 | H_0) . \quad (1)$$

Then, the test is said to have significance level  $\alpha$ , if the size is less than or equal to  $\alpha$ :

$$\Pr(\text{test rejects } H_0 | H_0) \leq \alpha . \quad (2)$$

**Sources:**

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#Definition\\_of\\_terms](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms).

#### 4.3.9 Critical value

**Definition:** In a statistical hypothesis test ( $\rightarrow$  I/4.3.1), the critical value ( $\rightarrow$  I/4.3.9)  $t_{\text{crit}}$  is that value of the test statistic ( $\rightarrow$  I/4.3.5)  $T(Y)$  which partitions the set of possible test statistics into “acceptance region” and “rejection region” based on a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ . Put differently, if the observed test statistic  $t_{\text{obs}} = T(y)$  computed from actually measured data ( $\rightarrow$  I/1.1.5)  $y$  is as extreme or more extreme than the critical value, the test rejects the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$  in favor of the alternative hypothesis ( $\rightarrow$  I/4.3.3).

**Sources:**

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#Definition\\_of\\_terms](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms).

#### 4.3.10 p-value

**Definition:** Let there be a statistical test ( $\rightarrow$  I/4.3.1) of the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0$  and the alternative hypothesis ( $\rightarrow$  I/4.3.3)  $H_1$  using the test statistic ( $\rightarrow$  I/4.3.5)  $T(Y)$ . Let  $y$  be the measured data ( $\rightarrow$  I/1.1.5) and let  $t_{\text{obs}} = T(y)$  be the observed test statistic computed from  $y$ . Moreover, assume that  $F_T(t)$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) (CDF) of the distribution of  $T(Y)$  under  $H_0$ .

Then, the p-value is the probability of obtaining a test statistic more extreme than or as extreme as  $t_{\text{obs}}$ , given that the null hypothesis  $H_0$  is true:

- $p = F_T(t_{\text{obs}})$ , if  $H_1$  is a left-sided one-tailed hypothesis ( $\rightarrow$  I/4.2.4);
- $p = 1 - F_T(t_{\text{obs}})$ , if  $H_1$  is a right-sided one-tailed hypothesis ( $\rightarrow$  I/4.2.4);
- $p = 2 \cdot \min([F_T(t_{\text{obs}}), 1 - F_T(t_{\text{obs}})])$ , if  $H_1$  is a two-tailed hypothesis ( $\rightarrow$  I/4.2.4).

**Sources:**

- Wikipedia (2021): “Statistical hypothesis testing”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: [https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#Definition\\_of\\_terms](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms).

#### 4.3.11 Distribution of p-value under null hypothesis

**Theorem:** Under the null hypothesis ( $\rightarrow$  I/4.3.2), the p-value ( $\rightarrow$  I/4.3.10) in a statistical test ( $\rightarrow$  I/4.3.1) follows a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$p \sim \mathcal{U}(0, 1) . \quad (1)$$

**Proof:** Without loss of generality, consider a left-sided one-tailed hypothesis test ( $\rightarrow$  I/4.2.4). Then, the p-value is a function of the test statistic ( $\rightarrow$  I/4.3.10)

$$\begin{aligned} P &= F_T(T) \\ p &= F_T(t_{\text{obs}}) \end{aligned} \quad (2)$$

where  $t_{\text{obs}}$  is the observed test statistic ( $\rightarrow$  I/4.3.5) and  $F_T(t)$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the test statistic ( $\rightarrow$  I/4.3.5) under the null hypothesis ( $\rightarrow$  I/4.3.2).

Then, we can obtain the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the p-value ( $\rightarrow$  I/4.3.10) as

$$\begin{aligned} F_P(p) &= \Pr(P < p) \\ &= \Pr(F_T(T) < p) \\ &= \Pr(T < F_T^{-1}(p)) \\ &= F_T(F_T^{-1}(p)) \\ &= p \end{aligned} \quad (3)$$

which is the cumulative distribution function of a continuous uniform distribution ( $\rightarrow$  II/3.1.4) over the interval  $[0, 1]$ :

$$F_X(x) = \int_{-\infty}^x \mathcal{U}(z; 0, 1) dz = x \quad \text{where } 0 \leq x \leq 1. \quad (4)$$

■

#### Sources:

- jll (2018): “Why are p-values uniformly distributed under the null hypothesis?”; in: *StackExchange Cross Validated*, retrieved on 2022-03-18; URL: <https://stats.stackexchange.com/a/345763/270304>.

#### 4.3.12 Minimum detectable effect

**Definition:** Consider a hypothesis test ( $\rightarrow$  I/4.3.1) concerning parameter  $\theta$ . We have  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$ , where  $\{\Theta_0, \Theta_1\}$  is a partition of  $\Theta$ . Let the probability of rejecting the null hypothesis for the test with  $n$  observations be  $\kappa_n(\theta)$ . The number of samples  $n$  is fixed.

The effect size for a parameter  $\theta \in \Theta_1$  is the distance of  $\theta$  from the null parameter space  $\Theta_0$ ,  $d(\theta, \Theta_0)$ . It can be defined as

$$d(\theta, \Theta_0) = \inf_{\theta_0 \in \Theta_0} \|\theta - \theta_0\|. \quad (1)$$

Let the test have a given significance level ( $\rightarrow$  I/4.3.8)  $\alpha$  and desired power ( $\rightarrow$  I/4.3.7)  $1 - \beta$ . The minimum detectable effect,  $\delta$ , is the smallest  $\delta$  so that the hypothesis test simultaneously has size ( $\rightarrow$  I/4.3.6) less than or equal to  $\alpha$  and power at least  $1 - \beta$ :

$$\left( \sup_{\theta \in \Theta_0} \kappa_n(\theta) \leq \alpha \right) \wedge \left( \inf_{\theta: d(\theta, \Theta_0) \leq \delta} \kappa_n(\theta) \geq 1 - \beta \right). \quad (2)$$

#### Sources:

- Sylvain Chabé-Ferret (2025): “Power Analysis”; in: *Statistical Tools for Causal Inference*, retrieved on 2025-07-05; URL: <https://chabefer.github.io/STCI/Power.html>.

### 4.3.13 Minimum required sample size

**Definition:** Consider a hypothesis test ( $\rightarrow$  I/4.3.1) concerning parameter  $\theta$ . We have  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$ , where  $\{\Theta_0, \Theta_1\}$  is a partition of  $\Theta$ . Let the probability of rejecting the null hypothesis for the test with  $n$  observations be  $\kappa_n(\theta)$ .

Given a pre-specified significance level ( $\rightarrow$  I/4.3.8)  $\alpha \in [0, 1]$  and a desired power ( $\rightarrow$  I/4.3.7)  $1 - \beta \in [0, 1]$ , the test must satisfy two conditions:

1. significance constraint: The probability of a type I error must not exceed  $\alpha$ :

$$\sup_{\theta \in \Theta_0} \kappa_n(\theta) \leq \alpha . \quad (1)$$

2. power constraint: The power of the test must be at least  $1 - \beta$  for all parameters corresponding to a minimum detectable effect ( $\rightarrow$  I/4.3.12),  $\delta$ . This is represented by a specified subset of the alternative hypothesis space,  $\Theta_{1,\delta} \subset \Theta_1$ :

$$\inf_{\theta \in \Theta_{1,\delta}} \kappa_n(\theta) \geq 1 - \beta . \quad (2)$$

The minimum required sample size,  $n_m$ , is the smallest integer  $n$  so that the hypothesis test simultaneously satisfies both the significance and power constraint.

**Sources:**

- Sylvain Chabé-Ferret (2025): “Power Analysis”; in: *Statistical Tools for Causal Inference*, retrieved on 2025-07-05; URL: <https://chabefer.github.io/STCI/Power.html>.

## 5 Bayesian statistics

### 5.1 Probabilistic modeling

#### 5.1.1 Generative model

**Definition:** Consider measured data ( $\rightarrow$  I/1.1.5)  $y$  and some unknown latent parameters ( $\rightarrow$  I/1.5.7)  $\theta$ . A statement about the distribution ( $\rightarrow$  I/1.5.1) of  $y$  given  $\theta$  is called a generative model  $m$

$$m : y \sim \mathcal{D}(\theta) , \quad (1)$$

where  $\mathcal{D}$  denotes an arbitrary probability distribution ( $\rightarrow$  I/1.5.1) and  $\theta$  are the parameters of this distribution.

**Sources:**

- Friston et al. (2008): “Bayesian decoding of brain images”; in: *NeuroImage*, vol. 39, pp. 181-205;  
URL: <https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203>; DOI: 10.1016/j.neuroim

#### 5.1.2 Likelihood function

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the probability density function ( $\rightarrow$  I/1.7.1) of the distribution of  $y$  given  $\theta$  is called the likelihood function of  $m$ :

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) . \quad (1)$$

#### 5.1.3 Prior distribution

**Definition:** Consider measured data  $y$  and some unknown latent parameters  $\theta$ . A distribution ( $\rightarrow$  I/1.5.1) of  $\theta$  unconditional on  $y$  is called a prior distribution:

$$\theta \sim \mathcal{D}(\lambda) . \quad (1)$$

The parameters  $\lambda$  of this distribution are called the prior hyperparameters and the probability density function ( $\rightarrow$  I/1.7.1) is called the prior density:

$$p(\theta|m) = \mathcal{D}(\theta; \lambda) . \quad (2)$$

#### 5.1.4 Prior predictive distribution

**Definition:** Consider a full probability model ( $\rightarrow$  I/5.1.6) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta)$ . Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of any data point ( $\rightarrow$  I/1.1.5)  $y_{\text{new}}$ , accounting for the prior distribution, is called the prior predictive distribution:

$$p(y_{\text{new}}) = \int p(y_{\text{new}}|\theta) p(\theta) d\theta . \quad (1)$$

### 5.1.5 Prior predictive distribution is marginal of joint likelihood

**Theorem:** The prior predictive distribution ( $\rightarrow$  I/5.1.4) is the marginal distribution ( $\rightarrow$  I/1.5.3) of the joint likelihood ( $\rightarrow$  I/5.1.15) of the new data ( $\rightarrow$  I/1.1.5)  $y_{\text{new}}$ , unconditional on the measured data  $y$ :

$$p(y_{\text{new}}) = \int p(y_{\text{new}}, \theta) d\theta \quad (1)$$

**Proof:** The prior predictive distribution ( $\rightarrow$  I/5.1.4) is defined as the marginal distribution ( $\rightarrow$  I/1.5.3) of new data  $y_{\text{new}}$ , predicted based on the prior distribution ( $\rightarrow$  I/5.1.3) before seeing the measured data  $y$ :

$$p(y_{\text{new}}) = \int p(y_{\text{new}}|\theta) p(\theta) d\theta . \quad (2)$$

By using the law of conditional probability ( $\rightarrow$  I/1.3.4), we can write the integrand as

$$p(y_{\text{new}}|\theta) p(\theta) = p(y_{\text{new}}, \theta) \quad (3)$$

which is the joint likelihood ( $\rightarrow$  I/5.1.7). Thus, expression (2) can be written as:

$$p(y_{\text{new}}) = \int p(y_{\text{new}}, \theta) d\theta . \quad (4)$$

■

### 5.1.6 Full probability model

**Definition:** Consider measured data  $y$  and some unknown latent parameters  $\theta$ . The combination of a generative model ( $\rightarrow$  I/5.1.1) for  $y$  in terms of the parameters  $\theta$  and a prior distribution ( $\rightarrow$  I/5.1.3) on  $\theta$  in terms of hyperparameters  $\lambda$  is called a full probability model  $m$ :

$$m : y \sim \mathcal{D}_1(\theta), \theta \sim \mathcal{D}_2(\lambda) . \quad (1)$$

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Probability and inference”; in: *Bayesian Data Analysis*, ch. 1, p. 3; URL: <http://www.stat.columbia.edu/~gelman/book/>.

### 5.1.7 Joint likelihood

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$  and a prior distribution ( $\rightarrow$  I/5.1.3) on  $\theta$ . Then, the joint probability ( $\rightarrow$  I/1.3.2) distribution ( $\rightarrow$  I/1.5.1) of  $y$  and  $\theta$  is called the joint likelihood:

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (1)$$



### 5.1.8 Joint likelihood is product of likelihood and prior

**Theorem:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$  and a prior distribution ( $\rightarrow$  I/5.1.3) on  $\theta$ . Then, the joint likelihood ( $\rightarrow$  I/5.1.7) is equal to the product of likelihood function ( $\rightarrow$  I/5.1.2) and prior density ( $\rightarrow$  I/5.1.3):

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (1)$$

**Proof:** The joint likelihood ( $\rightarrow$  I/5.1.7) is defined as the joint probability ( $\rightarrow$  I/1.3.2) distribution ( $\rightarrow$  I/1.5.1) of data  $y$  and parameters  $\theta$ :

$$p(y, \theta|m) . \quad (2)$$

Applying the law of conditional probability ( $\rightarrow$  I/1.3.4), we have:

$$\begin{aligned} p(y|\theta, m) &= \frac{p(y, \theta|m)}{p(\theta|m)} \\ &\Leftrightarrow \\ p(y, \theta|m) &= p(y|\theta, m) p(\theta|m) . \end{aligned} \quad (3)$$

■

### 5.1.9 Posterior distribution

**Definition:** Consider measured data  $y$  and some unknown latent parameters  $\theta$ . The distribution ( $\rightarrow$  I/1.5.1) of  $\theta$  conditional ( $\rightarrow$  I/1.5.4) on  $y$  is called the posterior distribution:

$$\theta|y \sim \mathcal{D}(\phi) . \quad (1)$$

The parameters  $\phi$  of this distribution are called the posterior hyperparameters and the probability density function ( $\rightarrow$  I/1.7.1) is called the posterior density:

$$p(\theta|y, m) = \mathcal{D}(\theta; \phi) . \quad (2)$$

### 5.1.10 Posterior density is proportional to joint likelihood

**Theorem:** In a full probability model ( $\rightarrow$  I/5.1.6)  $m$  describing measured data  $y$  using model parameters  $\theta$ , the posterior density ( $\rightarrow$  I/5.1.9) over the model parameters is proportional to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(\theta|y, m) \propto p(y, \theta|m) . \quad (1)$$

**Proof:** In a full probability model ( $\rightarrow$  I/5.1.6), the posterior distribution ( $\rightarrow$  I/5.1.9) can be expressed using Bayes' theorem ( $\rightarrow$  I/5.3.2):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (2)$$

Applying the law of conditional probability ( $\rightarrow$  I/1.3.4) to the numerator, we have:

$$p(\theta|y, m) = \frac{p(y, \theta|m)}{p(y|m)} . \quad (3)$$

Because the denominator does not depend on  $\theta$ , it is constant in  $\theta$  and thus acts a proportionality factor between the posterior distribution and the joint likelihood:

$$p(\theta|y, m) \propto p(y, \theta|m) . \quad (4)$$

■

### 5.1.11 Combined posterior distribution from independent data

**Theorem:** Let  $p(\theta|y_1)$  and  $p(\theta|y_2)$  be posterior distributions ( $\rightarrow$  I/5.1.9), obtained using the same prior distribution ( $\rightarrow$  I/5.1.3) from conditionally independent ( $\rightarrow$  I/1.3.7) data sets  $y_1$  and  $y_2$ :

$$p(y_1, y_2|\theta) = p(y_1|\theta) \cdot p(y_2|\theta) . \quad (1)$$

Then, the combined posterior distribution ( $\rightarrow$  I/1.5.1) is proportional to the product of the individual posterior densities ( $\rightarrow$  I/1.7.1), divided by the prior density:

$$p(\theta|y_1, y_2) \propto \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} . \quad (2)$$

**Proof:** Since  $p(\theta|y_1)$  and  $p(\theta|y_2)$  are posterior distributions ( $\rightarrow$  I/5.1.9), Bayes' theorem ( $\rightarrow$  I/5.3.2) holds for them:

$$\begin{aligned} p(\theta|y_1) &= \frac{p(y_1|\theta) \cdot p(\theta)}{p(y_1)} \\ p(\theta|y_2) &= \frac{p(y_2|\theta) \cdot p(\theta)}{p(y_2)} . \end{aligned} \quad (3)$$

Moreover, Bayes' theorem must also hold for the combined posterior distribution ( $\rightarrow$  I/5.1.10):

$$p(\theta|y_1, y_2) = \frac{p(y_1, y_2|\theta) \cdot p(\theta)}{p(y_1, y_2)} . \quad (4)$$

With that, we can express the combined posterior distribution as follows:

$$\begin{aligned} p(\theta|y_1, y_2) &\stackrel{(4)}{=} \frac{p(y_1, y_2|\theta) \cdot p(\theta)}{p(y_1, y_2)} \\ &\stackrel{(1)}{=} p(y_1|\theta) \cdot p(y_2|\theta) \cdot \frac{p(\theta)}{p(y_1, y_2)} \\ &\stackrel{(3)}{=} \frac{p(\theta|y_1) \cdot p(y_1)}{p(\theta)} \cdot \frac{p(\theta|y_2) \cdot p(y_2)}{p(\theta)} \cdot \frac{p(\theta)}{p(y_1, y_2)} \\ &= \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} \cdot \frac{p(y_1) \cdot p(y_2)}{p(y_1, y_2)} . \end{aligned} \quad (5)$$

Note that the second fraction does not depend on  $\theta$  and thus, the posterior distribution over  $\theta$  is proportional to the first fraction:

$$p(\theta|y_1, y_2) \propto \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} . \quad (6)$$

■

### 5.1.12 Posterior predictive distribution

**Definition:** Consider a full probability model ( $\rightarrow$  I/5.1.6) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta)$  and posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\theta|y)$  based on measured data ( $\rightarrow$  I/1.1.5)  $y$ . Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of new data  $y_{\text{new}}$ , predicted based on the posterior distribution, is called the posterior predictive distribution:

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}|\theta) p(\theta|y) d\theta . \quad (1)$$

### 5.1.13 Posterior predictive distribution is marginal of joint likelihood

**Theorem:** The posterior predictive distribution ( $\rightarrow$  I/5.1.12) is the marginal distribution ( $\rightarrow$  I/1.5.3) of the joint likelihood ( $\rightarrow$  I/5.1.15) of the new data ( $\rightarrow$  I/1.1.5)  $y_{\text{new}}$ , conditional on the measured data  $y$ :

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}, \theta|y) d\theta \quad (1)$$

**Proof:** The posterior predictive distribution ( $\rightarrow$  I/5.1.12) is defined as the marginal distribution ( $\rightarrow$  I/1.5.3) of new data  $y_{\text{new}}$ , predicted based on the posterior distribution ( $\rightarrow$  I/5.1.9) obtained from the measured data  $y$ :

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}|\theta) p(\theta|y) d\theta . \quad (2)$$

We notice that  $y_{\text{new}}$  is independent ( $\rightarrow$  I/1.3.9) of  $y$ , so we can write:

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}|\theta, y) p(\theta|y) d\theta . \quad (3)$$

By using the law of conditional probability ( $\rightarrow$  I/1.3.4), we can write the integrand as:

$$p(y_{\text{new}}|\theta, y) p(\theta|y) = p(y_{\text{new}}, \theta|y) . \quad (4)$$

This is the posterior ( $\rightarrow$  I/5.1.9) joint likelihood ( $\rightarrow$  I/5.1.7). Thus, expression (2) can be written as:

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}, \theta|y) d\theta . \quad (5)$$

■

### 5.1.14 Maximum-a-posteriori estimation

**Definition:** Consider a posterior distribution ( $\rightarrow$  I/5.1.9) of an unknown parameter  $\theta$ , given measured data  $y$ , parametrized by posterior hyperparameters  $\phi$ :

$$\theta|y \sim \mathcal{D}(\phi) . \quad (1)$$

Then, the value of  $\theta$  at which the posterior density ( $\rightarrow$  I/5.1.9) attains its maximum is called the “maximum-a-posteriori estimate”, “MAP estimate” or “posterior mode” of  $\theta$ :

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \mathcal{D}(\theta; \phi) . \quad (2)$$

**Sources:**

- Wikipedia (2023): “Maximum a posteriori estimation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-01; URL: [https://en.wikipedia.org/wiki/Maximum\\_a\\_posteriori\\_estimation#Description](https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation#Description).

### 5.1.15 Marginal likelihood

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$  and a prior distribution ( $\rightarrow$  I/5.1.3) on  $\theta$ . Then, the marginal probability ( $\rightarrow$  I/1.3.3) distribution ( $\rightarrow$  I/1.5.1) of  $y$  across the parameter space  $\Theta$  is called the marginal likelihood:

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta . \quad (1)$$

### 5.1.16 Marginal likelihood is integral of joint likelihood

**Theorem:** In a full probability model ( $\rightarrow$  I/5.1.6)  $m$  describing measured data  $y$  using model parameters  $\theta$ , the marginal likelihood ( $\rightarrow$  I/5.1.15) is the integral of the joint likelihood ( $\rightarrow$  I/5.1.7) across the parameter space  $\Theta$

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta \quad (1)$$

and related to likelihood function ( $\rightarrow$  I/5.1.2) and prior distribution ( $\rightarrow$  I/5.1.3) as follows:

$$p(y|m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (2)$$

**Proof:** In a full probability model ( $\rightarrow$  I/5.1.6), the marginal likelihood ( $\rightarrow$  I/5.1.15) is defined as the marginal probability of the data  $y$ , given only the model  $m$ :

$$p(y|m) . \quad (3)$$

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), this can be obtained by integrating the joint likelihood ( $\rightarrow$  I/5.1.7) function over the entire parameter space:

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta . \quad (4)$$

Applying the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand can also be written as the product of likelihood function ( $\rightarrow$  I/5.1.2) and prior density ( $\rightarrow$  I/5.1.3):

$$p(y|m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (5)$$

■

## 5.2 Prior distributions

### 5.2.1 Flat vs. hard vs. soft

**Definition:** Let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) for the parameter  $\theta$  of a generative model ( $\rightarrow$  I/5.1.1)  $m$ . Then,

- the distribution is called a “flat prior”, if its precision ( $\rightarrow$  I/1.13.13) is zero or variance ( $\rightarrow$  I/1.13.1) is infinite;
- the distribution is called a “hard prior”, if its precision ( $\rightarrow$  I/1.13.13) is infinite or variance ( $\rightarrow$  I/1.13.1) is zero;
- the distribution is called a “soft prior”, if its precision ( $\rightarrow$  I/1.13.13) and variance ( $\rightarrow$  I/1.13.1) are non-zero and finite.

**Sources:**

- Friston et al. (2002): “Classical and Bayesian Inference in Neuroimaging: Theory”; in: *NeuroImage*, vol. 16, iss. 2, pp. 465-483, fn. 1; URL: <https://www.sciencedirect.com/science/article/pii/S1053811902910906>; DOI: 10.1006/nimg.2002.1090.
- Friston et al. (2002): “Classical and Bayesian Inference in Neuroimaging: Applications”; in: *NeuroImage*, vol. 16, iss. 2, pp. 484-512, fn. 10; URL: <https://www.sciencedirect.com/science/article/pii/S1053811902910918>; DOI: 10.1006/nimg.2002.1091.

### 5.2.2 Uniform vs. non-uniform

**Definition:** Let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) for the parameter  $\theta$  of a generative model ( $\rightarrow$  I/5.1.1)  $m$  where  $\theta$  belongs to the parameter space  $\Theta$ . Then,

- the distribution is called a “uniform prior”, if its density ( $\rightarrow$  I/1.7.1) or mass ( $\rightarrow$  I/1.6.1) is constant over  $\Theta$ ;
- the distribution is called a “non-uniform prior”, if its density ( $\rightarrow$  I/1.7.1) or mass ( $\rightarrow$  I/1.6.1) is not constant over  $\Theta$ .

**Sources:**

- Wikipedia (2020): “Lindley’s paradox”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Lindley%27s\\_paradox#Bayesian\\_approach](https://en.wikipedia.org/wiki/Lindley%27s_paradox#Bayesian_approach).

### 5.2.3 Informative vs. non-informative

**Definition:** Let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) for the parameter  $\theta$  of a generative model ( $\rightarrow$  I/5.1.1)  $m$ . Then,

- the distribution is called an “informative prior”, if it biases the parameter towards particular values;

- the distribution is called a “weakly informative prior”, if it mildly influences the posterior distribution ( $\rightarrow$  I/5.1.10);
- the distribution is called a “non-informative prior”, if it does not influence ( $\rightarrow$  I/5.1.10) the posterior hyperparameters ( $\rightarrow$  I/5.1.9).

**Sources:**

- Soch J, Allefeld C, Haynes JD (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469-489, eq. 15, p. 473; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage.2016.07.047.

### 5.2.4 Empirical vs. non-empirical

**Definition:** Let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) for the parameter  $\theta$  of a generative model ( $\rightarrow$  I/5.1.1)  $m$ . Then,

- the distribution is called an “empirical prior”, if it has been derived from empirical data ( $\rightarrow$  I/5.1.10);
- the distribution is called a “theoretical prior”, if it was specified without regard to empirical data.

**Sources:**

- Soch J, Allefeld C, Haynes JD (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469-489, eq. 13, p. 473; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage.2016.07.047.

### 5.2.5 Conjugate vs. non-conjugate

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Then,

- the prior distribution ( $\rightarrow$  I/5.1.3) is called “conjugate”, if it, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1);
- the prior distribution is called “non-conjugate”, if this is not the case.

**Sources:**

- Wikipedia (2020): “Conjugate prior”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: [https://en.wikipedia.org/wiki/Conjugate\\_prior](https://en.wikipedia.org/wiki/Conjugate_prior).

### 5.2.6 Maximum entropy priors

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|\lambda, m)$  using prior hyperparameters ( $\rightarrow$  I/5.1.3)  $\lambda$ . Then, the prior distribution is called a “maximum entropy prior”, if

1) when  $\theta$  is a discrete random variable ( $\rightarrow$  I/1.2.6), it maximizes the entropy ( $\rightarrow$  I/2.1.1) of the prior probability mass function ( $\rightarrow$  I/1.6.1):

$$\lambda_{\text{maxent}} = \arg \max_{\lambda} H[p(\theta|\lambda, m)] ; \quad (1)$$

2) when  $\theta$  is a continuous random variable ( $\rightarrow$  I/1.2.6), it maximizes the differential entropy ( $\rightarrow$  I/2.2.1) of the prior probability density function ( $\rightarrow$  I/1.7.1):

$$\lambda_{\text{maxent}} = \arg \max_{\lambda} h[p(\theta|\lambda, m)] . \quad (2)$$

**Sources:**

- Wikipedia (2020): “Prior probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: [https://en.wikipedia.org/wiki/Prior\\_probability#Uninformative\\_priors](https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors).

### 5.2.7 Empirical Bayes priors

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|\lambda, m)$  using prior hyperparameters ( $\rightarrow$  I/5.1.3)  $\lambda$ . Let  $p(y|\lambda, m)$  be the marginal likelihood ( $\rightarrow$  I/5.1.15) when integrating the parameters out of the joint likelihood ( $\rightarrow$  I/5.1.16). Then, the prior distribution is called an “Empirical Bayes ( $\rightarrow$  I/5.3.4) prior”, if it maximizes the logarithmized marginal likelihood:

$$\lambda_{\text{EB}} = \arg \max_{\lambda} \log p(y|\lambda, m) . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Empirical Bayes method”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: [https://en.wikipedia.org/wiki/Empirical\\_Bayes\\_method#Introduction](https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction).

### 5.2.8 Reference priors

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|\lambda, m)$  using prior hyperparameters ( $\rightarrow$  I/5.1.3)  $\lambda$ . Let  $p(\theta|y, \lambda, m)$  be the posterior distribution ( $\rightarrow$  I/5.1.9) that is proportional to the the joint likelihood ( $\rightarrow$  I/5.1.10). Then, the prior distribution is called a “reference prior”, if it maximizes the expected ( $\rightarrow$  I/1.12.1) Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of the posterior distribution relative to the prior distribution:

$$\lambda_{\text{ref}} = \arg \max_{\lambda} \langle \text{KL} [p(\theta|y, \lambda, m) || p(\theta|\lambda, m)] \rangle . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Prior probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: [https://en.wikipedia.org/wiki/Prior\\_probability#Uninformative\\_priors](https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors).

## 5.3 Bayesian inference

### 5.3.1 Odds ratios

**Definition:** Let  $A_1$ ,  $A_2$  and  $B$  be arbitrary statements about random variables ( $\rightarrow$  I/1.2.2) where  $A_1$  and  $A_2$  are mutually exclusive ( $\rightarrow$  I/1.3.10), but not necessarily collectively exhaustive ( $\rightarrow$  I/1.4.12). Then, the “odds” or “prior ( $\rightarrow$  I/5.1.3) odds ratio” in favor of  $A_1$  against  $A_2$  is defined as the ratio of the probabilities of  $A_1$  and  $A_2$ , unconditional on  $B$

$$\frac{p(A_1)}{p(A_2)} \quad (1)$$

and the “posterior ( $\rightarrow$  I/5.1.9) odds ratio” in favor of  $A_1$  against  $A_2$  is defined as the ratio of the probabilities of  $A_1$  and  $A_2$ , conditional on  $B$

$$\frac{p(A_1|B)}{p(A_2|B)} \quad (2)$$

where  $p(\cdot)$  is a probability mass function ( $\rightarrow$  I/1.6.1) or probability density function ( $\rightarrow$  I/1.7.1). Prior and posterior odds ratio are related to each other via the Bayes factor ( $\rightarrow$  IV/3.3.1) through Bayes’ rule ( $\rightarrow$  I/5.3.3).

#### Sources:

- Wikipedia (2025): “Odds ratio”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-02-06; URL: [https://en.wikipedia.org/wiki/Odds\\_ratio](https://en.wikipedia.org/wiki/Odds_ratio).

### 5.3.2 Bayes’ theorem

**Theorem:** Let  $A$  and  $B$  be two arbitrary statements about random variables ( $\rightarrow$  I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, the conditional probability that  $A$  is true, given that  $B$  is true, is equal to

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} . \quad (1)$$

**Proof:** The conditional probability ( $\rightarrow$  I/1.3.4) is defined as the ratio of joint probability ( $\rightarrow$  I/1.3.2), i.e. the probability of both statements being true, and marginal probability ( $\rightarrow$  I/1.3.3), i.e. the probability of only the second one being true:

$$p(A|B) = \frac{p(A, B)}{p(B)} . \quad (2)$$

It can also be written down for the reverse situation, i.e. to calculate the probability that  $B$  is true, given that  $A$  is true:

$$p(B|A) = \frac{p(A, B)}{p(A)} . \quad (3)$$

Both equations can be rearranged for the joint probability

$$p(A|B)p(B) \stackrel{(2)}{=} p(A, B) \stackrel{(3)}{=} p(B|A)p(A) \quad (4)$$

from which Bayes’ theorem can be directly derived:

$$p(A|B) \stackrel{(4)}{=} \frac{p(B|A)p(A)}{p(B)} . \quad (5)$$

■

#### Sources:



- Koch, Karl-Rudolf (2007): “Rules of Probability”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, pp. 6/13, eqs. 2.12/2.38; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

### 5.3.3 Bayes’ rule

**Theorem:** Let  $A_1$ ,  $A_2$  and  $B$  be arbitrary statements about random variables ( $\rightarrow$  I/1.2.2) where  $A_1$  and  $A_2$  are mutually exclusive. Then, Bayes’ rule states that the posterior odds ( $\rightarrow$  I/5.3.1) are equal to the Bayes factor ( $\rightarrow$  IV/3.3.1) times the prior odds ( $\rightarrow$  I/5.3.1), i.e.

$$\frac{p(A_1|B)}{p(A_2|B)} = \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} . \quad (1)$$

**Proof:** Using Bayes’ theorem ( $\rightarrow$  I/5.3.2), the conditional probabilities ( $\rightarrow$  I/1.3.4) on the left are given by

$$p(A_1|B) = \frac{p(B|A_1) \cdot p(A_1)}{p(B)} \quad (2)$$

$$p(A_2|B) = \frac{p(B|A_2) \cdot p(A_2)}{p(B)} . \quad (3)$$

Dividing the two conditional probabilities by each other

$$\begin{aligned} \frac{p(A_1|B)}{p(A_2|B)} &= \frac{p(B|A_1) \cdot p(A_1)/p(B)}{p(B|A_2) \cdot p(A_2)/p(B)} \\ &= \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} , \end{aligned} \quad (4)$$

one obtains the posterior odds ratio as given by the theorem. ■

#### Sources:

- Wikipedia (2019): “Bayes’ theorem”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-06; URL: [https://en.wikipedia.org/wiki/Bayes%27\\_theorem#Bayes%E2%80%99\\_rule](https://en.wikipedia.org/wiki/Bayes%27_theorem#Bayes%E2%80%99_rule).

### 5.3.4 Empirical Bayes

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with model parameters  $\theta$  and hyper-parameters  $\lambda$  implying the likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, \lambda, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|\lambda, m)$ . Then, an Empirical Bayes treatment of  $m$ , also referred to as “type II maximum likelihood ( $\rightarrow$  I/4.1.3)” or “evidence ( $\rightarrow$  IV/3.1.3) approximation”, consists in

1) evaluating the marginal likelihood ( $\rightarrow$  I/5.1.15) of the model  $m$

$$p(y|\lambda, m) = \int p(y|\theta, \lambda, m) (\theta|\lambda, m) d\theta , \quad (1)$$

2) maximizing the log model evidence ( $\rightarrow$  IV/3.1.3) with respect to  $\lambda$

$$\hat{\lambda} = \arg \max_{\lambda} \log p(y|\lambda, m) \quad (2)$$

3) and using the prior distribution ( $\rightarrow$  I/5.1.3) at this maximum

$$p(\theta|m) = p(\theta|\hat{\lambda}, m) \quad (3)$$

for Bayesian inference ( $\rightarrow$  I/5.3.2), i.e. obtaining the posterior distribution ( $\rightarrow$  I/5.1.10) and computing the marginal likelihood ( $\rightarrow$  I/5.1.16).

#### Sources:

- Wikipedia (2021): “Empirical Bayes method”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: [https://en.wikipedia.org/wiki/Empirical\\_Bayes\\_method#Introduction](https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction).
- Bishop CM (2006): “The Evidence Approximation”; in: *Pattern Recognition for Machine Learning*, ch. 3.5, pp. 165-172; URL: <https://www.springer.com/gp/book/9780387310732>.

### 5.3.5 Variational Bayes

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with model parameters ( $\rightarrow$  I/1.5.7)  $\theta \in \Theta$  implying the likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Then, a Variational Bayes treatment of  $m$ , also referred to as “approximate inference” or “variational inference”, consists in

1) constructing an approximate posterior distribution ( $\rightarrow$  I/5.1.9)

$$q(\theta) \approx p(\theta|y, m), \quad (1)$$

2) evaluating the variational free energy ( $\rightarrow$  IV/3.1.11)

$$F_m[q(\theta)] = \int_{\Theta} q(\theta) \log \frac{p(\theta|y, m)}{q(\theta)} d\theta \quad (2)$$

3) and maximizing this function with respect to  $q(\theta)$

$$\hat{q}(\theta) = \arg \max_q F_m[q(\theta)] \quad (3)$$

for Bayesian inference ( $\rightarrow$  I/5.3.2), i.e. obtaining the posterior distribution ( $\rightarrow$  I/5.1.9) (from eq. (3)) and approximating the marginal likelihood ( $\rightarrow$  I/5.1.15) (by plugging eq. (3) into eq. (2)).

#### Sources:

- Wikipedia (2021): “Variational Bayesian methods”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: [https://en.wikipedia.org/wiki/Variational\\_Bayesian\\_methods#Evidence\\_lower\\_bound](https://en.wikipedia.org/wiki/Variational_Bayesian_methods#Evidence_lower_bound).
- Bishop CM (2006): “Variational Inference”; in: *Pattern Recognition for Machine Learning*, ch. 10.1, pp. 462-474; URL: <https://www.springer.com/gp/book/9780387310732>.
- Penny W, Flandin G, Trujillo-Barreto N (2007): “Bayesian Comparison of Spatially Regularised General Linear Models”; in: *Human Brain Mapping*, vol. 28, pp. 275–293, eqs. 2-9; URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327>; DOI: 10.1002/hbm.20327.

- Ostwald D, Kirilina E, Starke L, Blankenburg F (2014): “A tutorial on variational Bayes for latent linear stochastic time-series models”; in: *Journal of Mathematical Psychology*, vol. 60, pp. 1-14, eqs. 10-17; URL: <https://www.sciencedirect.com/science/article/abs/pii/S0022249614000352>; DOI: 10.1016/j.jmp.2014.04.003.

### 5.3.6 Decomposition of the free energy

**Theorem:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m) = p(y|\theta)$ , prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m) = p(\theta)$ , joint likelihood ( $\rightarrow$  I/5.1.7)  $p(y, \theta|m) = p(y, \theta)$ , true posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\theta|y, m) = p(\theta|y)$  and approximate posterior ( $\rightarrow$  I/5.3.5) distribution  $q(\theta|m) = q(\theta) \approx p(\theta|y)$ . Then, the variational free energy ( $\rightarrow$  IV/3.1.11) can be decomposed as

1) the difference between log model evidence and KL divergence of approximate from true posterior

$$F[q(\theta)] = \log p(y) - \text{KL}[q(\theta)||p(\theta|y)] , \quad (1)$$

2) the difference of expected log-likelihood and KL divergence of approximate posterior from prior

$$F[q(\theta)] = \langle \log p(y|\theta) \rangle_{q(\theta)} - \text{KL}[q(\theta)||p(\theta)] \quad (2)$$

3) the sum of expected joint log-likelihood and approximate posterior differential entropy

$$F[q(\theta)] = \langle \log p(y, \theta) \rangle_{q(\theta)} - h[q(\theta)] \quad (3)$$

where  $p(y|m) = p(y)$  is the marginal likelihood ( $\rightarrow$  I/5.1.15),  $\langle \cdot \rangle_{p(x)}$  denotes an expectation ( $\rightarrow$  I/1.12.1) with respect to the density ( $\rightarrow$  I/1.7.1)  $p(x)$ ,  $\text{KL}[\cdot||\cdot]$  denotes the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) and  $h[\cdot]$  denotes the differential entropy ( $\rightarrow$  I/2.2.1).

**Proof:** The log model evidence ( $\rightarrow$  IV/3.1.3) is defined as

$$\log p(y) = \log \int_{\Theta} p(y, \theta) d\theta \quad (4)$$

and using the approximate posterior density  $q(\theta)$ , this can be rewritten as

$$\begin{aligned} \log p(y) &= \log \int_{\Theta} \frac{q(\theta)}{q(\theta)} p(y, \theta) d\theta \\ &= \log \left\langle \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} . \end{aligned} \quad (5)$$

By Jensen's inequality ( $\rightarrow$  I/1.12.11) and because the logarithm is a concave function ( $\rightarrow$  I/2.5.4), we have:

$$\log \left\langle \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} \geq \left\langle \log \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} . \quad (6)$$

The right-hand side of this equation is referred to as the free energy ( $\rightarrow$  IV/3.1.11):

$$F[q(\theta)] = \left\langle \log \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} . \quad (7)$$

With that, we can show the above identities:

1) the difference between log model evidence and KL divergence of approximate from true posterior

$$\begin{aligned}
 F[q(\theta)] &= \left\langle \log \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} \\
 &= \left\langle \log \frac{p(\theta|y)p(y)}{q(\theta)} \right\rangle_{q(\theta)} \\
 &= \left\langle \log p(y) - \log \frac{q(\theta)}{p(\theta|y)} \right\rangle_{q(\theta)} \\
 &= \langle \log p(y) \rangle_{q(\theta)} - \int_{\Theta} q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \\
 &= \log p(y) - \text{KL}[q(\theta)||p(\theta|y)]
 \end{aligned} \tag{8}$$

where the first term is the log model evidence ( $\rightarrow$  IV/3.1.3) (= log marginal likelihood) and the second term can be seen as an approximation error (= divergence of approximate from true posterior distribution);

2) the difference of expected log-likelihood and KL divergence of approximate posterior from prior

$$\begin{aligned}
 F[q(\theta)] &= \left\langle \log \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} \\
 &= \left\langle \log \frac{p(y|\theta)p(\theta)}{q(\theta)} \right\rangle_{q(\theta)} \\
 &= \left\langle \log p(y|\theta) - \log \frac{q(\theta)}{p(\theta)} \right\rangle_{q(\theta)} \\
 &= \int_{\Theta} q(\theta) \log p(y|\theta) d\theta - \int_{\Theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \\
 &= \langle \log p(y|\theta) \rangle_{q(\theta)} - \text{KL}[q(\theta)||p(\theta)]
 \end{aligned} \tag{9}$$

where the first term can be seen as an accuracy term ( $\rightarrow$  IV/3.1.6) (= posterior expected log-likelihood) and the second term can be seen as a complexity penalty ( $\rightarrow$  IV/3.1.6) (= divergence of posterior from prior distribution);

3) the sum of expected joint log-likelihood and approximate posterior entropy

$$\begin{aligned}
 F[q(\theta)] &= \left\langle \log \frac{p(y, \theta)}{q(\theta)} \right\rangle_{q(\theta)} \\
 &= \langle \log p(y, \theta) - \log q(\theta) \rangle_{q(\theta)} \\
 &= \int_{\Theta} q(\theta) \log p(y, \theta) d\theta - \int_{\Theta} q(\theta) \log q(\theta) d\theta \\
 &= \langle \log p(y, \theta) \rangle_{q(\theta)} + h[q(\theta)]
 \end{aligned} \tag{10}$$

where the first term represents the negative energy (= posterior expected joint log-likelihood) and the second term presents the posterior entropy ( $\rightarrow$  I/2.2.1) (= differential entropy of the approximate posterior distribution).

**Sources:**

- Zeidman et al. (2023): “A primer on Variational Laplace (VL)” ; in: *NeuroImage*, vol. 279, art. 120310, p. 5, eqs. 12-14; URL: <https://www.sciencedirect.com/science/article/pii/S1053811923004615>; DOI: 10.1016/j.neuroimage.2023.120310.

**5.3.7 Free energy is lower bound on log model evidence**

**Theorem:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m) = p(y|\theta)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m) = p(\theta)$ . Then, under a variational Bayesian ( $\rightarrow$  I/5.3.5) treatment using the approximate posterior ( $\rightarrow$  I/5.3.5) distribution  $q(\theta|m) = q(\theta) \approx p(\theta|y)$ , the variational free energy ( $\rightarrow$  IV/3.1.11) is a lower bound on the log model evidence ( $\rightarrow$  IV/3.1.3):

$$F[q(\theta)] \leq \log p(y) = \log \int_{\Theta} p(y, \theta|m) d\theta . \quad (1)$$

**Proof:** Using a decomposition of the variational free energy ( $\rightarrow$  I/5.3.6), it can be shown that the free energy is equal to the difference between the log model evidence ( $\rightarrow$  IV/3.1.3) and the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of approximate from true posterior distribution:

$$F[q(\theta)] = \log p(y) - \text{KL}[q(\theta)||p(\theta|y)] \quad (2)$$

Since the KL divergence is zero or positive for any two distributions ( $\rightarrow$  I/2.5.4)

$$\text{KL}[P||Q] \geq 0 , \quad (3)$$

the free energy must be smaller than or equal to the log model evidence:

$$F[q(\theta)] \leq \log p(y) . \quad (4)$$

**Sources:**

- Zeidman et al. (2023): “A primer on Variational Laplace (VL)” ; in: *NeuroImage*, vol. 279, art. 120310, pp. 4-5, eqs. 9-11; URL: <https://www.sciencedirect.com/science/article/pii/S1053811923004615>; DOI: 10.1016/j.neuroimage.2023.120310.

## 6 Machine learning

### 6.1 Scoring rules

#### 6.1.1 Scoring rule

**Definition:** A scoring rule is any extended real-valued function  $\mathbf{S} : \mathcal{Q} \times \Omega \rightarrow \mathbb{R}$  where  $\mathcal{Q}$  is a family of probability distributions over the space  $\Omega$ , such that  $\mathbf{S}(Q, \cdot)$  is  $\mathcal{Q}$ -quasi-integrable for all  $Q \in \mathcal{Q}$ . Output of the function  $\mathbf{S}(Q, y)$  represents the loss or penalty when the forecast  $Q \in \mathcal{Q}$  is issued and the observation  $y \in \Omega$  is realized.

**Sources:**

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.
- Wikipedia (2024): “Scoring rule”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-02-28; URL: [https://en.wikipedia.org/wiki/Scoring\\_rule](https://en.wikipedia.org/wiki/Scoring_rule).

#### 6.1.2 Proper scoring rule

**Definition:** A scoring rule ( $\rightarrow$  I/6.1.1)  $\mathbf{S}$  is called a proper scoring rule, if and only if

$$\max_{Q \in \mathcal{Q}} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = \mathbb{E}_{Y \sim P}[\mathbf{S}(P, Y)] . \quad (1)$$

In other words, score function  $\mathbf{S}$  is a proper scoring rule, if it is maximized when the forecaster gives exactly the ground truth distribution  $P(Y)$  as its probabilistic forecast  $Q \in \mathcal{Q}$ .

**Sources:**

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.

#### 6.1.3 Strictly proper scoring rule

**Definition:** A scoring rule ( $\rightarrow$  I/6.1.1)  $\mathbf{S}$  is called a strictly proper scoring rule, if and only if

- $\mathbf{S}$  is a proper scoring rule ( $\rightarrow$  I/6.1.2), and
- $\arg \max_{Q \in \mathcal{Q}} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = P$  is the unique maximizer of  $\mathbf{S}$  in  $\mathcal{Q}$ .

In other words, a strictly proper scoring rule is maximized only when the the forecaster gives exactly the ground truth distribution  $P(Y)$  as its probabilistic forecast  $Q \in \mathcal{Q}$ .

**Sources:**

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.

### 6.1.4 Log probability scoring rule

**Definition:** A log (probability) scoring rule ( $\rightarrow$  I/6.1.1)  $S(q, y)$  is as a scoring rule that measures the quality of a probabilistic forecast in decision theory. Formally, it can be defined in discrete or continuous form as follows:

1) Log scoring rule for binary classification:

$$S(q, y) = \begin{cases} \log q, & \text{if } y = 1 \\ \log(1 - q), & \text{if } y = 0 \end{cases} \quad (1)$$

which can be expressed as

$$S(q, y) = y \log q + (1 - y) \log(1 - q) \quad (2)$$

Note that the expressions given above have slightly different domains. For the first equation, the domain is  $D_1 = ([0, 1) \times \{0\}) \cup ((0, 1] \times \{1\})$ , while for the second equation, the domain is  $D_2 = (0, 1) \times \{0, 1\}$ .

2) Log scoring rule for multiclass classification:

$$S(q, y) = \sum_k y_k \log q_k(x) = \log q_{y^*}(x) \quad (3)$$

where  $y^*$  is the true class and  $q$  is the predicted probability distribution over the classes. We have  $y_k = 1$ , if the true class is  $k$  and  $y_k = 0$  otherwise.

3) Log scoring rule for regression (continuous case):

$$S(q, y) = \log q(y) \quad (4)$$

where  $q$  is the predicted probability distribution over the continuous space and  $y$  is the true value.

#### Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.

### 6.1.5 Log probability is strictly proper scoring rule

**Theorem:** The log (probability) scoring rule ( $\rightarrow$  I/6.1.4) is a strictly proper scoring rule ( $\rightarrow$  I/6.1.3).

**Proof:** We will show that all versions of the log probability scoring rule (binary/multiclass/regression) are strictly proper scoring rules.

1) Binary log probability scoring rule:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = P(Y = 1) \log q + P(Y = 0) \log(1 - q) \quad (1)$$

Let  $p$  be the true probability of the event  $Y = 1$ . Then, the expected score is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = p \log q + (1 - p) \log(1 - q) \quad (2)$$

To find the maxima, take the derivative with respect to  $q$  and set it to zero:

$$\begin{aligned}
\frac{\partial}{\partial q} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= \frac{p}{q} - \frac{1-p}{1-q} \\
0 &= \frac{p - pq - q + pq}{q(1-q)} \\
0 &= \frac{p-q}{q(1-q)} \\
\Rightarrow p - q &= 0 \\
\Rightarrow p &= q
\end{aligned} \tag{3}$$

Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$\begin{aligned}
\frac{\partial^2}{\partial q^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= -\frac{p}{q^2} - \frac{1-p}{(1-q)^2} \\
&= -\left( \underbrace{\frac{p}{q^2}}_{>0} + \underbrace{\frac{1-p}{(1-q)^2}}_{>0} \right) < 0
\end{aligned} \tag{4}$$

Except for the cases  $q = 0$  and  $q = 1$ , the second derivative is always negative, which means that the function is concave and the maximum is unique. For  $q = 1$ , maximum is achieved only if  $p = 1$ , and similarly for  $q = 0$ , maximum is achieved only if  $p = 0$ . Therefore,  $p = q$  is the only maximizer and the log probability scoring rule for binary classification is strictly proper.

2a) Multiclass log probability scoring rule (Proof 1):

$$S(q, y) = \sum_k^K y_k \log q_k(x) \tag{5}$$

Let  $p_k$  be the true probability of the event  $Y = k$ . Since  $q_k$  is the predicted probability for class  $k$ , we know that  $\sum_i q_i = 1$ . Then, the expected score is:

$$\begin{aligned}
\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= \sum_k P(Y = k|x) \log(q_k(x)) \\
&= p_1 \log(q_1(x)) + p_2 \log(q_2(x)) + \dots + p_K \log(q_K(x)) \\
&= p_1 \log(q_1(x)) + p_2 \log(q_2(x)) + \dots + p_K \log\left(1 - \sum_{i \neq K} q_i(x)\right)
\end{aligned} \tag{6}$$

Taking the derivative with respect to  $q_j$  and setting it to zero:

$$\begin{aligned}
\frac{\partial}{\partial q_j} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= \frac{p_j}{q_j} - \frac{p_K}{1 - \sum_{i \neq K} q_i(x)} \\
0 &= \frac{p_j}{q_j} - \frac{p_K}{q_K} \\
\Rightarrow \frac{p_j}{q_j} &= \frac{p_K}{q_K}
\end{aligned} \tag{7}$$



This equality holds for any  $j$ :

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} = \dots = \frac{p_K}{q_K} = \lambda \quad (8)$$

Each  $q_i$  can be represented as a constant multiple of  $p_i$  as follows:  $q_i = \lambda p_i$

$$\begin{aligned} \sum_i q_i &= 1 \\ \sum_i \lambda p_i &= 1 \\ \lambda \sum_i p_i &= 1 \\ \lambda &= 1 \end{aligned} \quad (9)$$

Since  $\lambda = 1$ , we have  $p_i = q_i$  for all  $i$ . Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$\begin{aligned} \frac{\partial^2}{\partial q_j^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= -\frac{p_j}{q_j^2} - \frac{p_K}{(1 - \sum_{i \neq K} q_i(x))^2} \\ &= -\left( \underbrace{\frac{p_j}{q_j^2}}_{> 0} + \underbrace{\frac{p_K}{(1 - \sum_{i \neq K} q_i(x))^2}}_{> 0} \right) < 0 \end{aligned} \quad (10)$$

Except for the cases  $q_j = 0$  and  $q_j = 1$ , the second derivative is always negative, which means that the function is concave and the maximum is unique. For  $q_j = 1$ , maximum is achieved only if  $p_j = 1$ , and similarly for  $q_j = 0$  maximum is achieved only if  $p_j = 0$ . Therefore,  $p_j = q_j$  is the only maximizer and the log probability scoring rule for multiclass classification is strictly proper.

2b) Multiclass log probability scoring rule (Proof 2):

Alternatively, we can solve the optimization problem with Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(q, \lambda) = \sum_k P(Y = k|x) \log(q_k(x)) + \lambda \left( 1 - \sum_k q_k(x) \right) \quad (11)$$

Taking the derivative with respect to  $q_j$  and setting it to zero:

$$\begin{aligned} \frac{\partial}{\partial q_j} \mathcal{L}(q, \lambda) &= \frac{p_j}{q_j} - \lambda \\ 0 &= \frac{p_j}{q_j} - \lambda \\ \Rightarrow \frac{p_j}{q_j} &= \lambda \end{aligned} \quad (12)$$

The rest of the proof follows as in the first proof.

3) Continuous log probability scoring rule:

$$S(q, y) = \log q(y) \quad (13)$$

Let  $p(y)$  be the true probability density function ( $\rightarrow$  I/1.7.1) of the event  $Y = y$ . Then, the expected ( $\rightarrow$  I/1.12.1) score ( $\rightarrow$  I/6.1.1) is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = \int p(y) \log q(y) dy \quad (14)$$

Let  $X = \frac{q(y)}{p(y)}$  and  $\phi = \log(\cdot)$  (a concave function). By Jensen's inequality, we know that  $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ , if  $f$  is concave. Therefore, we have:

$$\begin{aligned} \int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log \int p(y) \frac{q(y)}{p(y)} dy \\ \int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log \int q(y) dy \\ \int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log(1) \\ \int p(y) \log \frac{q(y)}{p(y)} dy &\leq 0 \end{aligned} \quad (15)$$

The same result can be obtained by using the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1). The Kullback-Leibler divergence is always non-negative ( $\rightarrow$  I/2.5.3), therefore  $E - CE = KL \geq 0$ . The resulting expression is  $-KL$ , which is always non-positive. It is maximized only when  $q(y) = p(y)$  which means that the log probability scoring rule for continuous classification is strictly proper.

An alternative argument for uniqueness of the maximum point can be proposed as follows:  $\int p(y) \log \frac{q(y)}{p(y)} dy$  can be equal to 0 in two cases: Either  $\frac{q(y)}{p(y)}$  is equal to 1 for each value or the expression  $\log(\frac{q(y)}{p(y)})$  takes positive and negative values summing up to 0 at the end. The second case cannot occur, because it means that there exists a  $y_0$  such that  $q(y_0) > p(y_0)$ , implying that Jensen's inequality is violated. Therefore, the maximum is achieved, if and only if  $q = p$ . ■

#### Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): "Proper/Strictly Proper Scoring Rule"; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.

#### 6.1.6 Brier scoring rule

**Definition:** A Brier scoring rule ( $\rightarrow$  I/6.1.1)  $S(q, y)$  is as a scoring rule that measures the quality of a probabilistic forecast in decision theory. Formally, it can be defined for binary or multiclass classification as follows:

1) Brier scoring rule for binary classification:

$$S(q, y) = -(q - y)^2 \quad (1)$$

$q$  represents the predicted probability of the positive class ( $Y = 1$ ) and  $y$  is the true class label. Since we want the output of the scoring rule to be maximized when the predicted probability is close to the true class label, we use the negative of the squared difference between the predicted probability and the true class label.

2) Brier scoring rule for multiclass classification:

$$S(q, y) = - \sum_k (q_k - y_k)^2 = -(q_{y^*} - 1)^2 - \sum_{k \neq y^*} q_k^2 \quad (2)$$

where  $q_k$  is the predicted probability of class  $k$  and  $y^*$  is the true class label. Similar to the log probability scoring rule, we have  $y_k = 1$ , if the true class is  $k$  and  $y_k = 0$  otherwise.

3) Regression (continuous case):

Although there is no direct version of Brier score for regression, we can use the squared error loss as a scoring rule for regression problems as well.

#### Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.

### 6.1.7 Brier scoring rule is strictly proper scoring rule

**Theorem:** The brier scoring rule ( $\rightarrow$  I/6.1.6) is a strictly proper scoring rule ( $\rightarrow$  I/6.1.3).

**Proof:** We will show that both versions of the brier scoring rule (binary/multiclass) are strictly proper scoring rules.

1) Brier scoring rule for binary classification:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -P(Y = 1)(q - 1)^2 + P(Y = 0) - q^2 \quad (1)$$

Let  $p$  be the true probability of the event  $Y = 1$ . Then, the expected score is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -p(q - 1)^2 - (1 - p)q^2 \quad (2)$$

To find the maxima, take the derivative with respect to  $q$  and set it to zero:

$$\begin{aligned} \frac{\partial}{\partial q} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= -2p(q - 1) - 2(1 - p)q \\ &= -2pq + 2p - 2q + 2pq \\ &= 2p - 2q \\ 0 &= 2p - 2q \\ \Rightarrow p &= q \end{aligned} \quad (3)$$

We need to check the second derivative to see if it is a maximum (for the properness condition) and if it is the only maximizer (for the strictness condition):

$$\frac{\partial^2}{\partial q^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -2 < 0 \quad (4)$$

The second derivative is always negative which means that the function is concave and the maximum is unique. Therefore,  $p = q$  is the only maximizer and the Brier scoring rule for binary classification is strictly proper.

2) Brier scoring rule for multiclass classification:

$$\begin{aligned}
\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= \sum_k P(Y = k) \left[ - \sum_i (q_i - y_i)^2 \right] \\
&= \sum_k P(Y = k) \left[ - (q_k - 1)^2 - \sum_{i \neq k} q_i^2 \right] \\
&= \sum_k P(Y = k) \left[ - (q_k - 1)^2 + q_k^2 - \sum_i q_i^2 \right] \\
&= \sum_k P(Y = k) \left[ - q_k^2 - 1 + 2q_k + q_k^2 - \sum_i q_i^2 \right] \\
&= \sum_k P(Y = k) \left[ 2q_k - 1 - \sum_i q_i^2 \right] \\
&= \sum_k P(Y = k) (2q_k - 1) - \sum_k P(Y = k) \left( \sum_i q_i^2 \right) \\
&= \sum_k P(Y = k) (2q_k - 1) - \sum_i q_i^2 \underbrace{\left( \sum_k P(Y = k) \right)}_1 \\
&= \sum_k P(Y = k) (2q_k - 1) - \sum_i q_i^2 \\
&= \sum_k P(Y = k) (2q_k - 1) - q_k^2
\end{aligned} \tag{5}$$

Similar to what we did for log probability ( $\rightarrow$  I/6.1.5), this expression can be rewritten as follows (replacing  $q_K$  with  $1 - \sum_{i \neq K} q_i$ ):

$$\begin{aligned}
\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= p_1(2q_1 - 1) - q_1^2 + p_2(2q_2 - 1) - q_2^2 + \dots + p_K(2q_K - 1) - q_K^2 \\
&= p_1(2q_1 - 1) - q_1^2 + p_2(2q_2 - 1) - q_2^2 + \dots + p_K \left( 1 - 2 \sum_{i \neq K} q_i \right) - \left( 1 - \sum_{i \neq K} q_i \right)^2
\end{aligned} \tag{6}$$

Taking the derivative with respect to  $q_j$  and setting it to zero, we obtain:

$$\begin{aligned}
\frac{\partial}{\partial q_j} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= 2p_j - 2q_j - 2p_K + 2 \left( 1 - \sum_{i \neq K} q_i \right) \\
&= 2p_j - 2q_j - 2p_K + 2q_K \\
&= (p_j - q_j) + (q_K - p_K) \\
(p_j - q_j) &= (p_K - q_K)
\end{aligned} \tag{7}$$

We know that  $\sum_i p_i = 1$  and  $\sum_i q_i = 1$ , therefore:

$$\begin{aligned}
p_1 - q_1 &= p_2 - q_2 = \dots = p_K - q_K = \lambda \\
\sum_i p_i - q_i &= K \cdot \lambda = 0 \\
&\Rightarrow \lambda = 0 \quad \text{since } K \neq 0 \\
&\Rightarrow p_i = q_i \quad \text{for all } i = 1, \dots, K
\end{aligned} \tag{8}$$

Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$\frac{\partial^2}{\partial q_j^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -2 - 2 = -4 < 0 \tag{9}$$

The second derivative is always negative which means that the function is concave and the maximum is unique. Therefore,  $p = q$  is the only maximizer and the Brier scoring rule for multiclass classification is strictly proper.

■

#### Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: <https://trustworthyml.io/>; DOI: 10.48550/arXiv.2310.08215.



## Chapter II

# Probability Distributions

# 1 Univariate discrete distributions

## 1.0.1 Definition

**Definition:** Let  $X$  be a discrete random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be uniformly distributed with minimum  $a$  and maximum  $b$

$$X \sim \mathcal{U}(a, b) , \quad (1)$$

if and only if each integer between and including  $a$  and  $b$  occurs with the same probability.

**Sources:**

- Wikipedia (2020): “Discrete uniform distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: [https://en.wikipedia.org/wiki/Discrete\\_uniform\\_distribution](https://en.wikipedia.org/wiki/Discrete_uniform_distribution).

## 1.0.2 Probability mass function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a discrete uniform distribution ( $\rightarrow$  II/1.0.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \frac{1}{b - a + 1} \quad \text{where} \quad x \in \{a, a + 1, \dots, b - 1, b\} . \quad (2)$$

**Proof:** A discrete uniform variable is defined as ( $\rightarrow$  II/1.0.1) having the same probability for each integer between and including  $a$  and  $b$ . The number of integers between and including  $a$  and  $b$  is

$$n = b - a + 1 \quad (3)$$

and because the sum across all probabilities ( $\rightarrow$  I/1.6.1) is

$$\sum_{x=a}^b f_X(x) = 1 , \quad (4)$$

we have

$$f_X(x) = \frac{1}{n} = \frac{1}{b - a + 1} . \quad (5)$$

■

## 1.0.3 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a discrete uniform distribution ( $\rightarrow$  II/1.0.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is



$$F_X(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{\lfloor x \rfloor - a + 1}{b - a + 1}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b. \end{cases} \quad (2)$$

**Proof:** The probability mass function of the discrete uniform distribution ( $\rightarrow$  II/1.0.2) is

$$\mathcal{U}(x; a, b) = \frac{1}{b - a + 1} \quad \text{where } x \in \{a, a + 1, \dots, b - 1, b\}. \quad (3)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^x \mathcal{U}(z; a, b) \, dz \quad (4)$$

From (3), it follows that the cumulative probability increases step-wise by  $1/n$  at each integer between and including  $a$  and  $b$  where

$$n = b - a + 1 \quad (5)$$

is the number of integers between and including  $a$  and  $b$ . This can be expressed by noting that

$$F_X(x) \stackrel{(3)}{=} \frac{\lfloor x \rfloor - a + 1}{n}, \quad \text{if } a \leq x \leq b. \quad (6)$$

Also, because  $\Pr(X < a) = 0$ , we have

$$F_X(x) \stackrel{(4)}{=} \int_{-\infty}^x 0 \, dz = 0, \quad \text{if } x < a \quad (7)$$

and because  $\Pr(X > b) = 0$ , we have

$$\begin{aligned} F_X(x) &\stackrel{(4)}{=} \int_{-\infty}^x \mathcal{U}(z; a, b) \, dz \\ &= \int_{-\infty}^b \mathcal{U}(z; a, b) \, dz + \int_b^x \mathcal{U}(z; a, b) \, dz \\ &= F_X(b) + \int_b^x 0 \, dz \stackrel{(6)}{=} 1 + 0 \\ &= 1, \quad \text{if } x > b. \end{aligned} \quad (8)$$

This completes the proof. ■

#### 1.0.4 Quantile function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a discrete uniform distribution ( $\rightarrow$  II/1.0.1):

$$X \sim \mathcal{U}(a, b). \quad (1)$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \begin{cases} -\infty, & \text{if } p = 0 \\ a(1-p) + (b+1)p - 1, & \text{when } p \in \left\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{b-a}{n}, 1\right\} \end{cases} \quad (2)$$

with  $n = b - a + 1$ .

**Proof:** The cumulative distribution function of the discrete uniform distribution ( $\rightarrow$  II/1.0.3) is:

$$F_X(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{\lfloor x \rfloor - a + 1}{b - a + 1}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases} \quad (3)$$

The quantile function ( $\rightarrow$  I/1.11.1)  $Q_X(p)$  is defined as the smallest  $x$ , such that  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} \quad (4)$$

Because the CDF only returns ( $\rightarrow$  II/1.0.3) multiples of  $1/n$  with  $n = b - a + 1$ , the quantile function ( $\rightarrow$  I/1.11.1) is only defined for such values. First, we have  $Q_X(p) = -\infty$ , if  $p = 0$ . Second, since the cumulative probability increases step-wise ( $\rightarrow$  II/1.0.3) by  $1/n$  at each integer between and including  $a$  and  $b$ , the minimum  $x$  at which

$$F_X(x) = \frac{c}{n} \quad \text{where } c \in \{1, \dots, n\} \quad (5)$$

is given by

$$Q_X\left(\frac{c}{n}\right) = a + \frac{c}{n} \cdot n - 1 \quad (6)$$

Substituting  $p = c/n$  and  $n = b - a + 1$ , we can finally show:

$$\begin{aligned} Q_X(p) &= a + p \cdot (b - a + 1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1-p) + (b+1)p - 1 \end{aligned} \quad (7)$$

■

### 1.0.5 Shannon entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a discrete uniform distribution ( $\rightarrow$  II/1.0.1):

$$X \sim \mathcal{U}(a, b) \quad (1)$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  in nats is

$$H(X) = \ln(b - a + 1) \quad (2)$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (3)$$

Entropy is measured in nats by setting  $b = e$ . Then, with the probability mass function of the discrete uniform distribution ( $\rightarrow$  II/1.0.2), we have:

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log_e p(x) \\ &= - \sum_{x=a}^b p(x) \cdot \ln p(x) \\ &= - \sum_{x=a}^b \frac{1}{b-a+1} \cdot \ln \frac{1}{b-a+1} \\ &= -(b-a+1) \cdot \frac{1}{b-a+1} \cdot \ln \frac{1}{b-a+1} \\ &= - \ln \frac{1}{b-a+1} \\ &= \ln(b-a+1) . \end{aligned} \quad (4)$$

■

### 1.0.6 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two discrete uniform distributions ( $\rightarrow$  II/??)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \mathcal{U}(a_1, b_1) \\ Q : X &\sim \mathcal{U}(a_2, b_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = \ln \frac{b_2 - a_2 + 1}{b_1 - a_1 + 1} . \quad (2)$$

**Proof:** The KL divergence for a discrete random variable ( $\rightarrow$  I/2.5.1) with possible values  $\mathcal{X}$  is given by

$$\text{KL}[P \parallel Q] = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} . \quad (3)$$

This means that the KL divergence of  $P$  from  $Q$  is only defined, if for all  $x \in \mathcal{X}$ ,  $q(x) = 0$  implies  $p(x) = 0$ . Thus,  $\text{KL}[P \parallel Q]$  only exists, if  $a_2 \leq a_1$  and  $b_1 \leq b_2$ , i.e. if  $P$  only places non-zero probability where  $Q$  also places non-zero probability, such that  $q(x)$  is not zero for any  $x \in \mathcal{X}$  where  $p(x)$  is positive.

If this requirement is fulfilled, we can write

$$\text{KL}[P \parallel Q] = \sum_{x=-\infty}^{a_1} p(x) \ln \frac{p(x)}{q(x)} + \sum_{x=a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} + \sum_{x=b_1}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} \quad (4)$$

and because  $p(x) = 0$  for any  $x < a_1$  and any  $x > b_1$ , we have

$$\text{KL}[P \parallel Q] = \sum_{x=-\infty}^{a_1} 0 \cdot \ln \frac{0}{q(x)} + \sum_{x=a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} + \sum_{x=b_1}^{+\infty} 0 \cdot \ln \frac{0}{q(x)}. \quad (5)$$

Now,  $(0 \cdot \ln 0)$  is taken to be 0 by convention ( $\rightarrow$  I/2.1.1), such that

$$\text{KL}[P \parallel Q] = \sum_{x=a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} \quad (6)$$

and we can use the probability mass function of the discrete uniform distribution ( $\rightarrow$  II/1.0.2) to evaluate:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \sum_{x=a_1}^{b_1} \frac{1}{b_1 - a_1 + 1} \cdot \ln \frac{\frac{1}{b_1 - a_1 + 1}}{\frac{1}{b_2 - a_2 + 1}} \\ &= \frac{1}{b_1 - a_1 + 1} \cdot \ln \frac{b_2 - a_2 + 1}{b_1 - a_1 + 1} \sum_{x=a_1}^{b_1} 1 \\ &= \frac{1}{b_1 - a_1 + 1} \cdot \ln \frac{b_2 - a_2 + 1}{b_1 - a_1 + 1} \cdot (b_1 - a_1 + 1) \\ &= \ln \frac{b_2 - a_2 + 1}{b_1 - a_1 + 1}. \end{aligned} \quad (7)$$

■

### 1.0.7 Maximum entropy distribution

**Theorem:** The discrete uniform distribution ( $\rightarrow$  II/1.0.1) maximizes (Shannon) entropy ( $\rightarrow$  I/2.1.1) for a random variable ( $\rightarrow$  I/1.2.2) with finite support.

**Proof:** A random variable with finite support is a discrete random variable ( $\rightarrow$  I/1.2.6). Let  $X$  be such a random variable. Without loss of generality, we can assume that the possible values of the  $X$  can be enumerated from 1 to  $n$ .

Let  $g(x)$  be the discrete uniform distribution with minimum  $a = 1$  and maximum  $b = n$  which assigns to equal probability to all  $n$  possible values and let  $f(x)$  be an arbitrary discrete ( $\rightarrow$  I/1.2.6) probability distribution ( $\rightarrow$  I/1.5.1) on the set  $\{1, 2, \dots, n-1, n\}$ .

For a discrete random variable ( $\rightarrow$  I/1.2.6)  $X$  with set of possible values  $\mathcal{X}$  and probability mass function ( $\rightarrow$  I/1.6.1)  $p(x)$ , the Shannon entropy ( $\rightarrow$  I/2.1.1) is defined as:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (1)$$

Consider the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of distribution  $f(x)$  from distribution  $g(x)$  which is non-negative ( $\rightarrow$  I/2.5.2):

$$\begin{aligned}
0 \leq \text{KL}[f||g] &= \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \\
&= \sum_{x \in \mathcal{X}} f(x) \log f(x) - \sum_{x \in \mathcal{X}} f(x) \log g(x) \\
&\stackrel{(1)}{=} -H[f(x)] - \sum_{x \in \mathcal{X}} f(x) \log g(x) .
\end{aligned} \tag{2}$$

By plugging the probability mass function of the discrete uniform distribution ( $\rightarrow \text{II}/1.0.2$ ) into the second term, we obtain:

$$\begin{aligned}
\sum_{x \in \mathcal{X}} f(x) \log g(x) &= \sum_{x=1}^n f(x) \log \frac{1}{n-1+1} \\
&= \log \frac{1}{n} \sum_{x=1}^n f(x) \\
&= -\log(n) .
\end{aligned} \tag{3}$$

This is actually the negative of the entropy of the discrete uniform distribution ( $\rightarrow \text{II}/1.0.5$ ), such that:

$$\sum_{x \in \mathcal{X}} f(x) \log g(x) = -H[\mathcal{U}(1, n)] = -H[g(x)] . \tag{4}$$

Combining (2) with (4), we can show that

$$\begin{aligned}
0 &\leq \text{KL}[f||g] \\
0 &\leq -H[f(x)] - (-H[g(x)]) \\
H[g(x)] &\geq H[f(x)]
\end{aligned} \tag{5}$$

which means that the entropy ( $\rightarrow \text{I}/2.1.1$ ) of the discrete uniform distribution ( $\rightarrow \text{II}/1.0.1$ )  $\mathcal{U}(a, b)$  will be larger than or equal to any other distribution ( $\rightarrow \text{I}/1.5.1$ ) defined on the same set of values  $\{a, \dots, b\}$ . ■

#### Sources:

- Probability Fact (2023): “The entropy of a distribution with finite domain”; in: *Twitter*, retrieved on 2023-08-18; URL: <https://twitter.com/ProbFact/status/1673787091610750980>.

## 1.1 Bernoulli distribution

### 1.1.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow \text{I}/1.2.2$ ). Then,  $X$  is said to follow a Bernoulli distribution with success probability  $p$

$$X \sim \text{Bern}(p) , \tag{1}$$

if  $X = 1$  with probability ( $\rightarrow$  I/1.3.1)  $p$  and  $X = 0$  with probability ( $\rightarrow$  I/1.3.1)  $q = 1 - p$ .

**Sources:**

- Wikipedia (2020): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Bernoulli\\_distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution).

### 1.1.2 Special case of categorical distribution

**Theorem:** The Bernoulli distribution ( $\rightarrow$  II/1.1.1) with success probability  $p$  is a special case of the categorical distribution ( $\rightarrow$  II/2.1.1) with category probabilities  $p_1 = p$  and  $p_2 = 1 - p$ :

$$X \sim \text{Cat}([p, 1 - p]) \quad \Rightarrow \quad X \sim \text{Bern}(p) . \quad (1)$$

**Proof:** The probability mass function of the categorical distribution ( $\rightarrow$  II/2.1.2), where  $x$  is a  $1 \times k$  vector and  $e_1, \dots, e_k$  are the  $1 \times k$  elementary row vectors, is as follows:

$$\text{Cat}(x; [p_1, \dots, p_k]) = \begin{cases} p_1 , & \text{if } x = e_1 \\ \vdots & \vdots \\ p_k , & \text{if } x = e_k . \end{cases} \quad (2)$$

If we let  $k = 2$ ,  $p_1 = p$  and  $p_2 = 1 - p$ , we obtain

$$\text{Cat}(x; [p, 1 - p]) = \begin{cases} p , & \text{if } x = [1, 0] \\ 1 - p , & \text{if } x = [0, 1] . \end{cases} \quad (3)$$

Writing this in terms of the first entry  $x_1 \in \{0, 1\}$ , we get

$$\text{Cat}(x; [p, 1 - p]) = \begin{cases} p , & \text{if } x_1 = 1 \\ 1 - p , & \text{if } x_1 = 0 \end{cases} \quad (4)$$

which is equivalent to the probability mass function of the Bernoulli distribution ( $\rightarrow$  II/1.1.3). ■

### 1.1.3 Probability mass function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Bernoulli distribution ( $\rightarrow$  II/1.1.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \begin{cases} p , & \text{if } x = 1 \\ 1 - p , & \text{if } x = 0 . \end{cases} \quad (2)$$

**Proof:** This follows directly from the definition of the Bernoulli distribution ( $\rightarrow$  II/1.1.1). ■

### 1.1.4 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Bernoulli distribution ( $\rightarrow$  II/1.1.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = p . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average of all possible values:

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x) . \quad (3)$$

Since there are only two possible outcomes for a Bernoulli random variable ( $\rightarrow$  II/1.1.3), we have:

$$\begin{aligned} E(X) &= 0 \cdot \Pr(X = 0) + 1 \cdot \Pr(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p . \end{aligned} \quad (4)$$

■

#### Sources:

- Wikipedia (2020): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: [https://en.wikipedia.org/wiki/Bernoulli\\_distribution#Mean](https://en.wikipedia.org/wiki/Bernoulli_distribution#Mean).

### 1.1.5 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Bernoulli distribution ( $\rightarrow$  II/1.1.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.12.1) of  $X$  is

$$\text{Var}(X) = p(1 - p) . \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is the probability-weighted average of the squared deviation from the expected value ( $\rightarrow$  I/1.12.1) across all possible values

$$\text{Var}(X) = \sum_{x \in \mathcal{X}} (x - E(X))^2 \cdot \Pr(X = x) \quad (3)$$

and can also be written in terms of the expected values ( $\rightarrow$  I/1.13.4):

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (4)$$

The mean of a Bernoulli random variable ( $\rightarrow$  II/1.1.4) is

$$X \sim \text{Bern}(p) \quad \Rightarrow \quad E(X) = p \quad (5)$$

and the mean of a squared Bernoulli random variable is

$$E(X^2) = 0^2 \cdot \Pr(X = 0) + 1^2 \cdot \Pr(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p. \quad (6)$$

Combining (4), (5) and (6), we have:

$$\text{Var}(X) = p - p^2 = p(1 - p). \quad (7)$$

■

#### Sources:

- Wikipedia (2022): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-20; URL: [https://en.wikipedia.org/wiki/Bernoulli\\_distribution#Variance](https://en.wikipedia.org/wiki/Bernoulli_distribution#Variance).

### 1.1.6 Range of variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Bernoulli distribution ( $\rightarrow$  II/1.1.1):

$$X \sim \text{Bern}(p). \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is necessarily between 0 and 1/4:

$$0 \leq \text{Var}(X) \leq \frac{1}{4}. \quad (2)$$

**Proof:** The variance of a Bernoulli random variable ( $\rightarrow$  II/1.1.5) is

$$X \sim \text{Bern}(p) \quad \Rightarrow \quad \text{Var}(X) = p(1 - p) \quad (3)$$

which can also be understood as a function of the success probability ( $\rightarrow$  II/1.1.1)  $p$ :

$$\text{Var}(X) = \text{Var}(p) = -p^2 + p. \quad (4)$$

The first derivative of this function is

$$\frac{d\text{Var}(p)}{dp} = -2p + 1 \quad (5)$$

and setting this derivative to zero

$$\begin{aligned} \frac{d\text{Var}(p_M)}{dp} &= 0 \\ 0 &= -2p_M + 1 \\ p_M &= \frac{1}{2}, \end{aligned} \quad (6)$$

we obtain the maximum possible variance

$$\max[\text{Var}(X)] = \text{Var}(p_M) = -\left(\frac{1}{2}\right)^2 + \frac{1}{2} = \frac{1}{4}. \quad (7)$$

The function  $\text{Var}(p)$  is monotonically increasing for  $0 < p < p_M$  as  $d\text{Var}(p)/dp > 0$  in this interval and it is monotonically decreasing for  $p_M < p < 1$  as  $d\text{Var}(p)/dp < 0$  in this interval. Moreover, as variance is always non-negative ( $\rightarrow$  I/1.13.5), the minimum variance is



$$\min [\text{Var}(X)] = \text{Var}(0) = \text{Var}(1) = 0 . \quad (8)$$

Thus, we have:

$$\text{Var}(p) \in \left[ 0, \frac{1}{4} \right] . \quad (9)$$

■

#### Sources:

- Wikipedia (2022): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-27; URL: [https://en.wikipedia.org/wiki/Bernoulli\\_distribution#Variance](https://en.wikipedia.org/wiki/Bernoulli_distribution#Variance).

### 1.1.7 Shannon entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Bernoulli distribution ( $\rightarrow$  II/1.1.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  in bits is

$$H(X) = -p \log_2 p - (1 - p) \log_2 (1 - p) . \quad (2)$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (3)$$

Entropy is measured in bits by setting  $b = 2$ . Since there are only two possible outcomes for a Bernoulli random variable ( $\rightarrow$  II/1.1.3), we have:

$$\begin{aligned} H(X) &= -\Pr(X = 0) \cdot \log_2 \Pr(X = 0) - \Pr(X = 1) \cdot \log_2 \Pr(X = 1) \\ &= -p \log_2 p - (1 - p) \log_2 (1 - p) . \end{aligned} \quad (4)$$

■

#### Sources:

- Wikipedia (2022): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-02; URL: [https://en.wikipedia.org/wiki/Bernoulli\\_distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution).
- Wikipedia (2022): “Binary entropy function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-02; URL: [https://en.wikipedia.org/wiki/Binary\\_entropy\\_function](https://en.wikipedia.org/wiki/Binary_entropy_function).

### 1.1.8 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two Bernoulli distributions ( $\rightarrow$  II/1.1.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \text{Bern}(p_1) \\ Q : X &\sim \text{Bern}(p_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1(1-p_2)}{p_2(1-p_1)} . \quad (2)$$

**Proof:** The KL divergence for a discrete random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \quad (3)$$

which, applied to the Bernoulli distributions ( $\rightarrow$  II/1.1.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \sum_{x \in \{0,1\}} p(x) \ln \frac{p(x)}{q(x)} \\ &= p(X=0) \cdot \ln \frac{p(X=0)}{q(X=0)} + p(X=1) \cdot \ln \frac{p(X=1)}{q(X=1)} . \end{aligned} \quad (4)$$

Using the probability mass function of the Bernoulli distribution ( $\rightarrow$  II/1.1.3), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= (1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1}{p_2} \\ &= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1}{p_2} - p_1 \cdot \ln \frac{1-p_1}{1-p_2} \\ &= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \left( \ln \frac{p_1}{p_2} + \ln \frac{1-p_2}{1-p_1} \right) \\ &= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1(1-p_2)}{p_2(1-p_1)} \end{aligned} \quad (5)$$

■

## 1.2 Binomial distribution

### 1.2.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to follow a binomial distribution with number of trials  $n$  and success probability  $p$

$$X \sim \text{Bin}(n, p) , \quad (1)$$

if  $X$  is the number of successes observed in  $n$  independent ( $\rightarrow$  I/1.3.6) trials, where each trial has two possible outcomes ( $\rightarrow$  II/1.1.1) (success/failure) and the probability of success and failure are identical across trials ( $p/q = 1-p$ ).

#### Sources:

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution](https://en.wikipedia.org/wiki/Binomial_distribution).

### 1.2.2 Special case of multinomial distribution

**Theorem:** The binomial distribution ( $\rightarrow$  II/1.2.1) with sample size  $n$  and success probability  $p$  is a special case of the multinomial distribution ( $\rightarrow$  II/2.2.1) with sample size  $n$  category probabilities  $p_1 = p$  and  $p_2 = 1 - p$ :

$$X \sim \text{Mult}(n, [p, 1 - p]) \Rightarrow X \sim \text{Bin}(p) . \quad (1)$$

**Proof:** The probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2), where  $x$  is a  $1 \times k$  vector with  $x_i \in \{0, 1, \dots, n\}$ , is as follows:

$$\text{Mult}(x; n, [p_1, \dots, p_k]) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} . \quad (2)$$

If we let  $k = 2$ ,  $p_1 = p$  and  $p_2 = 1 - p$ , we obtain

$$\text{Mult}(x; n, [p, 1 - p]) = \binom{n}{x_1, x_2} p^{x_1} (1 - p)^{n - x_1} . \quad (3)$$

Recognizing that  $\sum_{i=1}^k x_i = n$ , such that  $x_2 = n - x_1$ , we get

$$\begin{aligned} \text{Mult}(x; n, [p, 1 - p]) &= \binom{n}{x_1, n - x_1} p^{x_1} (1 - p)^{n - x_1} \\ &= \binom{n}{x_1} p^{x_1} (1 - p)^{n - x_1} \end{aligned} \quad (4)$$

which is equivalent to the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3). ■

### 1.2.3 Probability mass function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \binom{n}{x} p^x (1 - p)^{n - x} . \quad (2)$$

**Proof:** A binomial variable ( $\rightarrow$  II/1.2.1) is defined as the number of successes observed in  $n$  independent ( $\rightarrow$  I/1.3.6) trials, where each trial has two possible outcomes ( $\rightarrow$  II/1.1.1) (success/failure) and the probability ( $\rightarrow$  I/1.3.1) of success and failure are identical across trials ( $p, q = 1 - p$ ).

If one has obtained  $x$  successes in  $n$  trials, one has also obtained  $(n - x)$  failures. The probability of a particular series of  $x$  successes and  $(n - x)$  failures, when order does matter, is

$$p^x (1 - p)^{n - x} . \quad (3)$$

When order does not matter, there is a number of series consisting of  $x$  successes and  $(n - x)$  failures. This number is equal to the number of possibilities in which  $x$  objects can be chosen from  $n$  objects which is given by the binomial coefficient:

$$\binom{n}{x} . \quad (4)$$

In order to obtain the probability of  $x$  successes and  $(n - x)$  failures, when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (5)$$

which is equivalent to the expression above. ■

#### 1.2.4 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \sum_{k=0}^x \binom{n}{k} p^k (1 - p)^{n-k} . \quad (2)$$

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) is defined as

$$F_X(x) = \Pr(X \leq x) \quad (3)$$

which, in terms of the probability mass function, is given by ( $\rightarrow$  I/1.8.5)

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) . \quad (4)$$

The probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3) is

$$f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} , \quad (5)$$

so that the cumulative distribution function of the binomial distribution becomes

$$\begin{aligned} F_X(x) &\stackrel{(4)}{=} \sum_{k=0}^x f_X(k) \\ &\stackrel{(5)}{=} \sum_{k=0}^x \binom{n}{k} p^k (1 - p)^{n-k} . \end{aligned} \quad (6)$$

**Sources:** ■

- Wikipedia (2025): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-02-06; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Cumulative\\_distribution\\_function](https://en.wikipedia.org/wiki/Binomial_distribution#Cumulative_distribution_function).

### 1.2.5 Probability-generating function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$  is

$$G_X(z) = (q + pz)^n \quad (2)$$

where  $q = 1 - p$ .

**Proof:** The probability-generating function ( $\rightarrow$  I/1.10.1) of  $X$  is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x) z^x \quad (3)$$

With the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3)

$$f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} , \quad (4)$$

we obtain:

$$\begin{aligned} G_X(z) &= \sum_{x=0}^n \binom{n}{x} p^x (1 - p)^{n-x} z^x \\ &= \sum_{x=0}^n \binom{n}{x} (pz)^x (1 - p)^{n-x} . \end{aligned} \quad (5)$$

According to the binomial theorem

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k , \quad (6)$$

the sum in equation (5) equals

$$G_X(z) = ((1 - p) + (pz))^n \quad (7)$$

which is equivalent to the result in (2). ■

#### Sources:

- ProofWiki (2022): “Probability Generating Function of Binomial Distribution”; in: *ProofWiki*, retrieved on 2022-10-11; URL: [https://proofwiki.org/wiki/Probability\\_Generating\\_Function\\_of\\_Binomial\\_Distribution](https://proofwiki.org/wiki/Probability_Generating_Function_of_Binomial_Distribution).

### 1.2.6 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = np . \quad (2)$$

**Proof:** By definition, a binomial random variable ( $\rightarrow$  II/1.2.1) is the sum of  $n$  independent and identical ( $\rightarrow$  I/1.2.8) Bernoulli trials ( $\rightarrow$  II/1.1.1) with success probability  $p$ . Therefore, the expected value is

$$E(X) = E(X_1 + \dots + X_n) \quad (3)$$

and because the expected value is a linear operator ( $\rightarrow$  I/1.12.5), this is equal to

$$E(X) = E(X_1) + \dots + E(X_n) = \sum_{i=1}^n E(X_i) . \quad (4)$$

With the expected value of the Bernoulli distribution ( $\rightarrow$  II/1.1.4), we have:

$$E(X) = \sum_{i=1}^n p = np . \quad (5)$$

■

#### Sources:

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Expected\\_value\\_and\\_variance](https://en.wikipedia.org/wiki/Binomial_distribution#Expected_value_and_variance).

### 1.2.7 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = np(1 - p) . \quad (2)$$

**Proof:** By definition, a binomial random variable ( $\rightarrow$  II/1.2.1) is the sum of  $n$  independent and identical ( $\rightarrow$  I/1.2.8) Bernoulli trials ( $\rightarrow$  II/1.1.1) with success probability  $p$ . Therefore, the variance is

$$\text{Var}(X) = \text{Var}(X_1 + \dots + X_n) \quad (3)$$

and because variances add up under independence ( $\rightarrow$  I/1.13.11), this is equal to

$$\text{Var}(X) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = \sum_{i=1}^n \text{Var}(X_i) . \quad (4)$$

With the variance of the Bernoulli distribution ( $\rightarrow$  II/1.1.5), we have:

$$\text{Var}(X) = \sum_{i=1}^n p(1-p) = np(1-p) . \quad (5)$$

#### Sources:

- Wikipedia (2022): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-20; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Expected\\_value\\_and\\_variance](https://en.wikipedia.org/wiki/Binomial_distribution#Expected_value_and_variance).

### 1.2.8 Range of variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is necessarily between 0 and  $n/4$ :

$$0 \leq \text{Var}(X) \leq \frac{n}{4} . \quad (2)$$

**Proof:** By definition, a binomial random variable ( $\rightarrow$  II/1.2.1) is the sum of  $n$  independent and identical ( $\rightarrow$  I/1.2.8) Bernoulli trials ( $\rightarrow$  II/1.1.1) with success probability  $p$ . Therefore, the variance is

$$\text{Var}(X) = \text{Var}(X_1 + \dots + X_n) \quad (3)$$

and because variances add up under independence ( $\rightarrow$  I/1.13.11), this is equal to

$$\text{Var}(X) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = \sum_{i=1}^n \text{Var}(X_i) . \quad (4)$$

As the variance of a Bernoulli random variable is always between 0 and  $1/4$  ( $\rightarrow$  II/1.1.6)

$$0 \leq \text{Var}(X_i) \leq \frac{1}{4} \quad \text{for all } i = 1, \dots, n , \quad (5)$$

the minimum variance of  $X$  is

$$\min [\text{Var}(X)] = n \cdot 0 = 0 \quad (6)$$

and the maximum variance of  $X$  is

$$\max [\text{Var}(X)] = n \cdot \frac{1}{4} = \frac{n}{4} . \quad (7)$$

Thus, we have:

$$\text{Var}(X) \in \left[0, \frac{n}{4}\right] . \quad (8)$$

### 1.2.9 Shannon entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  in bits is

$$H(X) = n \cdot H_{\text{bern}}(p) - E_{\text{lbc}}(n, p) \quad (2)$$

where  $H_{\text{bern}}(p)$  is the binary entropy function, i.e. the (Shannon) entropy of the Bernoulli distribution ( $\rightarrow$  II/1.1.7) with success probability  $p$

$$H_{\text{bern}}(p) = -p \cdot \log_2 p - (1 - p) \log_2(1 - p) \quad (3)$$

and  $E_{\text{lbc}}(n, p)$  is the expected value ( $\rightarrow$  I/1.12.1) of the logarithmized binomial coefficient ( $\rightarrow$  II/1.2.3) with superset size  $n$

$$E_{\text{lbc}}(n, p) = E \left[ \log_2 \binom{n}{X} \right] \quad \text{where } X \sim \text{Bin}(n, p) . \quad (4)$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (5)$$

Entropy is measured in bits by setting  $b = 2$ . Then, with the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), we have:

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} f_X(x) \cdot \log_2 f_X(x) \\ &= - \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} \cdot \log_2 \left[ \binom{n}{x} p^x (1-p)^{n-x} \right] \\ &= - \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} \cdot \left[ \log_2 \binom{n}{x} + x \cdot \log_2 p + (n-x) \cdot \log_2(1-p) \right] \\ &= - \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} \cdot \left[ \log_2 \binom{n}{x} + x \cdot \log_2 p + n \cdot \log_2(1-p) - x \cdot \log_2(1-p) \right] . \end{aligned} \quad (6)$$

Since the first factor in the sum corresponds to the probability mass ( $\rightarrow$  I/1.6.1) of  $X = x$ , we can rewrite this as the sum of the expected values ( $\rightarrow$  I/1.12.1) of the functions ( $\rightarrow$  I/1.12.9) of the discrete random variable ( $\rightarrow$  I/1.2.6)  $x$  in the square bracket:

$$\begin{aligned} H(X) &= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - \langle x \cdot \log_2 p \rangle_{p(x)} - \langle n \cdot \log_2(1-p) \rangle_{p(x)} + \langle x \cdot \log_2(1-p) \rangle_{p(x)} \\ &= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - \log_2 p \cdot \langle x \rangle_{p(x)} - n \cdot \log_2(1-p) + \log_2(1-p) \cdot \langle x \rangle_{p(x)} . \end{aligned} \quad (7)$$



Using the expected value of the binomial distribution ( $\rightarrow$  II/1.2.6), i.e.  $X \sim \text{Bin}(n, p) \Rightarrow \langle x \rangle = np$ , this gives:

$$\begin{aligned} H(X) &= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - np \cdot \log_2 p - n \cdot \log_2(1-p) + np \cdot \log_2(1-p) \\ &= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} + n [-p \cdot \log_2 p - (1-p) \log_2(1-p)] . \end{aligned} \quad (8)$$

Finally, we note that the first term is the negative expected value ( $\rightarrow$  I/1.12.1) of the logarithm of a binomial coefficient ( $\rightarrow$  II/1.2.3) and that the term in square brackets is the entropy of the Bernoulli distribution ( $\rightarrow$  II/1.2.9), such that we finally get:

$$H(X) = n \cdot H_{\text{bern}}(p) - E_{\text{lbc}}(n, p) . \quad (9)$$

Note that, because  $0 \leq H_{\text{bern}}(p) \leq 1$ , we have  $0 \leq n \cdot H_{\text{bern}}(p) \leq n$ , and because the entropy is non-negative ( $\rightarrow$  I/2.1.2), it must hold that  $n \geq E_{\text{lbc}}(n, p) \geq 0$ . ■

### 1.2.10 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two binomial distributions ( $\rightarrow$  II/1.2.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \text{Bin}(n, p_1) \\ Q : X &\sim \text{Bin}(n, p_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P || Q] = np_1 \cdot \ln \frac{p_1}{p_2} + n(1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} . \quad (2)$$

**Proof:** The KL divergence for a discrete random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P || Q] = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \quad (3)$$

which, applied to the binomial distributions ( $\rightarrow$  II/1.2.1) in (1), yields

$$\begin{aligned} \text{KL}[P || Q] &= \sum_{x=0}^n p(x) \ln \frac{p(x)}{q(x)} \\ &= p(X=0) \cdot \ln \frac{p(X=0)}{q(X=0)} + \dots + p(X=n) \cdot \ln \frac{p(X=n)}{q(X=n)} . \end{aligned} \quad (4)$$

Using the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), this becomes:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} \cdot \ln \frac{\binom{n}{x} p_1^x (1-p_1)^{n-x}}{\binom{n}{x} p_2^x (1-p_2)^{n-x}} \\
&= \sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} \cdot \left[ x \cdot \ln \frac{p_1}{p_2} + (n-x) \cdot \ln \frac{1-p_1}{1-p_2} \right] \\
&= \ln \frac{p_1}{p_2} \cdot \sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} x + \ln \frac{1-p_1}{1-p_2} \cdot \sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} (n-x) .
\end{aligned} \tag{5}$$

We can now see that some terms in this sum are expected values ( $\rightarrow$  I/1.12.1) with respect to binomial distributions ( $\rightarrow$  II/1.2.1):

$$\begin{aligned}
\sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} x &= \text{E}[x]_{\text{Bin}(n, p_1)} \\
\sum_{x=0}^n \binom{n}{x} p_1^x (1-p_1)^{n-x} (n-x) &= \text{E}[n-x]_{\text{Bin}(n, p_1)} .
\end{aligned} \tag{6}$$

Using the expected value of the binomial distribution ( $\rightarrow$  II/1.2.6), these can be simplified to

$$\begin{aligned}
\text{E}[x]_{\text{Bin}(n, p_1)} &= np_1 \\
\text{E}[n-x]_{\text{Bin}(n, p_1)} &= n - np_1 ,
\end{aligned} \tag{7}$$

such that the Kullback-Leibler divergence finally becomes:

$$\text{KL}[P \parallel Q] = np_1 \cdot \ln \frac{p_1}{p_2} + n(1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} . \tag{8}$$

■

#### Sources:

- PSPACEhard (2017): “Kullback-Leibler divergence for binomial distributions P and Q”; in: *Stack-Exchange Mathematics*, retrieved on 2023-10-20; URL: <https://math.stackexchange.com/a/2215384/480910>.

#### 1.2.11 Conditional binomial

**Theorem:** Let  $X$  and  $Y$  be two random variables ( $\rightarrow$  I/1.2.2) where  $Y$  is binomially distributed ( $\rightarrow$  II/1.2.1) conditional on ( $\rightarrow$  I/1.5.4)  $X$

$$Y|X \sim \text{Bin}(X, q) \tag{1}$$

and  $X$  also follows a binomial distribution ( $\rightarrow$  II/1.2.1), but with different success frequency ( $\rightarrow$  II/1.2.1):

$$X \sim \text{Bin}(n, p) . \tag{2}$$

Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of  $Y$  unconditional on  $X$  is again a binomial distribution ( $\rightarrow$  II/1.2.1):

$$Y \sim \text{Bin}(n, pq) . \quad (3)$$

**Proof:** We are interested in the probability that  $Y$  equals a number  $m$ . According to the law of marginal probability ( $\rightarrow$  I/1.3.3) or the law of total probability ( $\rightarrow$  I/1.4.11), this probability can be expressed as:

$$\Pr(Y = m) = \sum_{k=0}^{\infty} \Pr(Y = m|X = k) \cdot \Pr(X = k) . \quad (4)$$

Since, by definitions (2) and (1),  $\Pr(X = k) = 0$  when  $k > n$  and  $\Pr(Y = m|X = k) = 0$  when  $k < m$ , we have:

$$\Pr(Y = m) = \sum_{k=m}^n \Pr(Y = m|X = k) \cdot \Pr(X = k) . \quad (5)$$

Now we can take the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3) and plug it in for the terms in the sum of (5) to get:

$$\Pr(Y = m) = \sum_{k=m}^n \binom{k}{m} q^m (1-q)^{k-m} \cdot \binom{n}{k} p^k (1-p)^{n-k} . \quad (6)$$

Applying the binomial coefficient identity  $\binom{n}{k} \binom{k}{m} = \binom{n}{m} \binom{n-m}{k-m}$  and rearranging the terms, we have:

$$\Pr(Y = m) = \sum_{k=m}^n \binom{n}{m} \binom{n-m}{k-m} p^k q^m (1-p)^{n-k} (1-q)^{k-m} . \quad (7)$$

Now we partition  $p^k = p^m \cdot p^{k-m}$  and pull all terms dependent on  $k$  out of the sum:

$$\begin{aligned} \Pr(Y = m) &= \binom{n}{m} p^m q^m \sum_{k=m}^n \binom{n-m}{k-m} p^{k-m} (1-p)^{n-k} (1-q)^{k-m} \\ &= \binom{n}{m} (pq)^m \sum_{k=m}^n \binom{n-m}{k-m} (p(1-q))^{k-m} (1-p)^{n-k} . \end{aligned} \quad (8)$$

Then we substitute  $i = k - m$ , such that  $k = i + m$ :

$$\Pr(Y = m) = \binom{n}{m} (pq)^m \sum_{i=0}^{n-m} \binom{n-m}{i} (p-pq)^i (1-p)^{n-m-i} . \quad (9)$$

According to the binomial theorem

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k , \quad (10)$$

the sum in equation (9) is equal to

$$\sum_{i=0}^{n-m} \binom{n-m}{i} (p-pq)^i (1-p)^{n-m-i} = ((p-pq) + (1-p))^{n-m} . \quad (11)$$

Thus, (9) develops into

$$\begin{aligned}\Pr(Y = m) &= \binom{n}{m} (pq)^m (p - pq + 1 - p)^{n-m} \\ &= \binom{n}{m} (pq)^m (1 - pq)^{n-m}\end{aligned}\tag{12}$$

which is the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3) with parameters  $n$  and  $pq$ , such that

$$Y \sim \text{Bin}(n, pq) . \tag{13}$$

■

#### Sources:

- Wikipedia (2022): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-07; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Conditional\\_binomials](https://en.wikipedia.org/wiki/Binomial_distribution#Conditional_binomials).

## 1.3 Beta-binomial distribution

### 1.3.1 Definition

**Definition:** Let  $p$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1)

$$p \sim \text{Bet}(\alpha, \beta) \tag{1}$$

and let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a binomial distribution ( $\rightarrow$  II/1.2.1) conditional on  $p$

$$X \mid p \sim \text{Bin}(n, p) . \tag{2}$$

Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of  $X$  is called a beta-binomial distribution

$$X \sim \text{BetBin}(n, \alpha, \beta) \tag{3}$$

with number of trials ( $\rightarrow$  II/1.2.1)  $n$  and shape parameters ( $\rightarrow$  II/3.9.1)  $\alpha$  and  $\beta$ .

#### Sources:

- Wikipedia (2022): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: [https://en.wikipedia.org/wiki/Beta-binomial\\_distribution#Motivation\\_and\\_derivation](https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation).

### 1.3.2 Probability mass function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta-binomial distribution ( $\rightarrow$  II/1.3.1):

$$X \sim \text{BetBin}(n, \alpha, \beta) . \tag{1}$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \binom{n}{x} \cdot \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \quad (2)$$

where  $B(x, y)$  is the beta function.

**Proof:** A beta-binomial random variable ( $\rightarrow$  II/1.3.1) is defined as a binomial variate ( $\rightarrow$  II/1.2.1) for which the success probability is following a beta distribution ( $\rightarrow$  II/3.9.1):

$$\begin{aligned} X \mid p &\sim \text{Bin}(n, p) \\ p &\sim \text{Bet}(\alpha, \beta) . \end{aligned} \quad (3)$$

Thus, we can combine the law of marginal probability ( $\rightarrow$  I/1.3.3) and the law of conditional probability ( $\rightarrow$  I/1.3.4) to derive the probability ( $\rightarrow$  I/1.3.1) of  $X$  as

$$\begin{aligned} p(x) &= \int_{\mathcal{P}} p(x, p) \, dp \\ &= \int_{\mathcal{P}} p(x|p) p(p) \, dp \end{aligned} \quad (4)$$

where  $\mathcal{P} = [0, 1]$  is the set of possible values of  $p$ . Now, we can plug in the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3) and the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) to get

$$\begin{aligned} p(x) &= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \cdot \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} \, dp \\ &= \binom{n}{x} \cdot \frac{1}{B(\alpha, \beta)} \int_0^1 p^{\alpha+x-1} (1-p)^{\beta+n-x-1} \, dp \\ &= \binom{n}{x} \cdot \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \int_0^1 \frac{1}{B(\alpha + x, \beta + n - x)} p^{\alpha+x-1} (1-p)^{\beta+n-x-1} \, dp . \end{aligned} \quad (5)$$

Finally, we recognize that the integrand is equal to the probability density function of a beta distribution ( $\rightarrow$  II/3.9.5) and because probability density integrates to one ( $\rightarrow$  I/1.7.1), we have

$$p(x) = \binom{n}{x} \cdot \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} = f_X(x) . \quad (6)$$

This completes the proof. ■

#### Sources:

- Wikipedia (2022): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: [https://en.wikipedia.org/wiki/Beta-binomial\\_distribution#As\\_a\\_compound\\_distribution](https://en.wikipedia.org/wiki/Beta-binomial_distribution#As_a_compound_distribution).

### 1.3.3 Probability mass function in terms of gamma function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta-binomial distribution ( $\rightarrow$  II/1.3.1):

$$X \sim \text{BetBin}(n, \alpha, \beta) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  can be expressed as

$$f_X(x) = \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)} \quad (2)$$

where  $\Gamma(x)$  is the gamma function.

**Proof:** The probability mass function of the beta-binomial distribution ( $\rightarrow$  II/1.3.2) is given by

$$f_X(x) = \binom{n}{x} \cdot \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)} . \quad (3)$$

Note that the binomial coefficient can be expressed in terms of factorials

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} , \quad (4)$$

that factorials are related to the gamma function via  $n! = \Gamma(n+1)$

$$\frac{n!}{x!(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \quad (5)$$

and that the beta function is related to the gamma function via

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} . \quad (6)$$

Applying (4), (5) and (6) to (3), we get

$$f_X(x) = \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)} . \quad (7)$$

This completes the proof. ■

#### Sources:

- Wikipedia (2022): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: [https://en.wikipedia.org/wiki/Beta-binomial\\_distribution#As\\_a\\_compound\\_distribution](https://en.wikipedia.org/wiki/Beta-binomial_distribution#As_a_compound_distribution).

### 1.3.4 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta-binomial distribution ( $\rightarrow$  II/1.3.1):

$$X \sim \text{BetBin}(n, \alpha, \beta) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \frac{1}{B(\alpha, \beta)} \cdot \frac{\Gamma(n+1)}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^x \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(i+1) \cdot \Gamma(n-i+1)} \quad (2)$$

where  $B(x, y)$  is the beta function and  $\Gamma(x)$  is the gamma function.

**Proof:** The cumulative distribution function ( $\rightarrow$  I/1.8.1) is defined as

$$F_X(x) = \Pr(X \leq x) \quad (3)$$

which, for a discrete random variable ( $\rightarrow$  I/1.2.6), evaluates to

$$F_X(x) = \sum_{i=-\infty}^x f_X(i) . \quad (4)$$

With the probability mass function of the beta-binomial distribution ( $\rightarrow$  II/1.3.2), this becomes

$$F_X(x) = \sum_{i=0}^x \binom{n}{i} \cdot \frac{B(\alpha+i, \beta+n-i)}{B(\alpha, \beta)} . \quad (5)$$

Using the expression of binomial coefficients in terms of factorials

$$\binom{n}{k} = \frac{n!}{k! (n-k)!} , \quad (6)$$

the relationship between factorials and the gamma function

$$n! = \Gamma(n+1) \quad (7)$$

and the link between gamma function and beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha+\beta)} , \quad (8)$$

equation (5) can be further developped as follows:

$$\begin{aligned} F_X(x) &\stackrel{(6)}{=} \frac{1}{B(\alpha, \beta)} \cdot \sum_{i=0}^x \frac{n!}{i! (n-i)!} \cdot B(\alpha+i, \beta+n-i) \\ &\stackrel{(8)}{=} \frac{1}{B(\alpha, \beta)} \cdot \sum_{i=0}^x \frac{n!}{i! (n-i)!} \cdot \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(\alpha+\beta+n)} \\ &= \frac{1}{B(\alpha, \beta)} \cdot \frac{n!}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^x \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{i! (n-i)!} \\ &\stackrel{(7)}{=} \frac{1}{B(\alpha, \beta)} \cdot \frac{\Gamma(n+1)}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^x \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(i+1) \cdot \Gamma(n-i+1)} . \end{aligned} \quad (9)$$

This completes the proof. ■

## 1.4 Poisson distribution

### 1.4.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to follow a Poisson distribution with rate  $\lambda$

$$X \sim \text{Poiss}(\lambda) , \quad (1)$$

if and only if its probability mass function ( $\rightarrow$  I/1.6.1) is given by

$$\text{Poiss}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (2)$$

where  $x \in \mathbb{N}_0$  and  $\lambda > 0$ .

**Sources:**

- Wikipedia (2020): “Poisson distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-25; URL: [https://en.wikipedia.org/wiki/Poisson\\_distribution#Definitions](https://en.wikipedia.org/wiki/Poisson_distribution#Definitions).

### 1.4.2 Probability mass function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Poisson distribution ( $\rightarrow$  II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}_0 . \quad (2)$$

**Proof:** This follows directly from the definition of the Poisson distribution ( $\rightarrow$  II/1.4.1). ■

### 1.4.3 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Poisson distribution ( $\rightarrow$  II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$\mathbb{E}(X) = \lambda . \quad (2)$$

**Proof:** The expected value of a discrete random variable ( $\rightarrow$  I/1.12.1) is defined as

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) , \quad (3)$$

such that, with the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), we have:



$$\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{x}{x!} \lambda^x \\
&= \lambda e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} .
\end{aligned} \tag{4}$$

Substituting  $z = x - 1$ , such that  $x = z + 1$ , we get:

$$E(X) = \lambda e^{-\lambda} \cdot \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} . \tag{5}$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} , \tag{6}$$

the expected value of  $X$  finally becomes

$$\begin{aligned}
E(X) &= \lambda e^{-\lambda} \cdot e^{\lambda} \\
&= \lambda .
\end{aligned} \tag{7}$$

■

#### Sources:

- ProofWiki (2020): “Expectation of Poisson Distribution”; in: *ProofWiki*, retrieved on 2020-08-19; URL: [https://proofwiki.org/wiki/Expectation\\_of\\_Poisson\\_Distribution](https://proofwiki.org/wiki/Expectation_of_Poisson_Distribution).

#### 1.4.4 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Poisson distribution ( $\rightarrow$  II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \tag{1}$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \lambda . \tag{2}$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) can be expressed in terms of expected values ( $\rightarrow$  I/1.13.4) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \tag{3}$$

The expected value of a Poisson random variable ( $\rightarrow$  II/1.4.3) is

$$E(X) = \lambda . \tag{4}$$

Let us now consider the expectation ( $\rightarrow$  I/1.12.1) of  $X(X-1)$  which is defined as

$$E[X(X-1)] = \sum_{x \in \mathcal{X}} x(x-1) \cdot f_X(x), \quad (5)$$

such that, with the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), we have:

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=2}^{\infty} x(x-1) \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \cdot \sum_{x=2}^{\infty} x(x-1) \cdot \frac{\lambda^x}{x \cdot (x-1) \cdot (x-2)!} \\ &= \lambda^2 \cdot e^{-\lambda} \cdot \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}. \end{aligned} \quad (6)$$

Substituting  $z = x - 2$ , such that  $x = z + 2$ , we get:

$$E[X(X-1)] = \lambda^2 \cdot e^{-\lambda} \cdot \sum_{z=0}^{\infty} \frac{\lambda^z}{z!}. \quad (7)$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad (8)$$

the expected value of  $X(X-1)$  finally becomes

$$E[X(X-1)] = \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda^2. \quad (9)$$

Note that this expectation can be written as

$$E[X(X-1)] = E(X^2 - X) = E(X^2) - E(X), \quad (10)$$

such that, with (9) and (4), we have:

$$E(X^2) - E(X) = \lambda^2 \quad \Rightarrow \quad E(X^2) = \lambda^2 + \lambda. \quad (11)$$

Plugging (11) and (4) into (3), the variance of a Poisson random variable finally becomes

$$\text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda. \quad (12)$$

■

#### Sources:

- jbststatistics (2013): “The Poisson Distribution: Mathematically Deriving the Mean and Variance”; in: *YouTube*, retrieved on 2021-04-29; URL: [https://www.youtube.com/watch?v=65n\\_v92JZeE](https://www.youtube.com/watch?v=65n_v92JZeE).

## 1.4.5 Shannon entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Poisson distribution ( $\rightarrow$  II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \quad (1)$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  in nats is

$$H(X) = \lambda(1 - \ln \lambda) + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} . \quad (2)$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (3)$$

Entropy is measured in nats by setting  $b = e$ . Then, with the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), we have:

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log_e p(x) \\ &= - \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \cdot \ln \frac{\lambda^x e^{-\lambda}}{x!} \\ &= -e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} [\ln \lambda^x + \ln e^{-\lambda} - \ln x!] \\ &= -e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln \lambda^x}{x!} - e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln e^{-\lambda}}{x!} + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} \\ &= -e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x x \cdot \ln \lambda}{x!} - e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x (-\lambda) \cdot \ln e}{x!} + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} \\ &= -e^{-\lambda} \ln \lambda \sum_{x=0}^{\infty} \frac{\lambda^x \cdot x}{x!} + e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} . \end{aligned} \quad (4)$$

Using the relation  $x!/x = (x-1)!$ , the first term can be developed as follows

$$\begin{aligned}
e^{-\lambda} \ln \lambda \sum_{x=0}^{\infty} \frac{\lambda^x \cdot x}{x!} &= e^{-\lambda} \cdot \ln \lambda \cdot \sum_{x=0}^{\infty} \lambda \cdot \frac{\lambda^x}{\lambda} \cdot \frac{x}{x!} \\
&= e^{-\lambda} \cdot \lambda \cdot \ln \lambda \cdot \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda} \cdot \lambda \cdot \ln \lambda \cdot \left( \frac{\lambda^{-1}}{(-1)!} + \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \right) \\
&= e^{-\lambda} \cdot \lambda \cdot \ln \lambda \cdot \left( \frac{1}{\lambda} \cdot \frac{0}{0!} + \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \right) \\
&= e^{-\lambda} \cdot \lambda \cdot \ln \lambda \cdot \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} ,
\end{aligned} \tag{5}$$

such that equation (4) becomes

$$H(X) = -e^{-\lambda} \lambda \ln \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} + e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} . \tag{6}$$

Using the power series form of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} , \tag{7}$$

equation (6) finally becomes

$$\begin{aligned}
H(X) &= -e^{-\lambda} \lambda \ln \lambda e^{\lambda} + e^{-\lambda} \lambda e^{\lambda} + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} \\
&= -\lambda \ln \lambda + \lambda + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} \\
&= \lambda (1 - \ln \lambda) + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \ln x!}{x!} .
\end{aligned} \tag{8}$$

■

## 2 Multivariate discrete distributions

### 2.1 Categorical distribution

#### 2.1.1 Definition

**Definition:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to follow a categorical distribution with category probabilities  $p_1, \dots, p_k$

$$X \sim \text{Cat}([p_1, \dots, p_k]) , \quad (1)$$

if  $X = e_i$  with probability ( $\rightarrow$  I/1.3.1)  $p_i$  for all  $i = 1, \dots, k$ , where  $e_i$  is the  $i$ -th elementary row vector, i.e. a  $1 \times k$  vector of zeros with a one in  $i$ -th position.

**Sources:**

- Wikipedia (2020): “Categorical distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Categorical\\_distribution](https://en.wikipedia.org/wiki/Categorical_distribution).

#### 2.1.2 Probability mass function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a categorical distribution ( $\rightarrow$  II/2.1.1):

$$X \sim \text{Cat}([p_1, \dots, p_k]) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \begin{cases} p_1 , & \text{if } x = e_1 \\ \vdots & \vdots \\ p_k , & \text{if } x = e_k . \end{cases} \quad (2)$$

where  $e_1, \dots, e_k$  are the  $1 \times k$  elementary row vectors.

**Proof:** This follows directly from the definition of the categorical distribution ( $\rightarrow$  II/2.1.1). ■

#### 2.1.3 Mean

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a categorical distribution ( $\rightarrow$  II/2.1.1):

$$X \sim \text{Cat}([p_1, \dots, p_k]) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = [p_1, \dots, p_k] . \quad (2)$$

**Proof:** If we conceive the outcome of a categorical distribution ( $\rightarrow$  II/2.1.1) to be a  $1 \times k$  vector, then the elementary row vectors  $e_1 = [1, 0, \dots, 0]$ , ...,  $e_k = [0, \dots, 0, 1]$  are all the possible outcomes and they occur with probabilities  $\Pr(X = e_1) = p_1$ , ...,  $\Pr(X = e_k) = p_k$ . Consequently, the expected value ( $\rightarrow$  I/1.12.1) is

$$\begin{aligned}
E(X) &= \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x) \\
&= \sum_{i=1}^k e_i \cdot \Pr(X = e_i) \\
&= \sum_{i=1}^k e_i \cdot p_i \\
&= [p_1, \dots, p_k] .
\end{aligned} \tag{3}$$

■

### 2.1.4 Covariance

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a categorical distribution ( $\rightarrow$  II/2.1.1):

$$X \sim \text{Cat}(n, p) . \tag{1}$$

Then, the covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  is

$$\text{Cov}(X) = \text{diag}(p) - pp^T . \tag{2}$$

**Proof:** The categorical distribution ( $\rightarrow$  II/2.1.1) is a special case of the multinomial distribution ( $\rightarrow$  II/2.2.1) in which  $n = 1$ :

$$X \sim \text{Mult}(n, p) \quad \text{and} \quad n = 1 \quad \Rightarrow \quad X \sim \text{Cat}(p) . \tag{3}$$

The covariance matrix of the multinomial distribution ( $\rightarrow$  II/2.2.5) is

$$\text{Cov}(X) = n (\text{diag}(p) - pp^T) , \tag{4}$$

thus the covariance matrix of the categorical distribution is

$$\text{Cov}(X) = \text{diag}(p) - pp^T . \tag{5}$$

■

### 2.1.5 Shannon entropy

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a categorical distribution ( $\rightarrow$  II/2.1.1):

$$X \sim \text{Cat}(p) . \tag{1}$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  is

$$H(X) = - \sum_{i=1}^k p_i \cdot \log p_i . \tag{2}$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (3)$$

Since there are  $k$  possible values for a categorical random vector ( $\rightarrow$  II/2.1.1) with probabilities given by the entries ( $\rightarrow$  II/2.1.2) of the  $1 \times k$  vector  $p$ , we have:

$$\begin{aligned} H(X) &= -\Pr(X = e_1) \cdot \log \Pr(X = e_1) - \dots - \Pr(X = e_k) \cdot \log \Pr(X = e_k) \\ H(X) &= - \sum_{i=1}^k \Pr(X = e_i) \cdot \log \Pr(X = e_i) \\ H(X) &= - \sum_{i=1}^k p_i \cdot \log p_i . \end{aligned} \quad (4)$$

■

## 2.2 Multinomial distribution

### 2.2.1 Definition

**Definition:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to follow a multinomial distribution with number of trials  $n$  and category probabilities  $p_1, \dots, p_k$

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) , \quad (1)$$

if  $X$  are the numbers of observations belonging to  $k$  distinct categories in  $n$  independent ( $\rightarrow$  I/1.3.6) trials, where each trial has  $k$  possible outcomes ( $\rightarrow$  II/2.1.1) and the category probabilities are identical across trials.

#### Sources:

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Multinomial\\_distribution](https://en.wikipedia.org/wiki/Multinomial_distribution).

### 2.2.2 Probability mass function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, the probability mass function ( $\rightarrow$  I/1.6.1) of  $X$  is

$$f_X(x) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} . \quad (2)$$

**Proof:** A multinomial variable ( $\rightarrow$  II/2.2.1) is defined as a vector of the numbers of observations belonging to  $k$  distinct categories in  $n$  independent ( $\rightarrow$  I/1.3.6) trials, where each trial has  $k$  possible outcomes ( $\rightarrow$  II/2.1.1) and the category probabilities ( $\rightarrow$  I/1.3.1) are identical across trials.

Since the individual trials are independent ( $\rightarrow$  II/2.2.1) and joint probability factorizes under independence ( $\rightarrow$  I/1.3.9), the probability of a particular series of  $x_1$  observations for category 1,  $x_2$  observations for category 2, ... etc., when order does matter, is

$$\prod_{i=1}^k p_i^{x_i} . \quad (3)$$

When order does not matter, there is a number of series consisting of  $x_1$  observations for category 1,  $x_2$  observations for category 2, ... etc. This number is equal to the number of possibilities in which  $x_1$  category 1 objects,  $x_2$  category 2 objects, ... etc. can be distributed in a sequence of  $n$  objects which is given by the multinomial coefficient that can be expressed in terms of factorials:

$$\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! \cdot \dots \cdot x_k!} . \quad (4)$$

In order to obtain the probability of  $x_1$  observations for category 1,  $x_2$  observations for category 2, ... etc., when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x|n, [p_1, \dots, p_k]) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \quad (5)$$

which is equivalent to the expression above. ■

### 2.2.3 Cumulative distribution function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, 1) the joint cumulative distribution function ( $\rightarrow$  I/1.8.11) of  $X$  is

$$F_X(x) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \quad (2)$$

and 2) the marginal ( $\rightarrow$  I/1.5.3) cumulative distribution function ( $\rightarrow$  I/1.8.1) of each  $X_i$  is

$$F_{X_i}(x_i) = \sum_{k=0}^{x_i} \binom{n}{k} p_i^k (1 - p_i)^{n-k} . \quad (3)$$

**Proof:** The probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2) is

$$f_X(x) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} , \quad (4)$$

1) The joint cumulative distribution function (CDF) of a random vector ( $\rightarrow$  I/1.8.11) is defined as

$$F_X(x) = \Pr(X_1 \leq x_1, \dots, X_n \leq x_n) \quad (5)$$

which, in terms of the probability mass function, is given by ( $\rightarrow$  I/1.8.5)



$$F_X(x) = \sum_{t_1 \leq x_1} \dots \sum_{t_n \leq x_n} f_X(t_1, \dots, t_n) . \quad (6)$$

For a given set of counts  $x_1, \dots, x_k$  satisfying  $\sum_{i=1}^k x_i = n$ , we define the set of eligible values  $t_1, \dots, t_k$  for the CDF:

$$\mathcal{T}(x) = \left\{ (t_1, \dots, t_k) \mid t_i \in \mathbb{N}_0, 0 \leq t_i \leq x_i, \sum_{i=1}^k t_i = n \right\} . \quad (7)$$

Since any eligible set of counts  $t_1, \dots, t_k$  must add up to  $n$ , this set is equivalent to

$$\mathcal{T}(x) = \{(x_1, \dots, x_k)\} . \quad (8)$$

Therefore, the CDF is equal to the probability mass function at  $t = x$ :

$$F_X(x) \stackrel{(6)}{=} \sum_{t \in \mathcal{T}(x)} f_X(t_1, \dots, t_k) = f_X(x_1, \dots, x_k) . \quad (9)$$

so that the joint cumulative distribution function of the multinomial distribution becomes

$$F_X(x) \stackrel{(4)}{=} \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} . \quad (10)$$

2) The marginal distributions for the multinomial distribution ( $\rightarrow$  II/2.2.7) are

$$X_i \sim \text{Bin}(n, p_i) \quad \text{for all } i = 1, \dots, k . \quad (11)$$

Thus, the marginal cumulative distribution function of any entry  $X_i$  is equal to the cumulative distribution function of the binomial distribution ( $\rightarrow$  II/1.2.4):

$$F_{X_i}(x_i) = \sum_{k=0}^{x_i} \binom{n}{k} p_i^k (1 - p_i)^{n-k} . \quad (12)$$

■

## 2.2.4 Mean

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$\mathbb{E}(X) = [np_1, \dots, np_k] . \quad (2)$$

**Proof:** By definition, a multinomial random variable ( $\rightarrow$  II/2.2.1) is the sum of  $n$  independent and identical categorical trials ( $\rightarrow$  II/2.1.1) with category probabilities  $p_1, \dots, p_k$ . Therefore, the expected value is

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \dots + X_n) \quad (3)$$

and because the expected value is a linear operator ( $\rightarrow$  I/1.12.5), this is equal to

$$\begin{aligned} E(X) &= E(X_1) + \dots + E(X_n) \\ &= \sum_{i=1}^n E(X_i) . \end{aligned} \quad (4)$$

With the expected value of the categorical distribution ( $\rightarrow$  II/2.1.3), we have:

$$E(X) = \sum_{i=1}^n [p_1, \dots, p_k] = n \cdot [p_1, \dots, p_k] = [np_1, \dots, np_k] . \quad (5)$$

■

### 2.2.5 Covariance

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$[X_1, \dots, X_k] = X \sim \text{Mult}(n, p), \quad n \in \mathbb{N}, \quad p = [p_1, \dots, p_k]^T . \quad (1)$$

Then, the covariance matrix ( $\rightarrow$  I/1.16.1) of  $X$  is

$$\text{Cov}(X) = n (\text{diag}(p) - pp^T) . \quad (2)$$

**Proof:** We first observe that the sample space ( $\rightarrow$  I/1.1.2) of each coordinate  $X_i$  is  $\{0, 1, \dots, n\}$  and  $X_i$  is the sum of independent draws of category  $i$ , which is drawn with probability  $p_i$ . Thus each coordinate follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X_i \stackrel{\text{i.i.d.}}{\sim} \text{Bin}(n, p_i), \quad i = 1, \dots, k , \quad (3)$$

which has the variance ( $\rightarrow$  II/1.2.7)  $\text{Var}(X_i) = np_i(1 - p_i) = n(p_i - p_i^2)$ , constituting the elements of the main diagonal in  $\text{Cov}(X)$  in (2). To prove  $\text{Cov}(X_i, X_j) = -np_i p_j$  for  $i \neq j$  (which constitutes the off-diagonal elements of the covariance matrix), we first recognize that

$$X_i = \sum_{k=1}^n \mathbb{I}_i(k), \quad \text{with} \quad \mathbb{I}_i(k) = \begin{cases} 1 & \text{if } k\text{-th draw was of category } i, \\ 0 & \text{otherwise} , \end{cases} \quad (4)$$

where the indicator function  $\mathbb{I}_i$  is a Bernoulli-distributed ( $\rightarrow$  II/1.1.1) random variable with the expected value ( $\rightarrow$  II/1.1.4)  $p_i$ . Then, we have

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \text{Cov} \left( \sum_{k=1}^n \mathbb{I}_i(k), \sum_{l=1}^n \mathbb{I}_j(l) \right) \\
&= \sum_{k=1}^n \sum_{l=1}^n \text{Cov}(\mathbb{I}_i(k), \mathbb{I}_j(l)) \\
&= \sum_{k=1}^n \left[ \text{Cov}(\mathbb{I}_i(k), \mathbb{I}_j(k)) + \underbrace{\sum_{\substack{l=1 \\ l \neq k}}^n \text{Cov}(\mathbb{I}_i(k), \mathbb{I}_j(l))}_{=0} \right] \\
&\stackrel{i \neq j}{=} \sum_{k=1}^n \left( \text{E} \left( \underbrace{\mathbb{I}_i(k) \mathbb{I}_j(k)}_{=0} \right) - \text{E}(\mathbb{I}_i(k)) \text{E}(\mathbb{I}_j(k)) \right) \\
&= - \sum_{k=1}^n \text{E}(\mathbb{I}_i(k)) \text{E}(\mathbb{I}_j(k)) \\
&= -np_i p_j,
\end{aligned} \tag{5}$$

as desired. ■

#### Sources:

- Tutz G (2012): “Multinomial Response Models”; in: *Regression for Categorical Data*, pp. 209ff.;  
URL: <https://www.cambridge.org/core/books/regression-for-categorical-data/B71F71F2A484E2DF88256>  
DOI: 10.1017/CBO9780511842061.

### 2.2.6 Shannon entropy

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$X \sim \text{Mult}(n, p). \tag{1}$$

Then, the (Shannon) entropy ( $\rightarrow$  I/2.1.1) of  $X$  is

$$H(X) = n \cdot H_{\text{cat}}(p) - E_{\text{lmc}}(n, p) \tag{2}$$

where  $H_{\text{cat}}(p)$  is the categorical entropy function, i.e. the (Shannon) entropy of the categorical distribution ( $\rightarrow$  II/2.1.5) with category probabilities  $p$

$$H_{\text{cat}}(p) = - \sum_{i=1}^k p_i \cdot \log p_i \tag{3}$$

and  $E_{\text{lmc}}(n, p)$  is the expected value ( $\rightarrow$  I/1.12.1) of the logarithmized multinomial coefficient ( $\rightarrow$  II/2.2.2) with superset size  $n$

$$E_{\text{lmc}}(n, p) = \text{E} \left[ \log \binom{n}{X_1, \dots, X_k} \right] \quad \text{where } X \sim \text{Mult}(n, p). \tag{4}$$

**Proof:** The entropy ( $\rightarrow$  I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) . \quad (5)$$

The probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2) is

$$f_X(x) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \quad (6)$$

Let  $\mathcal{X}_{n,k}$  be the set of all vectors  $x \in \mathbb{N}^{1 \times k}$  satisfying  $\sum_{i=1}^k x_i = n$ . Then, we have:

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \log f_X(x) \\ &= - \sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \log \left[ \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \right] \\ &= - \sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \left[ \log \binom{n}{x_1, \dots, x_k} + \sum_{i=1}^k x_i \cdot \log p_i \right] . \end{aligned} \quad (7)$$

Since the first factor in the sum corresponds to the probability mass ( $\rightarrow$  I/1.6.1) of  $X = x$ , we can rewrite this as the sum of the expected values ( $\rightarrow$  I/1.12.1) of the functions ( $\rightarrow$  I/1.12.9) of the discrete random variable ( $\rightarrow$  I/1.2.6)  $x$  in the square bracket:

$$\begin{aligned} H(X) &= - \left\langle \log \binom{n}{x_1, \dots, x_k} \right\rangle_{p(x)} - \left\langle \sum_{i=1}^k x_i \cdot \log p_i \right\rangle_{p(x)} \\ &= - \left\langle \log \binom{n}{x_1, \dots, x_k} \right\rangle_{p(x)} - \sum_{i=1}^k \langle x_i \cdot \log p_i \rangle_{p(x)} . \end{aligned} \quad (8)$$

Using the expected value of the multinomial distribution ( $\rightarrow$  II/2.2.4), i.e.  $X \sim \text{Mult}(n, p) \Rightarrow \langle x_i \rangle = np_i$ , this gives:

$$\begin{aligned} H(X) &= - \left\langle \log \binom{n}{x_1, \dots, x_k} \right\rangle_{p(x)} - \sum_{i=1}^k np_i \cdot \log p_i \\ &= - \left\langle \log \binom{n}{x_1, \dots, x_k} \right\rangle_{p(x)} - n \sum_{i=1}^k p_i \cdot \log p_i . \end{aligned} \quad (9)$$

Finally, we note that the first term is the negative expected value ( $\rightarrow$  I/1.12.1) of the logarithm of a multinomial coefficient ( $\rightarrow$  II/2.2.2) and that the second term is the entropy of the categorical distribution ( $\rightarrow$  II/2.1.5), such that we finally get:

$$H(X) = n \cdot H_{\text{cat}}(p) - E_{\text{lmc}}(n, p) . \quad (10)$$

■

### 2.2.7 Marginal distributions

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of any entry  $X_i$  is a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X_i \sim \text{Bin}(n, p_i) \quad \text{for all } i = 1, \dots, k . \quad (2)$$

**Proof:** The entries of a multinomial ( $\rightarrow$  II/2.2.1) random vector  $X$  are the numbers of observations belonging to  $k$  distinct categories in  $n$  independent ( $\rightarrow$  I/1.3.6) trials where  $p_1, \dots, p_k$  are the category probabilities ( $\rightarrow$  II/2.2.1) identical across trials.

Let us define for each category  $i = 1, \dots, k$  and each trial  $j = 1, \dots, n$ :

$$Y_{ij} = \begin{cases} 0 , & \text{if } X_i = 0 \text{ in trial } j \\ 1 , & \text{if } X_i = 1 \text{ in trial } j . \end{cases} \quad (3)$$

Since  $\Pr(Y_{ij} = 1) = p_i$  and  $\Pr(Y_{ij} = 0) = 1 - p_i$ ,  $Y_{ij}$  is a Bernoulli ( $\rightarrow$  II/1.1.1) random variable with success probability ( $\rightarrow$  II/1.1.1)  $p_i$ :

$$Y_{ij} \sim \text{Bern}(p_i) . \quad (4)$$

Moreover, each  $X_i$  is the sum of the corresponding  $Y_{ij}$  over all trials

$$X_i = \sum_{j=1}^n Y_{ij} , \quad (5)$$

i.e.  $X_i$  is a sum of independent Bernoulli random variables which, by definition, follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$X_i \sim \text{Bin}(n, p_i) . \quad (6)$$

■

### 3 Univariate continuous distributions

#### 3.1 Continuous uniform distribution

##### 3.1.1 Definition

**Definition:** Let  $X$  be a continuous random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be uniformly distributed with minimum  $a$  and maximum  $b$

$$X \sim \mathcal{U}(a, b) , \quad (1)$$

if and only if each value between and including  $a$  and  $b$  occurs with the same probability.

**Sources:**

- Wikipedia (2020): “Uniform distribution (continuous)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: [https://en.wikipedia.org/wiki/Uniform\\_distribution\\_\(continuous\)](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)).

##### 3.1.2 Standard uniform distribution

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be standard uniformly distributed, if  $X$  follows a continuous uniform distribution ( $\rightarrow$  II/3.1.1) with minimum  $a = 0$  and maximum  $b = 1$ :

$$X \sim \mathcal{U}(0, 1) . \quad (1)$$

**Sources:**

- Wikipedia (2021): “Continuous uniform distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: [https://en.wikipedia.org/wiki/Continuous\\_uniform\\_distribution#Standard\\_uniform](https://en.wikipedia.org/wiki/Continuous_uniform_distribution#Standard_uniform).

##### 3.1.3 Probability density function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \begin{cases} \frac{1}{b-a} , & \text{if } a \leq x \leq b \\ 0 , & \text{otherwise .} \end{cases} \quad (2)$$

**Proof:** A continuous uniform variable is defined as ( $\rightarrow$  II/3.1.1) having a constant probability density between minimum  $a$  and maximum  $b$ . Therefore,

$$\begin{aligned} f_X(x) &\propto 1 \quad \text{for all } x \in [a, b] \quad \text{and} \\ f_X(x) &= 0, \quad \text{if } x < a \quad \text{or } x > b . \end{aligned} \quad (3)$$

To ensure that  $f_X(x)$  is a proper probability density function ( $\rightarrow$  I/1.7.1), the integral over all non-zero probabilities has to sum to 1. Therefore,

$$f_X(x) = \frac{1}{c(a, b)} \quad \text{for all } x \in [a, b] \quad (4)$$

where the normalization factor  $c(a, b)$  is specified, such that

$$\frac{1}{c(a, b)} \int_a^b 1 \, dx = 1 . \quad (5)$$

Solving this for  $c(a, b)$ , we obtain:

$$\begin{aligned} \int_a^b 1 \, dx &= c(a, b) \\ [x]_a^b &= c(a, b) \\ c(a, b) &= b - a . \end{aligned} \quad (6)$$

■

### 3.1.4 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < a \\ \frac{x-a}{b-a} , & \text{if } a \leq x \leq b \\ 1 , & \text{if } x > b . \end{cases} \quad (2)$$

**Proof:** The probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3) is:

$$\mathcal{U}(x; a, b) = \begin{cases} \frac{1}{b-a} , & \text{if } a \leq x \leq b \\ 0 , & \text{otherwise .} \end{cases} \quad (3)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^x \mathcal{U}(z; a, b) \, dz \quad (4)$$

First of all, if  $x < a$ , we have

$$F_X(x) = \int_{-\infty}^x 0 \, dz = 0 . \quad (5)$$

Moreover, if  $a \leq x \leq b$ , we have using (3)

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^a \mathcal{U}(z; a, b) \, dz + \int_a^x \mathcal{U}(z; a, b) \, dz \\
&= \int_{-\infty}^a 0 \, dz + \int_a^x \frac{1}{b-a} \, dz \\
&= 0 + \frac{1}{b-a} [z]_a^x \\
&= \frac{x-a}{b-a} .
\end{aligned} \tag{6}$$

Finally, if  $x > b$ , we have

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^b \mathcal{U}(z; a, b) \, dz + \int_b^x \mathcal{U}(z; a, b) \, dz \\
&= F_X(b) + \int_b^x 0 \, dz \\
&= \frac{b-a}{b-a} + 0 \\
&= 1 .
\end{aligned} \tag{7}$$

This completes the proof. ■

### 3.1.5 Quantile function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \tag{1}$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \begin{cases} -\infty , & \text{if } p = 0 \\ bp + a(1-p) , & \text{if } p > 0 . \end{cases} \tag{2}$$

**Proof:** The cumulative distribution function of the continuous uniform distribution ( $\rightarrow$  II/3.1.4) is:

$$F_X(x) = \begin{cases} 0 , & \text{if } x < a \\ \frac{x-a}{b-a} , & \text{if } a \leq x \leq b \\ 1 , & \text{if } x > b . \end{cases} \tag{3}$$

The quantile function ( $\rightarrow$  I/1.11.1)  $Q_X(p)$  is defined as the smallest  $x$ , such that  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \tag{4}$$

Thus, we have  $Q_X(p) = -\infty$ , if  $p = 0$ . When  $p > 0$ , it holds that ( $\rightarrow$  I/1.11.2)



$$Q_X(p) = F_X^{-1}(x) . \quad (5)$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{x - a}{b - a} \\ x &= p(b - a) + a \\ x &= bp + a(1 - p) . \end{aligned} \quad (6)$$

■

### 3.1.6 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \frac{1}{2}(a + b) . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx . \quad (3)$$

With the probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3), this becomes:

$$\begin{aligned} E(X) &= \int_a^b x \cdot \frac{1}{b - a} \, dx \\ &= \left[ \frac{1}{2} \frac{x^2}{b - a} \right]_a^b \\ &= \frac{1}{2} \frac{b^2 - a^2}{b - a} \\ &= \frac{1}{2} \frac{(b + a)(b - a)}{b - a} \\ &= \frac{1}{2}(a + b) . \end{aligned} \quad (4)$$

■

### 3.1.7 Median

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = \frac{1}{2}(a + b) . \quad (2)$$

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function ( $\rightarrow$  I/1.8.1) is  $1/2$ :

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the continuous uniform distribution ( $\rightarrow$  II/3.1.4) is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < a \\ \frac{x-a}{b-a} , & \text{if } a \leq x \leq b \\ 1 , & \text{if } x > b . \end{cases} \quad (4)$$

Thus, the inverse CDF ( $\rightarrow$  II/3.1.5) is

$$x = bp + a(1 - p) . \quad (5)$$

Setting  $p = 1/2$ , we obtain:

$$\text{median}(X) = b \cdot \frac{1}{2} + a \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{2}(a + b) . \quad (6)$$

■

### 3.1.8 Mode

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) \in [a, b] . \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3) is:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Since the PDF attains its only non-zero value whenever  $a \leq x \leq b$ ,

$$\max_x f_X(x) = \frac{1}{b-a}, \quad (5)$$

any value in the interval  $[a, b]$  may be considered the mode of  $X$ .

■

### 3.1.9 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b). \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \frac{1}{12}(b-a)^2. \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is the probability-weighted average of the squared deviation from the mean ( $\rightarrow$  I/1.12.1):

$$\text{Var}(X) = \int_{\mathbb{R}} (x - E(X))^2 \cdot f_X(x) \, dx. \quad (3)$$

With the expected value ( $\rightarrow$  II/3.1.6) and probability density function ( $\rightarrow$  II/3.1.3) of the continuous uniform distribution, this reads:

$$\begin{aligned}
\text{Var}(X) &= \int_a^b \left( x - \frac{1}{2}(a+b) \right)^2 \cdot \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \cdot \int_a^b \left( x - \frac{a+b}{2} \right)^2 dx \\
&= \frac{1}{b-a} \cdot \left[ \frac{1}{3} \left( x - \frac{a+b}{2} \right)^3 \right]_a^b \\
&= \frac{1}{3(b-a)} \cdot \left[ \left( \frac{2x - (a+b)}{2} \right)^3 \right]_a^b \\
&= \frac{1}{3(b-a)} \cdot \left[ \frac{1}{8} (2x - a - b)^3 \right]_a^b \\
&= \frac{1}{24(b-a)} \cdot [(2x - a - b)^3]_a^b \\
&= \frac{1}{24(b-a)} \cdot [(2b - a - b)^3 - (2a - a - b)^3] \\
&= \frac{1}{24(b-a)} \cdot [(b-a)^3 - (a-b)^3] \\
&= \frac{1}{24(b-a)} \cdot [(b-a)^3 + (-1)^3(a-b)^3] \\
&= \frac{1}{24(b-a)} \cdot [(b-a)^3 + (b-a)^3] \\
&= \frac{2(b-a)^3}{24(b-a)} \\
&= \frac{1}{12} (b-a)^2 .
\end{aligned} \tag{4}$$

■

### 3.1.10 Differential entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a continuous uniform distribution ( $\rightarrow$  II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \tag{1}$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = \ln(b-a) . \tag{2}$$

**Proof:** The differential entropy ( $\rightarrow$  I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \tag{3}$$

To measure  $h(X)$  in nats, we set  $b = e$ , such that

$$h(X) = - \int_{\mathcal{X}} p(x) \ln p(x) dx . \quad (4)$$

With the probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3), the differential entropy of  $X$  is:

$$\begin{aligned} h(X) &= - \int_a^b \frac{1}{b-a} \ln \left( \frac{1}{b-a} \right) dx \\ &= \frac{1}{b-a} \cdot \int_a^b \ln(b-a) dx \\ &= \frac{1}{b-a} \cdot [x \cdot \ln(b-a)]_a^b \\ &= \frac{1}{b-a} \cdot [b \cdot \ln(b-a) - a \cdot \ln(b-a)] \\ &= \frac{1}{b-a} (b-a) \ln(b-a) \\ &= \ln(b-a) . \end{aligned} \quad (5)$$

■

### 3.1.11 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two continuous uniform distributions ( $\rightarrow$  II/3.1.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \mathcal{U}(a_1, b_1) \\ Q : X &\sim \mathcal{U}(a_2, b_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P || Q] = \ln \frac{b_2 - a_2}{b_1 - a_1} . \quad (2)$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P || Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx . \quad (3)$$

This means that the KL divergence of  $P$  from  $Q$  is only defined, if for all  $x \in \mathcal{X}$ ,  $q(x) = 0$  implies  $p(x) = 0$ . Thus,  $\text{KL}[P || Q]$  only exists, if  $a_2 \leq a_1$  and  $b_1 \leq b_2$ , i.e. if  $P$  only places non-zero probability where  $Q$  also places non-zero probability, such that  $q(x)$  is not zero for any  $x \in \mathcal{X}$  where  $p(x)$  is positive.

If this requirement is fulfilled, we can write

$$\text{KL}[P || Q] = \int_{-\infty}^{a_1} p(x) \ln \frac{p(x)}{q(x)} dx + \int_{a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} dx + \int_{b_1}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx \quad (4)$$

and because  $p(x) = 0$  for any  $x < a_1$  and any  $x > b_1$ , we have

$$\text{KL}[P || Q] = \int_{-\infty}^{a_1} 0 \cdot \ln \frac{0}{q(x)} dx + \int_{a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} dx + \int_{b_1}^{+\infty} 0 \cdot \ln \frac{0}{q(x)} dx . \quad (5)$$

Now,  $(0 \cdot \ln 0)$  is taken to be zero by convention ( $\rightarrow$  I/2.1.1), such that

$$\text{KL}[P || Q] = \int_{a_1}^{b_1} p(x) \ln \frac{p(x)}{q(x)} dx \quad (6)$$

and we can use the probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3) to evaluate:

$$\begin{aligned} \text{KL}[P || Q] &= \int_{a_1}^{b_1} \frac{1}{b_1 - a_1} \ln \frac{\frac{1}{b_1 - a_1}}{\frac{1}{b_2 - a_2}} dx \\ &= \frac{1}{b_1 - a_1} \ln \frac{b_2 - a_2}{b_1 - a_1} \int_{a_1}^{b_1} dx \\ &= \frac{1}{b_1 - a_1} \ln \frac{b_2 - a_2}{b_1 - a_1} [x]_{a_1}^{b_1} \\ &= \frac{1}{b_1 - a_1} \ln \frac{b_2 - a_2}{b_1 - a_1} (b_1 - a_1) \\ &= \ln \frac{b_2 - a_2}{b_1 - a_1} . \end{aligned} \quad (7)$$

■

### 3.1.12 Maximum entropy distribution

**Theorem:** The continuous uniform distribution ( $\rightarrow$  II/3.1.1) maximizes differential entropy ( $\rightarrow$  I/2.2.1) for a random variable ( $\rightarrow$  I/1.2.2) with a fixed range.

**Proof:** Without loss of generality, let us assume that the random variable  $X$  is in the following range:  $a \leq X \leq b$ .

Let  $g(x)$  be the probability density function ( $\rightarrow$  I/1.7.1) of a continuous uniform distribution ( $\rightarrow$  II/3.1.1) with minimum  $a$  and maximum  $b$  and let  $f(x)$  be an arbitrary probability density function ( $\rightarrow$  I/1.7.1) defined on the same support  $\mathcal{X} = [a, b]$ .

For a random variable ( $\rightarrow$  I/1.2.2)  $X$  with set of possible values  $\mathcal{X}$  and probability density function ( $\rightarrow$  I/1.7.1)  $p(x)$ , the differential entropy ( $\rightarrow$  I/2.2.1) is defined as:

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx \quad (1)$$

Consider the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of distribution  $f(x)$  from distribution  $g(x)$  which is non-negative ( $\rightarrow$  I/2.5.2):

$$\begin{aligned} 0 \leq \text{KL}[f||g] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx \\ &= \int_{\mathcal{X}} f(x) \log f(x) dx - \int_{\mathcal{X}} f(x) \log g(x) dx \\ &\stackrel{(1)}{=} -h[f(x)] - \int_{\mathcal{X}} f(x) \log g(x) dx . \end{aligned} \quad (2)$$

By plugging the probability density function of the continuous uniform distribution ( $\rightarrow$  II/3.1.3) into the second term, we obtain:

$$\begin{aligned} \int_{\mathcal{X}} f(x) \log g(x) \, dx &= \int_{\mathcal{X}} f(x) \log \frac{1}{b-a} \, dx \\ &= \log \frac{1}{b-a} \int_{\mathcal{X}} f(x) \, dx \\ &= -\log(b-a) . \end{aligned} \quad (3)$$

This is actually the negative of the differential entropy of the continuous uniform distribution ( $\rightarrow$  II/3.1.10), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) \, dx = -h[\mathcal{U}(a, b)] = -h[g(x)] . \quad (4)$$

Combining (2) with (4), we can show that

$$\begin{aligned} 0 &\leq \text{KL}[f||g] \\ 0 &\leq -h[f(x)] - (-h[g(x)]) \\ h[g(x)] &\geq h[f(x)] \end{aligned} \quad (5)$$

which means that the differential entropy ( $\rightarrow$  I/2.2.1) of the continuous uniform distribution ( $\rightarrow$  II/3.1.1)  $\mathcal{U}(a, b)$  will be larger than or equal to any other distribution ( $\rightarrow$  I/1.5.1) defined in the same range. ■

#### Sources:

- Wikipedia (2023): “Maximum entropy probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-08-25; URL: [https://en.wikipedia.org/wiki/Maximum\\_entropy\\_probability\\_distribution#Uniform\\_and\\_piecewise\\_uniform\\_distributions](https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution#Uniform_and_piecewise_uniform_distributions).

## 3.2 Normal distribution

### 3.2.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be normally distributed with mean  $\mu$  and variance  $\sigma^2$  (or, standard deviation  $\sigma$ )

$$X \sim \mathcal{N}(\mu, \sigma^2) , \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (2)$$

where  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ .

#### Sources:

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution](https://en.wikipedia.org/wiki/Normal_distribution).

### 3.2.2 Special case of multivariate normal distribution

**Theorem:** The normal distribution ( $\rightarrow$  II/3.2.1) is a special case of the multivariate normal distribution ( $\rightarrow$  II/4.1.1) with number of variables  $n = 1$ , i.e. random vector ( $\rightarrow$  I/1.2.3)  $x \in \mathbb{R}$ , mean  $\mu \in \mathbb{R}$  and covariance matrix  $\Sigma = \sigma^2$ .

**Proof:** The probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) is

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] . \quad (1)$$

Setting  $n = 1$ , such that  $x, \mu \in \mathbb{R}$ , and  $\Sigma = \sigma^2$ , we obtain

$$\begin{aligned} \mathcal{N}(x; \mu, \sigma^2) &= \frac{1}{\sqrt{(2\pi)^1 |\sigma^2|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T (\sigma^2)^{-1} (x - \mu) \right] \\ &= \frac{1}{\sqrt{(2\pi) \sigma^2}} \cdot \exp \left[ -\frac{1}{2 \sigma^2} (x - \mu)^2 \right] \\ &= \frac{1}{\sqrt{2\pi} \sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \end{aligned} \quad (2)$$

which is equivalent to the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10). ■

#### Sources:

- Wikipedia (2022): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-08-19; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution](https://en.wikipedia.org/wiki/Multivariate_normal_distribution).

### 3.2.3 Standard normal distribution

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be standard normally distributed, if  $X$  follows a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu = 0$  and variance  $\sigma^2 = 1$ :

$$X \sim \mathcal{N}(0, 1) . \quad (1)$$

#### Sources:

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-26; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Standard\\_normal\\_distribution](https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution).

### 3.2.4 Relationship to standard normal distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ :

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantity  $Z = (X - \mu)/\sigma$  will have a standard normal distribution ( $\rightarrow$  II/3.2.3) with mean 0 and variance 1:



$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) . \quad (2)$$

**Proof:** Note that  $Z$  is a function of  $X$

$$Z = g(X) = \frac{X - \mu}{\sigma} \quad (3)$$

with the inverse function

$$X = g^{-1}(Z) = \sigma Z + \mu . \quad (4)$$

Because  $\sigma$  is positive,  $g(X)$  is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function ( $\rightarrow$  I/1.8.3) as

$$F_Y(y) = \begin{cases} 0 , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 1 , & \text{if } y > \max(\mathcal{Y}) . \end{cases} \quad (5)$$

The cumulative distribution function of the normally distributed ( $\rightarrow$  II/3.2.12)  $X$  is

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{t - \mu}{\sigma} \right)^2 \right] dt . \quad (6)$$

Applying (5) to (6), we have:

$$\begin{aligned} F_Z(z) &\stackrel{(5)}{=} F_X(g^{-1}(z)) \\ &\stackrel{(6)}{=} \int_{-\infty}^{\sigma z + \mu} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{t - \mu}{\sigma} \right)^2 \right] dt . \end{aligned} \quad (7)$$

Substituting  $s = (t - \mu)/\sigma$ , such that  $t = \sigma s + \mu$ , we obtain

$$\begin{aligned} F_Z(z) &= \int_{(-\infty - \mu)/\sigma}^{(\sigma z + \mu - \mu)/\sigma} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{(\sigma s + \mu) - \mu}{\sigma} \right)^2 \right] d(\sigma s + \mu) \\ &= \int_{-\infty}^z \frac{\sigma}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} s^2 \right] ds \\ &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \cdot \exp \left[ -\frac{1}{2} s^2 \right] ds \end{aligned} \quad (8)$$

which is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3).

■

### 3.2.5 Relationship to standard normal distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ :

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantity  $Z = (X - \mu)/\sigma$  will have a standard normal distribution ( $\rightarrow$  II/3.2.3) with mean 0 and variance 1:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) . \quad (2)$$

**Proof:** Note that  $Z$  is a function of  $X$

$$Z = g(X) = \frac{X - \mu}{\sigma} \quad (3)$$

with the inverse function

$$X = g^{-1}(Z) = \sigma Z + \mu . \quad (4)$$

Because  $\sigma$  is positive,  $g(X)$  is strictly increasing and we can calculate the probability density function of a strictly increasing function ( $\rightarrow$  I/1.7.3) as

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (5)$$

where  $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$ . With the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), we have

$$\begin{aligned} f_Z(z) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{g^{-1}(z) - \mu}{\sigma} \right)^2 \right] \cdot \frac{dg^{-1}(z)}{dz} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{(\sigma z + \mu) - \mu}{\sigma} \right)^2 \right] \cdot \frac{d(\sigma z + \mu)}{dz} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} z^2 \right] \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp \left[ -\frac{1}{2} z^2 \right] \end{aligned} \quad (6)$$

which is the probability density function ( $\rightarrow$  I/1.7.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3). ■

### 3.2.6 Relationship to standard normal distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ :

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantity  $Z = (X - \mu)/\sigma$  will have a standard normal distribution ( $\rightarrow$  II/3.2.3) with mean 0 and variance 1:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) . \quad (2)$$

**Proof:** The linear transformation theorem for multivariate normal distribution ( $\rightarrow$  II/4.1.14) states

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \quad (3)$$

where  $x$  is an  $n \times 1$  random vector ( $\rightarrow$  I/1.2.3) following a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with mean  $\mu$  and covariance  $\Sigma$ ,  $A$  is an  $m \times n$  matrix and  $b$  is an  $m \times 1$  vector. Note that

$$Z = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma} \quad (4)$$

is a special case of (3) with  $x = X$ ,  $\mu = \mu$ ,  $\Sigma = \sigma^2$ ,  $A = 1/\sigma$  and  $b = \mu/\sigma$ . Applying theorem (3) to  $Z$  as a function of  $X$ , we have

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X}{\sigma} - \frac{\mu}{\sigma} \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{1}{\sigma} \cdot \sigma^2 \cdot \frac{1}{\sigma}\right) \quad (5)$$

which results in the distribution:

$$Z \sim \mathcal{N}(0, 1) . \quad (6)$$

■

### 3.2.7 Relationship to chi-squared distribution

**Theorem:** Let  $X_1, \dots, X_n$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) where each of them is following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ :

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for } i = 1, \dots, n . \quad (1)$$

Define the sample mean ( $\rightarrow$  I/1.12.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

and the unbiased sample variance ( $\rightarrow$  I/1.13.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 . \quad (3)$$

Then, the sampling distribution ( $\rightarrow$  I/1.5.6) of the sample variance is given by a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $n - 1$  degrees of freedom:

$$V = (n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1) . \quad (4)$$

**Proof:** Consider the random variable ( $\rightarrow$  I/1.2.2)  $U_i$  defined as

$$U_i = \frac{X_i - \mu}{\sigma} \quad (5)$$

which follows a standard normal distribution ( $\rightarrow$  II/3.2.4)

$$U_i \sim \mathcal{N}(0, 1) . \quad (6)$$

Then, the sum of squared random variables  $U_i$  can be rewritten as

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left( \frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \sum_{i=1}^n \frac{(\bar{X} - \mu)^2}{\sigma^2} + 2 \sum_{i=1}^n \frac{(X_i - \bar{X})(\bar{X} - \mu)}{\sigma^2} \\ &= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 + 2 \frac{(\bar{X} - \mu)}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}) . \end{aligned} \quad (7)$$

Because the following sum is zero

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X}) &= \sum_{i=1}^n X_i - n\bar{X} \\ &= \sum_{i=1}^n X_i - n \cdot \frac{1}{n} \sum_{i=1}^n X_i \\ &= \sum_{i=1}^n X_i - \sum_{i=1}^n X_i \\ &= 0 , \end{aligned} \quad (8)$$

the third term disappears, i.e.

$$\sum_{i=1}^n U_i^2 = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 . \quad (9)$$

Cochran's theorem states that, if a sum of squared standard normal ( $\rightarrow$  II/3.2.3) random variables ( $\rightarrow$  I/1.2.2) can be written as a sum of squared forms

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{j=1}^m Q_j \quad \text{where} \quad Q_j = \sum_{k=1}^n \sum_{l=1}^n U_k B_{kl}^{(j)} U_l \\ &\quad \text{with} \quad \sum_{j=1}^m B^{(j)} = I_n \\ &\quad \text{and} \quad r_j = \text{rank}(B^{(j)}) , \end{aligned} \quad (10)$$

then the terms  $Q_j$  are independent ( $\rightarrow$  I/1.3.6) and each term  $Q_j$  follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $r_j$  degrees of freedom:

$$Q_j \sim \chi^2(r_j) . \quad (11)$$

We observe that (9) can be represented as

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &= Q_1 + Q_2 = \sum_{i=1}^n \left( U_i - \frac{1}{n} \sum_{j=1}^n U_j \right)^2 + \frac{1}{n} \left( \sum_{i=1}^n U_i \right)^2 \end{aligned} \quad (12)$$

where, with the  $n \times n$  matrix of ones  $J_n$ , the matrices  $B^{(j)}$  are

$$B^{(1)} = I_n - \frac{J_n}{n} \quad \text{and} \quad B^{(2)} = \frac{J_n}{n} . \quad (13)$$

Because all columns of  $B^{(2)}$  are identical, it has rank  $r_2 = 1$ . Because the  $n$  columns of  $B^{(1)}$  add up to zero, it has rank  $r_1 = n - 1$ . Thus, the conditions of Cochran's theorem are met and the squared form

$$Q_1 = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 = (n-1) \frac{1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = (n-1) \frac{s^2}{\sigma^2} \quad (14)$$

follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $n - 1$  degrees of freedom:

$$(n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1) . \quad (15)$$

■

#### Sources:

- Glen-b (2014): "Why is the sampling distribution of variance a chi-squared distribution?"; in: *StackExchange CrossValidated*, retrieved on 2021-05-20; URL: <https://stats.stackexchange.com/questions/121662/why-is-the-sampling-distribution-of-variance-a-chi-squared-distribution>.
- Wikipedia (2021): "Cochran's theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: [https://en.wikipedia.org/wiki/Cochran%27s\\_theorem#Sample\\_mean\\_and\\_sample\\_variance](https://en.wikipedia.org/wiki/Cochran%27s_theorem#Sample_mean_and_sample_variance).

#### 3.2.8 Relationship to t-distribution

**Theorem:** Let  $X_1, \dots, X_n$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) where each of them is following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ :

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for } i = 1, \dots, n . \quad (1)$$

Define the sample mean ( $\rightarrow$  I/1.12.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

and the unbiased sample variance ( $\rightarrow$  I/1.13.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 . \quad (3)$$

Then, subtracting  $\mu$  from the sample mean ( $\rightarrow$  I/1.12.1), dividing by the sample standard deviation ( $\rightarrow$  I/1.19.1) and multiplying with  $\sqrt{n}$  results in a quantity that follows a t-distribution ( $\rightarrow$  II/3.3.1) with  $n - 1$  degrees of freedom:

$$t = \sqrt{n} \frac{\bar{X} - \mu}{s} \sim t(n-1) . \quad (4)$$

**Proof:** Note that  $\bar{X}$  is a linear combination of  $X_1, \dots, X_n$ :

$$\bar{X} = \frac{1}{n}X_1 + \dots + \frac{1}{n}X_n . \quad (5)$$

Because the linear combination of independent normal random variables is also normally distributed ( $\rightarrow$  II/3.2.26), we have:

$$\bar{X} \sim \mathcal{N}\left(\frac{1}{n}n\mu, \left(\frac{1}{n}\right)^2 n\sigma^2\right) = \mathcal{N}(\mu, \sigma^2/n) . \quad (6)$$

Let  $Z = \sqrt{n}(\bar{X} - \mu)/\sigma$ . Because  $Z$  is a linear transformation ( $\rightarrow$  II/4.1.14) of  $\bar{X}$ , it also follows a normal distribution:

$$Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}\left(\frac{\sqrt{n}}{\sigma}(\mu - \mu), \left(\frac{\sqrt{n}}{\sigma}\right)^2 \sigma^2/n\right) = \mathcal{N}(0, 1) . \quad (7)$$

Let  $V = (n-1)s^2/\sigma^2$ . We know that this function of the sample variance follows a chi-squared distribution ( $\rightarrow$  II/3.2.7) with  $n - 1$  degrees of freedom:

$$V = (n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1) . \quad (8)$$

Observe that  $t$  is the ratio of a standard normal random variable ( $\rightarrow$  II/3.2.3) and the square root of a chi-squared random variable ( $\rightarrow$  II/3.7.1), divided by its degrees of freedom:

$$t = \sqrt{n} \frac{\bar{X} - \mu}{s} = \frac{\sqrt{n} \frac{\bar{X} - \mu}{\sigma}}{\sqrt{(n-1) \frac{s^2}{\sigma^2} / (n-1)}} = \frac{Z}{\sqrt{V/(n-1)}} . \quad (9)$$

Thus, by definition of the t-distribution ( $\rightarrow$  II/3.3.1), this ratio follows a t-distribution with  $n - 1$  degrees of freedom:

$$t \sim t(n-1) . \quad (10)$$

#### Sources:

- Wikipedia (2021): “Student’s t-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-27; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Characterization](https://en.wikipedia.org/wiki/Student%27s_t-distribution#Characterization).
- Wikipedia (2021): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-27; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Operations\\_on\\_multiple\\_independent\\_normal\\_variables](https://en.wikipedia.org/wiki/Normal_distribution#Operations_on_multiple_independent_normal_variables).

### 3.2.9 Gaussian integral

**Theorem:** The definite integral of  $\exp[-x^2]$  from  $-\infty$  to  $+\infty$  is equal to the square root of  $\pi$ :

$$\int_{-\infty}^{+\infty} \exp[-x^2] \, dx = \sqrt{\pi} . \quad (1)$$

**Proof:** Let

$$I = \int_0^{\infty} \exp[-x^2] \, dx \quad (2)$$

and

$$I_P = \int_0^P \exp[-x^2] \, dx = \int_0^P \exp[-y^2] \, dy . \quad (3)$$

Then, we have

$$\lim_{P \rightarrow \infty} I_P = I \quad (4)$$

and

$$\lim_{P \rightarrow \infty} I_P^2 = I^2 . \quad (5)$$

Moreover, we can write

$$\begin{aligned} I_P^2 &\stackrel{(3)}{=} \left( \int_0^P \exp[-x^2] \, dx \right) \left( \int_0^P \exp[-y^2] \, dy \right) \\ &= \int_0^P \int_0^P \exp[-(x^2 + y^2)] \, dx \, dy \\ &= \iint_{S_P} \exp[-(x^2 + y^2)] \, dx \, dy \end{aligned} \quad (6)$$

where  $S_P$  is the square with corners  $(0, 0)$ ,  $(0, P)$ ,  $(P, P)$  and  $(P, 0)$ . For this integral, we can write down the following inequality

$$\iint_{C_1} \exp[-(x^2 + y^2)] \, dx \, dy \leq I_P^2 \leq \iint_{C_2} \exp[-(x^2 + y^2)] \, dx \, dy \quad (7)$$

where  $C_1$  and  $C_2$  are the regions in the first quadrant bounded by circles with center at  $(0, 0)$  and going through the points  $(0, P)$  and  $(P, P)$ , respectively. The radii of these two circles are  $r_1 = \sqrt{P^2} = P$  and  $r_2 = \sqrt{2P^2} = P\sqrt{2}$ , such that we can rewrite equation (7) using polar coordinates as

$$\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp[-r^2] \, r \, dr \, d\theta \leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp[-r^2] \, r \, dr \, d\theta . \quad (8)$$

Solving the definite integrals yields:

$$\begin{aligned}
\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp[-r^2] r \, dr \, d\theta &\leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp[-r^2] r \, dr \, d\theta \\
\int_0^{\frac{\pi}{2}} \left[ -\frac{1}{2} \exp[-r^2] \right]_0^{r_1} d\theta &\leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \left[ -\frac{1}{2} \exp[-r^2] \right]_0^{r_2} d\theta \\
-\frac{1}{2} \int_0^{\frac{\pi}{2}} (\exp[-r_1^2] - 1) \, d\theta &\leq I_P^2 \leq -\frac{1}{2} \int_0^{\frac{\pi}{2}} (\exp[-r_2^2] - 1) \, d\theta \\
-\frac{1}{2} [(\exp[-r_1^2] - 1) \theta]_0^{\frac{\pi}{2}} &\leq I_P^2 \leq -\frac{1}{2} [(\exp[-r_2^2] - 1) \theta]_0^{\frac{\pi}{2}} \\
\frac{1}{2} (1 - \exp[-r_1^2]) \frac{\pi}{2} &\leq I_P^2 \leq \frac{1}{2} (1 - \exp[-r_2^2]) \frac{\pi}{2} \\
\frac{\pi}{4} (1 - \exp[-P^2]) &\leq I_P^2 \leq \frac{\pi}{4} (1 - \exp[-2P^2])
\end{aligned} \tag{9}$$

Calculating the limit for  $P \rightarrow \infty$ , we obtain

$$\begin{aligned}
\lim_{P \rightarrow \infty} \frac{\pi}{4} (1 - \exp[-P^2]) &\leq \lim_{P \rightarrow \infty} I_P^2 \leq \lim_{P \rightarrow \infty} \frac{\pi}{4} (1 - \exp[-2P^2]) \\
\frac{\pi}{4} &\leq I^2 \leq \frac{\pi}{4},
\end{aligned} \tag{10}$$

such that we have a preliminary result for  $I$ :

$$I^2 = \frac{\pi}{4} \quad \Rightarrow \quad I = \frac{\sqrt{\pi}}{2}. \tag{11}$$

Because the integrand in (1) is an even function, we can calculate the final result as follows:

$$\begin{aligned}
\int_{-\infty}^{+\infty} \exp[-x^2] \, dx &= 2 \int_0^{\infty} \exp[-x^2] \, dx \\
&\stackrel{(11)}{=} 2 \frac{\sqrt{\pi}}{2} \\
&= \sqrt{\pi}.
\end{aligned} \tag{12}$$

■

#### Sources:

- ProofWiki (2020): “Gaussian Integral”; in: *ProofWiki*, retrieved on 2020-11-25; URL: [https://proofwiki.org/wiki/Gaussian\\_Integral](https://proofwiki.org/wiki/Gaussian_Integral).
- ProofWiki (2020): “Integral to Infinity of Exponential of minus t squared”; in: *ProofWiki*, retrieved on 2020-11-25; URL: [https://proofwiki.org/wiki/Integral\\_to\\_Infinity\\_of\\_Exponential\\_of\\_-t%5E2](https://proofwiki.org/wiki/Integral_to_Infinity_of_Exponential_of_-t%5E2).

### 3.2.10 Probability density function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{1}$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is



$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (2)$$

**Proof:** This follows directly from the definition of the normal distribution ( $\rightarrow$  II/3.2.1). ■

### 3.2.11 Moment-generating function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] . \quad (2)$$

**Proof:** The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (3)$$

and the moment-generating function ( $\rightarrow$  I/1.9.1) is defined as

$$M_X(t) = \mathbb{E} [e^{tX}] . \quad (4)$$

Using the expected value for continuous random variables ( $\rightarrow$  I/1.12.1), the moment-generating function of  $X$  therefore is

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} \exp[tx] \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp \left[ tx - \frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] dx . \end{aligned} \quad (5)$$

Substituting  $u = (x - \mu)/(\sqrt{2}\sigma)$ , i.e.  $x = \sqrt{2}\sigma u + \mu$ , we have

$$\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(+\infty-\mu)/(\sqrt{2}\sigma)} \exp \left[ t \left( \sqrt{2}\sigma u + \mu \right) - \frac{1}{2} \left( \frac{\sqrt{2}\sigma u + \mu - \mu}{\sigma} \right)^2 \right] d \left( \sqrt{2}\sigma u + \mu \right) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp \left[ \left( \sqrt{2}\sigma u + \mu \right) t - u^2 \right] du \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp \left[ \sqrt{2}\sigma u t - u^2 \right] du \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp \left[ - \left( u^2 - \sqrt{2}\sigma u t \right) \right] du \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp \left[ - \left( u - \frac{\sqrt{2}}{2}\sigma t \right)^2 + \frac{1}{2}\sigma^2 t^2 \right] du \\
&= \frac{\exp \left[ \mu t + \frac{1}{2}\sigma^2 t^2 \right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp \left[ - \left( u - \frac{\sqrt{2}}{2}\sigma t \right)^2 \right] du
\end{aligned} \tag{6}$$

Now substituting  $v = u - \sqrt{2}/2 \sigma t$ , i.e.  $u = v + \sqrt{2}/2 \sigma t$ , we have

$$\begin{aligned}
M_X(t) &= \frac{\exp \left[ \mu t + \frac{1}{2}\sigma^2 t^2 \right]}{\sqrt{\pi}} \int_{-\infty - \sqrt{2}/2 \sigma t}^{+\infty - \sqrt{2}/2 \sigma t} \exp \left[ -v^2 \right] d \left( v + \sqrt{2}/2 \sigma t \right) \\
&= \frac{\exp \left[ \mu t + \frac{1}{2}\sigma^2 t^2 \right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp \left[ -v^2 \right] dv .
\end{aligned} \tag{7}$$

With the Gaussian integral ( $\rightarrow$  II/3.2.9)

$$\int_{-\infty}^{+\infty} \exp \left[ -x^2 \right] dx = \sqrt{\pi} , \tag{8}$$

this finally becomes

$$M_X(t) = \exp \left[ \mu t + \frac{1}{2}\sigma^2 t^2 \right] . \tag{9}$$

■

#### Sources:

- ProofWiki (2020): “Moment Generating Function of Gaussian Distribution”; in: *ProofWiki*, retrieved on 2020-03-03; URL: [https://proofwiki.org/wiki/Moment\\_Generating\\_Function\\_of\\_Gaussian\\_Distribution](https://proofwiki.org/wiki/Moment_Generating_Function_of_Gaussian_Distribution).

### 3.2.12 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right] \quad (2)$$

where  $\operatorname{erf}(x)$  is the error function defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt . \quad (3)$$

**Proof:** The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (4)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x \mathcal{N}(z; \mu, \sigma^2) dz \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{z - \mu}{\sigma} \right)^2 \right] dz \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x \exp \left[ -\left( \frac{z - \mu}{\sqrt{2}\sigma} \right)^2 \right] dz . \end{aligned} \quad (5)$$

Substituting  $t = (z - \mu)/(\sqrt{2}\sigma)$ , i.e.  $z = \sqrt{2}\sigma t + \mu$ , this becomes:

$$\begin{aligned} F_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty - \mu)/(\sqrt{2}\sigma)}^{(x - \mu)/(\sqrt{2}\sigma)} \exp(-t^2) d(\sqrt{2}\sigma t + \mu) \\ &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^0 \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt . \end{aligned} \quad (6)$$

Applying (3) to (6), we have:

$$\begin{aligned} F_X(x) &= \frac{1}{2} \lim_{x \rightarrow \infty} \operatorname{erf}(x) + \frac{1}{2} \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \\ &= \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \\ &= \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right] . \end{aligned} \quad (7)$$

**Sources:**

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Cumulative\\_distribution\\_function](https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function).
- Wikipedia (2020): “Error function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: [https://en.wikipedia.org/wiki/Error\\_function](https://en.wikipedia.org/wiki/Error_function).

**3.2.13 Cumulative distribution function without error function**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2). \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  can be expressed as

$$F_X(x) = \Phi_{\mu, \sigma}(x) = \varphi\left(\frac{x - \mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x - \mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2} \quad (2)$$

where  $\varphi(x)$  is the probability density function ( $\rightarrow$  I/1.7.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3) and  $n!!$  is a double factorial.

**Proof:**

1) First, consider the standard normal distribution ( $\rightarrow$  II/3.2.3)  $\mathcal{N}(0, 1)$  which has the probability density function ( $\rightarrow$  II/3.2.10)

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}. \quad (3)$$

Let  $T(x)$  be the indefinite integral of this function. It can be obtained using infinitely repeated integration by parts as follows:

$$\begin{aligned} T(x) &= \int \varphi(x) \, dx \\ &= \int \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \, dx \\ &= \frac{1}{\sqrt{2\pi}} \int 1 \cdot e^{-\frac{1}{2}x^2} \, dx \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left[ x \cdot e^{-\frac{1}{2}x^2} + \int x^2 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left[ x \cdot e^{-\frac{1}{2}x^2} + \left[ \frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{3}x^4 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left[ x \cdot e^{-\frac{1}{2}x^2} + \left[ \frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \left[ \frac{1}{15}x^5 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{15}x^6 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \right] \\ &= \dots \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left[ \sum_{i=1}^n \left( \frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \int \left( \frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right] \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left[ \sum_{i=1}^{\infty} \left( \frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \lim_{n \rightarrow \infty} \int \left( \frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right]. \end{aligned} \quad (4)$$

Since  $(2n - 1)!!$  grows faster than  $x^{2n}$ , it holds that

$$\frac{1}{\sqrt{2\pi}} \cdot \lim_{n \rightarrow \infty} \int \left( \frac{x^{2n}}{(2n - 1)!!} \cdot e^{-\frac{1}{2}x^2} \right) dx = \int 0 dx = c \quad (5)$$

for constant  $c$ , such that the indefinite integral becomes

$$\begin{aligned} T(x) &= \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{\infty} \left( \frac{x^{2i-1}}{(2i - 1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + c \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i - 1)!!} + c \\ &\stackrel{(3)}{=} \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i - 1)!!} + c. \end{aligned} \quad (6)$$

2) Next, let  $\Phi(x)$  be the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3):

$$\Phi(x) = \int_{-\infty}^x \varphi(x) dx. \quad (7)$$

It can be obtained by matching  $T(0)$  to  $\Phi(0)$  which is  $1/2$ , because the standard normal distribution is symmetric around zero:

$$\begin{aligned} T(0) &= \varphi(0) \cdot \sum_{i=1}^{\infty} \frac{0^{2i-1}}{(2i - 1)!!} + c = \frac{1}{2} = \Phi(0) \\ &\Leftrightarrow c = \frac{1}{2} \\ \Rightarrow \Phi(x) &= \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i - 1)!!} + \frac{1}{2}. \end{aligned} \quad (8)$$

3) Finally, the cumulative distribution functions ( $\rightarrow$  I/1.8.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3) and the general normal distribution ( $\rightarrow$  II/3.2.1) are related to each other ( $\rightarrow$  II/3.2.4) as

$$\Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (9)$$

Combining (9) with (8), we have:

$$\Phi_{\mu,\sigma}(x) = \varphi\left(\frac{x - \mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x - \mu}{\sigma}\right)^{2i-1}}{(2i - 1)!!} + \frac{1}{2}. \quad (10)$$

■

#### Sources:

- Soch J (2015): “Solution for the Indefinite Integral of the Standard Normal Probability Density Function”; in: *arXiv stat.OT*, 1512.04858; URL: <https://arxiv.org/abs/1512.04858>.
- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Cumulative\\_distribution\\_function](https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function).

### 3.2.14 Probability of being within standard deviations from mean

**Theorem:** (also called “68-95-99.7 rule”) Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1) with mean ( $\rightarrow$  I/1.12.1)  $\mu$  and variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ . Then, about 68%, 95% and 99.7% of the values of  $X$  will fall within 1, 2 and 3 standard deviations ( $\rightarrow$  I/1.19.1) from the mean ( $\rightarrow$  I/1.12.1), respectively:

$$\begin{aligned}\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &\approx 68\% \\ \Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95\% \\ \Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99.7\% .\end{aligned}\tag{1}$$

**Proof:** The cumulative distribution function of a normally distributed ( $\rightarrow$  II/3.2.12) random variable  $X$  is

$$F_X(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right]\tag{2}$$

where  $\operatorname{erf}(x)$  is the error function defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt\tag{3}$$

which exhibits a point-symmetry property:

$$\operatorname{erf}(-x) = -\operatorname{erf}(x) .\tag{4}$$

Thus, the probability that  $X$  falls between  $\mu - a \cdot \sigma$  and  $\mu + a \cdot \sigma$  is equal to:

$$\begin{aligned}p(a) &= \Pr(\mu - a\sigma \leq X \leq \mu + a\sigma) \\ &= F_X(\mu + a\sigma) - F_X(\mu - a\sigma) \\ &\stackrel{(2)}{=} \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma} \right) \right] - \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma} \right) \right] \\ &= \frac{1}{2} \left[ \operatorname{erf} \left( \frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma} \right) - \operatorname{erf} \left( \frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma} \right) \right] \\ &= \frac{1}{2} \left[ \operatorname{erf} \left( \frac{a}{\sqrt{2}} \right) - \operatorname{erf} \left( -\frac{a}{\sqrt{2}} \right) \right] \\ &\stackrel{(4)}{=} \frac{1}{2} \left[ \operatorname{erf} \left( \frac{a}{\sqrt{2}} \right) + \operatorname{erf} \left( \frac{a}{\sqrt{2}} \right) \right] \\ &= \operatorname{erf} \left( \frac{a}{\sqrt{2}} \right)\end{aligned}\tag{5}$$

With that, we can use numerical implementations of the error function to calculate:

$$\begin{aligned}\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &= p(1) = 68.27\% \\ \Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= p(2) = 95.45\% \\ \Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= p(3) = 99.73\% .\end{aligned}\tag{6}$$

**Sources:**

- Wikipedia (2022): “68-95-99.7 rule”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-05.08; URL: [https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7\\_rule](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule).

**3.2.15 Quantile function**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distributions ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p - 1) + \mu \quad (2)$$

where  $\operatorname{erf}^{-1}(x)$  is the inverse error function.

**Proof:** The cumulative distribution function of the normal distribution ( $\rightarrow$  II/3.2.12) is:

$$F_X(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right] . \quad (3)$$

Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF ( $\rightarrow$  I/1.11.2):

$$Q_X(p) = F_X^{-1}(x) . \quad (4)$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right] \\ 2p - 1 &= \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \\ \operatorname{erf}^{-1}(2p - 1) &= \frac{x - \mu}{\sqrt{2}\sigma} \\ x &= \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p - 1) + \mu . \end{aligned} \quad (5)$$

**Sources:**

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Quantile\\_function](https://en.wikipedia.org/wiki/Normal_distribution#Quantile_function).

**3.2.16 Mean**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \mu . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx . \quad (3)$$

With the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), this reads:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] dx . \end{aligned} \quad (4)$$

Substituting  $z = x - \mu$ , we have:

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z + \mu) \cdot \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] d(z + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z + \mu) \cdot \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] dz \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left( \int_{-\infty}^{+\infty} z \cdot \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] dz + \mu \int_{-\infty}^{+\infty} \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] dz \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left( \int_{-\infty}^{+\infty} z \cdot \exp \left[ -\frac{1}{2\sigma^2} \cdot z^2 \right] dz + \mu \int_{-\infty}^{+\infty} \exp \left[ -\frac{1}{2\sigma^2} \cdot z^2 \right] dz \right) . \end{aligned} \quad (5)$$

The general antiderivatives are

$$\begin{aligned} \int x \cdot \exp [-ax^2] dx &= -\frac{1}{2a} \cdot \exp [-ax^2] \\ \int \exp [-ax^2] dx &= \frac{1}{2} \sqrt{\frac{\pi}{a}} \cdot \operatorname{erf} [\sqrt{a}x] \end{aligned} \quad (6)$$

where  $\operatorname{erf}(x)$  is the error function. Using this, the integrals can be calculated as:

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left( \left[ -\sigma^2 \cdot \exp \left[ -\frac{1}{2\sigma^2} \cdot z^2 \right] \right]_{-\infty}^{+\infty} + \mu \left[ \sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[ \frac{1}{\sqrt{2}\sigma} z \right] \right]_{-\infty}^{+\infty} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left( \left( \lim_{z \rightarrow +\infty} \left( -\sigma^2 \cdot \exp \left[ -\frac{1}{2\sigma^2} \cdot z^2 \right] \right) - \lim_{z \rightarrow -\infty} \left( -\sigma^2 \cdot \exp \left[ -\frac{1}{2\sigma^2} \cdot z^2 \right] \right) \right) \right. \\ &\quad \left. + \mu \left( \lim_{z \rightarrow +\infty} \left( \sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[ \frac{1}{\sqrt{2}\sigma} z \right] \right) - \lim_{z \rightarrow -\infty} \left( \sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[ \frac{1}{\sqrt{2}\sigma} z \right] \right) \right) \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left( [0 - 0] + \mu \left[ \sqrt{\frac{\pi}{2}} \sigma - \left( -\sqrt{\frac{\pi}{2}} \sigma \right) \right] \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2\sqrt{\frac{\pi}{2}} \sigma \\ &= \mu . \end{aligned} \quad (7)$$



**Sources:**

- Papadopoulos, Alecos (2013): “How to derive the mean and variance of Gaussian random variable?”; in: *StackExchange Mathematics*, retrieved on 2020-01-09; URL: <https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable>.

**3.2.17 Median**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = \mu . \quad (2)$$

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function ( $\rightarrow$  I/1.8.1) is  $1/2$ :

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the normal distribution ( $\rightarrow$  II/3.2.12) is

$$F_X(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right] \quad (4)$$

where  $\text{erf}(x)$  is the error function. Thus, the inverse CDF is

$$x = \sqrt{2}\sigma \cdot \text{erf}^{-1}(2p - 1) + \mu \quad (5)$$

where  $\text{erf}^{-1}(x)$  is the inverse error function. Setting  $p = 1/2$ , we obtain:

$$\text{median}(X) = \sqrt{2}\sigma \cdot \text{erf}^{-1}(0) + \mu = \mu . \quad (6)$$

**3.2.18 Mode**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) = \mu . \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (4)$$

The first two derivatives of this function are:

$$f'_X(x) = \frac{df_X(x)}{dx} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x + \mu) \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (5)$$

$$f''_X(x) = \frac{d^2f_X(x)}{dx^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x + \mu)^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (6)$$

We now calculate the root of the first derivative (5):

$$\begin{aligned} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x + \mu) \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \\ 0 &= -x + \mu \\ x &= \mu . \end{aligned} \quad (7)$$

By plugging this value into the second derivative (6),

$$\begin{aligned} f''_X(\mu) &= -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp(0) + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (0)^2 \cdot \exp(0) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 , \end{aligned} \quad (8)$$

we confirm that it is in fact a maximum which shows that

$$\text{mode}(X) = \mu . \quad (9)$$

■

### 3.2.19 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \sigma^2 . \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) is the probability-weighted average of the squared deviation from the mean ( $\rightarrow$  I/1.12.1):

$$\text{Var}(X) = \int_{\mathbb{R}} (x - E(X))^2 \cdot f_X(x) dx . \quad (3)$$

With the expected value ( $\rightarrow$  II/3.2.16) and probability density function ( $\rightarrow$  II/3.2.10) of the normal distribution, this reads:

$$\begin{aligned}\text{Var}(X) &= \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] dx .\end{aligned}\quad (4)$$

Substituting  $z = x - \mu$ , we have:

$$\begin{aligned}\text{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} z^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] d(z + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} z^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 \right] dz .\end{aligned}\quad (5)$$

Now substituting  $z = \sqrt{2}\sigma x$ , we have:

$$\begin{aligned}\text{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{\sqrt{2}\sigma x}{\sigma} \right)^2 \right] d(\sqrt{2}\sigma x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp [-x^2] dx \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} dx .\end{aligned}\quad (6)$$

Since the integrand is symmetric with respect to  $x = 0$ , we can write:

$$\text{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} x^2 \cdot e^{-x^2} dx .\quad (7)$$

If we define  $z = x^2$ , then  $x = \sqrt{z}$  and  $dx = 1/2 z^{-1/2} dz$ . Substituting this into the integral

$$\text{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z \cdot e^{-z} \cdot \frac{1}{2} z^{-1/2} dz = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z^{\frac{3}{2}-1} \cdot e^{-z} dz\quad (8)$$

and using the definition of the gamma function

$$\Gamma(x) = \int_0^{\infty} z^{x-1} \cdot e^{-z} dz ,\quad (9)$$

we can finally show that

$$\text{Var}(X) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = \sigma^2 .\quad (10)$$

■

**Sources:**

- Papadopoulos, Alecos (2013): “How to derive the mean and variance of Gaussian random variable?”; in: *StackExchange Mathematics*, retrieved on 2020-01-09; URL: <https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable>.
- Wikipedia (2025): “Particular values of the gamma function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-12-19; URL: [https://en.wikipedia.org/wiki/Particular\\_values\\_of\\_the\\_gamma\\_function#Integers\\_and\\_half-integers](https://en.wikipedia.org/wiki/Particular_values_of_the_gamma_function#Integers_and_half-integers).

### 3.2.20 Full width at half maximum

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the full width at half maximum ( $\rightarrow$  I/1.19.4) (FWHM) of  $X$  is

$$\text{FWHM}(X) = 2\sqrt{2 \ln 2} \sigma . \quad (2)$$

**Proof:** The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (3)$$

and the mode of the normal distribution ( $\rightarrow$  II/3.2.18) is

$$\text{mode}(X) = \mu , \quad (4)$$

such that

$$f_{\max} = f_X(\text{mode}(X)) \stackrel{(4)}{=} f_X(\mu) \stackrel{(3)}{=} \frac{1}{\sqrt{2\pi}\sigma} . \quad (5)$$

The FWHM bounds satisfy the equation ( $\rightarrow$  I/1.19.4)

$$f_X(x_{\text{FWHM}}) = \frac{1}{2} f_{\max} \stackrel{(5)}{=} \frac{1}{2\sqrt{2\pi}\sigma} . \quad (6)$$

Using (3), we can develop this equation as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 \right] &= \frac{1}{2\sqrt{2\pi}\sigma} \\
\exp \left[ -\frac{1}{2} \left( \frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 \right] &= \frac{1}{2} \\
-\frac{1}{2} \left( \frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 &= \ln \frac{1}{2} \\
\left( \frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 &= -2 \ln \frac{1}{2} \\
\frac{x_{\text{FWHM}} - \mu}{\sigma} &= \pm \sqrt{2 \ln 2} \\
x_{\text{FWHM}} - \mu &= \pm \sqrt{2 \ln 2} \sigma \\
x_{\text{FWHM}} &= \pm \sqrt{2 \ln 2} \sigma + \mu .
\end{aligned} \tag{7}$$

This implies the following two solutions for  $x_{\text{FWHM}}$

$$\begin{aligned}
x_1 &= \mu - \sqrt{2 \ln 2} \sigma \\
x_2 &= \mu + \sqrt{2 \ln 2} \sigma ,
\end{aligned} \tag{8}$$

such that the full width at half maximum ( $\rightarrow$  I/1.19.4) of  $X$  is

$$\begin{aligned}
\text{FWHM}(X) &= \Delta x = x_2 - x_1 \\
&\stackrel{(8)}{=} \left( \mu + \sqrt{2 \ln 2} \sigma \right) - \left( \mu - \sqrt{2 \ln 2} \sigma \right) \\
&= 2\sqrt{2 \ln 2} \sigma .
\end{aligned} \tag{9}$$

■

#### Sources:

- Wikipedia (2020): “Full width at half maximum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: [https://en.wikipedia.org/wiki/Full\\_width\\_at\\_half\\_maximum](https://en.wikipedia.org/wiki/Full_width_at_half_maximum).

#### 3.2.21 Extreme points

**Theorem:** The probability density function ( $\rightarrow$  I/1.7.1) of the normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$  has a maximum at  $x = \mu$  and no other extrema. Consequently, the normal distribution ( $\rightarrow$  II/3.2.1) is a unimodal probability distribution ( $\rightarrow$  I/1.5.5).

**Proof:** The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \tag{1}$$

The first two derivatives of this function ( $\rightarrow$  II/3.2.18) are:

$$f'_X(x) = \frac{df_X(x)}{dx} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x + \mu) \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (2)$$

$$f''_X(x) = \frac{d^2f_X(x)}{dx^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x + \mu)^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (3)$$

The first derivative is zero, if and only if

$$-x + \mu = 0 \quad \Leftrightarrow \quad x = \mu . \quad (4)$$

Since the second derivative is negative at this value

$$f''_X(\mu) = -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 , \quad (5)$$

there is a maximum at  $x = \mu$ . From (2), it can be seen that  $f'_X(x)$  is positive for  $x < \mu$  and negative for  $x > \mu$ . Thus, there are no further extrema and  $\mathcal{N}(\mu, \sigma^2)$  is unimodal ( $\rightarrow$  II/3.2.18). ■

#### Sources:

- Wikipedia (2021): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Symmetries\\_and\\_derivatives](https://en.wikipedia.org/wiki/Normal_distribution#Symmetries_and_derivatives).

### 3.2.22 Inflection points

**Theorem:** The probability density function ( $\rightarrow$  I/1.7.1) of the normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$  has two inflection points at  $x = \mu - \sigma$  and  $x = \mu + \sigma$ , i.e. exactly one standard deviation ( $\rightarrow$  I/1.19.1) away from the expected value ( $\rightarrow$  I/1.12.1).

**Proof:** The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (1)$$

The first three derivatives of this function are:

$$f'_X(x) = \frac{df_X(x)}{dx} = \frac{1}{\sqrt{2\pi}\sigma} \cdot \left( -\frac{x - \mu}{\sigma^2} \right) \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \quad (2)$$

$$\begin{aligned} f''_X(x) &= \frac{d^2f_X(x)}{dx^2} = \frac{1}{\sqrt{2\pi}\sigma} \cdot \left( -\frac{1}{\sigma^2} \right) \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] + \frac{1}{\sqrt{2\pi}\sigma} \cdot \left( \frac{x - \mu}{\sigma^2} \right)^2 \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[ \left( \frac{x - \mu}{\sigma^2} \right)^2 - \frac{1}{\sigma^2} \right] \cdot \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \end{aligned} \quad (3)$$

$$\begin{aligned}
f_X'''(x) &= \frac{d^3 f_X(x)}{dx^3} = \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[ \frac{2}{\sigma^2} \left( \frac{x-\mu}{\sigma^2} \right) \right] \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] - \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[ \left( \frac{x-\mu}{\sigma^2} \right)^2 - \frac{1}{\sigma^2} \right] \cdot \left( \frac{x-\mu}{\sigma^2} \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[ -\left( \frac{x-\mu}{\sigma^2} \right)^3 + 3 \left( \frac{x-\mu}{\sigma^4} \right) \right] \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right].
\end{aligned} \tag{4}$$

The second derivative is zero, if and only if

$$\begin{aligned}
0 &= \left[ \left( \frac{x-\mu}{\sigma^2} \right)^2 - \frac{1}{\sigma^2} \right] \\
0 &= \frac{x^2}{\sigma^4} - \frac{2\mu x}{\sigma^4} + \frac{\mu^2}{\sigma^4} - \frac{1}{\sigma^2} \\
0 &= x^2 - 2\mu x + (\mu^2 - \sigma^2) \\
x_{1/2} &= -\frac{-2\mu}{2} \pm \sqrt{\left( \frac{-2\mu}{2} \right)^2 - (\mu^2 - \sigma^2)} \\
x_{1/2} &= \mu \pm \sqrt{\mu^2 - \mu^2 + \sigma^2} \\
x_{1/2} &= \mu \pm \sigma.
\end{aligned} \tag{5}$$

Since the third derivative is non-zero at this value

$$\begin{aligned}
f_X'''(\mu \pm \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[ -\left( \frac{\pm\sigma}{\sigma^2} \right)^3 + 3 \left( \frac{\pm\sigma}{\sigma^4} \right) \right] \cdot \exp \left[ -\frac{1}{2} \left( \frac{\pm\sigma}{\sigma} \right)^2 \right] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left( \pm \frac{2}{\sigma^3} \right) \cdot \exp \left( -\frac{1}{2} \right) \neq 0,
\end{aligned} \tag{6}$$

there are inflection points at  $x_{1/2} = \mu \pm \sigma$ . Because  $\mu$  is the mean and  $\sigma^2$  is the variance of a normal distribution ( $\rightarrow$  II/3.2.1), these points are exactly one standard deviation ( $\rightarrow$  I/1.19.1) away from the mean.

■

#### Sources:

- Wikipedia (2021): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: [https://en.wikipedia.org/wiki/Normal\\_distribution#Symmetries\\_and\\_derivatives](https://en.wikipedia.org/wiki/Normal_distribution#Symmetries_and_derivatives).

### 3.2.23 Differential entropy

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{1}$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  is

$$h(X) = \frac{1}{2} \ln(2\pi\sigma^2 e) . \quad (2)$$

**Proof:** The differential entropy ( $\rightarrow$  I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \quad (3)$$

To measure  $h(X)$  in nats, we set  $b = e$ , such that ( $\rightarrow$  I/1.12.1)

$$h(X) = -E[\ln p(x)] . \quad (4)$$

With the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), the differential entropy of  $X$  is:

$$\begin{aligned} h(X) &= -E \left[ \ln \left( \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \right) \right] \\ &= -E \left[ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} E \left[ \left( \frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot E[(x-\mu)^2] . \end{aligned} \quad (5)$$

Note that  $E[(x-\mu)^2]$  corresponds to the variance ( $\rightarrow$  I/1.21.8) of  $X$  and the variance of the normal distribution ( $\rightarrow$  II/3.2.19) is  $\sigma^2$ . Thus, we can proceed:

$$\begin{aligned} h(X) &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \sigma^2 \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2} \ln(2\pi\sigma^2 e) . \end{aligned} \quad (6)$$

■

#### Sources:

- Wang, Peng-Hua (2012): “Differential Entropy”; in: *National Taipei University*; URL: <https://web.ntpu.edu.tw/~phwang/teaching/2012s/IT/slides/chap08.pdf>.

#### 3.2.24 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two normal distributions ( $\rightarrow$  II/3.2.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ Q : X &\sim \mathcal{N}(\mu_2, \sigma_2^2) . \end{aligned} \quad (1)$$



Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[ \frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right]. \quad (2)$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the normal distributions ( $\rightarrow$  II/3.2.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{-\infty}^{+\infty} \mathcal{N}(x; \mu_1, \sigma_1^2) \ln \frac{\mathcal{N}(x; \mu_1, \sigma_1^2)}{\mathcal{N}(x; \mu_2, \sigma_2^2)} dx \\ &= \left\langle \ln \frac{\mathcal{N}(x; \mu_1, \sigma_1^2)}{\mathcal{N}(x; \mu_2, \sigma_2^2)} \right\rangle_{p(x)}. \end{aligned} \quad (4)$$

Using the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2\pi}\sigma_1} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu_1}{\sigma_1} \right)^2 \right]}{\frac{1}{\sqrt{2\pi}\sigma_2} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu_2}{\sigma_2} \right)^2 \right]} \right\rangle_{p(x)} \\ &= \left\langle \ln \left( \sqrt{\frac{\sigma_2^2}{\sigma_1^2}} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu_1}{\sigma_1} \right)^2 + \frac{1}{2} \left( \frac{x-\mu_2}{\sigma_2} \right)^2 \right] \right) \right\rangle_{p(x)} \\ &= \left\langle \frac{1}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} \left( \frac{x-\mu_1}{\sigma_1} \right)^2 + \frac{1}{2} \left( \frac{x-\mu_2}{\sigma_2} \right)^2 \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle - \left( \frac{x-\mu_1}{\sigma_1} \right)^2 + \left( \frac{x-\mu_2}{\sigma_2} \right)^2 - \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle - \frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{x^2 - 2\mu_2 x + \mu_2^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)}. \end{aligned} \quad (5)$$

Because the expected value ( $\rightarrow$  I/1.12.1) is a linear operator ( $\rightarrow$  I/1.12.5), the expectation can be moved into the sum:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left[ - \frac{\langle (x-\mu_1)^2 \rangle}{\sigma_1^2} + \frac{\langle x^2 - 2\mu_2 x + \mu_2^2 \rangle}{\sigma_2^2} - \left\langle \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle \right] \\ &= \frac{1}{2} \left[ - \frac{\langle (x-\mu_1)^2 \rangle}{\sigma_1^2} + \frac{\langle x^2 \rangle - \langle 2\mu_2 x \rangle + \langle \mu_2^2 \rangle}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right]. \end{aligned} \quad (6)$$

The first expectation corresponds to the variance ( $\rightarrow$  I/1.13.1)

$$\langle (X - \mu)^2 \rangle = \text{E}[(X - \text{E}(X))^2] = \text{Var}(X) \quad (7)$$

and the variance of a normally distributed random variable ( $\rightarrow$  II/3.2.19) is

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \text{Var}(X) = \sigma^2 . \quad (8)$$

Additionally applying the raw moments of the normal distribution ( $\rightarrow$  II/3.2.11)

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \langle x \rangle = \mu \quad \text{and} \quad \langle x^2 \rangle = \mu^2 + \sigma^2 , \quad (9)$$

the Kullback-Leibler divergence in (6) becomes

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left[ -\frac{\sigma_1^2}{\sigma_2^2} + \frac{\mu_1^2 + \sigma_1^2 - 2\mu_2\mu_1 + \mu_2^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right] \\ &= \frac{1}{2} \left[ \frac{\mu_1^2 - 2\mu_1\mu_2 + \mu_2^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] \\ &= \frac{1}{2} \left[ \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] \end{aligned} \quad (10)$$

which is equivalent to (2). ■

### 3.2.25 Maximum entropy distribution

**Theorem:** The normal distribution ( $\rightarrow$  II/3.2.1) maximizes differential entropy ( $\rightarrow$  I/2.2.1) for a random variable ( $\rightarrow$  I/1.2.2) with fixed variance ( $\rightarrow$  I/1.13.1).

**Proof:** For a random variable ( $\rightarrow$  I/1.2.2)  $X$  with set of possible values  $\mathcal{X}$  and probability density function ( $\rightarrow$  I/1.7.1)  $p(x)$ , the differential entropy ( $\rightarrow$  I/2.2.1) is defined as:

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) \, dx . \quad (1)$$

Let  $g(x)$  be the probability density function ( $\rightarrow$  I/1.7.1) of a normal distribution ( $\rightarrow$  II/3.2.1) with mean ( $\rightarrow$  I/1.12.1)  $\mu$  and variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$  and let  $f(x)$  be an arbitrary probability density function ( $\rightarrow$  I/1.7.1) with the same variance ( $\rightarrow$  I/1.13.1). Since differential entropy ( $\rightarrow$  I/2.2.1) is translation-invariant ( $\rightarrow$  I/2.2.3), we can assume that  $f(x)$  has the same mean as  $g(x)$ .

Consider the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of distribution  $f(x)$  from distribution  $g(x)$  which is non-negative ( $\rightarrow$  I/2.5.2):

$$\begin{aligned} 0 \leq \text{KL}[f \parallel g] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \, dx \\ &= \int_{\mathcal{X}} f(x) \log f(x) \, dx - \int_{\mathcal{X}} f(x) \log g(x) \, dx \\ &\stackrel{(1)}{=} -h[f(x)] - \int_{\mathcal{X}} f(x) \log g(x) \, dx . \end{aligned} \quad (2)$$

By plugging the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) into the second term, we obtain:

$$\begin{aligned}
\int_{\mathcal{X}} f(x) \log g(x) dx &= \int_{\mathcal{X}} f(x) \log \left( \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \right) dx \\
&= \int_{\mathcal{X}} f(x) \log \left( \frac{1}{\sqrt{2\pi}\sigma^2} \right) dx + \int_{\mathcal{X}} f(x) \log \left( \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right] \right) dx \quad (3) \\
&= -\frac{1}{2} \log(2\pi\sigma^2) \int_{\mathcal{X}} f(x) dx - \frac{\log(e)}{2\sigma^2} \int_{\mathcal{X}} f(x)(x-\mu)^2 dx .
\end{aligned}$$

Because the entire integral over a probability density function is one ( $\rightarrow$  I/1.7.1) and the second central moment is equal to the variance ( $\rightarrow$  I/1.21.8), we have:

$$\begin{aligned}
\int_{\mathcal{X}} f(x) \log g(x) dx &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{\log(e)\sigma^2}{2\sigma^2} \\
&= -\frac{1}{2} [\log(2\pi\sigma^2) + \log(e)] \quad (4) \\
&= -\frac{1}{2} \log(2\pi\sigma^2 e) .
\end{aligned}$$

This is actually the negative of the differential entropy of the normal distribution ( $\rightarrow$  II/3.2.23), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) dx = -h[\mathcal{N}(\mu, \sigma^2)] = -h[g(x)] . \quad (5)$$

Combining (2) with (5), we can show that

$$\begin{aligned}
0 &\leq \text{KL}[f||g] \\
0 &\leq -h[f(x)] - (-h[g(x)]) \quad (6) \\
h[g(x)] &\geq h[f(x)]
\end{aligned}$$

which means that the differential entropy ( $\rightarrow$  I/2.2.1) of the normal distribution ( $\rightarrow$  II/3.2.1)  $\mathcal{N}(\mu, \sigma^2)$  will be larger than or equal to any other distribution ( $\rightarrow$  I/1.5.1) with the same variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ . ■

#### Sources:

- Wikipedia (2021): “Differential entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: [https://en.wikipedia.org/wiki/Differential\\_entropy#Maximization\\_in\\_the\\_normal\\_distribution](https://en.wikipedia.org/wiki/Differential_entropy#Maximization_in_the_normal_distribution).

### 3.2.26 Linear combination of independent normals

**Theorem:** Let  $X_1, \dots, X_n$  be independent ( $\rightarrow$  I/1.3.6) normally distributed ( $\rightarrow$  II/3.2.1) random variables ( $\rightarrow$  I/1.2.2) with means ( $\rightarrow$  I/1.12.1)  $\mu_1, \dots, \mu_n$  and variances ( $\rightarrow$  I/1.13.1)  $\sigma_1^2, \dots, \sigma_n^2$ :

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{for } i = 1, \dots, n . \quad (1)$$

Then, any linear combination of those random variables

$$Y = \sum_{i=1}^n a_i X_i \quad \text{where} \quad a_1, \dots, a_n \in \mathbb{R} \quad (2)$$

also follows a normal distribution

$$Y \sim \mathcal{N} \left( \sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2 \right) \quad (3)$$

with mean and variance which are functions of the individual means and variances.

**Proof:** A set of  $n$  independent normal random variables  $X_1, \dots, X_n$  is equivalent ( $\rightarrow$  II/4.1.17) to an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3)  $x$  following a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with a diagonal covariance matrix ( $\rightarrow$  I/1.16.1). Therefore, we can write

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, n \quad \Rightarrow \quad x = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) \quad (4)$$

with mean vector and covariance matrix

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} = \text{diag}([\sigma_1^2, \dots, \sigma_n^2]) \quad (5)$$

Thus, we can apply the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \quad (6)$$

with the constant matrix and vector

$$A = [a_1, \dots, a_n] \quad \text{and} \quad b = 0 \quad (7)$$

This implies the following distribution for the linear combination given by equation (2):

$$Y = Ax + b \sim \mathcal{N}(A\mu, A\Sigma A^T) \quad (8)$$

Finally, we note that

$$\begin{aligned} A\mu &= [a_1, \dots, a_n] \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \sum_{i=1}^n a_i \mu_i \quad \text{and} \\ A\Sigma A^T &= [a_1, \dots, a_n] \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^n a_i^2 \sigma_i^2. \end{aligned} \quad (9)$$



### 3.2.27 Normal and uncorrelated does not imply independent

**Theorem:** Consider two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ . If each of them is normally distributed ( $\rightarrow$  II/3.2.1) and both are uncorrelated ( $\rightarrow$  I/1.17.1), then  $X$  and  $Y$  are not necessarily independent ( $\rightarrow$  I/1.3.6).

**Proof:** As an example, let  $V$  follow a Bernoulli distribution ( $\rightarrow$  II/1.1.1) with success probability ( $\rightarrow$  II/1.1.1)  $1/2$  and let  $W$  be defined as a transformation of  $V$ :

$$\begin{aligned} V &\sim \text{Bern}\left(\frac{1}{2}\right) \\ W &= 2V - 1 . \end{aligned} \tag{1}$$

By definition of the Bernoulli distribution ( $\rightarrow$  II/1.1.1), it follows that

$$p(V = 0) = p(V = 1) = \frac{1}{2} \quad \Rightarrow \quad p(W = -1) = p(W = +1) = \frac{1}{2} . \tag{2}$$

Moreover, let  $X$  follow a standard normal distribution ( $\rightarrow$  II/3.2.3) and let  $Y$  be defined as a combination of  $X$  and  $W$ :

$$\begin{aligned} X &\sim \mathcal{N}(0, 1) \\ Y &= WX . \end{aligned} \tag{3}$$

Then, by the nature of the random variable ( $\rightarrow$  I/1.2.2)  $W$ , it follows that

$$p(W = -1) = p(W = +1) = \frac{1}{2} \quad \Rightarrow \quad p(Y = -X) = p(Y = +X) = \frac{1}{2} . \tag{4}$$

Since the negative of a standard normal ( $\rightarrow$  II/3.2.3) random variable is also standard normally distributed ( $\rightarrow$  II/3.2.26),

$$X \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad -X \sim \mathcal{N}(0, 1) , \tag{5}$$

we can calculate the probability density function ( $\rightarrow$  I/1.7.1) belonging to the mixture distribution of  $Y$  as follows:

$$\begin{aligned} p(y) &= p(y|Y = -X) \cdot p(Y = -X) + p(y|Y = +X) \cdot p(Y = +X) \\ &\stackrel{(4)}{=} \mathcal{N}(y; 0, 1) \cdot \frac{1}{2} + \mathcal{N}(y; 0, 1) \cdot \frac{1}{2} \\ &= \mathcal{N}(y; 0, 1) \end{aligned} \tag{6}$$

where we have used the law of marginal probability ( $\rightarrow$  I/1.3.3) in the first line and  $\mathcal{N}(x; \mu, \sigma^2)$  denotes the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10). Thus,  $Y$  is also standard normally distributed ( $\rightarrow$  II/3.2.3):

$$Y \sim \mathcal{N}(0, 1) . \tag{7}$$

This means that both  $X$  and  $Y$  have expected value zero:

$$E(X) = E(Y) = 0 . \quad (8)$$

With that, we can start to work out the covariance of  $X$  and  $Y$ :

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &\stackrel{(8)}{=} E[XY] \\ &\stackrel{(3)}{=} E[XWX] \\ &= E[WX^2] . \end{aligned} \quad (9)$$

Since  $W$  and  $X$  are independent ( $\rightarrow$  I/1.3.6) by construction, their expected values factorize ( $\rightarrow$  I/1.12.7):

$$\begin{aligned} \text{Cov}(X, Y) &= E[W] \cdot E[X^2] \\ &= ((-1) \cdot p(W = -1) + (+1) \cdot p(W = +1)) \cdot E[X^2] \\ &\stackrel{(2)}{=} \left( (-1) \cdot \frac{1}{2} + (+1) \cdot \frac{1}{2} \right) \cdot E[X^2] \\ &= 0 \cdot E[X^2] \\ &= 0 . \end{aligned} \quad (10)$$

Thus,  $X$  and  $Y$  are uncorrelated ( $\rightarrow$  I/1.17.1):

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = 0 . \quad (11)$$

Yet,  $X$  and  $Y$  are not independent ( $\rightarrow$  I/1.3.6), since the marginal density ( $\rightarrow$  I/1.5.3) of  $Y$  is

$$p(y) = \mathcal{N}(y; 0, 1) , \quad (12)$$

but the conditional density ( $\rightarrow$  I/1.5.4) of  $Y$  given  $X$  is

$$p(y|x) = \begin{cases} 1/2 , & \text{if } y = -x \\ 1/2 , & \text{if } y = +x \\ 0 , & \text{otherwise} \end{cases} , \quad (13)$$

thus violating the behavior of probability under independence ( $\rightarrow$  I/1.3.9):

$$p(Y) \neq p(Y|X) . \quad (14)$$

Therefore,  $X$  and  $Y$  defined by (3) and (1) constitute an example for two random variables ( $\rightarrow$  I/1.2.2) that are normally distributed ( $\rightarrow$  II/3.2.1) and uncorrelated ( $\rightarrow$  I/1.17.1), but not independent ( $\rightarrow$  I/1.3.6).

■

#### Sources:

- Wikipedia (2024): “Misconceptions about the normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-04; URL: [https://en.wikipedia.org/wiki/Misconceptions\\_about\\_the\\_normal\\_distribution#A\\_symmetric\\_example](https://en.wikipedia.org/wiki/Misconceptions_about_the_normal_distribution#A_symmetric_example).

**3.2.28 Marginally normal does not imply jointly normal**

**Theorem:** Consider two random variables ( $\rightarrow$  I/1.2.2)  $X$  and  $Y$ . If the marginal distribution ( $\rightarrow$  I/1.5.3) of each of them is a normal distribution ( $\rightarrow$  II/3.2.1), then the joint distribution ( $\rightarrow$  I/1.5.2)  $X$  and  $Y$  is not necessarily a (multivariate) normal distribution ( $\rightarrow$  II/4.1.1).

**Proof:** Consider the example used to show that normally distributed and uncorrelated does not imply independent ( $\rightarrow$  II/3.2.27). This is characterized by the random variables ( $\rightarrow$  I/1.2.2)

$$\begin{aligned} V &\sim \text{Bern}\left(\frac{1}{2}\right) \\ W &= 2V - 1 . \end{aligned} \tag{1}$$

and

$$\begin{aligned} X &\sim \mathcal{N}(0, 1) \\ Y &= WX . \end{aligned} \tag{2}$$

Under these conditions, it can be shown that ( $\rightarrow$  II/3.2.27)

$$X \sim \mathcal{N}(0, 1) \quad \text{and} \quad Y \sim \mathcal{N}(0, 1) . \tag{3}$$

The linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \tag{4}$$

implies that, for bivariate normal random variables ( $\rightarrow$  II/4.2.1)  $X_1$  and  $X_2$ ,

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) . \tag{5}$$

any linear combination of  $X_1$  and  $X_2$  with non-zero coefficients

$$Z = aX_1 + bX_2, \quad a \neq 0, \quad b \neq 0 \tag{6}$$

follows a univariate normal distribution ( $\rightarrow$  II/4.2.6):

$$Z \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\sigma_{12} + b^2\sigma_2^2) . \tag{7}$$

Consider the sum of  $X$  and  $Y$  defined by (2):

$$Z = X + Y = aX + bY \quad \text{with} \quad a = b = 1 . \tag{8}$$

If  $X$  and  $Y$  were bivariate normally distributed ( $\rightarrow$  II/4.2.1), then this sum should be univariate normally distributed ( $\rightarrow$  II/3.2.1). However, this sum cannot be normally distributed, since

$$\Pr(X + Y = 0) = \frac{1}{2} \quad \text{and} \quad \Pr(X + Y = 2X) = \frac{1}{2} , \tag{9}$$

because

$$Y = \begin{cases} X, & \text{with probability } 1/2 \\ -X, & \text{with probability } 1/2 \end{cases}. \quad (10)$$

Thus,  $X$  and  $Y$  are not following a bivariate normal distribution ( $\rightarrow$  II/4.2.1). Therefore,  $X$  and  $Y$  defined by (2) and (1) constitute an example for two random variables ( $\rightarrow$  I/1.2.2) that are marginally normal ( $\rightarrow$  II/3.2.1), but not jointly normal ( $\rightarrow$  II/4.1.1). ■

#### Sources:

- Wikipedia (2024): “Misconceptions about the normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-11; URL: [https://en.wikipedia.org/wiki/Misconceptions\\_about\\_the\\_normal\\_distribution#A\\_symmetric\\_example](https://en.wikipedia.org/wiki/Misconceptions_about_the_normal_distribution#A_symmetric_example).

### 3.3 t-distribution

#### 3.3.1 Definition

**Definition:** Let  $Z$  and  $V$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following a standard normal distribution ( $\rightarrow$  II/3.2.3) and a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $\nu$  degrees of freedom ( $\rightarrow$  I/1.5.12), respectively:

$$\begin{aligned} Z &\sim \mathcal{N}(0, 1) \\ V &\sim \chi^2(\nu). \end{aligned} \quad (1)$$

Then, the ratio of  $Z$  to the square root of  $V$ , divided by the respective degrees of freedom, is said to be  $t$ -distributed with degrees of freedom  $\nu$ :

$$Y = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu). \quad (2)$$

The  $t$ -distribution is also called “Student’s  $t$ -distribution”, after William S. Gosset a.k.a. “Student”.

#### Sources:

- Wikipedia (2021): “Student’s  $t$ -distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-21; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Characterization](https://en.wikipedia.org/wiki/Student%27s_t-distribution#Characterization).

#### 3.3.2 Special case of multivariate $t$ -distribution

**Theorem:** The  $t$ -distribution ( $\rightarrow$  II/3.3.1) is a special case of the multivariate  $t$ -distribution ( $\rightarrow$  II/4.3.1) with number of variables  $n = 1$ , i.e. random vector ( $\rightarrow$  I/1.2.3)  $x \in \mathbb{R}$ , mean  $\mu = 0$  and covariance matrix  $\Sigma = 1$ .

**Proof:** The probability density function of the multivariate  $t$ -distribution ( $\rightarrow$  II/4.3.2) is

$$t(x; \mu, \Sigma, \nu) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu + n]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^{-(\nu+n)/2}. \quad (1)$$

Setting  $n = 1$ , such that  $x \in \mathbb{R}$ , as well as  $\mu = 0$  and  $\Sigma = 1$ , we obtain



$$\begin{aligned}
t(x; 0, 1, \nu) &= \sqrt{\frac{1}{(\nu\pi)^1|1|}} \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{1}{\nu}(x-0)^T 1^{-1}(x-0) \right]^{-(\nu+1)/2} \\
&= \sqrt{\frac{1}{\nu\pi}} \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{x^2}{\nu} \right]^{-(\nu+1)/2} \\
&= \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \left[ 1 + \frac{x^2}{\nu} \right]^{-\frac{\nu+1}{2}} .
\end{aligned} \tag{2}$$

which is equivalent to the probability density function of the t-distribution ( $\rightarrow$  II/3.3.5).

■

#### Sources:

- Wikipedia (2022): “Multivariate t-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-08-25; URL: [https://en.wikipedia.org/wiki/Multivariate\\_t-distribution#Derivation](https://en.wikipedia.org/wiki/Multivariate_t-distribution#Derivation).

### 3.3.3 Non-standardized t-distribution

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Student’s t-distribution ( $\rightarrow$  II/3.3.1) with  $\nu$  degrees of freedom. Then, the random variable ( $\rightarrow$  I/1.2.2)

$$Y = \sigma X + \mu \tag{1}$$

is said to follow a non-standardized t-distribution with non-centrality  $\mu$ , scale  $\sigma^2$  and degrees of freedom  $\nu$ :

$$Y \sim \text{nst}(\mu, \sigma^2, \nu) . \tag{2}$$

#### Sources:

- Wikipedia (2021): “Student’s t-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-20; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Generalized\\_Student's\\_t-distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution#Generalized_Student's_t-distribution).

### 3.3.4 Relationship to non-standardized t-distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a non-standardized t-distribution ( $\rightarrow$  II/3.3.3) with mean  $\mu$ , scale  $\sigma^2$  and degrees of freedom  $\nu$ :

$$X \sim \text{nst}(\mu, \sigma^2, \nu) . \tag{1}$$

Then, subtracting the mean and dividing by the square root of the scale results in a random variable ( $\rightarrow$  I/1.2.2) following a t-distribution ( $\rightarrow$  II/3.3.1) with degrees of freedom  $\nu$ :

$$Y = \frac{X - \mu}{\sigma} \sim t(\nu) . \tag{2}$$

**Proof:** The non-standardized t-distribution is a special case of the multivariate t-distribution ( $\rightarrow$  II/4.3.1) in which the mean vector and scale matrix are scalars:

$$X \sim \text{nst}(\mu, \sigma^2, \nu) \Rightarrow X \sim t(\mu, \sigma^2, \nu). \quad (3)$$

Therefore, we can apply the linear transformation theorem for the multivariate t-distribution for an  $n$ -dimensional random vector  $x$ :

$$x \sim t(\mu, \Sigma, \nu) \Rightarrow y = Ax + b \sim t(A\mu + b, A\Sigma A^T, \nu). \quad (4)$$

Comparing with equation (2), we have  $A = 1/\sigma$ ,  $b = -\mu/\sigma$  and the variable  $Y$  is distributed as:

$$\begin{aligned} Y &= \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma} \\ &\sim t\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \left(\frac{1}{\sigma}\right)^2 \sigma^2, \nu\right) \\ &= t(0, 1, \nu). \end{aligned} \quad (5)$$

Plugging  $\mu = 0$ ,  $\Sigma = 1$  and  $n = 1$  into the probability density function of the multivariate t-distribution ( $\rightarrow$  II/4.3.2),

$$p(x) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu + n]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right], \quad (6)$$

we get

$$p(x) = \sqrt{\frac{1}{\nu\pi}} \frac{\Gamma([\nu + 1]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{x^2}{\nu} \right] \quad (7)$$

which is the probability density function of Student's t-distribution ( $\rightarrow$  II/3.3.5) with  $\nu$  degrees of freedom. ■

### 3.3.5 Probability density function

**Theorem:** Let  $T$  be a random variable ( $\rightarrow$  I/1.2.2) following a t-distribution ( $\rightarrow$  II/3.3.1):

$$T \sim t(\nu). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $T$  is

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \cdot \sqrt{\nu\pi}} \cdot \left(\frac{t^2}{\nu} + 1\right)^{-\frac{\nu+1}{2}}. \quad (2)$$

**Proof:** A t-distributed random variable ( $\rightarrow$  II/3.3.1) is defined as the ratio of a standard normal random variable ( $\rightarrow$  II/3.2.3) and the square root of a chi-squared random variable ( $\rightarrow$  II/3.7.1), divided by its degrees of freedom ( $\rightarrow$  I/1.5.12)

$$X \sim \mathcal{N}(0, 1), Y \sim \chi^2(\nu) \Rightarrow T = \frac{X}{\sqrt{Y/\nu}} \sim t(\nu) \quad (3)$$

where  $X$  and  $Y$  are independent of each other ( $\rightarrow$  I/1.3.6).

The probability density function ( $\rightarrow$  II/3.2.10) of the standard normal distribution ( $\rightarrow$  II/3.2.3) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \quad (4)$$

and the probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3) is

$$f_Y(y) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot y^{\frac{\nu}{2}-1} \cdot e^{-\frac{y}{2}}. \quad (5)$$

Define the random variables  $T$  and  $W$  as functions of  $X$  and  $Y$

$$\begin{aligned} T &= X \cdot \sqrt{\frac{\nu}{Y}} \\ W &= Y, \end{aligned} \quad (6)$$

such that the inverse functions  $X$  and  $Y$  in terms of  $T$  and  $W$  are

$$\begin{aligned} X &= T \cdot \sqrt{\frac{W}{\nu}} \\ Y &= W. \end{aligned} \quad (7)$$

This implies the following Jacobian matrix and determinant:

$$\begin{aligned} J &= \begin{bmatrix} \frac{dX}{dT} & \frac{dX}{dW} \\ \frac{dY}{dT} & \frac{dY}{dW} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{W}{\nu}} & \frac{T}{2\nu\sqrt{W/\nu}} \\ 0 & 1 \end{bmatrix} \\ |J| &= \sqrt{\frac{W}{\nu}}. \end{aligned} \quad (8)$$

Because  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), the joint density ( $\rightarrow$  I/1.5.2) of  $X$  and  $Y$  is equal to the product ( $\rightarrow$  I/1.3.9) of the marginal densities ( $\rightarrow$  I/1.5.3):

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y). \quad (9)$$

With the probability density function of an invertible function ( $\rightarrow$  I/1.7.5), the joint density ( $\rightarrow$  I/1.5.2) of  $T$  and  $W$  can be derived as:

$$f_{T,W}(t, w) = f_{X,Y}(x, y) \cdot |J|. \quad (10)$$

Substituting (7) into (4) and (5), and then with (8) into (10), we get:

$$\begin{aligned} f_{T,W}(t, w) &= f_X\left(t \cdot \sqrt{\frac{w}{\nu}}\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\left(t \cdot \sqrt{\frac{w}{\nu}}\right)^2}{2}} \cdot \frac{1}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \sqrt{\frac{w}{\nu}} \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu+1}{2}-1} \cdot e^{-\frac{w}{2}\left(\frac{t^2}{\nu}+1\right)}. \end{aligned} \quad (11)$$

The marginal density ( $\rightarrow$  I/1.5.3) of  $T$  can now be obtained by integrating out ( $\rightarrow$  I/1.3.3)  $W$ :

$$\begin{aligned}
 f_T(t) &= \int_0^\infty f_{T,W}(t, w) \, dw \\
 &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \int_0^\infty w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)w\right] \, dw \\
 &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)\right]^{(\nu+1)/2}} \cdot \int_0^\infty \frac{\left[\frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)\right]^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \cdot w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)w\right] \, dw
 \end{aligned} \tag{12}$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution ( $\rightarrow$  II/3.4.7) with

$$a = \frac{\nu+1}{2} \quad \text{and} \quad b = \frac{1}{2}\left(\frac{t^2}{\nu} + 1\right), \tag{13}$$

and because a probability density function integrates to one ( $\rightarrow$  I/1.7.1), we finally have:

$$\begin{aligned}
 f_T(t) &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)\right]^{(\nu+1)/2}} \\
 &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \cdot \sqrt{\nu\pi}} \cdot \left(\frac{t^2}{\nu} + 1\right)^{-\frac{\nu+1}{2}}.
 \end{aligned} \tag{14}$$

■

#### Sources:

- Computation Empire (2021): “Student’s t Distribution: Derivation of PDF”; in: *You Tube*, retrieved on 2021-10-11; URL: <https://www.youtube.com/watch?v=6BraaGEVRY8>.

## 3.4 Gamma distribution

### 3.4.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to follow a gamma distribution with shape ( $\rightarrow$  I/1.5.11)  $a$  and rate ( $\rightarrow$  I/1.5.10)  $b$

$$X \sim \text{Gam}(a, b), \tag{1}$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\text{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx], \quad x > 0 \tag{2}$$

where  $a > 0$  and  $b > 0$ ,  $\Gamma(x)$  is the gamma function and the density is zero, if  $x \leq 0$ .

#### Sources:

- Koch, Karl-Rudolf (2007): “Gamma Distribution”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 47, eq. 2.172; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

### 3.4.2 Special case of Wishart distribution

**Theorem:** The gamma distribution ( $\rightarrow$  II/3.4.1) is a special case of the Wishart distribution ( $\rightarrow$  II/5.2.1) where the number of columns of the random matrix ( $\rightarrow$  I/1.2.4) is  $p = 1$ .

**Proof:** Let  $X$  be a  $p \times p$  positive-definite symmetric matrix, such that  $X$  follows a Wishart distribution ( $\rightarrow$  II/5.2.1):

$$Y \sim \mathcal{W}(V, n) . \quad (1)$$

Then,  $Y$  is described by the probability density function

$$p(Y) = \frac{1}{\Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2^{np}|V|^n}} \cdot |X|^{(n-p-1)/2} \cdot \exp\left[-\frac{1}{2}\text{tr}(V^{-1}X)\right] \quad (2)$$

where  $|A|$  is a matrix determinant,  $A^{-1}$  is a matrix inverse and  $\Gamma_p(x)$  is the multivariate gamma function of order  $p$ . If  $p = 1$ , then  $\Gamma_p(x) = \Gamma(x)$  is the ordinary gamma function,  $x = X$  and  $v = V$  are real numbers. Thus, the probability density function ( $\rightarrow$  I/1.7.1) of  $x$  can be developed as

$$\begin{aligned} p(x) &= \frac{1}{\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2^n v^n}} \cdot x^{(n-2)/2} \cdot \exp\left[-\frac{1}{2}\text{tr}(v^{-1}x)\right] \\ &= \frac{(2v)^{-n/2}}{\Gamma\left(\frac{n}{2}\right)} \cdot x^{n/2-1} \cdot \exp\left[-\frac{1}{2v}x\right] \end{aligned} \quad (3)$$

Finally, substituting  $a = \frac{n}{2}$  and  $b = \frac{1}{2v}$ , we get

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \quad (4)$$

which is the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7). ■

### 3.4.3 Standard gamma distribution

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to have a standard gamma distribution, if  $X$  follows a gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a > 0$  and rate  $b = 1$ :

$$X \sim \text{Gam}(a, 1) . \quad (1)$$

#### Sources:

- JoramSoch (2017): “Gamma-distributed random numbers”; in: *MACS – a new SPM toolbox for model assessment, comparison and selection*, retrieved on 2020-05-26; URL: [https://github.com/JoramSoch/MACS/blob/master/MD\\_gamrnd.m](https://github.com/JoramSoch/MACS/blob/master/MD_gamrnd.m); DOI: 10.5281/zenodo.845404.
- NIST/SEMATECH (2012): “Gamma distribution”; in: *e-Handbook of Statistical Methods*, ch. 1.3.6.6.11; URL: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda366b.htm>; DOI: 10.18434/M

### 3.4.4 Relationship to standard gamma distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a$  and rate  $b$ :

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the quantity  $Y = bX$  will have a standard gamma distribution ( $\rightarrow$  II/3.4.3) with shape  $a$  and rate 1:

$$Y = bX \sim \text{Gam}(a, 1) . \quad (2)$$

**Proof:** Note that  $Y$  is a function of  $X$

$$Y = g(X) = bX \quad (3)$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{b}Y . \quad (4)$$

Because  $b$  must be positive,  $g(X)$  is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function ( $\rightarrow$  I/1.8.3) as

$$F_Y(y) = \begin{cases} 0 , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) , & \text{if } y \in \mathcal{Y} \\ 1 , & \text{if } y > \max(\mathcal{Y}) . \end{cases} \quad (5)$$

The cumulative distribution function of the gamma-distributed ( $\rightarrow$  II/3.4.9)  $X$  is

$$F_X(x) = \int_{-\infty}^x \frac{b^a}{\Gamma(a)} t^{a-1} \exp[-bt] dt . \quad (6)$$

Applying (5) to (6), we have:

$$\begin{aligned} F_Y(y) &\stackrel{(5)}{=} F_X(g^{-1}(y)) \\ &\stackrel{(6)}{=} \int_{-\infty}^{y/b} \frac{b^a}{\Gamma(a)} t^{a-1} \exp[-bt] dt . \end{aligned} \quad (7)$$

Substituting  $s = bt$ , such that  $t = s/b$ , we obtain

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{b(y/b)} \frac{b^a}{\Gamma(a)} \left(\frac{s}{b}\right)^{a-1} \exp\left[-b\left(\frac{s}{b}\right)\right] d\left(\frac{s}{b}\right) \\ &= \int_{-\infty}^y \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}b} s^{a-1} \exp[-s] ds \\ &= \int_{-\infty}^y \frac{1}{\Gamma(a)} s^{a-1} \exp[-s] ds \end{aligned} \quad (8)$$

which is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the standard gamma distribution ( $\rightarrow$  II/3.4.3).

■

### 3.4.5 Relationship to standard gamma distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a$  and rate  $b$ :

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the quantity  $Y = bX$  will have a standard gamma distribution ( $\rightarrow$  II/3.4.3) with shape  $a$  and rate 1:

$$Y = bX \sim \text{Gam}(a, 1) . \quad (2)$$

**Proof:** Note that  $Y$  is a function of  $X$

$$Y = g(X) = bX \quad (3)$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{b}Y . \quad (4)$$

Because  $b$  must be positive,  $g(X)$  is strictly increasing and we can calculate the probability density function of a strictly increasing function ( $\rightarrow$  I/1.7.3) as

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (5)$$

where  $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$ . With the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we have

$$\begin{aligned} f_Y(y) &= \frac{b^a}{\Gamma(a)} [g^{-1}(y)]^{a-1} \exp[-b g^{-1}(y)] \cdot \frac{dg^{-1}(y)}{dy} \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{b}y\right)^{a-1} \exp\left[-b\left(\frac{1}{b}y\right)\right] \cdot \frac{d\left(\frac{1}{b}y\right)}{dy} \\ &= \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}} y^{a-1} \exp[-y] \cdot \frac{1}{b} \\ &= \frac{1}{\Gamma(a)} y^{a-1} \exp[-y] \end{aligned} \quad (6)$$

which is the probability density function ( $\rightarrow$  I/1.7.1) of the standard gamma distribution ( $\rightarrow$  II/3.4.3). ■

### 3.4.6 Scaling of a gamma random variable

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a$  and rate  $b$ :

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the quantity  $Y = cX$  will also be gamma-distributed with shape  $a$  and rate  $b/c$ :

$$Y = cX \sim \text{Gam}\left(a, \frac{b}{c}\right). \quad (2)$$

**Proof:** Note that  $Y$  is a function of  $X$

$$Y = g(X) = cX \quad (3)$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{c}Y. \quad (4)$$

Because the parameters of a gamma distribution are positive ( $\rightarrow$  II/3.4.1),  $c$  must also be positive. Thus,  $g(X)$  is strictly increasing and we can calculate the probability density function of a strictly increasing function ( $\rightarrow$  I/1.7.3) as

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}, & \text{if } y \in \mathcal{Y} \\ 0, & \text{if } y \notin \mathcal{Y} \end{cases} \quad (5)$$

The probability density function of the gamma-distributed ( $\rightarrow$  II/3.4.7)  $X$  is

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx]. \quad (6)$$

Applying (5) to (6), we have:

$$\begin{aligned} f_Y(y) &= \frac{b^a}{\Gamma(a)} [g^{-1}(y)]^{a-1} \exp[-bg^{-1}(y)] \frac{dg^{-1}(y)}{dy} \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{c}y\right)^{a-1} \exp\left[-b\left(\frac{1}{c}y\right)\right] \frac{d\left(\frac{1}{c}y\right)}{dy} \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{c}\right)^a \left(\frac{1}{c}\right)^{-1} y^{a-1} \exp\left[-\frac{b}{c}y\right] \frac{1}{c} \\ &= \frac{(b/c)^a}{\Gamma(a)} y^{a-1} \exp\left[-\frac{b}{c}y\right] \end{aligned} \quad (7)$$

which is the probability density function ( $\rightarrow$  I/1.7.1) of a gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a$  and rate  $b/c$ . ■

### 3.4.7 Probability density function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is



$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] . \quad (2)$$

**Proof:** This follows directly from the definition of the gamma distribution ( $\rightarrow$  II/3.4.1). ■

### 3.4.8 Moment-generating function

**Theorem:** Let  $X$  follow a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = \left(1 - \frac{t}{b}\right)^{-a} . \quad (2)$$

**Proof:** The moment-generating function of a random variable ( $\rightarrow$  I/1.9.1)  $X$  is defined as:

$$M_X(t) = \mathbb{E} [e^{tX}] , \quad t \in \mathbb{R} . \quad (3)$$

Applying the law of the unconscious statistician ( $\rightarrow$  I/1.12.9), we have:

$$M_X(t) = \int_{\mathbb{R}} e^{tx} \cdot f_X(x) dx . \quad (4)$$

With the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we have:

$$M_X(t) = \int_{\mathbb{R}} \exp[tx] \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx . \quad (5)$$

Now we summarize the two exponential functions inside the integral:

$$\begin{aligned} M_X(t) &= \int_{\mathbb{R}} \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] dx \\ &= \int_{\mathbb{R}} \frac{(b-t)^a}{(b-t)^a} \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] dx \\ &= \int_{\mathbb{R}} \frac{b^a}{(b-t)^a} \cdot \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] dx \\ &= \left(\frac{b}{b-t}\right)^a \int_{\mathbb{R}} \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] dx . \end{aligned} \quad (6)$$

The integrand is equal to the probability density function of a gamma distribution ( $\rightarrow$  II/3.4.7):

$$M_X(t) = \left(\frac{b}{b-t}\right)^a \int_{\mathbb{R}} \text{Gam}(x; a, b-t) dx . \quad (7)$$

Because the entire probability density integrates to one ( $\rightarrow$  I/1.7.1), we finally have:

$$M_X(t) = \left(\frac{b}{b-t}\right)^a = \left(\frac{b-t}{b}\right)^{-a} = \left(\frac{b}{b} - \frac{t}{b}\right)^{-a} = \left(1 - \frac{t}{b}\right)^{-a} . \quad (8)$$
■

### 3.4.9 Cumulative distribution function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} \quad (2)$$

where  $\Gamma(x)$  is the gamma function and  $\gamma(s, x)$  is the lower incomplete gamma function.

**Proof:** The probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) is:

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] . \quad (3)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$\begin{aligned} F_X(x) &= \int_0^x \text{Gam}(z; a, b) dz \\ &= \int_0^x \frac{b^a}{\Gamma(a)} z^{a-1} \exp[-bz] dz \\ &= \frac{b^a}{\Gamma(a)} \int_0^x z^{a-1} \exp[-bz] dz . \end{aligned} \quad (4)$$

Substituting  $t = bz$ , i.e.  $z = t/b$ , this becomes:

$$\begin{aligned} F_X(x) &= \frac{b^a}{\Gamma(a)} \int_{b \cdot 0}^{bx} \left(\frac{t}{b}\right)^{a-1} \exp\left[-b\left(\frac{t}{b}\right)\right] d\left(\frac{t}{b}\right) \\ &= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{b^{a-1}} \cdot \frac{1}{b} \int_0^{bx} t^{a-1} \exp[-t] dt \\ &= \frac{1}{\Gamma(a)} \int_0^{bx} t^{a-1} \exp[-t] dt . \end{aligned} \quad (5)$$

With the definition of the lower incomplete gamma function

$$\gamma(s, x) = \int_0^x t^{s-1} \exp[-t] dt , \quad (6)$$

we arrive at the final result given by equation (2):

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} . \quad (7)$$

■

#### Sources:

- Wikipedia (2020): “Incomplete gamma function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-29; URL: [https://en.wikipedia.org/wiki/Incomplete\\_gamma\\_function#Definition](https://en.wikipedia.org/wiki/Incomplete_gamma_function#Definition).

### 3.4.10 Quantile function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \begin{cases} -\infty , & \text{if } p = 0 \\ \gamma^{-1}(a, \Gamma(a) \cdot p)/b , & \text{if } p > 0 \end{cases} \quad (2)$$

where  $\gamma^{-1}(s, y)$  is the inverse of the lower incomplete gamma function  $\gamma(s, x)$

**Proof:** The cumulative distribution function of the gamma distribution ( $\rightarrow$  II/3.4.9) is:

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ \frac{\gamma(a, bx)}{\Gamma(a)} , & \text{if } x \geq 0 . \end{cases} \quad (3)$$

The quantile function ( $\rightarrow$  I/1.11.1)  $Q_X(p)$  is defined as the smallest  $x$ , such that  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \quad (4)$$

Thus, we have  $Q_X(p) = -\infty$ , if  $p = 0$ . When  $p > 0$ , it holds that ( $\rightarrow$  I/1.11.2)

$$Q_X(p) = F_X^{-1}(x) . \quad (5)$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{\gamma(a, bx)}{\Gamma(a)} \\ \Gamma(a) \cdot p &= \gamma(a, bx) \\ \gamma^{-1}(a, \Gamma(a) \cdot p) &= bx \\ x &= \frac{\gamma^{-1}(a, \Gamma(a) \cdot p)}{b} . \end{aligned} \quad (6)$$

■

#### Sources:

- Wikipedia (2020): “Incomplete gamma function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Incomplete\\_gamma\\_function#Definition](https://en.wikipedia.org/wiki/Incomplete_gamma_function#Definition).

### 3.4.11 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \frac{a}{b} . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx . \quad (3)$$

With the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), this reads:

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{1}{b} \cdot \frac{b^{a+1}}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] dx . \end{aligned} \quad (4)$$

Employing the relation  $\Gamma(x+1) = \Gamma(x) \cdot x$ , we have

$$E(X) = \int_0^\infty \frac{a}{b} \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] dx \quad (5)$$

and again using the density of the gamma distribution ( $\rightarrow$  II/3.4.7), we get

$$\begin{aligned} E(X) &= \frac{a}{b} \int_0^\infty \text{Gam}(x; a+1, b) dx \\ &= \frac{a}{b} . \end{aligned} \quad (6)$$

■

#### Sources:

- Turlapaty, Anish (2013): “Gamma random variable: mean & variance”; in: *YouTube*, retrieved on 2020-05-19; URL: <https://www.youtube.com/watch?v=Sy4wP-Y2dmA>.

#### 3.4.12 Median

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = \frac{1}{b} \gamma_a^{-1} \left( \frac{\Gamma(a)}{2} \right) \quad (2)$$

where  $\Gamma(x)$  is the gamma function and  $\gamma_s^{-1}(y)$  is the inverse function of the lower incomplete gamma function  $\gamma(s, x)$  in which the first argument is  $s$ .

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function ( $\rightarrow$  I/1.8.1) is 1/2:

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the gamma distribution ( $\rightarrow$  II/3.4.9) is

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} \quad (4)$$

where the lower incomplete gamma function  $\gamma(s, x)$  is given by

$$\gamma(s, x) = \int_0^x t^{s-1} \exp(-t) dt . \quad (5)$$

Thus, the inverse CDF, i.e. the quantile function ( $\rightarrow$  I/1.11.1), is

$$x = \frac{1}{b} \gamma_a^{-1}(p\Gamma(a)) \quad (6)$$

where  $\gamma_a^{-1}(y)$  is the inverse function of  $\gamma(a, x)$ . Setting  $p = 1/2$ , we obtain:

$$\text{median}(X) = \frac{1}{b} \gamma_a^{-1} \left( \frac{\Gamma(a)}{2} \right) . \quad (7)$$

■

#### Sources:

- Wikipedia (2025): “Gamma distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-10-24; URL: [https://en.wikipedia.org/wiki/Gamma\\_distribution#Median\\_approximations\\_and\\_bounds](https://en.wikipedia.org/wiki/Gamma_distribution#Median_approximations_and_bounds).

#### 3.4.13 Mode

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) = \begin{cases} \frac{a-1}{b} , & \text{if } a \geq 1 \\ 0 , & \text{if } a < 1 . \end{cases} \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) is:

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] . \quad (4)$$

The first two derivatives of this function are:

$$\begin{aligned} f'_X(x) &= \frac{df_X(x)}{dx} = \frac{b^a}{\Gamma(a)} (-bx^{a-1} \exp[-bx] + (a-1)x^{a-2} \exp[-bx]) \\ &= \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \left( \frac{a-1}{x} - b \right) \end{aligned} \quad (5)$$

$$\begin{aligned}
f_X''(x) &= \frac{d^2 f_X(x)}{dx^2} = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \left( -\frac{a-1}{x^2} \right) + \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \left( \frac{a-1}{x} - b \right)^2 \\
&= \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \left( \left( \frac{a-1}{x} - b \right)^2 - \frac{a-1}{x^2} \right).
\end{aligned} \tag{6}$$

We now calculate the root of the first derivative (5):

$$\begin{aligned}
f_X'(x) = 0 &= \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \left( \frac{a-1}{x} - b \right) \\
&\Leftrightarrow \\
0 &= \frac{a-1}{x} - b \\
bx &= a-1 \\
x &= \frac{a-1}{b}.
\end{aligned} \tag{7}$$

By plugging this value into the second derivative (6), we find:

$$\begin{aligned}
f_X''\left(\frac{a-1}{b}\right) &= \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{b}\right)^{a-1} \exp\left[-b\left(\frac{a-1}{b}\right)\right] \left( \left( \frac{a-1}{\frac{a-1}{b}} - b \right)^2 - \frac{a-1}{\left(\frac{a-1}{b}\right)^2} \right) \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{b}\right)^{a-1} \exp[-(a-1)] \left( (b-b)^2 - \frac{b^2}{a-1} \right) \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{b}\right)^{a-1} \exp[-(a-1)] \left( -\frac{b^2}{a-1} \right).
\end{aligned} \tag{8}$$

Thus, if  $a \geq 1$ , then  $f_X'((a-1)/b)$  is negative, such that  $f_X(x)$  reaches its maximum at  $x = (a-1)/b$ . However, if  $a < 1$ , this value is negative and not in the support of the random variable ( $\rightarrow$  II/3.4.1)  $X$ . In this case, the mode of  $X$  is 0, since  $f_X(x)$  tends towards infinity at  $x = 0$ :

$$\begin{aligned}
\lim_{\substack{x \rightarrow 0 \\ a < 1}} f_X(x) &= \lim_{\substack{x \rightarrow 0 \\ a < 1}} \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \\
&= \frac{b^a}{\Gamma(a)} \lim_{\substack{x \rightarrow 0 \\ a < 1}} x^{a-1} = \infty.
\end{aligned} \tag{9}$$

■

### 3.4.14 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b). \tag{1}$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \frac{a}{b^2} . \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) can be expressed in terms of expected values ( $\rightarrow$  I/1.13.4) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

The expected value of a gamma random variable ( $\rightarrow$  II/3.4.11) is

$$E(X) = \frac{a}{b} . \quad (4)$$

With the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), the expected value of a squared gamma random variable is

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{1}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] dx . \end{aligned} \quad (5)$$

Twice-applying the relation  $\Gamma(x+1) = \Gamma(x) \cdot x$ , we have

$$E(X^2) = \int_0^\infty \frac{a(a+1)}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a+2)} x^{(a+2)-1} \exp[-bx] dx \quad (6)$$

and again using the density of the gamma distribution ( $\rightarrow$  II/3.4.7), we get

$$\begin{aligned} E(X^2) &= \frac{a(a+1)}{b^2} \int_0^\infty \text{Gam}(x; a+2, b) dx \\ &= \frac{a^2 + a}{b^2} . \end{aligned} \quad (7)$$

Plugging (7) and (4) into (3), the variance of a gamma random variable finally becomes

$$\begin{aligned} \text{Var}(X) &= \frac{a^2 + a}{b^2} - \left(\frac{a}{b}\right)^2 \\ &= \frac{a}{b^2} . \end{aligned} \quad (8)$$

■

#### Sources:

- Turlapaty, Anish (2013): “Gamma random variable: mean & variance”; in: *YouTube*, retrieved on 2020-05-19; URL: <https://www.youtube.com/watch?v=Sy4wP-Y2dmA>.

### 3.4.15 Logarithmic expectation

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the expectation ( $\rightarrow$  I/1.12.1) of the natural logarithm of  $X$  is

$$\mathbb{E}(\ln X) = \psi(a) - \ln(b) \quad (2)$$

where  $\psi(x)$  is the digamma function.

**Proof:** Let  $Y = \ln(X)$ , such that  $\mathbb{E}(Y) = \mathbb{E}(\ln X)$  and consider the special case that  $b = 1$ . In this case, the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) is

$$f_X(x) = \frac{1}{\Gamma(a)} x^{a-1} \exp[-x] . \quad (3)$$

Multiplying this function with  $dx$ , we obtain

$$f_X(x) dx = \frac{1}{\Gamma(a)} x^a \exp[-x] \frac{dx}{x} . \quad (4)$$

Substituting  $y = \ln x$ , i.e.  $x = e^y$ , such that  $dx/dy = x$ , i.e.  $dx/x = dy$ , we get

$$\begin{aligned} f_Y(y) dy &= \frac{1}{\Gamma(a)} (e^y)^a \exp[-e^y] dy \\ &= \frac{1}{\Gamma(a)} \exp[ay - e^y] dy . \end{aligned} \quad (5)$$

Because  $f_Y(y)$  integrates to one, we have

$$\begin{aligned} 1 &= \int_{\mathbb{R}} f_Y(y) dy \\ 1 &= \int_{\mathbb{R}} \frac{1}{\Gamma(a)} \exp[ay - e^y] dy \\ \Gamma(a) &= \int_{\mathbb{R}} \exp[ay - e^y] dy . \end{aligned} \quad (6)$$

Note that the integrand in (6) is differentiable with respect to  $a$ :

$$\begin{aligned} \frac{d}{da} \exp[ay - e^y] dy &= y \exp[ay - e^y] dy \\ &\stackrel{(5)}{=} \Gamma(a) y f_Y(y) dy . \end{aligned} \quad (7)$$

Now we can calculate the expected value of  $Y = \ln(X)$ :



$$\begin{aligned}
E(Y) &= \int_{\mathbb{R}} y f_Y(y) dy \\
&\stackrel{(7)}{=} \frac{1}{\Gamma(a)} \int_{\mathbb{R}} \frac{d}{da} \exp[ay - e^y] dy \\
&= \frac{1}{\Gamma(a)} \frac{d}{da} \int_{\mathbb{R}} \exp[ay - e^y] dy \\
&\stackrel{(6)}{=} \frac{1}{\Gamma(a)} \frac{d}{da} \Gamma(a) \\
&= \frac{\Gamma'(a)}{\Gamma(a)} .
\end{aligned} \tag{8}$$

Using the derivative of a logarithmized function

$$\frac{d}{dx} \ln f(x) = \frac{f'(x)}{f(x)} \tag{9}$$

and the definition of the digamma function

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) , \tag{10}$$

we have

$$E(Y) = \psi(a) . \tag{11}$$

Finally, noting that  $1/b$  acts as a scaling parameter ( $\rightarrow$  II/3.4.4) on a gamma-distributed ( $\rightarrow$  II/3.4.1) random variable ( $\rightarrow$  I/1.2.2),

$$X \sim \text{Gam}(a, 1) \quad \Rightarrow \quad \frac{1}{b} X \sim \text{Gam}(a, b) , \tag{12}$$

and that a scaling parameter acts additively on the logarithmic expectation of a random variable,

$$E[\ln(cX)] = E[\ln(X) + \ln(c)] = E[\ln(X)] + \ln(c) , \tag{13}$$

it follows that

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad E(\ln X) = \psi(a) - \ln(b) . \tag{14}$$

■

#### Sources:

- whuber (2018): “What is the expected value of the logarithm of Gamma distribution?”; in: *StackExchange CrossValidated*, retrieved on 2020-05-25; URL: <https://stats.stackexchange.com/questions/370880/what-is-the-expected-value-of-the-logarithm-of-gamma-distribution>.

### 3.4.16 Expectation of $x \ln x$

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $(X \cdot \ln X)$  is

$$\mathbb{E}(X \ln X) = \frac{a}{b} [\psi(a+1) - \ln(b)] . \quad (2)$$

**Proof:** With the definition of the expected value ( $\rightarrow$  I/1.12.1), the law of the unconscious statistician ( $\rightarrow$  I/1.12.9) and the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we have:

$$\begin{aligned} \mathbb{E}(X \ln X) &= \int_0^\infty x \ln x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx \\ &= \frac{1}{\Gamma(a)} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{b} x^a \exp[-bx] dx \\ &= \frac{\Gamma(a+1)}{\Gamma(a)b} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] dx \end{aligned} \quad (3)$$

The integral now corresponds to the logarithmic expectation of a gamma distribution ( $\rightarrow$  II/3.4.15) with shape  $a+1$  and rate  $b$

$$\mathbb{E}(\ln Y) \quad \text{where} \quad Y \sim \text{Gam}(a+1, b) \quad (4)$$

which is given by ( $\rightarrow$  II/3.4.15)

$$\mathbb{E}(\ln Y) = \psi(a+1) - \ln(b) \quad (5)$$

where  $\psi(x)$  is the digamma function. Additionally employing the relation

$$\Gamma(x+1) = \Gamma(x) \cdot x \quad \Leftrightarrow \quad \frac{\Gamma(x+1)}{\Gamma(x)} = x , \quad (6)$$

the expression in equation (3) develops into:

$$\mathbb{E}(X \ln X) = \frac{a}{b} [\psi(a+1) - \ln(b)] . \quad (7)$$

■

#### Sources:

- gunes (2020): “What is the expected value of  $x \log(x)$  of the gamma distribution?”; in: *StackExchange CrossValidated*, retrieved on 2020-10-15; URL: <https://stats.stackexchange.com/questions/457357/what-is-the-expected-value-of-x-logx-of-the-gamma-distribution>.

**3.4.17 Differential entropy**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Gam}(a, b) \quad (1)$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  in nats is

$$h(X) = a + \ln \Gamma(a) + (1 - a) \cdot \psi(a) + \ln b . \quad (2)$$

**Proof:** The differential entropy ( $\rightarrow$  I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \quad (3)$$

To measure  $h(X)$  in nats, we set  $b = e$ , such that ( $\rightarrow$  I/1.12.1)

$$h(X) = -E[\ln p(x)] . \quad (4)$$

With the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), the differential entropy of  $X$  is:

$$\begin{aligned} h(X) &= -E \left[ \ln \left( \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \right) \right] \\ &= -E[a \cdot \ln b - \ln \Gamma(a) + (a-1) \ln x - bx] \\ &= -a \cdot \ln b + \ln \Gamma(a) - (a-1) \cdot E(\ln x) + b \cdot E(x) . \end{aligned} \quad (5)$$

Using the mean ( $\rightarrow$  II/3.4.11) and logarithmic expectation ( $\rightarrow$  II/3.4.15) of the gamma distribution ( $\rightarrow$  II/3.4.1)

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad E(X) = \frac{a}{b} \quad \text{and} \quad E(\ln X) = \psi(a) - \ln(b) , \quad (6)$$

the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  becomes:

$$\begin{aligned} h(X) &= -a \cdot \ln b + \ln \Gamma(a) - (a-1) \cdot (\psi(a) - \ln b) + b \cdot \frac{a}{b} \\ &= -a \cdot \ln b + \ln \Gamma(a) + (1-a) \cdot \psi(a) + a \cdot \ln b - \ln b + a \\ &= a + \ln \Gamma(a) + (1-a) \cdot \psi(a) - \ln b . \end{aligned} \quad (7)$$

■

**Sources:**

- Wikipedia (2021): “Gamma distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-14; URL: [https://en.wikipedia.org/wiki/Gamma\\_distribution#Information\\_entropy](https://en.wikipedia.org/wiki/Gamma_distribution#Information_entropy).

### 3.4.18 Kullback-Leibler divergence

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Assume two gamma distributions ( $\rightarrow$  II/3.4.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \text{Gam}(a_1, b_1) \\ Q : X &\sim \text{Gam}(a_2, b_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} . \quad (2)$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the gamma distributions ( $\rightarrow$  II/3.4.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{-\infty}^{+\infty} \text{Gam}(x; a_1, b_1) \ln \frac{\text{Gam}(x; a_1, b_1)}{\text{Gam}(x; a_2, b_2)} dx \\ &= \left\langle \ln \frac{\text{Gam}(x; a_1, b_1)}{\text{Gam}(x; a_2, b_2)} \right\rangle_{p(x)} . \end{aligned} \quad (4)$$

Using the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{b_1^{a_1}}{\Gamma(a_1)} x^{a_1-1} \exp[-b_1 x]}{\frac{b_2^{a_2}}{\Gamma(a_2)} x^{a_2-1} \exp[-b_2 x]} \right\rangle_{p(x)} \\ &= \left\langle \ln \left( \frac{b_1^{a_1}}{b_2^{a_2}} \cdot \frac{\Gamma(a_2)}{\Gamma(a_1)} \cdot x^{a_1-a_2} \cdot \exp[-(b_1 - b_2)x] \right) \right\rangle_{p(x)} \\ &= \langle a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \ln x - (b_1 - b_2) \cdot x \rangle_{p(x)} . \end{aligned} \quad (5)$$

Using the mean of the gamma distribution ( $\rightarrow$  II/3.4.11) and the expected value of a logarithmized gamma variate ( $\rightarrow$  II/3.4.15)

$$\begin{aligned} x \sim \text{Gam}(a, b) \quad \Rightarrow \quad \langle x \rangle &= \frac{a}{b} \quad \text{and} \\ \langle \ln x \rangle &= \psi(a) - \ln(b) , \end{aligned} \quad (6)$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot (\psi(a_1) - \ln(b_1)) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \\ &= a_2 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \psi(a_1) - (b_1 - b_2) \cdot \frac{a_1}{b_1} . \end{aligned} \quad (7)$$

Finally, combining the logarithms, we get:

$$\text{KL}[P || Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1}. \quad (8)$$

■

**Sources:**

- Penny, William D. (2001): “KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities”; in: *University College, London*; URL: <https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps>.

### 3.5 Exponential distribution

#### 3.5.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to be exponentially distributed with rate (or, inverse scale)  $\lambda$

$$X \sim \text{Exp}(\lambda), \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\text{Exp}(x; \lambda) = \lambda \exp[-\lambda x], \quad x \geq 0 \quad (2)$$

where  $\lambda > 0$ , and the density is zero, if  $x < 0$ .

**Sources:**

- Wikipedia (2020): “Exponential distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-08; URL: [https://en.wikipedia.org/wiki/Exponential\\_distribution#Definitions](https://en.wikipedia.org/wiki/Exponential_distribution#Definitions).

#### 3.5.2 Special case of gamma distribution

**Theorem:** The exponential distribution ( $\rightarrow$  II/3.5.1) is a special case of the gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $a = 1$  and rate  $b = \lambda$ .

**Proof:** The probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) is

$$\text{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx]. \quad (1)$$

Setting  $a = 1$  and  $b = \lambda$ , we obtain

$$\begin{aligned} \text{Gam}(x; 1, \lambda) &= \frac{\lambda^1}{\Gamma(1)} x^{1-1} \exp[-\lambda x] \\ &= \frac{x^0}{\Gamma(1)} \lambda \exp[-\lambda x] \\ &= \lambda \exp[-\lambda x] \end{aligned} \quad (2)$$

which is equivalent to the probability density function of the exponential distribution ( $\rightarrow$  II/3.5.3).

■

### 3.5.3 Probability density function

**Theorem:** Let  $X$  be a non-negative random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \lambda \exp[-\lambda x] . \quad (2)$$

**Proof:** This follows directly from the definition of the exponential distribution ( $\rightarrow$  II/3.5.1). ■

### 3.5.4 Moment-generating function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the moment generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = \frac{\lambda}{\lambda - t} \quad (2)$$

which is well-defined for  $t < \lambda$ .

**Proof:** Suppose  $X$  follows an exponential distribution ( $\rightarrow$  II/3.5.1) with rate  $\lambda$ ; that is,  $X \sim \text{Exp}(\lambda)$ . Then, the probability density function ( $\rightarrow$  II/3.5.3) is given by

$$f_X(x) = \lambda e^{-\lambda x} \quad (3)$$

and the moment-generating function ( $\rightarrow$  I/1.9.1) is defined as

$$M_X(t) = \text{E} [e^{tX}] . \quad (4)$$

Using the definition of expected value for continuous random variables ( $\rightarrow$  I/1.12.1), the moment-generating function of  $X$  is thus:

$$\begin{aligned} M_X(t) &= \int_0^{\infty} e^{tx} \cdot f_X(x) dx \\ &= \int_0^{\infty} e^{tx} \cdot \lambda e^{-\lambda x} dx \\ &= \int_0^{\infty} \lambda e^{x(t-\lambda)} dx \\ &= \frac{\lambda}{t-\lambda} e^{x(t-\lambda)} \Big|_{x=0}^{x=\infty} \\ &= \lim_{x \rightarrow \infty} \left[ \frac{\lambda}{t-\lambda} e^{x(t-\lambda)} \right] - \frac{\lambda}{t-\lambda} \\ &= \frac{\lambda}{t-\lambda} \left[ \lim_{x \rightarrow \infty} e^{x(t-\lambda)} - 1 \right] . \end{aligned} \quad (5)$$

Note that  $t$  cannot be equal to  $\lambda$ , else  $M_X(t)$  is undefined. Further, if  $t > \lambda$ , then  $\lim_{x \rightarrow \infty} e^{x(t-\lambda)} = \infty$ , which implies that  $M_X(t)$  diverges for  $t \geq \lambda$ . So, we must restrict the domain of  $M_X(t)$  to  $t < \lambda$ . Assuming this, we can further simplify (5):

$$\begin{aligned} M_X(t) &= \frac{\lambda}{t - \lambda} \left[ \lim_{x \rightarrow \infty} e^{x(t-\lambda)} - 1 \right] \\ &= \frac{\lambda}{t - \lambda} [0 - 1] \\ &= \frac{\lambda}{\lambda - t} . \end{aligned} \tag{6}$$

This completes the proof of (2). ■

### 3.5.5 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \tag{1}$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \tag{2}$$

**Proof:** The probability density function of the exponential distribution ( $\rightarrow$  II/3.5.3) is:

$$\text{Exp}(x; \lambda) = \begin{cases} 0 , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \tag{3}$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^x \text{Exp}(z; \lambda) dz . \tag{4}$$

If  $x < 0$ , we have:

$$F_X(x) = \int_{-\infty}^x 0 dz = 0 . \tag{5}$$

If  $x \geq 0$ , we have using (3):

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^0 \text{Exp}(z; \lambda) \, dz + \int_0^x \text{Exp}(z; \lambda) \, dz \\
&= \int_{-\infty}^0 0 \, dz + \int_0^x \lambda \exp[-\lambda z] \, dz \\
&= 0 + \lambda \left[ -\frac{1}{\lambda} \exp[-\lambda z] \right]_0^x \\
&= \lambda \left[ \left( -\frac{1}{\lambda} \exp[-\lambda x] \right) - \left( -\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\
&= 1 - \exp[-\lambda x] .
\end{aligned} \tag{6}$$

■

### 3.5.6 Quantile function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \tag{1}$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \begin{cases} -\infty , & \text{if } p = 0 \\ -\frac{\ln(1-p)}{\lambda} , & \text{if } p > 0 . \end{cases} \tag{2}$$

**Proof:** The cumulative distribution function of the exponential distribution ( $\rightarrow$  II/3.5.5) is:

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \tag{3}$$

The quantile function ( $\rightarrow$  I/1.11.1)  $Q_X(p)$  is defined as the smallest  $x$ , such that  $F_X(x) = p$ :

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \tag{4}$$

Thus, we have  $Q_X(p) = -\infty$ , if  $p = 0$ . When  $p > 0$ , it holds that ( $\rightarrow$  I/1.11.2)

$$Q_X(p) = F_X^{-1}(p) . \tag{5}$$

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= 1 - \exp[-\lambda x] \\
\exp[-\lambda x] &= 1 - p \\
-\lambda x &= \ln(1 - p) \\
x &= -\frac{\ln(1 - p)}{\lambda} .
\end{aligned} \tag{6}$$

■



### 3.5.7 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$\text{E}(X) = \frac{1}{\lambda} . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$\text{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx . \quad (3)$$

With the probability density function of the exponential distribution ( $\rightarrow$  II/3.5.3), this reads:

$$\begin{aligned} \text{E}(X) &= \int_0^{+\infty} x \cdot \lambda \exp(-\lambda x) \, dx \\ &= \lambda \int_0^{+\infty} x \cdot \exp(-\lambda x) \, dx . \end{aligned} \quad (4)$$

Using the following anti-derivative

$$\int x \cdot \exp(-\lambda x) \, dx = \left( -\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) , \quad (5)$$

the expected value becomes

$$\begin{aligned} \text{E}(X) &= \lambda \left[ \left( -\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \right]_0^{+\infty} \\ &= \lambda \left[ \lim_{x \rightarrow \infty} \left( -\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) - \left( -\frac{1}{\lambda} \cdot 0 - \frac{1}{\lambda^2} \right) \exp(-\lambda \cdot 0) \right] \\ &= \lambda \left[ 0 + \frac{1}{\lambda^2} \right] \\ &= \frac{1}{\lambda} . \end{aligned} \quad (6)$$

■

#### Sources:

- Koch, Karl-Rudolf (2007): “Expected Value”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 39, eq. 2.142a; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

### 3.5.8 Median

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = \frac{\ln 2}{\lambda} . \quad (2)$$

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function ( $\rightarrow$  I/1.8.1) is  $1/2$ :

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the exponential distribution ( $\rightarrow$  II/3.5.5) is

$$F_X(x) = 1 - \exp[-\lambda x], \quad x \geq 0 . \quad (4)$$

Thus, the inverse CDF is

$$x = -\frac{\ln(1-p)}{\lambda} \quad (5)$$

and setting  $p = 1/2$ , we obtain:

$$\text{median}(X) = -\frac{\ln(1 - \frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda} . \quad (6)$$

■

### 3.5.9 Mode

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) = 0 . \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the exponential distribution ( $\rightarrow$  II/3.5.3) is:

$$f_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ \lambda e^{-\lambda x} , & \text{if } x \geq 0 . \end{cases} \quad (4)$$

Since

$$f_X(0) = \lambda \quad (5)$$

and

$$0 < e^{-\lambda x} < 1 \quad \text{for any } x > 0, \quad (6)$$

it follows that

$$\text{mode}(X) = 0. \quad (7)$$

■

### 3.5.10 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda). \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \frac{1}{\lambda^2}. \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) of a random variable is defined as

$$\text{Var}(X) = E[(X - E(X))^2] \quad (3)$$

which, partitioned into expected values ( $\rightarrow$  I/1.13.4), reads:

$$\text{Var}(X) = E[X^2] - E[X]^2. \quad (4)$$

The expected value of the exponential distribution ( $\rightarrow$  II/3.5.7) is:

$$E[X] = \frac{1}{\lambda} \quad (5)$$

The second moment  $E[X^2]$  can be derived as follows:

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{+\infty} x^2 \cdot f_X(x) \, dx \\ &= \int_0^{+\infty} x^2 \cdot \lambda \exp(-\lambda x) \, dx \\ &= \lambda \int_0^{+\infty} x^2 \cdot \exp(-\lambda x) \, dx \end{aligned} \quad (6)$$

Using the following anti-derivative

$$\begin{aligned}
\int_0^{+\infty} x^2 \cdot \exp(-\lambda x) dx &= \left[ -\frac{1}{\lambda} x^2 \cdot \exp(-\lambda x) \right]_0^{+\infty} - \int_0^{+\infty} 2x \left( -\frac{1}{\lambda} \exp(-\lambda x) \right) dx \\
&= \left[ -\frac{1}{\lambda} x^2 \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left( \left[ \frac{1}{\lambda^2} 2x \cdot \exp(-\lambda x) \right]_0^{+\infty} - \int_0^{+\infty} 2 \left( \frac{1}{\lambda^2} \exp(-\lambda x) \right) dx \right) \\
&= \left[ -\frac{x^2}{\lambda} \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left( \left[ \frac{2x}{\lambda^2} \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left[ -\frac{2}{\lambda^3} \cdot \exp(-\lambda x) \right]_0^{+\infty} \right) \\
&= \left[ \left( -\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) \right]_0^{+\infty},
\end{aligned} \tag{7}$$

the second moment becomes

$$\begin{aligned}
E[X^2] &= \lambda \left[ \left( -\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) \right]_0^{+\infty} \\
&= \lambda \left[ \lim_{x \rightarrow \infty} \left( -\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) - \left( 0 - 0 - \frac{2}{\lambda^3} \right) \exp(-\lambda \cdot 0) \right] \\
&= \lambda \left[ 0 + \frac{2}{\lambda^3} \right] \\
&= \frac{2}{\lambda^2}.
\end{aligned} \tag{8}$$

Plugging (8) and (5) into (4), we have:

$$\begin{aligned}
\text{Var}(X) &= E[X^2] - E[X]^2 \\
&= \frac{2}{\lambda^2} - \left( \frac{1}{\lambda} \right)^2 \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}.
\end{aligned} \tag{9}$$

■

#### Sources:

- Taboga, Marco (2023): “Exponential distribution”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2023-01-23; URL: <https://www.statlect.com/probability-distributions/exponential-distribution>.
- Wikipedia (2023): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-01-23; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

#### 3.5.11 Skewness

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an exponential distribution ( $\rightarrow$  II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then the skewness ( $\rightarrow$  I/1.14.1) of  $X$  is

$$\text{Skew}(X) = 2 . \quad (2)$$

**Proof:** To compute the skewness of  $X$ , we partition the skewness into expected values ( $\rightarrow$  I/1.14.3)

$$\text{Skew}(X) = \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \quad (3)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $X$ , respectively. Since  $X$  follows an exponential distribution ( $\rightarrow$  II/3.5.1), the mean ( $\rightarrow$  II/3.5.7) of  $X$  is given by

$$\mu = E(X) = \frac{1}{\lambda} \quad (4)$$

and the standard deviation ( $\rightarrow$  II/3.5.10) of  $X$  is given by

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} . \quad (5)$$

Substituting (4) and (5) into (3) gives:

$$\begin{aligned} \text{Skew}(X) &= \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \\ &= \frac{E(X^3)}{\sigma^3} - \frac{3\mu\sigma^2 + \mu^3}{\sigma^3} \\ &= \frac{E(X^3)}{\left(\frac{1}{\lambda}\right)^3} - \frac{3\left(\frac{1}{\lambda}\right)\left(\frac{1}{\lambda}\right)^2 + \left(\frac{1}{\lambda}\right)^3}{\left(\frac{1}{\lambda}\right)^3} \\ &= \lambda^3 \cdot E(X^3) - \frac{\frac{3}{\lambda^3} + \frac{1}{\lambda^3}}{\frac{1}{\lambda^3}} \\ &= \lambda^3 \cdot E(X^3) - 4 . \end{aligned} \quad (6)$$

Thus, the remaining work is to compute  $E(X^3)$ . To do this, we use the moment-generating function of the exponential distribution ( $\rightarrow$  II/3.5.4) to calculate

$$E(X^3) = M_X'''(0) \quad (7)$$

based on the relationship between raw moment and moment-generating function ( $\rightarrow$  I/1.21.2). First, we differentiate the moment-generating function of the exponential distribution ( $\rightarrow$  II/3.5.4)

$$M_X(t) = \frac{\lambda}{\lambda - t} = \lambda(\lambda - t)^{-1} \quad (8)$$

with respect to  $t$ . Using the chain rule gives:

$$\begin{aligned} M_X'(t) &= -1 \cdot \lambda(\lambda - t)^{-2} \cdot (-1) \\ &= \lambda(\lambda - t)^{-2} . \end{aligned} \quad (9)$$

We continue using the chain rule to obtain the second derivative:

$$\begin{aligned} M_X''(t) &= -2 \cdot \lambda(\lambda - t)^{-3} \cdot (-1) \\ &= 2\lambda(\lambda - t)^{-3} . \end{aligned} \tag{10}$$

Finally, one more application of the chain rule gives us the third derivative:

$$\begin{aligned} M_X'''(t) &= -3 \cdot 2\lambda(\lambda - t)^{-4} \cdot (-1) \\ &= 6\lambda(\lambda - t)^{-4} \\ &= \frac{6\lambda}{(\lambda - t)^4} . \end{aligned} \tag{11}$$

Applying (7), together with (11), yields

$$\begin{aligned} E(X^3) &= M_X'''(0) \\ &= \frac{6\lambda}{(\lambda - 0)^4} \\ &= \frac{6\lambda}{\lambda^4} \\ &= \frac{6}{\lambda^3} . \end{aligned} \tag{12}$$

We now substitute (12) into (6), giving

$$\begin{aligned} \text{Skew}(X) &= \lambda^3 \cdot E(X^3) - 4 \\ &= \lambda^3 \cdot \left( \frac{6}{\lambda^3} \right) - 4 \\ &= 6 - 4 \\ &= 2 . \end{aligned} \tag{13}$$

This completes the proof of (2). ■

## 3.6 Log-normal distribution

### 3.6.1 Definition

**Definition:** Let  $\ln X$  be a random variable ( $\rightarrow$  I/1.2.2) following a normal distribution ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$  (or, standard deviation  $\sigma$ ):

$$Y = \ln(X) \sim \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the exponential function of  $Y$  is said to have a log-normal distribution with location parameter  $\mu$  and scale parameter  $\sigma$

$$X = \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2) \tag{2}$$

where  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ .

**Sources:**

- Wikipedia (2022): “Log-normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-02-07; URL: [https://en.wikipedia.org/wiki/Log-normal\\_distribution](https://en.wikipedia.org/wiki/Log-normal_distribution).

### 3.6.2 Probability density function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is given by:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] . \quad (2)$$

**Proof:** A log-normally distributed random variable ( $\rightarrow$  II/3.6.1) is defined as the exponential function of a normal random variable ( $\rightarrow$  II/3.2.1):

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad X = \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (3)$$

The probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) is

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} \right] . \quad (4)$$

Writing  $X$  as a function of  $Y$ , we have

$$X = g(Y) = \exp(Y) \quad (5)$$

with the inverse function

$$Y = g^{-1}(X) = \ln(X) . \quad (6)$$

Because the derivative of  $\exp(Y)$  is always positive,  $g(Y)$  is strictly increasing and we can calculate the probability density function of a strictly increasing function ( $\rightarrow$  I/1.7.3) as

$$f_X(x) = \begin{cases} f_Y(g^{-1}(x)) \frac{dg^{-1}(x)}{dx} , & \text{if } x \in \mathcal{X} \\ 0 , & \text{if } x \notin \mathcal{X} \end{cases} \quad (7)$$

where  $\mathcal{X} = \{x = g(y) : y \in \mathcal{Y}\}$ . With the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), we have

$$\begin{aligned}
f_X(x) &= f_Y(g^{-1}(x)) \cdot \frac{dg^{-1}(x)}{dx} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{g^{-1}(x) - \mu}{\sigma} \right)^2 \right] \cdot \frac{dg^{-1}(x)}{dx} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{(\ln x) - \mu}{\sigma} \right)^2 \right] \cdot \frac{d(\ln x)}{dx} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{\ln x - \mu}{\sigma} \right)^2 \right] \cdot \frac{1}{x} \\
&= \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right]
\end{aligned} \tag{8}$$

which is the probability density function ( $\rightarrow$  I/1.7.1) of the log-normal distribution ( $\rightarrow$  II/3.6.1). ■

#### Sources:

- Taboga, Marco (2021): “Log-normal distribution”; in: *Lectures on probability and statistics*, retrieved on 2022-02-13; URL: <https://www.statlect.com/probability-distributions/log-normal-distribution>.

### 3.6.3 Cumulative distribution function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \right] \tag{2}$$

where  $\operatorname{erf}(x)$  is the error function defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt . \tag{3}$$

**Proof:** The probability density function of the log-normal distribution ( $\rightarrow$  II/3.6.2) is:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp \left[ -\left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right)^2 \right] . \tag{4}$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:



$$\begin{aligned}
F_X(x) &= \int_{-\infty}^x \ln \mathcal{N}(z; \mu, \sigma^2) dz \\
&= \int_{-\infty}^x \frac{1}{z\sigma\sqrt{2\pi}} \cdot \exp \left[ - \left( \frac{\ln z - \mu}{\sqrt{2}\sigma} \right)^2 \right] dz \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \frac{1}{z} \cdot \exp \left[ - \left( \frac{\ln z - \mu}{\sqrt{2}\sigma} \right)^2 \right] dz .
\end{aligned} \tag{5}$$

From this point forward, the proof is similar to the derivation of the cumulative distribution function for the normal distribution ( $\rightarrow$  II/3.2.12). Substituting  $t = (\ln z - \mu)/(\sqrt{2}\sigma)$ , i.e.  $\ln z = \sqrt{2}\sigma t + \mu$ ,  $z = \exp(\sqrt{2}\sigma t + \mu)$  this becomes:

$$\begin{aligned}
F_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(\ln x - \mu)/(\sqrt{2}\sigma)} \frac{1}{\exp(\sqrt{2}\sigma t + \mu)} \cdot \exp(-t^2) d \left[ \exp(\sqrt{2}\sigma t + \mu) \right] \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{\ln x - \mu}{\sqrt{2}\sigma}} \frac{1}{\exp(\sqrt{2}\sigma t + \mu)} \cdot \exp(-t^2) \cdot \exp(\sqrt{2}\sigma t + \mu) dt \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{\ln x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^0 \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{\ln x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\
&= \frac{1}{\sqrt{\pi}} \int_0^{\infty} \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{\ln x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt .
\end{aligned} \tag{6}$$

Applying (3) to (6), we have:

$$\begin{aligned}
F_X(x) &= \frac{1}{2} \lim_{x \rightarrow \infty} \operatorname{erf}(x) + \frac{1}{2} \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \\
&= \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \\
&= \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \right] .
\end{aligned} \tag{7}$$

■

#### Sources:

- skdhfgeq2134 (2015): “How to derive the cdf of a lognormal distribution from its pdf”; in: *StackExchange*, retrieved on 2022-06-29; URL: <https://stats.stackexchange.com/questions/151398/how-to-derive-t-151404#151404>.

#### 3.6.4 Quantile function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the quantile function ( $\rightarrow$  I/1.11.1) of  $X$  is

$$Q_X(p) = \exp(\mu + \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p - 1)) \quad (2)$$

where  $\operatorname{erf}^{-1}(x)$  is the inverse error function.

**Proof:** The cumulative distribution function of the log-normal distribution ( $\rightarrow$  II/3.6.3) is:

$$F_X(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \right]. \quad (3)$$

From this point forward, the proof is similar to the derivation of the quantile function for the normal distribution ( $\rightarrow$  II/3.2.15). Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF ( $\rightarrow$  I/1.11.2):

$$Q_X(p) = F_X^{-1}(x). \quad (4)$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \right] \\ 2p - 1 &= \operatorname{erf} \left( \frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \\ \operatorname{erf}^{-1}(2p - 1) &= \frac{\ln x - \mu}{\sqrt{2}\sigma} \\ x &= \exp(\mu + \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p - 1)). \end{aligned} \quad (5)$$

■

#### Sources:

- Wikipedia (2022): “Log-normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-08; URL: [https://en.wikipedia.org/wiki/Log-normal\\_distribution#Mode,\\_median,\\_quantiles](https://en.wikipedia.org/wiki/Log-normal_distribution#Mode,_median,_quantiles).

### 3.6.5 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2). \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$\mathbb{E}(X) = \exp \left( \mu + \frac{1}{2}\sigma^2 \right). \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$\mathbb{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx. \quad (3)$$

With the probability density function of the log-normal distribution ( $\rightarrow$  II/3.6.2), this is:

$$\begin{aligned}
E(X) &= \int_0^{+\infty} x \cdot \frac{1}{x\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{+\infty} \exp\left[-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2}\right] dx .
\end{aligned} \tag{4}$$

Substituting  $z = \frac{\ln x - \mu}{\sigma}$ , i.e.  $x = \exp(\mu + \sigma z)$ , we have:

$$\begin{aligned}
E(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{(-\infty - \mu)/(\sigma)}^{(\ln x - \mu)/(\sigma)} \exp\left(-\frac{1}{2}z^2\right) d[\exp(\mu + \sigma z)] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2\right) \sigma \exp(\mu + \sigma z) dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2 + \sigma z + \mu\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 2\sigma z - 2\mu)\right] dz .
\end{aligned} \tag{5}$$

Now multiplying with  $\exp(\frac{1}{2}\sigma^2)$  and  $\exp(-\frac{1}{2}\sigma^2)$ , we have:

$$\begin{aligned}
E(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 2\sigma z + \sigma^2 - 2\mu - \sigma^2)\right] dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 2\sigma z + \sigma^2)\right] \exp\left(\mu + \frac{1}{2}\sigma^2\right) dz \\
&= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z - \sigma)^2\right] dz .
\end{aligned} \tag{6}$$

The probability density function of a normal distribution ( $\rightarrow$  II/3.2.10) is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \tag{7}$$

and, with unit variance  $\sigma^2 = 1$ , this reads:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x - \mu)^2\right] . \tag{8}$$

Using the definition of the probability density function ( $\rightarrow$  I/1.7.1), we get

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x - \mu)^2\right] dx = 1 \tag{9}$$

and applying (9) to (6), we have:

$$E(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) . \tag{10}$$

■

**Sources:**

- Taboga, Marco (2022): “Log-normal distribution”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-10-01; URL: <https://www.statlect.com/probability-distributions/log-normal-distribution>.

**3.6.6 Median**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = e^\mu . \quad (2)$$

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function is  $1/2$ :

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the lognormal distribution ( $\rightarrow$  II/3.6.3) is

$$F_X(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{\ln(x) - \mu}{\sigma\sqrt{2}} \right) \right] \quad (4)$$

where  $\text{erf}(x)$  is the error function defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt . \quad (5)$$

Thus, the inverse cumulative distribution function is

$$\begin{aligned} \ln(x) &= \sigma\sqrt{2} \cdot \text{erf}^{-1}(2p - 1) + \mu \\ x &= \exp \left[ \sigma\sqrt{2} \cdot \text{erf}^{-1}(2p - 1) + \mu \right] \end{aligned} \quad (6)$$

where  $\text{erf}^{-1}(x)$  is the inverse error function. Setting  $p = 1/2$ , we obtain:

$$\begin{aligned} \ln [\text{median}(X)] &= \sigma\sqrt{2} \cdot \text{erf}^{-1}(0) + \mu \\ \text{median}(X) &= e^\mu . \end{aligned} \quad (7)$$

■

**3.6.7 Mode**

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) = e^{(\mu - \sigma^2)} . \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the log-normal distribution ( $\rightarrow$  II/3.6.2) is:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] . \quad (4)$$

The first two derivatives of this function are:

$$f'_X(x) = -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \cdot \left( 1 + \frac{\ln x - \mu}{\sigma^2} \right) \quad (5)$$

$$\begin{aligned} f''_X(x) &= \frac{1}{\sqrt{2\pi}\sigma^2 x^3} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \cdot (\ln x - \mu) \cdot \left( 1 + \frac{\ln x - \mu}{\sigma^2} \right) \\ &+ \frac{\sqrt{2}}{\sqrt{\pi}x^3} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \cdot \left( 1 + \frac{\ln x - \mu}{\sigma^2} \right) \\ &- \frac{1}{\sqrt{2\pi}\sigma^2 x^3} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] . \end{aligned} \quad (6)$$

We now calculate the root of the first derivative (5):

$$\begin{aligned} f'_X(x) = 0 &= -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \cdot \left( 1 + \frac{\ln x - \mu}{\sigma^2} \right) \\ -1 &= \frac{\ln x - \mu}{\sigma^2} \\ x &= e^{(\mu - \sigma^2)} . \end{aligned} \quad (7)$$

By plugging this value into the second derivative (6),

$$\begin{aligned} f''_X(e^{(\mu - \sigma^2)}) &= \frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \cdot \exp \left[ -\frac{\sigma^2}{2} \right] \cdot (\sigma^2) \cdot (0) \\ &+ \frac{\sqrt{2}}{\sqrt{\pi}(e^{(\mu - \sigma^2)})^3} \cdot \exp \left[ -\frac{\sigma^2}{2} \right] \cdot (0) \\ &- \frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \cdot \exp \left[ -\frac{\sigma^2}{2} \right] \\ &= -\frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \cdot \exp \left[ -\frac{\sigma^2}{2} \right] < 0 , \end{aligned} \quad (8)$$

we confirm that it is a maximum, showing that

$$\text{mode}(X) = e^{(\mu - \sigma^2)} . \quad (9)$$

■

#### Sources:

- Wikipedia (2022): “Log-normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-02-12; URL: [https://en.wikipedia.org/wiki/Log-normal\\_distribution#Mode](https://en.wikipedia.org/wiki/Log-normal_distribution#Mode).
- Mdoc (2015): “Mode of lognormal distribution”; in: *Mathematics Stack Exchange*, retrieved on 2022-02-12; URL: <https://math.stackexchange.com/questions/1321221/mode-of-lognormal-distribution/1321626>.

### 3.6.8 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2). \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) . \quad (2)$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) of a random variable is defined as

$$\text{Var}(X) = \text{E}[(X - \text{E}(X))^2] \quad (3)$$

which, partitioned into expected values ( $\rightarrow$  I/1.13.4), reads:

$$\text{Var}(X) = \text{E}[X^2] - \text{E}[X]^2 . \quad (4)$$

The expected value of the log-normal distribution ( $\rightarrow$  II/3.6.5) is:

$$\text{E}[X] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (5)$$

The second moment  $\text{E}[X^2]$  can be derived as follows:

$$\begin{aligned} \text{E}[X^2] &= \int_{-\infty}^{+\infty} x^2 \cdot f_X(x) \, dx \\ &= \int_0^{+\infty} x^2 \cdot \frac{1}{x\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2}\right] \, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{+\infty} x \cdot \exp\left[-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2}\right] \, dx \end{aligned} \quad (6)$$

Substituting  $z = \frac{\ln x - \mu}{\sigma}$ , i.e.  $x = \exp(\mu + \sigma z)$ , we have:

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{(-\infty-\mu)/(\sigma)}^{(\ln x - \mu)/(\sigma)} \exp(\mu + \sigma z) \exp\left(-\frac{1}{2}z^2\right) d[\exp(\mu + \sigma z)] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2\right) \sigma \exp(2\mu + 2\sigma z) dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 4\sigma z - 4\mu)\right] dz
\end{aligned} \tag{7}$$

Now multiplying the integrand by  $\exp(2\sigma^2)$  and  $\exp(-2\sigma^2)$ , this becomes:

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 4\sigma z + 4\sigma^2 - 4\sigma^2 - 4\mu)\right] dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(z^2 - 4\sigma z + 4\sigma^2)\right] \exp(2\sigma^2 + 2\mu) dz \\
&= \exp(2\sigma^2 + 2\mu) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z - 2\sigma)^2\right] dz
\end{aligned} \tag{8}$$

The probability density function of a normal distribution ( $\rightarrow$  II/3.2.10) is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \tag{9}$$

and, with  $\mu = 2\sigma$  and unit variance, this reads:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x - 2\sigma)^2\right]. \tag{10}$$

Using the definition of the probability density function ( $\rightarrow$  I/1.7.1), we get

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x - 2\sigma)^2\right] dx = 1 \tag{11}$$

and applying (11) to (8), we have:

$$E[X]^2 = \exp(2\sigma^2 + 2\mu). \tag{12}$$

Finally, plugging (12) and (5) into (4), we have:

$$\begin{aligned}
\text{Var}(X) &= E[X^2] - E[X]^2 \\
&= \exp(2\sigma^2 + 2\mu) - \left[\exp\left(\mu + \frac{1}{2}\sigma^2\right)\right]^2 \\
&= \exp(2\sigma^2 + 2\mu) - \exp(2\mu + \sigma^2).
\end{aligned} \tag{13}$$

#### Sources:

- Taboga, Marco (2022): “Log-normal distribution”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-10-01; URL: <https://www.statlect.com/probability-distributions/log-normal-distribution>.
- Wikipedia (2022): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-01; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

■

### 3.6.9 Product of independent log-normals

**Theorem:** Let  $X_1, \dots, X_n$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following log-normal distributions ( $\rightarrow$  II/3.6.1):

$$X_i \sim \ln \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, n. \quad (1)$$

Then, the product of these random variables also follows a log-normal distribution ( $\rightarrow$  II/3.6.1)

$$Z = \prod_{i=1}^n X_i \sim \mathcal{N}(\mu, \sigma^2) \quad (2)$$

where the log-normal ( $\rightarrow$  II/3.6.1) parameters are given by

$$\mu = \sum_{i=1}^n \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^n \sigma_i^2. \quad (3)$$

**Proof:** A random variable follows a log-normal distribution, if and only if its natural logarithm follows a normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad \ln X \sim \mathcal{N}(\mu, \sigma^2). \quad (4)$$

Thus, from (1), we have

$$Y_i = \ln X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad (5)$$

and from (2), we have

$$\ln Z = \ln \left( \prod_{i=1}^n X_i \right) = \sum_{i=1}^n \ln X_i = \sum_{i=1}^n Y_i. \quad (6)$$

This means that the logarithm of the product of independent log-normal ( $\rightarrow$  II/3.6.1) random variables is a sum of independent normal ( $\rightarrow$  II/3.2.1) random variables. Because any linear combination of independent normal random variables is again normally distributed ( $\rightarrow$  II/3.2.26), this sum follows a normal distribution. Combining this with (6) and (5), we have:

$$\ln Z = \sum_{i=1}^n Y_i \sim \mathcal{N} \left( \sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2 \right). \quad (7)$$

If a random variable follows a normal distribution, then its exponential follows a log-normal distribution with the same parameters ( $\rightarrow$  II/3.6.1):

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2). \quad (8)$$

Thus, from (7), we have

$$Z = \exp(\ln Z) \sim \ln \mathcal{N} \left( \sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2 \right) \quad (9)$$

which is equivalent to (2) and (3). ■



**Sources:**

- drhab (2018): “If  $X_i$  is a log-normal r.v., show product of all  $X_i$  is also a log-normal r.v.”; in: *StackExchange Mathematics*, retrieved on 2024-11-15; URL: <https://math.stackexchange.com/a/2838307/480910>.

**3.6.10 Geometric mean of independent log-normals**

**Theorem:** Let  $X_1, \dots, X_n$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following log-normal distributions ( $\rightarrow$  II/3.6.1):

$$X_i \sim \ln \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, n. \quad (1)$$

Then, the geometric mean of these random variables also follows a log-normal distribution ( $\rightarrow$  II/3.6.1):

$$Z = \sqrt[n]{\prod_{i=1}^n X_i} \sim \mathcal{N}(\mu, \sigma^2) \quad (2)$$

where the log-normal ( $\rightarrow$  II/3.6.1) parameters are given by

$$\mu = \frac{1}{n} \sum_{i=1}^n \mu_i \quad \text{and} \quad \sigma^2 = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2. \quad (3)$$

**Proof:** A random variable follows a log-normal distribution, if and only if its natural logarithm follows a normal distribution ( $\rightarrow$  II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad \ln X \sim \mathcal{N}(\mu, \sigma^2). \quad (4)$$

Thus, from (1), we have

$$Y_i = \ln X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad (5)$$

and from (2), we have

$$\ln Z = \ln \left( \sqrt[n]{\prod_{i=1}^n X_i} \right) = \frac{1}{n} \sum_{i=1}^n \ln X_i = \frac{1}{n} \sum_{i=1}^n Y_i. \quad (6)$$

This means that the logarithm of the geometric mean of independent log-normal ( $\rightarrow$  II/3.6.1) random variables is the arithmetic mean of independent normal ( $\rightarrow$  II/3.2.1) random variables. This average, like any linear combination of independent normal random variables, is again normally distributed ( $\rightarrow$  II/3.2.26). Thus, combining (6) and (5), we have:

$$\ln Z = \frac{1}{n} \sum_{i=1}^n Y_i \sim \mathcal{N} \left( \frac{1}{n} \sum_{i=1}^n \mu_i, \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \right). \quad (7)$$

If a random variable [follows a normal distribution, then its exponential follows a log-normal distribution with the same parameters]:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2). \quad (8)$$

Thus, from (7), we have

$$Z = \exp(\ln Z) \sim \ln \mathcal{N} \left( \frac{1}{n} \sum_{i=1}^n \mu_i, \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \right) \quad (9)$$

which is equivalent to (2) and (3). ■

#### Sources:

- Probability Fact (2022): “The geometric mean of independent log-normal random variables has a log-normal distribution”; in: *X*, retrieved on 2024-11-22; URL: <https://x.com/ProbFact/status/1592989704646848512>.

## 3.7 Chi-squared distribution

### 3.7.1 Definition

**Definition:** Let  $X_1, \dots, X_k$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) where each of them is following a standard normal distribution ( $\rightarrow$  II/3.2.3):

$$X_i \sim \mathcal{N}(0, 1) \quad \text{for } i = 1, \dots, n. \quad (1)$$

Then, the sum of their squares follows a chi-squared distribution with  $k$  degrees of freedom:

$$Y = \sum_{i=1}^k X_i^2 \sim \chi^2(k) \quad \text{where } k > 0. \quad (2)$$

The probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3) with  $k$  degrees of freedom is

$$\chi^2(x; k) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} \exp[-x/2] \quad (3)$$

where  $k > 0$  and the density is zero, if  $x \leq 0$ .

#### Sources:

- Wikipedia (2020): “Chi-square distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-12; URL: [https://en.wikipedia.org/wiki/Chi-square\\_distribution#Definitions](https://en.wikipedia.org/wiki/Chi-square_distribution#Definitions).
- Robert V. Hogg, Joseph W. McKean, Allen T. Craig (2018): “The Chi-Squared-Distribution”; in: *Introduction to Mathematical Statistics*, Pearson, Boston, 2019, p. 178, eq. 3.3.7; URL: <https://www.pearson.com/store/p/introduction-to-mathematical-statistics/P100000843744>.

### 3.7.2 Special case of gamma distribution

**Theorem:** The chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $k$  degrees of freedom is a special case of the gamma distribution ( $\rightarrow$  II/3.4.1) with shape  $k/2$  and rate  $1/2$ :

$$X \sim \text{Gam} \left( \frac{k}{2}, \frac{1}{2} \right) \quad \Rightarrow \quad X \sim \chi^2(k). \quad (1)$$

**Proof:** The probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) for  $x > 0$ , where  $\alpha$  is the shape parameter and  $\beta$  is the rate parameter, is as follows:

$$\text{Gam}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp[-\beta x] \quad (2)$$

If we let  $\alpha = k/2$  and  $\beta = 1/2$ , we obtain

$$\text{Gam}\left(x; \frac{k}{2}, \frac{1}{2}\right) = \frac{x^{k/2-1} \exp[-x/2]}{\Gamma(k/2) 2^{k/2}} = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} \exp[-x/2] \quad (3)$$

which is equivalent to the probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3). ■

### 3.7.3 Probability density function

**Theorem:** Let  $Y$  be a random variable ( $\rightarrow$  I/1.2.2) following a chi-squared distribution ( $\rightarrow$  II/3.7.1):

$$Y \sim \chi^2(k) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $Y$  is

$$f_Y(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} \exp[-y/2] . \quad (2)$$

**Proof:** A chi-square-distributed random variable ( $\rightarrow$  II/3.7.1) with  $k$  degrees of freedom is defined as the sum of  $k$  squared standard normal random variables ( $\rightarrow$  II/3.2.3):

$$X_1, \dots, X_k \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad Y = \sum_{i=1}^k X_i^2 \sim \chi^2(k) . \quad (3)$$

Let  $x_1, \dots, x_k$  be values of  $X_1, \dots, X_k$  and consider  $x = (x_1, \dots, x_k)$  to be a point in  $k$ -dimensional space. Define

$$y = \sum_{i=1}^k x_i^2 \quad (4)$$

and let  $f_Y(y)$  and  $F_Y(y)$  be the probability density function ( $\rightarrow$  I/1.7.1) and cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y$ . Because the PDF is the first derivative of the CDF ( $\rightarrow$  I/1.7.7), we can write:

$$F_Y(y) = \frac{F_Y(y)}{dy} dy = f_Y(y) dy . \quad (5)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $Y$  can be expressed as

$$f_Y(y) dy = \int_V \prod_{i=1}^k (\mathcal{N}(x_i; 0, 1) dx_i) \quad (6)$$

where  $\mathcal{N}(x_i; 0, 1)$  is the probability density function ( $\rightarrow$  I/1.7.1) of the standard normal distribution ( $\rightarrow$  II/3.2.3) and  $V$  is the elemental shell volume at  $y(x)$ , which is proportional to the  $(k-1)$ -dimensional surface in  $k$ -space for which equation (4) is fulfilled. Using the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), equation (6) can be developed as follows:

$$\begin{aligned}
f_Y(y) dy &= \int_V \prod_{i=1}^k \left( \frac{1}{\sqrt{2\pi}} \cdot \exp \left[ -\frac{1}{2} x_i^2 \right] dx_i \right) \\
&= \int_V \frac{\exp \left[ -\frac{1}{2} (x_1^2 + \dots + x_k^2) \right]}{(2\pi)^{k/2}} dx_1 \dots dx_k \\
&= \frac{1}{(2\pi)^{k/2}} \int_V \exp \left[ -\frac{y}{2} \right] dx_1 \dots dx_k .
\end{aligned} \tag{7}$$

Because  $y$  is constant within the set  $V$ , it can be moved out of the integral:

$$f_Y(y) dy = \frac{\exp [-y/2]}{(2\pi)^{k/2}} \int_V dx_1 \dots dx_k . \tag{8}$$

Now, the integral is simply the surface area of the  $(k-1)$ -dimensional sphere with radius  $r = \sqrt{y}$ , which is

$$A = 2r^{k-1} \frac{\pi^{k/2}}{\Gamma(k/2)} , \tag{9}$$

times the infinitesimal thickness of the sphere, which is

$$\frac{dr}{dy} = \frac{1}{2} y^{-1/2} \Leftrightarrow dr = \frac{dy}{2y^{1/2}} . \tag{10}$$

Substituting (9) and (10) into (8), we have:

$$\begin{aligned}
f_Y(y) dy &= \frac{\exp [-y/2]}{(2\pi)^{k/2}} \cdot A dr \\
&= \frac{\exp [-y/2]}{(2\pi)^{k/2}} \cdot 2r^{k-1} \frac{\pi^{k/2}}{\Gamma(k/2)} \cdot \frac{dy}{2y^{1/2}} \\
&= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot \frac{2\sqrt{y}^{k-1}}{2\sqrt{y}} \cdot \exp [-y/2] dy \\
&= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot y^{\frac{k}{2}-1} \cdot \exp \left[ -\frac{y}{2} \right] dy .
\end{aligned} \tag{11}$$

From this, we get the final result in (2):

$$f_Y(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} \exp [-y/2] . \tag{12}$$

■

### Sources:

- Wikipedia (2020): “Proofs related to chi-squared distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Proofs\\_related\\_to\\_chi-squared\\_distribution#Derivation\\_of\\_the\\_pdf\\_for\\_k\\_degrees\\_of\\_freedom](https://en.wikipedia.org/wiki/Proofs_related_to_chi-squared_distribution#Derivation_of_the_pdf_for_k_degrees_of_freedom).
- Wikipedia (2020): “n-sphere”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/N-sphere#Volume\\_and\\_surface\\_area](https://en.wikipedia.org/wiki/N-sphere#Volume_and_surface_area).

### 3.7.4 Raw moments

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $k$  degrees of freedom:

$$X \sim \chi^2(k) . \quad (1)$$

Then, if  $m > -k/2$ , the raw moment ( $\rightarrow$  I/1.21.3)  $E(X^m)$  exists and is equal to:

$$E(X^m) = 2^m \frac{\Gamma\left(\frac{k}{2} + m\right)}{\Gamma\left(\frac{k}{2}\right)} . \quad (2)$$

**Proof:** Combining the definition of the raw moment ( $\rightarrow$  I/1.21.3) with the probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3), we have:

$$\begin{aligned} E(X^m) &= \int_0^\infty x^m \frac{1}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} x^{k/2-1} \exp[-x/2] dx \\ &= \frac{1}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_0^\infty x^{(k/2)+m-1} \exp[-x/2] dx . \end{aligned} \quad (3)$$

Now, we substitute  $u = x/2$ , such that  $x = 2u$ . As a result, we obtain:

$$\begin{aligned} E(X^m) &= \frac{1}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_0^\infty 2^{(k/2)+m-1} u^{(k/2)+m-1} \exp[-u] d(2u) \\ &= \frac{2^{(k/2)+m}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_0^\infty u^{(k/2)+m-1} \exp[-u] du \\ &= \frac{2^m}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty u^{(k/2)+m-1} \exp[-u] du . \end{aligned} \quad (4)$$

With the definition of the gamma function as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \quad x > 0 , \quad (5)$$

this leads to the desired result when  $m > -k/2$ . Observe that, if  $m$  is a non-negative integer, then  $m > -k/2$  is always true. Therefore, all moments ( $\rightarrow$  I/1.21.1) of a chi-squared distribution ( $\rightarrow$  II/3.7.1) exist and the  $m$ -th raw moment is given by the equation above. ■

#### Sources:

- Robert V. Hogg, Joseph W. McKean, Allen T. Craig (2018): “The 2-Distribution”; in: *Introduction to Mathematical Statistics*, Pearson, Boston, 2019, p. 179, eq. 3.3.8; URL: <https://www.pearson.com/store/p/introduction-to-mathematical-statistics/P100000843744>.

### 3.8 F-distribution

#### 3.8.1 Definition

**Definition:** Let  $X_1$  and  $X_2$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $\nu_1$  and  $\nu_2$  degrees of freedom ( $\rightarrow$  I/1.5.12), respectively:

$$\begin{aligned} X_1 &\sim \chi^2(\nu_1) \\ X_2 &\sim \chi^2(\nu_2) . \end{aligned} \quad (1)$$

Then, the ratio of  $X_1$  to  $X_2$ , divided by their respective degrees of freedom, is said to be  $F$ -distributed with numerator degrees of freedom  $\nu_1$  and denominator degrees of freedom  $\nu_2$ :

$$Y = \frac{X_1/\nu_1}{X_2/\nu_2} \sim F(\nu_1, \nu_2) \quad \text{where } \nu_1, \nu_2 > 0 . \quad (2)$$

The  $F$ -distribution is also called “Snedecor’s  $F$ -distribution” or “Fisher–Snedecor distribution”, after Ronald A. Fisher and George W. Snedecor.

#### Sources:

- Wikipedia (2021): “F-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-21; URL: <https://en.wikipedia.org/wiki/F-distribution#Characterization>.

#### 3.8.2 Probability density function

**Theorem:** Let  $F$  be a random variable ( $\rightarrow$  I/1.2.2) following an  $F$ -distribution ( $\rightarrow$  II/3.8.1):

$$F \sim F(u, v) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $F$  is

$$f_F(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right)} \cdot \left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1} \cdot \left(\frac{u}{v}f + 1\right)^{-\frac{u+v}{2}} . \quad (2)$$

**Proof:** An  $F$ -distributed random variable ( $\rightarrow$  II/3.8.1) is defined as the ratio of two chi-squared random variables ( $\rightarrow$  II/3.7.1), divided by their degrees of freedom ( $\rightarrow$  I/1.5.12)

$$X \sim \chi^2(u), Y \sim \chi^2(v) \quad \Rightarrow \quad F = \frac{X/u}{Y/v} \sim F(u, v) \quad (3)$$

where  $X$  and  $Y$  are independent of each other ( $\rightarrow$  I/1.3.6).

The probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3) is

$$f_X(x) = \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot x^{\frac{u}{2}-1} \cdot e^{-\frac{x}{2}} . \quad (4)$$

Define the random variables  $F$  and  $W$  as functions of  $X$  and  $Y$

$$\begin{aligned} F &= \frac{X/u}{Y/v} \\ W &= Y , \end{aligned} \quad (5)$$

such that the inverse functions  $X$  and  $Y$  in terms of  $F$  and  $W$  are

$$\begin{aligned} X &= \frac{u}{v}FW \\ Y &= W. \end{aligned} \quad (6)$$

This implies the following Jacobian matrix and determinant:

$$\begin{aligned} J &= \begin{bmatrix} \frac{dX}{dF} & \frac{dX}{dW} \\ \frac{dY}{dF} & \frac{dY}{dW} \end{bmatrix} = \begin{bmatrix} \frac{u}{v}W & \frac{u}{v}F \\ 0 & 1 \end{bmatrix} \\ |J| &= \frac{u}{v}W. \end{aligned} \quad (7)$$

Because  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), the joint density ( $\rightarrow$  I/1.5.2) of  $X$  and  $Y$  is equal to the product ( $\rightarrow$  I/1.3.9) of the marginal densities ( $\rightarrow$  I/1.5.3):

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y). \quad (8)$$

With the probability density function of an invertible function ( $\rightarrow$  I/1.7.5), the joint density ( $\rightarrow$  I/1.5.2) of  $F$  and  $W$  can be derived as:

$$f_{F,W}(f, w) = f_{X,Y}(x, y) \cdot |J|. \quad (9)$$

Substituting (6) into (4), and then with (7) into (9), we get:

$$\begin{aligned} f_{F,W}(f, w) &= f_X\left(\frac{u}{v}fw\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot \left(\frac{u}{v}fw\right)^{\frac{u}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{u}{v}fw\right)} \cdot \frac{1}{\Gamma\left(\frac{v}{2}\right) \cdot 2^{v/2}} \cdot w^{\frac{v}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \frac{u}{v}w \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot w^{\frac{u+v}{2}-1} \cdot e^{-\frac{w}{2}\left(\frac{u}{v}f+1\right)}. \end{aligned} \quad (10)$$

The marginal density ( $\rightarrow$  I/1.5.3) of  $F$  can now be obtained by integrating out ( $\rightarrow$  I/1.3.3)  $W$ :

$$\begin{aligned} f_F(f) &= \int_0^\infty f_{F,W}(f, w) dw \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \int_0^\infty w^{\frac{u+v}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{u}{v}f+1\right)w\right] dw \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \frac{\Gamma\left(\frac{u+v}{2}\right)}{\left[\frac{1}{2}\left(\frac{u}{v}f+1\right)\right]^{(u+v)/2}} \cdot \int_0^\infty \frac{\left[\frac{1}{2}\left(\frac{u}{v}f+1\right)\right]^{(u+v)/2}}{\Gamma\left(\frac{u+v}{2}\right)} \cdot w^{\frac{u+v}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{u}{v}f+1\right)w\right] dw \end{aligned} \quad (11)$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution ( $\rightarrow$  II/3.4.7) with

$$a = \frac{u+v}{2} \quad \text{and} \quad b = \frac{1}{2}\left(\frac{u}{v}f+1\right), \quad (12)$$

and because a probability density function integrates to one ( $\rightarrow$  I/1.7.1), we finally have:

$$\begin{aligned} f_F(f) &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \frac{\Gamma\left(\frac{u+v}{2}\right)}{\left[\frac{1}{2}\left(\frac{u}{v}f + 1\right)\right]^{(u+v)/2}} \\ &= \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right)} \cdot \left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1} \cdot \left(\frac{u}{v}f + 1\right)^{-\frac{u+v}{2}}. \end{aligned} \quad (13)$$

■

#### Sources:

- statisticsmatt (2018): “Statistical Distributions: Derive the F Distribution”; in: *You Tube*, retrieved on 2021-10-11; URL: <https://www.youtube.com/watch?v=AmHiOKYmHkI>.

### 3.9 Beta distribution

#### 3.9.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to follow a beta distribution with shape parameters  $\alpha$  and  $\beta$

$$X \sim \text{Bet}(\alpha, \beta), \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\text{Bet}(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (2)$$

where  $\alpha > 0$  and  $\beta > 0$ ,  $B(x, y)$  is the beta function and the density is zero, if  $x \notin [0, 1]$ .

#### Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Definitions](https://en.wikipedia.org/wiki/Beta_distribution#Definitions).

#### 3.9.2 Special case of Dirichlet distribution

**Theorem:** The beta distribution ( $\rightarrow$  II/3.9.1) with shape parameters  $\alpha$  and  $\beta$  is a special case of the Dirichlet distribution ( $\rightarrow$  II/2.2.1) with concentration parameters  $\alpha_1 = \alpha$  and  $\alpha_2 = \beta$ :

$$X \sim \text{Dir}([\alpha, \beta]) \quad \Rightarrow \quad X \sim \text{Bet}(\alpha, \beta). \quad (1)$$

**Proof:** The probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2), where  $x$  is a  $1 \times k$  vector with  $x_i \in [0, 1]$  for  $i = 1, \dots, k$ , is as follows:

$$\text{Dir}(x; [\alpha_1, \dots, \alpha_k]) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1}. \quad (2)$$

If we let  $k = 2$ ,  $\alpha_1 = \alpha$  and  $\alpha_2 = \beta$ , we obtain



$$\text{Dir}(x; [\alpha, \beta]) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} . \quad (3)$$

Applying the beta function  $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ , we have

$$\text{Dir}(x; [\alpha, \beta]) = \frac{1}{B(\alpha, \beta)} x_1^{\alpha-1} x_2^{\beta-1} . \quad (4)$$

Recognizing that  $\sum_{i=1}^k x_i = 1$ , such that  $x_2 = 1 - x_1$ , we get

$$\text{Dir}(x; [\alpha, \beta]) = \frac{1}{B(\alpha, \beta)} x_1^{\alpha-1} (1 - x_1)^{\beta-1} \quad (5)$$

which is equivalent to the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5). ■

### 3.9.3 Relationship to chi-squared distribution

**Theorem:** Let  $X$  and  $Y$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following chi-squared distributions ( $\rightarrow$  II/3.7.1):

$$X \sim \chi^2(m) \quad \text{and} \quad Y \sim \chi^2(n) . \quad (1)$$

Then, the quantity  $X/(X+Y)$  follows a beta distribution ( $\rightarrow$  II/3.9.1):

$$\frac{X}{X+Y} \sim \text{Bet} \left( \frac{m}{2}, \frac{n}{2} \right) . \quad (2)$$

**Proof:** The probability density function of the chi-squared distribution ( $\rightarrow$  II/3.7.3) is

$$X \sim \chi^2(u) \quad \Rightarrow \quad f_X(x) = \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot x^{\frac{u}{2}-1} \cdot e^{-\frac{x}{2}} . \quad (3)$$

Define the random variables  $Z$  and  $W$  as functions of  $X$  and  $Y$

$$\begin{aligned} Z &= \frac{X}{X+Y} \\ W &= Y , \end{aligned} \quad (4)$$

such that the inverse functions  $X$  and  $Y$  in terms of  $Z$  and  $W$  are

$$\begin{aligned} X &= \frac{ZW}{1-Z} \\ Y &= W . \end{aligned} \quad (5)$$

This implies the following Jacobian matrix and determinant:

$$\begin{aligned} J &= \begin{bmatrix} \frac{dX}{dZ} & \frac{dX}{dW} \\ \frac{dY}{dZ} & \frac{dY}{dW} \end{bmatrix} = \begin{bmatrix} \frac{W}{(1-Z)^2} & \frac{Z}{1-Z} \\ 0 & 1 \end{bmatrix} \\ |J| &= \frac{W}{(1-Z)^2} . \end{aligned} \quad (6)$$

Because  $X$  and  $Y$  are independent ( $\rightarrow$  I/1.3.6), the joint density ( $\rightarrow$  I/1.5.2) of  $X$  and  $Y$  is equal to the product ( $\rightarrow$  I/1.3.9) of the marginal densities ( $\rightarrow$  I/1.5.3):

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) . \quad (7)$$

With the probability density function of an invertible function ( $\rightarrow$  I/1.7.5), the joint density ( $\rightarrow$  I/1.5.2) of  $Z$  and  $W$  can be derived as:

$$f_{Z,W}(z, w) = f_{X,Y}(x, y) \cdot |J| . \quad (8)$$

Substituting (5) into (3), and then with (6) into (8), we get:

$$\begin{aligned} f_{Z,W}(z, w) &= f_X\left(\frac{zw}{1-z}\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \cdot 2^{m/2}} \cdot \left(\frac{zw}{1-z}\right)^{\frac{m}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{zw}{1-z}\right)} \cdot \frac{1}{\Gamma\left(\frac{n}{2}\right) \cdot 2^{n/2}} \cdot w^{\frac{n}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \frac{w}{(1-z)^2} \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{m/2} 2^{n/2}} \cdot \left(\frac{z}{1-z}\right)^{\frac{m}{2}-1} \left(\frac{1}{(1-z)}\right)^2 \cdot w^{\frac{m}{2}+\frac{n}{2}-1} e^{-\frac{1}{2}\left(\frac{zw}{1-z} + \frac{w(1-z)}{1-z}\right)} \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{w}{1-z}\right)} . \end{aligned} \quad (9)$$

The marginal density ( $\rightarrow$  I/1.5.3) of  $Z$  can now be obtained by integrating out ( $\rightarrow$  I/1.3.3)  $W$ :

$$\begin{aligned} f_Z(z) &= \int_0^\infty f_{Z,W}(z, w) dw \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \int_0^\infty w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{w}{1-z}\right)} dw \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}} \\ &\quad \int_0^\infty \frac{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}}{\Gamma\left(\frac{m+n}{2}\right)} \cdot w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2(1-z)}w} dw . \end{aligned} \quad (10)$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution ( $\rightarrow$  II/3.4.7) with

$$a = \frac{m+n}{2} \quad \text{and} \quad b = \frac{1}{2(1-z)} , \quad (11)$$

and because a probability density function integrates to one ( $\rightarrow$  I/1.7.1), we have:

$$\begin{aligned}
f_Z(z) &= \frac{1}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}} \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right) \cdot 2^{(m+n)/2}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}+\frac{m+n}{2}-1} \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{\frac{n}{2}-1}.
\end{aligned} \tag{12}$$

With the definition of the beta function ( $\rightarrow$  II/3.9.8), this becomes

$$f_Z(z) = \frac{1}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{\frac{n}{2}-1} \tag{13}$$

which is the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) with parameters

$$\alpha = \frac{m}{2} \quad \text{and} \quad \beta = \frac{n}{2}, \tag{14}$$

such that

$$Z \sim \text{Bet}\left(\frac{m}{2}, \frac{n}{2}\right). \tag{15}$$

#### Sources:

- Probability Fact (2021): “If  $X \sim \text{chisq}(m)$  and  $Y \sim \text{chisq}(n)$  are independent”; in: *Twitter*, retrieved on 2022-10-17; URL: <https://twitter.com/ProbFact/status/1450492787854647300>.

### 3.9.4 Relationship to F-distribution

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an F-distribution ( $\rightarrow$  II/3.8.1):

$$X \sim F(d_1, d_2). \tag{1}$$

Then, the quantity

$$Y = \frac{d_1 X / d_2}{1 + d_1 X / d_2} = \frac{d_1 X}{d_2 + d_1 X} \tag{2}$$

follows a beta distribution ( $\rightarrow$  II/3.9.1):

$$Y \sim \text{Bet}\left(\frac{d_1}{2}, \frac{d_2}{2}\right). \tag{3}$$

**Proof:** We denote  $Y = g(X)$ . The first derivative of the  $g$  is:

$$\begin{aligned}
\frac{dg(x)}{dx} &= \frac{d}{dx} \left( \frac{d_1 x}{d_2 + d_1 x} \right) \\
&= \frac{d_1 d_2 + d_1^2 x - d_1^2 x}{(d_2 + d_1 x)^2} \\
&= \frac{d_1 d_2}{(d_2 + d_1 x)^2}.
\end{aligned} \tag{4}$$

This derivative is positive for all  $x \geq 0$ , such that  $Y$  is a strictly increasing function of  $X$ . This means we can derive the distribution of  $Y$  by applying the probability density function of a strictly increasing function of a continuous random variable ( $\rightarrow$  I/1.7.3):

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}, & \text{if } y \in \mathcal{Y} \\ 0, & \text{if } y \notin \mathcal{Y}. \end{cases} \quad (5)$$

where  $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$ . Because  $g : \mathbb{R}_{\geq 0} \rightarrow [0, 1]$ , we have  $\mathcal{X} = \mathbb{R}_{\geq 0}$  and  $\mathcal{Y} = [0, 1]$ , such that

$$f_Y(y) = 0 \quad \text{for } y \notin [0, 1]. \quad (6)$$

The inverse function of  $g$  is:

$$\begin{aligned} g(x) = y &= \frac{d_1 x}{d_2 + d_1 x} \\ d_2 y + d_1 x y &= d_1 x \\ d_1 x - d_1 x y &= d_2 y \\ x(d_1 - d_1 y) &= d_2 y \\ x &= \frac{d_2 y}{d_1 - d_1 y} = g^{-1}(y). \end{aligned} \quad (7)$$

The derivative of  $g^{-1}$  with respect to  $y$  is:

$$\begin{aligned} \frac{dg^{-1}(y)}{dy} &= \frac{d}{dy} \left( \frac{d_2 y}{d_1 - d_1 y} \right) \\ &= \frac{d_1 d_2 + d_1 d_2 y - (-d_1 d_2 y)}{(d_1 - d_1 y)^2} \\ &= \frac{d_1 d_2}{d_1^2 - 2d_1^2 y + d_1^2 y^2} \\ &= \frac{d_2}{d_1 - 2d_1 y + d_1 y^2} \\ &= \frac{d_2}{d_1(1 - 2y + y^2)} \\ &= \frac{d_2}{d_1} \cdot \frac{1}{(1 - y)^2}. \end{aligned} \quad (8)$$

The probability density function of the F-distribution ( $\rightarrow$  II/3.8.2) with degrees of freedom ( $\rightarrow$  I/1.5.12)  $d_1$  and  $d_2$  is:

$$f_X(x) = \frac{\Gamma\left(\frac{d_1 + d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right) \cdot \Gamma\left(\frac{d_2}{2}\right)} \cdot \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} \cdot x^{\frac{d_1}{2} - 1} \cdot \left(\frac{d_1}{d_2} x + 1\right)^{-\frac{d_2 + d_2}{2}}. \quad (9)$$

With that, we have everything that we need to derive the distribution of  $Y$  for  $y \in \mathcal{Y}$ . Combining (5), (9), (7) and (8),  $f_Y(y)$  for  $y \in [0, 1]$  becomes:

$$\begin{aligned}
f_Y(y) &= f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} \\
&= \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right) \cdot \Gamma\left(\frac{d_2}{2}\right)} \cdot \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} \cdot \left(\frac{d_2 y}{d_1 - d_1 y}\right)^{\frac{d_1}{2}-1} \cdot \left(\frac{d_1}{d_2} \left(\frac{d_2 y}{d_1 - d_1 y}\right) + 1\right)^{-\frac{d_1+d_2}{2}} \cdot \\
&\quad \left(\frac{d_2}{d_1} \cdot \frac{1}{(1-y)^2}\right) \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} \cdot d_2^{\frac{d_1}{2}-1} \cdot y^{\frac{d_1}{2}-1} \cdot d_1^{-(\frac{d_1}{2}-1)} \cdot (1-y)^{-(\frac{d_1}{2}-1)} \cdot \\
&\quad \left(\frac{d_1}{d_2} \left(\frac{d_2 y}{d_1 - d_1 y}\right) + 1\right)^{-\frac{d_1+d_2}{2}} \cdot \left(\frac{d_2}{d_1} \cdot \frac{1}{(1-y)^2}\right) \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot \left(\frac{d_1}{d_2}\right) \cdot \left(\frac{d_2}{d_1}\right) \cdot y^{\frac{d_1}{2}-1} \cdot (1-y)^{-\frac{d_1}{2}+1} \cdot (1-y)^{-2} \cdot \\
&\quad \left(\frac{d_1 d_2 y}{d_2(d_1 - d_1 y)} + \frac{d_2(d_1 - d_1 y)}{d_2(d_1 - d_1 y)}\right)^{-\frac{d_1+d_2}{2}} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot y^{\frac{d_1}{2}-1} \cdot (1-y)^{-\frac{d_1}{2}-1} \cdot \left(\frac{d_1 d_2}{d_2(d_1 - d_1 y)}\right)^{-\frac{d_1+d_2}{2}} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot y^{\frac{d_1}{2}-1} \cdot (1-y)^{-\frac{d_1}{2}-1} \cdot \left(\frac{1}{1-y}\right)^{-\frac{d_1+d_2}{2}} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot y^{\frac{d_1}{2}-1} \cdot (1-y)^{-\frac{d_1}{2}-1} \cdot (1-y)^{\frac{d_1+d_2}{2}} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \cdot y^{\frac{d_1}{2}-1} \cdot (1-y)^{\frac{d_2}{2}-1} .
\end{aligned} \tag{10}$$

This is the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) with parameters

$$\alpha = \frac{d_1}{2} \quad \text{and} \quad \beta = \frac{d_2}{2} , \tag{11}$$

such that

$$Y \sim \text{Bet}\left(\frac{d_1}{2}, \frac{d_2}{2}\right) . \tag{12}$$

■

#### Sources:

- Wikipedia (2025): “F-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-04-04; URL: [https://en.wikipedia.org/wiki/F-distribution#Related\\_distributions](https://en.wikipedia.org/wiki/F-distribution#Related_distributions).
- Wikipedia (2025): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-04-04; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Related\\_distributions](https://en.wikipedia.org/wiki/Beta_distribution#Related_distributions).

### 3.9.5 Probability density function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} . \quad (2)$$

**Proof:** This follows directly from the definition of the beta distribution ( $\rightarrow$  II/3.9.1). ■

### 3.9.6 Moment-generating function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = 1 + \sum_{n=1}^{\infty} \left( \prod_{m=0}^{n-1} \frac{\alpha + m}{\alpha + \beta + m} \right) \frac{t^n}{n!} . \quad (2)$$

**Proof:** The probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) is

$$f_X(x) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (3)$$

and the moment-generating function ( $\rightarrow$  I/1.9.1) is defined as

$$M_X(t) = \text{E} [e^{tX}] . \quad (4)$$

Using the expected value for continuous random variables ( $\rightarrow$  I/1.12.1), the moment-generating function of  $X$  therefore is

$$\begin{aligned} M_X(t) &= \int_0^1 \exp[tx] \cdot \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{1}{\text{B}(\alpha, \beta)} \int_0^1 e^{tx} x^{\alpha-1} (1-x)^{\beta-1} dx . \end{aligned} \quad (5)$$

With the relationship between beta function and gamma function

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (6)$$

and the integral representation of the confluent hypergeometric function (Kummer's function of the first kind)

$${}_1F_1(a, b, z) = \frac{\Gamma(b)}{\Gamma(a) \Gamma(b-a)} \int_0^1 e^{zu} u^{a-1} (1-u)^{(b-a)-1} du , \quad (7)$$

the moment-generating function can be written as

$$M_X(t) = {}_1F_1(\alpha, \alpha + \beta, t) . \quad (8)$$

Note that the series equation for the confluent hypergeometric function (Kummer's function of the first kind) is

$${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{a^{\bar{n}}}{b^{\bar{n}}} \frac{z^n}{n!} \quad (9)$$

where  $m^{\bar{n}}$  is the rising factorial

$$m^{\bar{n}} = \prod_{i=0}^{n-1} (m + i) , \quad (10)$$

so that the moment-generating function can be written as

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\alpha^{\bar{n}}}{(\alpha + \beta)^{\bar{n}}} \frac{t^n}{n!} . \quad (11)$$

Applying the rising factorial equation (10) and using  $m^{\bar{0}} = x^0 = 0! = 1$ , we finally have:

$$M_X(t) = 1 + \sum_{n=1}^{\infty} \left( \prod_{m=0}^{n-1} \frac{\alpha + m}{\alpha + \beta + m} \right) \frac{t^n}{n!} . \quad (12)$$

■

#### Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Moment\\_generating\\_function](https://en.wikipedia.org/wiki/Beta_distribution#Moment_generating_function).
- Wikipedia (2020): “Confluent hypergeometric function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Confluent\\_hypergeometric\\_function#Kummer's\\_equation](https://en.wikipedia.org/wiki/Confluent_hypergeometric_function#Kummer's_equation).

### 3.9.7 Cumulative distribution function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.4.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the cumulative distribution function ( $\rightarrow$  I/1.8.1) of  $X$  is

$$F_X(x) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} \quad (2)$$

where  $B(a, b)$  is the beta function and  $B(x; a, b)$  is the incomplete gamma function.

**Proof:** The probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) is:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} . \quad (3)$$

Thus, the cumulative distribution function ( $\rightarrow$  I/1.8.1) is:

$$\begin{aligned} F_X(x) &= \int_0^x \text{Bet}(z; \alpha, \beta) \, dz \\ &= \int_0^x \frac{1}{B(\alpha, \beta)} z^{\alpha-1} (1-z)^{\beta-1} \, dz \\ &= \frac{1}{B(\alpha, \beta)} \int_0^x z^{\alpha-1} (1-z)^{\beta-1} \, dz . \end{aligned} \quad (4)$$

With the definition of the incomplete beta function

$$B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} \, dt , \quad (5)$$

we arrive at the final result given by equation (2):

$$F_X(x) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} . \quad (6)$$

■

#### Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Cumulative\\_distribution\\_function](https://en.wikipedia.org/wiki/Beta_distribution#Cumulative_distribution_function).
- Wikipedia (2020): “Beta function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: [https://en.wikipedia.org/wiki/Beta\\_function#Incomplete\\_beta\\_function](https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function).

### 3.9.8 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \frac{\alpha}{\alpha + \beta} . \quad (2)$$

**Proof:** The expected value ( $\rightarrow$  I/1.12.1) is the probability-weighted average over all possible values:

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, dx . \quad (3)$$

The probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \quad (4)$$

where the beta function is given by a ratio gamma functions:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} . \quad (5)$$

Combining (3), (4) and (5), we have:



$$\begin{aligned}
E(X) &= \int_0^1 x \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\
&= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + 1 + \beta)} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx .
\end{aligned} \tag{6}$$

Employing the relation  $\Gamma(x + 1) = \Gamma(x) \cdot x$ , we have

$$\begin{aligned}
E(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{\alpha \cdot \Gamma(\alpha)}{(\alpha + \beta) \cdot \Gamma(\alpha + \beta)} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx \\
&= \frac{\alpha}{\alpha + \beta} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx
\end{aligned} \tag{7}$$

and again using the density of the beta distribution ( $\rightarrow$  II/3.9.5), we get

$$\begin{aligned}
E(X) &= \frac{\alpha}{\alpha + \beta} \int_0^1 \text{Bet}(x; \alpha + 1, \beta) dx \\
&= \frac{\alpha}{\alpha + \beta} .
\end{aligned} \tag{8}$$

■

#### Sources:

- Boer Commander (2020): “Beta Distribution Mean and Variance Proof”; in: *YouTube*, retrieved on 2021-04-29; URL: <https://www.youtube.com/watch?v=3OgCcnpZtZ8>.

### 3.9.9 Median

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1):

$$X \sim \text{Bet}(\alpha, \beta) . \tag{1}$$

Then, the median ( $\rightarrow$  I/1.18.1) of  $X$  is

$$\text{median}(X) = B_{\alpha, \beta}^{-1} \left( \frac{B(\alpha, \beta)}{2} \right) = I_{1/2}^{-1}(\alpha, \beta) \tag{2}$$

where  $B(a, b)$  is the beta function,  $B_{a, b}^{-1}(y)$  is the inverse function of the incomplete beta function  $B(x; a, b)$  and  $I_y^{-1}(a, b)$  is the inverse function of the regularized incomplete beta function.

**Proof:** The median ( $\rightarrow$  I/1.18.1) is the value at which the cumulative distribution function ( $\rightarrow$  I/1.8.1) is 1/2:

$$F_X(\text{median}(X)) = \frac{1}{2} . \tag{3}$$

The cumulative distribution function of the beta distribution ( $\rightarrow$  II/3.9.7) is

$$F_X(x) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} \tag{4}$$

where the incomplete beta function  $B(x; a, b)$  is given by

$$B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt . \quad (5)$$

Thus, the inverse CDF, i.e. the quantile function ( $\rightarrow$  I/1.11.1), is

$$x = B_{\alpha, \beta}^{-1}(p B(\alpha, \beta)) \quad (6)$$

where  $B_{a,b}^{-1}(y)$  is the inverse function of  $B(x; a, b)$ . Setting  $p = 1/2$ , we obtain:

$$\text{median}(X) = B_{\alpha, \beta}^{-1} \left( \frac{B(\alpha, \beta)}{2} \right) . \quad (7)$$

Alternatively, the cumulative distribution function may be written as

$$F_X(x) = I_x(a, b) \quad (8)$$

using the regularized incomplete beta function

$$I_x(a, b) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} \quad (9)$$

in which case the inverse CDF is

$$x = I_p^{-1}(a, b) , \quad (10)$$

such that the median becomes

$$\text{median}(X) = I_{1/2}^{-1}(\alpha, \beta) . \quad (11)$$

■

#### Sources:

- Wikipedia (2025): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-10-24; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Median](https://en.wikipedia.org/wiki/Beta_distribution#Median).

#### 3.9.10 Mode

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) \in \begin{cases} \{0, 1\} , & \text{if } \alpha < 1 \text{ and } \beta < 1 \\ [0, 1] , & \text{if } \alpha = 1 \text{ and } \beta = 1 \end{cases} \quad (2)$$

and

$$\text{mode}(X) = \begin{cases} 0 \text{ or } 1 , & \text{if } \alpha < 1 \text{ or } \beta < 1 \text{ (but not } \alpha < 1 \text{ and } \beta < 1) \\ \frac{\alpha-1}{\alpha+\beta-2} , & \text{if } \alpha \geq 1 \text{ and } \beta \geq 1 \text{ (but not } \alpha = 1 \text{ and } \beta = 1) . \end{cases} \quad (3)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (4)$$

The probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) is:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} . \quad (5)$$

1) If  $\alpha < 1$ , then 0 is the mode, since

$$\lim_{\substack{x \rightarrow 0 \\ \alpha < 1}} f_X(x) = \infty , \quad \text{because} \quad \lim_{x \rightarrow 0} x^{\alpha-1} = \infty \quad \text{for} \quad \alpha < 1 , \quad (6)$$

and if  $\beta < 1$ , then 1 is the mode, since

$$\lim_{\substack{x \rightarrow 1 \\ \beta < 1}} f_X(x) = \infty , \quad \text{because} \quad \lim_{x \rightarrow 1} (1-x)^{\beta-1} = \infty \quad \text{for} \quad \beta < 1 . \quad (7)$$

2) If both  $\alpha < 1$  and  $\beta < 1$ , then

$$\lim_{x \rightarrow 0} f_X(x) = \infty \quad \text{and} \quad \lim_{x \rightarrow 1} f_X(x) = \infty , \quad (8)$$

so any value from the set  $\{0, 1\}$  may be considered the mode.

3) If both  $\alpha = 1$  and  $\beta = 1$ , then

$$\begin{aligned} f_X(x) &= \frac{1}{B(1, 1)} x^{1-1} (1-x)^{1-1} \\ &= \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^0 (1-x)^0 \\ &= 1 = \text{const.} , \end{aligned} \quad (9)$$

i.e. the distribution becomes equivalent to the (standard) continuous uniform distribution ( $\rightarrow$  II/3.1.1) with parameters  $a = 0$  and  $b = 1$  which has a constant probability density function ( $\rightarrow$  II/3.1.3). Thus, any value from the interval  $[0, 1]$  may be considered the mode ( $\rightarrow$  II/3.1.8).

4) For the remaining cases, we must analyze the probability density function. The first two derivatives of this function are:

$$\begin{aligned} f'_X(x) &= \frac{df_X(x)}{dx} = \frac{1}{B(\alpha, \beta)} [-(\beta-1)x^{\alpha-1}(1-x)^{\beta-2} + (\alpha-1)x^{\alpha-2}(1-x)^{\beta-1}] \\ &= \frac{1}{B(\alpha, \beta)} [(\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta-1)x^{\alpha-1}(1-x)^{\beta-2}] \end{aligned} \quad (10)$$

$$\begin{aligned} f''_X(x) &= \frac{d^2 f_X(x)}{dx^2} = \frac{1}{B(\alpha, \beta)} [(\alpha-1)((\alpha-2)x^{\alpha-3}(1-x)^{\beta-1} - (\beta-1)x^{\alpha-2}(1-x)^{\beta-2}) - \\ &\quad (\beta-1)((\alpha-1)x^{\alpha-2}(1-x)^{\beta-2} - (\beta-2)x^{\alpha-1}(1-x)^{\beta-3})] \\ &= \frac{1}{B(\alpha, \beta)} [(\alpha-1)(\alpha-2)x^{\alpha-3}(1-x)^{\beta-1} - 2(\alpha-1)(\beta-1)x^{\alpha-2}(1-x)^{\beta-2} + \\ &\quad (\beta-1)(\beta-2)x^{\alpha-1}(1-x)^{\beta-3}] . \end{aligned} \quad (11)$$

We now calculate the root of the first derivative (10):

$$\begin{aligned}
 f'_X(x) = 0 &= \frac{1}{B(\alpha, \beta)} [(\alpha - 1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta - 1)x^{\alpha-1}(1-x)^{\beta-2}] \\
 &\Leftrightarrow \\
 (\beta - 1)x^{\alpha-1}(1-x)^{\beta-2} &= (\alpha - 1)x^{\alpha-2}(1-x)^{\beta-1} \\
 (\beta - 1)x &= (\alpha - 1)(1-x) \\
 x[(\beta - 1) + (\alpha - 1)] &= \alpha - 1 \\
 x &= \frac{\alpha - 1}{\alpha + \beta - 2}.
 \end{aligned} \tag{12}$$

Note that for this quantity, we have

$$\begin{aligned}
 \frac{\alpha - 1}{\alpha + \beta - 2} &< 0, \quad \text{if } \alpha < 1 \quad \text{and} \quad \beta > 2 - \alpha \\
 \frac{\alpha - 1}{\alpha + \beta - 2} &> 1, \quad \text{if } \beta < 1 \quad \text{and} \quad \alpha > 2 - \beta.
 \end{aligned} \tag{13}$$

Also note that the following holds:

$$1 - x = 1 - \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{\alpha + \beta - 2}{\alpha + \beta - 2} - \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{\beta - 1}{\alpha + \beta - 2}. \tag{14}$$

By plugging  $x$  and  $1 - x$  into the second derivative (11), we find:

$$\begin{aligned}
 f''_X\left(\frac{\alpha - 1}{\alpha + \beta - 2}\right) &= \frac{1}{B(\alpha, \beta)} \left[ (\alpha - 1)(\alpha - 2) \left(\frac{\alpha - 1}{\alpha + \beta - 2}\right)^{\alpha-3} \left(\frac{\beta - 1}{\alpha + \beta - 2}\right)^{\beta-1} - \right. \\
 &\quad 2(\alpha - 1)(\beta - 1) \left(\frac{\alpha - 1}{\alpha + \beta - 2}\right)^{\alpha-2} \left(\frac{\beta - 1}{\alpha + \beta - 2}\right)^{\beta-2} + \\
 &\quad \left. (\beta - 1)(\beta - 2) \left(\frac{\alpha - 1}{\alpha + \beta - 2}\right)^{\alpha-1} \left(\frac{\beta - 1}{\alpha + \beta - 2}\right)^{\beta-3} \right].
 \end{aligned} \tag{15}$$

Multiplying with factors which are certainly positive, we can focus on those parts of the second derivative which determine its sign:

$$\begin{aligned}
 \frac{f''_X(x) \cdot B(\alpha, \beta)}{\left(\frac{\alpha-1}{\alpha+\beta-2}\right)^{\alpha-3} \cdot \left(\frac{\beta-1}{\alpha+\beta-2}\right)^{\beta-3}} &= (\alpha - 1)(\alpha - 2) \left(\frac{\beta - 1}{\alpha + \beta - 2}\right)^2 - \\
 &\quad 2(\alpha - 1)(\beta - 1) \left(\frac{\alpha - 1}{\alpha + \beta - 2}\right) \left(\frac{\beta - 1}{\alpha + \beta - 2}\right) + \\
 &\quad (\beta - 1)(\beta - 2) \left(\frac{\alpha - 1}{\alpha + \beta - 2}\right)^2.
 \end{aligned} \tag{16}$$

Further multiplying with or dividing by terms which are necessarily positive and thus do not change the sign of the function value, we get:

$$\begin{aligned}
\frac{f_X''(x) \cdot B(\alpha, \beta)}{\left(\frac{\alpha-1}{\alpha+\beta-2}\right)^{\alpha-3} \cdot \left(\frac{\beta-1}{\alpha+\beta-2}\right)^{\beta-3}} \cdot \frac{(\alpha+\beta-2)^2}{(\alpha-1)(\beta-1)} &= (\alpha-2)(\beta-1) - 2(\alpha-1)(\beta-1) + (\alpha-1)(\beta-2) \\
&= (\alpha-1)[(\beta-2) - (\beta-1)] + (\beta-1)[(\alpha-1) - (\alpha-2)] \\
&= -(\alpha-1) - (\beta-1) \\
&< 0, \quad \text{if } \alpha > 1 \quad \text{and} \quad \beta > 1.
\end{aligned} \tag{17}$$

Thus,  $f'_X(x)$  is negative for  $x = \frac{\alpha-1}{\alpha+\beta-2}$ , demonstrating that this is a maximum. To summarize:

- If  $\alpha < 1$  and  $\beta < 1$ , then  $f_X(x)$  diverges at both ends and both values from the set  $\{0, 1\}$  can be seen as the mode of  $X$ .
- If  $\alpha < 1$  or  $\beta < 1$  (but not  $\alpha < 1$  and  $\beta < 1$ ), then the mode of  $X$  is 0 or 1, because  $f_X(x)$  tends towards infinity at  $x = 0$  or  $x = 1$ .
- If  $\alpha = 1$  and  $\beta = 1$ , then  $f_X(x)$  is constant and any value in the interval  $[0, 1]$  can be seen as the mode of  $X$ .
- If  $\alpha \geq 1$  and  $\beta \geq 1$  (but not  $\alpha = 1$  and  $\beta = 1$ ), then  $0 < x = \frac{\alpha-1}{\alpha+\beta-2} < 1$  and  $f'_X(x) = 0$  and  $f'_X(x) < 0$ , such that  $f_X(x)$  reaches its maximum at  $\frac{\alpha-1}{\alpha+\beta-2} = \text{mode}(X)$ .

■

#### Sources:

- Wikipedia (2025): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-10-24; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Mode](https://en.wikipedia.org/wiki/Beta_distribution#Mode).

#### 3.9.11 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a beta distribution ( $\rightarrow$  II/3.9.1):

$$X \sim \text{Bet}(\alpha, \beta). \tag{1}$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1) \cdot (\alpha+\beta)^2}. \tag{2}$$

**Proof:** The variance ( $\rightarrow$  I/1.13.1) can be expressed in terms of expected values ( $\rightarrow$  I/1.13.4) as

$$\text{Var}(X) = E(X^2) - E(X)^2. \tag{3}$$

The expected value of a beta random variable ( $\rightarrow$  II/3.9.8) is

$$E(X) = \frac{\alpha}{\alpha+\beta}. \tag{4}$$

The probability density function of the beta distribution ( $\rightarrow$  II/3.9.5) is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \tag{5}$$

where the beta function is given by a ratio gamma functions:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} . \quad (6)$$

Therefore, the expected value of a squared beta random variable becomes

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha + 2 + \beta)} \int_0^1 \frac{\Gamma(\alpha + 2 + \beta)}{\Gamma(\alpha + 2) \cdot \Gamma(\beta)} x^{(\alpha+2)-1} (1-x)^{\beta-1} dx . \end{aligned} \quad (7)$$

Twice-applying the relation  $\Gamma(x + 1) = \Gamma(x) \cdot x$ , we have

$$\begin{aligned} E(X^2) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{(\alpha + 1) \cdot \alpha \cdot \Gamma(\alpha)}{(\alpha + \beta + 1) \cdot (\alpha + \beta) \cdot \Gamma(\alpha + \beta)} \int_0^1 \frac{\Gamma(\alpha + 2 + \beta)}{\Gamma(\alpha + 2) \cdot \Gamma(\beta)} x^{(\alpha+2)-1} (1-x)^{\beta-1} dx \\ &= \frac{(\alpha + 1) \cdot \alpha}{(\alpha + \beta + 1) \cdot (\alpha + \beta)} \int_0^1 \frac{\Gamma(\alpha + 2 + \beta)}{\Gamma(\alpha + 2) \cdot \Gamma(\beta)} x^{(\alpha+2)-1} (1-x)^{\beta-1} dx \end{aligned} \quad (8)$$

and again using the density of the beta distribution ( $\rightarrow$  II/3.9.5), we get

$$\begin{aligned} E(X^2) &= \frac{(\alpha + 1) \cdot \alpha}{(\alpha + \beta + 1) \cdot (\alpha + \beta)} \int_0^1 \text{Bet}(x; \alpha + 2, \beta) dx \\ &= \frac{(\alpha + 1) \cdot \alpha}{(\alpha + \beta + 1) \cdot (\alpha + \beta)} . \end{aligned} \quad (9)$$

Plugging (9) and (4) into (3), the variance of a beta random variable finally becomes

$$\begin{aligned} \text{Var}(X) &= \frac{(\alpha + 1) \cdot \alpha}{(\alpha + \beta + 1) \cdot (\alpha + \beta)} - \left( \frac{\alpha}{\alpha + \beta} \right)^2 \\ &= \frac{(\alpha^2 + \alpha) \cdot (\alpha + \beta)}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2} - \frac{\alpha^2 \cdot (\alpha + \beta + 1)}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2} \\ &= \frac{(\alpha^3 + \alpha^2\beta + \alpha^2 + \alpha\beta) - (\alpha^3 + \alpha^2\beta + \alpha^2)}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2} \\ &= \frac{\alpha\beta}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2} . \end{aligned} \quad (10)$$

■

#### Sources:

- Boer Commander (2020): “Beta Distribution Mean and Variance Proof”; in: *YouTube*, retrieved on 2021-04-29; URL: <https://www.youtube.com/watch?v=3OgCcnPZtZ8>.

## 3.10 Wald distribution

### 3.10.1 Definition

**Definition:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  is said to follow a Wald distribution with drift rate  $\gamma$  and threshold  $\alpha$

$$X \sim \text{Wald}(\gamma, \alpha), \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\text{Wald}(x; \gamma, \alpha) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha - \gamma x)^2}{2x}\right) \quad (2)$$

where  $\gamma > 0$ ,  $\alpha > 0$ , and the density is zero if  $x \leq 0$ .

#### Sources:

- Anders, R., Alario, F.-X., and van Maanen, L. (2016): “The Shifted Wald Distribution for Response Time Data Analysis”; in: *Psychological Methods*, vol. 21, no. 3, pp. 309-327; URL: <https://dx.doi.org/10.1037/met0000066>; DOI: 10.1037/met0000066.

### 3.10.2 Probability density function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a Wald distribution ( $\rightarrow$  II/3.10.1):

$$X \sim \text{Wald}(\gamma, \alpha). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha - \gamma x)^2}{2x}\right). \quad (2)$$

**Proof:** This follows directly from the definition of the Wald distribution ( $\rightarrow$  II/3.10.1). ■

### 3.10.3 Moment-generating function

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a Wald distribution ( $\rightarrow$  II/3.10.1):

$$X \sim \text{Wald}(\gamma, \alpha). \quad (1)$$

Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right]. \quad (2)$$

**Proof:** The probability density function of the Wald distribution ( $\rightarrow$  II/3.10.2) is

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha - \gamma x)^2}{2x}\right) \quad (3)$$

and the moment-generating function ( $\rightarrow$  I/1.9.1) is defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]. \quad (4)$$

Using the definition of expected value for continuous random variables ( $\rightarrow$  I/1.12.1), the moment-generating function of  $X$  therefore is

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \cdot \frac{\alpha}{\sqrt{2\pi x^3}} \cdot \exp\left[-\frac{(\alpha - \gamma x)^2}{2x}\right] dx \\ &= \frac{\alpha}{\sqrt{2\pi}} \int_0^\infty x^{-3/2} \cdot \exp\left[tx - \frac{(\alpha - \gamma x)^2}{2x}\right] dx. \end{aligned} \quad (5)$$

To evaluate this integral, we will need two identities about modified Bessel functions of the second kind<sup>1</sup>, denoted  $K_p$ . The function  $K_p$  (for  $p \in \mathbb{R}$ ) is one of the two linearly independent solutions of the differential equation

$$x^2 \frac{d^2 y}{dx^2} + x \frac{dy}{dx} - (x^2 + p^2)y = 0. \quad (6)$$

The first of these identities<sup>2</sup> gives an explicit solution for  $K_{-1/2}$ :

$$K_{-1/2}(x) = \sqrt{\frac{\pi}{2x}} e^{-x}. \quad (7)$$

The second of these identities<sup>3</sup> gives an integral representation of  $K_p$ :

$$K_p(\sqrt{ab}) = \frac{1}{2} \left(\frac{a}{b}\right)^{p/2} \int_0^\infty x^{p-1} \cdot \exp\left[-\frac{1}{2} \left(ax + \frac{b}{x}\right)\right] dx. \quad (8)$$

Starting from (5), we can expand the binomial term and rearrange the moment generating function into the following form:

$$\begin{aligned} M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \int_0^\infty x^{-3/2} \cdot \exp\left[tx - \frac{\alpha^2}{2x} + \alpha\gamma - \frac{\gamma^2 x}{2}\right] dx \\ &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp\left[\left(t - \frac{\gamma^2}{2}\right)x - \frac{\alpha^2}{2x}\right] dx \\ &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp\left[-\frac{1}{2}(\gamma^2 - 2t)x - \frac{1}{2} \cdot \frac{\alpha^2}{x}\right] dx. \end{aligned} \quad (9)$$

The integral now has the form of the integral in (8) with  $p = -1/2$ ,  $a = \gamma^2 - 2t$ , and  $b = \alpha^2$ . This allows us to write the moment-generating function in terms of the modified Bessel function  $K_{-1/2}$ :

$$M_X(t) = \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2 \left(\frac{\gamma^2 - 2t}{\alpha^2}\right)^{1/4} \cdot K_{-1/2}\left(\sqrt{\alpha^2(\gamma^2 - 2t)}\right). \quad (10)$$

Combining with (7) and simplifying gives

---

<sup>1</sup><https://dlmf.nist.gov/10.25>

<sup>2</sup><https://dlmf.nist.gov/10.39.2>

<sup>3</sup><https://dlmf.nist.gov/10.32.10>



$$\begin{aligned}
M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2 \left( \frac{\gamma^2 - 2t}{\alpha^2} \right)^{1/4} \cdot \sqrt{\frac{\pi}{2\sqrt{\alpha^2(\gamma^2 - 2t)}}} \cdot \exp \left[ -\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&= \frac{\alpha}{\sqrt{2} \cdot \sqrt{\pi}} \cdot e^{\alpha\gamma} \cdot 2 \cdot \frac{(\gamma^2 - 2t)^{1/4}}{\sqrt{\alpha}} \cdot \frac{\sqrt{\pi}}{\sqrt{2} \cdot \sqrt{\alpha} \cdot (\gamma^2 - 2t)^{1/4}} \cdot \exp \left[ -\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \quad (11) \\
&= e^{\alpha\gamma} \cdot \exp \left[ -\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right].
\end{aligned}$$

This finishes the proof of (2). ■

#### Sources:

- Siegrist, K. (2020): “The Wald Distribution”; in: *Random: Probability, Mathematical Statistics, Stochastic Processes*, retrieved on 2020-09-13; URL: <https://www.randomservices.org/random/special/Wald.html>.
- National Institute of Standards and Technology (2020): “NIST Digital Library of Mathematical Functions”, retrieved on 2020-09-13; URL: <https://dlmf.nist.gov>.

#### 3.10.4 Mean

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a Wald distribution ( $\rightarrow$  II/3.10.1):

$$X \sim \text{Wald}(\gamma, \alpha). \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \frac{\alpha}{\gamma}. \quad (2)$$

**Proof:** The mean or expected value  $E(X)$  is the first moment ( $\rightarrow$  I/1.21.1) of  $X$ , so we can use ( $\rightarrow$  I/1.21.2) the moment-generating function of the Wald distribution ( $\rightarrow$  II/3.10.3) to calculate

$$E(X) = M'_X(0). \quad (3)$$

First we differentiate

$$M_X(t) = \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \quad (4)$$

with respect to  $t$ . Using the chain rule gives

$$\begin{aligned}
M'_X(t) &= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} (\alpha^2(\gamma^2 - 2t))^{-1/2} \cdot -2\alpha^2 \\
&= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}}.
\end{aligned} \quad (5)$$

Evaluating (5) at  $t = 0$  gives the desired result:

$$\begin{aligned}
M'_X(0) &= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2(0))}} \\
&= \exp \left[ \alpha\gamma - \sqrt{\alpha^2 \cdot \gamma^2} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2 \cdot \gamma^2}} \\
&= \exp[0] \cdot \frac{\alpha^2}{\alpha\gamma} \\
&= \frac{\alpha}{\gamma} .
\end{aligned} \tag{6}$$

■

### 3.10.5 Variance

**Theorem:** Let  $X$  be a positive random variable ( $\rightarrow$  I/1.2.2) following a Wald distribution ( $\rightarrow$  II/3.10.1):

$$X \sim \text{Wald}(\gamma, \alpha) . \tag{1}$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \frac{\alpha}{\gamma^3} . \tag{2}$$

**Proof:** To compute the variance of  $X$ , we partition the variance into expected values ( $\rightarrow$  I/1.13.4):

$$\text{Var}(X) = E(X^2) - E(X)^2. \tag{3}$$

We then use the moment-generating function of the Wald distribution ( $\rightarrow$  II/3.10.3) to calculate

$$E(X^2) = M''_X(0) . \tag{4}$$

First we differentiate

$$M_X(t) = \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \tag{5}$$

with respect to  $t$ . Using the chain rule gives

$$\begin{aligned}
M'_X(t) &= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} (\alpha^2(\gamma^2 - 2t))^{-1/2} \cdot -2\alpha^2 \\
&= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}} \\
&= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1/2} .
\end{aligned} \tag{6}$$

Now we use the product rule to obtain the second derivative:

$$\begin{aligned}
M_X''(t) &= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1/2} \cdot -\frac{1}{2} (\alpha^2(\gamma^2 - 2t))^{-1/2} \cdot -2\alpha^2 \\
&\quad + \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} (\gamma^2 - 2t)^{-3/2} \cdot -2 \\
&= \alpha^2 \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1} \\
&\quad + \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-3/2} \\
&= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \left[ \frac{\alpha}{\gamma^2 - 2t} + \frac{1}{\sqrt{(\gamma^2 - 2t)^3}} \right].
\end{aligned} \tag{7}$$

Applying (4) yields

$$\begin{aligned}
E(X^2) &= M_X''(0) \\
&= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))} \right] \left[ \frac{\alpha}{\gamma^2 - 2(0)} + \frac{1}{\sqrt{(\gamma^2 - 2(0))^3}} \right] \\
&= \alpha \cdot \exp [\alpha\gamma - \alpha\gamma] \cdot \left[ \frac{\alpha}{\gamma^2} + \frac{1}{\gamma^3} \right] \\
&= \frac{\alpha^2}{\gamma^2} + \frac{\alpha}{\gamma^3}.
\end{aligned} \tag{8}$$

Since the mean of a Wald distribution ( $\rightarrow$  II/3.10.4) is given by  $E(X) = \alpha/\gamma$ , we can apply (3) to show

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - E(X)^2 \\
&= \frac{\alpha^2}{\gamma^2} + \frac{\alpha}{\gamma^3} - \left( \frac{\alpha}{\gamma} \right)^2 \\
&= \frac{\alpha}{\gamma^3}
\end{aligned} \tag{9}$$

which completes the proof of (2). ■

### 3.10.6 Skewness

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following a Wald distribution ( $\rightarrow$  II/3.10.1):

$$X \sim \text{Wald}(\gamma, \alpha). \tag{1}$$

Then the skewness ( $\rightarrow$  I/1.14.1) of  $X$  is

$$\text{Skew}(X) = \frac{3}{\sqrt{\alpha\gamma}}. \tag{2}$$

**Proof:** To compute the skewness of  $X$ , we partition the skewness into expected values ( $\rightarrow$  I/1.14.3):

$$\text{Skew}(X) = \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}, \quad (3)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $X$ , respectively. Since  $X$  follows a Wald distribution, the mean ( $\rightarrow$  II/3.10.4) of  $X$  is given by

$$\mu = E(X) = \frac{\alpha}{\gamma} \quad (4)$$

and the standard deviation ( $\rightarrow$  II/3.10.5) of  $X$  is given by

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\frac{\alpha}{\gamma^3}}. \quad (5)$$

Substituting (4) and (5) into (3) gives:

$$\begin{aligned} \text{Skew}(X) &= \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \\ &= \frac{E(X^3) - 3\left(\frac{\alpha}{\gamma}\right)\left(\frac{\alpha}{\gamma^3}\right) - \left(\frac{\alpha}{\gamma}\right)^3}{\left(\sqrt{\frac{\alpha}{\gamma^3}}\right)^3} \\ &= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[ E(X^3) - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3} \right]. \end{aligned} \quad (6)$$

Thus, the remaining work is to compute  $E(X^3)$ . To do this, we use the moment-generating function of the Wald distribution ( $\rightarrow$  II/3.10.3) to calculate

$$E(X^3) = M_X'''(0) \quad (7)$$

based on the relationship between raw moment and moment-generating function ( $\rightarrow$  I/1.21.2). First, we differentiate the moment-generating function

$$M_X(t) = \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \quad (8)$$

with respect to  $t$ . Using the chain rule, we have:

$$\begin{aligned} M_X'(t) &= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} (\alpha^2(\gamma^2 - 2t))^{-1/2} \cdot -2\alpha^2 \\ &= \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}} \\ &= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1/2}. \end{aligned} \quad (9)$$

Now we use the product rule to obtain the second derivative:

$$\begin{aligned}
M_X''(t) &= \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1/2} \cdot -\frac{1}{2} (\alpha^2(\gamma^2 - 2t))^{-1/2} \cdot -2\alpha^2 \\
&\quad + \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} (\gamma^2 - 2t)^{-3/2} \cdot -2 \\
&= \alpha^2 \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-1} \\
&\quad + \alpha \cdot \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot (\gamma^2 - 2t)^{-3/2} \\
&= \frac{\alpha^2}{\gamma^2 - 2t} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] + \frac{\alpha}{(\gamma^2 - 2t)^{3/2}} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] .
\end{aligned} \tag{10}$$

Finally, one more application of the chain rule will give us the third derivative. To start, we will decompose the second derivative obtained in (10) as

$$M''(t) = f(t) + g(t) \tag{11}$$

where

$$f(t) = \frac{\alpha^2}{\gamma^2 - 2t} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \tag{12}$$

and

$$g(t) = \frac{\alpha}{(\gamma^2 - 2t)^{3/2}} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] . \tag{13}$$

With this decomposition,  $M_X'(t) = f'(t) + g'(t)$ . Applying the product rule to  $f$  gives:

$$\begin{aligned}
f'(t) &= 2\alpha^2(\gamma^2 - 2t)^{-2} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&\quad + \alpha^2(\gamma^2 - 2t)^{-1} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} [\alpha^2(\gamma^2 - 2t)]^{-1/2} \cdot -2\alpha^2 \\
&= \frac{2\alpha^2}{(\gamma^2 - 2t)^2} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&\quad + \frac{\alpha^3}{(\gamma^2 - 2t)^{3/2}} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] .
\end{aligned} \tag{14}$$

Similarly, applying the product rule to  $g$  gives:

$$\begin{aligned}
g'(t) &= -\frac{3}{2}\alpha(\gamma^2 - 2t)^{-5/2}(-2) \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&\quad + \alpha(\gamma^2 - 2t)^{-3/2} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} [\alpha^2(\gamma^2 - 2t)]^{-1/2} \cdot -2\alpha^2 \\
&= \frac{3\alpha}{(\gamma^2 - 2t)^{5/2}} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&\quad + \frac{\alpha^2}{(\gamma^2 - 2t)^2} \exp \left[ \alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] .
\end{aligned} \tag{15}$$

Applying (7), together with (14) and (15), yields

$$\begin{aligned}
E(X^3) &= M_X'''(0) \\
&= f'(0) + g'(0) \\
&= \left[ \frac{2\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} \right] + \left[ \frac{3\alpha}{\gamma^5} + \frac{\alpha^2}{\gamma^4} \right] \\
&= \frac{3\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} + \frac{3\alpha}{\gamma^5} .
\end{aligned} \tag{16}$$

We now substitute (16) into (6), giving

$$\begin{aligned}
\text{Skew}(X) &= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[ E(X^3) - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3} \right] \\
&= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[ \frac{3\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} + \frac{3\alpha}{\gamma^5} - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3} \right] \\
&= \frac{\gamma^{9/2}}{\alpha^{3/2}} \cdot \frac{3\alpha}{\gamma^5} \\
&= \frac{3}{\alpha^{1/2} \cdot \gamma^{1/2}} \\
&= \frac{3}{\sqrt{\alpha\gamma}} .
\end{aligned} \tag{17}$$

This completes the proof of (2). ■

### 3.10.7 Method of moments

**Theorem:** Let  $y = \{y_1, \dots, y_n\}$  be a set of observed data independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a Wald distribution ( $\rightarrow$  II/3.10.1) with drift rate  $\gamma$  and threshold  $\alpha$ :

$$y_i \sim \text{Wald}(\gamma, \alpha), \quad i = 1, \dots, n . \tag{1}$$

Then, the method-of-moments estimates ( $\rightarrow$  I/4.1.8) for the parameters  $\gamma$  and  $\alpha$  are given by

$$\begin{aligned}
\hat{\gamma} &= \sqrt{\frac{\bar{y}}{\bar{v}}} \\
\hat{\alpha} &= \sqrt{\frac{\bar{y}^3}{\bar{v}}}
\end{aligned} \tag{2}$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2) and  $\bar{v}$  is the unbiased sample variance ( $\rightarrow$  I/1.13.2):

$$\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\
\bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 .
\end{aligned} \tag{3}$$

**Proof:** The mean ( $\rightarrow$  II/3.10.4) and variance ( $\rightarrow$  II/3.10.5) of the Wald distribution ( $\rightarrow$  II/3.10.1) in terms of the parameters  $\gamma$  and  $\alpha$  are given by

$$\begin{aligned} E(X) &= \frac{\alpha}{\gamma} \\ \text{Var}(X) &= \frac{\alpha}{\gamma^3} . \end{aligned} \tag{4}$$

Thus, matching the moments ( $\rightarrow$  I/4.1.8) requires us to solve the following system of equations for  $\gamma$  and  $\alpha$ :

$$\begin{aligned} \bar{y} &= \frac{\alpha}{\gamma} \\ \bar{v} &= \frac{\alpha}{\gamma^3} . \end{aligned} \tag{5}$$

To this end, our first step is to express the second equation of (5) as follows:

$$\begin{aligned} \bar{v} &= \frac{\alpha}{\gamma^3} \\ &= \frac{\alpha}{\gamma} \cdot \gamma^{-2} \\ &= \bar{y} \cdot \gamma^{-2} . \end{aligned} \tag{6}$$

Rearranging (6) gives

$$\gamma^2 = \frac{\bar{y}}{\bar{v}} , \tag{7}$$

or equivalently,

$$\gamma = \sqrt{\frac{\bar{y}}{\bar{v}}} . \tag{8}$$

Our final step is to solve the first equation of (5) for  $\alpha$  and substitute (8) for  $\gamma$ :

$$\begin{aligned} \alpha &= \bar{y} \cdot \gamma \\ &= \bar{y} \cdot \sqrt{\frac{\bar{y}}{\bar{v}}} \\ &= \sqrt{\bar{y}^2} \cdot \sqrt{\frac{\bar{y}}{\bar{v}}} \\ &= \sqrt{\frac{\bar{y}^3}{\bar{v}}} . \end{aligned} \tag{9}$$

Together, (8) and (9) constitute the method-of-moment estimates of  $\gamma$  and  $\alpha$ .

■

### 3.11 ex-Gaussian distribution

#### 3.11.1 Definition

**Definition:** Let  $A$  be a random variable ( $\rightarrow$  I/1.2.2) that is normally distributed ( $\rightarrow$  II/3.2.1) with mean  $\mu$  and variance  $\sigma^2$ , and let  $B$  be a random variable that is exponentially distributed ( $\rightarrow$  II/3.5.1) with rate  $\lambda$ . Suppose further that  $A$  and  $B$  are independent ( $\rightarrow$  I/1.3.6). Then the sum  $X = A + B$  is said to have an exponentially-modified Gaussian (i.e., ex-Gaussian) distribution, with parameters  $\mu$ ,  $\sigma$ , and  $\lambda$ ; that is,

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda), \quad (1)$$

where  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ , and  $\lambda > 0$ .

#### Sources:

- Luce, R. D. (1986): “Response Times: Their Role in Inferring Elementary Mental Organization”, 35-36; URL: <https://global.oup.com/academic/product/response-times-9780195036428>.

#### 3.11.2 Probability density function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda). \quad (1)$$

Then the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(t) = \frac{\lambda}{\sqrt{2\pi}} \exp \left[ \frac{\lambda^2 \sigma^2}{2} - \lambda(t - \mu) \right] \cdot \int_{-\infty}^{\frac{t-\mu}{\sigma} - \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 \right] dy. \quad (2)$$

**Proof:** Suppose  $X$  follows an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1). Then  $X = A + B$ , where  $A$  and  $B$  are independent ( $\rightarrow$  I/1.3.6),  $A$  is normally distributed ( $\rightarrow$  II/3.2.1) with mean ( $\rightarrow$  II/3.2.16)  $\mu$  and variance ( $\rightarrow$  II/3.2.19)  $\sigma^2$ , and  $B$  is exponentially distributed ( $\rightarrow$  II/3.5.1) with rate  $\lambda$ . Then, the probability density function ( $\rightarrow$  II/3.2.10) for  $A$  is given by

$$f_A(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{t - \mu}{\sigma} \right)^2 \right], \quad (3)$$

and the probability density function ( $\rightarrow$  II/3.5.3) for  $B$  is given by

$$f_B(t) = \begin{cases} \lambda \exp[-\lambda t], & \text{if } t \geq 0 \\ 0, & \text{if } t < 0. \end{cases} \quad (4)$$

Thus, the probability density function for the sum ( $\rightarrow$  I/1.7.2)  $X = A + B$  is given by taking the convolution of  $f_A$  and  $f_B$ :



$$\begin{aligned}
f_X(t) &= \int_{-\infty}^{\infty} f_A(x)f_B(t-x)dx \\
&= \int_{-\infty}^t f_A(x)f_B(t-x)dx + \int_t^{\infty} f_A(x)f_B(t-x)dx \\
&= \int_{-\infty}^t f_A(x)f_B(t-x)dx ,
\end{aligned} \tag{5}$$

which follows from the fact that  $f_B(t-x) = 0$  for  $x > t$ . From here, we substitute the expressions (3) and (4) for the probability density functions  $f_A$  and  $f_B$  in (5):

$$\begin{aligned}
f_X(t) &= \int_{-\infty}^t \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \cdot \lambda \exp[-\lambda(t-x)]dx \\
&= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^t \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \cdot \exp[-\lambda t + \lambda x]dx \\
&= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^t \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \cdot \exp[-\lambda t] \cdot \exp[\lambda x]dx \\
&= \frac{\lambda \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^t \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 + \lambda x \right] dx .
\end{aligned} \tag{6}$$

We can further simplify the integrand with a substitution; to this end, let

$$y = g(x) = \frac{x-\mu}{\sigma} - \lambda\sigma \tag{7}$$

This gives the following three identities:

$$\frac{dy}{dx} = \frac{1}{\sigma} , \quad \text{or equivalently,} \quad dx = \sigma dy , \tag{8}$$

$$\frac{x-\mu}{\sigma} = y + \lambda\sigma , \quad \text{and} \tag{9}$$

$$x = y\sigma + \lambda\sigma^2 + \mu . \tag{10}$$

Substituting these identities into (6) gives

$$\begin{aligned}
f_X(t) &= \frac{\lambda \exp[-\lambda t]}{\sigma \sqrt{2\pi}} \int_{-\infty}^{g(t)} \exp \left[ -\frac{1}{2}(y + \lambda\sigma)^2 + \lambda(y\sigma + \lambda\sigma^2 + \mu) \right] \sigma dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}(y^2 + 2y\lambda\sigma + \lambda^2\sigma^2) + \lambda y\sigma + \lambda^2\sigma^2 + \lambda\mu \right] dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 - y\lambda\sigma - \frac{\lambda^2\sigma^2}{2} + \lambda y\sigma + \lambda^2\sigma^2 + \lambda\mu \right] dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 \right] \cdot \exp \left[ \frac{\lambda^2\sigma^2}{2} + \lambda\mu \right] dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \cdot \exp \left[ \frac{\lambda^2\sigma^2}{2} + \lambda\mu \right] \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 \right] \cdot dy \\
&= \frac{\lambda}{\sqrt{2\pi}} \cdot \exp \left[ -\lambda t + \frac{\lambda^2\sigma^2}{2} + \lambda\mu \right] \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 \right] \cdot dy \\
&= \frac{\lambda}{\sqrt{2\pi}} \cdot \exp \left[ \frac{\lambda^2\sigma^2}{2} - \lambda(t - \mu) \right] \int_{-\infty}^{\frac{x-\mu}{\sigma} + \lambda\sigma} \exp \left[ -\frac{1}{2}y^2 \right] \cdot dy .
\end{aligned} \tag{11}$$

This finishes the proof of (2). ■

### 3.11.3 Moment-generating function

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) . \tag{1}$$

Then, the moment generating function ( $\rightarrow$  I/1.9.1) of  $X$  is

$$M_X(t) = \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] . \tag{2}$$

**Proof:** Suppose  $X$  follows an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1). Then,  $X = A + B$  where  $A$  and  $B$  are independent ( $\rightarrow$  I/1.3.6),  $A$  is normally distributed ( $\rightarrow$  II/3.2.1) with mean ( $\rightarrow$  II/3.2.16)  $\mu$  and variance ( $\rightarrow$  II/3.2.19)  $\sigma^2$ , and  $B$  is exponentially distributed ( $\rightarrow$  II/3.5.1) with rate  $\lambda$ . Then the moment generating function ( $\rightarrow$  II/3.2.11) for  $A$  is given by

$$M_A(t) = \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \tag{3}$$

and the moment generating function ( $\rightarrow$  II/3.5.4) for  $B$  is given by

$$M_B(t) = \frac{\lambda}{\lambda - t} . \tag{4}$$

By definition,  $X$  is a linear combination of independent random variables  $A$  and  $B$ , so the moment generating function ( $\rightarrow$  I/1.9.5) of  $X$  is the product of  $M_A(t)$  and  $M_B(t)$ . That is,

$$\begin{aligned}
M_X(t) &= M_A(t) \cdot M_B(t) \\
&= \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \cdot \left( \frac{\lambda}{\lambda - t} \right) \\
&= \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] .
\end{aligned} \tag{5}$$

This finishes the proof of (2). ■

### 3.11.4 Mean

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) . \tag{1}$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = \mu + \frac{1}{\lambda} . \tag{2}$$

**Proof:** The mean or expected value  $E(X)$  is the first raw moment ( $\rightarrow$  I/1.21.1) of  $X$ , so we can use ( $\rightarrow$  I/1.21.2) the moment-generating function of the ex-Gaussian distribution ( $\rightarrow$  II/3.11.3) to calculate

$$E(X) = M'_X(0) . \tag{3}$$

First, we differentiate

$$M_X(t) = \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \tag{4}$$

with respect to  $t$ . Using the product rule and chain rule gives:

$$\begin{aligned}
M'_X(t) &= \frac{\lambda}{(\lambda - t)^2} \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] + \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] (\mu + \sigma^2 t) \\
&= \left( \frac{\lambda}{\lambda - t} \right) \cdot \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] .
\end{aligned} \tag{5}$$

Evaluating (5) at  $t = 0$  gives the desired result:

$$\begin{aligned}
M'_X(0) &= \left( \frac{\lambda}{\lambda - 0} \right) \cdot \exp \left[ \mu \cdot 0 + \frac{1}{2} \sigma^2 \cdot 0^2 \right] \cdot \left[ \frac{1}{\lambda - 0} + \mu + \sigma^2 \cdot 0 \right] \\
&= 1 \cdot 1 \cdot \left[ \frac{1}{\lambda} + \mu \right] \\
&= \mu + \frac{1}{\lambda} .
\end{aligned} \tag{6}$$

■

### 3.11.5 Variance

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) . \quad (1)$$

Then, the variance ( $\rightarrow$  I/1.13.1) of  $X$  is

$$\text{Var}(X) = \sigma^2 + \frac{1}{\lambda^2} . \quad (2)$$

**Proof:** To compute the variance of  $X$ , we partition the variance into expected values ( $\rightarrow$  I/1.13.4):

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

We then use the moment-generating function of the ex-Gaussian distribution ( $\rightarrow$  II/3.11.3) to calculate

$$E(X^2) = M_X''(0) \quad (4)$$

based on the relationship between raw moment and moment-generating function ( $\rightarrow$  I/1.21.2). First, we differentiate

$$M_X(t) = \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \quad (5)$$

with respect to  $t$ . Using the product rule and chain rule gives:

$$\begin{aligned} M_X'(t) &= \frac{\lambda}{(\lambda - t)^2} \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] + \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] (\mu + \sigma^2 t) \\ &= \left( \frac{\lambda}{\lambda - t} \right) \cdot \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] \\ &= M_X(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] . \end{aligned} \quad (6)$$

We now use the product rule to obtain the second derivative:

$$\begin{aligned} M_X''(t) &= M_X'(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] + M_X(t) \cdot \left[ \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \\ &\stackrel{(6)}{=} M_X(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right]^2 + M_X(t) \cdot \left[ \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \\ &= M_X(t) \cdot \left[ \left( \frac{1}{\lambda - t} + \mu + \sigma^2 t \right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \\ &= \left( \frac{\lambda}{\lambda - t} \right) \cdot \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \cdot \left[ \left( \frac{1}{\lambda - t} + \mu + \sigma^2 t \right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \end{aligned} \quad (7)$$

Applying (4) yields

$$\begin{aligned}
E(X^2) &= M_X''(0) \\
&= \left( \frac{\lambda}{\lambda - 0} \right) \cdot \exp \left[ \mu \cdot 0 + \frac{1}{2} \sigma^2 \cdot 0^2 \right] \cdot \left[ \left( \frac{1}{\lambda - 0} + \mu + \sigma^2 \cdot 0 \right)^2 + \frac{1}{(\lambda - 0)^2} + \sigma^2 \right] \\
&= 1 \cdot 1 \cdot \left[ \left( \frac{1}{\lambda} + \mu \right)^2 + \frac{1}{\lambda^2} + \sigma^2 \right] \\
&= \frac{1}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \frac{1}{\lambda^2} + \sigma^2 \\
&= \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 .
\end{aligned} \tag{8}$$

Since the mean of an ex-Gaussian distribution ( $\rightarrow$  II/3.11.4) is given by

$$E(X) = \mu + \frac{1}{\lambda} , \tag{9}$$

we can apply (3) to show

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - E(X)^2 \\
&= \left[ \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 \right] - \left( \mu + \frac{1}{\lambda} \right)^2 \\
&= \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 - \mu^2 - \frac{2\mu}{\lambda} - \frac{1}{\lambda^2} \\
&= \sigma^2 + \frac{1}{\lambda^2} .
\end{aligned} \tag{10}$$

This completes the proof of (2). ■

### 3.11.6 Skewness

**Theorem:** Let  $X$  be a random variable ( $\rightarrow$  I/1.2.2) following an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) . \tag{1}$$

Then the skewness ( $\rightarrow$  I/1.14.1) of  $X$  is

$$\text{Skew}(X) = \frac{2}{\lambda^3 \left( \sigma^2 + \frac{1}{\lambda^2} \right)^{\frac{3}{2}}} . \tag{2}$$

**Proof:**

To compute the skewness of  $X$ , we partition the skewness into expected values ( $\rightarrow$  I/1.14.3):

$$\text{Skew}(X) = \frac{E(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} , \tag{3}$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $X$ , respectively. To prevent confusion between the labels used for the ex-Gaussian parameters in (1) and the mean and standard deviation of  $X$ , we rewrite (3) as

$$\text{Skew}(X) = \frac{E(X^3) - 3 \cdot E(X) \cdot \text{Var}(X) - E(X)^3}{\text{Var}(X)^{\frac{3}{2}}}. \quad (4)$$

Since  $X$  follows an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1), the mean ( $\rightarrow$  II/3.11.4) of  $X$  is given by

$$E(X) = \mu + \frac{1}{\lambda} \quad (5)$$

and the variance ( $\rightarrow$  II/3.11.5) of  $X$  is given by

$$\text{Var}(X) = \sigma^2 + \frac{1}{\lambda^2}. \quad (6)$$

Thus, the primary work is to compute  $E(X^3)$ . To do this, we use the moment-generating function of the ex-Gaussian distribution ( $\rightarrow$  II/3.11.3) to calculate

$$E(X^3) = M_X'''(0) \quad (7)$$

based on the relationship between raw moment and moment-generating function ( $\rightarrow$  I/1.21.2).

First, we differentiate the moment-generating function of the ex-Gaussian distribution ( $\rightarrow$  II/3.11.3)

$$M_X(t) = \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \quad (8)$$

with respect to  $t$ . Using the product rule and chain rule, we have:

$$\begin{aligned} M_X'(t) &= \frac{\lambda}{(\lambda - t)^2} \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] + \left( \frac{\lambda}{\lambda - t} \right) \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] (\mu + \sigma^2 t) \\ &= \left( \frac{\lambda}{\lambda - t} \right) \cdot \exp \left[ \mu t + \frac{1}{2} \sigma^2 t^2 \right] \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] \\ &= M_X(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right]. \end{aligned} \quad (9)$$

We then use the product rule to obtain the second derivative:

$$\begin{aligned} M_X''(t) &= M_X'(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right] + M_X(t) \cdot \left[ \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \\ &= M_X(t) \cdot \left[ \frac{1}{\lambda - t} + \mu + \sigma^2 t \right]^2 + M_X(t) \cdot \left[ \frac{1}{(\lambda - t)^2} + \sigma^2 \right] \\ &= M_X(t) \cdot \left[ \left( \frac{1}{\lambda - t} + \mu + \sigma^2 t \right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2 \right]. \end{aligned} \quad (10)$$

Finally, we use the product rule and chain rule to obtain the third derivative:

$$M_X'''(t) = M_X'(t) \left[ \left( \frac{1}{\lambda - t} + \mu + \sigma^2 t \right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2 \right] + M_X(t) \left[ 2 \left( \frac{1}{\lambda - t} + \mu + \sigma^2 t \right) \left( \frac{1}{(\lambda - t)^2} + \sigma^2 \right) + \frac{1}{(\lambda - t)^3} \right] \quad (11)$$

Applying (7), together with (11), yields:

$$\begin{aligned}
 E(X^3) &= M_X'''(0) \\
 &= M_X'(0) \left[ \left( \frac{1}{\lambda} + \mu \right)^2 + \frac{1}{\lambda^2} + \sigma^2 \right] + M_X(0) \left[ 2 \left( \frac{1}{\lambda} + \mu \right) \left( \frac{1}{\lambda^2} + \sigma^2 \right) + \frac{2}{\lambda^3} \right] \\
 &= \left( \mu + \frac{1}{\lambda} \right) \left( \frac{1}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \frac{1}{\lambda^2} + \sigma^2 \right) + \left( \frac{2}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 + \frac{2}{\lambda^3} \right) \\
 &= \left( \mu + \frac{1}{\lambda} \right) \left( \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 \right) + \left( \frac{4}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 \right) \\
 &= \frac{2\mu}{\lambda^2} + \frac{2\mu^2}{\lambda} + \mu^3 + \mu\sigma^2 + \frac{2}{\lambda^3} + \frac{2\mu}{\lambda^2} + \frac{\mu^2}{\lambda} + \frac{\sigma^2}{\lambda} + \frac{4}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 \\
 &= \frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3.
 \end{aligned} \tag{12}$$

We now substitute (12), (5), and (6) into the numerator of (4), giving

$$\begin{aligned}
 E(X^3) - 3 \cdot E(X) \cdot \text{Var}(X) - E(X)^3 &= \left( \frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3 \right) - 3 \left( \mu + \frac{1}{\lambda} \right) \left( \sigma^2 + \frac{1}{\lambda^2} \right) - \left( \mu + \frac{1}{\lambda} \right)^3 \\
 &= \frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3 - 3\mu\sigma^2 - \frac{3\mu}{\lambda^2} - \frac{3\sigma^2}{\lambda} - \frac{3}{\lambda^3} - \mu^3 - \frac{3}{\lambda} \\
 &= \frac{2}{\lambda^3}.
 \end{aligned} \tag{13}$$

Thus, we have:

$$\begin{aligned}
 \text{Skew}(X) &= \frac{E(X^3) - 3 \cdot E(X) \cdot \text{Var}(X) - E(X)^3}{\text{Var}(X)^{\frac{3}{2}}} \\
 &= \frac{\frac{2}{\lambda^3}}{\left( \sigma^2 + \frac{1}{\lambda^2} \right)^{\frac{3}{2}}} \\
 &= \frac{2}{\lambda^3 \left( \sigma^2 + \frac{1}{\lambda^2} \right)^{\frac{3}{2}}}.
 \end{aligned} \tag{14}$$

This completes the proof of (2). ■

### 3.11.7 Method of moments

**Theorem:** Let  $y = \{y_1, \dots, y_n\}$  be a set of observed data independent and identically distributed ( $\rightarrow$  I/1.2.8) according to an ex-Gaussian distribution ( $\rightarrow$  II/3.11.1) with parameters  $\mu$ ,  $\sigma$ , and  $\lambda$ :

$$y_i \sim \text{ex-Gaussian}(\mu, \sigma, \lambda), \quad i = 1, \dots, n. \tag{1}$$

Then, the method-of-moments estimates ( $\rightarrow$  I/4.1.8) for the parameters  $\mu$ ,  $\sigma$ , and  $\lambda$  are given by

$$\begin{aligned}
\hat{\mu} &= \bar{y} - \sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} \\
\hat{\sigma} &= \sqrt{\bar{v} \cdot \left(1 - \sqrt[3]{\frac{\bar{s}^2}{4}}\right)} \\
\hat{\lambda} &= \sqrt[3]{\frac{2}{\bar{s} \cdot \bar{v}^{3/2}}} ,
\end{aligned} \tag{2}$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2),  $\bar{v}$  is the sample variance ( $\rightarrow$  I/1.13.2) and  $\bar{s}$  is the sample skewness ( $\rightarrow$  I/1.14.2)

$$\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\
\bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\
\bar{s} &= \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^3}{\left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2\right]^{3/2}} .
\end{aligned} \tag{3}$$

**Proof:** The mean ( $\rightarrow$  II/3.11.4), variance ( $\rightarrow$  II/3.11.5), and skewness ( $\rightarrow$  II/3.11.6) of the ex-Gaussian distribution ( $\rightarrow$  II/3.11.1) in terms of the parameters  $\mu$ ,  $\sigma$ , and  $\lambda$  are given by

$$\begin{aligned}
E(X) &= \mu + \frac{1}{\lambda} \\
\text{Var}(X) &= \sigma^2 + \frac{1}{\lambda^2} \\
\text{Skew}(X) &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2}\right)^{3/2}} .
\end{aligned} \tag{4}$$

Thus, matching the moments ( $\rightarrow$  I/4.1.8) requires us to solve the following system of equations for  $\mu$ ,  $\sigma$ , and  $\lambda$ :

$$\begin{aligned}
\bar{y} &= \mu + \frac{1}{\lambda} \\
\bar{v} &= \sigma^2 + \frac{1}{\lambda^2} \\
\bar{s} &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2}\right)^{3/2}} .
\end{aligned} \tag{5}$$

To this end, our first step is to substitute the second equation of (5) into the third equation:

$$\begin{aligned}
\bar{s} &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2}\right)^{3/2}} \\
&= \frac{2}{\lambda^3 \cdot \bar{v}^{3/2}} .
\end{aligned} \tag{6}$$



Re-expressing (6) in terms of  $\lambda^3$  and taking the cube root gives:

$$\lambda = \sqrt[3]{\frac{2}{\bar{s} \cdot \bar{v}^{3/2}}} . \quad (7)$$

Next, we solve the first equation of (5) for  $\mu$  and substitute (7):

$$\begin{aligned} \mu &= \bar{y} - \frac{1}{\lambda} \\ &= \bar{y} - \sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} . \end{aligned} \quad (8)$$

Finally, we solve the second equation of (5) for  $\sigma$ :

$$\sigma^2 = \bar{v} - \frac{1}{\lambda^2} . \quad (9)$$

Taking the square root gives and substituting (7) gives:

$$\begin{aligned} \sigma &= \sqrt{\bar{v} - \frac{1}{\lambda^2}} \\ &= \sqrt{\bar{v} - \left( \sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} \right)^2} \\ &= \sqrt{\bar{v} - \bar{v} \cdot \sqrt[3]{\frac{\bar{s}^2}{4}}} \\ &= \sqrt{\bar{v} \cdot \left( 1 - \sqrt[3]{\frac{\bar{s}^2}{4}} \right)} . \end{aligned} \quad (10)$$

Together, (8), (10), and (7) constitute the method-of-moment estimates of  $\mu$ ,  $\sigma$ , and  $\lambda$ .

■

## 4 Multivariate continuous distributions

### 4.1 Multivariate normal distribution

#### 4.1.1 Definition

**Definition:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to be multivariate normally distributed with mean  $\mu$  and covariance  $\Sigma$

$$X \sim \mathcal{N}(\mu, \Sigma), \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \quad (2)$$

where  $\mu$  is an  $n$ -dimensional real vector and  $\Sigma$  is an  $n \times n$  positive-definite matrix.

#### Sources:

- Koch KR (2007): “Multivariate Normal Distribution”; in: *Introduction to Bayesian Statistics*, ch. 2.5.1, pp. 51-53, eq. 2.195; URL: <https://www.springer.com/gp/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

#### 4.1.2 Special case of matrix-normal distribution

**Theorem:** The multivariate normal distribution ( $\rightarrow$  II/4.1.1) is a special case of the matrix-normal distribution ( $\rightarrow$  II/5.1.1) with number of variables  $p = 1$ , i.e. random matrix ( $\rightarrow$  I/1.2.2)  $X = x \in \mathbb{R}^{n \times 1}$ , mean  $M = \mu \in \mathbb{R}^{n \times 1}$ , covariance across rows  $U = \Sigma$  and covariance across columns  $V = 1$ .

**Proof:** The probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3) is

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right]. \quad (1)$$

Setting  $p = 1$ ,  $X = x$ ,  $M = \mu$ ,  $U = \Sigma$  and  $V = 1$ , we obtain

$$\begin{aligned} \mathcal{MN}(x; \mu, \Sigma, 1) &= \frac{1}{\sqrt{(2\pi)^n |1|^n |\Sigma|^1}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (1^{-1} (x - \mu)^T \Sigma^{-1} (x - \mu)) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \end{aligned} \quad (2)$$

which is equivalent to the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4). ■

#### Sources:

- Wikipedia (2022): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-31; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Definition](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Definition).

### 4.1.3 Relationship to chi-squared distribution

**Theorem:** Let  $x$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) following a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with zero mean ( $\rightarrow$  I/1.12.16) and arbitrary covariance matrix ( $\rightarrow$  II/4.1.8)  $\Sigma$ :

$$x \sim \mathcal{N}(0, \Sigma) . \quad (1)$$

Then, the quadratic form of  $x$ , weighted by  $\Sigma$ , follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $n$  degrees of freedom ( $\rightarrow$  I/1.5.12):

$$y = x^T \Sigma^{-1} x \sim \chi^2(n) . \quad (2)$$

**Proof:** Define a new random vector ( $\rightarrow$  I/1.2.3)  $z$  as

$$z = \Sigma^{-1/2} x . \quad (3)$$

where  $\Sigma^{-1/2}$  is the matrix square root of  $\Sigma$ . This matrix must exist, because  $\Sigma$  is a covariance matrix ( $\rightarrow$  I/1.16.1) and thus positive semi-definite ( $\rightarrow$  I/1.16.5). Due to the linear transformation theorem ( $\rightarrow$  II/4.1.14),  $z$  is distributed as

$$\begin{aligned} z &\sim \mathcal{N}\left(\Sigma^{-1/2}0, \Sigma^{-1/2}\Sigma\Sigma^{-1/2T}\right) \\ &\sim \mathcal{N}\left(\Sigma^{-1/2}0, \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\Sigma^{-1/2}\right) \\ &\sim \mathcal{N}(0, I_n) , \end{aligned} \quad (4)$$

i.e. each entry of this vector follows ( $\rightarrow$  II/4.1.15) a standard normal distribution ( $\rightarrow$  II/3.2.3):

$$z_i \sim \mathcal{N}(0, 1) \quad \text{for all } i = 1, \dots, n . \quad (5)$$

We further observe that  $y$  can be represented in terms of  $z$

$$y = x^T \Sigma^{-1} x = (x^T \Sigma^{-1/2}) (\Sigma^{-1/2} x) = z^T z , \quad (6)$$

thus  $z$  is a sum of  $n$  squared standard normally distributed ( $\rightarrow$  II/3.2.3) random variables ( $\rightarrow$  I/1.2.2)

$$y = \sum_{i=1}^n z_i^2 \quad \text{where all } z_i \sim \mathcal{N}(0, 1) \quad (7)$$

which, by definition, is chi-squared distributed ( $\rightarrow$  II/3.7.1) with  $n$  degrees of freedom:

$$y \sim \chi^2(n) . \quad (8)$$

■

#### Sources:

- Koch KR (2007): “Chi-Squared Distribution”; in: *Introduction to Bayesian Statistics*, ch. 2.4.5, pp. 48-49, eq. 2.180; URL: <https://www.springer.com/gp/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

#### 4.1.4 Probability density function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right] . \quad (2)$$

**Proof:** This follows directly from the definition of the multivariate normal distribution ( $\rightarrow$  II/4.1.1). ■

#### 4.1.5 Moment-generating function

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the moment-generating function ( $\rightarrow$  I/1.9.1) of  $x$  is

$$M_x(t) = \exp \left[ t^T \mu + \frac{1}{2} t^T \Sigma t \right] . \quad (2)$$

**Proof:** The moment-generating function of a random vector ( $\rightarrow$  I/1.9.1)  $X$  is defined as:

$$M_X(t) = \mathbb{E} \left[ e^{t^T X} \right], \quad t \in \mathbb{R}^n . \quad (3)$$

Applying the law of the unconscious statistician ( $\rightarrow$  I/1.12.9), we have:

$$M_x(t) = \int_{\mathcal{X}} e^{t^T x} \cdot f_X(x) \, dx . \quad (4)$$

With the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), we have:

$$M_x(t) = \int_{\mathbb{R}^n} \exp [t^T x] \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right] \, dx . \quad (5)$$

Now we summarize the two exponential functions inside the integral:

$$\begin{aligned}
M_x(t) &= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) + t^T x \right] dx \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu - 2t^T x) \right] dx \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x^T \Sigma^{-1} x - 2(\mu + \Sigma t)^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu) \right] dx \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} ((x - [\mu + \Sigma t])^T \Sigma^{-1} (x - [\mu + \Sigma t]) - 2t^T \mu - t^T \Sigma t) \right] dx \\
&= \exp \left[ t^T \mu + \frac{1}{2} t^T \Sigma t \right] \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - [\mu + \Sigma t])^T \Sigma^{-1} (x - [\mu + \Sigma t]) \right] dx .
\end{aligned} \tag{6}$$

The integrand is equal to the probability density function of a multivariate normal distribution ( $\rightarrow$  II/4.1.4):

$$M_x(t) = \exp \left[ t^T \mu + \frac{1}{2} t^T \Sigma t \right] \int_{\mathbb{R}^n} \mathcal{N}(x; \mu + \Sigma t, \Sigma) dx . \tag{7}$$

Because the entire probability density integrates to one ( $\rightarrow$  I/1.7.1), we finally have:

$$M_x(t) = \exp \left[ t^T \mu + \frac{1}{2} t^T \Sigma t \right] . \tag{8}$$

■

#### 4.1.6 Mean

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \tag{1}$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $x$  is

$$E(x) = \mu . \tag{2}$$

**Proof:**

1) First, consider a set of independent ( $\rightarrow$  I/1.3.6) and standard normally ( $\rightarrow$  II/3.2.3) distributed random variables ( $\rightarrow$  I/1.2.2):

$$z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad i = 1, \dots, n . \tag{3}$$

Then, these variables together form a multivariate normally ( $\rightarrow$  II/4.1.17) distributed random vector ( $\rightarrow$  I/1.2.3):

$$z \sim \mathcal{N}(0_n, I_n) . \tag{4}$$

By definition, the expected value of a random vector is equal to the vector of all expected values ( $\rightarrow$  I/1.12.16):

$$\mathbb{E}(z) = \mathbb{E} \left( \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \right) = \begin{bmatrix} \mathbb{E}(z_1) \\ \vdots \\ \mathbb{E}(z_n) \end{bmatrix} . \quad (5)$$

Because the expected value of all its entries is zero ( $\rightarrow$  II/3.2.16), the expected value of the random vector is

$$\mathbb{E}(z) = \begin{bmatrix} \mathbb{E}(z_1) \\ \vdots \\ \mathbb{E}(z_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = 0_n . \quad (6)$$

2) Next, consider an  $n \times n$  matrix  $A$  solving the equation  $AA^T = \Sigma$ . Such a matrix exists, because  $\Sigma$  is defined to be positive definite ( $\rightarrow$  II/4.1.1). Then,  $x$  can be represented as a linear transformation ( $\rightarrow$  II/4.1.14) of  $z$ :

$$x = Az + \mu \sim \mathcal{N}(A0_n + \mu, AI_nA^T) = \mathcal{N}(\mu, \Sigma) . \quad (7)$$

Thus, the expected value ( $\rightarrow$  I/1.12.1) of  $x$  can be written as:

$$\mathbb{E}(x) = \mathbb{E}(Az + \mu) . \quad (8)$$

With the linearity of the expected value ( $\rightarrow$  I/1.12.5), this becomes:

$$\begin{aligned} \mathbb{E}(x) &= \mathbb{E}(Az + \mu) \\ &= \mathbb{E}(Az) + \mathbb{E}(\mu) \\ &= A \mathbb{E}(z) + \mu \\ &\stackrel{(6)}{=} A 0_n + \mu \\ &= \mu . \end{aligned} \quad (9)$$

■

#### Sources:

- Taboga, Marco (2021): “Multivariate normal distribution”; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-09-15; URL: <https://www.statlect.com/probability-distributions/multivariate-normal-distribution>.

#### 4.1.7 Mode

**Theorem:** Let  $X$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the mode ( $\rightarrow$  I/1.18.3) of  $X$  is

$$\text{mode}(X) = \mu . \quad (2)$$

**Proof:** The mode ( $\rightarrow$  I/1.18.3) is the value which maximizes the probability density function ( $\rightarrow$  I/1.7.1):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) is:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] . \quad (4)$$

The gradient of this function  $\nabla_x f_X(x) \in \mathbb{R}^n$  is:

$$\begin{aligned} \nabla_x f_X(x) &= f_X(x) \cdot \nabla_x \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \\ &= f_X(x) \cdot \left[ -\frac{1}{2} (2 \Sigma^{-1} (x - \mu)) \right] \\ &= f_X(x) \cdot [-\Sigma^{-1} (x - \mu)] . \end{aligned} \quad (5)$$

The Hessian of this function  $\nabla_x^2 f_X(x) \in \mathbb{R}^{n \times n}$  is:

$$\begin{aligned} \nabla_x^2 f_X(x) &= \nabla_x f_X(x) \cdot [-\Sigma^{-1} (x - \mu)]^T + f_X(x) \cdot \nabla_x [-\Sigma^{-1} (x - \mu)] \\ &= f_X(x) \cdot [-\Sigma^{-1} (x - \mu)] \cdot [-(x - \mu)^T \Sigma^{-1}] + f_X(x) \cdot [-\Sigma^{-1}] \\ &= f_X(x) \cdot [\Sigma^{-1} (x - \mu)(x - \mu)^T \Sigma^{-1} - \Sigma^{-1}] . \end{aligned} \quad (6)$$

Setting the gradient (5) to zero, we calculate the root:

$$\begin{aligned} \nabla_x f_X(x) = 0 &= f_X(x) \cdot [-\Sigma^{-1} (x - \mu)] \\ \Leftrightarrow 0 &= -\Sigma^{-1} (x - \mu) \\ \Leftrightarrow x &= \mu . \end{aligned} \quad (7)$$

Plugging this value into the Hessian (6), we obtain:

$$\begin{aligned} \nabla_x^2 f_X(\mu) &= f_X(\mu) \cdot [\Sigma^{-1} (\mu - \mu)(\mu - \mu)^T \Sigma^{-1} - \Sigma^{-1}] \\ &= -f_X(\mu) \cdot \Sigma^{-1} . \end{aligned} \quad (8)$$

Since  $\Sigma$  is positive-definite ( $\rightarrow$  II/4.1.1),  $\Sigma^{-1}$  is also positive-definite ( $\rightarrow$  II/4.1.1) and since  $f_X(\mu)$  is positive,  $-f_X(\mu) \cdot \Sigma^{-1}$  is a negative-definite matrix, so that  $f_X(\mu)$  is a maximum. Taken together, this shows that

$$\text{mode}(X) = \mu . \quad (9)$$

■

### 4.1.8 Covariance

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the covariance matrix ( $\rightarrow$  I/1.16.1) of  $x$  is

$$\text{Cov}(x) = \Sigma . \quad (2)$$

**Proof:**

1) First, consider a set of independent ( $\rightarrow$  I/1.3.6) and standard normally ( $\rightarrow$  II/3.2.3) distributed random variables ( $\rightarrow$  I/1.2.2):

$$z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad i = 1, \dots, n . \quad (3)$$

Then, these variables together form a multivariate normally ( $\rightarrow$  II/4.1.17) distributed random vector ( $\rightarrow$  I/1.2.3):

$$z \sim \mathcal{N}(0_n, I_n) . \quad (4)$$

Because the covariance is zero for independent random variables ( $\rightarrow$  I/1.15.6), we have

$$\text{Cov}(z_i, z_j) = 0 \quad \text{for all } i \neq j . \quad (5)$$

Moreover, as the variance of all entries of the vector is one ( $\rightarrow$  II/3.2.19), we have

$$\text{Var}(z_i) = 1 \quad \text{for all } i = 1, \dots, n . \quad (6)$$

Taking (5) and (6) together, the covariance matrix ( $\rightarrow$  I/1.16.1) of  $z$  is

$$\text{Cov}(z) = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = I_n . \quad (7)$$

2) Next, consider an  $n \times n$  matrix  $A$  solving the equation  $AA^T = \Sigma$ . Such a matrix exists, because  $\Sigma$  is defined to be positive definite ( $\rightarrow$  II/4.1.1). Then,  $x$  can be represented as a linear transformation ( $\rightarrow$  II/4.1.14) of  $z$ :

$$x = Az + \mu \sim \mathcal{N}(A0_n + \mu, AI_nA^T) = \mathcal{N}(\mu, \Sigma) . \quad (8)$$

Thus, the covariance ( $\rightarrow$  I/1.15.1) of  $x$  can be written as:

$$\text{Cov}(x) = \text{Cov}(Az + \mu) . \quad (9)$$

With the invariance of the covariance matrix under addition ( $\rightarrow$  I/1.16.6)

$$\text{Cov}(x + a) = \text{Cov}(x) \quad (10)$$

and the scaling of the covariance matrix upon multiplication ( $\rightarrow$  I/1.16.7)

$$\text{Cov}(Ax) = A\text{Cov}(x)A^T , \quad (11)$$



this becomes:

$$\begin{aligned}
 \text{Cov}(x) &= \text{Cov}(Az + \mu) \\
 &\stackrel{(10)}{=} \text{Cov}(Az) \\
 &\stackrel{(11)}{=} A \text{Cov}(z) A^T \\
 &\stackrel{(7)}{=} A I_n A^T \\
 &= A A^T \\
 &= \Sigma .
 \end{aligned} \tag{12}$$

■

#### Sources:

- Rosenfeld, Meni (2016): “Deriving the Covariance of Multivariate Gaussian”; in: *StackExchange Mathematics*, retrieved on 2022-09-15; URL: <https://math.stackexchange.com/questions/1905977/deriving-the-covariance-of-multivariate-gaussian>.

#### 4.1.9 Expectation of a quadratic form

**Theorem:** Let  $X$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) . \tag{1}$$

Then, the expectation ( $\rightarrow$  I/1.12.1) of the quadratic form  $X^T A X$  is

$$\mathbb{E} [X^T A X] = \mu^T A \mu + \text{tr}(A \Sigma) \tag{2}$$

where  $A$  is an  $n \times n$  matrix.

**Proof:** For any random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^n$ , this expectation is a function of mean and covariance ( $\rightarrow$  I/1.12.17):

$$\mathbb{E} [X^T A X] = \mathbb{E}(X)^T A X + \text{tr}(A \text{Cov}(X)) . \tag{3}$$

For a multivariate normal random vector ( $\rightarrow$  II/4.1.1)  $X \sim \mathcal{N}(\mu, \Sigma)$ , mean ( $\rightarrow$  II/4.1.6) and covariance ( $\rightarrow$  II/4.1.8) are

$$\mathbb{E}(X) = \mu \quad \text{and} \quad \text{Cov}(X) = \Sigma . \tag{4}$$

Combing (3) with (4), we have:

$$\mathbb{E} [X^T A X] = \mu^T A \mu + \text{tr}(A \Sigma) \tag{5}$$

■

### 4.1.10 Differential entropy

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $x$  in nats is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n . \quad (2)$$

**Proof:** The differential entropy ( $\rightarrow$  I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \quad (3)$$

To measure  $h(X)$  in nats, we set  $b = e$ , such that ( $\rightarrow$  I/1.12.1)

$$h(X) = -\mathbb{E} [\ln p(x)] . \quad (4)$$

With the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), the differential entropy of  $x$  is:

$$\begin{aligned} h(x) &= -\mathbb{E} \left[ \ln \left( \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) \right] \\ &= -\mathbb{E} \left[ -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \\ &= \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mathbb{E} [(x - \mu)^T \Sigma^{-1} (x - \mu)] . \end{aligned} \quad (5)$$

The last term can be evaluated as

$$\begin{aligned} \mathbb{E} [(x - \mu)^T \Sigma^{-1} (x - \mu)] &= \mathbb{E} [\text{tr} ((x - \mu)^T \Sigma^{-1} (x - \mu))] \\ &= \mathbb{E} [\text{tr} (\Sigma^{-1} (x - \mu) (x - \mu)^T)] \\ &= \text{tr} (\Sigma^{-1} \mathbb{E} [(x - \mu) (x - \mu)^T]) \\ &= \text{tr} (\Sigma^{-1} \Sigma) \\ &= \text{tr} (I_n) \\ &= n , \end{aligned} \quad (6)$$

such that the differential entropy is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n . \quad (7)$$

■

### Sources:

- Kiuahnm (2018): “Entropy of the multivariate Gaussian”; in: *StackExchange Mathematics*, retrieved on 2020-05-14; URL: <https://math.stackexchange.com/questions/2029707/entropy-of-the-multivariate-ga>

#### 4.1.11 Mutual information

**Theorem:** Let  $X \in \mathbb{R}^n$  and  $Y \in \mathbb{R}^m$  be random vectors ( $\rightarrow$  I/1.2.3) that are jointly multivariate normal ( $\rightarrow$  II/4.1.1):

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{bmatrix} \right) . \quad (1)$$

Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is

$$I(X, Y) = \frac{1}{2} \ln \left[ \frac{|\Sigma_1| |\Sigma_2|}{|\Sigma|} \right] \quad (2)$$

where  $\Sigma \in \mathbb{R}^{p \times p}$  is the covariance matrix ( $\rightarrow$  I/1.16.1) of the random vector ( $\rightarrow$  I/1.2.3)  $\begin{bmatrix} X \\ Y \end{bmatrix} \in \mathbb{R}^p$ ,

with  $p = n + m$ .

**Proof:** Mutual information can be written in terms of marginal and joint differential entropy ( $\rightarrow$  I/2.4.3):

$$I(X, Y) = h(X) + h(Y) - h(X, Y) . \quad (3)$$

The marginal distributions of the multivariate normal distribution are also multivariate normal ( $\rightarrow$  II/4.1.15)

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \Rightarrow X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) , \quad (4)$$

such that the marginals ( $\rightarrow$  I/1.5.3) of  $X$  and  $Y$  are:

$$X \sim \mathcal{N}(\mu_1, \Sigma_1) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_2, \Sigma_2) . \quad (5)$$

The differential entropy of the multivariate normal distribution ( $\rightarrow$  II/4.1.10) is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n \quad (6)$$

where  $|\Sigma|$  is the determinant of  $\Sigma$ . Combining (3) with (6), we get:

$$\begin{aligned}
I(X, Y) &\stackrel{(3)}{=} h(X) + h(Y) - h(X, Y) \\
&\stackrel{(5)}{=} h[\mathcal{N}(\mu_1, \Sigma_1)] + h[\mathcal{N}(\mu_2, \Sigma_2)] - h[\mathcal{N}(\mu, \Sigma)] \\
&\stackrel{(6)}{=} \left[ \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma_1| + \frac{1}{2} n \right] + \left[ \frac{m}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma_2| + \frac{1}{2} m \right] - \left[ \frac{p}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} p \right] \\
&= \left( \frac{n+m-p}{2} \ln(2\pi) + \frac{1}{2} (n+m-p) \right) + \left( \frac{1}{2} \ln |\Sigma_1| + \frac{1}{2} \ln |\Sigma_2| - \frac{1}{2} \ln |\Sigma| \right) \\
&= \frac{1}{2} (\ln |\Sigma_1| + \ln |\Sigma_2| - \ln |\Sigma|) \\
&= \frac{1}{2} \ln \left[ \frac{|\Sigma_1| |\Sigma_2|}{|\Sigma|} \right].
\end{aligned} \tag{7}$$

■

**Sources:**

- a06e (2019): “Mutual information between subsets of variables in the multivariate normal distribution”; in: *StackExchange Cross Validated*, retrieved on 2024-11-01; URL: <https://stats.stackexchange.com/a/438613/270304>.

**4.1.12 Kullback-Leibler divergence**

**Theorem:** Let  $x$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Assume two multivariate normal distributions ( $\rightarrow$  II/4.1.1)  $P$  and  $Q$  specifying the probability distribution of  $x$  as

$$\begin{aligned}
P : x &\sim \mathcal{N}(\mu_1, \Sigma_1) \\
Q : x &\sim \mathcal{N}(\mu_2, \Sigma_2).
\end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P || Q] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right]. \tag{2}$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P || Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \tag{3}$$

which, applied to the multivariate normal distributions ( $\rightarrow$  II/4.1.1) in (1), yields

$$\begin{aligned}
\text{KL}[P || Q] &= \int_{\mathbb{R}^n} \mathcal{N}(x; \mu_1, \Sigma_1) \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} dx \\
&= \left\langle \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} \right\rangle_{p(x)}.
\end{aligned} \tag{4}$$

Using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), this becomes:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \right]}{\frac{1}{\sqrt{(2\pi)^n |\Sigma_2|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right]} \right\rangle_{p(x)} \\
&= \left\langle \frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)} \quad (5) \\
&= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)}.
\end{aligned}$$

Now, using the fact that  $x = \text{tr}(x)$ , if  $a$  is scalar, and the trace property  $\text{tr}(ABC) = \text{tr}(BCA)$ , we have:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (x - \mu_2)(x - \mu_2)^T] \right\rangle_{p(x)} \\
&= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T)] \right\rangle_{p(x)}. \quad (6)
\end{aligned}$$

Because trace function and expected value are both linear operators ( $\rightarrow$  I/1.12.20), the expectation can be moved inside the trace:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \frac{1}{2} \left( \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)}] + \text{tr} [\Sigma_2^{-1} \langle xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T \rangle_{p(x)}] \right) \\
&= \frac{1}{2} \left( \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)}] + \text{tr} [\Sigma_2^{-1} (\langle xx^T \rangle_{p(x)} - \langle 2\mu_2 x^T \rangle_{p(x)} + \langle \mu_2 \mu_2^T \rangle_{p(x)})] \right) \quad (7)
\end{aligned}$$

Using the expectation of a linear form for the multivariate normal distribution ( $\rightarrow$  II/4.1.14)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle Ax \rangle = A\mu \quad (8)$$

and the expectation of a quadratic form for the multivariate normal distribution ( $\rightarrow$  I/1.12.17)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle x^T A x \rangle = \mu^T A \mu + \text{tr}(A \Sigma), \quad (9)$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \frac{1}{2} \left( \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\Sigma_1 + \mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left( \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [I_n] + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left( \ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\mu_1^T \Sigma_2^{-1} \mu_1 - 2\mu_1^T \Sigma_2^{-1} \mu_2 + \mu_2^T \Sigma_2^{-1} \mu_2] \right) \quad (10) \\
&= \frac{1}{2} \left[ \ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right].
\end{aligned}$$

Finally, rearranging the terms, we get:

$$\text{KL}[P || Q] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right]. \quad (11)$$

■

**Sources:**

- Duchi, John (2014): “Derivations for Linear Algebra and Optimization”; in: *University of California, Berkeley*; URL: [http://www.eecs.berkeley.edu/~jduchi/projects/general\\_notes.pdf](http://www.eecs.berkeley.edu/~jduchi/projects/general_notes.pdf).

**4.1.13 Maximum entropy distribution**

**Theorem:** The multivariate normal distribution ( $\rightarrow$  II/4.1.1) maximizes differential entropy ( $\rightarrow$  I/2.2.1) for a random vector ( $\rightarrow$  I/1.2.3) with fixed covariance matrix ( $\rightarrow$  I/1.16.1).

**Proof:** For a random vector ( $\rightarrow$  I/1.2.2)  $X$  with set of possible values  $\mathcal{X}$  and probability density function ( $\rightarrow$  I/1.7.1)  $p(x)$ , the differential entropy ( $\rightarrow$  I/2.2.1) is defined as:

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx. \quad (1)$$

Let  $g(x)$  be the probability density function ( $\rightarrow$  I/1.7.1) of a multivariate normal distribution ( $\rightarrow$  II/3.2.1) with mean ( $\rightarrow$  I/1.12.1)  $\mu$  and covariance ( $\rightarrow$  I/1.16.1)  $\Sigma$  and let  $f(x)$  be an arbitrary probability density function ( $\rightarrow$  I/1.7.1) with the same covariance ( $\rightarrow$  I/1.16.1). Since differential entropy ( $\rightarrow$  I/2.2.1) is translation-invariant ( $\rightarrow$  I/2.2.3), we can assume that  $f(x)$  has the same mean as  $g(x)$ .

Consider the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of distribution  $f(x)$  from distribution  $g(x)$  which is non-negative ( $\rightarrow$  I/2.5.2):

$$\begin{aligned} 0 \leq \text{KL}[f||g] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx \\ &= \int_{\mathcal{X}} f(x) \log f(x) dx - \int_{\mathcal{X}} f(x) \log g(x) dx \\ &\stackrel{(1)}{=} -h[f(x)] - \int_{\mathcal{X}} f(x) \log g(x) dx. \end{aligned} \quad (2)$$

By plugging the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) into the second term, we obtain:

$$\begin{aligned} \int_{\mathcal{X}} f(x) \log g(x) dx &= \int_{\mathcal{X}} f(x) \log \left( \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) dx \\ &= \int_{\mathcal{X}} f(x) \log \left( \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \right) dx + \int_{\mathcal{X}} f(x) \log \left( \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) dx \\ &= \left( -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| \right) \int_{\mathcal{X}} f(x) dx - \frac{1}{2} \int_{\mathcal{X}} f(x) [(x - \mu)^T \Sigma^{-1} (x - \mu)] dx. \end{aligned} \quad (3)$$

Because the entire integral over a probability density function is one ( $\rightarrow$  I/1.7.1) and with the definition of the expected value ( $\rightarrow$  I/1.12.1), this becomes:

$$\int_{\mathcal{X}} f(x) \log g(x) dx = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \langle (x - \mu)^T \Sigma^{-1} (x - \mu) \rangle_{f(x)} . \quad (4)$$

Using the expectation of a trace ( $\rightarrow$  I/1.12.20) and the definition of the covariance matrix ( $\rightarrow$  I/1.16.1), the second term can be developed as follows:

$$\begin{aligned} \langle (x - \mu)^T \Sigma^{-1} (x - \mu) \rangle_{f(x)} &= \langle \text{tr} [(x - \mu)^T \Sigma^{-1} (x - \mu)] \rangle_{f(x)} \\ &= \langle \text{tr} [\Sigma^{-1} (x - \mu)(x - \mu)^T] \rangle_{f(x)} \\ &= \text{tr} \left[ \Sigma^{-1} \langle (x - \mu)(x - \mu)^T \rangle_{f(x)} \right] \\ &= \text{tr} [\Sigma^{-1} \Sigma] \\ &= \text{tr} [I_n] \\ &= n . \end{aligned} \quad (5)$$

Thus, the second term in (2) is equal to

$$\int_{\mathcal{X}} f(x) \log g(x) dx = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{n}{2} . \quad (6)$$

This is actually the negative of the differential entropy of the multivariate normal distribution ( $\rightarrow$  II/3.2.23), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) dx = -h[\mathcal{N}(\mu, \Sigma)] = -h[g(x)] . \quad (7)$$

Combining (2) with (7), we can show that

$$\begin{aligned} 0 &\leq \text{KL}[f||g] \\ 0 &\leq -h[f(x)] - (-h[g(x)]) \\ h[g(x)] &\geq h[f(x)] \end{aligned} \quad (8)$$

which means that the differential entropy ( $\rightarrow$  I/2.2.1) of the multivariate normal distribution ( $\rightarrow$  II/4.1.1)  $\mathcal{N}(\mu, \Sigma)$  will be larger than or equal to any other distribution ( $\rightarrow$  I/1.5.1) with the same covariance matrix ( $\rightarrow$  I/1.16.1)  $\Sigma$ . ■

#### Sources:

- Wikipedia (2025): “Maximum entropy probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-05-23; URL: [https://en.wikipedia.org/wiki/Maximum\\_entropy\\_probability\\_distribution#Other\\_examples](https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution#Other_examples).

#### 4.1.14 Linear transformation

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, any linear transformation of  $x$  is also multivariate normally distributed:

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) . \quad (2)$$

**Proof:** The moment-generating function of a random vector ( $\rightarrow$  I/1.9.1)  $x$  is

$$M_x(t) = \mathbb{E}(\exp[t^T x]) \quad (3)$$

and therefore the moment-generating function of the random vector  $y$  is given by

$$\begin{aligned} M_y(t) &\stackrel{(2)}{=} \mathbb{E}(\exp[t^T(Ax + b)]) \\ &= \mathbb{E}(\exp[t^T Ax] \cdot \exp[t^T b]) \\ &= \exp[t^T b] \cdot \mathbb{E}(\exp[t^T Ax]) \\ &\stackrel{(3)}{=} \exp[t^T b] \cdot M_x(A^T t) . \end{aligned} \quad (4)$$

The moment-generating function of the multivariate normal distribution ( $\rightarrow$  II/4.1.5) is

$$M_x(t) = \exp\left[t^T \mu + \frac{1}{2} t^T \Sigma t\right] \quad (5)$$

and therefore the moment-generating function of the random vector  $y$  becomes

$$\begin{aligned} M_y(t) &\stackrel{(4)}{=} \exp[t^T b] \cdot M_x(A^T t) \\ &\stackrel{(5)}{=} \exp[t^T b] \cdot \exp\left[t^T A\mu + \frac{1}{2} t^T A\Sigma A^T t\right] \\ &= \exp\left[t^T (A\mu + b) + \frac{1}{2} t^T A\Sigma A^T t\right] . \end{aligned} \quad (6)$$

Because moment-generating function and probability density function of a random variable are equivalent, this demonstrates that  $y$  is following a multivariate normal distribution with mean  $A\mu + b$  and covariance  $A\Sigma A^T$ . ■

#### Sources:

- Taboga, Marco (2010): “Linear combinations of normal random variables”; in: *Lectures on probability and statistics*, retrieved on 2019-08-27; URL: <https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations>.

#### 4.1.15 Marginal distributions

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of any subset vector  $x_s$  is also a multivariate normal distribution

$$x_s \sim \mathcal{N}(\mu_s, \Sigma_s) \quad (2)$$



where  $\mu_s$  drops the irrelevant variables (the ones not in the subset, i.e. marginalized out) from the mean vector  $\mu$  and  $\Sigma_s$  drops the corresponding rows and columns from the covariance matrix  $\Sigma$ .

**Proof:** Define an  $m \times n$  subset matrix  $S$  such that  $s_{ij} = 1$ , if the  $j$ -th element in  $x_s$  corresponds to the  $i$ -th element in  $x$ , and  $s_{ij} = 0$  otherwise. Then,

$$x_s = Sx \quad (3)$$

and we can apply the linear transformation theorem ( $\rightarrow$  II/4.1.14) to give

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^T). \quad (4)$$

Finally, we see that  $S\mu = \mu_s$  and  $S\Sigma S^T = \Sigma_s$ . ■

#### 4.1.16 Conditional distributions

**Theorem:** Let  $x$  follow a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma). \quad (1)$$

Then, the conditional distribution ( $\rightarrow$  I/1.5.4) of any subset vector  $x_1$ , given the complement vector  $x_2$ , is also a multivariate normal distribution

$$x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \quad (2)$$

where the conditional mean ( $\rightarrow$  I/1.12.1) and covariance ( $\rightarrow$  I/1.15.1) are

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \quad (3)$$

with block-wise mean and covariance defined as

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \end{aligned} \quad (4)$$

**Proof:** Without loss of generality, we assume that, in parallel to (4),

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5)$$

where  $x_1$  is an  $n_1 \times 1$  vector,  $x_2$  is an  $n_2 \times 1$  vector and  $x$  is an  $n_1 + n_2 = n \times 1$  vector.

By construction, the joint distribution ( $\rightarrow$  I/1.5.2) of  $x_1$  and  $x_2$  is:

$$x_1, x_2 \sim \mathcal{N}(\mu, \Sigma) . \quad (6)$$

Moreover, the marginal distribution ( $\rightarrow$  I/1.5.3) of  $x_2$  follows from ( $\rightarrow$  II/4.1.15) (1) and (4) as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) . \quad (7)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), it holds that

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \quad (8)$$

Applying (6) and (7) to (8), we have:

$$p(x_1|x_2) = \frac{\mathcal{N}(x; \mu, \Sigma)}{\mathcal{N}(x_2; \mu_2, \Sigma_{22})} . \quad (9)$$

Using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), this becomes:

$$\begin{aligned} p(x_1|x_2) &= \frac{1/\sqrt{(2\pi)^n|\Sigma|} \cdot \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]}{1/\sqrt{(2\pi)^{n_2}|\Sigma_{22}|} \cdot \exp\left[-\frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right]} \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + \frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right] . \end{aligned} \quad (10)$$

Writing the inverse of  $\Sigma$  as

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \quad (11)$$

and applying (4) to (10), we get:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\quad \exp\left[-\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^T \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \right. \\ &\quad \left. + \frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1}(x_2 - \mu_2)\right] . \end{aligned} \quad (12)$$

Multiplying out within the exponent of (12), we have

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\quad \exp\left[-\frac{1}{2}\left((x_1 - \mu_1)^T \Sigma^{11}(x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma^{12}(x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma^{22}(x_2 - \mu_2)\right) \right. \\ &\quad \left. + \frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1}(x_2 - \mu_2)\right] \end{aligned} \quad (13)$$

where we have used the fact that  $\Sigma^{21T} = \Sigma^{12}$ , because  $\Sigma^{-1}$  is a symmetric matrix.

The inverse of a block matrix is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}, \quad (14)$$

thus the inverse of  $\Sigma$  in (11) is

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}. \quad (15)$$

Plugging this into (13), we have:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp \left[ -\frac{1}{2} \left( (x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\ &\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\ &\quad (x_2 - \mu_2)^T [\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}] (x_2 - \mu_2)) \\ &\quad \left. + \frac{1}{2} ((x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right]. \end{aligned} \quad (16)$$

Eliminating some terms, we have:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp \left[ -\frac{1}{2} \left( (x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\ &\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\ &\quad \left. (x_2 - \mu_2)^T \Sigma_{22}^{-1} \Sigma_{21} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) \right). \end{aligned} \quad (17)$$

Rearranging the terms, we have

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[ -\frac{1}{2} \cdot \right. \\ &\quad \left. [(x_1 - \mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} [(x_1 - \mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)] \right] \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[ -\frac{1}{2} \cdot \right. \\ &\quad \left. [x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2))]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} [x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2))] \right] \end{aligned} \quad (18)$$

where we have used the fact that  $\Sigma_{21} = \Sigma_{12}^T$ , because  $\Sigma$  is a covariance matrix ( $\rightarrow$  I/1.16.1).

The determinant of a block matrix is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C|, \quad (19)$$

such that we have for  $\Sigma$  that

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|. \quad (20)$$

With this and  $n - n_2 = n_1$ , we finally arrive at

$$p(x_1|x_2) = \frac{1}{\sqrt{(2\pi)^{n_1} |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|}} \cdot \exp \left[ -\frac{1}{2} \cdot \left[ x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) \right]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[ x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) \right] \right] \quad (21)$$

which is the probability density function of a multivariate normal distribution ( $\rightarrow$  II/4.1.4)

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \quad (22)$$

with the mean ( $\rightarrow$  I/1.12.1)  $\mu_{1|2}$  and covariance ( $\rightarrow$  I/1.15.1)  $\Sigma_{1|2}$  given by (3).

■

#### Sources:

- Wang, Ruye (2006): “Marginal and conditional distributions of multivariate normal distribution”; in: *Computer Image Processing and Analysis*; URL: <http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html>.
- Wikipedia (2020): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Conditional\\_distributions](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions).

#### 4.1.17 Conditions for independence

**Theorem:** Let  $x$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) following a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma). \quad (1)$$

Then, the components of  $x$  are statistically independent ( $\rightarrow$  I/1.3.6), if and only if the covariance matrix ( $\rightarrow$  I/1.16.1) is a diagonal matrix:

$$p(x) = p(x_1) \cdot \dots \cdot p(x_n) \quad \Leftrightarrow \quad \Sigma = \text{diag}([\sigma_1^2, \dots, \sigma_n^2]). \quad (2)$$

**Proof:** The marginal distribution of one entry from a multivariate normal random vector is a univariate normal distribution ( $\rightarrow$  II/4.1.15) where mean ( $\rightarrow$  I/1.12.1) and variance ( $\rightarrow$  I/1.13.1) are equal to the corresponding entries of the mean vector and covariance matrix:

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad x_i \sim \mathcal{N}(\mu_i, \sigma_{ii}^2) . \quad (3)$$

The probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) is

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \quad (4)$$

and the probability density function of the univariate normal distribution ( $\rightarrow$  II/3.2.10) is

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2 \right] . \quad (5)$$

1) Let

$$p(x) = p(x_1) \cdot \dots \cdot p(x_n) . \quad (6)$$

Then, because the covariance is zero under independence ( $\rightarrow$  I/1.15.6), the entries of the covariance matrix are the pair-wise covariances ( $\rightarrow$  I/1.16.1) and the covariance matrix of the multivariate normal distribution ( $\rightarrow$  II/4.1.8) is equal to  $\Sigma$ , we have

$$\Sigma = \text{diag} ([\text{Cov}(x_1, x_1), \dots, \text{Cov}(x_n, x_n)]) . \quad (7)$$

Moreover, because the diagonal entries of the covariance matrix are the element-wise variances ( $\rightarrow$  I/1.15.5) and the variance of the univariate normal distribution ( $\rightarrow$  II/3.2.19) is equal to  $\sigma^2$  and equating  $\sigma^2_{\_i} = \sigma^2_{\_ii}$  from (3), we get

$$\Sigma = \text{diag} ([\sigma_1^2, \dots, \sigma_n^2]) . \quad (8)$$

2) Let

$$\Sigma = \text{diag} ([\sigma_1^2, \dots, \sigma_n^2]) . \quad (9)$$

Then, we have

$$\begin{aligned} p(x) &\stackrel{(4)}{=} \frac{1}{\sqrt{(2\pi)^n |\text{diag} ([\sigma_1^2, \dots, \sigma_n^2])|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \text{diag} ([\sigma_1^2, \dots, \sigma_n^2])^{-1} (x - \mu) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \text{diag} ([1/\sigma_1^2, \dots, 1/\sigma_n^2]) (x - \mu) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right] \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2 \right] \end{aligned} \quad (10)$$

which implies that

$$p(x) = p(x_1) \cdot \dots \cdot p(x_n) . \quad (11)$$

■

#### 4.1.18 Independence of products

**Theorem:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) following a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (1)$$

and consider two matrices  $A \in \mathbb{R}^{k \times n}$  and  $B \in \mathbb{R}^{l \times n}$ . Then,  $AX$  and  $BX$  are independent ( $\rightarrow$  I/1.3.6), if and only if the cross-matrix product, weighted with the covariance matrix ( $\rightarrow$  II/4.1.8) is equal to the zero matrix:

$$AX \text{ and } BX \text{ ind.} \Leftrightarrow A\Sigma B^T = 0_{kl} . \quad (2)$$

**Proof:** Define a new random vector ( $\rightarrow$  I/1.2.3)  $C$  as

$$C = \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{(k+l) \times n} . \quad (3)$$

Then, due to the linear transformation theorem ( $\rightarrow$  II/4.1.14), we have

$$CX = \begin{bmatrix} AX \\ BX \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} A\mu \\ B\mu \end{bmatrix}, C\Sigma C^T \right) \quad (4)$$

with the combined covariance matrix ( $\rightarrow$  I/1.16.1)

$$C\Sigma C^T = \begin{bmatrix} A\Sigma A^T & A\Sigma B^T \\ B\Sigma A^T & B\Sigma B^T \end{bmatrix} . \quad (5)$$

We know that the necessary and sufficient condition for two components of a multivariate normal random vector to be independent is that their entries in the covariance matrix are zero ( $\rightarrow$  II/4.1.17). Thus,  $AX$  and  $BX$  are independent ( $\rightarrow$  I/1.3.6), if and only if

$$A\Sigma B^T = (B\Sigma A^T)^T = 0_{kl} \quad (6)$$

where  $0_{kl}$  is the  $k \times l$  zero matrix. This proves the result in (2).

■

#### Sources:

- jld (2018): “Understanding t-test for linear regression”; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: <https://stats.stackexchange.com/a/344008>.

### 4.1.19 Drawing samples

**Theorem:** Let  $X \in \mathbb{R}^n$  be a random vector ( $\rightarrow$  I/1.2.3) with all entries independently following a standard normal distribution ( $\rightarrow$  II/3.2.3). Moreover, let  $A$  be an  $n \times n$  matrix, such that  $AA^T = \Sigma$ . Then,  $Y = \mu + AX$  follows a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with mean ( $\rightarrow$  I/1.12.16)  $\mu$  and covariance ( $\rightarrow$  I/1.16.1)  $\Sigma$ :

$$Y = \mu + AX \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

**Proof:** If all entries of  $X$  are independent and standard normally distributed ( $\rightarrow$  II/3.2.3)

$$X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for all } i = 1, \dots, n , \quad (2)$$

this implies a multivariate normal distribution with diagonal covariance matrix ( $\rightarrow$  II/4.1.17):

$$X \sim \mathcal{N}(0_n, I_n) . \quad (3)$$

Thus, with the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14), it follows that

$$\begin{aligned} Y = \mu + AX &\sim \mathcal{N}(A0_n + \mu, AI_nA^T) \\ &\sim \mathcal{N}(\mu, AA^T) \\ &\sim \mathcal{N}(\mu, \Sigma) . \end{aligned} \quad (4)$$

Thus, given  $X$  defined by (2),  $Y$  defined by (1) is a sample ( $\rightarrow$  I/1.1.6) from  $\mathcal{N}(\mu, \Sigma)$ . ■

#### Sources:

- Wikipedia (2025): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-05-23; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Drawing\\_values\\_from\\_the\\_distribution](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Drawing_values_from_the_distribution).

## 4.2 Bivariate normal distribution

### 4.2.1 Definition

**Definition:** Let  $X$  be an  $2 \times 1$  random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to have a bivariate normal distribution, if  $X$  follows a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (1)$$

with means ( $\rightarrow$  I/1.12.1)  $\mu_1$  and  $\mu_2$ , variances ( $\rightarrow$  I/1.13.1)  $\sigma_1^2$  and  $\sigma_2^2$  and covariance ( $\rightarrow$  I/1.15.1)  $\sigma_{12}$ :

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} . \quad (2)$$

#### Sources:

- Wikipedia (2023): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-09-22; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Bivariate\\_case](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case).

#### 4.2.2 Probability density function

**Theorem:** Let  $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$  follow a bivariate normal distribution ( $\rightarrow$  II/4.2.1):

$$X \sim \mathcal{N} \left( \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is:

$$f_X(x) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp \left[ -\frac{1}{2} \frac{\sigma_2^2(x_1 - \mu_1)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) + \sigma_1^2(x_2 - \mu_2)^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \right]. \quad (2)$$

**Proof:** The probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) for an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3)  $x$  is:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]. \quad (3)$$

Plugging in  $n = 2$ ,  $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$  and  $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ , we obtain:

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^2 \left| \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right|}} \cdot \exp \left[ -\frac{1}{2} \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^T \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \right] \\ &= \frac{1}{2\pi \left| \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right|^{\frac{1}{2}}} \cdot \exp \left[ -\frac{1}{2} \begin{bmatrix} (x_1 - \mu_1) & (x_2 - \mu_2) \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} (x_1 - \mu_1) \\ (x_2 - \mu_2) \end{bmatrix} \right]. \end{aligned} \quad (4)$$

Using the determinant of a  $2 \times 2$  matrix

$$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc \quad (5)$$



and the inverse of of a  $2 \times 2$  matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \quad (6)$$

the probability density function ( $\rightarrow$  I/1.7.1) becomes:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp \left[ -\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \begin{bmatrix} (x_1 - \mu_1) & (x_2 - \mu_2) \end{bmatrix} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix} \begin{bmatrix} (x_1 - \mu_1) \\ (x_2 - \mu_2) \end{bmatrix} \right] \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp \left[ -\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \begin{bmatrix} \sigma_2^2(x_1 - \mu_1) - \sigma_{12}(x_2 - \mu_2) & \sigma_1^2(x_2 - \mu_2) - \sigma_{12}(x_1 - \mu_1) \end{bmatrix} \begin{bmatrix} (x_1 - \mu_1) \\ (x_2 - \mu_2) \end{bmatrix} \right] \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp \left[ -\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} (\sigma_2^2(x_1 - \mu_1)^2 - \sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) + \sigma_1^2(x_2 - \mu_2)^2 - \sigma_{12}(x_2 - \mu_2)(x_1 - \mu_1)) \right] \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp \left[ -\frac{1}{2} \frac{\sigma_2^2(x_1 - \mu_1)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) + \sigma_1^2(x_2 - \mu_2)^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \right]. \end{aligned} \quad (7)$$

■

#### 4.2.3 Probability density function in terms of correlation coefficient

**Theorem:** Let  $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$  follow a bivariate normal distribution ( $\rightarrow$  II/4.2.1):

$$X \sim \mathcal{N} \left( \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \right] \quad (2)$$

where  $\rho$  is the correlation ( $\rightarrow$  I/1.17.1) between  $X_1$  and  $X_2$ .

**Proof:** Since  $X$  follows a special case of the multivariate normal distribution, its covariance matrix is ( $\rightarrow$  II/4.1.8)

$$\text{Cov}(X) = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \quad (3)$$

and the covariance matrix can be decomposed into correlation matrix and standard deviations ( $\rightarrow$  I/1.16.8):

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}.
\end{aligned} \tag{4}$$

The determinant of this matrix is

$$\begin{aligned}
|\Sigma| &= \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \\
&= \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \cdot \left| \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right| \cdot \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \\
&= (\sigma_1 \sigma_2)(1 - \rho^2)(\sigma_1 \sigma_2) \\
&= \sigma_1^2 \sigma_2^2 (1 - \rho^2)
\end{aligned} \tag{5}$$

and the inverse of this matrix is

$$\begin{aligned}
\Sigma^{-1} &= \left( \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}^{-1} \\
&= \frac{1}{1 - \rho^2} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix}.
\end{aligned} \tag{6}$$

The probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) for an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3)  $x$  is:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]. \tag{7}$$

Plugging in  $n = 2$ ,  $\mu$  from (1) and  $\Sigma$  from (5) and (6), the probability density function becomes:

$$\begin{aligned}
f_X(x) &= \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2 (1-\rho^2)}} \cdot \exp \left[ -\frac{1}{2} \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^T \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \right] \\
&= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp \left[ -\frac{1}{2(1-\rho^2)} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} & \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} \\ \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \right] \\
&= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp \left[ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x_1-\mu_1}{\sigma_1} - \rho \frac{x_2-\mu_2}{\sigma_2} \right) \left( \frac{x_2-\mu_2}{\sigma_2} - \rho \frac{x_1-\mu_1}{\sigma_1} \right) \right] \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} \\ \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \right] \\
&= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \left( \frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1 \sigma_2} + \left( \frac{x_2-\mu_2}{\sigma_2} \right)^2 \right) \right]. \tag{8}
\end{aligned}$$

■

**Sources:**

- Wikipedia (2023): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-09-29; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Bivariate\\_case](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case).

**4.2.4 Construction from standard normal distributions**

**Theorem:** Let  $Y_1$  and  $Y_2$  be independent ( $\rightarrow$  I/1.3.6) random variables ( $\rightarrow$  I/1.2.2) following a standard normal distribution ( $\rightarrow$  II/3.2.3):

$$Y_1 \sim \mathcal{N}(0, 1) \quad \text{and} \quad Y_2 \sim \mathcal{N}(0, 1) \quad \text{ind.} \tag{1}$$

and define the two random variables ( $\rightarrow$  I/1.2.2)  $X_1$  and  $X_2$

$$\begin{aligned}
X_1 &= \sigma_1 Y_1 + \mu_1 \\
X_2 &= \sigma_2 \left( \rho Y_1 + \sqrt{1-\rho^2} Y_2 \right) + \mu_2
\end{aligned} \tag{2}$$

where  $\mu_1, \mu_2 \in \mathbb{R}$ ,  $\sigma_1, \sigma_2 \in \mathbb{R}_{>0}$  and  $\rho \in [-1, +1]$ . Then, the random vector ( $\rightarrow$  I/1.2.3) with entries  $X_1$  and  $X_2$  follows a bivariate normal distribution ( $\rightarrow$  II/4.2.1)

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \right) \tag{3}$$

where mean and covariance ( $\rightarrow$  II/4.2.1) are functions of the parameters  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$ ,  $\sigma_2$  and  $\rho$ .

**Proof:** The probability density function of the standard normal distribution ( $\rightarrow$  II/3.2.10) is

$$Y \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad f_Y(y) = \frac{1}{\sqrt{2\pi}} \cdot \exp \left[ -\frac{1}{2} y^2 \right]. \tag{4}$$

The random variables  $X_1$  and  $X_2$  are functions of  $Y_1$  and  $Y_2$

$$\begin{aligned} X_1 &= \sigma_1 Y_1 + \mu_1 \\ X_2 &= \sigma_2 \left( \rho Y_1 + \sqrt{1 - \rho^2} Y_2 \right) + \mu_2, \end{aligned} \quad (5)$$

such that the inverse functions  $Y_1$  and  $Y_2$  in terms of  $X_1$  and  $X_2$  are

$$\begin{aligned} Y_1 &= (X_1 - \mu_1) / \sigma_1 \\ Y_2 &= [(X_2 - \mu_2) / \sigma_2 - \rho Y_1] / \sqrt{1 - \rho^2} \\ &= [(X_2 - \mu_2) / \sigma_2 - \rho(X_1 - \mu_1) / \sigma_1] / \sqrt{1 - \rho^2}. \end{aligned} \quad (6)$$

This implies the following Jacobian matrix

$$J = \begin{bmatrix} \frac{dY_1}{dX_1} & \frac{dY_1}{dX_2} \\ \frac{dY_2}{dX_1} & \frac{dY_2}{dX_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 \\ \frac{\rho}{\sigma_1 \sqrt{1 - \rho^2}} & \frac{1}{\sigma_2 \sqrt{1 - \rho^2}} \end{bmatrix} \quad (7)$$

and determinant of the Jacobian matrix

$$\begin{aligned} |J| &= \frac{1}{\sigma_1} \cdot \frac{1}{\sigma_2 \sqrt{1 - \rho^2}} - 0 \cdot \frac{\rho}{\sigma_1 \sqrt{1 - \rho^2}} \\ &= \frac{1}{\sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \\ &= \frac{1}{\sqrt{\sigma_1^2 \sigma_2^2 (1 - \rho^2)}}. \end{aligned} \quad (8)$$

Because  $Y_1$  and  $Y_2$  are independent ( $\rightarrow$  I/1.3.6), the joint density ( $\rightarrow$  I/1.5.2) of  $Y_1$  and  $Y_2$  is equal to the product ( $\rightarrow$  I/1.3.9) of the marginal densities ( $\rightarrow$  I/1.5.3):

$$f_{Y_1, Y_2}(y_1, y_2) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2). \quad (9)$$

Substituting (4) into (9), we get:

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} y_1^2 \right] \cdot \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} y_2^2 \right] \\ &= \frac{1}{\sqrt{(2\pi)^2}} \cdot \exp \left[ -\frac{1}{2} (y_1^2 + y_2^2) \right]. \end{aligned} \quad (10)$$

With the probability density function of an invertible function ( $\rightarrow$  I/1.7.5), the joint density ( $\rightarrow$  I/1.5.2) of  $X_1$  and  $X_2$  can be derived as:

$$f_{X_1, X_2}(x_1, x_2) = f_{Y_1, Y_2}(y_1, y_2) \cdot |J|. \quad (11)$$

Substituting (10), (6) and (8) into (11), we get:

$$\begin{aligned}
f_{X_1, X_2}(x_1, x_2) &= \frac{1}{\sqrt{(2\pi)^2}} \cdot \exp \left[ -\frac{1}{2} \left( \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left( \left[ \frac{x_2 - \mu_2}{\sigma_2} - \rho \frac{x_1 - \mu_1}{\sigma_1} \right] / \sqrt{1 - \rho^2} \right)^2 \right) \right] \cdot \frac{1}{\sqrt{\sigma_1^2 \sigma_2^2 (1 - \rho^2)}} \\
&= \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2 (1 - \rho^2)}} \cdot \exp \left[ -\frac{1}{2} \left( \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \frac{1}{1 - \rho^2} \left[ \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right) \right] \\
&= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \cdot \exp \left[ -\frac{1}{2(1 - \rho^2)} \left( \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \right]
\end{aligned} \tag{12}$$

This is the probability density function of the bivariate normal distribution ( $\rightarrow$  II/4.2.3) with parameters

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{13}$$

Thus, it follows that

$$X \sim \mathcal{N}(\mu, \Sigma). \tag{14}$$

■

#### Sources:

- Ostwald, Dirk (2023): “Normalverteilungen”; in: *Allgemeines Lineares Modell*, Einheit (4), Folie 15; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/4\\_Normalverteilungen.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/4_Normalverteilungen.pdf).
- Wikipedia (2025): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-05-23; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Bivariate\\_case](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case).

#### 4.2.5 Mutual information

**Theorem:** Let  $X$  and  $Y$  follow a bivariate normal distribution ( $\rightarrow$  II/4.2.1):

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right). \tag{1}$$

Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$  is

$$I(X, Y) = -\frac{1}{2} \ln(1 - \rho^2) \tag{2}$$

where  $\rho$  is the correlation ( $\rightarrow$  I/1.17.1) of  $X$  and  $Y$ .

**Proof:** Mutual information can be written in terms of marginal and joint differential entropy ( $\rightarrow$  I/2.4.3):

$$I(X, Y) = h(X) + h(Y) - h(X, Y). \tag{3}$$

The marginal distributions of the multivariate normal distribution are also multivariate normal ( $\rightarrow$  II/4.1.15)

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \Rightarrow X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) , \quad (4)$$

such that the marginals ( $\rightarrow$  I/1.5.3) of the bivariate normal distribution ( $\rightarrow$  II/4.2.1) are univariate normal distributions ( $\rightarrow$  II/3.2.1):

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right) \Rightarrow X \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_2, \sigma_2^2) . \quad (5)$$

The differential entropy of the univariate normal distribution ( $\rightarrow$  II/3.2.23) is

$$h(X) = \frac{1}{2} \ln(2\pi\sigma^2 e) \quad (6)$$

and the differential entropy of the multivariate normal distribution ( $\rightarrow$  II/4.1.10) is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n \quad (7)$$

where  $|\Sigma|$  is the determinant of the covariance matrix ( $\rightarrow$  I/1.16.1)  $\Sigma$ . A two-dimensional covariance matrix can be rewritten in terms of correlations ( $\rightarrow$  I/1.16.8) as follows:

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} . \end{aligned} \quad (8)$$

Combining (3) with (6) and (7), applying  $n = 2$ , we get:

$$\begin{aligned}
I(X, Y) &\stackrel{(3)}{=} h(X) + h(Y) - h(X, Y) \\
&\stackrel{(5)}{=} h[\mathcal{N}(\mu_1, \sigma_1^2)] + h[\mathcal{N}(\mu_2, \sigma_2^2)] - h[\mathcal{N}(\mu, \Sigma)] \\
&\stackrel{(8)}{=} \left[ \frac{1}{2} \ln(2\pi\sigma_1^2 e) \right] + \left[ \frac{1}{2} \ln(2\pi\sigma_2^2 e) \right] - \left[ \frac{2}{2} \ln(2\pi) + \frac{1}{2} \ln \left| \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right| + \frac{1}{2} \cdot 2 \right] \\
&= \left( \frac{2}{2} \ln(2\pi) + \frac{2}{2} \ln(e) - \ln(2\pi) - 1 \right) + \left( \frac{1}{2} \ln(\sigma_1^2) + \frac{1}{2} \ln(\sigma_2^2) - \frac{1}{2} \ln \left| \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right| \right) \\
&= \frac{1}{2} [\ln(\sigma_1^2) + \ln(\sigma_2^2) - \ln(\sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2)] \\
&= \frac{1}{2} \ln \left[ \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2} \right] \\
&= \frac{1}{2} \ln \left[ \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2(1 - \rho^2)} \right] \\
&= \frac{1}{2} \ln \left[ \frac{1}{1 - \rho^2} \right] \\
&= -\frac{1}{2} \ln(1 - \rho^2) .
\end{aligned} \tag{9}$$

#### Sources:

- Krafft, Peter (2013): “Correlation and Mutual Information”; in: *Princeton University Department of Computer Science: Laboratory for Intelligent Probabilistic Systems*, February 13, 2013; URL: <https://lips.cs.princeton.edu/correlation-and-mutual-information/>.

#### 4.2.6 Linear combination

**Theorem:** Let  $X$  and  $Y$  follow a bivariate normal distribution ( $\rightarrow$  II/4.2.1):

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right) . \tag{1}$$

Then, any linear combination of  $X$  and  $Y$  follows a univariate normal distribution ( $\rightarrow$  II/3.2.1):

$$Z = aX + bY \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\sigma_{12} + b^2\sigma_2^2) . \tag{2}$$

**Proof:** The linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14) states that

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad Y = AX + c \sim \mathcal{N}(A\mu + c, A\Sigma A^T) \tag{3}$$

where  $X \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  and  $c \in \mathbb{R}^n$ . In the present case, we have

$$X \in \mathbb{R}^2 \quad \text{and} \quad A = \begin{bmatrix} a & b \end{bmatrix} \in \mathbb{R}^{1 \times 2} \quad \text{and} \quad c = 0 \in \mathbb{R} , \tag{4}$$

such that

$$Z = A \begin{bmatrix} X \\ Y \end{bmatrix} + c = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + 0 = aX + bY . \quad (5)$$

Combining (3), (1) and (5), it follows that

$$\begin{aligned} Z &\sim \mathcal{N} \left( \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \right) \\ &\sim \mathcal{N} \left( a\mu_1 + b\mu_2, \begin{bmatrix} a\sigma_1^2 + b\sigma_{12} & a\sigma_{12} + b\sigma_2^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \right) \\ &\sim \mathcal{N} (a\mu_1 + b\mu_2, (a^2\sigma_1^2 + ab\sigma_{12}) + (ab\sigma_{12} + b^2\sigma_2^2)) \\ &\sim \mathcal{N} (a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\sigma_{12} + b^2\sigma_2^2) . \end{aligned} \quad (6)$$

■

#### Sources:

- Wikipedia (2024): “Misconceptions about the normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-10-25; URL: [https://en.wikipedia.org/wiki/Misconceptions\\_about\\_the\\_normal\\_distribution](https://en.wikipedia.org/wiki/Misconceptions_about_the_normal_distribution).

## 4.3 Multivariate t-distribution

### 4.3.1 Definition

**Definition:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to follow a multivariate  $t$ -distribution with mean  $\mu$ , scale matrix  $\Sigma$  and degrees of freedom  $\nu$

$$X \sim t(\mu, \Sigma, \nu) , \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$t(x; \mu, \Sigma, \nu) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu + n]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^{-(\nu+n)/2} \quad (2)$$

where  $\mu$  is an  $n$ -dimensional real vector,  $\Sigma$  is an  $n \times n$  positive-definite matrix and  $\nu > 0$ .

#### Sources:

- Koch KR (2007): “Multivariate t-Distribution”; in: *Introduction to Bayesian Statistics*, ch. 2.5.2, pp. 53-55; URL: <https://www.springer.com/gp/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.



### 4.3.2 Probability density function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a multivariate t-distribution ( $\rightarrow$  II/4.3.1):

$$X \sim t(\mu, \Sigma, \nu) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)} \left[ 1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^{-(\nu+n)/2} . \quad (2)$$

**Proof:** This follows directly from the definition of the multivariate t-distribution ( $\rightarrow$  II/4.3.1). ■

### 4.3.3 Relationship to F-distribution

**Theorem:** Let  $X$  be a  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) following a multivariate t-distribution ( $\rightarrow$  II/4.3.1) with mean  $\mu$ , scale matrix  $\Sigma$  and degrees of freedom  $\nu$ :

$$X \sim t(\mu, \Sigma, \nu) . \quad (1)$$

Then, the centered, weighted and standardized quadratic form of  $X$  follows an F-distribution ( $\rightarrow$  II/3.8.1) with degrees of freedom  $n$  and  $\nu$ :

$$(X - \mu)^T \Sigma^{-1} (X - \mu) / n \sim F(n, \nu) . \quad (2)$$

**Proof:** The linear transformation theorem for the multivariate t-distribution states

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad y = Ax + b \sim t(A\mu + b, A\Sigma A^T, \nu) \quad (3)$$

where  $x$  is an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) following a multivariate t-distribution ( $\rightarrow$  II/4.3.1),  $A$  is an  $m \times n$  matrix and  $b$  is an  $m \times 1$  vector. Define the following quantities

$$\begin{aligned} Y &= \Sigma^{-1/2} (X - \mu) = \Sigma^{-1/2} X - \Sigma^{-1/2} \mu \\ Z &= Y^T Y / n = (X - \mu)^T \Sigma^{-1} (X - \mu) / n \end{aligned} \quad (4)$$

where  $\Sigma^{-1/2}$  is a matrix square root of the inverse of  $\Sigma$ . Then, applying (3) to (4) with (1), one obtains the distribution of  $Y$  as

$$\begin{aligned} Y &\sim t(\Sigma^{-1/2} \mu - \Sigma^{-1/2} \mu, \Sigma^{-1/2} \Sigma \Sigma^{-1/2}, \nu) \\ &= t(0_n, \Sigma^{-1/2} \Sigma^{1/2} \Sigma^{1/2} \Sigma^{-1/2}, \nu) \\ &= t(0_n, I_n, \nu) , \end{aligned} \quad (5)$$

i.e. the marginal distributions ( $\rightarrow$  I/1.5.3) of the individual entries of  $Y$  are univariate t-distributions ( $\rightarrow$  II/3.3.1) with  $\nu$  degrees of freedom:

$$Y_i \sim t(\nu), \quad i = 1, \dots, n . \quad (6)$$

Note that, when  $X$  follows a t-distribution with  $n$  degrees of freedom, this is equivalent to ( $\rightarrow$  II/3.3.1) an expression of  $X$  in terms of a standard normal ( $\rightarrow$  II/3.2.3) random variable  $Z$  and a chi-squared ( $\rightarrow$  II/3.7.1) random variable  $V$ :

$$X \sim t(n) \quad \Leftrightarrow \quad X = \frac{Z}{\sqrt{V/n}} \quad \text{with independent} \quad Z \sim \mathcal{N}(0, 1) \quad \text{and} \quad V \sim \chi^2(n). \quad (7)$$

With that,  $Z$  from (4) can be rewritten as follows:

$$\begin{aligned} Z &\stackrel{(4)}{=} Y^T Y / n \\ &= \frac{1}{n} \sum_{i=1}^n Y_i^2 \\ &\stackrel{(7)}{=} \frac{1}{n} \sum_{i=1}^n \left( \frac{Z_i}{\sqrt{V/\nu}} \right)^2 \\ &= \frac{(\sum_{i=1}^n Z_i^2) / n}{V/\nu}. \end{aligned} \quad (8)$$

Because by definition, the sum of squared standard normal random variables follows a chi-squared distribution ( $\rightarrow$  II/3.7.1)

$$X_i \sim \mathcal{N}(0, 1), \quad i = 1, \dots, n \quad \Rightarrow \quad \sum_{i=1}^n X_i^2 \sim \chi^2(n), \quad (9)$$

the quantity  $Z$  becomes a ratio of the following form

$$Z = \frac{W/n}{V/\nu} \quad \text{with} \quad W \sim \chi^2(n) \quad \text{and} \quad V \sim \chi^2(\nu), \quad (10)$$

such that  $Z$ , by definition, follows an F-distribution ( $\rightarrow$  II/3.8.1):

$$Z = \frac{W/n}{V/\nu} \sim F(n, \nu). \quad (11)$$

■

#### Sources:

- Lin, Pi-Erh (1972): “Some Characterizations of the Multivariate t Distribution”; in: *Journal of Multivariate Analysis*, vol. 2, pp. 339-344, Lemma 2; URL: <https://core.ac.uk/download/pdf/81139018.pdf>; DOI: 10.1016/0047-259X(72)90021-8.
- Nadarajah, Saralees; Kotz, Samuel (2005): “Mathematical Properties of the Multivariate t Distribution”; in: *Acta Applicandae Mathematicae*, vol. 89, pp. 53-84, page 56; URL: <https://link.springer.com/content/pdf/10.1007/s10440-005-9003-4.pdf>; DOI: 10.1007/s10440-005-9003-4.

## 4.4 Normal-gamma distribution

### 4.4.1 Definition

**Definition:** Let  $X$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) and let  $Y$  be a positive random variable ( $\rightarrow$  I/1.2.2). Then,  $X$  and  $Y$  are said to follow a normal-gamma distribution

$$X, Y \sim \text{NG}(\mu, \Lambda, a, b) , \quad (1)$$

if the distribution of  $X$  conditional on  $Y$  is a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with mean vector  $\mu$  and covariance matrix  $(y\Lambda)^{-1}$  and  $Y$  follows a gamma distribution ( $\rightarrow$  II/3.4.1) with shape parameter  $a$  and rate parameter  $b$ :

$$\begin{aligned} X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\ Y &\sim \text{Gam}(a, b) . \end{aligned} \quad (2)$$

The  $n \times n$  matrix  $\Lambda$  is referred to as the precision matrix ( $\rightarrow$  I/1.16.12) of the normal-gamma distribution.

#### Sources:

- Koch KR (2007): “Normal-Gamma Distribution”; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: <https://www.springer.com/gp/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

#### 4.4.2 Special case of normal-Wishart distribution

**Theorem:** The normal-gamma distribution ( $\rightarrow$  II/4.4.1) is a special case of the normal-Wishart distribution ( $\rightarrow$  II/5.3.1) where the number of columns of the random matrices ( $\rightarrow$  I/1.2.4) is  $p = 1$ .

**Proof:** Let  $X$  be an  $n \times p$  real matrix and let  $Y$  be a  $p \times p$  positive-definite symmetric matrix, such that  $X$  and  $Y$  jointly follow a normal-Wishart distribution ( $\rightarrow$  II/5.3.1):

$$X, Y \sim \text{NW}(M, U, V, \nu) . \quad (1)$$

Then,  $X$  and  $Y$  are described by the probability density function ( $\rightarrow$  II/5.3.2)

$$\begin{aligned} p(X, Y) = & \frac{1}{\sqrt{(2\pi)^{np}|U|^p|V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \\ & \exp \left[ -\frac{1}{2} \text{tr} \left( Y [(X - M)^T U^{-1} (X - M) + V^{-1}] \right) \right] \end{aligned} \quad (2)$$

where  $|A|$  is a matrix determinant,  $A^{-1}$  is a matrix inverse and  $\Gamma_p(x)$  is the multivariate gamma function of order  $p$ . If  $p = 1$ , then  $\Gamma_p(x) = \Gamma(x)$  is the ordinary gamma function,  $x = X$  is a column vector and  $y = Y$  is a real number. Thus, the probability density function ( $\rightarrow$  I/1.7.1) of  $x$  and  $y$  can be developed as

$$\begin{aligned}
p(x, y) &= \frac{1}{\sqrt{(2\pi)^n |U| |V|^\nu}} \cdot \frac{\sqrt{2^{-\nu}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{(\nu+n-2)/2} \cdot \\
&\quad \exp \left[ -\frac{1}{2} \text{tr} \left( y \left[ (x - M)^T U^{-1} (x - M) + V^{-1} \right] \right) \right] \\
&= \sqrt{\frac{|U^{-1}|}{(2\pi)^n}} \cdot \frac{\sqrt{(2|V|)^{-\nu}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{\frac{\nu}{2} + \frac{n}{2} - 1} \cdot \\
&\quad \exp \left[ -\frac{1}{2} \left( y \left[ (x - M)^T U^{-1} (x - M) + 2(2V)^{-1} \right] \right) \right] \\
&= \sqrt{\frac{|U^{-1}|}{(2\pi)^n}} \cdot \frac{\left(\frac{1}{2|V|}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{\frac{\nu}{2} + \frac{n}{2} - 1} \cdot \\
&\quad \exp \left[ -\frac{y}{2} \left( (x - M)^T U^{-1} (x - M) + 2 \left( \frac{1}{2V} \right) \right) \right]
\end{aligned} \tag{3}$$

In the matrix-normal distribution ( $\rightarrow$  II/5.1.1), we have  $M \in \mathbb{R}^{n \times p}$ ,  $U \in \mathbb{R}^{n \times n}$ ,  $V \in \mathbb{R}^{p \times p}$  and  $\nu \in \mathbb{R}$ . Thus, with  $p = 1$ ,  $M$  becomes a column vector and  $V$  becomes a real number, such that  $V = |V| = 1/V^{-1}$ . Finally, substituting  $\mu = M$ ,  $\Lambda = U^{-1}$ ,  $a = \frac{\nu}{2}$  and  $b = \frac{1}{2V}$ , we get

$$p(x, y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a + \frac{n}{2} - 1} \exp \left[ -\frac{y}{2} \left( (x - \mu)^T \Lambda (x - \mu) + 2b \right) \right] \tag{4}$$

which is the probability density function of the normal-gamma distribution ( $\rightarrow$  II/4.4.3). ■

#### 4.4.3 Probability density function

**Theorem:** Let  $x$  and  $y$  follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \tag{1}$$

Then, the joint probability ( $\rightarrow$  I/1.3.2) density function ( $\rightarrow$  I/1.7.1) of  $x$  and  $y$  is

$$p(x, y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a + \frac{n}{2} - 1} \exp \left[ -\frac{y}{2} \left( (x - \mu)^T \Lambda (x - \mu) + 2b \right) \right] . \tag{2}$$

**Proof:** The normal-gamma distribution ( $\rightarrow$  II/4.4.1) is defined as  $X$  conditional on  $Y$  following a multivariate distribution ( $\rightarrow$  II/4.1.1) and  $Y$  following a gamma distribution ( $\rightarrow$  II/3.4.1):

$$\begin{aligned}
X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\
Y &\sim \text{Gam}(a, b) .
\end{aligned} \tag{3}$$

Thus, using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) and the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we have the following probabilities:

$$\begin{aligned}
p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\
&= \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[ -\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \\
p(y) &= \text{Gam}(y; a, b) \\
&= \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] .
\end{aligned} \tag{4}$$

The law of conditional probability ( $\rightarrow$  I/1.3.4) implies that

$$p(x, y) = p(x|y) p(y) , \tag{5}$$

such that the normal-gamma density function becomes:

$$p(x, y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[ -\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] . \tag{6}$$

Using the relation  $|yA| = y^n|A|$  for an  $n \times n$  matrix  $A$  and rearranging the terms, we have:

$$p(x, y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp \left[ -\frac{y}{2} ((x - \mu)^T \Lambda (x - \mu) + 2b) \right] . \tag{7}$$

■

#### Sources:

- Koch KR (2007): “Normal-Gamma Distribution”; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: <https://www.springer.com/gp/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

#### 4.4.4 Mean

**Theorem:** Let  $x \in \mathbb{R}^n$  and  $y > 0$  follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \tag{1}$$

Then, the expected value ( $\rightarrow$  I/1.12.1) of  $x$  and  $y$  is

$$\mathbb{E}[(x, y)] = \left( \mu, \frac{a}{b} \right) . \tag{2}$$

**Proof:** Consider the random vector ( $\rightarrow$  I/1.2.3)

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ y \end{bmatrix} . \tag{3}$$

According to the expected value of a random vector ( $\rightarrow$  I/1.12.16), its expected value is

$$\mathbb{E} \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} \mathbb{E}(x_1) \\ \vdots \\ \mathbb{E}(x_n) \\ \mathbb{E}(y) \end{bmatrix} = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix} . \quad (4)$$

When  $x$  and  $y$  are jointly normal-gamma distributed, then ( $\rightarrow$  II/4.4.1) by definition  $x$  follows a multivariate normal distribution ( $\rightarrow$  II/4.1.1) conditional on  $y$  and  $y$  follows a univariate gamma distribution ( $\rightarrow$  II/3.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) \quad \Leftrightarrow \quad x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \quad \wedge \quad y \sim \text{Gam}(a, b) . \quad (5)$$

Thus, with the expected value of the multivariate normal distribution ( $\rightarrow$  II/4.1.6) and the law of conditional probability ( $\rightarrow$  I/1.3.4),  $\mathbb{E}(x)$  becomes

$$\begin{aligned} \mathbb{E}(x) &= \iint x \cdot p(x, y) \, dx \, dy \\ &= \iint x \cdot p(x|y) \cdot p(y) \, dx \, dy \\ &= \int p(y) \int x \cdot p(x|y) \, dx \, dy \\ &= \int p(y) \langle x \rangle_{\mathcal{N}(\mu, (y\Lambda)^{-1})} \, dy \\ &= \int p(y) \cdot \mu \, dy \\ &= \mu \int p(y) \, dy \\ &= \mu , \end{aligned} \quad (6)$$

and with the expected value of the gamma distribution ( $\rightarrow$  II/3.4.11),  $\mathbb{E}(y)$  becomes

$$\begin{aligned} \mathbb{E}(y) &= \int y \cdot p(y) \, dy \\ &= \langle y \rangle_{\text{Gam}(a, b)} \\ &= \frac{a}{b} . \end{aligned} \quad (7)$$

Thus, the expectation of the random vector in equations (3) and (4) is

$$\mathbb{E} \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} \mu \\ a/b \end{bmatrix} , \quad (8)$$

as indicated by equation (2).

■

## 4.4.5 Covariance

**Theorem:** Let  $x \in \mathbb{R}^n$  and  $y > 0$  follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then,

1) the covariance ( $\rightarrow$  I/1.15.1) of  $x$ , conditional ( $\rightarrow$  I/1.5.4) on  $y$  is

$$\text{Cov}(x|y) = \frac{1}{y} \Lambda^{-1} ; \quad (2)$$

2) the covariance ( $\rightarrow$  I/1.15.1) of  $x$ , unconditional ( $\rightarrow$  I/1.5.3) on  $y$  is

$$\text{Cov}(x) = \frac{b}{a-1} \Lambda^{-1} ; \quad (3)$$

3) the variance ( $\rightarrow$  I/1.13.1) of  $y$  is

$$\text{Var}(y) = \frac{a}{b^2} . \quad (4)$$

**Proof:**

1) According to the definition of the normal-gamma distribution ( $\rightarrow$  II/4.4.1), the distribution of  $x$  given  $y$  is a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) . \quad (5)$$

The covariance of the multivariate normal distribution ( $\rightarrow$  II/4.1.8) is

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \text{Cov}(x) = \Sigma , \quad (6)$$

such that we have:

$$\text{Cov}(x|y) = (y\Lambda)^{-1} = \frac{1}{y} \Lambda^{-1} . \quad (7)$$

2) The marginal distribution of the normal-gamma distribution ( $\rightarrow$  II/4.4.8) with respect to  $x$  is a multivariate t-distribution ( $\rightarrow$  II/4.3.1):

$$x \sim t\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) . \quad (8)$$

The covariance of the multivariate t-distribution is

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad \text{Cov}(x) = \frac{\nu}{\nu-2} \Sigma , \quad (9)$$

such that we have:

$$\text{Cov}(x) = \frac{2a}{2a-2} \left(\frac{a}{b}\Lambda\right)^{-1} = \frac{a}{a-1} \frac{b}{a} \Lambda^{-1} = \frac{b}{a-1} \Lambda^{-1} . \quad (10)$$

3) The marginal distribution of the normal-gamma distribution ( $\rightarrow$  II/4.4.8) with respect to  $y$  is a univariate gamma distribution ( $\rightarrow$  II/3.4.1):

$$y \sim \text{Gam}(a, b) . \quad (11)$$

The variance of the gamma distribution ( $\rightarrow$  II/3.4.14) is

$$x \sim \text{Gam}(a, b) \quad \Rightarrow \quad \text{Var}(x) = \frac{a}{b^2}, \quad (12)$$

such that we have:

$$\text{Var}(y) = \frac{a}{b^2}. \quad (13)$$

■

#### 4.4.6 Differential entropy

**Theorem:** Let  $x$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) and let  $y$  be a positive random variable ( $\rightarrow$  I/1.2.2). Assume that  $x$  and  $y$  are jointly normal-gamma distributed:

$$(x, y) \sim \text{NG}(\mu, \Lambda^{-1}, a, b) \quad (1)$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $x$  in nats is

$$\begin{aligned} h(x, y) = & \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n \\ & + a + \ln \Gamma(a) - \frac{n-2+2a}{2} \psi(a) + \frac{n-2}{2} \ln b. \end{aligned} \quad (2)$$

**Proof:** The probability density function of the normal-gamma distribution ( $\rightarrow$  II/4.4.3) is

$$p(x, y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b). \quad (3)$$

The differential entropy of the multivariate normal distribution ( $\rightarrow$  II/4.1.10) is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2}n \quad (4)$$

and the differential entropy of the univariate gamma distribution ( $\rightarrow$  II/3.4.17) is

$$h(y) = a + \ln \Gamma(a) + (1-a) \cdot \psi(a) - \ln b \quad (5)$$

where  $\Gamma(x)$  is the gamma function and  $\psi(x)$  is the digamma function.

The differential entropy of a continuous random variable ( $\rightarrow$  I/2.2.1) in nats is given by

$$h(Z) = - \int_{\mathcal{Z}} p(z) \ln p(z) dz \quad (6)$$

which, applied to the normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $x$  and  $y$ , yields

$$h(x, y) = - \int_0^\infty \int_{\mathbb{R}^n} p(x, y) \ln p(x, y) dx dy. \quad (7)$$

Using the law of conditional probability ( $\rightarrow$  I/1.3.4), this can be evaluated as follows:



$$\begin{aligned}
h(x, y) &= - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(x|y) p(y) dx dy \\
&= - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(x|y) dx dy - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(y) dx dy \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y) \ln p(x|y) dx dy + \int_0^\infty p(y) \ln p(y) \int_{\mathbb{R}^n} p(x|y) dx dy \\
&= \langle h(x|y) \rangle_{p(y)} + h(y) .
\end{aligned} \tag{8}$$

In other words, the differential entropy of the normal-gamma distribution over  $x$  and  $y$  is equal to the sum of a multivariate normal entropy regarding  $x$  conditional on  $y$ , expected over  $y$ , and a univariate gamma entropy regarding  $y$ .

From equations (3) and (4), the first term becomes

$$\begin{aligned}
\langle h(x|y) \rangle_{p(y)} &= \left\langle \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |(y\Lambda)^{-1}| + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \left\langle \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |(y\Lambda)| + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \left\langle \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(y^n |\Lambda|) + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n - \left\langle \frac{n}{2} \ln y \right\rangle_{p(y)}
\end{aligned} \tag{9}$$

and using the relation ( $\rightarrow$  II/3.4.15)  $y \sim \text{Gam}(a, b) \Rightarrow \langle \ln y \rangle = \psi(a) - \ln(b)$ , we have

$$\langle h(x|y) \rangle_{p(y)} = \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n - \frac{n}{2} \psi(a) + \frac{n}{2} \ln b . \tag{10}$$

By plugging (10) and (5) into (8), one arrives at the differential entropy given by (2). ■

#### 4.4.7 Kullback-Leibler divergence

**Theorem:** Let  $x$  be an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3) and let  $y$  be a positive random variable ( $\rightarrow$  I/1.2.2). Assume two normal-gamma distributions ( $\rightarrow$  II/4.4.1)  $P$  and  $Q$  specifying the joint distribution of  $x$  and  $y$  as

$$\begin{aligned}
P : (x, y) &\sim \text{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\
Q : (x, y) &\sim \text{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) .
\end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\begin{aligned}
\text{KL}[P || Q] &= \frac{1}{2} \frac{a_1}{b_1} [(\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1)] + \frac{1}{2} \text{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \\
&\quad + a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} .
\end{aligned} \tag{2}$$

**Proof:** The probability density function of the normal-gamma distribution ( $\rightarrow$  II/4.4.3) is

$$p(x, y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) . \quad (3)$$

The Kullback-Leibler divergence of the multivariate normal distribution ( $\rightarrow$  II/4.1.12) is

$$\text{KL}[P || Q] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (4)$$

and the Kullback-Leibler divergence of the univariate gamma distribution ( $\rightarrow$  II/3.4.18) is

$$\text{KL}[P || Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} \quad (5)$$

where  $\Gamma(x)$  is the gamma function and  $\psi(x)$  is the digamma function.

The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P || Q] = \int_{\mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)} dz \quad (6)$$

which, applied to the normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $x$  and  $y$ , yields

$$\text{KL}[P || Q] = \int_0^\infty \int_{\mathbb{R}^n} p(x, y) \ln \frac{p(x, y)}{q(x, y)} dx dy . \quad (7)$$

Using the law of conditional probability ( $\rightarrow$  I/1.3.4), this can be evaluated as follows:

$$\begin{aligned} \text{KL}[P || Q] &= \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(x|y) p(y)}{q(x|y) q(y)} dx dy \\ &= \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(y)}{q(y)} dx dy \\ &= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty p(y) \ln \frac{p(y)}{q(y)} \int_{\mathbb{R}^n} p(x|y) dx dy \\ &= \langle \text{KL}[p(x|y) || q(x|y)] \rangle_{p(y)} + \text{KL}[p(y) || q(y)] . \end{aligned} \quad (8)$$

In other words, the KL divergence between two normal-gamma distributions over  $x$  and  $y$  is equal to the sum of a multivariate normal KL divergence regarding  $x$  conditional on  $y$ , expected over  $y$ , and a univariate gamma KL divergence regarding  $y$ .

From equations (3) and (4), the first term becomes

$$\begin{aligned} &\langle \text{KL}[p(x|y) || q(x|y)] \rangle_{p(y)} \\ &= \left\langle \frac{1}{2} \left[ (\mu_2 - \mu_1)^T (y\Lambda_2) (\mu_2 - \mu_1) + \text{tr}((y\Lambda_2)(y\Lambda_1)^{-1}) - \ln \frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - n \right] \right\rangle_{p(y)} \\ &= \left\langle \frac{y}{2} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2} \text{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \right\rangle_{p(y)} \end{aligned} \quad (9)$$

and using the relation ( $\rightarrow$  II/3.4.11)  $y \sim \text{Gam}(a, b) \Rightarrow \langle y \rangle = a/b$ , we have

$$\langle \text{KL}[p(x|y) || q(x|y)] \rangle_{p(y)} = \frac{1}{2} \frac{a_1}{b_1} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2} \text{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} . \quad (10)$$

By plugging (10) and (5) into (8), one arrives at the KL divergence given by (2). ■

#### Sources:

- Soch J, Allefeld A (2016): “Kullback-Leibler Divergence for the Normal-Gamma Distribution”; in: *arXiv math.ST*, 1611.01437; URL: <https://arxiv.org/abs/1611.01437>.

#### 4.4.8 Marginal distributions

**Theorem:** Let  $x$  and  $y$  follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then, the marginal distribution ( $\rightarrow$  I/1.5.3) of  $y$  is a gamma distribution ( $\rightarrow$  II/3.4.1)

$$y \sim \text{Gam}(a, b) \quad (2)$$

and the marginal distribution ( $\rightarrow$  I/1.5.3) of  $x$  is a multivariate t-distribution ( $\rightarrow$  II/4.3.1)

$$x \sim t \left( \mu, \left( \frac{a}{b} \Lambda \right)^{-1}, 2a \right) . \quad (3)$$

**Proof:** The probability density function of the normal-gamma distribution ( $\rightarrow$  II/4.4.3) is given by

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\ p(y) &= \text{Gam}(y; a, b) . \end{aligned} \quad (4)$$

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), the marginal distribution of  $y$  can be derived as

$$\begin{aligned} p(y) &= \int p(x, y) \, dx \\ &= \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \text{Gam}(y; a, b) \, dx \\ &= \text{Gam}(y; a, b) \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \, dx \\ &= \text{Gam}(y; a, b) \end{aligned} \quad (5)$$

which is the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) with shape parameter  $a$  and rate parameter  $b$ .

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), the marginal distribution of  $x$  can be derived as

$$\begin{aligned}
p(x) &= \int p(x, y) \, dy \\
&= \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \text{Gam}(y; a, b) \, dy \\
&= \int \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x - \mu)^T(y\Lambda)(x - \mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] \, dy \\
&= \int \sqrt{\frac{y^n|\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x - \mu)^T(y\Lambda)(x - \mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] \, dy \\
&= \int \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \cdot \exp\left[-\left(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right)y\right] \, dy \\
&= \int \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(a + \frac{n}{2})}{(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu))^{a+\frac{n}{2}}} \cdot \text{Gam}\left(y; a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right) \, dy \\
&= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(a + \frac{n}{2})}{(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu))^{a+\frac{n}{2}}} \int \text{Gam}\left(y; a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right) \, dy \\
&= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(a + \frac{n}{2})}{(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu))^{a+\frac{n}{2}}} \\
&= \frac{\sqrt{|\Lambda|}}{(2\pi)^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot b^a \cdot \left(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right)^{-(a+\frac{n}{2})} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(\frac{1}{b}\right)^{-a} \cdot \left(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right)^{-a} \cdot 2^{-\frac{n}{2}} \cdot \left(b + \frac{1}{2}(x - \mu)^T\Lambda(x - \mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(1 + \frac{1}{2b}(x - \mu)^T\Lambda(x - \mu)\right)^{-a} \cdot (2b + (x - \mu)^T\Lambda(x - \mu))^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(\frac{1}{2a}\right)^{-a} \cdot \left(2a + (x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-a} \cdot \left(\frac{b}{a}\right)^{-\frac{n}{2}} \cdot \left(2a + (x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n |\Lambda|}}{(2a)^{-a} \pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(2a + (x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-a} \cdot \left(2a + (x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n |\Lambda|}}{(2a)^{-a} \pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot (2a)^{-a} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-a} \cdot (2a)^{-\frac{n}{2}} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n |\Lambda|}}{(2a)^{\frac{n}{2}} \pi^{\frac{n}{2}}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-\frac{2a+n}{2}} \\
&= \sqrt{\frac{\left|\frac{a}{b}\Lambda\right|}{(2a\pi)^n}} \cdot \frac{\Gamma(\frac{2a+n}{2})}{\Gamma(\frac{2a}{2})} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T\left(\frac{a}{b}\Lambda\right)(x - \mu)\right)^{-\frac{2a+n}{2}}
\end{aligned}$$

(6)

which is the probability density function of a multivariate t-distribution ( $\rightarrow$  II/4.3.2) with mean vector  $\mu$ , shape matrix  $\left(\frac{a}{b}\Lambda\right)^{-1}$  and  $2a$  degrees of freedom.



#### 4.4.9 Conditional distributions

**Theorem:** Let  $x$  and  $y$  follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then,

1) the conditional distribution ( $\rightarrow$  I/1.5.4) of  $x$  given  $y$  is a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) ; \quad (2)$$

2) the conditional distribution ( $\rightarrow$  I/1.5.4) of a subset vector  $x_1$ , given the complement vector  $x_2$  and  $y$ , is also a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x_1|x_2, y \sim \mathcal{N}(\mu_{1|2}(y), \Sigma_{1|2}(y)) \quad (3)$$

with the conditional mean ( $\rightarrow$  I/1.12.1) and covariance ( $\rightarrow$  I/1.15.1)

$$\begin{aligned} \mu_{1|2}(y) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2}(y) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \quad (4)$$

where  $\mu_1, \mu_2$  and  $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}, \Sigma_{21}$  are block-wise components ( $\rightarrow$  II/4.1.16) of  $\mu$  and  $\Sigma(y) = (y\Lambda)^{-1}$ ;

3) the conditional distribution ( $\rightarrow$  I/1.5.4) of  $y$  given  $x$  is a gamma distribution ( $\rightarrow$  II/3.4.1)

$$y|x \sim \text{Gam}\left(a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu)\right) \quad (5)$$

where  $n$  is the dimensionality of  $x$ .

**Proof:**

1) This follows from the definition of the normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) . \end{aligned} \quad (6)$$

2) This follows from (2) and the conditional distributions of the multivariate normal distribution ( $\rightarrow$  II/4.1.16):

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \Sigma) \\ \Rightarrow x_1|x_2 &\sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} . \end{aligned} \quad (7)$$

3) The conditional density of  $y$  given  $x$  follows from Bayes' theorem ( $\rightarrow$  I/5.3.2) as

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} . \quad (8)$$

The conditional distribution ( $\rightarrow$  I/1.5.4) of  $x$  given  $y$  is a multivariate normal distribution ( $\rightarrow$  II/4.4.3)

$$p(x|y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[ -\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right], \quad (9)$$

the marginal distribution ( $\rightarrow$  I/1.5.3) of  $y$  is a gamma distribution ( $\rightarrow$  II/4.4.8)

$$p(y) = \text{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] \quad (10)$$

and the marginal distribution ( $\rightarrow$  I/1.5.3) of  $x$  is a multivariate t-distribution ( $\rightarrow$  II/4.4.8)

$$\begin{aligned} p(x) &= t \left( x; \mu, \left( \frac{a}{b} \Lambda \right)^{-1}, 2a \right) \\ &= \sqrt{\frac{\left| \frac{a}{b} \Lambda \right|}{(2a\pi)^n}} \cdot \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)} \cdot \left( 1 + \frac{1}{2a}(x - \mu)^T \left( \frac{a}{b} \Lambda \right) (x - \mu) \right)^{-\frac{2a+n}{2}} \\ &= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left( b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{-(a + \frac{n}{2})}. \end{aligned} \quad (11)$$

Plugging (9), (10) and (11) into (8), we obtain

$$\begin{aligned} p(y|x) &= \frac{\sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[ -\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by]}{\sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left( b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{-(a + \frac{n}{2})}} \\ &= y^{\frac{n}{2}} \cdot \exp \left[ -\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \cdot y^{a-1} \cdot \exp[-by] \cdot \frac{1}{\Gamma\left(a + \frac{n}{2}\right)} \cdot \left( b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{a + \frac{n}{2}} \\ &= \frac{\left( b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{a + \frac{n}{2}}}{\Gamma\left(a + \frac{n}{2}\right)} \cdot y^{a + \frac{n}{2} - 1} \cdot \exp \left[ - \left( b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right) y \right] \end{aligned} \quad (12)$$

which is the probability density function of a gamma distribution ( $\rightarrow$  II/3.4.7) with shape and rate parameters

$$a + \frac{n}{2} \quad \text{and} \quad b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu), \quad (13)$$

such that

$$p(y|x) = \text{Gam} \left( y; a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right). \quad (14)$$

■

#### 4.4.10 Drawing samples

**Theorem:** Let  $Z_1 \in \mathbb{R}^n$  be a random vector ( $\rightarrow$  I/1.2.3) with all entries independently following a standard normal distribution ( $\rightarrow$  II/3.2.3) and let  $Z_2 \in \mathbb{R}$  be a random variable ( $\rightarrow$  I/1.2.2) following a standard gamma distribution ( $\rightarrow$  II/3.4.3) with shape  $a$ . Moreover, let  $A \in \mathbb{R}^{n \times n}$  be a matrix, such that  $AA^T = \Lambda^{-1}$ .

Then,  $X = \mu + AZ_1/\sqrt{Z_2/b}$  and  $Y = Z_2/b$  jointly follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1) with mean vector ( $\rightarrow$  I/1.12.16)  $\mu$ , precision matrix ( $\rightarrow$  I/1.16.12)  $\Lambda$ , shape parameter  $a$  and rate parameter  $b$ :

$$\left( X = \mu + AZ_1/\sqrt{Z_2/b}, Y = Z_2/b \right) \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

**Proof:** If all entries of  $Z_1$  are independent and standard normally distributed ( $\rightarrow$  II/3.2.3)

$$z_{1i} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for all } i = 1, \dots, n , \quad (2)$$

this implies a multivariate normal distribution with diagonal covariance matrix ( $\rightarrow$  II/4.1.17):

$$Z_1 \sim \mathcal{N}(0_n, I_n) \quad (3)$$

where  $0_n$  is an  $n \times 1$  matrix of zeros and  $I_n$  is the  $n \times n$  identity matrix.

If the distribution of  $Z_2$  is a standard gamma distribution ( $\rightarrow$  II/3.4.3)

$$Z_2 \sim \text{Gam}(a, 1) , \quad (4)$$

then due to the relationship between gamma and standard gamma distribution ( $\rightarrow$  II/3.4.4), we have:

$$Y = \frac{Z_2}{b} \sim \text{Gam}(a, b) . \quad (5)$$

Moreover, using the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14), it follows that:

$$\begin{aligned} Z_1 &\sim \mathcal{N}(0_n, I_n) \\ X = \mu + \frac{1}{\sqrt{Z_2/b}} AZ_1 &\sim \mathcal{N} \left( \mu + \frac{1}{\sqrt{Z_2/b}} A 0_n, \left( \frac{1}{\sqrt{Z_2/b}} A \right) I_n \left( \frac{1}{\sqrt{Z_2/b}} A \right)^T \right) \\ X &\sim \mathcal{N} \left( \mu + 0_n, \left( \frac{1}{\sqrt{Y}} \right)^2 AA^T \right) \\ X &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) . \end{aligned} \quad (6)$$

Thus,  $Y$  follows a gamma distribution ( $\rightarrow$  II/3.4.1) and the distribution of  $X$  conditional on  $Y$  is a multivariate normal distribution ( $\rightarrow$  II/4.1.1):

$$\begin{aligned} X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\ Y &\sim \text{Gam}(a, b) . \end{aligned} \quad (7)$$

This means that, by definition ( $\rightarrow$  II/4.4.1),  $X$  and  $Y$  jointly follow a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$X, Y \sim \text{NG}(\mu, \Lambda, a, b) , \quad (8)$$

Thus, given  $Z_1$  defined by (2) and  $Z_2$  defined by (4),  $X$  and  $Y$  defined by (1) are a sample ( $\rightarrow$  I/1.1.6) from  $\text{NG}(\mu, \Lambda, a, b)$ . ■

#### Sources:

- Wikipedia (2022): “Normal-gamma distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: [https://en.wikipedia.org/wiki/Normal-gamma\\_distribution#Generating\\_normal-gamma\\_random\\_variates](https://en.wikipedia.org/wiki/Normal-gamma_distribution#Generating_normal-gamma_random_variates).

## 4.5 Dirichlet distribution

### 4.5.1 Definition

**Definition:** Let  $X$  be a  $1 \times k$  random vector ( $\rightarrow$  I/1.2.3). Then,  $X$  is said to follow a Dirichlet distribution with concentration parameters  $\alpha = [\alpha_1, \dots, \alpha_k]$

$$X \sim \text{Dir}(\alpha) , \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\text{Dir}(x; \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1} \quad (2)$$

where  $\alpha_i > 0$  for all  $i = 1, \dots, k$ , and the density is zero, if  $x_i \notin [0, 1]$  for any  $i = 1, \dots, k$  or  $\sum_{i=1}^k x_i \neq 1$ .

#### Sources:

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: [https://en.wikipedia.org/wiki/Dirichlet\\_distribution#Probability\\_density\\_function](https://en.wikipedia.org/wiki/Dirichlet_distribution#Probability_density_function).

### 4.5.2 Probability density function

**Theorem:** Let  $X$  be a random vector ( $\rightarrow$  I/1.2.3) following a Dirichlet distribution ( $\rightarrow$  II/4.5.1):

$$X \sim \text{Dir}(\alpha) . \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f_X(x) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1} . \quad (2)$$

**Proof:** This follows directly from the definition of the Dirichlet distribution ( $\rightarrow$  II/4.5.1). ■



### 4.5.3 Kullback-Leibler divergence

**Theorem:** Let  $X$  be an  $1 \times k$  random vector ( $\rightarrow$  I/1.2.3). Assume two Dirichlet distributions ( $\rightarrow$  II/4.5.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \text{Dir}(\alpha_1) \\ Q : X &\sim \text{Dir}(\alpha_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \left[ \psi(\alpha_{1i}) - \psi\left(\sum_{i=1}^k \alpha_{1i}\right) \right] . \quad (2)$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the Dirichlet distributions ( $\rightarrow$  II/4.1.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{\mathcal{X}^k} \text{Dir}(x; \alpha_1) \ln \frac{\text{Dir}(x; \alpha_1)}{\text{Dir}(x; \alpha_2)} dx \\ &= \left\langle \ln \frac{\text{Dir}(x; \alpha_1)}{\text{Dir}(x; \alpha_2)} \right\rangle_{p(x)} \end{aligned} \quad (4)$$

where  $\mathcal{X}^k$  is the set  $\left\{x \in \mathbb{R}^k \mid \sum_{i=1}^k x_i = 1, 0 \leq x_i \leq 1, i = 1, \dots, k\right\}$ .

Using the probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\prod_{i=1}^k \Gamma(\alpha_{1i})} \prod_{i=1}^k x_i^{\alpha_{1i}-1}}{\frac{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)}{\prod_{i=1}^k \Gamma(\alpha_{2i})} \prod_{i=1}^k x_i^{\alpha_{2i}-1}} \right\rangle_{p(x)} \\ &= \left\langle \ln \left( \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} \cdot \frac{\prod_{i=1}^k \Gamma(\alpha_{2i})}{\prod_{i=1}^k \Gamma(\alpha_{1i})} \cdot \prod_{i=1}^k x_i^{\alpha_{1i}-\alpha_{2i}} \right) \right\rangle_{p(x)} \\ &= \left\langle \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \cdot \ln(x_i) \right\rangle_{p(x)} \\ &= \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \cdot \langle \ln x_i \rangle_{p(x)} . \end{aligned} \quad (5)$$

Using the expected value of a logarithmized Dirichlet variate

$$X \sim \text{Dir}(\alpha) \Rightarrow \langle \ln x_i \rangle = \psi(\alpha_i) - \psi\left(\sum_{i=1}^k \alpha_i\right), \quad (6)$$

the Kullback-Leibler divergence from (5) becomes:

$$\text{KL}[P || Q] = \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \cdot \left[ \psi(\alpha_{1i}) - \psi\left(\sum_{i=1}^k \alpha_{1i}\right) \right] \quad (7)$$

■

#### Sources:

- Penny, William D. (2001): “KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities”; in: *University College, London*, p. 2, eqs. 8-9; URL: <https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps>.

#### 4.5.4 Exceedance probabilities

**Theorem:** Let  $r = [r_1, \dots, r_k]$  be a random vector ( $\rightarrow$  I/1.2.3) following a Dirichlet distribution ( $\rightarrow$  II/4.5.1) with concentration parameters  $\alpha = [\alpha_1, \dots, \alpha_k]$ :

$$r \sim \text{Dir}(\alpha). \quad (1)$$

1) If  $k = 2$ , then the exceedance probability ( $\rightarrow$  I/1.3.11) for  $r_1$  is

$$\varphi_1 = 1 - \frac{B\left(\frac{1}{2}; \alpha_1, \alpha_2\right)}{B(\alpha_1, \alpha_2)} \quad (2)$$

where  $B(x, y)$  is the beta function and  $B(x; a, b)$  is the incomplete beta function.

2) If  $k > 2$ , then the exceedance probability ( $\rightarrow$  I/1.3.11) for  $r_i$  is

$$\varphi_i = \int_0^\infty \prod_{j \neq i} \left( \frac{\gamma(\alpha_j, q_j)}{\Gamma(\alpha_j)} \right) \frac{q_i^{\alpha_i-1} \exp[-q_i]}{\Gamma(\alpha_i)} dq_i. \quad (3)$$

where  $\Gamma(x)$  is the gamma function and  $\gamma(s, x)$  is the lower incomplete gamma function.

**Proof:** In the context of the Dirichlet distribution ( $\rightarrow$  II/4.5.1), the exceedance probability ( $\rightarrow$  I/1.3.11) for a particular  $r_i$  is defined as:

$$\begin{aligned} \varphi_i &= p\left(\forall j \in \{1, \dots, k \mid j \neq i\} : r_i > r_j \mid \alpha\right) \\ &= p\left(\bigwedge_{j \neq i} r_i > r_j \mid \alpha\right). \end{aligned} \quad (4)$$

The probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2) is given by:

$$\text{Dir}(r; \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k r_i^{\alpha_i-1}. \quad (5)$$

Note that the probability density function is only calculated, if

$$r_i \in [0, 1] \quad \text{for } i = 1, \dots, k \quad \text{and} \quad \sum_{i=1}^k r_i = 1, \quad (6)$$

and defined to be zero otherwise ( $\rightarrow$  II/4.5.1).

1) If  $k = 2$ , the probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2) reduces to

$$p(r) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} r_1^{\alpha_1-1} r_2^{\alpha_2-1} \quad (7)$$

which is equivalent to the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5)

$$p(r_1) = \frac{r_1^{\alpha_1-1} (1 - r_1)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)} \quad (8)$$

with the beta function given by

$$B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x + y)}. \quad (9)$$

With (6), the exceedance probability for this bivariate case simplifies to

$$\varphi_1 = p(r_1 > r_2) = p(r_1 > 1 - r_1) = p(r_1 > 1/2) = \int_{\frac{1}{2}}^1 p(r_1) dr_1. \quad (10)$$

Using the cumulative distribution function of the beta distribution ( $\rightarrow$  II/3.9.7), it evaluates to

$$\varphi_1 = 1 - \int_0^{\frac{1}{2}} p(r_1) dr_1 = 1 - \frac{B(\frac{1}{2}; \alpha_1, \alpha_2)}{B(\alpha_1, \alpha_2)} \quad (11)$$

with the incomplete beta function

$$B(x; a, b) = \int_0^x x^{a-1} (1 - x)^{b-1} dx. \quad (12)$$

2) If  $k > 2$ , there is no similarly simple expression, because in general

$$\varphi_i = p(r_i = \max(r)) > p(r_i > 1/2) \quad \text{for } i = 1, \dots, k, \quad (13)$$

i.e. exceedance probabilities cannot be evaluated using a simple threshold on  $r_i$ , because  $r_i$  might be the maximal element in  $r$  without being larger than 1/2. Instead, we make use of the relationship between the Dirichlet and the gamma distribution which states that

$$\begin{aligned} Y_1 &\sim \text{Gam}(\alpha_1, \beta), \dots, Y_k \sim \text{Gam}(\alpha_k, \beta), Y_s = \sum_{i=1}^k Y_i \\ \Rightarrow X &= (X_1, \dots, X_k) = \left( \frac{Y_1}{Y_s}, \dots, \frac{Y_k}{Y_s} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_k). \end{aligned} \quad (14)$$

The probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7) is given by

$$\text{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \quad \text{for } x > 0. \quad (15)$$

Consider the gamma random variables ( $\rightarrow$  II/3.4.1)

$$q_1 \sim \text{Gam}(\alpha_1, 1), \dots, q_k \sim \text{Gam}(\alpha_k, 1), \quad q_s = \sum_{j=1}^k q_j \quad (16)$$

and the Dirichlet random vector ( $\rightarrow$  II/4.5.1)

$$r = (r_1, \dots, r_k) = \left( \frac{q_1}{q_s}, \dots, \frac{q_k}{q_s} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_k). \quad (17)$$

Obviously, it holds that

$$r_i > r_j \Leftrightarrow q_i > q_j \quad \text{for } i, j = 1, \dots, k \quad \text{with } j \neq i. \quad (18)$$

Therefore, consider the probability that  $q_i$  is larger than  $q_j$ , given  $q_i$  is known. This probability is equal to the probability that  $q_j$  is smaller than  $q_i$ , given  $q_i$  is known

$$p(q_i > q_j | q_i) = p(q_j < q_i | q_i) \quad (19)$$

which can be expressed in terms of the cumulative distribution function of the gamma distribution ( $\rightarrow$  II/3.4.9) as

$$p(q_j < q_i | q_i) = \int_0^{q_i} \text{Gam}(q_j; \alpha_j, 1) dq_j = \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \quad (20)$$

where  $\Gamma(x)$  is the gamma function and  $\gamma(s, x)$  is the lower incomplete gamma function. Since the gamma variates are independent of each other, these probabilities factorize:

$$p(\forall_{j \neq i} [q_i > q_j] | q_i) = \prod_{j \neq i} p(q_i > q_j | q_i) = \prod_{j \neq i} \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)}. \quad (21)$$

In order to obtain the exceedance probability  $\varphi_i$ , the dependency on  $q_i$  in this probability still has to be removed. From equations (4) and (18), it follows that

$$\varphi_i = p(\forall_{j \neq i} [r_i > r_j]) = p(\forall_{j \neq i} [q_i > q_j]). \quad (22)$$

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), we have

$$\varphi_i = \int_0^\infty p(\forall_{j \neq i} [q_i > q_j] | q_i) p(q_i) dq_i. \quad (23)$$

With (21) and (16), this becomes

$$\varphi_i = \int_0^\infty \prod_{j \neq i} (p(q_i > q_j | q_i)) \cdot \text{Gam}(q_i; \alpha_i, 1) dq_i. \quad (24)$$

And with (20) and (15), it becomes

$$\varphi_i = \int_0^\infty \prod_{j \neq i} \left( \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \right) \cdot \frac{q_i^{\alpha_i-1} \exp[-q_i]}{\Gamma(\alpha_i)} dq_i. \quad (25)$$

In other words, the exceedance probability ( $\rightarrow$  I/1.3.11) for one element from a Dirichlet-distributed ( $\rightarrow$  II/4.5.1) random vector ( $\rightarrow$  I/1.2.3) is an integral from zero to infinity where the first term in the integrand conforms to a product of gamma ( $\rightarrow$  II/3.4.1) cumulative distribution functions ( $\rightarrow$  I/1.8.1) and the second term is a gamma ( $\rightarrow$  II/3.4.1) probability density function ( $\rightarrow$  I/1.7.1).

■

**Sources:**

- Soch J, Allefeld C (2016): “Exceedance Probabilities for the Dirichlet Distribution”; in: *arXiv stat.AP*, 1611.01439; URL: <https://arxiv.org/abs/1611.01439>.

## 5 Matrix-variate continuous distributions

### 5.1 Matrix-normal distribution

#### 5.1.1 Definition

**Definition:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4). Then,  $X$  is said to be matrix-normally distributed with mean  $M$ , covariance ( $\rightarrow$  I/1.16.1) across rows  $U$  and covariance ( $\rightarrow$  I/1.16.1) across columns  $V$

$$X \sim \mathcal{MN}(M, U, V), \quad (1)$$

if and only if its probability density function ( $\rightarrow$  I/1.7.1) is given by

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right] \quad (2)$$

where  $M$  is an  $n \times p$  real matrix,  $U$  is an  $n \times n$  positive-definite matrix and  $V$  is a  $p \times p$  positive-definite matrix.

#### Sources:

- Wikipedia (2020): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Definition](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Definition).

#### 5.1.2 Equivalence to multivariate normal distribution

**Theorem:** The matrix  $X$  is matrix-normally distributed ( $\rightarrow$  II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V), \quad (1)$$

if and only if  $\text{vec}(X)$  is multivariate normally distributed ( $\rightarrow$  II/4.1.1)

$$\text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) \quad (2)$$

where  $\text{vec}(X)$  is the vectorization operator and  $\otimes$  is the Kronecker product.

**Proof:** The probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3) with  $n \times p$  mean  $M$ ,  $n \times n$  covariance across rows  $U$  and  $p \times p$  covariance across columns  $V$  is

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right]. \quad (3)$$

Using the trace property  $\text{tr}(ABC) = \text{tr}(BCA)$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} ((X - M)^T U^{-1} (X - M) V^{-1}) \right]. \quad (4)$$

Using the trace-vectorization relation  $\text{tr}(A^T B) = \text{vec}(A)^T \text{vec}(B)$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{vec}(X - M)^T \text{vec} (U^{-1} (X - M) V^{-1}) \right]. \quad (5)$$

Using the vectorization-Kronecker relation  $\text{vec}(ABC) = (C^T \otimes A) \text{vec}(B)$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{vec}(X - M)^T (V^{-1} \otimes U^{-1}) \text{vec}(X - M) \right]. \quad (6)$$

Using the Kronecker product property  $(A^{-1} \otimes B^{-1}) = (A \otimes B)^{-1}$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{vec}(X - M)^T (V \otimes U)^{-1} \text{vec}(X - M) \right]. \quad (7)$$

Using the vectorization property  $\text{vec}(A + B) = \text{vec}(A) + \text{vec}(B)$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[ -\frac{1}{2} [\text{vec}(X) - \text{vec}(M)]^T (V \otimes U)^{-1} [\text{vec}(X) - \text{vec}(M)] \right]. \quad (8)$$

Using the Kronecker-determinant relation  $|A \otimes B| = |A|^m |B|^n$ , we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V \otimes U|}} \cdot \exp \left[ -\frac{1}{2} [\text{vec}(X) - \text{vec}(M)]^T (V \otimes U)^{-1} [\text{vec}(X) - \text{vec}(M)] \right]. \quad (9)$$

This is the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4) with the  $np \times 1$  mean vector  $\text{vec}(M)$  and the  $np \times np$  covariance matrix  $V \otimes U$ :

$$\mathcal{MN}(X; M, U, V) = \mathcal{N}(\text{vec}(X); \text{vec}(M), V \otimes U). \quad (10)$$

By showing that the probability density functions ( $\rightarrow$  I/1.7.1) are identical, it is proven that the associated probability distributions ( $\rightarrow$  I/1.5.1) are equivalent. ■

#### Sources:

- Wikipedia (2020): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Proof](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Proof).

### 5.1.3 Probability density function

**Theorem:** Let  $X$  be a random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V). \quad (1)$$

Then, the probability density function ( $\rightarrow$  I/1.7.1) of  $X$  is

$$f(X) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1}(X - M)^T U^{-1}(X - M)) \right]. \quad (2)$$

**Proof:** This follows directly from the definition of the matrix-normal distribution ( $\rightarrow$  II/5.1.1). ■

### 5.1.4 Mean

**Theorem:** Let  $X$  be a random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then, the mean or expected value ( $\rightarrow$  I/1.12.1) of  $X$  is

$$E(X) = M . \quad (2)$$

**Proof:** When  $X$  follows a matrix-normal distribution ( $\rightarrow$  II/5.1.1), its vectorized version follows a multivariate normal distribution ( $\rightarrow$  II/5.1.2)

$$\text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) \quad (3)$$

and the expected value of this multivariate normal distribution is ( $\rightarrow$  II/4.1.6)

$$E[\text{vec}(X)] = \text{vec}(M) . \quad (4)$$

Since the expected value of a random matrix is calculated element-wise ( $\rightarrow$  I/1.12.19), we can invert the vectorization operator to get:

$$E[X] = M . \quad (5)$$

■

#### Sources:

- Wikipedia (2022): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-15; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Expected\\_values](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Expected_values).

### 5.1.5 Covariance

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then,

1) the covariance matrix ( $\rightarrow$  I/1.16.1) of each row of  $X$  is a scalar multiple of  $V$

$$\text{Cov}(x_{i,\bullet}^T) \propto V \quad \text{for all } i = 1, \dots, n ; \quad (2)$$

2) the covariance matrix ( $\rightarrow$  I/1.16.1) of each column of  $X$  is a scalar multiple of  $U$

$$\text{Cov}(x_{\bullet,j}) \propto U \quad \text{for all } j = 1, \dots, p . \quad (3)$$

#### Proof:

1) The marginal distribution ( $\rightarrow$  I/1.5.3) of a given row of  $X$  is a multivariate normal distribution ( $\rightarrow$  II/5.1.10)

$$x_{i,\bullet}^T \sim \mathcal{N}(m_{i,\bullet}^T, u_{ii}V) , \quad (4)$$



and the covariance of this multivariate normal distribution ( $\rightarrow$  II/4.1.8) is

$$\text{Cov}(x_{i,\bullet}^T) = u_{ii}V \propto V . \quad (5)$$

2) The marginal distribution ( $\rightarrow$  I/1.5.3) of a given column of  $X$  is a multivariate normal distribution ( $\rightarrow$  II/5.1.10)

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj}U) , \quad (6)$$

and the covariance of this multivariate normal distribution ( $\rightarrow$  II/4.1.8) is

$$\text{Cov}(x_{\bullet,j}) = v_{jj}U \propto U . \quad (7)$$

#### Sources:

- Wikipedia (2022): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-15; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Expected\\_values](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Expected_values).

#### 5.1.6 Differential entropy

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then, the differential entropy ( $\rightarrow$  I/2.2.1) of  $X$  in nats is

$$h(X) = \frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |V| + \frac{p}{2} \ln |U| + \frac{np}{2} . \quad (2)$$

**Proof:** The matrix-normal distribution is equivalent to the multivariate normal distribution ( $\rightarrow$  II/5.1.2),

$$X \sim \mathcal{MN}(M, U, V) \Leftrightarrow \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) , \quad (3)$$

and the differential entropy for the multivariate normal distribution ( $\rightarrow$  II/4.1.10) in nats is

$$X \sim \mathcal{N}(\mu, \Sigma) \Rightarrow h(X) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2}n \quad (4)$$

where  $X$  is an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3). Thus, we can plug the distribution parameters from (1) into the differential entropy in (4) using the relationship given by (3):

$$h(X) = \frac{np}{2} \ln(2\pi) + \frac{1}{2} \ln |V \otimes U| + \frac{1}{2}np . \quad (5)$$

Using the Kronecker product property

$$|A \otimes B| = |A|^m |B|^n \quad \text{where} \quad A \in \mathbb{R}^{n \times n} \quad \text{and} \quad B \in \mathbb{R}^{m \times m} , \quad (6)$$

the differential entropy from (5) becomes:

$$\begin{aligned} h(X) &= \frac{np}{2} \ln(2\pi) + \frac{1}{2} \ln (|V|^n |U|^p) + \frac{1}{2}np \\ &= \frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |V| + \frac{p}{2} \ln |U| + \frac{np}{2} . \end{aligned} \quad (7)$$



### 5.1.7 Kullback-Leibler divergence

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4). Assume two matrix-normal distributions ( $\rightarrow$  II/5.1.1)  $P$  and  $Q$  specifying the probability distribution of  $X$  as

$$\begin{aligned} P : X &\sim \mathcal{MN}(M_1, U_1, V_1) \\ Q : X &\sim \mathcal{MN}(M_2, U_2, V_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left[ \text{vec}(M_2 - M_1)^T \text{vec}(U_2^{-1}(M_2 - M_1)V_2^{-1}) \right. \\ &\quad \left. + \text{tr}((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] . \end{aligned} \quad (2)$$

**Proof:** The matrix-normal distribution is equivalent to the multivariate normal distribution ( $\rightarrow$  II/5.1.2),

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) , \quad (3)$$

and the Kullback-Leibler divergence for the multivariate normal distribution ( $\rightarrow$  II/4.1.12) is

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (4)$$

where  $X$  is an  $n$ -dimensional random vector ( $\rightarrow$  I/1.2.3).

Thus, we can plug the distribution parameters from (1) into the KL divergence in (4) using the relationship given by (3)

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left[ (\text{vec}(M_2) - \text{vec}(M_1))^T (V_2 \otimes U_2)^{-1} (\text{vec}(M_2) - \text{vec}(M_1)) \right. \\ &\quad \left. + \text{tr}((V_2 \otimes U_2)^{-1} (V_1 \otimes U_1)) - \ln \frac{|V_1 \otimes U_1|}{|V_2 \otimes U_2|} - np \right] . \end{aligned} \quad (5)$$

Using the vectorization operator and Kronecker product properties

$$\text{vec}(A) + \text{vec}(B) = \text{vec}(A + B) \quad (6)$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \quad (7)$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \quad (8)$$

$$|A \otimes B| = |A|^m |B|^n \quad \text{where} \quad A \in \mathbb{R}^{n \times n} \quad \text{and} \quad B \in \mathbb{R}^{m \times m} , \quad (9)$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{aligned} \text{KL}[P || Q] = & \frac{1}{2} \left[ \text{vec}(M_2 - M_1)^T (V_2^{-1} \otimes U_2^{-1}) \text{vec}(M_2 - M_1) \right. \\ & \left. + \text{tr}((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] . \end{aligned} \quad (10)$$

Using the relationship between Kronecker product and vectorization operator

$$(C^T \otimes A) \text{vec}(B) = \text{vec}(ABC) , \quad (11)$$

we finally have:

$$\begin{aligned} \text{KL}[P || Q] = & \frac{1}{2} \left[ \text{vec}(M_2 - M_1)^T \text{vec}(U_2^{-1}(M_2 - M_1)V_2^{-1}) \right. \\ & \left. + \text{tr}((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] . \end{aligned} \quad (12)$$

■

### 5.1.8 Transposition

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then, the transpose of  $X$  also has a matrix-normal distribution:

$$X^T \sim \mathcal{MN}(M^T, V, U) . \quad (2)$$

**Proof:** For a random vector ( $\rightarrow$  I/1.2.3)  $X \in \mathbb{R}^n$  with probability density function ( $\rightarrow$  I/1.7.1)  $f_X(x)$ , the probability density function of the invertible function ( $\rightarrow$  I/1.7.5)  $Y = g(X)$  is

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |J_{g^{-1}}(y)| , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \quad (3)$$

where  $|J_{g^{-1}}(y)|$  is the determinant of the Jacobian matrix

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ \vdots & \ddots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix} \quad (4)$$

and  $\mathcal{Y}$  is the set of possible outcomes of  $Y$ :

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} . \quad (5)$$

In the present case, we have  $Y = g(X) = X^T$  and  $X = g^{-1}(Y) = Y^T$  and all  $Y \in \mathcal{Y} = \mathbb{R}^{p \times n}$ . For the vectorized matrices  $X$  and  $Y$ , the Jacobian matrix is

$$J_{g^{-1}}(Y) = \begin{bmatrix} \frac{dx_{11}}{dy_{11}} & \frac{dx_{11}}{dy_{21}} & \cdots & \frac{dx_{11}}{dy_{pn}} \\ \frac{dx_{21}}{dy_{11}} & \frac{dx_{21}}{dy_{21}} & \cdots & \frac{dx_{21}}{dy_{pn}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dx_{np}}{dy_{11}} & \frac{dx_{np}}{dy_{21}} & \cdots & \frac{dx_{np}}{dy_{pn}} \end{bmatrix} \in \mathbb{R}^{np}. \quad (6)$$

Because by transposition,  $y_{ji} = x_{ij}$ , we have

$$\frac{dx_{ij}}{dy_{kl}} = \begin{cases} 1, & \text{if } k = j \text{ and } l = i \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Thus,  $J_{g^{-1}}(Y)$  is row-equivalent to  $I_{np}$  and  $|J_{g^{-1}}(Y)| = |I_{np}| = 1$ . Therefore, we have:

$$\begin{aligned} f_Y(Y) &= f_X(g^{-1}(Y)) |J_{g^{-1}}(Y)| \\ &= \mathcal{MN}(Y^T; M, U, V) |I_{np}| \\ &= \mathcal{MN}(Y^T; M, U, V). \end{aligned} \quad (8)$$

The probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3) is:

$$f(X) = \mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right]. \quad (9)$$

Thus, substituting  $X = Y^T$ , we get:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^n}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (Y^T - M)^T U^{-1} (Y^T - M)) \right]. \quad (10)$$

Using  $(A + B)^T = (A^T + B^T)$ , we have:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^n}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (Y - M^T) U^{-1} (Y - M^T)^T) \right]. \quad (11)$$

Using  $\text{tr}(ABC) = \text{tr}(CAB)$ , we obtain

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^n}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (U^{-1} (Y - M^T)^T V^{-1} (Y - M^T)) \right] \quad (12)$$

which is the probability density function of a matrix-normal distribution ( $\rightarrow$  II/5.1.3) with mean  $M^T$ , covariance across rows  $V$  and covariance across columns  $U$ :

$$f(Y) = \mathcal{MN}(Y; M^T, V, U). \quad (13)$$

■

### 5.1.9 Linear transformation

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then, a linear transformation of  $X$  is also matrix-normally distributed

$$Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^T, B^TVB) \quad (2)$$

where  $A$  is an  $r \times n$  matrix of full rank  $r \leq n$  and  $B$  is a  $p \times s$  matrix of full rank  $s \leq p$  and  $C$  is an  $r \times s$  matrix.

**Proof:** The matrix-normal distribution is equivalent to the multivariate normal distribution ( $\rightarrow$  II/5.1.2),

$$X \sim \mathcal{MN}(M, U, V) \Leftrightarrow \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) , \quad (3)$$

and the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14) states:

$$x \sim \mathcal{N}(\mu, \Sigma) \Rightarrow y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) . \quad (4)$$

The vectorization of  $Y = AXB + C$  is

$$\begin{aligned} \text{vec}(Y) &= \text{vec}(AXB + C) \\ &= \text{vec}(AXB) + \text{vec}(C) \\ &= (B^T \otimes A)\text{vec}(X) + \text{vec}(C) \end{aligned} \quad (5)$$

and the Kronecker product obeys

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) . \quad (6)$$

Using (3) and (4), we have

$$\begin{aligned} \text{vec}(Y) &\sim \mathcal{N}((B^T \otimes A)\text{vec}(M) + \text{vec}(C), (B^T \otimes A)(V \otimes U)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB) + \text{vec}(C), (B^TV \otimes AU)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB + C), B^TVB \otimes AUA^T) . \end{aligned} \quad (7)$$

Using (3), we finally have:

$$Y \sim \mathcal{MN}(AMB + C, AUA^T, B^TVB) . \quad (8)$$

■

### 5.1.10 Marginal distributions

**Theorem:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) following a matrix-normal distribution ( $\rightarrow$  II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then,

1) the marginal distribution ( $\rightarrow$  I/1.5.3) of any subset matrix  $X_{I,J}$ , obtained by dropping some rows and/or columns from  $X$ , is also a matrix-normal distribution ( $\rightarrow$  II/5.1.1)

$$X_{I,J} \sim \mathcal{MN}(M_{I,J}, U_{I,I}, V_{J,J}) \quad (2)$$

where  $I \subseteq \{1, \dots, n\}$  is an (ordered) subset of all row indices and  $J \subseteq \{1, \dots, p\}$  is an (ordered) subset of all column indices, such that  $M_{I,J}$  is the matrix dropping the irrelevant rows and columns (the ones not in the subset, i.e. marginalized out) from the mean matrix  $M$ ;  $U_{I,I}$  is the matrix dropping rows not in  $I$  from  $U$ ; and  $V_{J,J}$  is the matrix dropping columns not in  $J$  from  $V$ ;

2) the marginal distribution ( $\rightarrow$  I/1.5.3) of each row vector is a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x_{i,\bullet}^T \sim \mathcal{N}(m_{i,\bullet}^T, u_{ii}V) \quad (3)$$

where  $m_{i,\bullet}$  is the  $i$ -th row of  $M$  and  $u_{ii}$  is the  $i$ -th diagonal entry of  $U$ ;

3) the marginal distribution ( $\rightarrow$  I/1.5.3) of each column vector is a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj}U) \quad (4)$$

where  $m_{\bullet,j}$  is the  $j$ -th column of  $M$  and  $v_{jj}$  is the  $j$ -th diagonal entry of  $V$ ; and

4) the marginal distribution ( $\rightarrow$  I/1.5.3) of one element of  $X$  is a univariate normal distribution ( $\rightarrow$  II/3.2.1)

$$x_{ij} \sim \mathcal{N}(m_{ij}, u_{ii}v_{jj}) \quad (5)$$

where  $m_{ij}$  is the  $(i, j)$ -th entry of  $M$ .

**Proof:**

1) Define a selector matrix  $A$ , such that  $a_{ij} = 1$ , if the  $i$ -th row in the subset matrix should be the  $j$ -th row from the original matrix (and  $a_{ij} = 0$  otherwise)

$$A \in \mathbb{R}^{|I| \times n}, \quad \text{s.t.} \quad a_{ij} = \begin{cases} 1, & \text{if } I_i = j \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

and define a selector matrix  $B$ , such that  $b_{ij} = 1$ , if the  $j$ -th column in the subset matrix should be the  $i$ -th column from the original matrix (and  $b_{ij} = 0$  otherwise)

$$B \in \mathbb{R}^{p \times |J|}, \quad \text{s.t.} \quad b_{ij} = \begin{cases} 1, & \text{if } J_j = i \\ 0, & \text{otherwise} . \end{cases} \quad (7)$$

Then,  $X_{I,J}$  can be expressed as

$$X_{I,J} = AXB \quad (8)$$

and we can apply the linear transformation theorem ( $\rightarrow$  II/5.1.9) to give

$$X_{I,J} \sim \mathcal{MN}(AMB, AU A^T, B^T V B) . \quad (9)$$

Finally, we see that  $AMB = M_{I,J}$ ,  $AU A^T = U_{I,I}$  and  $B^T V B = V_{J,J}$ .

2) This is a special case of 1). Setting  $A$  to the  $i$ -th elementary row vector in  $n$  dimensions and  $B$  to the  $p \times p$  identity matrix

$$A = e_i, \quad B = I_p , \quad (10)$$

the  $i$ -th row of  $X$  can be expressed as

$$\begin{aligned} x_{i,\bullet} &= AXB = e_i X I_p = e_i X \\ &\stackrel{(9)}{\sim} \mathcal{MN}(m_{i,\bullet}, u_{ii}, V) . \end{aligned} \quad (11)$$

Thus, the transpose of the row vector is distributed as ( $\rightarrow$  II/5.1.8)

$$x_{i,\bullet}^T \sim \mathcal{MN}(m_{i,\bullet}^T, V, u_{ii}) \quad (12)$$

which is equivalent to a multivariate normal distribution ( $\rightarrow$  II/5.1.2):

$$x_{i,\bullet}^T \sim \mathcal{N}(m_{i,\bullet}^T, u_{ii} V) . \quad (13)$$

3) This is a special case of 1). Setting  $A$  to the  $n \times n$  identity matrix and  $B$  to the  $j$ -th elementary row vector in  $p$  dimensions

$$A = I_n, \quad B = e_j^T , \quad (14)$$

the  $j$ -th column of  $X$  can be expressed as

$$\begin{aligned} x_{\bullet,j} &= AXB = I_n X e_j^T = X e_j^T \\ &\stackrel{(9)}{\sim} \mathcal{MN}(m_{\bullet,j}, U, v_{jj}) \end{aligned} \quad (15)$$

which is equivalent to a multivariate normal distribution ( $\rightarrow$  II/5.1.2):

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj} U) . \quad (16)$$

4) This is a special case of 2) and 3). Setting  $A$  to the  $i$ -th elementary row vector in  $n$  dimensions and  $B$  to the  $j$ -th elementary row vector in  $p$  dimensions

$$A = e_i, \quad B = e_j^T , \quad (17)$$

the  $(i, j)$ -th entry of  $X$  can be expressed as

$$\begin{aligned} x_{ij} &= AXB = e_i X e_j^T \\ &\stackrel{(9)}{\sim} \mathcal{MN}(m_{ij}, u_{ii}, v_{jj}) . \end{aligned} \quad (18)$$

As  $x_{ij}$  is a scalar, this is equivalent to a univariate normal distribution ( $\rightarrow$  II/3.2.1) as a special case ( $\rightarrow$  II/3.2.2) of the matrix-normal distribution ( $\rightarrow$  II/4.1.2):

$$x_{ij} \sim \mathcal{N}(m_{ij}, u_{ii}v_{jj}) . \quad (19)$$

■

### 5.1.11 Redundancy of parameters

**Theorem:** The covariance parameters of the matrix-normal distribution ( $\rightarrow$  II/5.1.1) are redundant up to a scalar factor, i.e. the two probability distributions ( $\rightarrow$  I/1.5.1)

$$\begin{aligned} X &\sim \mathcal{MN}(M, U, V) \\ X &\sim \mathcal{MN}\left(M, a \cdot U, \frac{1}{a} \cdot V\right) \end{aligned} \quad (1)$$

are equivalent for any  $a \in \mathbb{R}$  with  $a > 0$  where  $X \in \mathbb{R}^{n \times p}$  is a random matrix ( $\rightarrow$  I/1.2.4) and  $U \in \mathbb{R}^{n \times n}$  and  $V \in \mathbb{R}^{p \times p}$  are positive-definite matrices.

**Proof:** Since  $U$  and  $V$  must be positive-definite in the matrix-normal distribution ( $\rightarrow$  II/5.1.1), the scalar  $a$  must be larger than zero. A random matrix follows a matrix-normal distribution, if and only if its vectorization is multivariate normally distributed ( $\rightarrow$  II/5.1.2)

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) \quad (2)$$

where  $\text{vec}(X)$  is the vectorization operator and  $\otimes$  is the Kronecker product. Thus, the second distribution in (1) is equivalent to

$$\begin{aligned} X \sim \mathcal{MN}\left(M, a \cdot U, \frac{1}{a} \cdot V\right) &\Leftrightarrow \text{vec}(X) \sim \mathcal{N}\left(\text{vec}(M), \frac{1}{a} V \otimes aU\right) \\ &\sim \mathcal{N}\left(\text{vec}(M), \frac{1}{a} (V \otimes aU)\right) \\ &\sim \mathcal{N}\left(\text{vec}(M), \frac{a}{a} (V \otimes U)\right) \\ &\sim \mathcal{N}(\text{vec}(M), V \otimes U) \end{aligned} \quad (3)$$

which proves the equivalence of the two distributions in (1).

■

#### Sources:

- Glanz, Hunter; Carvalho, Luis (2013): “An Expectation-Maximization Algorithm for the Matrix Normal Distribution”; in: *arXiv stat.ME*, sect. 2.1, p. 3; URL: <https://arxiv.org/abs/1309.6609>; DOI: 10.48550/arXiv.1309.6609.



### 5.1.12 Redundancy of parameters

**Theorem:** The covariance parameters of the matrix-normal distribution ( $\rightarrow$  II/5.1.1) are redundant up to a scalar factor, i.e. the two probability distributions ( $\rightarrow$  I/1.5.1)

$$\begin{aligned} X &\sim \mathcal{MN}(M, U, V) \\ X &\sim \mathcal{MN}\left(M, a \cdot U, \frac{1}{a} \cdot V\right) \end{aligned} \quad (1)$$

are equivalent for any  $a \in \mathbb{R}$  with  $a > 0$  where  $X \in \mathbb{R}^{n \times p}$  is a random matrix ( $\rightarrow$  I/1.2.4) and  $U \in \mathbb{R}^{n \times n}$  and  $V \in \mathbb{R}^{p \times p}$  are positive-definite matrices.

**Proof:** The probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3) is

$$\begin{aligned} X &\sim \mathcal{MN}(M, U, V) \\ \Rightarrow f(X) &= \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right]. \end{aligned} \quad (2)$$

Using the inverse matrix property  $(cA)^{-1} = (1/c)A^{-1}$  and the matrix determinant property  $|cA| = c^n |A|$ , the probability density function ( $\rightarrow$  I/1.7.1) of the second distribution in (1) becomes

$$\begin{aligned} p(X) &= \frac{1}{\sqrt{(2\pi)^{np} \left|\frac{1}{a}V\right|^n |aU|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( \left(\frac{1}{a}V\right)^{-1} (X - M)^T (aU)^{-1} (X - M) \right) \right] \\ &= \frac{1}{\sqrt{(2\pi)^{np} \left(\left(\frac{1}{a}\right)^p |V|\right)^n (a^n |U|)^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( \frac{a}{a} V^{-1} (X - M)^T U^{-1} (X - M) \right) \right] \\ &= \frac{1}{\sqrt{(2\pi)^{np} \left(\frac{1}{a}\right)^{np} a^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right] \\ &= \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1} (X - M)^T U^{-1} (X - M)) \right] \end{aligned} \quad (3)$$

which is equal to the probability density function ( $\rightarrow$  I/1.7.1) of the first distribution in (1). ■

#### Sources:

- Glanz, Hunter; Carvalho, Luis (2013): “An Expectation-Maximization Algorithm for the Matrix Normal Distribution”; in: *arXiv stat.ME*, sect. 2.1, p. 3; URL: <https://arxiv.org/abs/1309.6609>; DOI: 10.48550/arXiv.1309.6609.

### 5.1.13 Drawing samples

**Theorem:** Let  $X \in \mathbb{R}^{n \times p}$  be a random matrix ( $\rightarrow$  I/1.2.4) with all entries independently following a standard normal distribution ( $\rightarrow$  II/3.2.3). Moreover, let  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{p \times p}$ , such that  $AA^T = U$  and  $B^T B = V$ .

Then,  $Y = M + AXB$  follows a matrix-normal distribution ( $\rightarrow$  II/5.1.1) with mean ( $\rightarrow$  I/1.12.19)  $M$ , covariance ( $\rightarrow$  I/1.16.1) across rows  $U$  and covariance ( $\rightarrow$  I/1.16.1) across columns  $V$ :

$$Y = M + AXB \sim \mathcal{MN}(M, U, V) . \quad (1)$$

**Proof:** If all entries of  $X$  are independent and standard normally distributed ( $\rightarrow$  II/3.2.3)

$$x_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for all } i = 1, \dots, n \quad \text{and } j = 1, \dots, p , \quad (2)$$

this implies a multivariate normal distribution with diagonal covariance matrix ( $\rightarrow$  II/4.1.17)

$$\begin{aligned} \text{vec}(X) &\sim \mathcal{N}(\text{vec}(0_{np}), I_{np}) \\ &\sim \mathcal{N}(\text{vec}(0_{np}), I_p \otimes I_n) \end{aligned} \quad (3)$$

where  $0_{np}$  is an  $n \times p$  matrix of zeros and  $I_n$  is the  $n \times n$  identity matrix.

Due to the relationship between multivariate and matrix-normal distribution ( $\rightarrow$  II/5.1.2), we have:

$$X \sim \mathcal{MN}(0_{np}, I_n, I_p) . \quad (4)$$

Thus, with the linear transformation theorem for the matrix-normal distribution ( $\rightarrow$  II/5.1.9), it follows that

$$\begin{aligned} Y = M + AXB &\sim \mathcal{MN}(M + A0_{np}B, AI_nA^T, B^T I_p B) \\ &\sim \mathcal{MN}(M, AA^T, B^T B) \\ &\sim \mathcal{MN}(M, U, V) . \end{aligned} \quad (5)$$

Thus, given  $X$  defined by (2),  $Y$  defined by (1) is a sample ( $\rightarrow$  I/1.1.6) from  $\mathcal{MN}(M, U, V)$ . ■

#### Sources:

- Wikipedia (2021): “Matrix normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-07; URL: [https://en.wikipedia.org/wiki/Matrix\\_normal\\_distribution#Drawing\\_values\\_from\\_the\\_distribution](https://en.wikipedia.org/wiki/Matrix_normal_distribution#Drawing_values_from_the_distribution).

## 5.2 Wishart distribution

### 5.2.1 Definition

**Definition:** Let  $X$  be an  $n \times p$  matrix following a matrix-normal distribution ( $\rightarrow$  II/5.1.1) with mean zero, independence across rows and covariance across columns  $V$ :

$$X \sim \mathcal{MN}(0, I_n, V) . \quad (1)$$

Define the scatter matrix  $S$  as the product of the transpose of  $X$  with itself:

$$S = X^T X = \sum_{i=1}^n x_i^T x_i . \quad (2)$$

Then, the matrix  $S$  is said to follow a Wishart distribution with scale matrix  $V$  and degrees of freedom  $n$

$$S \sim \mathcal{W}(V, n) \quad (3)$$

where  $n > p - 1$  and  $V$  is a positive definite symmetric covariance matrix.

#### Sources:

- Wikipedia (2020): “Wishart distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Wishart\\_distribution#Definition](https://en.wikipedia.org/wiki/Wishart_distribution#Definition).

### 5.2.2 Kullback-Leibler divergence

**Theorem:** Let  $S$  be a  $p \times p$  random matrix ( $\rightarrow$  I/1.2.4). Assume two Wishart distributions ( $\rightarrow$  II/5.2.1)  $P$  and  $Q$  specifying the probability distribution of  $S$  as

$$\begin{aligned} P : S &\sim \mathcal{W}(V_1, n_1) \\ Q : S &\sim \mathcal{W}(V_2, n_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of  $P$  from  $Q$  is given by

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[ n_2 (\ln |V_2| - \ln |V_1|) + n_1 \text{tr}(V_2^{-1} V_1) + 2 \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} + (n_1 - n_2) \psi_p\left(\frac{n_1}{2}\right) - n_1 p \right] \quad (2)$$

where  $\Gamma_p(x)$  is the multivariate gamma function

$$\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^k \Gamma\left(x - \frac{j-1}{2}\right) \quad (3)$$

and  $\psi_p(x)$  is the multivariate digamma function

$$\psi_p(x) = \frac{d \ln \Gamma_p(x)}{dx} = \sum_{j=1}^k \psi\left(x - \frac{j-1}{2}\right) . \quad (4)$$

**Proof:** The KL divergence for a continuous random variable ( $\rightarrow$  I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (5)$$

which, applied to the Wishart distributions ( $\rightarrow$  II/5.2.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{\mathcal{S}^p} \mathcal{W}(S; V_1, n_1) \ln \frac{\mathcal{W}(S; V_1, n_1)}{\mathcal{W}(S; V_2, n_2)} dS \\ &= \left\langle \ln \frac{\mathcal{W}(S; \alpha_1)}{\mathcal{W}(S; \alpha_2)} \right\rangle_{p(S)} \end{aligned} \quad (6)$$

where  $\mathcal{S}^p$  is the set of all positive-definite symmetric  $p \times p$  matrices.

Using the probability density function of the Wishart distribution, this becomes:

$$\begin{aligned}
 \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2^{n_1 p} |V_1|^{n_1} \Gamma_p(\frac{n_1}{2})}} \cdot |S|^{(n_1 - p - 1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr}(V_1^{-1} S) \right]}{\frac{1}{\sqrt{2^{n_2 p} |V_2|^{n_2} \Gamma_p(\frac{n_2}{2})}} \cdot |S|^{(n_2 - p - 1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr}(V_2^{-1} S) \right]} \right\rangle_{p(S)} \\
 &= \left\langle \ln \left( \sqrt{2^{(n_2 - n_1)p}} \cdot \frac{|V_2|^{n_2}}{|V_1|^{n_1}} \cdot \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} \cdot |S|^{(n_1 - n_2)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr}(V_1^{-1} S) + \frac{1}{2} \text{tr}(V_2^{-1} S) \right] \right) \right\rangle_{p(S)} \\
 &= \left\langle \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} \right. \\
 &\quad \left. + \frac{n_1 - n_2}{2} \ln |S| - \frac{1}{2} \text{tr}(V_1^{-1} S) + \frac{1}{2} \text{tr}(V_2^{-1} S) \right\rangle_{p(S)} \\
 &= \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} \\
 &\quad + \frac{n_1 - n_2}{2} \langle \ln |S| \rangle_{p(S)} - \frac{1}{2} \langle \text{tr}(V_1^{-1} S) \rangle_{p(S)} + \frac{1}{2} \langle \text{tr}(V_2^{-1} S) \rangle_{p(S)} .
 \end{aligned} \tag{7}$$

Using the expected value of a Wishart random matrix

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle S \rangle = nV , \tag{8}$$

such that the expected value of the matrix trace ( $\rightarrow$  I/1.12.20) becomes

$$\langle \text{tr}(AS) \rangle = \text{tr}(\langle AS \rangle) = \text{tr}(A \langle S \rangle) = \text{tr}(A \cdot (nV)) = n \cdot \text{tr}(AV) , \tag{9}$$

and the expected value of a Wishart log-determinant

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle \ln |S| \rangle = \psi_p \left( \frac{n}{2} \right) + p \cdot \ln 2 + \ln |V| , \tag{10}$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{aligned}
 \text{KL}[P \parallel Q] &= \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} \\
 &\quad + \frac{n_1 - n_2}{2} \left[ \psi_p \left( \frac{n_1}{2} \right) + p \cdot \ln 2 + \ln |V_1| \right] - \frac{n_1}{2} \text{tr}(V_1^{-1} V_1) + \frac{n_1}{2} \text{tr}(V_2^{-1} V_1) \\
 &= \frac{n_2}{2} (\ln |V_2| - \ln |V_1|) + \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} + \frac{n_1 - n_2}{2} \psi_p \left( \frac{n_1}{2} \right) - \frac{n_1}{2} \text{tr}(I_p) + \frac{n_1}{2} \text{tr}(V_2^{-1} V_1) \\
 &= \frac{1}{2} \left[ n_2 (\ln |V_2| - \ln |V_1|) + n_1 \text{tr}(V_2^{-1} V_1) + 2 \ln \frac{\Gamma_p(\frac{n_2}{2})}{\Gamma_p(\frac{n_1}{2})} + (n_1 - n_2) \psi_p \left( \frac{n_1}{2} \right) - n_1 p \right] .
 \end{aligned} \tag{11}$$

■

**Sources:**

- Penny, William D. (2001): “KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities”; in: *University College, London*, pp. 2-3, eqs. 13/15; URL: <https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps>.
- Wikipedia (2021): “Wishart distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-02; URL: [https://en.wikipedia.org/wiki/Wishart\\_distribution#KL-divergence](https://en.wikipedia.org/wiki/Wishart_distribution#KL-divergence).

## 5.3 Normal-Wishart distribution

### 5.3.1 Definition

**Definition:** Let  $X$  be an  $n \times p$  random matrix ( $\rightarrow$  I/1.2.4) and let  $Y$  be a  $p \times p$  positive-definite symmetric matrix. Then,  $X$  and  $Y$  are said to follow a normal-Wishart distribution

$$X, Y \sim \text{NW}(M, U, V, \nu), \quad (1)$$

if the distribution of  $X$  conditional on  $Y$  is a matrix-normal distribution ( $\rightarrow$  II/5.1.1) with mean  $M$ , covariance across rows  $U$ , covariance across columns  $Y^{-1}$  and  $Y$  follows a Wishart distribution ( $\rightarrow$  II/5.2.1) with scale matrix  $V$  and degrees of freedom  $\nu$ :

$$\begin{aligned} X|Y &\sim \mathcal{MN}(M, U, Y^{-1}) \\ Y &\sim \mathcal{W}(V, \nu). \end{aligned} \quad (2)$$

The  $p \times p$  matrix  $Y$  can be seen as the precision matrix ( $\rightarrow$  I/1.16.12) across the columns of the  $n \times p$  matrix  $X$ .

#### Sources:

- Bishop, Christopher M. (2006): “Appendix B. Probability Distributions”; in: *Pattern Recognition for Machine Learning*, p. 690, eq. B.53; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 5.3.2 Probability density function

**Theorem:** Let  $X$  and  $Y$  follow a normal-Wishart distribution ( $\rightarrow$  II/5.3.1):

$$X, Y \sim \text{NW}(M, U, V, \nu). \quad (1)$$

Then, the joint probability ( $\rightarrow$  I/1.3.2) density function ( $\rightarrow$  I/1.7.1) of  $X$  and  $Y$  is

$$\begin{aligned} p(X, Y) &= \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \\ &\quad \exp \left[ -\frac{1}{2} \text{tr} \left( Y \left[ (X - M)^T U^{-1} (X - M) + V^{-1} \right] \right) \right]. \end{aligned} \quad (2)$$

**Proof:** The normal-Wishart distribution ( $\rightarrow$  II/5.3.1) is defined as  $X$  conditional on  $Y$  following a matrix-normal distribution ( $\rightarrow$  II/5.1.1) and  $Y$  following a Wishart distribution ( $\rightarrow$  II/5.2.1):

$$\begin{aligned} X|Y &\sim \mathcal{MN}(M, U, Y^{-1}) \\ Y &\sim \mathcal{W}(V, \nu) . \end{aligned} \quad (3)$$

Thus, using the probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3) and the probability density function of the Wishart distribution, we have the following probabilities:

$$\begin{aligned} p(X|Y) &= \mathcal{MN}(X; M, U, Y^{-1}) \\ &= \sqrt{\frac{|Y|^n}{(2\pi)^{np}|U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (Y(X - M)^T U^{-1}(X - M)) \right] \\ p(Y) &= \mathcal{W}(Y; V, \nu) \\ &= \frac{1}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{2^{\nu p}|V|^\nu}} \cdot |Y|^{(\nu-p-1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1}Y) \right] . \end{aligned} \quad (4)$$

The law of conditional probability ( $\rightarrow$  I/1.3.4) implies that

$$p(X, Y) = p(X|Y) p(Y) , \quad (5)$$

such that the normal-Wishart density function becomes:

$$\begin{aligned} p(X, Y) &= \mathcal{MN}(X; M, U, Y^{-1}) \cdot \mathcal{W}(Y; V, \nu) \\ &= \sqrt{\frac{|Y|^n}{(2\pi)^{np}|U|^p}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (Y(X - M)^T U^{-1}(X - M)) \right] \cdot \\ &\quad \frac{1}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{2^{\nu p}|V|^\nu}} \cdot |Y|^{(\nu-p-1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (V^{-1}Y) \right] \\ &= \frac{1}{\sqrt{(2\pi)^{np}|U|^p|V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \cdot \\ &\quad \exp \left[ -\frac{1}{2} \text{tr} (Y [(X - M)^T U^{-1}(X - M) + V^{-1}]) \right] . \end{aligned} \quad (6)$$

■

#### Sources:

- Bishop, Christopher M. (2006): “Appendix B. Probability Distributions”; in: *Pattern Recognition for Machine Learning*, p. 690, eq. B.53; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 5.3.3 Mean

**Theorem:** Let  $X \in \mathbb{R}^{n \times p}$  and  $Y \in \mathbb{R}^{p \times p}$  follow a normal-Wishart distribution ( $\rightarrow$  II/5.3.1):

$$X, Y \sim \text{NW}(M, U, V, \nu) . \quad (1)$$

Then, the expected value ( $\rightarrow$  I/1.12.1) of  $X$  and  $Y$  is

$$E[(X, Y)] = (M, \nu V) . \quad (2)$$

**Proof:** Consider the random matrix ( $\rightarrow$  I/1.2.4)

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \\ y_{11} & \dots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{p1} & \dots & y_{pp} \end{bmatrix} . \quad (3)$$

According to the expected value of a random matrix ( $\rightarrow$  I/1.12.19), its expected value is

$$E\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} E(x_{11}) & \dots & E(x_{1p}) \\ \vdots & \ddots & \vdots \\ E(x_{n1}) & \dots & E(x_{np}) \\ E(y_{11}) & \dots & E(y_{1p}) \\ \vdots & \ddots & \vdots \\ E(y_{p1}) & \dots & E(y_{pp}) \end{bmatrix} = \begin{bmatrix} E(X) \\ E(Y) \end{bmatrix} . \quad (4)$$

When  $X$  and  $Y$  are jointly normal-Wishart distributed, then ( $\rightarrow$  II/5.3.1) by definition  $X$  follows a matrix-normal distribution ( $\rightarrow$  II/5.1.1) conditional on  $Y$  and  $Y$  follows a Wishart distribution ( $\rightarrow$  II/5.2.1):

$$X, Y \sim \text{NW}(M, U, V, \nu) \Leftrightarrow X|Y \sim \mathcal{MN}(M, U, Y^{-1}) \quad \wedge \quad Y \sim \mathcal{W}(V, \nu) . \quad (5)$$

Thus, with the expected value of the matrix-variate normal distribution ( $\rightarrow$  II/5.1.4) and the law of conditional probability ( $\rightarrow$  I/1.3.4),  $E(X)$  becomes

$$\begin{aligned} E(X) &= \iint X \cdot p(X, Y) \, dX \, dY \\ &= \iint X \cdot p(X|Y) \cdot p(Y) \, dX \, dY \\ &= \int p(Y) \int X \cdot p(X|Y) \, dX \, dY \\ &= \int p(Y) \langle X \rangle_{\mathcal{MN}(M, U, Y^{-1})} \, dY \\ &= \int p(Y) \cdot M \, dY \\ &= M \int p(Y) \, dY \\ &= M , \end{aligned} \quad (6)$$

and with the expected value of the Wishart distribution,  $E(Y)$  becomes

$$\begin{aligned} E(Y) &= \int Y \cdot p(Y) \, dY \\ &= \langle Y \rangle_{\mathcal{W}(V, \nu)} \\ &= \nu V . \end{aligned} \tag{7}$$

Thus, the expectation of the random matrix in equations (3) and (4) is

$$E \left( \begin{bmatrix} X \\ Y \end{bmatrix} \right) = \begin{bmatrix} M \\ \nu V \end{bmatrix} , \tag{8}$$

as indicated by equation (2).

■



## Chapter III

# Statistical Models

# 1 Univariate normal data

## 1.1 Univariate Gaussian

### 1.1.1 Definition

**Definition:** A univariate Gaussian data set is given by a set of real numbers  $y = \{y_1, \dots, y_n\}$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a normal distribution ( $\rightarrow$  II/3.2.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ :

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

#### Sources:

- Bishop, Christopher M. (2006): “Example: The univariate Gaussian”; in: *Pattern Recognition for Machine Learning*, ch. 10.1.3, p. 470, eq. 10.21; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 1.1.2 Maximum likelihood estimation

**Theorem:** Let there be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1)  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) for mean  $\mu$  and variance  $\sigma^2$  are given by

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (2)$$

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for each observation is given by the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10)

$$p(y_i | \mu, \sigma^2) = \mathcal{N}(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \quad (3)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y | \mu, \sigma^2) = \prod_{i=1}^n p(y_i | \mu) = \sqrt{\frac{1}{(2\pi\sigma^2)^n}} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n \left( \frac{y_i - \mu}{\sigma} \right)^2 \right]. \quad (4)$$

This can be developed into

$$\begin{aligned}
p(y|\mu, \sigma^2) &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n \left( \frac{y_i^2 - 2y_i\mu + \mu^2}{\sigma^2} \right) \right] \\
&= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} (y^T y - 2n\bar{y}\mu + n\mu^2) \right]
\end{aligned} \tag{5}$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  is the mean of data points and  $y^T y = \sum_{i=1}^n y_i^2$  is the sum of squared data points. Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$\text{LL}(\mu, \sigma^2) = \log p(y|\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y^T y - 2n\bar{y}\mu + n\mu^2) . \tag{6}$$

The derivative of the log-likelihood function (6) with respect to  $\mu$  is

$$\frac{d\text{LL}(\mu, \sigma^2)}{d\mu} = \frac{n\bar{y}}{\sigma^2} - \frac{n\mu}{\sigma^2} = \frac{n}{\sigma^2} (\bar{y} - \mu) \tag{7}$$

and setting this derivative to zero gives the MLE for  $\mu$ :

$$\begin{aligned}
\frac{d\text{LL}(\hat{\mu}, \sigma^2)}{d\mu} &= 0 \\
0 &= \frac{n}{\sigma^2} (\bar{y} - \hat{\mu}) \\
0 &= \bar{y} - \hat{\mu} \\
\hat{\mu} &= \bar{y} \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i .
\end{aligned} \tag{8}$$

The derivative of the log-likelihood function (6) at  $\hat{\mu}$  with respect to  $\sigma^2$  is

$$\begin{aligned}
\frac{d\text{LL}(\hat{\mu}, \sigma^2)}{d\sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y^T y - 2n\bar{y}\hat{\mu} + n\hat{\mu}^2) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i^2 - 2y_i\hat{\mu} + \hat{\mu}^2) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \hat{\mu})^2
\end{aligned} \tag{9}$$

and setting this derivative to zero gives the MLE for  $\sigma^2$ :

$$\begin{aligned}
\frac{dLL(\hat{\mu}, \hat{\sigma}^2)}{d\sigma^2} &= 0 \\
0 &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\mu})^2 \\
\frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\mu})^2 \\
\frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{n}{2\hat{\sigma}^2} &= \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\mu})^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2
\end{aligned} \tag{10}$$

Together, (8) and (10) constitute the MLE for the univariate Gaussian. ■

#### Sources:

- Bishop CM (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, pp. 93-94, eqs. 2.121, 2.122; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 1.1.3 One-sample t-test

**Theorem:** Let

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \tag{2}$$

with sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  and sample variance ( $\rightarrow$  I/1.13.2)  $s^2$  follows a Student's t-distribution ( $\rightarrow$  II/3.3.1) with  $n - 1$  degrees of freedom ( $\rightarrow$  I/1.5.12)

$$t \sim t(n - 1) \tag{3}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu = \mu_0 . \tag{4}$$

**Proof:** The sample mean ( $\rightarrow$  I/1.12.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (5)$$

and the sample variance ( $\rightarrow$  I/1.13.2) is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 . \quad (6)$$

Using the linear combination formula for normal random variables ( $\rightarrow$  II/3.2.26), the sample mean follows a normal distribution ( $\rightarrow$  II/3.2.1) with the following parameters:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \sim \mathcal{N} \left( \frac{1}{n} n\mu, \left( \frac{1}{n} \right)^2 n\sigma^2 \right) = \mathcal{N} (\mu, \sigma^2/n) . \quad (7)$$

Again employing the linear combination theorem and applying the null hypothesis from (4), the distribution of  $Z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$  becomes standard normal ( $\rightarrow$  II/3.2.3)

$$Z = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma} \sim \mathcal{N} \left( \frac{\sqrt{n}}{\sigma}(\mu - \mu_0), \left( \frac{\sqrt{n}}{\sigma} \right)^2 \frac{\sigma^2}{n} \right) \stackrel{(4)}{=} \mathcal{N} (0, 1) . \quad (8)$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution ( $\rightarrow$  II/3.2.7), the distribution of  $V = (n-1) s^2/\sigma^2$  is

$$V = \frac{(n-1) s^2}{\sigma^2} \sim \chi^2 (n-1) . \quad (9)$$

Finally, since the ratio of a standard normal random variable and the square root of a chi-squared random variable follows a t-distribution ( $\rightarrow$  II/3.3.1), the distribution of the test statistic ( $\rightarrow$  I/4.3.5) is given by

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{Z}{\sqrt{V/(n-1)}} \sim t(n-1) . \quad (10)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) can be rejected when  $t$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the Student's t-distribution ( $\rightarrow$  II/3.3.1) with  $n-1$  degrees of freedom ( $\rightarrow$  I/1.5.12) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ .

■

#### Sources:

- Wikipedia (2021): “Student's t-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Derivation](https://en.wikipedia.org/wiki/Student%27s_t-distribution#Derivation).

#### 1.1.4 Two-sample t-test

**Theorem:** Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma^2), & i = 1, \dots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma^2), & i = 1, \dots, n_2 \end{aligned} \quad (1)$$

be two univariate Gaussian data sets ( $\rightarrow$  III/1.1.1) representing two groups of unequal size  $n_1$  and  $n_2$  with unknown means  $\mu_1$  and  $\mu_2$  and equal unknown variance  $\sigma^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

with sample means ( $\rightarrow$  I/1.12.2)  $\bar{y}_1$  and  $\bar{y}_2$  and pooled standard deviation ( $\rightarrow$  I/1.19.3)  $s_p$  follows a Student's t-distribution ( $\rightarrow$  II/3.3.1) with  $n_1 + n_2 - 2$  degrees of freedom ( $\rightarrow$  I/1.5.12)

$$t \sim t(n_1 + n_2 - 2) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu_1 - \mu_2 = \mu_\Delta . \quad (4)$$

**Proof:** The sample means ( $\rightarrow$  I/1.12.2) are given by

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \end{aligned} \quad (5)$$

and the pooled standard deviation ( $\rightarrow$  I/1.19.3) is given by

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (6)$$

with the sample variances ( $\rightarrow$  I/1.13.2)

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 \\ s_2^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 . \end{aligned} \quad (7)$$

Using the linear combination formula for normal random variables ( $\rightarrow$  II/3.2.26), the sample means follows normal distributions ( $\rightarrow$  II/3.2.1) with the following parameters:

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N} \left( \frac{1}{n_1} n_1 \mu_1, \left( \frac{1}{n_1} \right)^2 n_1 \sigma^2 \right) = \mathcal{N} (\mu_1, \sigma^2/n_1) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N} \left( \frac{1}{n_2} n_2 \mu_2, \left( \frac{1}{n_2} \right)^2 n_2 \sigma^2 \right) = \mathcal{N} (\mu_2, \sigma^2/n_2) . \end{aligned} \quad (8)$$

Again employing the linear combination theorem and applying the null hypothesis from (4), the distribution of  $Z = ((\bar{y}_1 - \bar{y}_2) - \mu_\Delta) / (\sigma \sqrt{1/n_1 + 1/n_2})$  becomes standard normal ( $\rightarrow$  II/3.2.3)

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N} \left( \frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \left( \frac{1}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2 \left( \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right) \right) \stackrel{(4)}{=} \mathcal{N}(0, 1) . \quad (9)$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution ( $\rightarrow$  II/3.2.7), the distribution of  $V = (n_1 + n_2 - 2) s_p^2 / \sigma^2$  is

$$V = \frac{(n_1 + n_2 - 2) s_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) . \quad (10)$$

Finally, since the ratio of a standard normal random variable and the square root of a chi-squared random variable follows a t-distribution ( $\rightarrow$  II/3.3.1), the distribution of the test statistic ( $\rightarrow$  I/4.3.5) is given by

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} \sim t(n_1 + n_2 - 2) . \quad (11)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) can be rejected when  $t$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the Student's t-distribution ( $\rightarrow$  II/3.3.1) with  $n_1 + n_2 - 2$  degrees of freedom ( $\rightarrow$  I/1.5.12) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ .

■

#### Sources:

- Wikipedia (2021): “Student’s t-distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Derivation](https://en.wikipedia.org/wiki/Student%27s_t-distribution#Derivation).
- Wikipedia (2021): “Student’s t-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-test#Equal\\_or\\_unequal\\_sample\\_sizes,\\_similar\\_variances\\_\(1/2\\_%3C\\_sX1/sX2\\_%3C\\_2\)](https://en.wikipedia.org/wiki/Student%27s_t-test#Equal_or_unequal_sample_sizes,_similar_variances_(1/2_%3C_sX1/sX2_%3C_2)).

#### 1.1.5 Paired t-test

**Theorem:** Let  $y_{i1}$  and  $y_{i2}$  with  $i = 1, \dots, n$  be paired observations, such that

$$y_{i1} \sim \mathcal{N}(y_{i2} + \mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

is a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown shift  $\mu$  and unknown variance  $\sigma^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \quad \text{where} \quad d_i = y_{i1} - y_{i2} \quad (2)$$

with sample mean ( $\rightarrow$  I/1.12.2)  $\bar{d}$  and sample variance ( $\rightarrow$  I/1.13.2)  $s_d^2$  follows a Student’s t-distribution ( $\rightarrow$  II/3.3.1) with  $n - 1$  degrees of freedom ( $\rightarrow$  I/1.5.12)

$$t \sim t(n - 1) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu = \mu_0 . \quad (4)$$

**Proof:** Define the pair-wise difference  $d_i = y_{i1} - y_{i2}$  which, as a linear transformation of a normal random variable ( $\rightarrow$  II/4.1.14), is distributed as

$$d_i = y_{i1} - y_{i2} \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (5)$$

Therefore,  $d_1, \dots, d_n$  satisfy the conditions of the one-sample t-test ( $\rightarrow$  III/1.1.3) which results in the test statistic given by (2). ■

#### Sources:

- Wikipedia (2021): “Student’s t-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-test#Dependent\\_t-test\\_for\\_paired\\_samples](https://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples).

### 1.1.6 F-test for equality of variances

**Theorem:** Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma_1^2), & i = 1, \dots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma_2^2), & i = 1, \dots, n_2 \end{aligned} \quad (1)$$

be two univariate Gaussian data sets ( $\rightarrow$  III/1.1.1) representing two groups of unequal size  $n_1$  and  $n_2$  with unknown means  $\mu_1$  and  $\mu_2$  and unknown variances  $\sigma_1^2$  and  $\sigma_2^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F = \frac{s_1^2}{s_2^2} = \frac{\frac{1}{n_1-1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2}{\frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2} \quad (2)$$

with sample means ( $\rightarrow$  I/1.12.2)  $\bar{y}_1$  and  $\bar{y}_2$  and sample variances ( $\rightarrow$  I/1.13.2)  $s_1^2$  and  $s_2^2$  follows an F-distribution ( $\rightarrow$  II/3.8.1) with numerator degrees of freedom  $n_1 - 1$  and denominator degrees of freedom  $n_2 - 1$

$$F \sim F(n_1 - 1, n_2 - 1) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the two variances ( $\rightarrow$  II/3.2.1) are equal:

$$H_0 : \sigma_1^2 = \sigma_2^2. \quad (4)$$

**Proof:** We know that, for a sample of normal random variables, the sample variance is following a chi-squared distribution ( $\rightarrow$  II/3.2.7):

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad \Rightarrow \quad V = (n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1). \quad (5)$$

Thus, we have:

$$\begin{aligned} V_1 &= (n_1 - 1) \frac{s_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{and} \\ V_2 &= (n_2 - 1) \frac{s_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1). \end{aligned} \quad (6)$$



Moreover, by definition, the ratio of two chi-squared random variables, divided by their degrees of freedom, is following an F-distribution ( $\rightarrow$  II/3.8.1):

$$X_1 \sim \chi^2(d_1), X_2 \sim \chi^2(d_2) \quad \Rightarrow \quad Y = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2) . \quad (7)$$

Thus, we have:

$$\begin{aligned} F &= \frac{V_1/(n_1 - 1)}{V_2/(n_2 - 1)} \\ &= \frac{(n_1 - 1) \frac{s_1^2}{\sigma_1^2} / (n_1 - 1)}{(n_2 - 1) \frac{s_2^2}{\sigma_2^2} / (n_2 - 1)} \\ &= \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \\ &\stackrel{(4)}{=} \frac{s_1^2}{s_2^2} . \end{aligned} \quad (8)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) of equal variances can be rejected when  $F$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the F-distribution ( $\rightarrow$  II/??) with degrees of freedom ( $\rightarrow$  I/1.5.12)  $n_1 - 1$  and  $n_2 - 1$  using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ . ■

#### Sources:

- Wikipedia (2024): “F-test of equality of variances”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-07-05; URL: [https://en.wikipedia.org/wiki/F-test\\_of\\_equality\\_of\\_variances#The\\_test](https://en.wikipedia.org/wiki/F-test_of_equality_of_variances#The_test).

### 1.1.7 Power analysis for one-sample t-test

**Theorem:** Let there be  $n$  i.i.d. ( $\rightarrow$  I/1.2.8) observations  $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$  with both  $\mu$  and  $\sigma^2$  unknown. Further, let  $z_x = \Phi^{-1}(1 - x)$ , where  $\Phi$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of a standard normal distribution ( $\rightarrow$  II/3.2.3). Equivalently, with  $Z \sim \mathcal{N}(0, 1)$ , let

$$z_x = \Phi^{-1}(1 - x) \quad \Leftrightarrow \quad \Phi(z_x) = 1 - x \quad \Leftrightarrow \quad \Pr(Z \leq z_x) = 1 - x \Leftrightarrow \Pr(z_x \leq Z) = x . \quad (1)$$

1) One-sided one-sample t-test:

We perform a one-sided ( $\rightarrow$  I/4.3.4) one-sample t-test ( $\rightarrow$  III/1.1.3) of the null hypothesis ( $\rightarrow$  I/4.3.2)  $H_0 : \mu \leq \mu_0$  vs.  $H_1 : \mu > \mu_0$ . The test has significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ . If  $H_1$  is true and  $\mu = \mu_0 + \delta$ , then the power ( $\rightarrow$  I/4.3.7) of the test is

$$\Phi \left( \frac{\delta \sqrt{n}}{\sigma} - z_\alpha \right) . \quad (2)$$

If we keep  $\alpha, n, \sigma^2$  the same as before and require that the power is at least  $1 - \beta$ , then the minimum detectable effect ( $\rightarrow$  I/4.3.12) is

$$\frac{(z_\alpha + z_\beta) \sigma}{\sqrt{n}} . \quad (3)$$

If we keep  $\alpha$ ,  $\sigma^2$ ,  $\delta$ ,  $\beta$  the same as before, then the minimum required sample size ( $\rightarrow$  I/4.3.13) is

$$\frac{\sigma^2(z_\alpha + z_\beta)^2}{\delta^2}. \quad (4)$$

2) Two-sided one-sample t-test:

We perform a two-sided ( $\rightarrow$  I/4.3.4) one-sample t-test ( $\rightarrow$  III/1.1.3) of hypothesis  $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$ . The test has size  $\alpha$ . If  $H_1$  is true and  $\mu = \mu_0 + \delta$  (assuming  $\delta > 0$  without loss of generality), then the power ( $\rightarrow$  I/4.3.7) of the test is

$$\Phi\left(\frac{\delta\sqrt{n}}{\sigma} - z_{\alpha/2}\right) + \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \approx \Phi\left(\frac{\delta\sqrt{n}}{\sigma} - z_{\alpha/2}\right). \quad (5)$$

If we keep  $\alpha$ ,  $n$ ,  $\sigma^2$  the same as before, and require that the power is at least  $1 - \beta$ , then the minimum detectable effect ( $\rightarrow$  I/4.3.12) is

$$\frac{(z_{\alpha/2} + z_\beta)\sigma}{\sqrt{n}}. \quad (6)$$

If we keep  $\alpha$ ,  $\sigma^2$ ,  $\delta$ ,  $\beta$  the same as before then the minimum required sample size ( $\rightarrow$  I/4.3.12) is

$$\frac{\sigma^2(z_{\alpha/2} + z_\beta)^2}{\delta^2}. \quad (7)$$

**Proof:** The  $t$ -statistic is

$$T = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}. \quad (8)$$

which has a  $t$ -distribution ( $\rightarrow$  III/1.1.3) with  $n - 1$  degrees of freedom ( $\rightarrow$  I/1.5.12). Unless  $n$  is very small, the  $t(n - 1)$  distribution is very close to a  $\mathcal{N}(0, 1)$  distribution and for this proof, we will assume that this is a good approximation.

1) One-sided one-sample t-test:

The test has size ( $\rightarrow$  I/4.3.6)  $\alpha$ , we reject the null, if

$$T > z_\alpha. \quad (9)$$

So the power ( $\rightarrow$  I/4.3.7) of the test is

$$\Pr\left(\frac{\bar{Y} - \mu_0}{s/\sqrt{n}} > z_\alpha\right). \quad (10)$$

If the true mean is  $\mu = \mu_0 + \delta$ , where  $\delta > 0$ , then the power is

$$\Pr\left(\frac{\bar{Y} - \mu_0 - \delta}{s/\sqrt{n}} + \frac{\delta\sqrt{n}}{s} > z_\alpha\right). \quad (11)$$

Now that we're subtracting the true mean, the first term in the sum has a  $t$ -distribution ( $\rightarrow$  II/3.3.1), which we approximate with a  $\mathcal{N}(0, 1)$  distribution, denoted as  $Z$ . We also approximate  $s$  with  $\sigma$ , since it should be a good approximation, and the error here doesn't make much difference to the calculation. The power is therefore approximately

$$\Pr\left(Z > z_\alpha - \frac{\delta\sqrt{n}}{\sigma}\right) = \Pr\left(Z < \frac{\delta\sqrt{n}}{\sigma} - z_\alpha\right) = \Phi\left(\frac{\delta\sqrt{n}}{\sigma} - z_\alpha\right). \quad (12)$$

To calculate the minimum detectable effect ( $\rightarrow$  I/4.3.12), we want the smallest  $\delta$ , such that the power is at least  $1 - \beta$ . If the power is at least  $1 - \beta$ , then

$$1 - \beta \leq \Phi \left( \frac{\delta\sqrt{n}}{\sigma} - z_{\alpha} \right) . \quad (13)$$

Using  $z_{\beta} = \Phi^{-1}(1 - \beta)$  and rearranging, we get

$$z_{\beta} + z_{\alpha} \leq \frac{\delta\sqrt{n}}{\sigma} . \quad (14)$$

Therefore,

$$\frac{(z_{\beta} + z_{\alpha})\sigma}{\sqrt{n}} \leq \delta , \quad (15)$$

and taking the minimum value for  $\delta$  gives the minimum detectable effect.

To calculate the minimum required sample size ( $\rightarrow$  I/4.3.13), we want the smallest  $n$ , such that the power is at least  $1 - \beta$ . Rearranging (13), we have

$$\frac{\sigma^2(z_{\alpha} + z_{\beta})^2}{\delta^2} \leq n . \quad (16)$$

2) Two-sided one-sample t-test:

The test has size ( $\rightarrow$  I/4.3.6)  $\alpha$ , we reject the null, if

$$|T| > z_{\alpha/2} . \quad (17)$$

So the power ( $\rightarrow$  I/4.3.7) of the test is

$$\Pr \left( \left( \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} > z_{\alpha/2} \right) \wedge \left( \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < -z_{\alpha/2} \right) \right) . \quad (18)$$

If the true mean is  $\mu = \mu_0 + \delta$  (where  $\delta > 0$  without loss of generality), then the power is

$$\Pr \left( \left( \frac{\bar{Y} - \mu_0 - \delta}{s/\sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma} > z_{\alpha/2} \right) \wedge \left( \frac{\bar{Y} - \mu_0 - \delta}{s/\sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma} < -z_{\alpha/2} \right) \right) . \quad (19)$$

As before, we use a  $\mathcal{N}(0, 1)$  distribution and we approximate  $s$  with  $\sigma$ . The power is therefore approximately

$$\Phi \left( \frac{\delta\sqrt{n}}{\sigma} - z_{\alpha/2} \right) + \Phi \left( -z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \right) . \quad (20)$$

If we have a large enough  $\delta$ , then the second term, which is the probability of rejecting the null hypothesis, but with a sign error on  $\delta$ , will be small. For the later calculations, we will set it to 0 and use the approximation

$$\Phi \left( \frac{\delta\sqrt{n}}{\sigma} - z_{\alpha/2} \right) \quad (21)$$

for the power. To calculate the minimum detectable effect ( $\rightarrow$  I/4.3.12), we want the smallest  $\delta$ , such that the power is at least  $1 - \beta$ . If the power is at least  $1 - \beta$ , then

$$1 - \beta \leq \Phi \left( \frac{\delta\sqrt{n}}{\sigma} - z_{\alpha/2} \right) . \quad (22)$$

Using  $z_\beta = \Phi^{-1}(1 - \beta)$  and rearranging, we get

$$z_\beta + z_{\alpha/2} \leq \frac{\delta\sqrt{n}}{\sigma}. \quad (23)$$

Therefore,

$$\frac{(z_\beta + z_{\alpha/2})\sigma}{\sqrt{n}} \leq \delta, \quad (24)$$

and taking the minimum value for  $\delta$  gives the minimum detectable effect.

To calculate the minimum required sample size ( $\rightarrow$  I/4.3.13), we want the smallest  $n$ , such that the power is at least  $1 - \beta$ . Rearranging (22), we have

$$\frac{\sigma^2(z_{\alpha/2} + z_\beta)^2}{\delta^2} \leq n. \quad (25)$$

■

#### Sources:

- Sylvain Chabé-Ferret (2025): “Power Analysis”; in: *Statistical Tools for Causal Inference*, retrieved on 2025-07-05; URL: <https://chabefer.github.io/STCI/Power.html>.

### 1.1.8 Conjugate prior distribution

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for this model is a normal-gamma distribution ( $\rightarrow$  II/4.4.1)

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (2)$$

where  $\tau = 1/\sigma^2$  is the inverse variance or precision.

**Proof:** By definition, a conjugate prior ( $\rightarrow$  I/5.2.5) is a prior distribution ( $\rightarrow$  I/5.1.3) that, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned} p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{(\sqrt{2\pi}\sigma^2)^n} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (3)$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}
p(y|\mu, \tau) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\
&= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{4}$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

Separating constant and variable terms, we have:

$$p(y|\mu, \tau) = \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]. \tag{5}$$

Expanding the product in the exponent, we have

$$\begin{aligned}
p(y|\mu, \tau) &= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2) \right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \left( \sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2 \right) \right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y^T y - 2\mu n\bar{y} + n\mu^2) \right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau n}{2} \left( \frac{1}{n} y^T y - 2\mu\bar{y} + \mu^2 \right) \right]
\end{aligned} \tag{6}$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  is the mean of data points and  $y^T y = \sum_{i=1}^n y_i^2$  is the sum of squared data points. Completing the square over  $\mu$ , finally gives

$$p(y|\mu, \tau) = \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau n}{2} \left( (\mu - \bar{y})^2 - \bar{y}^2 + \frac{1}{n} y^T y \right) \right] \tag{7}$$

In other words, the likelihood function ( $\rightarrow$  I/5.1.2) is proportional to a power of  $\tau$  times an exponential of  $\tau$  and an exponential of a squared form of  $\mu$ , weighted by  $\tau$ :

$$p(y|\mu, \tau) \propto \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y^T y - n\bar{y}^2) \right] \cdot \exp \left[ -\frac{\tau n}{2} (\mu - \bar{y})^2 \right]. \tag{8}$$

The same is true for a normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $\mu$  and  $\tau$

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \tag{9}$$

the probability density function of which ( $\rightarrow$  II/4.4.3)

$$p(\mu, \tau) = \sqrt{\frac{\tau \lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\tau \lambda_0}{2} (\mu - \mu_0)^2 \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \quad (10)$$

exhibits the same proportionality

$$p(\mu, \tau) \propto \tau^{a_0+1/2-1} \cdot \exp[-\tau b_0] \cdot \exp \left[ -\frac{\tau \lambda_0}{2} (\mu - \mu_0)^2 \right] \quad (11)$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Bishop CM (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, pp. 97-102, eq. 2.154; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 1.1.9 Posterior distribution

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.1.8) over the model parameters  $\mu$  and  $\tau = 1/\sigma^2$ :

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau \lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a normal-gamma distribution ( $\rightarrow$  II/4.4.1)

$$p(\mu, \tau | y) = \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) . \end{aligned} \quad (4)$$

**Proof:** According to Bayes’ theorem ( $\rightarrow$  I/5.3.2), the posterior distribution ( $\rightarrow$  I/5.1.9) is given by

$$p(\mu, \tau | y) = \frac{p(y | \mu, \tau) p(\mu, \tau)}{p(y)} . \quad (5)$$

Since  $p(y)$  is just a normalization factor, the posterior is proportional ( $\rightarrow$  I/5.1.10) to the numerator:

$$p(\mu, \tau | y) \propto p(y | \mu, \tau) p(\mu, \tau) = p(y, \mu, \tau) . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned}
 p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\
 &= \frac{1}{(\sqrt{2\pi}\sigma^2)^n} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right]
 \end{aligned} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}
 p(y|\mu, \tau) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\
 &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\
 &= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]
 \end{aligned} \tag{8}$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

Combining the likelihood function ( $\rightarrow$  I/5.1.2) (8) with the prior distribution ( $\rightarrow$  I/5.1.3) (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned}
 p(y, \mu, \tau) &= p(y|\mu, \tau) p(\mu, \tau) \\
 &= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \cdot \\
 &\quad \sqrt{\frac{\tau \lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\tau \lambda_0}{2} (\mu - \mu_0)^2 \right] \cdot \\
 &\quad \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] .
 \end{aligned} \tag{9}$$

Collecting identical variables gives:

$$\begin{aligned}
 p(y, \mu, \tau) &= \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
 &\quad \exp \left[ -\frac{\tau}{2} \left( \sum_{i=1}^n (y_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right) \right] .
 \end{aligned} \tag{10}$$

Expanding the products in the exponent ( $\rightarrow$  III/1.1.8) gives

$$p(y, \mu, \tau) = \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[ -\frac{\tau}{2} ((y^T y - 2\mu n \bar{y} + n\mu^2) + \lambda_0(\mu^2 - 2\mu\mu_0 + \mu_0^2)) \right] \quad (11)$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  and  $y^T y = \sum_{i=1}^n y_i^2$ , such that

$$p(y, \mu, \tau) = \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[ -\frac{\tau}{2} (\mu^2(\lambda_0 + n) - 2\mu(\lambda_0\mu_0 + n\bar{y}) + (y^T y + \lambda_0\mu_0^2)) \right] \quad (12)$$

Completing the square over  $\mu$ , we finally have

$$p(y, \mu, \tau) = \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[ -\frac{\tau \lambda_n}{2} (\mu - \mu_n)^2 - \frac{\tau}{2} (y^T y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) \right] \quad (13)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu_n &= \frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n . \end{aligned} \quad (14)$$

Ergo, the joint likelihood is proportional to

$$p(y, \mu, \tau) \propto \tau^{1/2} \cdot \exp \left[ -\frac{\tau \lambda_n}{2} (\mu - \mu_n)^2 \right] \cdot \tau^{a_n-1} \cdot \exp [-b_n \tau] \quad (15)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) . \end{aligned} \quad (16)$$

From the term in (13), we can isolate the posterior distribution over  $\mu$  given  $\tau$ :

$$p(\mu|\tau, y) = \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) . \quad (17)$$

From the remaining term, we can isolate the posterior distribution over  $\tau$ :

$$p(\tau|y) = \text{Gam}(\tau; a_n, b_n) . \quad (18)$$

Together, (17) and (18) constitute the joint ( $\rightarrow$  I/1.3.2) posterior distribution ( $\rightarrow$  I/5.1.9) of  $\mu$  and  $\tau$ .



**Sources:**

- Bishop CM (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, pp. 97-102, eq. 2.154; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

**1.1.10 Log model evidence****Theorem:** Let

$$m : y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.1.8) over the model parameters  $\mu$  and  $\tau = 1/\sigma^2$ :

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau \lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\log p(y|m) = -\frac{n}{2} \log(2\pi) + \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \quad (3)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) . \end{aligned} \quad (4)$$

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the model evidence ( $\rightarrow$  I/5.1.15) for this model is:

$$p(y|m) = \iint p(y|\mu, \tau) p(\mu, \tau) d\mu d\tau . \quad (5)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand is equivalent to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(y|m) = \iint p(y, \mu, \tau) d\mu d\tau . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned}
p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\
&= \frac{1}{(\sqrt{2\pi}\sigma^2)^n} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}
p(y|\mu, \tau) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\
&= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{8}$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

When deriving the posterior distribution ( $\rightarrow$  III/1.1.9)  $p(\mu, \tau|y)$ , the joint likelihood  $p(y, \mu, \tau)$  is obtained as

$$\begin{aligned}
p(y, \mu, \tau) &= \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[ -\frac{\tau \lambda_n}{2} (\mu - \mu_n)^2 - \frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] .
\end{aligned} \tag{9}$$

Using the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), we can rewrite this as

$$\begin{aligned}
p(y, \mu, \tau) &= \sqrt{\frac{\tau^n}{(2\pi)^n}} \sqrt{\frac{\tau \lambda_0}{2\pi}} \sqrt{\frac{2\pi}{\tau \lambda_n}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) \exp \left[ -\frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] .
\end{aligned} \tag{10}$$

Now,  $\mu$  can be integrated out easily:

$$\begin{aligned}
\int p(y, \mu, \tau) d\mu &= \sqrt{\frac{1}{(2\pi)^n}} \sqrt{\frac{\lambda_0}{\lambda_n}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0+n/2-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[ -\frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] .
\end{aligned} \tag{11}$$

Using the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we can rewrite this as

$$\int p(y, \mu, \tau) d\mu = \sqrt{\frac{1}{(2\pi)^n}} \sqrt{\frac{\lambda_0}{\lambda_n}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \text{Gam}(\tau; a_n, b_n) . \quad (12)$$

Finally,  $\tau$  can also be integrated out:

$$\iint p(y, \mu, \tau) d\mu d\tau = \sqrt{\frac{1}{(2\pi)^n}} \sqrt{\frac{\lambda_0}{\lambda_n}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} . \quad (13)$$

Thus, the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is given by

$$\log p(y|m) = -\frac{n}{2} \log(2\pi) + \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \quad (14)$$

#### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

#### 1.1.11 Accuracy and complexity

**Theorem:** Let

$$m : y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.1.8) over the model parameters  $\mu$  and  $\tau = 1/\sigma^2$ :

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau \lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, accuracy and complexity ( $\rightarrow$  IV/3.1.6) of this model are

$$\begin{aligned} \text{Acc}(m) &= -\frac{1}{2} \frac{a_n}{b_n} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n \lambda_n^{-1} + \frac{n}{2} (\psi(a_n) - \log(b_n)) - \frac{n}{2} \log(2\pi) \\ \text{Com}(m) &= \frac{1}{2} \frac{a_n}{b_n} [\lambda_0(\mu_0 - \mu_n)^2 - 2(b_n - b_0)] + \frac{1}{2} \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\ &\quad + a_0 \cdot \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \cdot \psi(a_n) \end{aligned} \quad (3)$$

where  $\mu_n$  and  $\lambda_n$  as well as  $a_n$  and  $b_n$  are the posterior hyperparameters for the univariate Gaussian ( $\rightarrow$  III/1.1.9) and  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Proof:** Model accuracy and complexity are defined as ( $\rightarrow$  IV/3.1.6)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\mu, \tau, m) \rangle_{p(\mu, \tau|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\mu, \tau|y, m) || p(\mu, \tau|m)] . \end{aligned} \quad (4)$$

The accuracy term is the expectation ( $\rightarrow$  I/1.12.1) of the log-likelihood function ( $\rightarrow$  I/4.1.2)  $\log p(y|\mu, \tau)$  with respect to the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\mu, \tau|y)$ . With the log-likelihood function for the univariate Gaussian ( $\rightarrow$  III/1.1.2) and the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9), the model accuracy of  $m$  evaluates to:

$$\begin{aligned}
\text{Acc}(m) &= \langle \log p(y|\mu, \tau) \rangle_{p(\mu, \tau|y)} \\
&= \left\langle \left\langle \log p(y|\mu, \tau) \right\rangle_{p(\mu|\tau, y)} \right\rangle_{p(\tau|y)} \\
&= \left\langle \left\langle \frac{n}{2} \log(\tau) - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y^T y - 2n\bar{y}\mu + n\mu^2) \right\rangle_{\mathcal{N}(\mu_n, (\tau\lambda_n)^{-1})} \right\rangle_{\text{Gam}(a_n, b_n)} \\
&= \left\langle \frac{n}{2} \log(\tau) - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} \right\rangle_{\text{Gam}(a_n, b_n)} \quad (5) \\
&= \frac{n}{2} (\psi(a_n) - \log(b_n)) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \frac{a_n}{b_n} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} \\
&= -\frac{1}{2} \frac{a_n}{b_n} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} + \frac{n}{2} (\psi(a_n) - \log(b_n)) - \frac{n}{2} \log(2\pi)
\end{aligned}$$

The complexity penalty is the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\mu, \tau|y)$  from the prior distribution ( $\rightarrow$  I/5.1.3)  $p(\mu, \tau)$ . With the prior distribution ( $\rightarrow$  III/1.1.8) given by (2), the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9) and the Kullback-Leibler divergence of the normal-gamma distribution ( $\rightarrow$  II/4.4.7), the model complexity of  $m$  evaluates to:

$$\begin{aligned}
\text{Com}(m) &= \text{KL} [p(\mu, \tau|y) || p(\mu, \tau)] \\
&= \text{KL} [\text{NG}(\mu_n, \lambda_n^{-1}, a_n, b_n) || \text{NG}(\mu_0, \lambda_0^{-1}, a_0, b_0)] \\
&= \frac{1}{2} \frac{a_n}{b_n} [\lambda_0(\mu_0 - \mu_n)^2] + \frac{1}{2} \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&\quad + a_0 \cdot \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \cdot \psi(a_n) - (b_n - b_0) \cdot \frac{a_n}{b_n} \quad (6) \\
&= \frac{1}{2} \frac{a_n}{b_n} [\lambda_0(\mu_0 - \mu_n)^2 - 2(b_n - b_0)] + \frac{1}{2} \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&\quad + a_0 \cdot \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \cdot \psi(a_n) .
\end{aligned}$$

A control calculation confirms that

$$\text{Acc}(m) - \text{Com}(m) = \text{LME}(m) \quad (7)$$

where  $\text{LME}(m)$  is the log model evidence for the univariate Gaussian ( $\rightarrow$  III/1.1.10).

■

### 1.1.12 Cross-validated log model evidence

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.1.8) as the prior distribution ( $\rightarrow$  I/5.1.3) on the mean parameter ( $\rightarrow$  I/1.12.1)  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)) and both imposing a gamma distribution ( $\rightarrow$  III/1.1.8) on the precision parameter ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ :

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \tau^{-1}), \quad \mu = 0, \quad \tau \sim \text{Gam}(a_0, b_0) \\ m_1 : y_i &\sim \mathcal{N}(\mu, \tau^{-1}), \quad \mu|\tau \sim \mathcal{N}(\mu_0, (\tau\lambda_0)^{-1}), \quad \tau \sim \text{Gam}(a_0, b_0) \end{aligned} \quad (2)$$

Then, the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are

$$\begin{aligned} \text{cvLME}(m_0) &= -\frac{n}{2} \log(2\pi) + S \cdot \log \Gamma\left(\frac{n}{2}\right) - S \cdot \log \Gamma\left(\frac{1}{2} \frac{S-1}{S} n\right) \\ &\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^T y) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} (y_1^{(i)T} y_1^{(i)}) \right] \\ \text{cvLME}(m_1) &= -\frac{n}{2} \log(2\pi) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) + S \cdot \log \Gamma\left(\frac{n}{2}\right) - S \cdot \log \Gamma\left(\frac{1}{2} \frac{S-1}{S} n\right) \\ &\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^T y - n\bar{y}^2) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} (y_1^{(i)T} y_1^{(i)} - n_1 \bar{y}_1^{(i)}) \right] \end{aligned} \quad (3)$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2),  $y_1^{(i)}$  are the training data in the  $i$ -th cross-validation fold with  $n_1$  data points and  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9).

**Proof:** For evaluation of the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME), we assume that  $n$  data points are divided into  $S \mid n$  data subsets without remainder. Then, the number of training data points  $n_1$  and test data points  $n_2$  are given by

$$\begin{aligned} n &= n_1 + n_2 \\ n_1 &= \frac{S-1}{S} n \\ n_2 &= \frac{1}{S} n, \end{aligned} \quad (4)$$

such that training data  $y_1$  and test data  $y_2$  in the  $i$ -th cross-validation fold are

$$\begin{aligned} y &= \{y_1, \dots, y_n\} \\ y_1^{(i)} &= \left\{ x \in y \mid x \notin y_2^{(i)} \right\} = y \setminus y_2^{(i)} \\ y_2^{(i)} &= \{y_{(i-1) \cdot n_2 + 1}, \dots, y_{i \cdot n_2}\} . \end{aligned} \quad (5)$$

First, we have a look at the alternative  $m_1$  assuming  $\mu \neq 0$ . To begin, the training data  $y_1^{(i)}$  are analyzed using a non-informative prior distribution ( $\rightarrow$  I/5.2.3) and applying the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9):

$$\begin{aligned}
\mu_0^{(1)} = 0 &\Rightarrow \mu_n^{(1)} = \frac{\lambda_0^{(1)} \mu_0^{(1)} + n_1 \bar{y}_1^{(i)}}{\lambda_0^{(1)} + n_1} = \bar{y}_1^{(i)} \\
\lambda_0^{(1)} = 0 &\Rightarrow \lambda_n^{(1)} = \lambda_0^{(1)} + n_1 = n_1 \\
a_0^{(1)} = 0 &\Rightarrow a_n^{(1)} = a_0^{(1)} + \frac{n_1}{2} = \frac{n_1}{2} \\
b_0^{(1)} = 0 &\Rightarrow b_n^{(1)} = b_0^{(1)} + \frac{1}{2} \left( y_1^{(i)T} y_1^{(i)} + \lambda_0^{(1)} \mu_0^{(1)2} - \lambda_n^{(1)} \mu_n^{(1)2} \right) = \frac{1}{2} \left( y_1^{(i)T} y_1^{(i)} - n_1 \left( \bar{y}_1^{(i)} \right)^2 \right).
\end{aligned} \tag{6}$$

This results in a posterior characterized by  $\mu_n^{(1)}$ ,  $\lambda_n^{(1)}$ ,  $a_n^{(1)}$  and  $b_n^{(1)}$ . Then, the test data  $y_2^{(i)}$  are analyzed using this posterior as an informative prior distribution ( $\rightarrow$  I/5.2.3), again applying the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9):

$$\begin{aligned}
\mu_0^{(2)} &= \mu_n^{(1)} = \bar{y}_1^{(i)} \\
\Rightarrow \mu_n^{(2)} &= \frac{\lambda_0^{(2)} \mu_0^{(2)} + n_2 \bar{y}_2^{(i)}}{\lambda_0^{(2)} + n_2} \\
&= \frac{n_1 \bar{y}_1^{(i)} + n_2 \bar{y}_2^{(i)}}{n_1 + n_2} \\
&= \frac{n \bar{y}}{n} = \bar{y} \\
\lambda_0^{(2)} &= \lambda_n^{(1)} = n_1 \\
\Rightarrow \lambda_n^{(2)} &= \lambda_0^{(2)} + n_2 = n_1 + n_2 = n \\
a_0^{(2)} &= a_n^{(1)} = \frac{n_1}{2} \\
\Rightarrow a_n^{(2)} &= a_0^{(2)} + \frac{n_2}{2} = \frac{n_1}{2} + \frac{n_2}{2} = \frac{n}{2} \\
b_0^{(2)} &= b_n^{(1)} = \frac{1}{2} \left( y_1^{(i)T} y_1^{(i)} - n_1 \bar{y}_1^{(i)2} \right) \\
\Rightarrow b_n^{(2)} &= b_0^{(2)} + \frac{1}{2} \left( y_2^{(i)T} y_2^{(i)} + \lambda_0^{(2)} \mu_0^{(2)2} - \lambda_n^{(2)} \mu_n^{(2)2} \right) \\
&= \frac{1}{2} \left( y_1^{(i)T} y_1^{(i)} - n_1 \left( \bar{y}_1^{(i)} \right)^2 \right) + \frac{1}{2} \left( y_2^{(i)T} y_2^{(i)} + n_1 \left( \bar{y}_1^{(i)} \right)^2 - n \bar{y}^2 \right) \\
&= \frac{1}{2} \left( y^T y - n \bar{y}^2 \right).
\end{aligned} \tag{7}$$

In the test data, we now have a prior characterized by  $\mu_0^{(2)}$ ,  $\lambda_0^{(2)}$ ,  $a_0^{(2)}$  and  $b_0^{(2)}$  and a posterior characterized by  $\mu_n^{(2)}$ ,  $\lambda_n^{(2)}$ ,  $a_n^{(2)}$  and  $b_n^{(2)}$ . Applying the log model evidence for the univariate Gaussian ( $\rightarrow$  III/1.1.10), the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) therefore follows as

$$\begin{aligned}
\text{oosLME}_i(m_1) &= -\frac{n_2}{2} \log(2\pi) + \frac{1}{2} \log \frac{\lambda_0^{(2)}}{\lambda_n^{(2)}} + \log \Gamma(a_n^{(2)}) - \log \Gamma(a_0^{(2)}) \\
&\quad + a_0^{(2)} \log b_0^{(2)} - a_n^{(2)} \log b_n^{(2)} \\
&= -\frac{n_2}{2} \log(2\pi) + \frac{1}{2} \log \left( \frac{n_1}{n} \right) + \log \Gamma \left( \frac{n}{2} \right) - \log \Gamma \left( \frac{n_1}{2} \right) \\
&\quad + \frac{n_1}{2} \log \left[ \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} - n_1 \bar{y}_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^{\text{T}} y - n \bar{y}^2) \right] \\
&= -\frac{n}{2S} \log(2\pi) + \frac{1}{2} \log \left( \frac{S-1}{S} \right) + \log \Gamma \left( \frac{n}{2} \right) - \log \Gamma \left( \frac{(S-1)n}{2S} \right) \\
&\quad + \frac{(S-1)n}{2S} \log \left[ \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} - n_1 \bar{y}_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^{\text{T}} y - n \bar{y}^2) \right].
\end{aligned} \tag{8}$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, such that the cvLME of  $m_1$  is:

$$\begin{aligned}
\text{cvLME}(m_1) &= \sum_{i=1}^S \text{oosLME}_i(m_1) \\
&= \sum_{i=1}^S \left( -\frac{n}{2S} \log(2\pi) + \frac{1}{2} \log \left( \frac{S-1}{S} \right) + \log \Gamma \left( \frac{n}{2} \right) - \log \Gamma \left( \frac{(S-1)n}{2S} \right) \right. \\
&\quad \left. + \frac{(S-1)n}{2S} \log \left[ \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} - n_1 \bar{y}_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^{\text{T}} y - n \bar{y}^2) \right] \right) \\
&= -\frac{n}{2} \log(2\pi) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) + S \cdot \log \Gamma \left( \frac{n}{2} \right) - S \cdot \log \Gamma \left( \frac{1}{2} \frac{S-1}{S} n \right) \\
&\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^{\text{T}} y - n \bar{y}^2) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} - n_1 \bar{y}_1^{(i)} \right) \right].
\end{aligned} \tag{9}$$

Next, we consider the null model  $m_0$  assuming  $\mu = 0$ . Because this model has no parameter for  $\mu$ , there are no hyperparameters  $\mu_0$  and  $\lambda_0$ , but only hyperparameters  $a_0$  and  $b_0$  for the parameter  $\tau$ . In the first step, the training data  $y_1^{(i)}$  are analyzed using a non-informative prior distribution ( $\rightarrow$  I/5.2.3) and applying the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9):

$$\begin{aligned}
a_0^{(1)} = 0 &\Rightarrow a_n^{(1)} = a_0^{(1)} + \frac{n_1}{2} = \frac{n_1}{2} \\
b_0^{(1)} = 0 &\Rightarrow b_n^{(1)} = b_0^{(1)} + \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} \right) = \frac{1}{2} \left( y_1^{(i)\text{T}} y_1^{(i)} \right).
\end{aligned} \tag{10}$$

In the second step, the test data  $y_2^{(i)}$  are analyzed using this posterior as an informative prior distribution ( $\rightarrow$  I/5.2.3), again applying the posterior distribution for the univariate Gaussian ( $\rightarrow$  III/1.1.9):

$$\begin{aligned}
a_0^{(2)} &= a_n^{(1)} = \frac{n_1}{2} \\
\Rightarrow a_n^{(2)} &= a_0^{(2)} + \frac{n_2}{2} = \frac{n_1}{2} + \frac{n_2}{2} = \frac{n}{2} \\
b_0^{(2)} &= b_n^{(1)} = \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) \\
\Rightarrow b_n^{(2)} &= b_0^{(2)} + \frac{1}{2} \left( y_2^{(i)\top} y_2^{(i)} \right) = \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) + \frac{1}{2} \left( y_2^{(i)\top} y_2^{(i)} \right) = \frac{1}{2} (y^\top y) .
\end{aligned} \tag{11}$$

In the test data, we now have a prior characterized by  $a_0^{(2)}$  and  $b_0^{(2)}$  and a posterior characterized  $a_n^{(2)}$  and  $b_n^{(2)}$ . Applying the log model evidence for the univariate Gaussian ( $\rightarrow$  III/1.1.10), the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) therefore follows as

$$\begin{aligned}
\text{oosLME}_i(m_1) &= -\frac{n_2}{2} \log(2\pi) + \log \Gamma(a_n^{(2)}) - \log \Gamma(a_0^{(2)}) \\
&\quad + a_0^{(2)} \log b_0^{(2)} - a_n^{(2)} \log b_n^{(2)} \\
&= -\frac{n_2}{2} \log(2\pi) + \log \Gamma\left(\frac{n}{2}\right) - \log \Gamma\left(\frac{n_1}{2}\right) \\
&\quad + \frac{n_1}{2} \log \left[ \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^\top y) \right] \\
&= -\frac{n}{2S} \log(2\pi) + \log \Gamma\left(\frac{n}{2}\right) - \log \Gamma\left(\frac{(S-1)n}{2S}\right) \\
&\quad + \frac{(S-1)n}{2S} \log \left[ \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^\top y) \right] .
\end{aligned} \tag{12}$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, the cvLME of  $m_0$  becomes:

$$\begin{aligned}
\text{cvLME}(m_0) &= \sum_{i=1}^S \text{oosLME}_i(m_0) \\
&= \sum_{i=1}^S \left( -\frac{n}{2S} \log(2\pi) + \log \Gamma\left(\frac{n}{2}\right) - \log \Gamma\left(\frac{(S-1)n}{2S}\right) \right. \\
&\quad \left. + \frac{(S-1)n}{2S} \log \left[ \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) \right] - \frac{n}{2} \log \left[ \frac{1}{2} (y^\top y) \right] \right) \\
&= -\frac{n}{2} \log(2\pi) + S \cdot \log \Gamma\left(\frac{n}{2}\right) - S \cdot \log \Gamma\left(\frac{1}{2} \frac{S-1}{S} n\right) \\
&\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^\top y) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} \left( y_1^{(i)\top} y_1^{(i)} \right) \right] .
\end{aligned} \tag{13}$$

Together, (13) and (9) conform to the results given in (3). ■



### 1.1.13 Cross-validated log Bayes factor

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.1.1) with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.1.8) as the prior distribution ( $\rightarrow$  I/5.1.3) on the mean parameter ( $\rightarrow$  I/1.12.1)  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)) and both imposing a gamma distribution ( $\rightarrow$  III/1.1.8) on the precision parameter ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ :

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \tau^{-1}), \quad \mu = 0, \quad \tau \sim \text{Gam}(a_0, b_0) \\ m_1 : y_i &\sim \mathcal{N}(\mu, \tau^{-1}), \quad \mu|\tau \sim \mathcal{N}(\mu_0, (\tau\lambda_0)^{-1}), \quad \tau \sim \text{Gam}(a_0, b_0) \end{aligned} \quad (2)$$

Then, the cross-validated ( $\rightarrow$  IV/3.1.9) log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_0$  is

$$\text{cvLBF}_{10} = \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{S \cdot n}{2} \left[ \log \left( 1 - \frac{n\bar{y}^2}{y^T y} \right) \right] + \frac{n_1}{2} \sum_{i=1}^S \left[ \log \left( 1 - \frac{n_1 \bar{y}_1^{(i)}}{y_1^{(i)T} y_1^{(i)}} \right) \right] \quad (3)$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2),  $y_1^{(i)}$  are the training data in the  $i$ -th cross-validation fold with  $n_1$  data points and  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9).

**Proof:** The relationship between log Bayes factor and log model evidences ( $\rightarrow$  IV/3.3.8) also holds for cross-validated log bayes factor ( $\rightarrow$  IV/3.3.6) (cvLBF) and cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME):

$$\text{cvLBF}_{12} = \text{cvLME}(m_1) - \text{cvLME}(m_2). \quad (4)$$

The cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are given by ( $\rightarrow$  III/1.1.12)

$$\begin{aligned} \text{cvLME}(m_0) &= -\frac{n}{2} \log(2\pi) + S \cdot \log \Gamma \left( \frac{n}{2} \right) - S \cdot \log \Gamma \left( \frac{1}{2} \frac{S-1}{S} n \right) \\ &\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^T y) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} (y_1^{(i)T} y_1^{(i)}) \right] \\ \text{cvLME}(m_1) &= -\frac{n}{2} \log(2\pi) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) + S \cdot \log \Gamma \left( \frac{n}{2} \right) - S \cdot \log \Gamma \left( \frac{1}{2} \frac{S-1}{S} n \right) \\ &\quad - \frac{S \cdot n}{2} \log \left[ \frac{1}{2} (y^T y - n\bar{y}^2) \right] + \frac{n_1}{2} \sum_{i=1}^S \log \left[ \frac{1}{2} (y_1^{(i)T} y_1^{(i)} - n_1 \bar{y}_1^{(i)}) \right]. \end{aligned} \quad (5)$$

Subtracting the two cvLMEs from each other, the cvLBF emerges as

$$\begin{aligned}
\text{cvLBF}_{10} &= \text{cvLME}(m_1) - \text{LME}(m_0) \\
&= + \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{S \cdot n}{2} \left[ \log \left( \frac{1}{2} (y^T y - n \bar{y}^2) \right) - \log \left( \frac{1}{2} (y^T y) \right) \right] \\
&\quad + \frac{n_1}{2} \sum_{i=1}^S \left[ \log \left( \frac{1}{2} (y_1^{(i)T} y_1^{(i)} - n_1 \bar{y}_1^{(i)}) \right) - \log \left( \frac{1}{2} (y_1^{(i)T} y_1^{(i)}) \right) \right] \\
&= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{S \cdot n}{2} \left[ \log \left( \frac{y^T y - n \bar{y}^2}{y^T y} \right) \right] + \frac{n_1}{2} \sum_{i=1}^S \left[ \log \left( \frac{y_1^{(i)T} y_1^{(i)} - n_1 \bar{y}_1^{(i)}}{y_1^{(i)T} y_1^{(i)}} \right) \right] \\
&= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{S \cdot n}{2} \left[ \log \left( 1 - \frac{n \bar{y}^2}{y^T y} \right) \right] + \frac{n_1}{2} \sum_{i=1}^S \left[ \log \left( 1 - \frac{n_1 \bar{y}_1^{(i)}}{y_1^{(i)T} y_1^{(i)}} \right) \right] .
\end{aligned} \tag{6}$$

■

## 1.2 Univariate Gaussian with known variance

### 1.2.1 Definition

**Definition:** A univariate Gaussian data set with known variance is given by a set of real numbers  $y = \{y_1, \dots, y_n\}$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a normal distribution ( $\rightarrow$  II/3.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ :

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n . \tag{1}$$

#### Sources:

- Bishop, Christopher M. (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 97, eq. 2.137; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 1.2.2 Maximum likelihood estimation

**Theorem:** Let there be univariate Gaussian data with known variance ( $\rightarrow$  III/1.2.1)  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n . \tag{1}$$

Then, the maximum likelihood estimate ( $\rightarrow$  I/4.1.3) for the mean  $\mu$  is given by

$$\hat{\mu} = \bar{y} \tag{2}$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \tag{3}$$

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for each observation is given by the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10)

$$p(y_i|\mu) = \mathcal{N}(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \quad (4)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\mu) = \prod_{i=1}^n p(y_i|\mu) = \sqrt{\frac{1}{(2\pi\sigma^2)^n}} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] . \quad (5)$$

This can be developed into

$$\begin{aligned} p(y|\mu) &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n \left( \frac{y_i^2 - 2y_i\mu + \mu^2}{\sigma^2} \right) \right] \\ &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} (y^T y - 2n\bar{y}\mu + n\mu^2) \right] \end{aligned} \quad (6)$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  is the mean ( $\rightarrow$  I/1.12.2) of data points ( $\rightarrow$  I/1.1.5) and  $y^T y = \sum_{i=1}^n y_i^2$  is the sum of squared data points ( $\rightarrow$  I/1.1.5). Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$\text{LL}(\mu) = \log p(y|\mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y^T y - 2n\bar{y}\mu + n\mu^2) . \quad (7)$$

The derivatives of the log-likelihood with respect to  $\mu$  are

$$\begin{aligned} \frac{d\text{LL}(\mu)}{d\mu} &= \frac{n\bar{y}}{\sigma^2} - \frac{n\mu}{\sigma^2} = \frac{n}{\sigma^2} (\bar{y} - \mu) \\ \frac{d^2\text{LL}(\mu)}{d\mu^2} &= -\frac{n}{\sigma^2} . \end{aligned} \quad (8)$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned} \frac{d\text{LL}(\hat{\mu})}{d\mu} &= 0 \\ 0 &= \frac{n}{\sigma^2} (\bar{y} - \hat{\mu}) \\ 0 &= \bar{y} - \hat{\mu} \\ \hat{\mu} &= \bar{y} \end{aligned} \quad (9)$$

Plugging this value into the second derivative, we confirm:

$$\frac{d^2\text{LL}(\hat{\mu})}{d\mu^2} = -\frac{n}{\sigma^2} < 0 . \quad (10)$$

This demonstrates that the estimate ( $\rightarrow$  I/3.1.1)  $\hat{\mu} = \bar{y}$  maximizes the likelihood ( $\rightarrow$  I/5.1.2)  $p(y|\mu)$ . ■

#### Sources:

- Bishop, Christopher M. (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 98, eq. 2.143; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 1.2.3 One-sample z-test

**Theorem:** Let

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$z = \sqrt{n} \frac{\bar{y} - \mu_0}{\sigma} \quad (2)$$

with sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  follows a standard normal distribution ( $\rightarrow$  II/3.2.3)

$$z \sim \mathcal{N}(0, 1) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu = \mu_0 . \quad (4)$$

**Proof:** The sample mean ( $\rightarrow$  I/1.12.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \quad (5)$$

Using the linear combination formula for normal random variables ( $\rightarrow$  II/3.2.26), the sample mean follows a normal distribution ( $\rightarrow$  II/3.2.1) with the following parameters:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \sim \mathcal{N} \left( \frac{1}{n} n \mu, \left( \frac{1}{n} \right)^2 n \sigma^2 \right) = \mathcal{N}(\mu, \sigma^2/n) . \quad (6)$$

Again employing the linear combination theorem, the distribution of  $z = \sqrt{n/\sigma^2}(\bar{y} - \mu_0)$  becomes

$$z = \sqrt{\frac{n}{\sigma^2}}(\bar{y} - \mu_0) \sim \mathcal{N} \left( \sqrt{\frac{n}{\sigma^2}}(\mu - \mu_0), \left( \sqrt{\frac{n}{\sigma^2}} \right)^2 \frac{\sigma^2}{n} \right) = \mathcal{N} \left( \sqrt{n} \frac{\mu - \mu_0}{\sigma}, 1 \right) , \quad (7)$$

such that, under the null hypothesis in (4), we have:

$$z \sim \mathcal{N}(0, 1), \quad \text{if } \mu = \mu_0 . \quad (8)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) can be rejected when  $z$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the standard normal distribution ( $\rightarrow$  II/3.2.3) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ . ■

#### Sources:

- Wikipedia (2021): “Z-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: [https://en.wikipedia.org/wiki/Z-test#Use\\_in\\_location\\_testing](https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing).
- Wikipedia (2021): “Gauß-Test”; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: <https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Einstichproben-Gau%C3%9F-Test>.

## 1.2.4 Two-sample z-test

**Theorem:** Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma_1^2), & i = 1, \dots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma_2^2), & i = 1, \dots, n_2 \end{aligned} \quad (1)$$

be two univariate Gaussian data sets ( $\rightarrow$  III/1.1.1) representing two groups of unequal size  $n_1$  and  $n_2$  with unknown means  $\mu_1$  and  $\mu_2$  and unknown variances  $\sigma_1^2$  and  $\sigma_2^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (2)$$

with sample means ( $\rightarrow$  I/1.12.2)  $\bar{y}_1$  and  $\bar{y}_2$  follows a standard normal distribution ( $\rightarrow$  II/3.2.3)

$$z \sim \mathcal{N}(0, 1) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu_1 - \mu_2 = \mu_\Delta . \quad (4)$$

**Proof:** The sample means ( $\rightarrow$  I/1.12.2) are given by

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} . \end{aligned} \quad (5)$$

Using the linear combination formula for normal random variables ( $\rightarrow$  II/3.2.26), the sample means follows normal distributions ( $\rightarrow$  II/3.2.1) with the following parameters:

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N} \left( \frac{1}{n_1} n_1 \mu_1, \left( \frac{1}{n_1} \right)^2 n_1 \sigma_1^2 \right) = \mathcal{N}(\mu_1, \sigma_1^2/n_1) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N} \left( \frac{1}{n_2} n_2 \mu_2, \left( \frac{1}{n_2} \right)^2 n_2 \sigma_2^2 \right) = \mathcal{N}(\mu_2, \sigma_2^2/n_2) . \end{aligned} \quad (6)$$

Again employing the linear combination theorem, the distribution of  $z = [(\bar{y}_1 - \bar{y}_2) - \mu_\Delta] / \sigma_\Delta$  becomes

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma_\Delta} \sim \mathcal{N} \left( \frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, \left( \frac{1}{\sigma_\Delta} \right)^2 \sigma_\Delta^2 \right) = \mathcal{N} \left( \frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, 1 \right) \quad (7)$$

where  $\sigma_\Delta$  is the pooled standard deviation ( $\rightarrow$  I/1.19.3)

$$\sigma_\Delta = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , \quad (8)$$

such that, under the null hypothesis in (4), we have:

$$z \sim \mathcal{N}(0, 1), \quad \text{if } \mu_\Delta = \mu_1 - \mu_2. \quad (9)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) can be rejected when  $z$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the standard normal distribution ( $\rightarrow$  II/3.2.3) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ . ■

#### Sources:

- Wikipedia (2021): “Z-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: [https://en.wikipedia.org/wiki/Z-test#Use\\_in\\_location\\_testing](https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing).
- Wikipedia (2021): “Gauß-Test”; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: [https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test\\_f%C3%BCr\\_unabh%C3%A4ngige\\_Stichproben](https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test_f%C3%BCr_unabh%C3%A4ngige_Stichproben).

### 1.2.5 Paired z-test

**Theorem:** Let  $y_{i1}$  and  $y_{i2}$  with  $i = 1, \dots, n$  be paired observations, such that

$$y_{i1} \sim \mathcal{N}(y_{i2} + \mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

is a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown shift  $\mu$  and known variance  $\sigma^2$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$z = \sqrt{n} \frac{\bar{d} - \mu_0}{\sigma} \quad \text{where} \quad d_i = y_{i1} - y_{i2} \quad (2)$$

with sample mean ( $\rightarrow$  I/1.12.2)  $\bar{d}$  follows a standard normal distribution ( $\rightarrow$  II/3.2.3)

$$z \sim \mathcal{N}(0, 1) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : \mu = \mu_0. \quad (4)$$

**Proof:** Define the pair-wise difference  $d_i = y_{i1} - y_{i2}$  which, as a linear transformation of a normal random variable ( $\rightarrow$  II/4.1.14), is distributed as

$$d_i = y_{i1} - y_{i2} \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (5)$$

Therefore,  $d_1, \dots, d_n$  satisfy the conditions of the one-sample z-test ( $\rightarrow$  III/1.2.3) which results in the test statistic given by (2). ■

#### Sources:

- Wikipedia (2021): “Z-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: [https://en.wikipedia.org/wiki/Z-test#Use\\_in\\_location\\_testing](https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing).
- Wikipedia (2021): “Gauß-Test”; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: [https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test\\_f%C3%BCr\\_abh%C3%A4ngige\\_\(verbundene\)\\_Stichproben](https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test_f%C3%BCr_abh%C3%A4ngige_(verbundene)_Stichproben).

### 1.2.6 Conjugate prior distribution

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for this model is a normal distribution ( $\rightarrow$  II/3.2.1)

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \quad (2)$$

with prior ( $\rightarrow$  I/5.1.3) mean ( $\rightarrow$  I/1.12.1)  $\mu_0$  and prior ( $\rightarrow$  I/5.1.3) precision ( $\rightarrow$  I/1.13.13)  $\lambda_0$ .

**Proof:** By definition, a conjugate prior ( $\rightarrow$  I/5.2.5) is a prior distribution ( $\rightarrow$  I/5.1.3) that, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \left( \sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (3)$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\ &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\ &= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (4)$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

Expanding the product in the exponent, we have

$$\begin{aligned}
p(y|\mu) &= \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2) \right] \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \left( \sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2 \right) \right] \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y^T y - 2\mu n\bar{y} + n\mu^2) \right] \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[ -\frac{\tau n}{2} \left( \frac{1}{n} y^T y - 2\mu\bar{y} + \mu^2 \right) \right]
\end{aligned} \tag{5}$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  is the mean of data points and  $y^T y = \sum_{i=1}^n y_i^2$  is the sum of squared data points. Completing the square over  $\mu$ , finally gives

$$p(y|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[ -\frac{\tau n}{2} \left( (\mu - \bar{y})^2 - \bar{y}^2 + \frac{1}{n} y^T y \right) \right] \tag{6}$$

In other words, the likelihood function ( $\rightarrow$  I/5.1.2) is proportional to an exponential of a squared form of  $\mu$ , weighted by some constant:

$$p(y|\mu) \propto \exp \left[ -\frac{\tau n}{2} (\mu - \bar{y})^2 \right]. \tag{7}$$

The same is true for a normal distribution ( $\rightarrow$  II/3.2.1) over  $\mu$

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \tag{8}$$

the probability density function of which ( $\rightarrow$  II/3.2.10)

$$p(\mu) = \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] \tag{9}$$

exhibits the same proportionality

$$p(\mu) \propto \exp \left[ -\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] \tag{10}$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Bishop, Christopher M. (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, pp. 97-98, eq. 2.138; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.



## 1.2.7 Posterior distribution

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume a normal distribution ( $\rightarrow$  III/1.2.6) over the model parameter  $\mu$ :

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a normal distribution ( $\rightarrow$  II/3.2.1)

$$p(\mu|y) = \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \quad (4)$$

with the sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  and the inverse variance or precision ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ .

**Proof:** According to Bayes' theorem ( $\rightarrow$  I/5.3.2), the posterior distribution ( $\rightarrow$  I/5.1.9) is given by

$$p(\mu|y) = \frac{p(y|\mu) p(\mu)}{p(y)} . \quad (5)$$

Since  $p(y)$  is just a normalization factor, the posterior is proportional ( $\rightarrow$  I/5.1.10) to the numerator:

$$p(\mu|y) \propto p(y|\mu) p(\mu) = p(y, \mu) . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \left( \sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\ &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\ &= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (8)$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

Combining the likelihood function ( $\rightarrow$  I/5.1.2) (8) with the prior distribution ( $\rightarrow$  I/5.1.3) (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, \mu) &= p(y|\mu) p(\mu) \\ &= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] . \end{aligned} \quad (9)$$

Rearranging the terms, we then have:

$$p(y, \mu) = \left( \frac{\tau}{2\pi} \right)^{n/2} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 - \frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] . \quad (10)$$

Expanding the products in the exponent ( $\rightarrow$  III/1.2.6) gives

$$\begin{aligned} p(y, \mu) &= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \left( \frac{\lambda_0}{2\pi} \right)^{\frac{1}{2}} \cdot \exp \left[ -\frac{1}{2} \left( \sum_{i=1}^n \tau (y_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right) \right] \\ &= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \left( \frac{\lambda_0}{2\pi} \right)^{\frac{1}{2}} \cdot \exp \left[ -\frac{1}{2} \left( \sum_{i=1}^n \tau (y_i^2 - 2y_i\mu + \mu^2) + \lambda_0 (\mu^2 - 2\mu\mu_0 + \mu_0^2) \right) \right] \\ &= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \left( \frac{\lambda_0}{2\pi} \right)^{\frac{1}{2}} \cdot \exp \left[ -\frac{1}{2} (\tau (y^T y - 2n\bar{y}\mu + n\mu^2) + \lambda_0 (\mu^2 - 2\mu\mu_0 + \mu_0^2)) \right] \\ &= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \left( \frac{\lambda_0}{2\pi} \right)^{\frac{1}{2}} \cdot \exp \left[ -\frac{1}{2} (\mu^2 (\tau n + \lambda_0) - 2\mu (\tau n\bar{y} + \lambda_0\mu_0) + (\tau y^T y + \lambda_0\mu_0^2)) \right] \end{aligned} \quad (11)$$

where  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  and  $y^T y = \sum_{i=1}^n y_i^2$ . Completing the square in  $\mu$  then yields

$$p(y, \mu) = \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \left( \frac{\lambda_0}{2\pi} \right)^{\frac{1}{2}} \cdot \exp \left[ -\frac{\lambda_n}{2} (\mu - \mu_n)^2 + f_n \right] \quad (12)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu_n &= \frac{\lambda_0\mu_0 + \tau n\bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \quad (13)$$

and the remaining independent term

$$f_n = -\frac{1}{2} (\tau y^T y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) . \quad (14)$$

Ergo, the joint likelihood in (12) is proportional to

$$p(y, \mu) \propto \exp \left[ -\frac{\lambda_n}{2} (\mu - \mu_n)^2 \right] , \quad (15)$$

such that the posterior distribution over  $\mu$  is given by

$$p(\mu|y) = \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) . \quad (16)$$

with the posterior hyperparameters given in (13). ■

### Sources:

- Bishop, Christopher M. (2006): “Bayesian inference for the Gaussian”; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 98, eqs. 2.139-2.142; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%202006.pdf>.

### 1.2.8 Log model evidence

**Theorem:** Let

$$m : y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume a normal distribution ( $\rightarrow$  III/1.2.6) over the model parameter  $\mu$ :

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\log p(y|m) = \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) . \quad (3)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \quad (4)$$

with the sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  and the inverse variance or precision ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ .

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the model evidence ( $\rightarrow$  I/5.1.15) for this model is:

$$p(y|m) = \int p(y|\mu) p(\mu) d\mu . \quad (5)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand is equivalent to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(y|m) = \int p(y, \mu) d\mu . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned}
p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] \\
&= \left( \sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}
p(y|\mu, \tau) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[ -\frac{\tau}{2} (y_i - \mu)^2 \right] \\
&= \left( \sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{8}$$

using the inverse variance or precision  $\tau = 1/\sigma^2$ .

When deriving the posterior distribution ( $\rightarrow$  III/1.2.7)  $p(\mu|y)$ , the joint likelihood  $p(y, \mu)$  is obtained as

$$\begin{aligned}
p(y, \mu) &= p(y|\mu) p(\mu) \\
&= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \exp \left[ -\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] \\
&= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_0}{2} (\mu^2 - 2\mu\mu_0 + \mu_0^2) - \frac{\tau}{2} (y^T y - 2\mu n\bar{y} + n\mu^2) \right] \\
&= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{1}{2} (\mu^2(\lambda_0 + \tau n) - 2\mu(\lambda_0\mu_0 + \tau n\bar{y}) + \lambda_0\mu_0^2 + \tau y^T y) \right] \\
&= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_0 + \tau n}{2} (\mu^2 - 2\mu\mu_n + \mu_n^2) - \frac{1}{2} (\tau y^T y + \lambda_0\mu_0^2 - (\lambda_0 + \tau n)\mu_n^2) \right] \\
&= \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[ -\frac{\lambda_n}{2} (\mu - \mu_n)^2 - \frac{1}{2} (\tau y^T y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) \right].
\end{aligned} \tag{9}$$

Using the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10), we can rewrite this as

$$p(y, \mu) = \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \sqrt{\frac{2\pi}{\lambda_n}} \cdot \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) \cdot \exp \left[ -\frac{1}{2} (\tau y^T y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) \right]. \tag{10}$$

Now,  $\mu$  can be integrated out using the properties of the probability density function ( $\rightarrow$  I/1.7.1):

$$p(y|m) = \int p(y, \mu) d\mu = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{\lambda_n}} \cdot \exp \left[ -\frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right]. \quad (11)$$

Thus, the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is given by

$$\log p(y|m) = \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2). \quad (12)$$

■

### 1.2.9 Accuracy and complexity

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume a statistical model ( $\rightarrow$  I/5.1.6) imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$ :

$$m : y_i \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}). \quad (2)$$

Then, accuracy and complexity ( $\rightarrow$  IV/3.1.6) of this model are

$$\begin{aligned} \text{Acc}(m) &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} \left[ \tau y^T y - 2\tau n \bar{y} \mu_n + \tau n \mu_n^2 + \frac{\tau n}{\lambda_n} \right] \\ \text{Com}(m) &= \frac{1}{2} \left[ \frac{\lambda_0}{\lambda_n} + \lambda_0 (\mu_0 - \mu_n)^2 - 1 + \log \left( \frac{\lambda_0}{\lambda_n} \right) \right] \end{aligned} \quad (3)$$

where  $\mu_n$  and  $\lambda_n$  are the posterior hyperparameters for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7),  $\tau = 1/\sigma^2$  is the inverse variance or precision ( $\rightarrow$  I/1.13.13) and  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2).

**Proof:** Model accuracy and complexity are defined as ( $\rightarrow$  IV/3.1.6)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\mu, m) \rangle_{p(\mu|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\mu|y, m) || p(\mu|m)] . \end{aligned} \quad (4)$$

The accuracy term is the expectation ( $\rightarrow$  I/1.12.1) of the log-likelihood function ( $\rightarrow$  I/4.1.2)  $\log p(y|\mu)$  with respect to the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\mu|y)$ . With the log-likelihood function for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.2) and the posterior distribution for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7), the model accuracy of  $m$  evaluates to:

$$\begin{aligned} \text{Acc}(m) &= \langle \log p(y|\mu) \rangle_{p(\mu|y)} \\ &= \left\langle \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{\tau}{2} (y^T y - 2n \bar{y} \mu + n \mu^2) \right\rangle_{\mathcal{N}(\mu_n, \lambda_n^{-1})} \\ &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} \left[ \tau y^T y - 2\tau n \bar{y} \mu_n + \tau n \mu_n^2 + \frac{\tau n}{\lambda_n} \right] . \end{aligned} \quad (5)$$

The complexity penalty is the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\mu|y)$  from the prior distribution ( $\rightarrow$  I/5.1.3)  $p(\mu)$ . With the prior distribution ( $\rightarrow$  III/1.2.6) given by (2), the posterior distribution for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7) and the Kullback-Leibler divergence of the normal distribution ( $\rightarrow$  II/3.2.24), the model complexity of  $m$  evaluates to:

$$\begin{aligned} \text{Com}(m) &= \text{KL} [p(\mu|y) || p(\mu)] \\ &= \text{KL} [\mathcal{N}(\mu_n, \lambda_n^{-1}) || \mathcal{N}(\mu_0, \lambda_0^{-1})] \\ &= \frac{1}{2} \left[ \frac{\lambda_0}{\lambda_n} + \lambda_0(\mu_0 - \mu_n)^2 - 1 + \log \left( \frac{\lambda_0}{\lambda_n} \right) \right] . \end{aligned} \quad (6)$$

A control calculation confirms that

$$\text{Acc}(m) - \text{Com}(m) = \text{LME}(m) \quad (7)$$

where  $\text{LME}(m)$  is the log model evidence for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.8).

■

### 1.2.10 Log Bayes factor

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu = 0 \\ m_1 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) . \end{aligned} \quad (2)$$

Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_0$  is

$$\text{LBF}_{10} = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \quad (3)$$

where  $\mu_n$  and  $\lambda_n$  are the posterior hyperparameters for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7) which are functions of the inverse variance or precision ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$  and the sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$ .

**Proof:** The log Bayes factor is equal to the difference of two log model evidences ( $\rightarrow$  IV/3.3.8):

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (4)$$

The LME of the alternative  $m_1$  is equal to the log model evidence for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.8):

$$\text{LME}(m_1) = \log p(y|m_1) = \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) . \quad (5)$$

Because the null model  $m_0$  has no free parameter, its log model evidence ( $\rightarrow$  IV/3.1.3) (logarithmized marginal likelihood ( $\rightarrow$  I/5.1.15)) is equal to the log-likelihood function for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.2) at the value  $\mu = 0$ :

$$\text{LME}(m_0) = \log p(y|\mu = 0) = \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} (\tau y^T y) . \quad (6)$$

Subtracting the two LMEs from each other, the LBF emerges as

$$\text{LBF}_{10} = \text{LME}(m_1) - \text{LME}(m_0) = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \quad (7)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/1.2.7)

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \quad (8)$$

with the sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  and the inverse variance or precision ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ . ■

### 1.2.11 Expectation of log Bayes factor

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu = 0 \\ m_1 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) . \end{aligned} \quad (2)$$

Then, under the null hypothesis ( $\rightarrow$  I/4.3.2) that  $m_0$  generated the data, the expectation ( $\rightarrow$  I/1.12.1) of the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  with  $\mu_0 = 0$  against  $m_0$  is

$$\langle \text{LBF}_{10} \rangle = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{\lambda_n - \lambda_0}{\lambda_n} \right) \quad (3)$$

where  $\lambda_n$  is the posterior precision for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7).

**Proof:** The log Bayes factor for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.10) is

$$\text{LBF}_{10} = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \quad (4)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/1.2.7)

$$\begin{aligned}\mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n\end{aligned}\tag{5}$$

with the sample mean ( $\rightarrow$  I/1.12.2)  $\bar{y}$  and the inverse variance or precision ( $\rightarrow$  I/1.13.13)  $\tau = 1/\sigma^2$ . Plugging  $\mu_n$  from (5) into (4), we obtain:

$$\begin{aligned}\text{LBF}_{10} &= \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left( \lambda_0 \mu_0^2 - \lambda_n \frac{(\lambda_0 \mu_0 + \tau n \bar{y})^2}{\lambda_n^2} \right) \\ &= \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left( \lambda_0 \mu_0^2 - \frac{1}{\lambda_n} (\lambda_0^2 \mu_0^2 - 2\tau n \lambda_0 \mu_0 \bar{y} + \tau^2 (n \bar{y})^2) \right)\end{aligned}\tag{6}$$

Because  $m_1$  uses a zero-mean prior distribution ( $\rightarrow$  I/5.1.3) with prior mean ( $\rightarrow$  I/1.12.1)  $\mu_0 = 0$  per construction, the log Bayes factor simplifies to:

$$\text{LBF}_{10} = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{\tau^2 (n \bar{y})^2}{\lambda_n} \right).\tag{7}$$

From (1), we know that the data are distributed as  $y_i \sim \mathcal{N}(\mu, \sigma^2)$ , such that we can derive the expectation ( $\rightarrow$  I/1.12.1) of  $(n \bar{y})^2$  as follows:

$$\begin{aligned}\langle (n \bar{y})^2 \rangle &= \left\langle \sum_{i=1}^n \sum_{j=1}^n y_i y_j \right\rangle = \langle n y_i^2 + (n^2 - n) [y_i y_j]_{i \neq j} \rangle \\ &= n(\mu^2 + \sigma^2) + (n^2 - n) \mu^2 \\ &= n^2 \mu^2 + n \sigma^2.\end{aligned}\tag{8}$$

Applying this expected value ( $\rightarrow$  I/1.12.1) to (7), the expected LBF emerges as:

$$\begin{aligned}\langle \text{LBF}_{10} \rangle &= \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{\tau^2 (n^2 \mu^2 + n \sigma^2)}{\lambda_n} \right) \\ &= \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{(\tau n \mu)^2 + \tau n}{\lambda_n} \right)\end{aligned}\tag{9}$$

Under the null hypothesis ( $\rightarrow$  I/4.3.2) that  $m_0$  generated the data, the unknown mean is  $\mu = 0$ , such that the log Bayes factor further simplifies to:

$$\langle \text{LBF}_{10} \rangle = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{\tau n}{\lambda_n} \right).\tag{10}$$

Finally, plugging  $\lambda_n$  from (5) into (10), we obtain:

$$\langle \text{LBF}_{10} \rangle = \frac{1}{2} \log \left( \frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left( \frac{\lambda_n - \lambda_0}{\lambda_n} \right).\tag{11}$$

■



## 1.2.12 Cross-validated log model evidence

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu = 0 \\ m_1 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) . \end{aligned} \quad (2)$$

Then, the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are

$$\begin{aligned} \text{cvLME}(m_0) &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} (\tau y^T y) \\ \text{cvLME}(m_1) &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \left[ y^T y + \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] \end{aligned} \quad (3)$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2),  $\tau = 1/\sigma^2$  is the inverse variance or precision ( $\rightarrow$  I/1.13.13),  $y_1^{(i)}$  are the training data in the  $i$ -th cross-validation fold with  $n_1$  data points and  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9).

**Proof:** For evaluation of the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME), we assume that  $n$  data points are divided into  $S \mid n$  data subsets without remainder. Then, the number of training data points  $n_1$  and test data points  $n_2$  are given by

$$\begin{aligned} n &= n_1 + n_2 \\ n_1 &= \frac{S-1}{S} n \\ n_2 &= \frac{1}{S} n , \end{aligned} \quad (4)$$

such that training data  $y_1$  and test data  $y_2$  in the  $i$ -th cross-validation fold are

$$\begin{aligned} y &= \{y_1, \dots, y_n\} \\ y_1^{(i)} &= \{x \in y \mid x \notin y_2^{(i)}\} = y \setminus y_2^{(i)} \\ y_2^{(i)} &= \{y_{(i-1) \cdot n_2 + 1}, \dots, y_{i \cdot n_2}\} . \end{aligned} \quad (5)$$

First, we consider the null model  $m_0$  assuming  $\mu = 0$ . Because this model has no free parameter, nothing is estimated from the training data and the assumed parameter value is applied to the test data. Consequently, the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) is equal to the log-likelihood function ( $\rightarrow$  III/1.2.2) of the test data at  $\mu = 0$ :

$$\text{oosLME}_i(m_0) = \log p\left(y_2^{(i)} \mid \mu = 0\right) = \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} \left[ \tau y_2^{(i)\text{T}} y_2^{(i)} \right] . \quad (6)$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, such that the cvLME of  $m_0$  is:

$$\begin{aligned} \text{cvLME}(m_0) &= \sum_{i=1}^S \text{oosLME}_i(m_0) \\ &= \sum_{i=1}^S \left( \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} \left[ \tau y_2^{(i)\text{T}} y_2^{(i)} \right] \right) \\ &= \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} [\tau y^{\text{T}} y] . \end{aligned} \quad (7)$$

Next, we have a look at the alternative  $m_1$  assuming  $\mu \neq 0$ . First, the training data  $y_1^{(i)}$  are analyzed using a non-informative prior distribution ( $\rightarrow$  I/5.2.3) and applying the posterior distribution for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7):

$$\begin{aligned} \mu_0^{(1)} &= 0 \\ \lambda_0^{(1)} &= 0 \\ \mu_n^{(1)} &= \frac{\tau n_1 \bar{y}_1^{(i)} + \lambda_0^{(1)} \mu_0^{(1)}}{\tau n_1 + \lambda_0^{(1)}} = \bar{y}_1^{(i)} \\ \lambda_n^{(1)} &= \tau n_1 + \lambda_0^{(1)} = \tau n_1 . \end{aligned} \quad (8)$$

This results in a posterior characterized by  $\mu_n^{(1)}$  and  $\lambda_n^{(1)}$ . Then, the test data  $y_2^{(i)}$  are analyzed using this posterior as an informative prior distribution ( $\rightarrow$  I/5.2.3), again applying the posterior distribution for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.7):

$$\begin{aligned} \mu_0^{(2)} &= \mu_n^{(1)} = \bar{y}_1^{(i)} \\ \lambda_0^{(2)} &= \lambda_n^{(1)} = \tau n_1 \\ \mu_n^{(2)} &= \frac{\tau n_2 \bar{y}_2^{(i)} + \lambda_0^{(2)} \mu_0^{(2)}}{\tau n_2 + \lambda_0^{(2)}} = \bar{y} \\ \lambda_n^{(2)} &= \tau n_2 + \lambda_0^{(2)} = \tau n . \end{aligned} \quad (9)$$

In the test data, we now have a prior characterized by  $\mu_0^{(2)}/\lambda_0^{(2)}$  and a posterior characterized  $\mu_n^{(2)}/\lambda_n^{(2)}$ . Applying the log model evidence for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.8), the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) therefore follows as

$$\begin{aligned} \text{oosLME}_i(m_1) &= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{\lambda_0^{(2)}}{\lambda_n^{(2)}}\right) - \frac{1}{2} \left[ \tau y_2^{(i)\text{T}} y_2^{(i)} + \lambda_0^{(2)} \mu_0^{(2)2} - \lambda_n^{(2)} \mu_n^{(2)2} \right] \\ &= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{n_1}{n}\right) - \frac{1}{2} \left[ \tau y_2^{(i)\text{T}} y_2^{(i)} + \frac{\tau}{n_1} \left( n_1 \bar{y}_1^{(i)} \right)^2 - \frac{\tau}{n} (n \bar{y})^2 \right] . \end{aligned} \quad (10)$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, the cvLME of  $m_1$  becomes:

$$\begin{aligned}
 \text{cvLME}(m_1) &= \sum_{i=1}^S \text{oosLME}_i(m_1) \\
 &= \sum_{i=1}^S \left( \frac{n_2}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left( \frac{n_1}{n} \right) - \frac{1}{2} \left[ \tau y_2^{(i)\text{T}} y_2^{(i)} + \frac{\tau}{n_1} \left( n_1 \bar{y}_1^{(i)} \right)^2 - \frac{\tau}{n} (n\bar{y})^2 \right] \right) \\
 &= \frac{S \cdot n_2}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left( \frac{n_1}{n} \right) - \frac{\tau}{2} \sum_{i=1}^S \left[ y_2^{(i)\text{T}} y_2^{(i)} + \frac{\left( n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right] \\
 &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \left[ y^{\text{T}} y + \sum_{i=1}^S \left( \frac{\left( n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right].
 \end{aligned} \tag{11}$$

Together, (7) and (11) conform to the results given in (3). ■

### 1.2.13 Cross-validated log Bayes factor

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned}
 m_0 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu = 0 \\
 m_1 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}).
 \end{aligned} \tag{2}$$

Then, the cross-validated ( $\rightarrow$  IV/3.1.9) log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_0$  is

$$\text{cvLBF}_{10} = \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{\left( n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \tag{3}$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2),  $\tau = 1/\sigma^2$  is the inverse variance or precision ( $\rightarrow$  I/1.13.13),  $y_1^{(i)}$  are the training data in the  $i$ -th cross-validation fold with  $n_1$  data points and  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9).

**Proof:** The relationship between log Bayes factor and log model evidences ( $\rightarrow$  IV/3.3.8) also holds for cross-validated log bayes factor ( $\rightarrow$  IV/3.3.6) (cvLBF) and cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME):

$$\text{cvLBF}_{12} = \text{cvLME}(m_1) - \text{cvLME}(m_2) . \quad (4)$$

The cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are given by ( $\rightarrow$  III/1.2.12)

$$\begin{aligned} \text{cvLME}(m_0) &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} (\tau y^T y) \\ \text{cvLME}(m_1) &= \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \left[ y^T y + \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] . \end{aligned} \quad (5)$$

Subtracting the two cvLMEs from each other, the cvLBF emerges as

$$\begin{aligned} \text{cvLBF}_{10} &= \text{cvLME}(m_1) - \text{LME}(m_0) \\ &= \left( \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \left[ y^T y + \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] \right) \\ &\quad - \left( \frac{n}{2} \log \left( \frac{\tau}{2\pi} \right) - \frac{1}{2} (\tau y^T y) \right) \\ &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) . \end{aligned} \quad (6)$$

■

#### 1.2.14 Expectation of cross-validated log Bayes factor

**Theorem:** Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

be a univariate Gaussian data set ( $\rightarrow$  III/1.2.1) with unknown mean  $\mu$  and known variance  $\sigma^2$ . Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $\mu$  is zero (null model ( $\rightarrow$  I/4.3.2)), the other imposing a normal distribution ( $\rightarrow$  III/1.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $\mu$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu = 0 \\ m_1 : y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) . \end{aligned} \quad (2)$$

Then, the expectation ( $\rightarrow$  I/1.12.1) of the cross-validated ( $\rightarrow$  IV/3.1.9) log Bayes factor ( $\rightarrow$  IV/3.3.6) (cvLBF) in favor of  $m_1$  against  $m_0$  is

$$\langle \text{cvLBF}_{10} \rangle = \frac{S}{2} \log \left( \frac{S-1}{S} \right) + \frac{1}{2} [\tau n \mu^2] \quad (3)$$

where  $\tau = 1/\sigma^2$  is the inverse variance or precision ( $\rightarrow$  I/1.13.13) and  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9).

**Proof:** The cross-validated log Bayes factor for the univariate Gaussian with known variance ( $\rightarrow$  III/1.2.13) is

$$\text{cvLBF}_{10} = \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \quad (4)$$

From (1), we know that the data are distributed as  $y_i \sim \mathcal{N}(\mu, \sigma^2)$ , such that we can derive the expectation ( $\rightarrow$  I/1.12.1) of  $(n\bar{y})^2$  and  $(n_1 \bar{y}_1^{(i)})^2$  as follows:

$$\begin{aligned} \langle (n\bar{y})^2 \rangle &= \left\langle \sum_{i=1}^n \sum_{j=1}^n y_i y_j \right\rangle = \langle n y_i^2 + (n^2 - n) [y_i y_j]_{i \neq j} \rangle \\ &= n(\mu^2 + \sigma^2) + (n^2 - n)\mu^2 \\ &= n^2 \mu^2 + n\sigma^2. \end{aligned} \quad (5)$$

Applying this expected value ( $\rightarrow$  I/1.12.1) to (4), the expected cvLBF emerges as:

$$\begin{aligned} \langle \text{cvLBF}_{10} \rangle &= \left\langle \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right\rangle \\ &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{\langle (n_1 \bar{y}_1^{(i)})^2 \rangle}{n_1} - \frac{\langle (n\bar{y})^2 \rangle}{n} \right) \\ &\stackrel{(5)}{=} \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left( \frac{n_1^2 \mu^2 + n_1 \sigma^2}{n_1} - \frac{n^2 \mu^2 + n \sigma^2}{n} \right) \\ &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S ([n_1 \mu^2 + \sigma^2] - [n \mu^2 + \sigma^2]) \\ &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S (n_1 - n) \mu^2 \end{aligned} \quad (6)$$

Because it holds that ( $\rightarrow$  III/1.2.12)  $n_1 + n_2 = n$  and  $n_2 = n/S$ , we finally have:

$$\begin{aligned} \langle \text{cvLBF}_{10} \rangle &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S (-n_2) \mu^2 \\ &= \frac{S}{2} \log \left( \frac{S-1}{S} \right) + \frac{1}{2} [\tau n \mu^2]. \end{aligned} \quad (7)$$

■

## 1.3 Analysis of variance

### 1.3.1 One-way ANOVA

**Definition:** Consider measurements  $y_{ij} \in \mathbb{R}$  from distinct objects  $j = 1, \dots, n_i$  in separate groups  $i = 1, \dots, k$ .

Then, in one-way analysis of variance (ANOVA), these measurements are assumed to come from normal distributions ( $\rightarrow$  II/3.2.1)

$$y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for all } i = 1, \dots, k \quad \text{and } j = 1, \dots, n_i \quad (1)$$

where

- $\mu_i$  is the expected value ( $\rightarrow$  I/1.12.1) in group  $i$  and
- $\sigma^2$  is the common variance ( $\rightarrow$  I/1.13.1) across groups.

Alternatively, the model may be written as

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned} \quad (2)$$

where  $\varepsilon_{ij}$  is the error term ( $\rightarrow$  III/1.4.1) belonging to observation  $j$  in category  $i$  and  $\varepsilon_{ij}$  are the independent and identically distributed ( $\rightarrow$  I/1.2.8).

**Sources:**

- Bortz, Jürgen (1977): “Einfaktorielle Varianzanalyse”; in: *Lehrbuch der Statistik. Für Sozialwissenschaftler*, ch. 12.1, pp. 528ff.; URL: <https://books.google.de/books?id=INCyBgAAQBAJ>.
- Denziloe (2018): “Derive the distribution of the ANOVA F-statistic under the alternative hypothesis”; in: *StackExchange Cross Validated*, retrieved on 2022-11-06; URL: <https://stats.stackexchange.com/questions/355594/derive-the-distribution-of-the-anova-f-statistic-under-the-alternative-hypothesis>.

### 1.3.2 Treatment sum of squares

**Definition:** Let there be an analysis of variance (ANOVA) model with one ( $\rightarrow$  III/1.3.1), two ( $\rightarrow$  III/1.3.8) or multiple factors influencing the measured data  $y$  (here, using the reparametrized version ( $\rightarrow$  III/1.3.7) of one-way ANOVA ( $\rightarrow$  III/1.3.1)):

$$y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) . \quad (1)$$

Then, the treatment sum of squares is defined as the explained sum of squares ( $\rightarrow$  III/1.5.8) (ESS) for each main effect, i.e. as the sum of squared deviations of the average for each level of the factor, from the average across all observations:

$$\text{SS}_{\text{treat}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 . \quad (2)$$

Here,  $\bar{y}_i$  is the mean for the  $i$ -th level of the factor (out of  $k$  levels), computed from  $n_i$  values  $y_{ij}$ , and  $\bar{y}$  is the mean across all values  $y_{ij}$ .

**Sources:**

- Wikipedia (2022): “Analysis of variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: [https://en.wikipedia.org/wiki/Analysis\\_of\\_variance#Partitioning\\_of\\_the\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares).

### 1.3.3 Ordinary least squares for one-way ANOVA

**Theorem:** Given the one-way analysis of variance ( $\rightarrow$  III/1.3.1) assumption

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, k, \quad j = 1, \dots, n_i, \quad (1)$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\mu}_i = \bar{y}_i \quad (2)$$

where  $\bar{y}_i$  is the sample mean ( $\rightarrow$  I/1.12.2) of all observations in group ( $\rightarrow$  III/1.3.1)  $i$ :

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}. \quad (3)$$

**Proof:** The residual sum of squares ( $\rightarrow$  III/1.5.9) for this model is

$$\text{RSS}(\mu) = \sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \quad (4)$$

and the derivatives of RSS with respect to  $\mu_i$  are

$$\begin{aligned} \frac{d\text{RSS}(\mu)}{d\mu_i} &= \sum_{j=1}^{n_i} \frac{d}{d\mu_i} (y_{ij} - \mu_i)^2 \\ &= \sum_{j=1}^{n_i} 2(y_{ij} - \mu_i)(-1) \\ &= 2 \sum_{j=1}^{n_i} (\mu_i - y_{ij}) \\ &= 2n_i\mu_i - 2 \sum_{j=1}^{n_i} y_{ij} \quad \text{for } i = 1, \dots, k. \end{aligned} \quad (5)$$

Setting these derivatives to zero, we obtain the estimates of  $\mu_i$ :

$$\begin{aligned} 0 &= 2n_i\hat{\mu}_i - 2 \sum_{j=1}^{n_i} y_{ij} \\ \hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \text{for } i = 1, \dots, k. \end{aligned} \quad (6)$$

■

### 1.3.4 Sums of squares in one-way ANOVA

**Theorem:** Given one-way analysis of variance ( $\rightarrow$  III/1.3.1),

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

sums of squares can be partitioned as follows

$$\text{SS}_{\text{tot}} = \text{SS}_{\text{treat}} + \text{SS}_{\text{res}} \quad (2)$$

where  $\text{SS}_{\text{tot}}$  is the total sum of squares ( $\rightarrow$  III/1.5.7),  $\text{SS}_{\text{treat}}$  is the treatment sum of squares ( $\rightarrow$  III/1.3.2) (equivalent to explained sum of squares ( $\rightarrow$  III/1.5.8)) and  $\text{SS}_{\text{res}}$  is the residual sum of squares ( $\rightarrow$  III/1.5.9).

**Proof:** The total sum of squares ( $\rightarrow$  III/1.5.7) for one-way ANOVA ( $\rightarrow$  III/1.3.1) is given by

$$\text{SS}_{\text{tot}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad (3)$$

where  $\bar{y}$  is the mean across all values  $y_{ij}$ . This can be rewritten as

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})] \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + 2 \sum_{i=1}^k (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i). \end{aligned} \quad (4)$$

Note that the following sum is zero

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \bar{y}_i = \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad (5)$$

so that the sum in (4) reduces to

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \quad (6)$$

With the treatment sum of squares ( $\rightarrow$  III/1.3.2) for one-way ANOVA ( $\rightarrow$  III/1.3.1)

$$\text{SS}_{\text{treat}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \quad (7)$$

and the residual sum of squares ( $\rightarrow$  III/1.5.9) for one-way ANOVA ( $\rightarrow$  III/1.3.1)

$$\text{SS}_{\text{res}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad (8)$$



we finally have:

$$SS_{\text{tot}} = SS_{\text{treat}} + SS_{\text{res}} . \quad (9)$$

■

#### Sources:

- Wikipedia (2022): “Analysis of variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: [https://en.wikipedia.org/wiki/Analysis\\_of\\_variance#Partitioning\\_of\\_the\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares).

### 1.3.5 F-test for main effect in one-way ANOVA

**Theorem:** Assume the one-way analysis of variance ( $\rightarrow$  III/1.3.1) model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, k, \quad j = 1, \dots, n_i . \quad (1)$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \quad (2)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F \sim F(k-1, n-k) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$\begin{aligned} H_0 : \mu_1 = \dots = \mu_k \\ H_1 : \mu_i \neq \mu_j \quad \text{for at least one } i, j \in \{1, \dots, k\}, i \neq j . \end{aligned} \quad (4)$$

**Proof:** Denote sample sizes as

$$\begin{aligned} n_i &= \text{number of samples in category } i \\ n &= \sum_{i=1}^k n_i \end{aligned} \quad (5)$$

and denote sample means as

$$\begin{aligned} \bar{y}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} . \end{aligned} \quad (6)$$

Let  $\mu$  be the common mean ( $\rightarrow$  I/1.12.1) according to  $H_0$  given by (4), i.e.  $\mu_1 = \dots = \mu_k = \mu$ . Under this null hypothesis, we have:

$$y_{ij} \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for all } i = 1, \dots, k, j = 1, \dots, n_i. \quad (7)$$

Thus, the random variable ( $\rightarrow$  I/1.2.2)  $U_{ij} = (y_{ij} - \mu)/\sigma$  follows a standard normal distribution ( $\rightarrow$  II/3.2.4)

$$U_{ij} = \frac{y_{ij} - \mu}{\sigma} \sim \mathcal{N}(0, 1). \quad (8)$$

Now consider the following sum:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{y_{ij} - \mu}{\sigma} \right)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) + (\bar{y} - \mu))^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + (\bar{y} - \mu)^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + 2(y_{ij} - \bar{y}_i)(\bar{y} - \mu) + 2(\bar{y}_i - \bar{y})(\bar{y} - \mu)] \end{aligned} \quad (9)$$

Because the following sum over  $j$  is zero for all  $i$

$$\begin{aligned} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) &= \sum_{j=1}^{n_i} y_{ij} - n_i \bar{y}_i \\ &= \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \\ &= 0, \quad i = 1, \dots, k \end{aligned} \quad (10)$$

and the following sum over  $i$  and  $j$  is also zero

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}) &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}) \\ &= \sum_{i=1}^k n_i \bar{y}_i - \bar{y} \sum_{i=1}^k n_i \\ &= \sum_{i=1}^k n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \bar{y} \cdot \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \\ &= 0, \end{aligned} \quad (11)$$

non-square products in (9) disappear and the sum reduces to

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left[ \left( \frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 + \left( \frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 + \left( \frac{\bar{y} - \mu}{\sigma} \right)^2 \right] \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{\bar{y} - \mu}{\sigma} \right)^2. \end{aligned} \quad (12)$$

Cochran's theorem states that, if a sum of squared standard normal ( $\rightarrow$  II/3.2.3) random variables ( $\rightarrow$  I/1.2.2) can be written as a sum of squared forms

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{j=1}^m Q_j \quad \text{where} \quad Q_j = U^T B^{(j)} U \\ &\quad \text{with} \quad \sum_{j=1}^m B^{(j)} = I_n \\ &\quad \text{and} \quad r_j = \text{rank}(B^{(j)}) , \end{aligned} \tag{13}$$

then the terms  $Q_j$  are independent ( $\rightarrow$  I/1.3.6) and each term  $Q_j$  follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $r_j$  degrees of freedom:

$$Q_j \sim \chi^2(r_j), \quad j = 1, \dots, m . \tag{14}$$

Let  $U$  be the  $n$ -dimensional column vector of all observations

$$U = \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix} \tag{15}$$

where the group-wise  $n_i \times 1$  column vectors are

$$u_1 = \begin{bmatrix} (y_{1,1} - \mu)/\sigma \\ \vdots \\ (y_{1,n_1} - \mu)/\sigma \end{bmatrix}, \quad \dots, \quad u_k = \begin{bmatrix} (y_{k,1} - \mu)/\sigma \\ \vdots \\ (y_{k,n_k} - \mu)/\sigma \end{bmatrix} . \tag{16}$$

Then, we observe that the sum in (12) can be represented in the form of (13) using the matrices

$$\begin{aligned} B^{(1)} &= I_n - \text{diag} \left( \frac{1}{n_1} J_{n_1}, \dots, \frac{1}{n_k} J_{n_k} \right) \\ B^{(2)} &= \text{diag} \left( \frac{1}{n_1} J_{n_1}, \dots, \frac{1}{n_k} J_{n_k} \right) - \frac{1}{n} J_n \\ B^{(3)} &= \frac{1}{n} J_n \end{aligned} \tag{17}$$

where  $J_n$  is an  $n \times n$  matrix of ones and  $\text{diag}(A_1, \dots, A_n)$  denotes a block-diagonal matrix composed of  $A_1, \dots, A_n$ . We observe that those matrices satisfy

$$\sum_{i=1}^k \sum_{j=1}^{n_i} U_{ij}^2 = Q_1 + Q_2 + Q_3 = U^T B^{(1)} U + U^T B^{(2)} U + U^T B^{(3)} U \tag{18}$$

as well as

$$B^{(1)} + B^{(2)} + B^{(3)} = I_n \tag{19}$$

and their ranks are:

$$\begin{aligned}
\text{rank}(B^{(1)}) &= n - k \\
\text{rank}(B^{(2)}) &= k - 1 \\
\text{rank}(B^{(3)}) &= 1 .
\end{aligned} \tag{20}$$

Let's write down the explained sum of squares ( $\rightarrow$  III/1.5.8) and the residual sum of squares ( $\rightarrow$  III/1.5.9) for one-way analysis of variance ( $\rightarrow$  III/1.3.1) as

$$\begin{aligned}
\text{ESS} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \\
\text{RSS} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 .
\end{aligned} \tag{21}$$

Then, using (12), (13), (14), (17) and (20), we find that

$$\begin{aligned}
\frac{\text{ESS}}{\sigma^2} &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 = Q_2 = U^T B^{(2)} U \sim \chi^2(k - 1) \\
\frac{\text{RSS}}{\sigma^2} &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 = Q_1 = U^T B^{(1)} U \sim \chi^2(n - k) .
\end{aligned} \tag{22}$$

Because  $\text{ESS}/\sigma^2$  and  $\text{RSS}/\sigma^2$  are also independent by (14), the F-statistic from (2) is equal to the ratio of two independent chi-squared distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}
F &= \frac{(\text{ESS}/\sigma^2)/(k - 1)}{(\text{RSS}/\sigma^2)/(n - k)} \\
&= \frac{\text{ESS}/(k - 1)}{\text{RSS}/(n - k)} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}
\end{aligned} \tag{23}$$

which, by definition of the F-distribution ( $\rightarrow$  II/3.8.1), is distributed as

$$F \sim F(k - 1, n - k) \tag{24}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the main effect. ■

#### Sources:

- Denziloe (2018): “Derive the distribution of the ANOVA F-statistic under the alternative hypothesis”; in: *StackExchange Cross Validated*, retrieved on 2022-11-06; URL: <https://stats.stackexchange.com/questions/355594/derive-the-distribution-of-the-anova-f-statistic-under-the-alternative-hypothesis>.

### 1.3.6 F-statistic in terms of OLS estimates

**Theorem:** Given the one-way analysis of variance ( $\rightarrow$  III/1.3.1) assumption

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (1)$$

1) the F-statistic for the main effect ( $\rightarrow$  III/1.3.5) can be expressed in terms of ordinary least squares parameter estimates ( $\rightarrow$  III/1.3.3) as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\hat{\mu}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2} \quad (2)$$

2) or, when using the reparametrized version of one-way ANOVA ( $\rightarrow$  III/1.3.7), the F-statistic can be expressed as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\delta}_i)^2} . \quad (3)$$

**Proof:** The F-statistic for the main effect in one-way ANOVA ( $\rightarrow$  III/1.3.5) is given in terms of the sample means ( $\rightarrow$  I/1.12.2) as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \quad (4)$$

where  $\bar{y}_i$  is the average of all values  $y_{ij}$  from category  $i$  and  $\bar{y}$  is the grand mean of all values  $y_{ij}$  from all categories  $i = 1, \dots, k$ .

1) The ordinary least squares estimates for one-way ANOVA ( $\rightarrow$  III/1.3.3) are

$$\hat{\mu}_i = \bar{y}_i , \quad (5)$$

such that

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\hat{\mu}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2} . \quad (6)$$

2) The OLS estimates for reparametrized one-way ANOVA ( $\rightarrow$  III/1.3.7) are

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ \hat{\delta}_i &= \bar{y}_i - \bar{y} , \end{aligned} \quad (7)$$

such that

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\delta}_i)^2} . \quad (8)$$

■

### 1.3.7 Reparametrization of one-way ANOVA

**Theorem:** The one-way analysis of variance ( $\rightarrow$  III/1.3.1) model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

can be rewritten using parameters  $\mu$  and  $\delta_i$  instead of  $\mu_i$

$$y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (2)$$

with the constraint

$$\sum_{i=1}^k \frac{n_i}{n} \delta_i = 0, \quad (3)$$

in which case

1) the model parameters are related to each other as

$$\delta_i = \mu_i - \mu, \quad i = 1, \dots, k; \quad (4)$$

2) the ordinary least squares estimates ( $\rightarrow$  III/1.3.3) are given by

$$\hat{\delta}_i = \bar{y}_i - \bar{y} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}; \quad (5)$$

3) the following sum of squares ( $\rightarrow$  III/1.3.4) is chi-square distributed ( $\rightarrow$  II/3.7.1)

$$\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2 \sim \chi^2(k-1); \quad (6)$$

4) and the following test statistic ( $\rightarrow$  I/4.3.5) is F-distributed ( $\rightarrow$  II/3.8.1)

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \sim F(k-1, n-k) \quad (7)$$

under the null hypothesis for the main effect ( $\rightarrow$  III/1.3.5)

$$H_0 : \delta_1 = \dots = \delta_k = 0. \quad (8)$$

**Proof:**

1) Equating (1) with (2), we get:

$$\begin{aligned} y_{ij} &= \mu + \delta_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} = y_{ij} \\ \mu + \delta_i &= \mu_i \\ \delta_i &= \mu_i - \mu. \end{aligned} \quad (9)$$

2) The residual sum of squares ( $\rightarrow$  III/1.5.9) for the reparametrized model is

$$\text{RSS}(\mu, \delta) = \sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ijk}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu - \delta_i)^2 \quad (10)$$

and the derivatives of RSS with respect to  $\mu$ ,  $\delta$  are

$$\begin{aligned}
 \frac{dRSS}{d\mu} &= \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{d}{d\mu} (y_{ij} - \mu - \delta_i)^2 \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} -2(y_{ij} - \mu - \delta_i) \\
 &= \sum_{i=1}^k \left( 2n_i\mu + 2n_i\delta_i - 2 \sum_{j=1}^{n_i} y_{ij} \right) \\
 &= 2n\mu + 2 \sum_{i=1}^k n_i\delta_i - 2 \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}
 \end{aligned} \tag{11}$$

$$\begin{aligned}
 \frac{dRSS}{d\delta_i} &= \sum_{j=1}^{n_i} \frac{d}{d\delta_i} (y_{ij} - \mu - \delta_i)^2 \\
 &= \sum_{j=1}^{n_i} -2(y_{ij} - \mu - \delta_i) \\
 &= 2n_i\mu + 2n_i\delta_i - 2 \sum_{j=1}^{n_i} y_{ij} .
 \end{aligned} \tag{12}$$

Setting these derivatives to zero, we obtain the estimates of  $\mu$  and  $\delta_i$ :

$$\begin{aligned}
 0 &= 2n\hat{\mu} + 2 \sum_{i=1}^k n_i\delta_i - 2 \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \\
 \hat{\mu} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^k \frac{n_i}{n} \delta_i \\
 &\stackrel{(3)}{=} \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \\
 &= \bar{y}
 \end{aligned} \tag{13}$$

$$\begin{aligned}
 0 &= 2n_i\hat{\mu} + 2n_i\hat{\delta}_i - 2 \sum_{j=1}^{n_i} y_{ij} \\
 \hat{\delta}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \hat{\mu} \\
 &\stackrel{(13)}{=} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \\
 &= \bar{y}_i - \bar{y} .
 \end{aligned} \tag{14}$$

3) Let  $U_{ij} = (y_{ij} - \mu - \delta_i)/\sigma$ , such that ( $\rightarrow$  II/3.2.4)  $U_{ij} \sim \mathcal{N}(0, 1)$  and consider the sum of all squared random variables ( $\rightarrow$  I/1.2.2)  $U_{ij}$ :

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \frac{y_{ij} - \mu - \delta_i}{\sigma} \right)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) - \delta_i + (\bar{y} - \mu)]^2. \end{aligned} \quad (15)$$

This square of sums, using a number of intermediate steps, can be developed ( $\rightarrow$  III/1.3.5) into a sum of squares:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} U_{ij}^2 &= \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + (\bar{y} - \mu)^2] \\ &= \frac{1}{\sigma^2} \left[ \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y} - \mu)^2 \right]. \end{aligned} \quad (16)$$

To this sum, Cochran's theorem for one-way analysis of variance can be applied ( $\rightarrow$  III/1.3.5), yielding the distributions:

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 &\sim \chi^2(n - k) \\ \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 &\stackrel{???}{=} \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2 \sim \chi^2(k - 1). \end{aligned} \quad (17)$$

4) The ratio of two chi-square distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2), divided by their degrees of freedom ( $\rightarrow$  I/1.5.12), is defined to be F-distributed ( $\rightarrow$  II/3.8.1), so that

$$\begin{aligned} F &= \frac{\left( \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2 \right) / (k - 1)}{\left( \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right) / (n - k)} \\ &= \frac{\frac{1}{k-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\ &= \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\hat{\delta}_i - \delta_i)^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\ &\stackrel{(8)}{=} \frac{\frac{1}{k-1} \sum_{i=1}^k n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \end{aligned} \quad (18)$$

follows the F-distribution

$$F \sim F(k - 1, n - k) \quad (19)$$

under the null hypothesis.



**Sources:**

- Wikipedia (2022): “Analysis of variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: [https://en.wikipedia.org/wiki/Analysis\\_of\\_variance#For\\_a\\_single\\_factor](https://en.wikipedia.org/wiki/Analysis_of_variance#For_a_single_factor).

**1.3.8 Two-way ANOVA**

**Definition:** Let there be two factors  $A$  and  $B$  with levels  $i = 1, \dots, a$  and  $j = 1, \dots, b$  that are used to group measurements  $y_{ijk} \in \mathbb{R}$  from distinct objects  $k = 1, \dots, n_{ij}$  into  $a \cdot b$  categories  $(i, j) \in \{1, \dots, a\} \times \{1, \dots, b\}$ .

Then, in two-way analysis of variance (ANOVA), these measurements are assumed to come from normal distributions ( $\rightarrow$  II/3.2.1)

$$y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2) \quad \text{for all } i = 1, \dots, a, \quad j = 1, \dots, b, \quad \text{and } k = 1, \dots, n_{ij} \quad (1)$$

with

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (2)$$

where

- $\mu$  is called the “grand mean”;
- $\alpha_i$  is the additive “main effect” of the  $i$ -th level of factor  $A$ ;
- $\beta_j$  is the additive “main effect” of the  $j$ -th level of factor  $B$ ;
- $\gamma_{ij}$  is the non-additive “interaction effect” of category  $(i, j)$ ;
- $\mu_{ij}$  is the expected value ( $\rightarrow$  I/1.12.1) in category  $(i, j)$ ; and
- $\sigma^2$  is common variance ( $\rightarrow$  I/1.13.1) across all categories.

Alternatively, the model may be written as

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \quad (3)$$

where  $\varepsilon_{ijk}$  is the error term ( $\rightarrow$  III/1.4.1) corresponding to observation  $k$  belonging to the  $i$ -th level of  $A$  and the  $j$ -th level of  $B$ .

As the two-way ANOVA model is underdetermined, the parameters of the model are additionally subject to the constraints

$$\begin{aligned} \sum_{i=1}^a w_{ij} \alpha_i &= 0 \quad \text{for all } j = 1, \dots, b \\ \sum_{j=1}^b w_{ij} \beta_j &= 0 \quad \text{for all } i = 1, \dots, a \\ \sum_{i=1}^a w_{ij} \gamma_{ij} &= 0 \quad \text{for all } j = 1, \dots, b \\ \sum_{j=1}^b w_{ij} \gamma_{ij} &= 0 \quad \text{for all } i = 1, \dots, a \end{aligned} \quad (4)$$

where the weights are  $w_{ij} = n_{ij}/n$  and the total sample size ( $\rightarrow$  I/1.1.7) is  $n = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$ .

**Sources:**

- Bortz, Jürgen (1977): “Zwei- und mehrfaktorielle Varianzanalyse”; in: *Lehrbuch der Statistik. Für Sozialwissenschaftler*, ch. 12.2, pp. 538ff.; URL: <https://books.google.de/books?id=INCyBgAAQBAJ>.
- ttd (2021): “Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J”; in: *StackExchange CrossValidated*, retrieved on 2022-11-06; URL: <https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1j-1-under-the-null-hypothesis>.

### 1.3.9 Interaction sum of squares

**Definition:** Let there be an analysis of variance (ANOVA) model with two ( $\rightarrow$  III/1.3.8) or more factors influencing the measured data  $y$  (here, using the standard formulation ( $\rightarrow$  III/1.3.11) of two-way ANOVA ( $\rightarrow$  III/1.3.8)):

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \quad (1)$$

Then, the interaction sum of squares is defined as the explained sum of squares ( $\rightarrow$  III/1.5.8) (ESS) for each interaction, i.e. as the sum of squared deviations of the average for each cell from the average across all observations, controlling for the treatment sums of squares ( $\rightarrow$  III/1.3.2) of the corresponding factors:

$$\begin{aligned} \text{SS}_{A \times B} &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet}] - [\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - [\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}])^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2. \end{aligned} \quad (2)$$

Here,  $\bar{y}_{ij\bullet}$  is the mean for the  $(i, j)$ -th cell (out of  $a \times b$  cells), computed from  $n_{ij}$  values  $y_{ijk}$ ,  $\bar{y}_{i\bullet\bullet}$  and  $\bar{y}_{\bullet j\bullet}$  are the level means for the two factors and  $\bar{y}_{\bullet\bullet\bullet}$  is the mean across all values  $y_{ijk}$ .

**Sources:**

- Nandy, Siddhartha (2018): “Two-Way Analysis of Variance”; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: <https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf>.

### 1.3.10 Ordinary least squares for two-way ANOVA

**Theorem:** Given the two-way analysis of variance ( $\rightarrow$  III/1.3.8) assumption

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}, \end{aligned} \quad (1)$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) and satisfying the constraints for the model parameters ( $\rightarrow$  III/1.3.8) are given by

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\bullet\bullet\bullet} \\
\hat{\alpha}_i &= \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \\
\hat{\beta}_j &= \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet} \\
\hat{\gamma}_{ij} &= \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}
\end{aligned} \tag{2}$$

where  $\bar{y}_{\bullet\bullet\bullet}$ ,  $\bar{y}_{i\bullet\bullet}$ ,  $\bar{y}_{\bullet j\bullet}$  and  $\bar{y}_{ij\bullet}$  are the following sample means ( $\rightarrow$  I/1.12.2):

$$\begin{aligned}
\bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}
\end{aligned} \tag{3}$$

with the sample size ( $\rightarrow$  I/1.1.7) numbers

$$\begin{aligned}
n_{ij} &\text{ -- number of samples in category } (i, j) \\
n_{i\bullet} &= \sum_{j=1}^b n_{ij} \\
n_{\bullet j} &= \sum_{i=1}^a n_{ij} \\
n &= \sum_{i=1}^a \sum_{j=1}^b n_{ij} .
\end{aligned} \tag{4}$$

**Proof:** In two-way ANOVA, model parameters are subject to the constraints ( $\rightarrow$  III/1.3.8)

$$\begin{aligned}
\sum_{i=1}^a w_{ij} \alpha_i &= 0 \quad \text{for all } j = 1, \dots, b \\
\sum_{j=1}^b w_{ij} \beta_j &= 0 \quad \text{for all } i = 1, \dots, a \\
\sum_{i=1}^a w_{ij} \gamma_{ij} &= 0 \quad \text{for all } j = 1, \dots, b \\
\sum_{j=1}^b w_{ij} \gamma_{ij} &= 0 \quad \text{for all } i = 1, \dots, a
\end{aligned} \tag{5}$$

where  $w_{ij} = n_{ij}/n$ . The residual sum of squares ( $\rightarrow$  III/1.5.9) for this model is

$$\text{RSS}(\mu, \alpha, \beta, \gamma) = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} \varepsilon_{ijk}^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \quad (6)$$

and the derivatives of RSS with respect to  $\mu$ ,  $\alpha$ ,  $\beta$  and  $\gamma$  are

$$\begin{aligned} \frac{d\text{RSS}}{d\mu} &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} \frac{d}{d\mu} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} -2(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}) \\ &= \sum_{i=1}^a \sum_{j=1}^b \left( 2n_{ij}\mu + 2n_{ij}(\alpha_i + \beta_j + \gamma_{ij}) - 2 \sum_{k=1}^{n_{ij}} y_{ijk} \right) \\ &= 2n\mu + 2 \left( \sum_{i=1}^a n_{i\bullet} \alpha_i + \sum_{j=1}^b n_{\bullet j} \beta_j + \sum_{i=1}^a \sum_{j=1}^b n_{ij} \gamma_{ij} \right) - 2 \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \end{aligned} \quad (7)$$

$$\begin{aligned} \frac{d\text{RSS}}{d\alpha_i} &= \sum_{j=1}^b \sum_{k=1}^{n_{ij}} \frac{d}{d\alpha_i} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \\ &= \sum_{j=1}^b \sum_{k=1}^{n_{ij}} -2(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}) \\ &= 2n_{i\bullet} \mu + 2n_{i\bullet} \alpha_i + 2 \left( \sum_{j=1}^b n_{ij} \beta_j + \sum_{j=1}^b n_{ij} \gamma_{ij} \right) - 2 \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \end{aligned} \quad (8)$$

$$\begin{aligned} \frac{d\text{RSS}}{d\beta_j} &= \sum_{i=1}^a \sum_{k=1}^{n_{ij}} \frac{d}{d\beta_j} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \\ &= \sum_{i=1}^a \sum_{k=1}^{n_{ij}} -2(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}) \\ &= 2n_{\bullet j} \mu + 2n_{\bullet j} \beta_j + 2 \left( \sum_{i=1}^a n_{ij} \alpha_i + \sum_{i=1}^a n_{ij} \gamma_{ij} \right) - 2 \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} \end{aligned} \quad (9)$$

$$\begin{aligned} \frac{d\text{RSS}}{d\gamma_{ij}} &= \sum_{k=1}^{n_{ij}} \frac{d}{d\gamma_{ij}} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \\ &= \sum_{k=1}^{n_{ij}} -2(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}) \\ &= 2n_{ij}(\mu + \alpha_i + \beta_j + \gamma_{ij}) - 2 \sum_{k=1}^{n_{ij}} y_{ijk} . \end{aligned} \quad (10)$$

Setting these derivatives to zero, we obtain the estimates of  $\mu$ ,  $\alpha_i$ ,  $\beta_j$  and  $\gamma_{ij}$ :

$$\begin{aligned}
 0 &= 2n\hat{\mu} + 2 \left( \sum_{i=1}^a n_{i\bullet} \alpha_i + \sum_{j=1}^b n_{\bullet j} \beta_j + \sum_{i=1}^a \sum_{j=1}^b n_{ij} \gamma_{ij} \right) - 2 \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \hat{\mu} &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{i=1}^a \frac{n_{i\bullet}}{n} \alpha_i - \sum_{j=1}^b \frac{n_{\bullet j}}{n} \beta_j - \sum_{i=1}^a \sum_{j=1}^b \frac{n_{ij}}{n} \gamma_{ij} \\
 &\stackrel{(4)}{=} \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{j=1}^b \sum_{i=1}^a \frac{n_{ij}}{n} \alpha_i - \sum_{i=1}^a \sum_{j=1}^b \frac{n_{ij}}{n} \beta_j - \sum_{i=1}^a \sum_{j=1}^b \frac{n_{ij}}{n} \gamma_{ij} \\
 &\stackrel{(5)}{=} \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 &\stackrel{(3)}{=} \bar{y}_{\bullet\bullet\bullet}
 \end{aligned} \tag{11}$$

$$\begin{aligned}
 0 &= 2n_{i\bullet} \hat{\mu} + 2n_{i\bullet} \hat{\alpha}_i + 2 \left( \sum_{j=1}^b n_{ij} \beta_j + \sum_{j=1}^b n_{ij} \gamma_{ij} \right) - 2 \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \hat{\alpha}_i &= \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \sum_{j=1}^b \frac{n_{ij}}{n_{i\bullet}} \beta_j - \sum_{j=1}^b \frac{n_{ij}}{n_{i\bullet}} \gamma_{ij} \\
 &= \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \frac{n}{n_{i\bullet}} \sum_{j=1}^b \frac{n_{ij}}{n} \beta_j - \frac{n}{n_{i\bullet}} \sum_{j=1}^b \frac{n_{ij}}{n} \gamma_{ij} \\
 &\stackrel{(5)}{=} \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 &\stackrel{(3)}{=} \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}
 \end{aligned} \tag{12}$$

$$\begin{aligned}
 0 &= 2n_{\bullet j} \hat{\mu} + 2n_{\bullet j} \hat{\beta}_j + 2 \left( \sum_{i=1}^a n_{ij} \alpha_i + \sum_{i=1}^a n_{ij} \gamma_{ij} \right) - 2 \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \hat{\beta}_j &= \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \sum_{i=1}^a \frac{n_{ij}}{n_{\bullet j}} \alpha_i - \sum_{i=1}^a \frac{n_{ij}}{n_{\bullet j}} \gamma_{ij} \\
 &= \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \frac{n}{n_{\bullet j}} \sum_{i=1}^a \frac{n_{ij}}{n} \alpha_i - \frac{n}{n_{\bullet j}} \sum_{i=1}^a \frac{n_{ij}}{n} \gamma_{ij} \\
 &\stackrel{(5)}{=} \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 &\stackrel{(3)}{=} \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}
 \end{aligned} \tag{13}$$

$$\begin{aligned}
0 &= 2n_{ij}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}) - 2 \sum_{k=1}^{n_{ij}} y_{ijk} \\
\hat{\gamma}_{ij} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu} \\
&= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} + \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
&\stackrel{(3)}{=} \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet} .
\end{aligned} \tag{14}$$

■

**Sources:**

- Olbricht, Gayla R. (2011): “Two-Way ANOVA: Interaction”; in: *Stat 512: Applied Regression Analysis*, Purdue University, Spring 2011, Lect. 27; URL: [https://www.stat.purdue.edu/~ghobbs/STAT\\_512/Lecture\\_Notes/ANOVA/Topic\\_27.pdf](https://www.stat.purdue.edu/~ghobbs/STAT_512/Lecture_Notes/ANOVA/Topic_27.pdf).

**1.3.11 Sums of squares in two-way ANOVA**

**Theorem:** Given two-way analysis of variance ( $\rightarrow$  III/1.3.8),

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

sums of squares can be partitioned as follows

$$\text{SS}_{\text{tot}} = \text{SS}_A + \text{SS}_B + \text{SS}_{A \times B} + \text{SS}_{\text{res}} \tag{2}$$

where  $\text{SS}_{\text{tot}}$  is the total sum of squares ( $\rightarrow$  III/1.5.7),  $\text{SS}_A$ ,  $\text{SS}_B$  and  $\text{SS}_{A \times B}$  are treatment ( $\rightarrow$  III/1.3.2) and interaction sum of squares ( $\rightarrow$  III/1.3.9) (summing into the explained sum of squares ( $\rightarrow$  III/1.5.8)) and  $\text{SS}_{\text{res}}$  is the residual sum of squares ( $\rightarrow$  III/1.5.9).

**Proof:** The total sum of squares ( $\rightarrow$  III/1.5.7) for two-way ANOVA ( $\rightarrow$  III/1.3.8) is given by

$$\text{SS}_{\text{tot}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 \tag{3}$$

where  $\bar{y}_{\bullet\bullet\bullet}$  is the mean across all values  $y_{ijk}$ . This can be rewritten as

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(y_{ijk} - \bar{y}_{ij\bullet}) + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) +$$

$$(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})]^2$$

It can be shown ( $\rightarrow$  III/1.3.12) that the following sums are all zero:

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet}) &= 0 \\
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) &= 0 \\
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) &= 0 \\
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}) &= 0 .
\end{aligned} \tag{5}$$

This means that the sum in (4) reduces to

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} \left[ (y_{ijk} - \bar{y}_{ij\bullet})^2 + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \right. \\
&\quad \left. (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \right] \\
&= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \\
&\quad \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 .
\end{aligned} \tag{6}$$

With the treatment sums of squares ( $\rightarrow$  III/1.3.2)

$$\begin{aligned}
SS_A &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 \\
SS_B &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 ,
\end{aligned} \tag{7}$$

the interaction sum of squares ( $\rightarrow$  III/1.3.9)

$$SS_{A \times B} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \tag{8}$$

and the residual sum of squares ( $\rightarrow$  III/1.5.9) for two-way ANOVA ( $\rightarrow$  III/1.3.8)

$$SS_{\text{res}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 , \tag{9}$$

we finally have:

$$SS_{\text{tot}} = SS_A + SS_B + SS_{A \times B} + SS_{\text{res}} . \quad (10)$$

■

**Sources:**

- Nandy, Siddhartha (2018): “Two-Way Analysis of Variance”; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: <https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf>.
- Wikipedia (2022): “Analysis of variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: [https://en.wikipedia.org/wiki/Analysis\\_of\\_variance#Partitioning\\_of\\_the\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares).

**1.3.12 Cochran’s theorem for two-way ANOVA**

**Theorem:** Assume the two-way analysis of variance ( $\rightarrow$  III/1.3.8) model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (1)$$

$$\varepsilon_{ijk} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}$$

under the well-known constraints for the model parameters ( $\rightarrow$  III/1.3.8)

$$\begin{aligned} \sum_{i=1}^a \frac{n_{ij}}{n} \alpha_i &= 0 \quad \text{for all } j = 1, \dots, b \\ \sum_{j=1}^b \frac{n_{ij}}{n} \beta_j &= 0 \quad \text{for all } i = 1, \dots, a \\ \sum_{i=1}^a \frac{n_{ij}}{n} \gamma_{ij} &= 0 \quad \text{for all } j = 1, \dots, b \\ \sum_{j=1}^b \frac{n_{ij}}{n} \gamma_{ij} &= 0 \quad \text{for all } i = 1, \dots, a . \end{aligned} \quad (2)$$

Then, the following sums of squares ( $\rightarrow$  III/1.3.11) are chi-square distributed ( $\rightarrow$  II/3.7.1)

$$\begin{aligned} \frac{1}{\sigma^2} n (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 &= \frac{SS_M}{\sigma^2} \sim \chi^2(1) \\ \frac{1}{\sigma^2} \sum_{i=1}^a n_{i\bullet} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 &= \frac{SS_A}{\sigma^2} \sim \chi^2(a-1) \\ \frac{1}{\sigma^2} \sum_{j=1}^b n_{\bullet j} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 &= \frac{SS_B}{\sigma^2} \sim \chi^2(b-1) \\ \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 &= \frac{SS_{A \times B}}{\sigma^2} \sim \chi^2((a-1)(b-1)) \\ \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 &= \frac{SS_{\text{res}}}{\sigma^2} \sim \chi^2(n - ab) . \end{aligned} \quad (3)$$



**Proof:** Denote sample sizes as

$$\begin{aligned}
 n_{ij} &= \text{number of samples in category } (i, j) \\
 n_{i\bullet} &= \sum_{j=1}^b n_{ij} \\
 n_{\bullet j} &= \sum_{i=1}^a n_{ij} \\
 n &= \sum_{i=1}^a \sum_{j=1}^b n_{ij}
 \end{aligned} \tag{4}$$

and denote sample means as

$$\begin{aligned}
 \bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} \\
 \bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} .
 \end{aligned} \tag{5}$$

According to the model given by (1), the observations are distributed as:

$$y_{ijk} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2) \quad \text{for all } i, j, k . \tag{6}$$

Thus, the random variable ( $\rightarrow$  I/1.2.2)  $U_{ijk} = (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma$  follows a standard normal distribution ( $\rightarrow$  II/3.2.4)

$$U_{ijk} = \frac{y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}}{\sigma} \sim \mathcal{N}(0, 1) . \tag{7}$$

Now consider the following sum

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} U_{ijk}^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} \left( \frac{y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}}{\sigma} \right)^2 \tag{8}$$

which can be rewritten as follows:

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} U_{ijk}^2 &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}) - \\
&\quad [\bar{y}_{\bullet\bullet\bullet} + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})] + \\
&\quad [\bar{y}_{\bullet\bullet\bullet} + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})]]^2 \\
&= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(y_{ijk} - [\bar{y}_{\bullet\bullet\bullet} + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})]) + \\
&\quad (\bar{y}_{\bullet\bullet\bullet} - \mu) + ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i) + ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j) + \\
&\quad ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij}))^2 \\
&= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(y_{ijk} - \bar{y}_{ij\bullet}) + (\bar{y}_{\bullet\bullet\bullet} - \mu) + ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i) + \\
&\quad ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j) + ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})]^2.
\end{aligned} \tag{9}$$

Note that the following sums are all zero:

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet}) &= \sum_{i=1}^a \sum_{j=1}^b \left[ \sum_{k=1}^{n_{ij}} y_{ijk} - n_{ij} \cdot \bar{y}_{ij\bullet} \right] \\
&\stackrel{(5)}{=} \sum_{i=1}^a \sum_{j=1}^b \left[ \sum_{k=1}^{n_{ij}} y_{ijk} - n_{ij} \cdot \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \right] \\
&= \sum_{i=1}^a \sum_{j=1}^b 0 = 0
\end{aligned} \tag{10}$$

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i) &= \sum_{i=1}^a n_{i\bullet\bullet} \cdot (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} - \alpha_i) \\
&= \sum_{i=1}^a n_{i\bullet\bullet} \cdot \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \sum_{i=1}^a n_{i\bullet\bullet} - \sum_{i=1}^a n_{i\bullet\bullet} \alpha_i \\
&\stackrel{(5)}{=} \sum_{i=1}^a n_{i\bullet\bullet} \cdot \frac{1}{n_{i\bullet\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - n \cdot \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{i=1}^a n_{i\bullet\bullet} \alpha_i \\
&= - \sum_{i=1}^a n_{i\bullet\bullet} \alpha_i \stackrel{(4)}{=} -n \sum_{i=1}^a \sum_{j=1}^b \frac{n_{ij}}{n} \alpha_i \stackrel{(2)}{=} 0
\end{aligned} \tag{11}$$

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet}] - \beta_j) &= \sum_{j=1}^b n_{\bullet j} \cdot (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet} - \beta_j) \\
&= \sum_{j=1}^b n_{\bullet j} \cdot \bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet} \sum_{j=1}^b n_{\bullet j} - \sum_{j=1}^b n_{\bullet j} \beta_j \\
&\stackrel{(5)}{=} \sum_{j=1}^b n_{\bullet j} \cdot \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} - n \cdot \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{j=1}^b n_{\bullet j} \beta_j \\
&= - \sum_{j=1}^b n_{\bullet j} \beta_j \stackrel{(4)}{=} -n \sum_{j=1}^b \sum_{i=1}^a \frac{n_{ij}}{n} \beta_j \stackrel{(2)}{=} 0
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij \bullet} - \bar{y}_{i \bullet \bullet} - \bar{y}_{\bullet j \bullet} + \bar{y}_{\bullet \bullet \bullet}] - \gamma_{ij}) \\
&= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(\bar{y}_{ij \bullet} - \bar{y}_{\bullet \bullet \bullet}) - (\bar{y}_{i \bullet \bullet} - \bar{y}_{\bullet \bullet \bullet}) - (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet}) - \gamma_{ij}] \\
&= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij \bullet} - \bar{y}_{\bullet \bullet \bullet} - \gamma_{ij}) - \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{i \bullet \bullet} - \bar{y}_{\bullet \bullet \bullet}) - \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet}) \\
&\stackrel{(12)}{=} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij \bullet} - \bar{y}_{\bullet \bullet \bullet} - \gamma_{ij}) - \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{i \bullet \bullet} - \bar{y}_{\bullet \bullet \bullet}) \\
&\stackrel{(11)}{=} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{ij \bullet} - \bar{y}_{\bullet \bullet \bullet} - \gamma_{ij}) \\
&= \sum_{i=1}^a \sum_{j=1}^b n_{ij} \bar{y}_{ij \bullet} - \bar{y}_{\bullet \bullet \bullet} \sum_{i=1}^a \sum_{j=1}^b n_{ij} - \sum_{i=1}^a \sum_{j=1}^b n_{ij} \gamma_{ij} \\
&\stackrel{(5)}{=} \sum_{i=1}^a \sum_{j=1}^b n_{ij} \cdot \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - n \cdot \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{i=1}^a \sum_{j=1}^b n_{ij} \gamma_{ij} \\
&= - \sum_{i=1}^a \sum_{j=1}^b n_{ij} \gamma_{ij} = -\frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \frac{n_{ij}}{n} \gamma_{ij} \stackrel{(2)}{=} 0.
\end{aligned} \tag{13}$$

Note further that  $\bar{y}_{\bullet \bullet \bullet}$  and  $\mu$  are not dependent on  $i, j$  and  $k$ :

$$\bar{y}_{\bullet \bullet \bullet} = \text{const.} \quad \text{and} \quad \mu = \text{const.} \tag{14}$$

Thus, all the non-square products in (9) disappear and the sum reduces to

$$\begin{aligned}
\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} U_{ijk}^2 &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} [(y_{ijk} - \bar{y}_{ij\bullet})^2 + (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 + ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 + \\
&\quad ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 + ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2] \\
&= \frac{1}{\sigma^2} \left[ \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 + \right. \\
&\quad \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 + \\
&\quad \left. \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \right]. \tag{15}
\end{aligned}$$

Cochran's theorem states that, if a sum of squared standard normal ( $\rightarrow$  II/3.2.3) random variables ( $\rightarrow$  I/1.2.2) can be written as a sum of squared forms

$$\begin{aligned}
\sum_{i=1}^n U_i^2 &= \sum_{j=1}^m Q_j \quad \text{where} \quad Q_j = U^T B^{(j)} U \\
&\quad \text{with} \quad \sum_{j=1}^m B^{(j)} = I_n \\
&\quad \text{and} \quad r_j = \text{rank}(B^{(j)}), \tag{16}
\end{aligned}$$

then the terms  $Q_j$  are independent ( $\rightarrow$  I/1.3.6) and each term  $Q_j$  follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $r_j$  degrees of freedom:

$$Q_j \sim \chi^2(r_j), \quad j = 1, \dots, m. \tag{17}$$

First, we define the  $n$ -dimensional vector  $U$ :

$$U = \begin{bmatrix} u_{1\bullet} \\ \vdots \\ u_{a\bullet} \end{bmatrix} \quad \text{where} \quad u_{i\bullet} = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{ib} \end{bmatrix} \quad \text{where} \quad u_{ij} = \begin{bmatrix} (y_{i,j,1} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma \\ \vdots \\ (y_{i,j,n_{ij}} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma \end{bmatrix}. \tag{18}$$

Next, we specify the  $n \times n$  matrices  $B$

$$\begin{aligned}
B^{(1)} &= I_n - \text{diag} \left[ \text{diag} \left( \frac{1}{n_{11}} J_{n_{11}}, \dots, \frac{1}{n_{1b}} J_{n_{1b}} \right), \dots, \text{diag} \left( \frac{1}{n_{a1}} J_{n_{a1}}, \dots, \frac{1}{n_{ab}} J_{n_{ab}} \right) \right] \\
B^{(2)} &= \frac{1}{n} J_n \\
B^{(3)} &= \text{diag} \left( \frac{1}{n_{1\bullet}} J_{n_{1\bullet}}, \dots, \frac{1}{n_{a\bullet}} J_{n_{a\bullet}} \right) - \frac{1}{n} J_n \\
B^{(4)} &= M_B - \frac{1}{n} J_n \\
B^{(5)} &= \text{diag} \left[ \text{diag} \left( \frac{1}{n_{11}} J_{n_{11}}, \dots, \frac{1}{n_{1b}} J_{n_{1b}} \right), \dots, \text{diag} \left( \frac{1}{n_{a1}} J_{n_{a1}}, \dots, \frac{1}{n_{ab}} J_{n_{ab}} \right) \right] \\
&\quad - \text{diag} \left( \frac{1}{n_{1\bullet}} J_{n_{1\bullet}}, \dots, \frac{1}{n_{a\bullet}} J_{n_{a\bullet}} \right) - M_B + \frac{1}{n} J_n
\end{aligned} \tag{19}$$

with the factor B matrix  $M_B$  given by

$$M_B = \begin{bmatrix} \text{diag} \left( \frac{1}{n_{\bullet 1}} J_{n_{11}, n_{11}}, \dots, \frac{1}{n_{\bullet b}} J_{n_{1b}, n_{1b}} \right) & \cdots & \text{diag} \left( \frac{1}{n_{\bullet 1}} J_{n_{11}, n_{a1}}, \dots, \frac{1}{n_{\bullet b}} J_{n_{1b}, n_{ab}} \right) \\ \vdots & \ddots & \vdots \\ \text{diag} \left( \frac{1}{n_{\bullet 1}} J_{n_{a1}, n_{11}}, \dots, \frac{1}{n_{\bullet b}} J_{n_{ab}, n_{1b}} \right) & \cdots & \text{diag} \left( \frac{1}{n_{\bullet 1}} J_{n_{a1}, n_{a1}}, \dots, \frac{1}{n_{\bullet b}} J_{n_{ab}, n_{ab}} \right) \end{bmatrix}. \tag{20}$$

where  $J_n$  is an  $n \times n$  matrix of ones,  $J_{n,m}$  is an  $n \times m$  matrix of ones and  $\text{diag}(A_1, \dots, A_n)$  denotes a block-diagonal matrix composed of  $A_1, \dots, A_n$ . We observe that those matrices satisfy

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} U_{ijk}^2 = \sum_{l=1}^5 Q_l = \sum_{l=1}^5 U^T B^{(l)} U \tag{21}$$

as well as

$$\sum_{l=1}^5 B^{(l)} = I_n \tag{22}$$

and their ranks are

$$\begin{aligned}
\text{rank}(B^{(1)}) &= n - ab \\
\text{rank}(B^{(2)}) &= 1 \\
\text{rank}(B^{(3)}) &= a - 1 \\
\text{rank}(B^{(4)}) &= b - 1 \\
\text{rank}(B^{(5)}) &= (a - 1)(b - 1).
\end{aligned} \tag{23}$$

Thus, the conditions for applying Cochran's theorem given by (16) are fulfilled and we can use (15), (17), (19) and (23) to conclude that

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 &= Q_2 = U^T B^{(2)} U \sim \chi^2(1) \\
\frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 &= Q_3 = U^T B^{(3)} U \sim \chi^2(a-1) \\
\frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 &= Q_4 = U^T B^{(4)} U \sim \chi^2(b-1) \\
\frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 &= Q_5 = U^T B^{(5)} U \sim \chi^2((a-1)(b-1)) \\
\frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 &= Q_1 = U^T B^{(1)} U \sim \chi^2(n-ab) .
\end{aligned} \tag{24}$$

Finally, we identify the terms  $Q$  with sums of squares in two-way ANOVA ( $\rightarrow$  III/1.3.11) and simplify them to reach the expressions given by (3):

$$\begin{aligned}
\frac{SS_M}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 = \frac{1}{\sigma^2} n (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 \\
\frac{SS_A}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 = \frac{1}{\sigma^2} \sum_{i=1}^a n_{i\bullet} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 \\
\frac{SS_B}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 = \frac{1}{\sigma^2} \sum_{j=1}^b n_{\bullet j} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 \\
\frac{SS_{A \times B}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 = \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \\
\frac{SS_{\text{res}}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 .
\end{aligned} \tag{25}$$

■

#### Sources:

- Nandy, Siddhartha (2018): “Two-Way Analysis of Variance”; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: <https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf>.

#### 1.3.13 F-test for main effect in two-way ANOVA

**Theorem:** Assume the two-way analysis of variance ( $\rightarrow$  III/1.3.8) model

$$\begin{aligned}
y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\
\varepsilon_{ijk} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij} .
\end{aligned} \tag{1}$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F_A = \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \quad (2)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F_A \sim F(a-1, n-ab) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the main effect ( $\rightarrow$  III/1.3.8) of factor A

$$\begin{aligned} H_0 : \alpha_1 = \dots = \alpha_a = 0 \\ H_1 : \alpha_i \neq 0 \quad \text{for at least one } i \in \{1, \dots, a\} \end{aligned} \quad (4)$$

and the test statistic ( $\rightarrow$  I/4.3.5)

$$F_B = \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \quad (5)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F_B \sim F(b-1, n-ab) \quad (6)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the main effect ( $\rightarrow$  III/1.3.8) of factor B

$$\begin{aligned} H_0 : \beta_1 = \dots = \beta_b = 0 \\ H_1 : \beta_j \neq 0 \quad \text{for at least one } j \in \{1, \dots, b\} . \end{aligned} \quad (7)$$

**Proof:** Applying Cochran's theorem for two-analysis of variance ( $\rightarrow$  III/1.3.12), we find that the following squared sums

$$\begin{aligned} \frac{SS_A}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 = \frac{1}{\sigma^2} \sum_{i=1}^a n_{i\bullet} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 \\ \frac{SS_B}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 = \frac{1}{\sigma^2} \sum_{j=1}^b n_{\bullet j} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 \\ \frac{SS_{\text{res}}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \end{aligned} \quad (8)$$

are independent ( $\rightarrow$  I/1.3.6) and chi-squared distributed ( $\rightarrow$  II/3.7.1):

$$\begin{aligned} \frac{SS_A}{\sigma^2} &\sim \chi^2(a-1) \\ \frac{SS_B}{\sigma^2} &\sim \chi^2(b-1) \\ \frac{SS_{\text{res}}}{\sigma^2} &\sim \chi^2(n-ab) . \end{aligned} \quad (9)$$

1) Thus, the F-statistic from (2) is equal to the ratio of two independent ( $\rightarrow$  I/1.3.6) chi-squared distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}
 F_A &= \frac{(SS_A/\sigma^2)/(a-1)}{(SS_{\text{res}}/\sigma^2)/(n-ab)} \\
 &= \frac{SS_A/(a-1)}{SS_{\text{res}}/(n-ab)} \\
 &\stackrel{(8)}{=} \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\
 &\stackrel{(4)}{=} \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2}
 \end{aligned} \tag{10}$$

which, by definition of the F-distribution ( $\rightarrow$  II/3.8.1), is distributed as

$$F_A \sim F(a-1, n-ab) \tag{11}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for main effect of  $A$ .

2) Similarly, the F-statistic from (5) is equal to the ratio of two independent chi-squared distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}
 F_B &= \frac{(SS_B/\sigma^2)/(b-1)}{(SS_{\text{res}}/\sigma^2)/(n-ab)} \\
 &= \frac{SS_B/(b-1)}{SS_{\text{res}}/(n-ab)} \\
 &\stackrel{(8)}{=} \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\
 &\stackrel{(7)}{=} \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2}
 \end{aligned} \tag{12}$$

which, by definition of the F-distribution ( $\rightarrow$  II/3.8.1), is distributed as

$$F_B \sim F(b-1, n-ab) \tag{13}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for main effect of  $B$ . ■

#### Sources:

- ttd (2021): “Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J”; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: <https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1j-1-under-the-null-hypothesis>.
- JohnK (2014): “In a two-way ANOVA, how can the F-statistic for one factor have a central distribution if the null is false for the other factor?”; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: <https://stats.stackexchange.com/questions/124166/in-a-two-way-anova-how-can-the-f>.



## 1.3.14 F-test for interaction in two-way ANOVA

**Theorem:** Assume the two-way analysis of variance ( $\rightarrow$  III/1.3.8) model

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij} . \end{aligned} \quad (1)$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F_{A \times B} = \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \quad (2)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F_{A \times B} \sim F((a-1)(b-1), n-ab) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the interaction effect ( $\rightarrow$  III/1.3.8) of factors A and B

$$\begin{aligned} H_0 : \gamma_{11} = \dots = \gamma_{ab} &= 0 \\ H_1 : \gamma_{ij} &\neq 0 \quad \text{for at least one } (i, j) \in \{1, \dots, a\} \times \{1, \dots, b\} . \end{aligned} \quad (4)$$

**Proof:** Applying Cochran's theorem for two-analysis of variance ( $\rightarrow$  III/1.3.12), we find that the following squared sums

$$\begin{aligned} \frac{SS_{A \times B}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \\ \frac{SS_{\text{res}}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \end{aligned} \quad (5)$$

are independent ( $\rightarrow$  I/1.3.6) and chi-squared distributed ( $\rightarrow$  II/3.7.1):

$$\begin{aligned} \frac{SS_{A \times B}}{\sigma^2} &\sim \chi^2((a-1)(b-1)) \\ \frac{SS_{\text{res}}}{\sigma^2} &\sim \chi^2(n-ab) . \end{aligned} \quad (6)$$

Thus, the F-statistic from (2) is equal to the ratio of two independent ( $\rightarrow$  I/1.3.6) chi-squared distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}
F_{A \times B} &= \frac{(\text{SS}_{A \times B} / \sigma^2) / ((a-1)(b-1))}{(\text{SS}_{\text{res}} / \sigma^2) / (n-ab)} \\
&= \frac{\text{SS}_{A \times B} / ((a-1)(b-1))}{\text{SS}_{\text{res}} / (n-ab)} \\
&\stackrel{(5)}{=} \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\
&\stackrel{(4)}{=} \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2}
\end{aligned} \tag{7}$$

which, by definition of the F-distribution ( $\rightarrow$  II/3.8.1), is distributed as

$$F_{A \times B} \sim F((a-1)(b-1), n-ab) \tag{8}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for an interaction of A and B.

■

#### Sources:

- Nandy, Siddhartha (2018): “Two-Way Analysis of Variance”; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: <https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf>.
- ttd (2021): “Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J”; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: <https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1-j-1-under-the-null-hypothesis>.

#### 1.3.15 F-test for grand mean in two-way ANOVA

**Theorem:** Assume the two-way analysis of variance ( $\rightarrow$  III/1.3.8) model

$$\begin{aligned}
y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\
\varepsilon_{ijk} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.
\end{aligned} \tag{1}$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F_M = \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \tag{2}$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F_M \sim F(1, n-ab) \tag{3}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the grand mean ( $\rightarrow$  III/1.3.8)

$$\begin{aligned}
H_0 : \mu &= 0 \\
H_1 : \mu &\neq 0.
\end{aligned} \tag{4}$$

**Proof:** Applying Cochran's theorem for two-analysis of variance ( $\rightarrow$  III/1.3.12), we find that the following squared sums

$$\begin{aligned}\frac{SS_M}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 = \frac{1}{\sigma^2} n(\bar{y}_{\bullet\bullet\bullet} - \mu)^2 \\ \frac{SS_{\text{res}}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2\end{aligned}\quad (5)$$

are independent ( $\rightarrow$  I/1.3.6) and chi-squared distributed ( $\rightarrow$  II/3.7.1):

$$\begin{aligned}\frac{SS_M}{\sigma^2} &\sim \chi^2(1) \\ \frac{SS_{\text{res}}}{\sigma^2} &\sim \chi^2(n - ab) .\end{aligned}\quad (6)$$

Thus, the F-statistic from (2) is equal to the ratio of two independent ( $\rightarrow$  I/1.3.6) chi-squared distributed ( $\rightarrow$  II/3.7.1) random variables ( $\rightarrow$  I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}F_M &= \frac{(SS_M/\sigma^2)/(1)}{(SS_{\text{res}}/\sigma^2)/(n - ab)} \\ &= \frac{SS_M/(1)}{SS_{\text{res}}/(n - ab)} \\ &\stackrel{(5)}{=} \frac{n(\bar{y}_{\bullet\bullet\bullet} - \mu)^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\ &\stackrel{(4)}{=} \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2}\end{aligned}\quad (7)$$

which, by definition of the F-distribution ( $\rightarrow$  II/3.8.1), is distributed as

$$F_M \sim F(1, n - ab) \quad (8)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) for the grand mean. ■

#### Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: <https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf>.
- Olbricht, Gayla R. (2011): "Two-Way ANOVA: Interaction"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Spring 2011, Lect. 27; URL: [https://www.stat.purdue.edu/~ghobbs/STAT\\_512/Lecture\\_Notes/ANOVA/Topic\\_27.pdf](https://www.stat.purdue.edu/~ghobbs/STAT_512/Lecture_Notes/ANOVA/Topic_27.pdf).

### 1.3.16 F-statistics in terms of OLS estimates

**Theorem:** Given the two-way analysis of variance ( $\rightarrow$  III/1.3.8) assumption

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (1)$$

the F-statistics for the grand mean ( $\rightarrow$  III/1.3.15), the main effects ( $\rightarrow$  III/1.3.13) and the interaction ( $\rightarrow$  III/1.3.14) can be expressed in terms of ordinary least squares parameter estimates ( $\rightarrow$  III/1.3.10) as

$$\begin{aligned} F_M &= \frac{n\hat{\mu}^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\ F_A &= \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} \hat{\alpha}_i^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\ F_B &= \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} \hat{\beta}_j^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\ F_{A \times B} &= \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} \hat{\gamma}_{ij}^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \end{aligned} \quad (2)$$

where the predicted values  $\hat{y}_{ijk}$  are given by

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}. \quad (3)$$

**Theorem:** The F-statistics for the grand mean ( $\rightarrow$  III/1.3.15), the main effects ( $\rightarrow$  III/1.3.13) and the interaction ( $\rightarrow$  III/1.3.14) in two-way ANOVA ( $\rightarrow$  III/1.3.8) are calculated as

$$\begin{aligned} F_M &= \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\ F_A &= \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\ F_B &= \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\ F_{A \times B} &= \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \end{aligned} \quad (4)$$

and the ordinary least squares estimates for two-way ANOVA ( $\rightarrow$  III/1.3.10) are

$$\begin{aligned} \hat{\mu} &= \bar{y}_{\bullet\bullet\bullet} \\ \hat{\alpha}_i &= \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \\ \hat{\beta}_j &= \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet} \\ \hat{\gamma}_{ij} &= \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet} \end{aligned} \quad (5)$$

where the sample means ( $\rightarrow$  I/1.12.2) are given by

$$\begin{aligned}
\bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^a \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} .
\end{aligned} \tag{6}$$

We first note that the predicted values can be evaluated as

$$\begin{aligned}
\hat{y}_{ijk} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} \\
&= \bar{y}_{\bullet\bullet\bullet} + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}) \\
&= \bar{y}_{i\bullet\bullet} + \bar{y}_{\bullet j\bullet} + \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} \\
&= \bar{y}_{ij\bullet} .
\end{aligned} \tag{7}$$

Substituting this (7) and the OLS estimates (5) into the F-formulas (4), we obtain:

$$\begin{aligned}
F_M &= \frac{n\hat{\mu}^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\
F_A &= \frac{\frac{1}{a-1} \sum_{i=1}^a n_{i\bullet} \hat{\alpha}_i^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\
F_B &= \frac{\frac{1}{b-1} \sum_{j=1}^b n_{\bullet j} \hat{\beta}_j^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} \\
F_{A \times B} &= \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b n_{ij} \hat{\gamma}_{ij}^2}{\frac{1}{n-ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2} .
\end{aligned} \tag{8}$$

■

## 1.4 Simple linear regression

### 1.4.1 Definition

**Definition:** Let  $y$  and  $x$  be two  $n \times 1$  vectors.

Then, a statement asserting a linear relationship between  $x$  and  $y$

$$y = \beta_0 + \beta_1 x + \varepsilon , \tag{1}$$

together with a statement asserting a normal distribution ( $\rightarrow$  II/4.1.1) for  $\varepsilon$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{2}$$

is called a univariate simple regression model or simply, “simple linear regression”.

- $y$  is called “dependent variable”, “measured data” or “signal”;
- $x$  is called “independent variable”, “predictor” or “covariate”;
- $V$  is called “covariance matrix” or “covariance structure”;
- $\beta_1$  is called “slope of the regression line ( $\rightarrow$  III/1.4.10)”;
- $\beta_0$  is called “intercept of the regression line ( $\rightarrow$  III/1.4.10)”;
- $\varepsilon$  is called “noise”, “errors” or “error terms”;
- $\sigma^2$  is called “noise variance” or “error variance”;
- $n$  is the number of observations.

When the covariance structure  $V$  is equal to the  $n \times n$  identity matrix, this is called simple linear regression with independent and identically distributed ( $\rightarrow$  I/1.2.8) (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) . \quad (3)$$

In this case, the linear regression model can also be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) . \quad (4)$$

Otherwise, it is called simple linear regression with correlated observations ( $\rightarrow$  I/1.17.1).

#### Sources:

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Fitting\\_the\\_regression\\_line](https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line).

### 1.4.2 Special case of multiple linear regression

**Theorem:** Simple linear regression ( $\rightarrow$  III/1.4.1) is a special case of multiple linear regression ( $\rightarrow$  III/1.5.1) with design matrix  $X$  and regression coefficients  $\beta$

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (1)$$

where  $1_n$  is an  $n \times 1$  vector of ones,  $x$  is the  $n \times 1$  single predictor variable,  $\beta_0$  is the intercept and  $\beta_1$  is the slope of the regression line ( $\rightarrow$  III/1.4.10).

**Proof:** Without loss of generality, consider the simple linear regression case with uncorrelated errors ( $\rightarrow$  III/1.4.1):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n . \quad (2)$$

In matrix notation and using the multivariate normal distribution ( $\rightarrow$  II/4.1.1), this can also be written as

$$\begin{aligned} y &= \beta_0 1_n + \beta_1 x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_n) \\ y &= \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_n) . \end{aligned} \quad (3)$$

Comparing with the multiple linear regression equations for uncorrelated errors ( $\rightarrow$  III/1.5.1), we finally note:

$$y = X\beta + \varepsilon \quad \text{with} \quad X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}. \quad (4)$$

In the case of correlated observations ( $\rightarrow$  III/1.4.1), the error distribution changes to ( $\rightarrow$  III/1.5.1):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V). \quad (5)$$

■

### 1.4.3 Ordinary least squares

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2),  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) between  $x$  and  $y$ .

**Proof:** The residual sum of squares ( $\rightarrow$  III/1.5.9) is defined as

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (3)$$

The derivatives of  $\text{RSS}(\beta_0, \beta_1)$  with respect to  $\beta_0$  and  $\beta_1$  are

$$\begin{aligned} \frac{\text{dRSS}(\beta_0, \beta_1)}{\text{d}\beta_0} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) \\ &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\text{dRSS}(\beta_0, \beta_1)}{\text{d}\beta_1} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \\ &= -2 \sum_{i=1}^n (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \end{aligned} \quad (4)$$

and setting these derivatives to zero

$$\begin{aligned}
0 &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
0 &= -2 \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)
\end{aligned} \tag{5}$$

yields the following equations:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^n x_i + \hat{\beta}_0 \cdot n &= \sum_{i=1}^n y_i \\
\hat{\beta}_1 \sum_{i=1}^n x_i^2 + \hat{\beta}_0 \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i .
\end{aligned} \tag{6}$$

From the first equation, we can derive the estimate for the intercept:

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \\
&= \bar{y} - \hat{\beta}_1 \bar{x} .
\end{aligned} \tag{7}$$

From the second equation, we can derive the estimate for the slope:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^n x_i^2 + \hat{\beta}_0 \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i \\
\hat{\beta}_1 \sum_{i=1}^n x_i^2 + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i &\stackrel{(7)}{=} \sum_{i=1}^n x_i y_i \\
\hat{\beta}_1 \left( \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} .
\end{aligned} \tag{8}$$

Note that the numerator can be rewritten as

$$\begin{aligned}
\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \\
&= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + \sum_{i=1}^n \bar{x} \bar{y} \\
&= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\
&= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})
\end{aligned} \tag{9}$$



and that the denominator can be rewritten as

$$\begin{aligned}
 \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\
 &= \sum_{i=1}^n x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
 &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \\
 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
 &= \sum_{i=1}^n (x_i - \bar{x})^2 .
 \end{aligned} \tag{10}$$

With (9) and (10), the estimate from (8) can be simplified as follows:

$$\begin{aligned}
 \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 &= \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\
 &= \frac{s_{xy}}{s_x^2} .
 \end{aligned} \tag{11}$$

Together, (7) and (11) constitute the ordinary least squares parameter estimates for simple linear regression. ■

#### Sources:

- Penny, William (2006): “Linear regression”; in: *Mathematics for Brain Imaging*, ch. 1.2.2, pp. 14-16, eqs. 1.24/1.25; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2021): “Proofs involving ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares#Derivation\\_of\\_simple\\_linear\\_regression\\_estimators](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Derivation_of_simple_linear_regression_estimators).

#### 1.4.4 Ordinary least squares

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \tag{1}$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}\end{aligned}\tag{2}$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2),  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) between  $x$  and  $y$ .

**Proof:** Simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}\tag{3}$$

and ordinary least squares estimates ( $\rightarrow$  III/1.5.3) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y.\tag{4}$$

Writing out equation (4), we have

$$\begin{aligned}\hat{\beta} &= \left( \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} y \\ &= \left( \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^T x \end{bmatrix} \right)^{-1} \begin{bmatrix} n\bar{y} \\ x^T y \end{bmatrix} \\ &= \frac{1}{nx^T x - (n\bar{x})^2} \begin{bmatrix} x^T x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ x^T y \end{bmatrix} \\ &= \frac{1}{nx^T x - (n\bar{x})^2} \begin{bmatrix} n\bar{y} x^T x - n\bar{x} x^T y \\ n x^T y - (n\bar{x})(n\bar{y}) \end{bmatrix}.\end{aligned}\tag{5}$$

Thus, the second entry of  $\hat{\beta}$  is equal to ( $\rightarrow$  III/1.4.3):

$$\begin{aligned}\hat{\beta}_1 &= \frac{n x^T y - (n\bar{x})(n\bar{y})}{nx^T x - (n\bar{x})^2} \\ &= \frac{x^T y - n\bar{x}\bar{y}}{x^T x - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{s_{xy}}{s_x^2}.\end{aligned}\tag{6}$$

Moreover, the first entry of  $\hat{\beta}$  is equal to:

$$\begin{aligned}
\hat{\beta}_0 &= \frac{n\bar{y}x^T x - n\bar{x}x^T y}{nx^T x - (n\bar{x})^2} \\
&= \frac{\bar{y}x^T x - \bar{x}x^T y}{x^T x - n\bar{x}^2} \\
&= \frac{\bar{y}x^T x - \bar{x}x^T y + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y}}{x^T x - n\bar{x}^2} \\
&= \frac{\bar{y}(x^T x - n\bar{x}^2) - \bar{x}(x^T y - n\bar{x}\bar{y})}{x^T x - n\bar{x}^2} \\
&= \frac{\bar{y}(x^T x - n\bar{x}^2)}{x^T x - n\bar{x}^2} - \frac{\bar{x}(x^T y - n\bar{x}\bar{y})}{x^T x - n\bar{x}^2} \\
&= \bar{y} - \bar{x} \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
&= \bar{y} - \hat{\beta}_1 \bar{x} .
\end{aligned} \tag{7}$$

■

#### 1.4.5 Expectation of estimates

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, the expected values ( $\rightarrow$  I/1.12.1) of the estimated parameters are

$$\begin{aligned}
E(\hat{\beta}_0) &= \beta_0 \\
E(\hat{\beta}_1) &= \beta_1
\end{aligned} \tag{2}$$

which means that the ordinary least squares solution ( $\rightarrow$  III/1.4.3) produces unbiased estimators ( $\rightarrow$  I/3.1.2).

**Proof:** According to the simple linear regression model in (1), the expectation of a single data point is

$$E(y_i) = \beta_0 + \beta_1 x_i . \tag{3}$$

The ordinary least squares estimates for simple linear regression ( $\rightarrow$  III/1.4.3) are given by

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} .
\end{aligned} \tag{4}$$

If we define the following quantity

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} , \tag{5}$$

we note that

$$\begin{aligned}\sum_{i=1}^n c_i &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i - n\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{n\bar{x} - n\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0 ,\end{aligned}\tag{6}$$

and

$$\begin{aligned}\sum_{i=1}^n c_i x_i &= \frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i^2 - \bar{x}x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1 .\end{aligned}\tag{7}$$

With (5), the estimate for the slope from (4) becomes

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum_{i=1}^n c_i (y_i - \bar{y}) \\ &= \sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i\end{aligned}\tag{8}$$

and with (3), (6) and (7), its expectation becomes:

$$\begin{aligned}\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left(\sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i\right) \\ &= \sum_{i=1}^n c_i \mathbb{E}(y_i) - \bar{y} \sum_{i=1}^n c_i \\ &= \beta_1 \sum_{i=1}^n c_i x_i + \beta_0 \sum_{i=1}^n c_i - \bar{y} \sum_{i=1}^n c_i \\ &= \beta_1 .\end{aligned}\tag{9}$$

Finally, with (3) and (9), the expectation of the intercept estimate from (4) becomes

$$\begin{aligned}
E(\hat{\beta}_0) &= E\left(\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i\right) \\
&= \frac{1}{n} \sum_{i=1}^n E(y_i) - E(\hat{\beta}_1) \cdot \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \cdot \bar{x} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0 .
\end{aligned} \tag{10}$$

#### Sources:

- Penny, William (2006): “Finding the uncertainty in estimating the slope”; in: *Mathematics for Brain Imaging*, ch. 1.2.4, pp. 18-20, eq. 1.37; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2021): “Proofs involving ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares#Unbiasedness\\_and\\_variance\\_of\\_%7F%22%60UNIQ--postMath-00000037-QINU%60%22%7F](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F%22%60UNIQ--postMath-00000037-QINU%60%22%7F).

#### 1.4.6 Variance of estimates

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, the variances ( $\rightarrow$  I/1.13.1) of the estimated parameters are

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \frac{x^T x}{n} \cdot \frac{\sigma^2}{(n-1)s_x^2} \\
\text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{(n-1)s_x^2}
\end{aligned} \tag{2}$$

where  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $x^T x$  is the sum of squared values of the covariate.

**Proof:** According to the simple linear regression model in (1), the variance of a single data point is

$$\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2 . \tag{3}$$

The ordinary least squares estimates for simple linear regression ( $\rightarrow$  III/1.4.3) are given by

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} .
\end{aligned} \tag{4}$$

If we define the following quantity

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (5)$$

we note that

$$\begin{aligned} \sum_{i=1}^n c_i^2 &= \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \\ &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned} \quad (6)$$

With (5), the estimate for the slope from (4) becomes

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum_{i=1}^n c_i (y_i - \bar{y}) \\ &= \sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i \end{aligned} \quad (7)$$

and with (3) and (6) as well as invariance ( $\rightarrow$  I/1.13.7), scaling ( $\rightarrow$  I/1.13.8) and additivity ( $\rightarrow$  I/1.13.11) of the variance, the variance of  $\hat{\beta}_1$  is:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var} \left( \sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i \right) \\ &= \text{Var} \left( \sum_{i=1}^n c_i y_i \right) \\ &= \sum_{i=1}^n c_i^2 \text{Var}(y_i) \\ &= \sigma^2 \sum_{i=1}^n c_i^2 \\ &= \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sigma^2}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sigma^2}{(n-1) s_x^2}. \end{aligned} \quad (8)$$

Finally, with (3) and (8), the variance of the intercept estimate from (4) becomes:

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i\right) \\
&= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) + \text{Var}\left(\hat{\beta}_1 \cdot \bar{x}\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \text{Var}(y_i) + \bar{x}^2 \cdot \text{Var}(\hat{\beta}_1) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 + \bar{x}^2 \frac{\sigma^2}{(n-1)s_x^2} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{(n-1)s_x^2}.
\end{aligned} \tag{9}$$

Applying the formula for the sample variance ( $\rightarrow$  I/1.13.2)  $s_x^2$ , we finally get:

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left( \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left( \frac{\frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left( \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \bar{x}^2\right) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left( \frac{\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + 2\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left( \frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \frac{x^T x}{n} \cdot \frac{\sigma^2}{(n-1)s_x^2}.
\end{aligned} \tag{10}$$

■

#### Sources:

- Penny, William (2006): “Finding the uncertainty in estimating the slope”; in: *Mathematics for Brain Imaging*, ch. 1.2.4, pp. 18-20, eq. 1.37; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2021): “Proofs involving ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares#Unbiasedness\\_and\\_variance\\_of\\_%7F%22%60UNIQ--postMath-00000037-QINU%60%22%7F](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F%22%60UNIQ--postMath-00000037-QINU%60%22%7F).

#### 1.4.7 Distribution of estimates

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, the estimated parameters are normally distributed ( $\rightarrow$  II/4.1.1) as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \frac{\sigma^2}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \quad (2)$$

where  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2) and  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$ .

**Proof:** Simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad (3)$$

such that (1) can also be written as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad (4)$$

and ordinary least squares estimates ( $\rightarrow$  III/1.5.3) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad (5)$$

From (4) and the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14), it follows that

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n). \quad (6)$$

From (5), in combination with (6) and the transformation theorem ( $\rightarrow$  II/4.1.14), it follows that

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}((X^T X)^{-1} X^T X\beta, \sigma^2 (X^T X)^{-1} X^T I_n X (X^T X)^{-1}) \\ &\sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}). \end{aligned} \quad (7)$$

Applying (3), the covariance matrix ( $\rightarrow$  II/4.1.1) can be further developed as follows:



$$\begin{aligned}
\sigma^2(X^T X)^{-1} &= \sigma^2 \left( \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \\
&= \sigma^2 \left( \begin{bmatrix} 1_n^T 1_n & 1_n^T x \\ x^T 1_n & x^T x \end{bmatrix} \right)^{-1} \\
&= \sigma^2 \left( \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^T x \end{bmatrix} \right)^{-1} \\
&= \frac{\sigma^2}{nx^T x - (n\bar{x})^2} \begin{bmatrix} x^T x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\
&= \frac{\sigma^2}{x^T x - n\bar{x}^2} \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.
\end{aligned} \tag{8}$$

Note that the denominator in the first factor is equal to

$$\begin{aligned}
x^T x - n\bar{x}^2 &= x^T x - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - 2n\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n \bar{x}^2 \\
&= \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^n (x_i^2 - \bar{x})^2 \\
&= (n-1) s_x^2.
\end{aligned} \tag{9}$$

Thus, combining (7), (8) and (9), we have

$$\hat{\beta} \sim \mathcal{N} \left( \beta, \frac{\sigma^2}{(n-1) s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \tag{10}$$

which is equivalent to equation (2). ■

#### Sources:

- Wikipedia (2021): “Proofs involving ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-09; URL: [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares#Unbiasedness\\_and\\_variance\\_of\\_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F).

### 1.4.8 Correlation of estimates

**Theorem:** In simple linear regression ( $\rightarrow$  III/1.4.1), when the independent variable  $x$  is mean-centered ( $\rightarrow$  I/1.12.1), the ordinary least squares ( $\rightarrow$  III/1.4.3) estimates for slope and intercept are uncorrelated ( $\rightarrow$  I/1.17.1).

**Proof:** The parameter estimates for simple linear regression are bivariate normally distributed under ordinary least squares ( $\rightarrow$  III/1.4.7):

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \frac{\sigma^2}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \quad (1)$$

Because the covariance matrix ( $\rightarrow$  I/1.16.1) of the multivariate normal distribution ( $\rightarrow$  II/4.1.1) contains the pairwise covariances of the random variables ( $\rightarrow$  I/1.2.2), we can deduce that the covariance ( $\rightarrow$  I/1.15.1) of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  is:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2} \quad (2)$$

where  $\sigma^2$  is the noise variance ( $\rightarrow$  III/1.4.1),  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $n$  is the number of observations. When  $x$  is mean-centered, we have  $\bar{x} = 0$ , such that:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0. \quad (3)$$

Because correlation is equal to covariance divided by standard deviations ( $\rightarrow$  I/1.17.1), we can conclude that the correlation of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  is also zero:

$$\text{Corr}(\hat{\beta}_0, \hat{\beta}_1) = 0. \quad (4)$$

■

### 1.4.9 Effects of mean-centering

**Theorem:** In simple linear regression ( $\rightarrow$  III/1.4.1), when the dependent variable  $y$  and/or the independent variable  $x$  are mean-centered ( $\rightarrow$  I/1.12.1), the ordinary least squares ( $\rightarrow$  III/1.4.3) estimate for the intercept changes, but that of the slope does not.

**Proof:**

1) Under unaltered  $y$  and  $x$ , ordinary least squares estimates for simple linear regression ( $\rightarrow$  III/1.4.3) are

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \quad (1)$$

with sample means ( $\rightarrow$  I/1.12.2)  $\bar{x}$  and  $\bar{y}$ , sample variance ( $\rightarrow$  I/1.13.2)  $s_x^2$  and sample covariance ( $\rightarrow$  I/1.15.2)  $s_{xy}$ , such that  $\hat{\beta}_0$  estimates “the mean  $y$  at  $x = 0$ ”.

2) Let  $\tilde{x}$  be the mean-centered covariate vector ( $\rightarrow$  III/1.4.1):

$$\tilde{x}_i = x_i - \bar{x} \quad \Rightarrow \quad \bar{\tilde{x}} = 0 . \quad (2)$$

Under this condition, the parameter estimates become

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{\tilde{x}} \\ &= \bar{y} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})(y_i - \bar{y})}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \quad (3)$$

and we can see that  $\hat{\beta}_1(\tilde{x}, y) = \hat{\beta}_1(x, y)$ , but  $\hat{\beta}_0(\tilde{x}, y) \neq \hat{\beta}_0(x, y)$ , specifically  $\beta_0$  now estimates “the mean  $y$  at the mean  $x$ ”.

3) Let  $\tilde{y}$  be the mean-centered data vector ( $\rightarrow$  III/1.4.1):

$$\tilde{y}_i = y_i - \bar{y} \quad \Rightarrow \quad \bar{\tilde{y}} = 0 . \quad (4)$$

Under this condition, the parameter estimates become

$$\begin{aligned} \hat{\beta}_0 &= \bar{\tilde{y}} - \hat{\beta}_1 \bar{x} \\ &= -\hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \quad (5)$$

and we can see that  $\hat{\beta}_1(x, \tilde{y}) = \hat{\beta}_1(x, y)$ , but  $\hat{\beta}_0(x, \tilde{y}) \neq \hat{\beta}_0(x, y)$ , specifically  $\beta_0$  now estimates “the mean  $x$ , multiplied with the negative slope”.

4) Finally, consider mean-centering both  $x$  and  $y$ :

$$\begin{aligned} \tilde{x}_i &= x_i - \bar{x} \quad \Rightarrow \quad \bar{\tilde{x}} = 0 \\ \tilde{y}_i &= y_i - \bar{y} \quad \Rightarrow \quad \bar{\tilde{y}} = 0 . \end{aligned} \quad (6)$$

Under this condition, the parameter estimates become

$$\begin{aligned} \hat{\beta}_0 &= \bar{\tilde{y}} - \hat{\beta}_1 \bar{\tilde{x}} \\ &= 0 \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \quad (7)$$

and we can see that  $\hat{\beta}_1(\tilde{x}, \tilde{y}) = \hat{\beta}_1(x, y)$ , but  $\hat{\beta}_0(\tilde{x}, \tilde{y}) \neq \hat{\beta}_0(x, y)$ , specifically  $\beta_0$  is now forced to become zero.

■

### 1.4.10 Regression line

**Definition:** Let there be a simple linear regression with independent observations ( $\rightarrow$  III/1.4.1) using dependent variable  $y$  and independent variable  $x$ :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) . \quad (1)$$

Then, given some parameters  $\beta_0, \beta_1 \in \mathbb{R}$ , the set

$$L(\beta_0, \beta_1) = \{(x, y) \in \mathbb{R}^2 \mid y = \beta_0 + \beta_1 x\} \quad (2)$$

is called a “regression line” and the set

$$L(\hat{\beta}_0, \hat{\beta}_1) \quad (3)$$

is called the “fitted regression line”, with estimated regression coefficients  $\hat{\beta}_0, \hat{\beta}_1$ , e.g. obtained via ordinary least squares ( $\rightarrow$  III/1.4.3).

**Sources:**

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Fitting\\_the\\_regression\\_line](https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line).

### 1.4.11 Regression line includes center of mass

**Theorem:** In simple linear regression ( $\rightarrow$  III/1.4.1), the regression line ( $\rightarrow$  III/1.4.10) estimated using ordinary least squares ( $\rightarrow$  III/1.4.3) includes the point  $M(\bar{x}, \bar{y})$ .

**Proof:** The fitted regression line ( $\rightarrow$  III/1.4.10) is described by the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{where} \quad x, y \in \mathbb{R} . \quad (1)$$

Plugging in the coordinates of  $M$  and the ordinary least squares estimate of the intercept ( $\rightarrow$  III/1.4.3), we obtain

$$\begin{aligned} \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \bar{y} &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \\ \bar{y} &= \bar{y} . \end{aligned} \quad (2)$$

which is a true statement. Thus, the regression line ( $\rightarrow$  III/1.4.10) goes through the center of mass point  $(\bar{x}, \bar{y})$ , if the model ( $\rightarrow$  III/1.4.1) includes an intercept term  $\beta_0$ . ■

**Sources:**

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Numerical\\_properties](https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties).

### 1.4.12 Projection of data point to regression line

**Theorem:** Consider simple linear regression ( $\rightarrow$  III/1.4.1) and an estimated regression line ( $\rightarrow$  III/1.4.10) specified by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{where} \quad x, y \in \mathbb{R} . \quad (1)$$

For any given data point  $O(x_o|y_o)$ , the point on the regression line  $P(x_p|y_p)$  that is closest to this data point is given by:

$$P(w | \hat{\beta}_0 + \hat{\beta}_1 w) \quad \text{with} \quad w = \frac{x_o + (y_o - \hat{\beta}_0)\hat{\beta}_1}{1 + \hat{\beta}_1^2} . \quad (2)$$

**Proof:** The intersection point of the regression line ( $\rightarrow$  III/1.4.10) with the y-axis is

$$S(0|\hat{\beta}_0) . \quad (3)$$

Let  $a$  be a vector describing the direction of the regression line, let  $b$  be the vector pointing from  $S$  to  $O$  and let  $p$  be the vector pointing from  $S$  to  $P$ .

Because  $\hat{\beta}_1$  is the slope of the regression line, we have

$$a = \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix} . \quad (4)$$

Moreover, with the points  $O$  and  $S$ , we have

$$b = \begin{pmatrix} x_o \\ y_o \end{pmatrix} - \begin{pmatrix} 0 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} x_o \\ y_o - \hat{\beta}_0 \end{pmatrix} . \quad (5)$$

Because  $P$  is located on the regression line,  $p$  is collinear with  $a$  and thus a scalar multiple of this vector:

$$p = w \cdot a . \quad (6)$$

Moreover, as  $P$  is the point on the regression line which is closest to  $O$ , this means that the vector  $b - p$  is orthogonal to  $a$ , such that the inner product of these two vectors is equal to zero:

$$a^T(b - p) = 0 . \quad (7)$$

Rearranging this equation gives

$$\begin{aligned} a^T(b - p) &= 0 \\ a^T(b - w \cdot a) &= 0 \\ a^T b - w \cdot a^T a &= 0 \\ w \cdot a^T a &= a^T b \\ w &= \frac{a^T b}{a^T a} . \end{aligned} \quad (8)$$

With (4) and (5),  $w$  can be calculated as

$$\begin{aligned}
w &= \frac{a^T b}{a^T a} \\
w &= \frac{\begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}^T \begin{pmatrix} x_o \\ y_o - \hat{\beta}_0 \end{pmatrix}}{\begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}^T \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}} \\
w &= \frac{x_o + (y_o - \hat{\beta}_0)\hat{\beta}_1}{1 + \hat{\beta}_1^2}
\end{aligned} \tag{9}$$

Finally, with the point  $S$  (3) and the vector  $p$  (6), the coordinates of  $P$  are obtained as

$$\begin{pmatrix} x_p \\ y_p \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\beta}_0 \end{pmatrix} + w \cdot \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} w \\ \hat{\beta}_0 + \hat{\beta}_1 w \end{pmatrix}. \tag{10}$$

Together, (10) and (9) constitute the proof of equation (2). ■

#### Sources:

- Penny, William (2006): “Projections”; in: *Mathematics for Brain Imaging*, ch. 1.4.10, pp. 34-35, eqs. 1.87/1.88; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).

#### 1.4.13 Sums of squares

**Theorem:** Under ordinary least squares ( $\rightarrow$  III/1.4.3) for simple linear regression ( $\rightarrow$  III/1.4.1), total ( $\rightarrow$  III/1.5.7), explained ( $\rightarrow$  III/1.5.8) and residual ( $\rightarrow$  III/1.5.9) sums of squares are given by

$$\begin{aligned}
\text{TSS} &= (n-1) s_y^2 \\
\text{ESS} &= (n-1) \frac{s_{xy}^2}{s_x^2} \\
\text{RSS} &= (n-1) \left( s_y^2 - \frac{s_{xy}^2}{s_x^2} \right)
\end{aligned} \tag{1}$$

where  $s_x^2$  and  $s_y^2$  are the sample variances ( $\rightarrow$  I/1.13.2) of  $x$  and  $y$  and  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) between  $x$  and  $y$ .

**Proof:** The ordinary least squares parameter estimates ( $\rightarrow$  III/1.4.3) are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \tag{2}$$

1) The total sum of squares ( $\rightarrow$  III/1.5.7) is defined as

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \tag{3}$$

which can be reformulated as follows:

$$\begin{aligned}
 \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
 &= (n-1) \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\
 &= (n-1) s_y^2 .
 \end{aligned} \tag{4}$$

2) The explained sum of squares ( $\rightarrow$  III/1.5.8) is defined as

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{5}$$

which, with the OLS parameter estimates, becomes:

$$\begin{aligned}
 \text{ESS} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\
 &\stackrel{(2)}{=} \sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\
 &= \sum_{i=1}^n \left( \hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\
 &\stackrel{(2)}{=} \sum_{i=1}^n \left( \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 \\
 &= \left( \frac{s_{xy}}{s_x^2} \right)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= \left( \frac{s_{xy}}{s_x^2} \right)^2 (n-1) s_x^2 \\
 &= (n-1) \frac{s_{xy}^2}{s_x^2} .
 \end{aligned} \tag{6}$$

3) The residual sum of squares ( $\rightarrow$  III/1.5.9) is defined as

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{7}$$

which, with the OLS parameter estimates, becomes:

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&\stackrel{(2)}{=} \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^n \left( (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\
&= \sum_{i=1}^n \left( (y_i - \bar{y})^2 - 2\hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 (x_i - \bar{x})^2 \right) \\
&= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= (n-1) s_y^2 - 2(n-1) \hat{\beta}_1 s_{xy} + (n-1) \hat{\beta}_1^2 s_x^2 \\
&\stackrel{(2)}{=} (n-1) s_y^2 - 2(n-1) \left( \frac{s_{xy}}{s_x^2} \right) s_{xy} + (n-1) \left( \frac{s_{xy}}{s_x^2} \right)^2 s_x^2 \\
&= (n-1) s_y^2 - (n-1) \frac{s_{xy}^2}{s_x^2} \\
&= (n-1) \left( s_y^2 - \frac{s_{xy}^2}{s_x^2} \right).
\end{aligned} \tag{8}$$

■

#### 1.4.14 Partition of sums of squares

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

where  $\beta_0$  and  $\beta_1$  are intercept and slope parameter ( $\rightarrow$  III/1.4.1), respectively. Then, it holds that

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{2}$$

where TSS is the total sum of squares ( $\rightarrow$  III/1.5.7), ESS is the explained sum of squares ( $\rightarrow$  III/1.5.8) and RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9).

**Proof:** For simple linear regression, total, explained and residual sum squares are given by ( $\rightarrow$  III/1.4.13)



$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
\text{ESS} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\
\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned} \tag{3}$$

where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the estimated regression coefficients obtained via ordinary least squares ( $\rightarrow$  III/1.4.3)

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{4}$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2) of  $x$  and  $y$ ,  $s_{xy}$  is the unbiased sample covariance ( $\rightarrow$  I/1.15.2) of  $x$  and  $y$  and  $s_x^2$  is the unbiased sample variance ( $\rightarrow$  I/1.15.2) of  $x$ :

$$\begin{aligned}
s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.
\end{aligned} \tag{5}$$

With that in mind, we start working out the total sum of squares:

$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^n ((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2) \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&\stackrel{(3)}{=} \text{ESS} + \text{RSS} + \sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).
\end{aligned} \tag{6}$$

Thus, what remains to be shown is that the following sum is zero:

$$\sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad (7)$$

Using the expression  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  for the fitted signal values ( $\rightarrow$  III/1.4.10), we proceed as follows:

$$\begin{aligned} \sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) \\ &\stackrel{(4)}{=} \sum_{i=1}^n 2(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y}) \\ &= \sum_{i=1}^n 2 \left( (y_i - \bar{y}) - (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) \right) (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) \\ &= 2 \sum_{i=1}^n \left( (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right) \hat{\beta}_1 (x_i - \bar{x}) \\ &= 2 \sum_{i=1}^n \left( (y_i - \hat{y}_i) \hat{\beta}_1 (x_i - \bar{x}) - \hat{\beta}_1 (x_i - \bar{x}) \hat{\beta}_1 (x_i - \bar{x}) \right) \\ &= 2 \left[ \hat{\beta}_1 \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) \right]. \end{aligned} \quad (8)$$

Next, we recognize the sample covariance and sample variance terms from (5):

$$\begin{aligned} \sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2 \left[ \hat{\beta}_1 (n-1) s_{xy} - \hat{\beta}_1^2 (n-1) s_x^2 \right] \\ &= 2(n-1) \left[ \hat{\beta}_1 s_{xy} - \hat{\beta}_1^2 s_x^2 \right]. \end{aligned} \quad (9)$$

Now, we can apply to functional form of the estimate  $\hat{\beta}_1$  from (4) to get:

$$\begin{aligned} \sum_{i=1}^n 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2(n-1) \left[ \left( \frac{s_{xy}}{s_x^2} \right) s_{xy} - \left( \frac{s_{xy}}{s_x^2} \right)^2 s_x^2 \right] \\ &= 2(n-1) \left[ \frac{s_{xy}^2}{s_x^2} - \frac{s_{xy}^2}{s_x^2} \right] \\ &= 2(n-1) \cdot 0 \\ &= 0. \end{aligned} \quad (10)$$

Plugging the result from (10) into (6), we finally get:

$$\text{TSS} = \text{ESS} + \text{RSS}. \quad (11)$$

**Sources:**



- Ostwald, Dirk (2023): “Korrelation”; in: *Allgemeines Lineares Modell*, Einheit (2), Folien 19-23; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/2\\_Korrelation.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/2_Korrelation.pdf).
- Manabu, Hayashi (2021): “TSS = RSS + ESS | Simple Linear Regression”; in: *You Tube*, retrieved on 2024-07-12; URL: <https://www.youtube.com/watch?v=N7pHym1L9b0>.

#### 1.4.15 Transformation matrices

**Theorem:** Under ordinary least squares ( $\rightarrow$  III/1.4.3) for simple linear regression ( $\rightarrow$  III/1.4.1), estimation ( $\rightarrow$  III/1.5.11), projection ( $\rightarrow$  III/1.5.12) and residual-forming ( $\rightarrow$  III/1.5.13) matrices are given by

$$\begin{aligned} E &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) 1_n^T - \bar{x} x^T \\ -\bar{x} 1_n^T + x^T \end{bmatrix} \\ P &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^T x/n) - \bar{x}(x_1 + x_n) + x_1x_n \\ \vdots & \ddots & \vdots \\ (x^T x/n) - \bar{x}(x_1 + x_n) + x_1x_n & \cdots & (x^T x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\ R &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (n-1)(x^T x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1x_n \\ \vdots & \ddots & \vdots \\ -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1x_n & \cdots & (n-1)(x^T x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix} \end{aligned} \quad (1)$$

where  $1_n$  is an  $n \times 1$  vector of ones,  $x$  is the  $n \times 1$  single predictor variable,  $\bar{x}$  is the sample mean ( $\rightarrow$  I/1.12.2) of  $x$  and  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$ .

**Proof:** Simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad (2)$$

such that the simple linear regression model can also be written as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \quad (3)$$

Moreover, we note the following equality ( $\rightarrow$  III/1.4.7):

$$x^T x - n\bar{x}^2 = (n-1)s_x^2. \quad (4)$$

1) The estimation matrix is given by ( $\rightarrow$  III/1.5.14)

$$E = (X^T X)^{-1} X^T \quad (5)$$

which is a  $2 \times n$  matrix and can be reformulated as follows:

$$\begin{aligned}
E &= (X^T X)^{-1} X^T \\
&= \left( \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \\
&= \left( \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^T x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \\
&= \frac{1}{nx^T x - (n\bar{x})^2} \begin{bmatrix} x^T x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \\
&= \frac{1}{x^T x - n\bar{x}^2} \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \\
&\stackrel{(4)}{=} \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) 1_n^T - \bar{x} x^T \\ -\bar{x} 1_n^T + x^T \end{bmatrix}.
\end{aligned} \tag{6}$$

2) The projection matrix is given by ( $\rightarrow$  III/1.5.14)

$$P = X(X^T X)^{-1} X^T = X E \tag{7}$$

which is an  $n \times n$  matrix and can be reformulated as follows:

$$\begin{aligned}
P = X E &= \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \\
&= \frac{1}{(n-1)s_x^2} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} (x^T x/n) - \bar{x}x_1 & \cdots & (x^T x/n) - \bar{x}x_n \\ -\bar{x} + x_1 & \cdots & -\bar{x} + x_n \end{bmatrix} \\
&= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^T x/n) - \bar{x}(x_1 + x_n) + x_1x_n \\ \vdots & \ddots & \vdots \\ (x^T x/n) - \bar{x}(x_1 + x_n) + x_1x_n & \cdots & (x^T x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix}.
\end{aligned} \tag{8}$$

3) The residual-forming matrix is given by ( $\rightarrow$  III/1.5.14)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P \tag{9}$$

which also is an  $n \times n$  matrix and can be reformulated as follows:

$$\begin{aligned}
R = I_n - P &= \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \\
&\stackrel{(4)}{=} \frac{1}{(n-1)s_x^2} \begin{bmatrix} x^T x - n\bar{x}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x^T x - n\bar{x}^2 \end{bmatrix} \\
&\quad - \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n & \cdots & (x^T x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\
&= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (n-1)(x^T x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n \\ \vdots & \ddots & \vdots \\ -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n & \cdots & (n-1)(x^T x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix}.
\end{aligned} \tag{10}$$

■

#### 1.4.16 Weighted least squares

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.4.1) with correlated observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \tag{1}$$

the parameters minimizing the weighted residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\begin{aligned}
\hat{\beta}_0 &= \frac{x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \\
\hat{\beta}_1 &= \frac{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} y - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} y}{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} x - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} x}
\end{aligned} \tag{2}$$

where  $\mathbf{1}_n$  is an  $n \times 1$  vector of ones.

**Proof:** Let there be an  $n \times n$  square matrix  $W$ , such that

$$W V W^T = I_n. \tag{3}$$

Since  $V$  is a covariance matrix and thus symmetric ( $\rightarrow$  I/1.15.4),  $W$  is also symmetric and can be expressed as the matrix square root of the inverse of  $V$ :

$$W V W = I_n \quad \Leftrightarrow \quad V = W^{-1} W^{-1} \quad \Leftrightarrow \quad V^{-1} = W W \quad \Leftrightarrow \quad W = V^{-1/2}. \tag{4}$$

Because  $\beta_0$  is a scalar, (1) may also be written as

$$y = \beta_0 \mathbf{1}_n + \beta_1 x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad (5)$$

Left-multiplying (5) with  $W$ , the linear transformation theorem ( $\rightarrow$  II/4.1.14) implies that

$$Wy = \beta_0 W\mathbf{1}_n + \beta_1 Wx + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WW^T). \quad (6)$$

Applying (3), we see that (6) is actually a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$\tilde{y} = \begin{bmatrix} \tilde{x}_0 & \tilde{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (7)$$

where  $\tilde{y} = Wy$ ,  $\tilde{x}_0 = W\mathbf{1}_n$ ,  $\tilde{x} = Wx$  and  $\tilde{\varepsilon} = W\varepsilon$ , such that we can apply the ordinary least squares solution ( $\rightarrow$  III/1.5.3) giving:

$$\begin{aligned} \hat{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\ &= \left( \begin{bmatrix} \tilde{x}_0^T \\ \tilde{x}^T \end{bmatrix} \begin{bmatrix} \tilde{x}_0 & \tilde{x} \end{bmatrix} \right)^{-1} \begin{bmatrix} \tilde{x}_0^T \\ \tilde{x}^T \end{bmatrix} \tilde{y} \\ &= \begin{bmatrix} \tilde{x}_0^T \tilde{x}_0 & \tilde{x}_0^T \tilde{x} \\ \tilde{x}^T \tilde{x}_0 & \tilde{x}^T \tilde{x} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}_0^T \\ \tilde{x}^T \end{bmatrix} \tilde{y}. \end{aligned} \quad (8)$$

Applying the inverse of a  $2 \times 2$  matrix, this reformulates to:

$$\begin{aligned} \hat{\beta} &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} & -\tilde{x}_0^T \tilde{x} \\ -\tilde{x}^T \tilde{x}_0 & \tilde{x}_0^T \tilde{x}_0 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}_0^T \\ \tilde{x}^T \end{bmatrix} \tilde{y} \\ &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} \tilde{x}_0^T - \tilde{x}_0^T \tilde{x} \tilde{x}^T \\ \tilde{x}_0^T \tilde{x}_0 \tilde{x}^T - \tilde{x}^T \tilde{x}_0 \tilde{x}_0^T \end{bmatrix} \tilde{y} \\ &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} \tilde{x}_0^T \tilde{y} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{y} \\ \tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{y} - \tilde{x}^T \tilde{x}_0 \tilde{x}_0^T \tilde{y} \end{bmatrix}. \end{aligned} \quad (9)$$

Reinserting  $\tilde{x}_0 = W\mathbf{1}_n$ ,  $\tilde{x} = Wx$  and  $W^T W = WW^T = V^{-1}$ , we finally have

$$\begin{aligned} \hat{\beta} &= \frac{1}{\mathbf{1}_n^T W^T W \mathbf{1}_n x^T W^T W x - \mathbf{1}_n^T W^T W x x^T W^T W \mathbf{1}_n} \begin{bmatrix} x^T W^T W x \mathbf{1}_n^T W^T W y - \mathbf{1}_n^T W^T W x x^T W^T W y \\ \mathbf{1}_n^T W^T W \mathbf{1}_n x^T W^T W y - x^T W^T W \mathbf{1}_n \mathbf{1}_n^T W^T W y \end{bmatrix} \\ &= \frac{1}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \begin{bmatrix} x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y \\ \mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} y - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} y \end{bmatrix} \\ &= \begin{bmatrix} \frac{x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \\ \frac{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} y - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} y}{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} x - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} x} \end{bmatrix} \end{aligned} \quad (10)$$

which corresponds to the weighted least squares solution (2).

■

### 1.4.17 Weighted least squares

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.4.1) with correlated observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad (1)$$

the parameters minimizing the weighted residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\begin{aligned} \hat{\beta}_0 &= \frac{x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \\ \hat{\beta}_1 &= \frac{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} y - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} y}{\mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} x - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} x} \end{aligned} \quad (2)$$

where  $\mathbf{1}_n$  is an  $n \times 1$  vector of ones.

**Proof:** Simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2) with

$$X = \begin{bmatrix} \mathbf{1}_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (3)$$

and weighted least squares estimates ( $\rightarrow$  III/1.5.21) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \quad (4)$$

Writing out equation (4), we have

$$\begin{aligned} \hat{\beta} &= \left( \begin{bmatrix} \mathbf{1}_n^T \\ x^T \end{bmatrix} V^{-1} \begin{bmatrix} \mathbf{1}_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{1}_n^T \\ x^T \end{bmatrix} V^{-1} y \\ &= \begin{bmatrix} \mathbf{1}_n^T V^{-1} \mathbf{1}_n & \mathbf{1}_n^T V^{-1} x \\ x^T V^{-1} \mathbf{1}_n & x^T V^{-1} x \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n^T V^{-1} y \\ x^T V^{-1} y \end{bmatrix} \\ &= \frac{1}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \begin{bmatrix} x^T V^{-1} x & -\mathbf{1}_n^T V^{-1} x \\ -x^T V^{-1} \mathbf{1}_n & \mathbf{1}_n^T V^{-1} \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mathbf{1}_n^T V^{-1} y \\ x^T V^{-1} y \end{bmatrix} \\ &= \frac{1}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n} \begin{bmatrix} x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y \\ \mathbf{1}_n^T V^{-1} \mathbf{1}_n x^T V^{-1} y - x^T V^{-1} \mathbf{1}_n \mathbf{1}_n^T V^{-1} y \end{bmatrix}. \end{aligned} \quad (5)$$

Thus, the first entry of  $\hat{\beta}$  is equal to ( $\rightarrow$  III/1.4.16):

$$\hat{\beta}_0 = \frac{x^T V^{-1} x \mathbf{1}_n^T V^{-1} y - \mathbf{1}_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x \mathbf{1}_n^T V^{-1} \mathbf{1}_n - \mathbf{1}_n^T V^{-1} x x^T V^{-1} \mathbf{1}_n}. \quad (6)$$

Moreover, the second entry of  $\hat{\beta}$  is equal to ( $\rightarrow$  III/1.4.16):

$$\hat{\beta}_1 = \frac{1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y}{1_n^T V^{-1} 1_n x^T V^{-1} x - x^T V^{-1} 1_n 1_n^T V^{-1} x} . \quad (7)$$

■

#### 1.4.18 Maximum likelihood estimation

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\beta_0$ ,  $\beta_1$  and  $\sigma^2$  are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2),  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) between  $x$  and  $y$ .

**Proof:** With the probability density function of the normal distribution ( $\rightarrow$  II/3.2.10) and probability under independence ( $\rightarrow$  I/1.3.6), the linear regression equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned} p(y|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n p(y_i|\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^n \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[ -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right] \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right] \end{aligned} \quad (3)$$

and the log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\begin{aligned} \text{LL}(\beta_0, \beta_1, \sigma^2) &= \log p(y|\beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 . \end{aligned} \quad (4)$$

The derivative of the log-likelihood function (4) with respect to  $\beta_0$  is



$$\frac{dLL(\beta_0, \beta_1, \sigma^2)}{d\beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \quad (5)$$

and setting this derivative to zero gives the MLE for  $\beta_0$ :

$$\begin{aligned} \frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\beta_0} &= 0 \\ 0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ 0 &= \sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i \\ \hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} . \end{aligned} \quad (6)$$

The derivative of the log-likelihood function (4) at  $\hat{\beta}_0$  with respect to  $\beta_1$  is

$$\frac{dLL(\hat{\beta}_0, \beta_1, \sigma^2)}{d\beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \beta_1 x_i^2) \quad (7)$$

and setting this derivative to zero gives the MLE for  $\beta_1$ :

$$\begin{aligned} \frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\beta_1} &= 0 \\ 0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \\ 0 &= \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\ 0 &\stackrel{(6)}{=} \sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\ 0 &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\ 0 &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} . \end{aligned} \quad (8)$$

The derivative of the log-likelihood function (4) at  $(\hat{\beta}_0, \hat{\beta}_1)$  with respect to  $\sigma^2$  is

$$\frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (9)$$

and setting this derivative to zero gives the MLE for  $\sigma^2$ :

$$\begin{aligned} \frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\hat{\sigma}^2} &= 0 \\ 0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ \frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \quad (10)$$

Together, (6), (8) and (10) constitute the MLE for simple linear regression. ■

#### 1.4.19 Maximum likelihood estimation

**Theorem:** Given a simple linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\beta_0$ ,  $\beta_1$  and  $\sigma^2$  are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2),  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of  $x$  and  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) between  $x$  and  $y$ .

**Proof:** Simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (3)$$

and maximum likelihood estimates ( $\rightarrow$  III/1.5.23) are given by

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta}). \end{aligned} \quad (4)$$

Under independent observations, the covariance matrix is

$$V = I_n, \quad \text{such that} \quad V^{-1} = I_n. \quad (5)$$

Thus, we can write out the estimate of  $\beta$

$$\begin{aligned} \hat{\beta} &= \left( \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} y \\ &= \left( \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} y \end{aligned} \quad (6)$$

which is equal to the ordinary least squares solution for simple linear regression ( $\rightarrow$  III/1.4.4):

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}. \end{aligned} \quad (7)$$

Additionally, we can write out the estimate of  $\sigma^2$ :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= \frac{1}{n} \left( y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right)^T \left( y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right) \\ &= \frac{1}{n} \left( y - \hat{\beta}_0 - \hat{\beta}_1 x \right)^T \left( y - \hat{\beta}_0 - \hat{\beta}_1 x \right) \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \quad (8)$$

■

#### 1.4.20 t-test for intercept parameter

**Theorem:** Consider a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

and the parameter estimates ( $\rightarrow$  III/1.4.18)

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2) of the  $x_i$  and  $y_i$ ,  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) of the  $x_i$  and  $y_i$  and  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of the  $x_i$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t_0 = \frac{\bar{y} - \hat{\beta}_1 \bar{x}}{\sqrt{\hat{\sigma}^2} \sigma_0} \quad (3)$$

with  $\sigma_0$  equal to the first diagonal element of the parameter covariance matrix ( $\rightarrow$  III/1.4.7)

$$\sigma_0 = \frac{x^T x / n}{(n-1) s_x^2} \quad \text{where} \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4)$$

follows a t-distribution ( $\rightarrow$  II/3.3.1)

$$t_0 \sim t(n-2) \quad (5)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the intercept parameter ( $\rightarrow$  III/1.4.1) is zero:

$$H_0 : \beta_0 = 0. \quad (6)$$

**Proof:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the contrast-based t-test ( $\rightarrow$  III/1.5.28) is based on the t-statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T V^{-1} X)^{-1} c}} \quad (7)$$

which follows a t-distribution ( $\rightarrow$  II/3.3.1) under the null hypothesis ( $\rightarrow$  I/4.3.2) that the scalar product of the contrast vector ( $\rightarrow$  III/1.5.26) and the regression coefficients ( $\rightarrow$  III/1.5.1) is zero:

$$t \sim t(n-p), \quad \text{if} \quad c^T \beta = 0. \quad (8)$$

Since simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2), in the present case we have the following quantities:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad c_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad X = \begin{bmatrix} 1_n & x \end{bmatrix}, \quad V = I_n. \quad (9)$$

Thus, we have the null hypothesis

$$H_0 : c_0^T \beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_0 = 0 \quad (10)$$

and the contrast estimate

$$c_0^T \hat{\beta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (11)$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations ( $\rightarrow$  III/1.4.7), we have identified the parameter covariance matrix as

$$(X^T X)^{-1} = \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}. \quad (12)$$

Plugging (9), (11), (12) and (2) into (7), the test statistic becomes

$$\begin{aligned} t_0 &= \frac{c_0^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c_0^T (X^T X)^{-1} c_0}} \\ &= \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^T}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 1 & 0 \end{bmatrix} (X^T X)^{-1} \begin{bmatrix} 1 & 0 \end{bmatrix}^T}} \\ &= \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^T}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 1 & 0 \end{bmatrix} \left( \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 \end{bmatrix}^T}} \\ &= \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2 \left( \frac{x^T x/n}{(n-1)s_x^2} \right)}} \\ &= \frac{\bar{y} - \hat{\beta}_1 \bar{x}}{\sqrt{\hat{\sigma}^2} \sigma_0}. \end{aligned} \quad (13)$$

Finally, because  $X = \begin{bmatrix} 1_n & x \end{bmatrix}$  is an  $n \times 2$  matrix, we have  $p = 2$ , such that from (8), it follows that

$$t_0 \sim t(n-2), \quad \text{if } \beta_0 = 0. \quad (14)$$

■

#### 1.4.21 t-test for slope parameter

**Theorem:** Consider a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

and the parameter estimates ( $\rightarrow$  III/1.4.18)

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2) of the  $x_i$  and  $y_i$ ,  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) of the  $x_i$  and  $y_i$  and  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of the  $x_i$ .

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t_1 = \frac{s_{xy}/s_x^2}{\sqrt{\hat{\sigma}^2} \sigma_1} \quad (3)$$

with  $\sigma_1$  equal to the first diagonal element of the parameter covariance matrix ( $\rightarrow$  III/1.4.7)

$$\sigma_1 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

follows a t-distribution ( $\rightarrow$  II/3.3.1)

$$t_1 \sim t(n-2) \quad (5)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the slope parameter ( $\rightarrow$  III/1.4.1) is zero:

$$H_0 : \beta_1 = 0 . \quad (6)$$

**Proof:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the contrast-based t-test ( $\rightarrow$  III/1.5.28) is based on the t-statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T V^{-1} X)^{-1} c}} \quad (7)$$

which follows a t-distribution ( $\rightarrow$  II/3.3.1) under the null hypothesis ( $\rightarrow$  I/4.3.2) that the scalar product of the contrast vector ( $\rightarrow$  III/1.5.26) and the regression coefficients ( $\rightarrow$  III/1.5.1) is zero:

$$t \sim t(n-p), \quad \text{if } c^T \beta = 0 . \quad (8)$$

Since simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2), in the present case we have the following quantities:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad c_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad X = \begin{bmatrix} 1_n & x \end{bmatrix}, \quad V = I_n . \quad (9)$$

Thus, we have the null hypothesis

$$H_0 : c_1^T \beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0 \quad (10)$$

and the contrast estimate

$$c_1^T \hat{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} . \quad (11)$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations ( $\rightarrow$  III/1.4.7), we have identified the parameter covariance matrix as

$$(X^T X)^{-1} = \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}. \quad (12)$$

Plugging (9), (11), (12) and (2) into (7), the test statistic becomes

$$\begin{aligned} t_1 &= \frac{c_1^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c_1^T (X^T X)^{-1} c_1}} \\ &= \frac{\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^T}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} (X^T X)^{-1} \begin{bmatrix} 0 & 1 \end{bmatrix}^T}} \\ &= \frac{\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^T}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \left( \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 0 & 1 \end{bmatrix}^T}} \\ &= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 \left( \frac{1}{(n-1)s_x^2} \right)}} \\ &= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \frac{s_{xy}/s_x^2}{\sqrt{\hat{\sigma}^2} \sigma_1}. \end{aligned} \quad (13)$$

Finally, because  $X = \begin{bmatrix} 1_n & x \end{bmatrix}$  is an  $n \times 2$  matrix, we have  $p = 2$ , such that from (8) it follows that

$$t_1 \sim t(n-2), \quad \text{if } \beta_1 = 0. \quad (14)$$

■

#### 1.4.22 F-test for model comparison

**Theorem:** Consider a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

and the parameter estimates ( $\rightarrow$  III/1.4.18)

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \end{aligned} \quad (2)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means ( $\rightarrow$  I/1.12.2) of the  $x_i$  and  $y_i$ ,  $s_{xy}$  is the sample covariance ( $\rightarrow$  I/1.15.2) of the  $x_i$  and  $y_i$  and  $s_x^2$  is the sample variance ( $\rightarrow$  I/1.13.2) of the  $x_i$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F = \frac{s_{xy}^2/s_x^2}{\hat{\sigma}^2/(n-1)} \quad (3)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F \sim F(1, n-2) \quad (4)$$

under the scenario that the data were generated using a model in which the slope parameter ( $\rightarrow$  III/1.4.1) is zero:

$$H_0 : \beta_1 = 0. \quad (5)$$

**Proof:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the contrast-based F-test ( $\rightarrow$  III/1.5.29) is based on the F-statistic ( $\rightarrow$  I/4.3.5)

$$F = \hat{\beta}^T C (\hat{\sigma}^2 C^T (X^T V^{-1} X)^{-1} C)^{-1} C^T \hat{\beta} / q \quad (6)$$

which follows an F-distribution ( $\rightarrow$  II/3.8.1) under the null hypothesis ( $\rightarrow$  I/4.3.2) that the product of the contrast matrix ( $\rightarrow$  III/1.5.27)  $C \in \mathbb{R}^{p \times q}$  and the regression coefficients ( $\rightarrow$  III/1.5.1) is a zero vector:

$$F \sim F(q, n-p), \quad \text{if } C^T \beta = 0_q = [0, \dots, 0]^T. \quad (7)$$

Since simple linear regression is a special case of multiple linear regression ( $\rightarrow$  III/1.4.2), we have the following quantities, if we want to compare the regression model against a model without the slope parameter:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad X = \begin{bmatrix} 1_n & x \end{bmatrix}, \quad V = I_n. \quad (8)$$

Thus, we have the null hypothesis

$$H_0 : C^T \beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0 \quad (9)$$

and the contrast estimate

$$C^T \hat{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \quad (10)$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations ( $\rightarrow$  III/1.4.7), we have identified the parameter covariance matrix as



$$(X^T X)^{-1} = \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}. \quad (11)$$

Plugging (8), (10), (11) and (2) into (6), the test statistic becomes

$$\begin{aligned} F &= \hat{\beta}^T C (\hat{\sigma}^2 C^T (X^T V^{-1} X)^{-1} C)^{-1} C^T \hat{\beta} / q \\ &= \left( \frac{s_{xy}}{s_x^2} \right) \left( \hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \left( \frac{1}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 0 & 1 \end{bmatrix}^T \right)^{-1} \left( \frac{s_{xy}}{s_x^2} \right) / 1 \\ &= \frac{s_{xy}^2 / (s_x^2)^2}{\hat{\sigma}^2 / ((n-1)s_x^2)} \\ &= \frac{s_{xy}^2 / s_x^2}{\hat{\sigma}^2 / (n-1)}. \end{aligned} \quad (12)$$

Finally, because  $C = \begin{bmatrix} 0 & 1 \end{bmatrix}^T \in \mathbb{R}^{2 \times 1}$  and  $X = \begin{bmatrix} 1_n & x \end{bmatrix} \in \mathbb{R}^{n \times 2}$ , we have  $p = 2$  and  $q = 1$ , such that from (7) it follows that

$$F \sim F(1, n-2), \quad \text{if } \beta_1 = 0. \quad (13)$$

■

### 1.4.23 Sum of residuals is zero

**Theorem:** In simple linear regression ( $\rightarrow$  III/1.4.1), the sum of the residuals ( $\rightarrow$  III/1.5.9) is zero when estimated using ordinary least squares ( $\rightarrow$  III/1.4.3).

**Proof:** The residuals are defined as the estimated error terms ( $\rightarrow$  III/1.4.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (1)$$

where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are parameter estimates obtained using ordinary least squares ( $\rightarrow$  III/1.4.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \quad (2)$$

With that, we can calculate the sum of the residuals:

$$\begin{aligned} \sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n y_i - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i \\ &= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} \\ &= 0. \end{aligned} \quad (3)$$

Thus, the sum of the residuals ( $\rightarrow$  III/1.5.9) is zero under ordinary least squares ( $\rightarrow$  III/1.4.3), if the model ( $\rightarrow$  III/1.4.1) includes an intercept term  $\beta_0$ . ■

**Sources:**

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Numerical\\_properties](https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties).

#### 1.4.24 Correlation with covariate is zero

**Theorem:** In simple linear regression ( $\rightarrow$  III/1.4.1), the residuals ( $\rightarrow$  III/1.5.9) and the covariate ( $\rightarrow$  III/1.4.1) are uncorrelated ( $\rightarrow$  I/1.17.1) when estimated using ordinary least squares ( $\rightarrow$  III/1.4.3).

**Proof:** The residuals are defined as the estimated error terms ( $\rightarrow$  III/1.4.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (1)$$

where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are parameter estimates obtained using ordinary least squares ( $\rightarrow$  III/1.4.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \quad (2)$$

With that, we can calculate the inner product of the covariate and the residuals vector:

$$\begin{aligned}
\sum_{i=1}^n x_i \hat{\varepsilon}_i &= \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \\
&= \sum_{i=1}^n (x_i y_i - x_i (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i^2) \\
&= \sum_{i=1}^n (x_i (y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} x_i - x_i^2)) \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \left( \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) \\
&= \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \right) - \hat{\beta}_1 \left( \sum_{i=1}^n x_i^2 - 2n \bar{x} \bar{x} + n \bar{x}^2 \right) \\
&= \left( \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \right) - \hat{\beta}_1 \left( \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 \right) \\
&= \sum_{i=1}^n (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x} \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i^2 - 2\bar{x} x_i + \bar{x}^2) \\
&= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= (n-1) s_{xy} - \frac{s_{xy}}{s_x^2} (n-1) s_x^2 \\
&= (n-1) s_{xy} - (n-1) s_{xy} \\
&= 0.
\end{aligned} \tag{3}$$

Because an inner product of zero also implies zero correlation ( $\rightarrow$  I/1.17.1), this demonstrates that residuals ( $\rightarrow$  III/1.5.9) and covariate ( $\rightarrow$  III/1.4.1) values are uncorrelated under ordinary least squares ( $\rightarrow$  III/1.4.3).

■

#### Sources:

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Numerical\\_properties](https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties).

#### 1.4.25 Residual variance in terms of sample variance

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, residual variance ( $\rightarrow$  IV/1.0.1) and sample variance ( $\rightarrow$  I/1.13.2) are related to each other via the correlation coefficient ( $\rightarrow$  I/1.17.1):

$$\hat{\sigma}^2 = (1 - r_{xy}^2) s_y^2. \quad (2)$$

**Proof:** The residual variance ( $\rightarrow$  IV/1.0.1) can be expressed in terms of the residual sum of squares ( $\rightarrow$  III/1.5.9):

$$\hat{\sigma}^2 = \frac{1}{n-1} \text{RSS}(\hat{\beta}_0, \hat{\beta}_1) \quad (3)$$

and the residual sum of squares for simple linear regression ( $\rightarrow$  III/1.4.13) is

$$\text{RSS}(\hat{\beta}_0, \hat{\beta}_1) = (n-1) \left( s_y^2 - \frac{s_{xy}^2}{s_x^2} \right). \quad (4)$$

Combining (3) and (4), we obtain:

$$\begin{aligned} \hat{\sigma}^2 &= \left( s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) \\ &= \left( 1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right) s_y^2 \\ &= \left( 1 - \left( \frac{s_{xy}}{s_x s_y} \right)^2 \right) s_y^2. \end{aligned} \quad (5)$$

Using the relationship between correlation, covariance and standard deviation ( $\rightarrow$  I/1.17.1)

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \quad (6)$$

which also holds for sample correlation, sample covariance ( $\rightarrow$  I/1.15.2) and sample standard deviation ( $\rightarrow$  I/1.19.1)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \quad (7)$$

we get the final result:

$$\hat{\sigma}^2 = (1 - r_{xy}^2) s_y^2. \quad (8)$$

■

#### Sources:

- Penny, William (2006): “Relation to correlation”; in: *Mathematics for Brain Imaging*, ch. 1.2.3, p. 18, eq. 1.28; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Numerical\\_properties](https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties).

### 1.4.26 Correlation coefficient in terms of slope estimate

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, correlation coefficient ( $\rightarrow$  I/1.17.4) and the estimated value of the slope parameter ( $\rightarrow$  III/1.4.1) are related to each other via the sample ( $\rightarrow$  I/1.13.2) standard deviations ( $\rightarrow$  I/1.19.1):

$$r_{xy} = \frac{s_x}{s_y} \hat{\beta}_1. \quad (2)$$

**Proof:** The ordinary least squares estimate of the slope ( $\rightarrow$  III/1.4.3) is given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \quad (3)$$

Using the relationship between covariance and correlation ( $\rightarrow$  I/1.15.8)

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y \quad (4)$$

which also holds for sample correlation ( $\rightarrow$  I/1.17.4) and sample covariance ( $\rightarrow$  I/1.15.2)

$$s_{xy} = s_x r_{xy} s_y, \quad (5)$$

we get the final result:

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\beta}_1 &= \frac{s_x r_{xy} s_y}{s_x^2} \\ \hat{\beta}_1 &= \frac{s_y}{s_x} r_{xy} \\ \Leftrightarrow r_{xy} &= \frac{s_x}{s_y} \hat{\beta}_1. \end{aligned} \quad (6)$$

■

#### Sources:

- Penny, William (2006): “Relation to correlation”; in: *Mathematics for Brain Imaging*, ch. 1.2.3, p. 18, eq. 1.27; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Fitting\\_the\\_regression\\_line](https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line).

### 1.4.27 Coefficient of determination in terms of correlation coefficient

**Theorem:** Assume a simple linear regression model ( $\rightarrow$  III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (1)$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.4.3). Then, the coefficient of determination ( $\rightarrow$  IV/1.1.1) is equal to the squared correlation coefficient ( $\rightarrow$  I/1.17.4) between  $x$  and  $y$ :

$$R^2 = r_{xy}^2 . \quad (2)$$

**Proof:** The ordinary least squares estimates for simple linear regression ( $\rightarrow$  III/1.4.3) are

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} . \end{aligned} \quad (3)$$

The coefficient of determination ( $\rightarrow$  IV/1.1.1)  $R^2$  is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data. This can be quantified as the ratio of explained sum of squares ( $\rightarrow$  III/1.5.8) to total sum of squares ( $\rightarrow$  III/1.5.7):

$$R^2 = \frac{\text{ESS}}{\text{TSS}} . \quad (4)$$

Using the explained and total sum of squares for simple linear regression ( $\rightarrow$  III/1.4.13), we have:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} . \end{aligned} \quad (5)$$

By applying (3), we can further develop the coefficient of determination:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n \left( \hat{\beta}_1 (x_i - \bar{x}) \right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{s_x^2}{s_y^2} \\ &= \left( \frac{s_x}{s_y} \hat{\beta}_1 \right)^2 . \end{aligned} \quad (6)$$

Using the relationship between correlation coefficient and slope estimate ( $\rightarrow$  III/1.4.26), we conclude:

$$R^2 = \left( \frac{s_x}{s_y} \hat{\beta}_1 \right)^2 = r_{xy}^2 . \quad (7)$$

■

**Sources:**

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression#Fitting\\_the\\_regression\\_line](https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line).
- Wikipedia (2021): “Coefficient of determination”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Coefficient\\_of\\_determination#As\\_squared\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Coefficient_of_determination#As_squared_correlation_coefficient).
- Wikipedia (2021): “Correlation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: [https://en.wikipedia.org/wiki/Correlation#Sample\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Correlation#Sample_correlation_coefficient).

## 1.5 Multiple linear regression

### 1.5.1 Definition

**Definition:** Let  $y$  be an  $n \times 1$  vector and let  $X$  be an  $n \times p$  matrix. Then, a statement asserting a linear combination of  $X$  into  $y$

$$y = X\beta + \varepsilon, \quad (1)$$

together with a statement asserting a normal distribution ( $\rightarrow$  II/4.1.1) for  $\varepsilon$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (2)$$

is called a univariate linear regression model or simply, “multiple linear regression”.

- $y$  is called “measured data”, “dependent variable” or “measurements”;
- $X$  is called “design matrix”, “set of independent variables” or “predictors”;
- $V$  is called “covariance matrix” or “covariance structure”;
- $\beta$  are called “regression coefficients” or “weights”;
- $\varepsilon$  is called “noise”, “errors” or “error terms”;
- $\sigma^2$  is called “noise variance” or “error variance”;
- $n$  is the number of observations;
- $p$  is the number of predictors.

Alternatively, the linear combination may also be written as

$$y = \sum_{i=1}^p \beta_i x_i + \varepsilon \quad (3)$$

or, when the model includes an intercept term, as

$$y = \beta_0 + \sum_{i=1}^p \beta_i x_i + \varepsilon \quad (4)$$

which is equivalent to adding a constant regressor  $x_0 = 1_n$  to the design matrix  $X$ .

When the covariance structure  $V$  is equal to the  $n \times n$  identity matrix, this is called multiple linear regression with independent and identically distributed ( $\rightarrow$  I/1.2.8) (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \quad (5)$$

Otherwise, it is called multiple linear regression with correlated observations ( $\rightarrow$  I/1.17.1).

**Sources:**

- Wikipedia (2020): “Linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: [https://en.wikipedia.org/wiki/Linear\\_regression#Simple\\_and\\_multiple\\_linear\\_regression](https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression).

### 1.5.2 Special case of general linear model

**Theorem:** Multiple linear regression ( $\rightarrow$  III/1.5.1) is a special case of the general linear model ( $\rightarrow$  III/2.2.1) with number of measurements  $v = 1$ , such that data matrix  $Y$ , regression coefficients  $B$ , noise matrix  $E$  and noise covariance  $\Sigma$  equate as

$$Y = y, \quad B = \beta, \quad E = \varepsilon \quad \text{and} \quad \Sigma = \sigma^2 \quad (1)$$

where  $y$ ,  $\beta$ ,  $\varepsilon$  and  $\sigma^2$  are the data vector, regression coefficients, noise vector and noise variance from multiple linear regression ( $\rightarrow$  III/1.5.1).

**Proof:** The linear regression model with correlated errors ( $\rightarrow$  III/1.5.1) is given by:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (2)$$

Because  $\varepsilon$  is an  $n$ -dimensional vector and  $\sigma^2$  is scalar, we have the following identities:

$$\begin{aligned} \text{vec}(\varepsilon) &= \varepsilon \\ \sigma^2 \otimes V &= \sigma^2 V . \end{aligned} \quad (3)$$

Thus, using the relationship between multivariate normal and matrix normal distribution ( $\rightarrow$  II/5.1.2), equation (2) can also be written as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{MN}(0, V, \sigma^2) . \quad (4)$$

Comparing with the general linear model with correlated observations ( $\rightarrow$  III/2.2.1)

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) , \quad (5)$$

we finally note the equivalences given in equation (1). ■

#### Sources:

- Wikipedia (2022): “General linear model”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-21; URL: [https://en.wikipedia.org/wiki/General\\_linear\\_model#Comparison\\_to\\_multiple\\_linear\\_regression](https://en.wikipedia.org/wiki/General_linear_model#Comparison_to_multiple_linear_regression).

### 1.5.3 Ordinary least squares

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) , \quad (1)$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y . \quad (2)$$



**Proof:** Let  $\hat{\beta}$  be the ordinary least squares (OLS) solution and let  $\hat{\varepsilon} = y - X\hat{\beta}$  be the resulting vector of residuals. Then, this vector must be orthogonal to the design matrix,

$$X^T \hat{\varepsilon} = 0, \quad (3)$$

because if it wasn't, there would be another solution  $\tilde{\beta}$  giving another vector  $\tilde{\varepsilon}$  with a smaller residual sum of squares. From (3), the OLS formula can be directly derived:

$$\begin{aligned} X^T \hat{\varepsilon} &= 0 \\ X^T (y - X\hat{\beta}) &= 0 \\ X^T y - X^T X \hat{\beta} &= 0 \\ X^T X \hat{\beta} &= X^T y \\ \hat{\beta} &= (X^T X)^{-1} X^T y. \end{aligned} \quad (4)$$

■

#### Sources:

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)” in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 10/11; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

#### 1.5.4 Ordinary least squares

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (1)$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad (2)$$

**Proof:** The residual sum of squares ( $\rightarrow$  III/1.5.9) is defined as

$$\text{RSS}(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) \quad (3)$$

which can be developed into

$$\begin{aligned} \text{RSS}(\beta) &= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X \beta \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta. \end{aligned} \quad (4)$$

The derivative of  $\text{RSS}(\beta)$  with respect to  $\beta$  is

$$\frac{d\text{RSS}(\beta)}{d\beta} = -2X^T y + 2X^T X \beta \quad (5)$$

and setting this derivative to zero, we obtain:

$$\begin{aligned}
\frac{d\text{RSS}(\hat{\beta})}{d\beta} &= 0 \\
0 &= -2X^T y + 2X^T X \hat{\beta} \\
X^T X \hat{\beta} &= X^T y \\
\hat{\beta} &= (X^T X)^{-1} X^T y .
\end{aligned} \tag{6}$$

Since the quadratic form  $y^T y$  in (4) is positive,  $\hat{\beta}$  minimizes  $\text{RSS}(\beta)$ . ■

#### Sources:

- Wikipedia (2020): “Proofs involving ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-03; URL: [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares#Least\\_squares\\_estimator\\_for\\_%CE%B2](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Least_squares_estimator_for_%CE%B2).
- ad (2015): “Derivation of the Least Squares Estimator for Beta in Matrix Notation”; in: *Economic Theory Blog*, retrieved on 2021-05-27; URL: [https://economictheoryblog.com/2015/02/19/ols\\_estimator/](https://economictheoryblog.com/2015/02/19/ols_estimator/).

### 1.5.5 Ordinary least squares

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

the parameters minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y . \tag{2}$$

**Proof:** We consider the sum of squared differences between  $y$  and  $X\beta$ :

$$\sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) . \tag{3}$$

First, we note that the residual vector  $\hat{\varepsilon}$  implied by the ordinary least squares solution  $\hat{\beta}$  is orthogonal to the columns of the design matrix, such that the result of their multiplication is the  $p$ -dimensional zero vector (where  $X \in \mathbb{R}^{n \times p}$ ):

$$\begin{aligned}
X^T (y - X\hat{\beta}) &= X^T y - X^T X \hat{\beta} \\
&= X^T y - X^T X (X^T X)^{-1} X^T y \\
&= X^T y - X^T y \\
&= 0_p .
\end{aligned} \tag{4}$$

Second, since  $X^T X$  is a positive semi-definite matrix ( $\rightarrow$  I/1.16.5), the following product is non-negative for each  $p$ -dimensional real vector  $z$ :

$$z^T X^T X z \geq 0 \quad \text{for each } z \in \mathbb{R}^p . \tag{5}$$

We continue developping the sum of squared differences from (3):

$$\begin{aligned}
 (y - X\beta)^T(y - X\beta) &= (y - X\hat{\beta} + X\hat{\beta} - X\beta)^T(y - X\hat{\beta} + X\hat{\beta} - X\beta) \\
 &= \left( (y - X\hat{\beta}) + X(\hat{\beta} - \beta) \right)^T \left( (y - X\hat{\beta}) + X(\hat{\beta} - \beta) \right) \\
 &= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (y - X\hat{\beta})^T X(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^T X^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \\
 &\stackrel{(4)}{=} (y - X\hat{\beta})^T(y - X\hat{\beta}) + 0_p^T(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^T 0_p + (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \\
 &= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) .
 \end{aligned} \tag{6}$$

By virtue of (5), the second term on the right-hand side must be non-zero:

$$(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \geq 0 . \tag{7}$$

Thus, the residual sum of squares must be greater than or equal to the first term

$$(y - X\beta)^T(y - X\beta) \geq (y - X\hat{\beta})^T(y - X\hat{\beta}) \tag{8}$$

and its minimum value is reached when the the second term is zero:

$$\begin{aligned}
 (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) &= 0 \\
 \Leftrightarrow (\hat{\beta} - \beta) &= 0 \\
 \Leftrightarrow \beta &= \hat{\beta} .
 \end{aligned} \tag{9}$$

Thus, the residual sum of squares is minimized when  $\beta = \hat{\beta}$ :

$$\hat{\beta} = (X^T X)^{-1} X^T y = \arg \min_{\beta} (y - X\beta)^T(y - X\beta) . \tag{10}$$

■

#### Sources:

- Ostwald, Dirk (2023): “Parameterschätzung”; in: *Allgemeines Lineares Modell*, Einheit (6), Folien 10-12; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/6\\_Parametersch%C3%A4tzung.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/6_Parametersch%C3%A4tzung.pdf).

### 1.5.6 Ordinary least squares for two regressors

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) in which the design matrix ( $\rightarrow$  III/1.5.1) has two columns:

$$y = X\beta + \varepsilon \quad \text{where} \quad y \in \mathbb{R}^{n \times 1} \quad \text{and} \quad X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \in \mathbb{R}^{n \times 2} . \tag{1}$$

Then,

1) the ordinary least squares ( $\rightarrow$  III/1.5.3) estimates for  $\beta_1$  and  $\beta_2$  are given by

$$\hat{\beta}_1 = \frac{x_2^T x_2 x_1^T y - x_1^T x_2 x_2^T y}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} \quad \text{and} \quad \hat{\beta}_2 = \frac{x_1^T x_1 x_2^T y - x_2^T x_1 x_1^T y}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} \tag{2}$$

2) and, if the two regressors are orthogonal to each other, they simplify to

$$\hat{\beta}_1 = \frac{x_1^T y}{x_1^T x_1} \quad \text{and} \quad \hat{\beta}_2 = \frac{x_2^T y}{x_2^T x_2} . \quad (3)$$

**Proof:** The model in (1) is a special case of multiple linear regression ( $\rightarrow$  III/1.5.1) and the ordinary least squares solution for multiple linear regression ( $\rightarrow$  III/1.5.3) is:

$$\hat{\beta} = (X^T X)^{-1} X^T y . \quad (4)$$

1) Plugging  $X = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$  into this equation, we obtain:

$$\begin{aligned} \hat{\beta} &= \left( \begin{bmatrix} x_1^T \\ x_2^T \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} x_1^T \\ x_2^T \end{bmatrix} y \\ &= \begin{pmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1^T y \\ x_2^T y \end{pmatrix} . \end{aligned} \quad (5)$$

Using the inverse of of a  $2 \times 2$  matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} , \quad (6)$$

this can be further developped into

$$\begin{aligned} \hat{\beta} &= \frac{1}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} \begin{pmatrix} x_2^T x_2 & -x_1^T x_2 \\ -x_2^T x_1 & x_1^T x_1 \end{pmatrix} \begin{pmatrix} x_1^T y \\ x_2^T y \end{pmatrix} \\ &= \frac{1}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} \begin{pmatrix} x_2^T x_2 x_1^T y - x_1^T x_2 x_2^T y \\ x_1^T x_1 x_2^T y - x_2^T x_1 x_1^T y \end{pmatrix} \end{aligned} \quad (7)$$

which can also be written as

$$\begin{aligned} \hat{\beta}_1 &= \frac{x_2^T x_2 x_1^T y - x_1^T x_2 x_2^T y}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} \\ \hat{\beta}_2 &= \frac{x_1^T x_1 x_2^T y - x_2^T x_1 x_1^T y}{x_1^T x_1 x_2^T x_2 - x_1^T x_2 x_2^T x_1} . \end{aligned} \quad (8)$$

2) If two regressors are orthogonal to each other, this means that the inner product of the corresponding vectors is zero:

$$x_1 \perp x_2 \quad \Leftrightarrow \quad x_1^T x_2 = x_2^T x_1 = 0 . \quad (9)$$

Applying this to equation (8), we obtain:

$$\begin{aligned}\hat{\beta}_1 &= \frac{x_2^T x_2 x_1^T y}{x_1^T x_1 x_2^T x_2} = \frac{x_1^T y}{x_1^T x_1} \\ \hat{\beta}_2 &= \frac{x_1^T x_1 x_2^T y}{x_1^T x_1 x_2^T x_2} = \frac{x_2^T y}{x_2^T x_2}.\end{aligned}\tag{10}$$

■

### 1.5.7 Total sum of squares

**Definition:** Let there be a multiple linear regression with independent observations ( $\rightarrow$  III/1.5.1) using measured data  $y$  and design matrix  $X$ :

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Then, the total sum of squares (TSS) is defined as the sum of squared deviations of the measured signal from the average signal:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \tag{2}$$

**Sources:**

- Wikipedia (2020): “Total sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: [https://en.wikipedia.org/wiki/Total\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Total_sum_of_squares).

### 1.5.8 Explained sum of squares

**Definition:** Let there be a multiple linear regression with independent observations ( $\rightarrow$  III/1.5.1) using measured data  $y$  and design matrix  $X$ :

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Then, the explained sum of squares (ESS) is defined as the sum of squared deviations of the fitted signal from the average signal:

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \tag{2}$$

with estimated regression coefficients  $\hat{\beta}$ , e.g. obtained via ordinary least squares ( $\rightarrow$  III/1.5.3).

**Sources:**

- Wikipedia (2020): “Explained sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: [https://en.wikipedia.org/wiki/Explained\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Explained_sum_of_squares).

### 1.5.9 Residual sum of squares

**Definition:** Let there be a multiple linear regression with independent observations ( $\rightarrow$  III/1.5.1) using measured data  $y$  and design matrix  $X$ :

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \quad (1)$$

Then, the residual sum of squares (RSS) is defined as the sum of squared deviations of the measured signal from the fitted signal:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad (2)$$

with estimated regression coefficients  $\hat{\beta}$ , e.g. obtained via ordinary least squares ( $\rightarrow$  III/1.5.3).

**Sources:**

- Wikipedia (2020): “Residual sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: [https://en.wikipedia.org/wiki/Residual\\_sum\\_of\\_squares](https://en.wikipedia.org/wiki/Residual_sum_of_squares).

### 1.5.10 Total, explained and residual sum of squares

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

and let  $X$  contain a constant regressor  $1_n$  modelling the intercept term. Then, it holds that

$$\text{TSS} = \text{ESS} + \text{RSS} \quad (2)$$

where TSS is the total sum of squares ( $\rightarrow$  III/1.5.7), ESS is the explained sum of squares ( $\rightarrow$  III/1.5.8) and RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9).

**Proof:** The total sum of squares ( $\rightarrow$  III/1.5.7) is given by

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

where  $\bar{y}$  is the mean across all  $y_i$ . The TSS can be rewritten as

$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^n (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^n ((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i))^2 \\
&= \sum_{i=1}^n ((\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i)^2 \\
&= \sum_{i=1}^n ((\hat{y}_i - \bar{y})^2 + 2\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i^2) \\
&= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i(x_i\hat{\beta} - \bar{y}) \\
&= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i \left( \sum_{j=1}^p x_{ij}\hat{\beta}_j \right) - 2 \sum_{i=1}^n \hat{\varepsilon}_i \bar{y} \\
&= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{j=1}^p \hat{\beta}_j \sum_{i=1}^n \hat{\varepsilon}_i x_{ij} - 2\bar{y} \sum_{i=1}^n \hat{\varepsilon}_i
\end{aligned} \tag{4}$$

The fact that the design matrix includes a constant regressor ensures that

$$\sum_{i=1}^n \hat{\varepsilon}_i = \hat{\varepsilon}^T \mathbf{1}_n = 0 \tag{5}$$

and because the residuals are orthogonal to the design matrix ( $\rightarrow$  III/1.5.3), we have

$$\sum_{i=1}^n \hat{\varepsilon}_i x_{ij} = \hat{\varepsilon}^T x_j = 0. \tag{6}$$

Applying (5) and (6) to (4), this becomes

$$\text{TSS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \tag{7}$$

and, with the definitions of explained ( $\rightarrow$  III/1.5.8) and residual sum of squares ( $\rightarrow$  III/1.5.9), it is

$$\text{TSS} = \text{ESS} + \text{RSS}. \tag{8}$$

■

#### Sources:

- Wikipedia (2020): “Partition of sums of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-09; URL: [https://en.wikipedia.org/wiki/Partition\\_of\\_sums\\_of\\_squares#Partitioning\\_the\\_sum\\_of\\_squares\\_in\\_linear\\_regression](https://en.wikipedia.org/wiki/Partition_of_sums_of_squares#Partitioning_the_sum_of_squares_in_linear_regression).

### 1.5.11 Estimation matrix

**Definition:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the estimation matrix is the matrix  $E$  that results in ordinary least squares ( $\rightarrow$  III/1.5.3) or weighted least squares ( $\rightarrow$  III/1.5.21) parameter estimates when right-multiplied with the measured data:

$$Ey = \hat{\beta} . \quad (1)$$

### 1.5.12 Projection matrix

**Definition:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the projection matrix is the matrix  $P$  that results in the fitted signal explained by estimated parameters ( $\rightarrow$  III/1.5.11) when right-multiplied with the measured data:

$$Py = \hat{y} = X\hat{\beta} . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: [https://en.wikipedia.org/wiki/Projection\\_matrix#Overview](https://en.wikipedia.org/wiki/Projection_matrix#Overview).

### 1.5.13 Residual-forming matrix

**Definition:** In multiple linear regression ( $\rightarrow$  III/1.5.1), the residual-forming matrix is the matrix  $R$  that results in the vector of residuals left over by estimated parameters ( $\rightarrow$  III/1.5.11) when right-multiplied with the measured data:

$$Ry = \hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} . \quad (1)$$

**Sources:**

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: [https://en.wikipedia.org/wiki/Projection\\_matrix#Properties](https://en.wikipedia.org/wiki/Projection_matrix#Properties).

### 1.5.14 Estimation, projection and residual-forming matrix

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.5.3). Then, the estimated parameters, fitted signal and residuals are given by

$$\begin{aligned} \hat{\beta} &= Ey \\ \hat{y} &= Py \\ \hat{\varepsilon} &= Ry \end{aligned} \quad (2)$$

where



$$\begin{aligned}
E &= (X^T X)^{-1} X^T \\
P &= X(X^T X)^{-1} X^T \\
R &= I_n - X(X^T X)^{-1} X^T
\end{aligned} \tag{3}$$

are the estimation matrix ( $\rightarrow$  III/1.5.11), projection matrix ( $\rightarrow$  III/1.5.12) and residual-forming matrix ( $\rightarrow$  III/1.5.13) and  $n$  is the number of observations.

**Proof:**

1) Ordinary least squares parameter estimates of  $\beta$  are defined as minimizing the residual sum of squares ( $\rightarrow$  III/1.5.9)

$$\hat{\beta} = \arg \min_{\beta} [(y - X\beta)^T (y - X\beta)] \tag{4}$$

and the solution to this ( $\rightarrow$  III/1.5.3) is given by

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T y \\
&\stackrel{(3)}{=} Ey .
\end{aligned} \tag{5}$$

2) The fitted signal is given by multiplying the design matrix with the estimated regression coefficients

$$\hat{y} = X\hat{\beta} \tag{6}$$

and using (5), this becomes

$$\begin{aligned}
\hat{y} &= X(X^T X)^{-1} X^T y \\
&\stackrel{(3)}{=} Py .
\end{aligned} \tag{7}$$

3) The residuals of the model are calculated by subtracting the fitted signal from the measured signal

$$\hat{\varepsilon} = y - \hat{y} \tag{8}$$

and using (7), this becomes

$$\begin{aligned}
\hat{\varepsilon} &= y - X(X^T X)^{-1} X^T y \\
&= (I_n - X(X^T X)^{-1} X^T) y \\
&\stackrel{(3)}{=} Ry .
\end{aligned} \tag{9}$$

■

**Sources:**

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slide 10; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

### 1.5.15 Symmetry of projection and residual-forming matrix

**Theorem:** The projection matrix ( $\rightarrow$  III/1.5.12) and the residual-forming matrix ( $\rightarrow$  III/1.5.13) are symmetric:

$$\begin{aligned} P^T &= P \\ R^T &= R . \end{aligned} \tag{1}$$

**Proof:** Let  $X$  be the design matrix from the linear regression model ( $\rightarrow$  III/1.5.1). Then, the matrix  $X^T X$  is symmetric, because

$$(X^T X)^T = X^T X^{TT} = X^T X . \tag{2}$$

Thus, the inverse of  $X^T X$  is also symmetric, i.e.

$$((X^T X)^{-1})^T = (X^T X)^{-1} . \tag{3}$$

1) The projection matrix for ordinary least squares is given by ( $\rightarrow$  III/1.5.14)

$$P = X(X^T X)^{-1} X^T , \tag{4}$$

such that

$$\begin{aligned} P^T &= (X(X^T X)^{-1} X^T)^T \\ &= X^{TT} ((X^T X)^{-1})^T X^T \\ &= X(X^T X)^{-1} X^T \\ &\stackrel{(4)}{=} P . \end{aligned} \tag{5}$$

2) The residual-forming matrix for ordinary least squares is given by ( $\rightarrow$  III/1.5.14)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P , \tag{6}$$

such that

$$\begin{aligned} R^T &= (I_n - P)^T \\ &= I_n^T - P^T \\ &\stackrel{(5)}{=} I_n - P \\ &\stackrel{(6)}{=} R . \end{aligned} \tag{7}$$

■

#### Sources:

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-12-22; URL: [https://en.wikipedia.org/wiki/Projection\\_matrix#Properties](https://en.wikipedia.org/wiki/Projection_matrix#Properties).

### 1.5.16 Idempotence of projection and residual-forming matrix

**Theorem:** The projection matrix ( $\rightarrow$  III/1.5.12) and the residual-forming matrix ( $\rightarrow$  III/1.5.13) are idempotent:

$$\begin{aligned} P^2 &= P \\ R^2 &= R. \end{aligned} \tag{1}$$

**Proof:**

1) The projection matrix for ordinary least squares is given by ( $\rightarrow$  III/1.5.14)

$$P = X(X^T X)^{-1} X^T, \tag{2}$$

such that

$$\begin{aligned} P^2 &= X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &\stackrel{(2)}{=} P. \end{aligned} \tag{3}$$

2) The residual-forming matrix for ordinary least squares is given by ( $\rightarrow$  III/1.5.14)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P, \tag{4}$$

such that

$$\begin{aligned} R^2 &= (I_n - P)(I_n - P) \\ &= I_n - P - P + P^2 \\ &\stackrel{(3)}{=} I_n - 2P + P \\ &= I_n - P \\ &\stackrel{(4)}{=} R. \end{aligned} \tag{5}$$

■

**Sources:**

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: [https://en.wikipedia.org/wiki/Projection\\_matrix#Properties](https://en.wikipedia.org/wiki/Projection_matrix#Properties).

### 1.5.17 Independence of estimated parameters and residuals

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and consider estimation using weighted least squares ( $\rightarrow$  III/1.5.21). Then, the estimated parameters and the vector of residuals ( $\rightarrow$  III/1.5.19) are independent from each other:

$$\begin{aligned}\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \quad \text{and} \\ \hat{\varepsilon} &= y - X \hat{\beta} \quad \text{ind.}\end{aligned}\tag{2}$$

**Proof:** Equation (1) implies the following distribution ( $\rightarrow$  III/1.5.19) for the random vector ( $\rightarrow$  I/1.2.3)  $y$ :

$$\begin{aligned}y &\sim \mathcal{N}(X\beta, \sigma^2 V) \\ &\sim \mathcal{N}(X\beta, \Sigma) \\ \text{with } \Sigma &= \sigma^2 V.\end{aligned}\tag{3}$$

Note that the estimated parameters and residuals can be written as projections from the same random vector ( $\rightarrow$  III/1.5.14)  $y$ :

$$\begin{aligned}\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ &= Ay \\ \text{with } A &= (X^T V^{-1} X)^{-1} X^T V^{-1}\end{aligned}\tag{4}$$

$$\begin{aligned}\hat{\varepsilon} &= y - X \hat{\beta} \\ &= (I_n - X(X^T V^{-1} X)^{-1} X^T V^{-1}) y \\ &= By \\ \text{with } B &= (I_n - X(X^T V^{-1} X)^{-1} X^T V^{-1}).\end{aligned}\tag{5}$$

Two projections  $AZ$  and  $BZ$  from the same multivariate normal ( $\rightarrow$  II/4.1.1) random vector ( $\rightarrow$  I/1.2.3)  $Z \sim \mathcal{N}(\mu, \Sigma)$  are independent, if and only if the following condition holds ( $\rightarrow$  II/4.1.18):

$$A \Sigma B^T = 0.\tag{6}$$

Combining (3), (4) and (5), we check whether this is fulfilled in the present case:

$$\begin{aligned}A \Sigma B^T &= (X^T V^{-1} X)^{-1} X^T V^{-1} (\sigma^2 V) (I_n - X(X^T V^{-1} X)^{-1} X^T V^{-1})^T \\ &= \sigma^2 [(X^T V^{-1} X)^{-1} X^T V^{-1} V - (X^T V^{-1} X)^{-1} X^T V^{-1} V V^{-1} X (X^T V^{-1} X)^{-1} X^T] \\ &= \sigma^2 [(X^T V^{-1} X)^{-1} X^T - (X^T V^{-1} X)^{-1} X^T] \\ &= \sigma^2 \cdot 0_{pn} \\ &= 0.\end{aligned}\tag{7}$$

This demonstrates that  $\hat{\beta}$  and  $\hat{\varepsilon}$  – and likewise, all pairs of terms separately derived ( $\rightarrow$  III/1.5.28) from  $\hat{\beta}$  and  $\hat{\varepsilon}$  – are statistically independent ( $\rightarrow$  I/1.3.6). ■

#### Sources:

- jld (2018): “Understanding t-test for linear regression”; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: <https://stats.stackexchange.com/a/344008>.

### 1.5.18 Distribution of OLS estimates, signal and residuals

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

and consider estimation using ordinary least squares ( $\rightarrow$  III/1.5.3). Then, the estimated parameters, fitted signal and residuals are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1}) \\ \hat{y} &\sim \mathcal{N}(X\beta, \sigma^2 P) \\ \hat{\varepsilon} &\sim \mathcal{N}(0, \sigma^2(I_n - P)) \end{aligned} \quad (2)$$

where  $P$  is the projection matrix ( $\rightarrow$  III/1.5.12) for ordinary least squares ( $\rightarrow$  III/1.5.3)

$$P = X(X^T X)^{-1} X^T. \quad (3)$$

**Proof:** We will use the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14):

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T). \quad (4)$$

The distributional assumption in (1) is equivalent to ( $\rightarrow$  II/4.1.17):

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \quad (5)$$

Applying (4) to (5), the measured data are distributed as

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n). \quad (6)$$

1) The parameter estimates from ordinary least squares ( $\rightarrow$  III/1.5.3) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (7)$$

and thus, by applying (4) to (7), they are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}([(X^T X)^{-1} X^T] X\beta, \sigma^2 [(X^T X)^{-1} X^T] I_n [X(X^T X)^{-1}]) \\ &\sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}). \end{aligned} \quad (8)$$

2) The fitted signal in multiple linear regression ( $\rightarrow$  III/1.5.14) is given by

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Py \quad (9)$$

and thus, by applying (4) to (9), they are distributed as

$$\begin{aligned} \hat{y} &\sim \mathcal{N}(X\beta, \sigma^2 X(X^T X)^{-1} X^T) \\ &\sim \mathcal{N}(X\beta, \sigma^2 P). \end{aligned} \quad (10)$$

3) The residuals of the linear regression model ( $\rightarrow$  III/1.5.14) are given by

$$\hat{\varepsilon} = y - X\hat{\beta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P) y \quad (11)$$

and thus, by applying (4) to (11), they are distributed as

$$\begin{aligned} \hat{\varepsilon} &\sim \mathcal{N} \left( [I_n - X(X^T X)^{-1} X^T] X\beta, \sigma^2 [I_n - P] I_n [I_n - P]^T \right) \\ &\sim \mathcal{N} \left( X\beta - X\beta, \sigma^2 [I_n - P] [I_n - P]^T \right) . \end{aligned} \quad (12)$$

Because the residual-forming matrix ( $\rightarrow$  III/1.5.13) is symmetric ( $\rightarrow$  III/1.5.15) and idempotent ( $\rightarrow$  III/1.5.16), this becomes:

$$\hat{\varepsilon} \sim \mathcal{N} (0, \sigma^2 (I_n - P)) . \quad (13)$$

■

#### Sources:

- Koch, Karl-Rudolf (2007): “Linear Model”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4, eqs. 4.2, 4.30; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): “Multiple Regression”; in: *Mathematics for Brain Imaging*, ch. 1.5, pp. 39-41, eqs. 1.106-1.110; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Ostwald, Dirk (2023): “Modellformulierung”; in: *Allgemeines Lineares Modell*, Einheit (5), Folie 14; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/5\\_Modellformulierung.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/5_Modellformulierung.pdf).
- Ostwald, Dirk (2023): “Parameterschätzung”; in: *Allgemeines Lineares Modell*, Einheit (6), Folien 10-12; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/6\\_Parametersch%C3%A4tzung.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/6_Parametersch%C3%A4tzung.pdf).

#### 1.5.19 Distribution of WLS estimates, signal and residuals

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

and consider estimation using weighted least squares ( $\rightarrow$  III/1.5.21). Then, the estimated parameters, fitted signal and residuals are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N} (\beta, \sigma^2 (X^T V^{-1} X)^{-1}) \\ \hat{y} &\sim \mathcal{N} (X\beta, \sigma^2 (P V)) \\ \hat{\varepsilon} &\sim \mathcal{N} (0, \sigma^2 (I_n - P) V) \end{aligned} \quad (2)$$

where  $P$  is the projection matrix ( $\rightarrow$  III/1.5.12) for weighted least squares ( $\rightarrow$  III/1.5.21)

$$P = X(X^T V^{-1} X)^{-1} X^T V^{-1} . \quad (3)$$

**Proof:** We will use the linear transformation theorem for the multivariate normal distribution ( $\rightarrow$  II/4.1.14):

$$x \sim \mathcal{N}(\mu, \Sigma) \Rightarrow y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T). \quad (4)$$

Applying (4) to (1), the measured data are distributed as

$$y \sim \mathcal{N}(X\beta, \sigma^2 V). \quad (5)$$

1) The parameter estimates from weighted least squares ( $\rightarrow$  III/1.5.21) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \quad (6)$$

and thus, by applying (4) to (6), they are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}([ (X^T V^{-1} X)^{-1} X^T V^{-1} ] X\beta, \sigma^2 [ (X^T V^{-1} X)^{-1} X^T V^{-1} ] V [ V^{-1} X (X^T V^{-1} X)^{-1} ]) \\ &\sim \mathcal{N}(\beta, \sigma^2 (X^T V^{-1} X)^{-1}). \end{aligned} \quad (7)$$

2) The fitted signal in multiple linear regression ( $\rightarrow$  III/1.5.14) is given by

$$\hat{y} = X\hat{\beta} = X(X^T V^{-1} X)^{-1} X^T V^{-1} y = Py \quad (8)$$

and thus, by applying (4) to (8), they are distributed as

$$\begin{aligned} \hat{y} &\sim \mathcal{N}(X\beta, \sigma^2 X(X^T V^{-1} X)^{-1} X^T) \\ &\sim \mathcal{N}(X\beta, \sigma^2 (PV)). \end{aligned} \quad (9)$$

3) The residuals of the linear regression model ( $\rightarrow$  III/1.5.14) are given by

$$\hat{\varepsilon} = y - X\hat{\beta} = (I_n - X(X^T V^{-1} X)^{-1} X^T V^{-1}) y = (I_n - P) y \quad (10)$$

and thus, by applying (4) to (10), they are distributed as

$$\begin{aligned} \hat{\varepsilon} &\sim \mathcal{N}([I_n - X(X^T V^{-1} X)^{-1} X^T V^{-1}] X\beta, \sigma^2 [I_n - P] V [I_n - P]^T) \\ &\sim \mathcal{N}(X\beta - X\beta, \sigma^2 [V - VP^T - PV + PVP^T]) \\ &\sim \mathcal{N}(0, \sigma^2 [V - VV^{-1} X(X^T V^{-1} X)^{-1} X^T - X(X^T V^{-1} X)^{-1} X^T V^{-1} V + PVP^T]) \\ &\sim \mathcal{N}(0, \sigma^2 [V - 2PV + X(X^T V^{-1} X)^{-1} X^T V^{-1} VV^{-1} X(X^T V^{-1} X)^{-1} X^T]) \\ &\sim \mathcal{N}(0, \sigma^2 [V - 2PV + PV]) \\ &\sim \mathcal{N}(0, \sigma^2 [V - PV]) \\ &\sim \mathcal{N}(0, \sigma^2 [I_n - P] V). \end{aligned} \quad (11)$$

■

#### Sources:

- Koch, Karl-Rudolf (2007): “Linear Model”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4, eqs. 4.2, 4.30; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): “Multiple Regression”; in: *Mathematics for Brain Imaging*, ch. 1.5, pp. 39-41, eqs. 1.106-1.110; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, eq. A.10; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.
- Soch J, Meyer AP, Allefeld C, Haynes JD (2017): “How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging”; in: *NeuroImage*, vol. 158, pp. 186-195, eq. A.2; URL: <https://www.sciencedirect.com/science/article/pii/S105381191730527X>; DOI: 10.1016/j.neuroimage.2017.06.056.

### 1.5.20 Distribution of residual sum of squares

**Theorem:** Assume a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

and consider estimation using weighted least squares ( $\rightarrow$  III/1.5.21). Then, the residual sum of squares ( $\rightarrow$  III/1.5.9)  $\hat{\varepsilon}^T \hat{\varepsilon}$ , divided by the true error variance ( $\rightarrow$  III/1.5.1)  $\sigma^2$ , follows a chi-squared distribution ( $\rightarrow$  II/3.7.1) with  $n - p$  degrees of freedom ( $\rightarrow$  I/1.5.12)

$$\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} \sim \chi^2(n - p) \quad (2)$$

where  $n$  and  $p$  are the dimensions of the  $n \times p$  design matrix ( $\rightarrow$  III/1.5.1)  $X$ .

**Proof:** Consider an  $n \times n$  square matrix  $W$ , such that

$$WVW^T = I_n. \quad (3)$$

Then, left-multiplying the regression model in (1) with  $W$  gives

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T) \quad (4)$$

which can be rewritten as

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (5)$$

where  $\tilde{y} = Wy$ ,  $\tilde{X} = WX$  and  $\tilde{\varepsilon} = W\varepsilon$ . This implies the distribution ( $\rightarrow$  II/4.1.14)

$$\tilde{y} \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I_n). \quad (6)$$

With that, we have obtained a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations. Cochran’s theorem for multivariate normal variables states that, for an  $n$ -dimensional normal random vector ( $\rightarrow$  II/4.1.1) whose covariance matrix ( $\rightarrow$  I/1.16.1) is a scalar multiple of the identity matrix, a specific squared form will follow a non-central chi-squared distribution where the degrees of freedom and the non-centrality parameter depend on the matrix in the quadratic form:

$$x \sim \mathcal{N}(\mu, \sigma^2 I_n) \quad \Rightarrow \quad y = x^T Ax / \sigma^2 \sim \chi^2(\text{rk}(A), \mu^T A \mu). \quad (7)$$

First, we formulate the residuals ( $\rightarrow$  III/1.5.14) in terms of transformed measurements  $\tilde{y}$ :



$$\begin{aligned}
\hat{\varepsilon} &= \tilde{y} - \tilde{X}\hat{\beta} & \text{where} & \quad \hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\
&= (I_n - \tilde{P})\tilde{y} & \text{where} & \quad \tilde{P} = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \\
&= \tilde{R}\tilde{y} & \text{where} & \quad \tilde{R} = I_n - \tilde{P} .
\end{aligned} \tag{8}$$

Next, we observe that the residual sum of squares can be represented as a quadratic form:

$$\frac{1}{\sigma^2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} = \tilde{y}^T \tilde{R}^T \tilde{R} \tilde{y} / \sigma^2 \tag{9}$$

Because the residual-forming matrix ( $\rightarrow$  III/1.5.13)  $\tilde{R}$  is symmetric ( $\rightarrow$  III/1.5.15) and idempotent ( $\rightarrow$  III/1.5.16), we have  $\tilde{R}^T = \tilde{R}$  and  $\tilde{R}^2 = \tilde{R}$ , such that:

$$\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} = \tilde{y}^T \tilde{R} \tilde{y} / \sigma^2 . \tag{10}$$

Since  $\tilde{R}$  is idempotent, its rank is equal to its trace, such that:

$$\text{rk}(\tilde{R}) = \text{tr}(\tilde{R}) . \tag{11}$$

With that, we can apply Cochran's theorem given by (7) which yields

$$\begin{aligned}
\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} &\sim \chi^2 \left( \text{rk}(\tilde{R}), \beta^T \tilde{X}^T \tilde{R} \tilde{X} \beta \right) \\
&\sim \chi^2 \left( \text{tr}(I_n - \tilde{P}), \beta^T \tilde{X}^T (I_n - \tilde{P}) \tilde{X} \beta \right) \\
&\sim \chi^2 \left( \text{tr}(I_n) - \text{tr}(\tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T), \beta^T (\tilde{X}^T \tilde{X} - \tilde{X}^T \tilde{X} (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T) \beta \right) \\
&\sim \chi^2 \left( \text{tr}(I_n) - \text{tr}(\tilde{X}^T \tilde{X} (\tilde{X}^T \tilde{X})^{-1}), \beta^T (\tilde{X}^T \tilde{X} - \tilde{X}^T \tilde{X}) \beta \right) \\
&\sim \chi^2 \left( \text{tr}(I_n) - \text{tr}(I_p), \beta^T 0_{pp} \beta \right) \\
&\sim \chi^2 (n - p, 0) .
\end{aligned} \tag{12}$$

Because a non-central chi-squared distribution with non-centrality parameter of zero reduces to the central chi-squared distribution, we obtain our final result:

$$\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} \sim \chi^2(n - p) . \tag{13}$$

■

#### Sources:

- Koch, Karl-Rudolf (2007): “Estimation of the Variance Factor in Traditional Statistics”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4.2.3, eq. 4.37; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): “Estimating error variance”; in: *Mathematics for Brain Imaging*, ch. 2.2, pp. 49-51, eqs. 2.4-2.8; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).
- Wikipedia (2022): “Ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-12-13; URL: [https://en.wikipedia.org/wiki/Ordinary\\_least\\_squares#Estimation](https://en.wikipedia.org/wiki/Ordinary_least_squares#Estimation).
- ocran (2022): “Why is RSS distributed chi square times n-p?”; in: *StackExchange Cross Validated*, retrieved on 2022-12-21; URL: <https://stats.stackexchange.com/a/20230>.

### 1.5.21 Weighted least squares

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad (1)$$

the parameters minimizing the weighted residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \quad (2)$$

**Proof:** Let there be an  $n \times n$  square matrix  $W$ , such that

$$WVW^T = I_n. \quad (3)$$

Since  $V$  is a covariance matrix and thus symmetric ( $\rightarrow$  I/1.15.4),  $W$  is also symmetric and can be expressed as the matrix square root of the inverse of  $V$ :

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2}. \quad (4)$$

Left-multiplying the linear regression equation (1) with  $W$ , the linear transformation theorem ( $\rightarrow$  II/4.1.14) implies that

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T). \quad (5)$$

Applying (3), we see that (5) is actually a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (6)$$

where  $\tilde{y} = Wy$ ,  $\tilde{X} = WX$  and  $\tilde{\varepsilon} = W\varepsilon$ , such that we can apply the ordinary least squares solution ( $\rightarrow$  III/1.5.3) giving

$$\begin{aligned} \hat{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\ &= ((WX)^T WX)^{-1} (WX)^T Wy \\ &= (X^T W^T WX)^{-1} X^T W^T Wy \\ &= (X^T WWX)^{-1} X^T WWy \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} y \end{aligned} \quad (7)$$

which corresponds to the weighted least squares solution (2). ■

#### Sources:

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 20/23; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.
- Wikipedia (2021): “Weighted least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-17; URL: [https://en.wikipedia.org/wiki/Weighted\\_least\\_squares#Motivation](https://en.wikipedia.org/wiki/Weighted_least_squares#Motivation).

## 1.5.22 Weighted least squares

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad (1)$$

the parameters minimizing the weighted residual sum of squares ( $\rightarrow$  III/1.5.9) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \quad (2)$$

**Proof:** Let there be an  $n \times n$  square matrix  $W$ , such that

$$WVW^T = I_n. \quad (3)$$

Since  $V$  is a covariance matrix and thus symmetric ( $\rightarrow$  I/1.15.4),  $W$  is also symmetric and can be expressed the matrix square root of the inverse of  $V$ :

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2}. \quad (4)$$

Left-multiplying the linear regression equation (1) with  $W$ , the linear transformation theorem ( $\rightarrow$  II/4.1.14) implies that

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T). \quad (5)$$

Applying (3), we see that (5) is actually a linear regression model ( $\rightarrow$  III/1.5.1) with independent observations

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \quad (6)$$

With this, we can express the weighted residual sum of squares ( $\rightarrow$  III/1.5.9) as

$$\text{wRSS}(\beta) = \sum_{i=1}^n (W\varepsilon)_i^2 = (W\varepsilon)^T (W\varepsilon) = (Wy - WX\beta)^T (Wy - WX\beta) \quad (7)$$

which can be developed into

$$\begin{aligned} \text{wRSS}(\beta) &= y^T W^T W y - y^T W^T W X \beta - \beta^T X^T W^T W y + \beta^T X^T W^T W X \beta \\ &= y^T W W y - 2\beta^T X^T W W y + \beta^T X^T W W X \beta \\ &\stackrel{(4)}{=} y^T V^{-1} y - 2\beta^T X^T V^{-1} y + \beta^T X^T V^{-1} X \beta. \end{aligned} \quad (8)$$

The derivative of  $\text{wRSS}(\beta)$  with respect to  $\beta$  is

$$\frac{d\text{wRSS}(\beta)}{d\beta} = -2X^T V^{-1} y + 2X^T V^{-1} X \beta \quad (9)$$

and setting this derivative to zero, we obtain:

$$\begin{aligned} \frac{d\text{wRSS}(\hat{\beta})}{d\beta} &= 0 \\ 0 &= -2X^T V^{-1} y + 2X^T V^{-1} X \hat{\beta} \\ X^T V^{-1} X \hat{\beta} &= X^T V^{-1} y \\ \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y. \end{aligned} \quad (10)$$

Since the quadratic form  $y^T V^{-1} y$  in (8) is positive,  $\hat{\beta}$  minimizes  $\text{wRSS}(\beta)$ . ■

### 1.5.23 Maximum likelihood estimation

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad (1)$$

the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\beta$  and  $\sigma^2$  are given by

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}). \end{aligned} \quad (2)$$

**Proof:** With the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), the linear regression equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned} p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right] \end{aligned} \quad (3)$$

and, using  $|\sigma^2 V| = (\sigma^2)^n |V|$ , the log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| \\ &\quad - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta). \end{aligned} \quad (4)$$

Substituting the precision matrix  $P = V^{-1}$  into (4) to ease notation, we have:

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|V|) \\ &\quad - \frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta). \end{aligned} \quad (5)$$

The derivative of the log-likelihood function (5) with respect to  $\beta$  is

$$\begin{aligned} \frac{d\text{LL}(\beta, \sigma^2)}{d\beta} &= \frac{d}{d\beta} \left( -\frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) \right) \\ &= \frac{1}{2\sigma^2} \frac{d}{d\beta} (2\beta^T X^T P y - \beta^T X^T P X \beta) \\ &= \frac{1}{2\sigma^2} (2X^T P y - 2X^T P X \beta) \\ &= \frac{1}{\sigma^2} (X^T P y - X^T P X \beta) \end{aligned} \quad (6)$$

and setting this derivative to zero gives the MLE for  $\beta$ :

$$\begin{aligned}
 \frac{dLL(\hat{\beta}, \sigma^2)}{d\beta} &= 0 \\
 0 &= \frac{1}{\sigma^2} \left( X^T P y - X^T P X \hat{\beta} \right) \\
 0 &= X^T P y - X^T P X \hat{\beta} \\
 X^T P X \hat{\beta} &= X^T P y \\
 \hat{\beta} &= (X^T P X)^{-1} X^T P y
 \end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at  $\hat{\beta}$  with respect to  $\sigma^2$  is

$$\begin{aligned}
 \frac{dLL(\hat{\beta}, \sigma^2)}{d\sigma^2} &= \frac{d}{d\sigma^2} \left( -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \right) \\
 &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\
 &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})
 \end{aligned} \tag{8}$$

and setting this derivative to zero gives the MLE for  $\sigma^2$ :

$$\begin{aligned}
 \frac{dLL(\hat{\beta}, \hat{\sigma}^2)}{d\sigma^2} &= 0 \\
 0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\
 \frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\
 \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{n}{2\hat{\sigma}^2} &= \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\
 \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})
 \end{aligned} \tag{9}$$

Together, (7) and (9) constitute the MLE for multiple linear regression.

■

#### 1.5.24 Maximum log-likelihood

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$  with correlation structure ( $\rightarrow$  I/1.17.6)  $V$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V). \tag{1}$$

Then, the maximum log-likelihood ( $\rightarrow$  I/4.1.4) for this model is

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] \quad (2)$$

under uncorrelated observations ( $\rightarrow$  III/1.5.1), i.e. if  $V = I_n$ , and

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{wRSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] - \frac{1}{2} \log |V|, \quad (3)$$

in the general case, i.e. if  $V \neq I_n$ , where RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9) and wRSS is the weighted residual sum of squares ( $\rightarrow$  III/1.5.22).

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for multiple linear regression is given by ( $\rightarrow$  III/1.5.23)

$$\begin{aligned} p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right], \end{aligned} \quad (4)$$

such that, with  $|\sigma^2 V| = (\sigma^2)^n |V|$ , the log-likelihood function ( $\rightarrow$  I/4.1.2) for this model becomes ( $\rightarrow$  III/1.5.23)

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta). \end{aligned} \quad (5)$$

The maximum likelihood estimate for the noise variance ( $\rightarrow$  III/1.5.23) is

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \quad (6)$$

which can also be expressed in terms of the (weighted) residual sum of squares ( $\rightarrow$  III/1.5.9) as

$$\frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) = \frac{1}{n} (Wy - WX\hat{\beta})^T (Wy - WX\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n (W\hat{\varepsilon})_i^2 = \frac{\text{wRSS}}{n} \quad (7)$$

where  $W = V^{-1/2}$ . Plugging (6) into (5), we obtain the maximum log-likelihood ( $\rightarrow$  I/4.1.4) as

$$\begin{aligned} \text{MLL}(m) &= \text{LL}(\hat{\beta}, \hat{\sigma}^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2} \log |V| - \frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left( \frac{\text{wRSS}}{n} \right) - \frac{1}{2} \log |V| - \frac{1}{2} \cdot \frac{n}{\text{wRSS}} \cdot \text{wRSS} \\ &= -\frac{n}{2} \log \left( \frac{\text{wRSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] - \frac{1}{2} \log |V| \end{aligned} \quad (8)$$

which proves the result in (3). Assuming  $V = I_n$ , we have

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\text{RSS}}{n} \quad (9)$$

and

$$\frac{1}{2} \log |V| = \frac{1}{2} \log |I_n| = \frac{1}{2} \log 1 = 0, \quad (10)$$

such that

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] \quad (11)$$

which proves the result in (2). This completes the proof. ■

#### Sources:

- Claeskens G, Hjort NL (2008): “Akaike’s information criterion”; in: *Model Selection and Model Averaging*, ex. 2.2, p. 66; URL: <https://www.cambridge.org/core/books/model-selection-and-model-averaging/E6F1EC77279D1223423BB64FC3A12C37>; DOI: 10.1017/CBO9780511790485.

#### 1.5.25 Log-likelihood ratio

**Theorem:** Let  $y = [y_1, \dots, y_n]^T$  be an  $n$ -dimensional data vector ( $\rightarrow$  I/1.1.5) and consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m_1$  with design matrix ( $\rightarrow$  III/1.5.1)  $X = [X_0, X_1] \in \mathbb{R}^{n \times p}$  as well as a reduced linear regression model ( $\rightarrow$  III/1.5.1)  $m_0$  with design matrix ( $\rightarrow$  III/1.5.1)  $X_0 \in \mathbb{R}^{n \times p_0}$ :

$$\begin{aligned} m_1 : y &= X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \\ m_0 : y &= X_0\beta_0 + \varepsilon_0, \quad \varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2 V). \end{aligned} \quad (1)$$

Both models use the same covariance matrix ( $\rightarrow$  III/1.5.1)  $V \in \mathbb{R}^{n \times n}$ , but entail potentially different regression coefficients ( $\rightarrow$  III/1.5.1)  $\beta, \beta_0$  and noise variances ( $\rightarrow$  III/1.5.1)  $\sigma^2, \sigma_0^2$ . Then, the log-likelihood ratio ( $\rightarrow$  I/4.1.7) for comparing  $m_0$  vs.  $m_1$  is given by

$$\ln \Lambda_{01} = \frac{n}{2} \ln \left( \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right) \quad (2)$$

where  $\hat{\sigma}^2$  and  $\hat{\sigma}_0^2$  are the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of the noise variance ( $\rightarrow$  III/1.5.1) based on the full model  $m_1$  and the reduced model  $m_0$ , respectively.

**Proof:** The likelihood ratio ( $\rightarrow$  I/4.1.6) between two models  $m_1$  and  $m_2$  with model parameters  $\theta_1$  and  $\theta_2$  and parameter spaces  $\Theta_1$  and  $\Theta_2$  is defined as the quotient of their maximized ( $\rightarrow$  I/4.1.3) likelihood functions ( $\rightarrow$  I/5.1.2):

$$\Lambda_{12} = \frac{\max_{\theta_1 \in \Theta_1} p(y|\theta_1, m_1)}{\max_{\theta_2 \in \Theta_2} p(y|\theta_2, m_2)}. \quad (3)$$

The likelihood function ( $\rightarrow$  I/5.1.2) of multiple linear regression ( $\rightarrow$  III/1.5.1) is a multivariate normal probability density function ( $\rightarrow$  II/4.1.4):

$$\begin{aligned} p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right] \\ &= \sqrt{\frac{1}{(2\pi\sigma^2)^n |V|}} \cdot \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right] \end{aligned} \quad (4)$$

and the maximum likelihood estimates for multiple linear regression ( $\rightarrow$  III/1.5.23) are given by

$$\begin{aligned}\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta}) .\end{aligned}\tag{5}$$

Thus, the likelihood ratio comparing  $m_0$  vs.  $m_1$  is equal to

$$\begin{aligned}\Lambda_{01} &= \frac{p(y|\hat{\beta}_0, \hat{\sigma}_0^2, m_0)}{p(y|\hat{\beta}, \hat{\sigma}^2, m_1)} \\ &= \frac{\sqrt{\frac{1}{(2\pi\hat{\sigma}_0^2)^n |V|}} \cdot \exp \left[ -\frac{1}{2\hat{\sigma}_0^2} (y - X_0 \hat{\beta}_0)^T V^{-1} (y - X_0 \hat{\beta}_0) \right]}{\sqrt{\frac{1}{(2\pi\hat{\sigma}^2)^n |V|}} \cdot \exp \left[ -\frac{1}{2\hat{\sigma}^2} (y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta}) \right]} \\ &= \frac{(2\pi)^{-n/2} (\hat{\sigma}_0^2)^{-n/2} |V|^{-n/2} \cdot \exp \left[ -\frac{1}{2\hat{\sigma}_0^2} (y - X_0 \hat{\beta}_0)^T V^{-1} (y - X_0 \hat{\beta}_0) \right]}{(2\pi)^{-n/2} (\hat{\sigma}^2)^{-n/2} |V|^{-n/2} \cdot \exp \left[ -\frac{1}{2\hat{\sigma}^2} (y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta}) \right]} \\ &= \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-n/2} \cdot \frac{\exp \left[ -\frac{n}{2} \frac{(y - X_0 \hat{\beta}_0)^T V^{-1} (y - X_0 \hat{\beta}_0)}{(y - X_0 \hat{\beta}_0)^T V^{-1} (y - X_0 \hat{\beta}_0)} \right]}{\exp \left[ -\frac{n}{2} \frac{(y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta})}{(y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta})} \right]} \\ &= \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-n/2} \cdot \frac{\exp \left[ -\frac{n}{2} \right]}{\exp \left[ -\frac{n}{2} \right]} \\ &= \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-n/2} .\end{aligned}\tag{6}$$

Logarithmizing both sides, the log-likelihood ratio is obtained as

$$\ln \Lambda_{01} = \ln \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-n/2} = -\frac{n}{2} \ln \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right) = \frac{n}{2} \ln \left( \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right) .\tag{7}$$

■

#### Sources:

- Ostwald, Dirk (2023): “F-Statistiken”; in: *Allgemeines Lineares Modell*, Einheit (8), Folien 20-22;  
URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/8\\_F\\_Statistiken-p-9972.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/8_F_Statistiken-p-9972.pdf).

#### 1.5.26 t-contrast

**Definition:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with  $n \times p$  design matrix  $X$  and  $p \times 1$  regression coefficients  $\beta$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) .\tag{1}$$

Then, a t-contrast is specified by a  $p \times 1$  vector  $c$  and it entails the null hypothesis ( $\rightarrow$  I/4.3.2) that the product of this vector and the regression coefficients is zero:

$$H_0 : c^T \beta = 0 .\tag{2}$$



Consequently, the alternative hypothesis ( $\rightarrow$  I/4.3.3) of a two-tailed t-test ( $\rightarrow$  I/4.2.4) is

$$H_1 : c^T \beta \neq 0 \quad (3)$$

and the alternative hypothesis ( $\rightarrow$  I/4.3.3) of a one-sided t-test ( $\rightarrow$  I/4.2.4) would be

$$H_1 : c^T \beta < 0 \quad \text{or} \quad H_1 : c^T \beta > 0 . \quad (4)$$

Here,  $c$  is called the “contrast vector” and  $c^T \beta$  is called the “contrast value”. With estimated regression coefficients,  $c^T \hat{\beta}$  is called the “estimated contrast value”.

#### Sources:

- Stephan, Klaas Enno (2010): “Classical (frequentist) inference”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 7/9; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

#### 1.5.27 F-contrast

**Definition:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with  $n \times p$  design matrix  $X$  and  $p \times 1$  regression coefficients  $\beta$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, an F-contrast is specified by a  $p \times q$  matrix  $C$ , yielding a  $q \times 1$  vector  $\gamma = C^T \beta$ , and it entails the null hypothesis ( $\rightarrow$  I/4.3.2) that each value in this vector is zero:

$$H_0 : \gamma_1 = 0 \wedge \dots \wedge \gamma_q = 0 . \quad (2)$$

Consequently, the alternative hypothesis ( $\rightarrow$  I/4.3.3) of the statistical test ( $\rightarrow$  I/4.3.1) would be that at least one entry of this vector is non-zero:

$$H_1 : \gamma_1 \neq 0 \vee \dots \vee \gamma_q \neq 0 . \quad (3)$$

Here,  $C$  is called the “contrast matrix” and  $C^T \beta$  are called the “contrast values”. With estimated regression coefficients,  $C^T \hat{\beta}$  are called the “estimated contrast values”.

#### Sources:

- Stephan, Klaas Enno (2010): “Classical (frequentist) inference”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 23/25; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

#### 1.5.28 Contrast-based t-test

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

and a t-contrast ( $\rightarrow$  III/1.5.26) on the model parameters

$$\gamma = c^T \beta \quad \text{where} \quad c \in \mathbb{R}^{p \times 1} . \quad (2)$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T V^{-1} X)^{-1} c}} \quad (3)$$

with the parameter estimates ( $\rightarrow$  III/1.5.23)

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \end{aligned} \quad (4)$$

follows a t-distribution ( $\rightarrow$  II/3.3.1)

$$t \sim t(n-p) \quad (5)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$\begin{aligned} H_0 &: c^T \beta = 0 \\ H_1 &: c^T \beta > 0 . \end{aligned} \quad (6)$$

**Proof:**

1) We know that the estimated regression coefficients in linear regression follow a multivariate normal distribution ( $\rightarrow$  III/1.5.19):

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T V^{-1} X)^{-1}) . \quad (7)$$

Thus, the estimated contrast value ( $\rightarrow$  III/1.5.26)  $\hat{\gamma} = c^T \hat{\beta}$  is distributed according to a univariate normal distribution ( $\rightarrow$  II/4.1.14):

$$\hat{\gamma} \sim \mathcal{N}(c^T \beta, \sigma^2 c^T (X^T V^{-1} X)^{-1} c) . \quad (8)$$

Now, define the random variable  $z$  by dividing  $\hat{\gamma}$  by its standard deviation:

$$z = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2 c^T (X^T V^{-1} X)^{-1} c}} . \quad (9)$$

Again applying the linear transformation theorem ( $\rightarrow$  II/4.1.14), this is distributed as

$$z \sim \mathcal{N}\left(\frac{c^T \beta}{\sqrt{\sigma^2 c^T (X^T V^{-1} X)^{-1} c}}, 1\right) \quad (10)$$

and thus follows a standard normal distribution ( $\rightarrow$  II/3.2.3) under the null hypothesis ( $\rightarrow$  I/4.3.2):

$$z \sim \mathcal{N}(0, 1), \quad \text{if } H_0 . \quad (11)$$

2) We also know that the residual sum of squares ( $\rightarrow$  III/1.5.9), divided the true error variance ( $\rightarrow$  III/1.5.1)

$$v = \frac{1}{\sigma^2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} = \frac{1}{\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \quad (12)$$

is following a chi-squared distribution ( $\rightarrow$  III/1.5.20):

$$v \sim \chi^2(n - p) . \quad (13)$$

3) Because the estimated regression coefficients and the vector of residuals are independent from each other ( $\rightarrow$  III/1.5.17)

$$\hat{\beta} \quad \text{and} \quad \hat{\varepsilon} \quad \text{ind.} \quad (14)$$

and thus, the estimated contrast values are also independent from the function of the residual sum of squares

$$z = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2 c^T (X^T V^{-1} X)^{-1} c}} \quad \text{and} \quad v = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} \quad \text{ind.} , \quad (15)$$

the following quantity is, by definition, t-distributed ( $\rightarrow$  II/3.3.1)

$$t = \frac{z}{\sqrt{v/(n - p)}} \sim t(n - p), \quad \text{if } H_0 \quad (16)$$

and the quantity can be evaluated as:

$$\begin{aligned} t &\stackrel{(16)}{=} \frac{z}{\sqrt{v/(n - p)}} \\ &\stackrel{(15)}{=} \frac{c^T \hat{\beta}}{\sqrt{\sigma^2 c^T (X^T V^{-1} X)^{-1} c}} \cdot \sqrt{\frac{n - p}{\hat{\varepsilon}^T \hat{\varepsilon} / \sigma^2}} \\ &= \frac{c^T \hat{\beta}}{\sqrt{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p} \cdot c^T (X^T V^{-1} X)^{-1} c}} \\ &\stackrel{(12)}{=} \frac{c^T \hat{\beta}}{\sqrt{\frac{(y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta})}{n - p} \cdot c^T (X^T V^{-1} X)^{-1} c}} \\ &\stackrel{(4)}{=} \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T V^{-1} X)^{-1} c}} . \end{aligned} \quad (17)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) in (6) can be rejected when  $t$  from (17) is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from Student's t-distribution ( $\rightarrow$  II/3.3.1) with  $n - p$  degrees of freedom ( $\rightarrow$  I/1.5.12) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ .

■

#### Sources:

- Stephan, Klaas Enno (2010): “Classical (frequentist) inference”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 7/9; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.
- Walter, Henrik (ed.) (2005): “Datenanalyse für funktionell bildgebende Verfahren”; in: *Funktionelle Bildgebung in Psychiatrie und Psychotherapie*, Schattauer, Stuttgart/New York, 2005, p. 40; URL: [https://books.google.de/books?id=edWzKAHi7jQC&source=gbs\\_navlinks\\_s](https://books.google.de/books?id=edWzKAHi7jQC&source=gbs_navlinks_s).
- jld (2018): “Understanding t-test for linear regression”; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: <https://stats.stackexchange.com/a/344008>.

- Soch, Joram (2020): “Distributional Transformation Improves Decoding Accuracy When Predicting Chronological Age From Structural MRI”; in: *Frontiers in Psychiatry*, vol. 11, art. 604268, eqs. 8/9; URL: <https://www.frontiersin.org/articles/10.3389/fpsy.2020.604268/full>; DOI: 10.3389/fpsy.2020.604268.

### 1.5.29 Contrast-based F-test

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

and an F-contrast ( $\rightarrow$  III/1.5.27) on the model parameters

$$\gamma = C^T \beta \quad \text{where} \quad C \in \mathbb{R}^{p \times q} . \quad (2)$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F = \hat{\beta}^T C (\hat{\sigma}^2 C^T (X^T V^{-1} X)^{-1} C)^{-1} C^T \hat{\beta} / q \quad (3)$$

with the parameter estimates ( $\rightarrow$  III/1.5.23)

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \end{aligned} \quad (4)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F \sim F(q, n-p) \quad (5)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2)

$$\begin{aligned} H_0 &: \gamma_1 = 0 \wedge \dots \wedge \gamma_q = 0 \\ H_1 &: \gamma_1 \neq 0 \vee \dots \vee \gamma_q \neq 0 . \end{aligned} \quad (6)$$

**Proof:**

1) We know that the estimated regression coefficients in linear regression follow a multivariate normal distribution ( $\rightarrow$  III/1.5.19):

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T V^{-1} X)^{-1}) . \quad (7)$$

Thus, the estimated contrast vector ( $\rightarrow$  III/1.5.27)  $\hat{\gamma} = C^T \hat{\beta}$  is also distributed according to a multivariate normal distribution ( $\rightarrow$  II/4.1.14):

$$\hat{\gamma} \sim \mathcal{N}(C^T \beta, \sigma^2 C^T (X^T V^{-1} X)^{-1} C) . \quad (8)$$

Substituting the noise variance  $\sigma^2$  with the noise precision  $\tau = 1/\sigma^2$ , we can also write this down as a conditional distribution ( $\rightarrow$  I/1.5.4):

$$\hat{\gamma} | \tau \sim \mathcal{N}(C^T \beta, (\tau Q)^{-1}) \quad \text{with} \quad Q = (C^T (X^T V^{-1} X)^{-1} C)^{-1} . \quad (9)$$

2) We also know that the residual sum of squares ( $\rightarrow$  III/1.5.9), divided the true error variance ( $\rightarrow$  III/1.5.1)

$$\frac{1}{\sigma^2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} = \frac{1}{\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \quad (10)$$

is following a chi-squared distribution ( $\rightarrow$  III/1.5.20):

$$\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} = \tau \hat{\varepsilon}^T \hat{\varepsilon} \sim \chi^2(n-p) . \quad (11)$$

The chi-squared distribution is a special case of the gamma distribution ( $\rightarrow$  II/3.7.2)

$$X \sim \chi^2(k) \quad \Rightarrow \quad X \sim \text{Gam} \left( \frac{k}{2}, \frac{1}{2} \right) \quad (12)$$

and the gamma distribution changes under multiplication ( $\rightarrow$  II/3.4.6) in the following way:

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad cX \sim \text{Gam} \left( a, \frac{b}{c} \right) . \quad (13)$$

Thus, combining (12) and (13) with (11), we obtain the marginal distribution ( $\rightarrow$  I/1.5.3) of  $\tau$  as:

$$\frac{1}{\hat{\varepsilon}^T \hat{\varepsilon}} (\tau \hat{\varepsilon}^T \hat{\varepsilon}) = \tau \sim \text{Gam} \left( \frac{n-p}{2}, \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{2} \right) . \quad (14)$$

3) Note that the joint distribution ( $\rightarrow$  I/1.5.2) of  $\hat{\gamma}$  and  $\tau$  is, following from (9) and (14) and by definition, a normal-gamma distribution ( $\rightarrow$  II/4.4.1):

$$\hat{\gamma}, \tau \sim \text{NG} \left( C^T \beta, Q, \frac{n-p}{2}, \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{2} \right) . \quad (15)$$

The marginal distribution of a normal-gamma distribution with respect to the normal random variable, is a multivariate t-distribution ( $\rightarrow$  II/4.4.8):

$$X, Y \sim \text{NG}(\mu, \Lambda, a, b) \quad \Rightarrow \quad X \sim t \left( \mu, \left( \frac{a}{b} \Lambda \right)^{-1}, 2a \right) . \quad (16)$$

Thus, the marginal distribution ( $\rightarrow$  I/1.5.3) of  $\hat{\gamma}$  is:

$$\hat{\gamma} \sim t \left( C^T \beta, \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} Q \right)^{-1}, n-p \right) . \quad (17)$$

4) Because of the following relationship between the multivariate t-distribution and the F-distribution ( $\rightarrow$  II/4.3.3)

$$X \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad (X - \mu)^T \Sigma^{-1} (X - \mu) / \nu \sim F(n, \nu) , \quad (18)$$

the following quantity is, by definition, F-distributed ( $\rightarrow$  II/3.8.1)

$$F = (\hat{\gamma} - C^T \beta)^T \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} Q \right) (\hat{\gamma} - C^T \beta) / q \sim F(q, n-p) \quad (19)$$

and under the null hypothesis ( $\rightarrow$  I/4.3.2) (6), it can be evaluated as:

$$\begin{aligned}
F &\stackrel{(19)}{=} (\hat{\gamma} - C^T \beta)^T \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} Q \right) (\hat{\gamma} - C^T \beta) / q \\
&\stackrel{(6)}{=} \hat{\gamma}^T \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} Q \right) \hat{\gamma} / q \\
&\stackrel{(2)}{=} \hat{\beta}^T C \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} Q \right) C^T \hat{\beta} / q \\
&\stackrel{(9)}{=} \hat{\beta}^T C \left( \frac{n-p}{\hat{\varepsilon}^T \hat{\varepsilon}} (C^T (X^T V^{-1} X)^{-1} C)^{-1} \right) C^T \hat{\beta} / q \\
&\stackrel{(10)}{=} \hat{\beta}^T C \left( \frac{n-p}{(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})} (C^T (X^T V^{-1} X)^{-1} C)^{-1} \right) C^T \hat{\beta} / q \\
&\stackrel{(4)}{=} \hat{\beta}^T C \left( \frac{1}{\hat{\sigma}^2} (C^T (X^T V^{-1} X)^{-1} C)^{-1} \right) C^T \hat{\beta} / q \\
&= \hat{\beta}^T C (\hat{\sigma}^2 C^T (X^T V^{-1} X)^{-1} C)^{-1} C^T \hat{\beta} / q .
\end{aligned} \tag{20}$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) in (6) can be rejected when  $F$  from (20) is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from Fisher's F-distribution ( $\rightarrow$  II/3.8.1) with  $q$  numerator and  $n - p$  denominator degrees of freedom ( $\rightarrow$  I/1.5.12) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ .

■

#### Sources:

- Stephan, Klaas Enno (2010): “Classical (frequentist) inference”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 23/25; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.
- Koch, Karl-Rudolf (2007): “Multivariate Distributions”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 2.5, eqs. 2.202, 2.213, 2.211; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.
- jld (2018): “Understanding t-test for linear regression”; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: <https://stats.stackexchange.com/a/344008>.
- Penny, William (2006): “Comparing nested GLMs”; in: *Mathematics for Brain Imaging*, ch. 2.3, pp. 51-52, eq. 2.9; URL: [https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi\\_course.pdf](https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf).

#### 1.5.30 t-test for single regressor

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

using the  $n \times p$  design matrix  $X$  and the parameter estimates ( $\rightarrow$  III/1.5.23)

$$\begin{aligned}
\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\
\hat{\sigma}^2 &= \frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) .
\end{aligned} \tag{2}$$

Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$t_j = \frac{\hat{\beta}_j}{\sqrt{(\hat{\varepsilon}^T V^{-1} \hat{\varepsilon}) / (n - p) \sigma_{jj}}} \quad (3)$$

with the  $n$ -dimensional vector of residuals ( $\rightarrow$  III/1.5.14)

$$\hat{\varepsilon} = y - X\hat{\beta} \quad (4)$$

and  $\sigma_{jj}$  equal to the  $j$ -th diagonal element of the parameter covariance matrix ( $\rightarrow$  III/1.5.19)

$$\sigma_{jj} = \left[ (X^T V^{-1} X)^{-1} \right]_{jj} \quad (5)$$

follows a t-distribution ( $\rightarrow$  II/3.3.1)

$$t_j \sim t(n - p) \quad (6)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the  $j$ -th regression coefficient ( $\rightarrow$  III/1.5.1) is zero:

$$H_0 : \beta_j = 0 . \quad (7)$$

**Proof:** This is a special case of the contrast-based t-test for multiple linear regression ( $\rightarrow$  III/1.5.28) based on the following t-statistic ( $\rightarrow$  II/3.3.1):

$$t = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T V^{-1} X)^{-1} c}} \sim t(n - p) . \quad (8)$$

In this special case, the contrast vector ( $\rightarrow$  III/1.5.26) is equal to the  $j$ -th elementary vector  $e_j$  (a  $p \times 1$  vector of zeros, with a single 1 in the  $j$ -th entry)

$$c = e_j = [0, \dots, 0, 1, 0, \dots, 0]^T , \quad (9)$$

such that the null hypothesis is given by

$$H_0 : c^T \beta = e_j^T \beta = \beta_j = 0 \quad (10)$$

and the test statistic becomes

$$\begin{aligned} t_j &= \frac{e_j^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 e_j^T (X^T V^{-1} X)^{-1} e_j}} \\ &= \frac{[0, \dots, 0, 1, 0, \dots, 0] \left[ \hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \hat{\beta}_j, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p \right]^T}{\sqrt{\frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) [0, \dots, 1, \dots, 0] (X^T V^{-1} X)^{-1} [0, \dots, 1, \dots, 0]^T}} \\ &= \frac{\hat{\beta}_j}{\sqrt{\frac{1}{n-p} (\hat{\varepsilon}^T V^{-1} \hat{\varepsilon}) \left[ (X^T V^{-1} X)^{-1} \right]_{jj}}} \\ &= \frac{\hat{\beta}_j}{\sqrt{(\hat{\varepsilon}^T V^{-1} \hat{\varepsilon}) / (n - p) \sigma_{jj}}} . \end{aligned} \quad (11)$$

■

**Sources:**

- Ostwald, Dirk (2023): “T-Statistiken”; in: *Allgemeines Lineares Modell*, Einheit (7), Folien 20, 27; URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/7\\_T\\_Statistiken-p-9968.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/7_T_Statistiken-p-9968.pdf).

**1.5.31 F-test for multiple regressors**

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

the design matrix and regression coefficients of which are partitioned as

$$\begin{aligned} X &= \begin{bmatrix} X_0 & X_1 \end{bmatrix} \in \mathbb{R}^{n \times p} \quad \text{where} \quad X_0 \in \mathbb{R}^{n \times p_0} \quad \text{and} \quad X_1 \in \mathbb{R}^{n \times p_1} \\ \beta &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \in \mathbb{R}^{p \times 1} \quad \text{where} \quad \beta_0 \in \mathbb{R}^{p_0 \times 1} \quad \text{and} \quad \beta_1 \in \mathbb{R}^{p_1 \times 1} \end{aligned} \quad (2)$$

with  $p = p_0 + p_1$ . Then, the test statistic ( $\rightarrow$  I/4.3.5)

$$F = \frac{(\hat{\varepsilon}_0^T V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^T V^{-1} \hat{\varepsilon})/p_1}{\hat{\varepsilon}^T V^{-1} \hat{\varepsilon}/(n - p)} \quad (3)$$

follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F \sim F(p_1, n - p) \quad (4)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that all regression coefficients ( $\rightarrow$  III/1.5.1)  $\beta_1$  are zero:

$$H_0 : \beta_1 = 0_{p_1} \quad \Leftrightarrow \quad \beta_{1j} = 0 \quad \text{for all} \quad j = 1, \dots, p_1. \quad (5)$$

In (3),  $\hat{\varepsilon}$  and  $\hat{\varepsilon}_0$  are the residual vectors ( $\rightarrow$  III/1.5.14) when using either the full design matrix  $X$  or the reduced design matrix  $X_0$ :

$$\begin{aligned} \hat{\varepsilon} &= y - X\hat{\beta} \quad \text{with} \quad \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\varepsilon}_0 &= y - X_0 \hat{\beta}_0 \quad \text{with} \quad \hat{\beta}_0 = (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y. \end{aligned} \quad (6)$$

**Proof:** This is a special case of the contrast-based F-test for multiple linear regression ( $\rightarrow$  III/1.5.29) based on the F-statistic ( $\rightarrow$  I/4.3.5)

$$F = \hat{\beta}^T C (\hat{\sigma}^2 C^T (X^T V^{-1} X)^{-1} C)^{-1} C^T \hat{\beta} / q \quad (7)$$

which follows an F-distribution ( $\rightarrow$  II/3.8.1) under the null hypothesis ( $\rightarrow$  I/4.3.2) that the product of the contrast matrix ( $\rightarrow$  III/1.5.27)  $C \in \mathbb{R}^{p \times q}$  and the regression coefficients ( $\rightarrow$  III/1.5.1) equals zero:



$$F \sim F(q, n - p), \quad \text{if } C^T \beta = 0_q = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (8)$$

In (7),  $\hat{\sigma}^2$  is an estimate of the noise variance ( $\rightarrow$  III/1.5.1) calculated as the weighted ( $\rightarrow$  III/1.5.21) residual sum of squares ( $\rightarrow$  III/1.5.9), divided by  $n - p$ :

$$\hat{\sigma}^2 = \frac{1}{n - p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}). \quad (9)$$

In the present case, in order to compare the full model specified by  $X$  against the reduced model specified by  $X_0$ , we have to define the contrast matrix ( $\rightarrow$  III/1.5.27) as a vertical concatenation of a zero matrix on the first  $p_0$  components and an identity matrix on the last  $p_1$  components of  $\beta$ ,

$$C_1 = \begin{bmatrix} 0_{p_0, p_1} \\ I_{p_1} \end{bmatrix} \in \mathbb{R}^{p \times p_1}, \quad (10)$$

i.e. specify an omnibus F-contrast that tests the alternative hypothesis ( $\rightarrow$  I/4.3.3) that any of the coefficients  $\beta_1$  associated with the regressors  $X_1$  is different from zero against the null hypothesis ( $\rightarrow$  I/4.3.2) that all those coefficients are zero:

$$\begin{aligned} H_0 : C_1^T \beta &= \begin{bmatrix} 0_{p_0, p_1} \\ I_{p_1} \end{bmatrix}^T \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0_{p_1} \quad \Leftrightarrow \quad \beta_{1j} = 0 \quad \text{for all } j = 1, \dots, p_1 \\ \Rightarrow H_1 &\Leftrightarrow \neg H_0 : C_1^T \beta = \beta_1 \neq 0_{p_1} \quad \Leftrightarrow \quad \beta_{1j} \neq 0 \quad \text{for at least one } j = 1, \dots, p_1. \end{aligned} \quad (11)$$

Thus, plugging  $C = C_1$  and  $q = p_1$  into (7) and noting that  $\hat{\sigma}^2$  from (9) is a scalar, we obtain:

$$\begin{aligned} F &= \hat{\beta}^T C_1 (\hat{\sigma}^2 C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} / p_1 \\ &\stackrel{(9)}{=} \frac{\hat{\beta}^T C_1 (C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} / p_1}{(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) / (n - p)}. \end{aligned} \quad (12)$$

Here, we take note of the fact that the denominator in (12) is already equal to the denominator in (3):

$$(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) / (n - p) \stackrel{(6)}{=} \hat{\varepsilon}^T V^{-1} \hat{\varepsilon} / (n - p). \quad (13)$$

Therefore, what remains to be shown is that the numerator in (12) is equal to the numerator in (3):

$$\hat{\beta}^T C_1 (C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} / p_1 = (\hat{\varepsilon}_0^T V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^T V^{-1} \hat{\varepsilon}) / p_1. \quad (14)$$

To do this, we start with the inner-most matrix:

$$\begin{aligned}
X^T V^{-1} X &\stackrel{(2)}{=} \begin{bmatrix} X_0 & X_1 \end{bmatrix}^T V^{-1} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \\
&= \begin{bmatrix} X_0^T \\ X_1^T \end{bmatrix} V^{-1} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \\
&= \begin{bmatrix} X_0^T V^{-1} X_0 & X_0^T V^{-1} X_1 \\ X_1^T V^{-1} X_0 & X_1^T V^{-1} X_1 \end{bmatrix}.
\end{aligned} \tag{15}$$

The inverse of a block matrix is:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}. \tag{16}$$

Note that, with the contrast matrix  $C_1$ , we only extract the lower-right part of the inverse block matrix, so that we have:

$$\begin{aligned}
(C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} &\stackrel{(10)}{=} \left( \begin{bmatrix} 0_{p_1, p_0} & I_{p_1} \end{bmatrix} (X^T V^{-1} X)^{-1} \begin{bmatrix} 0_{p_0, p_1} \\ I_{p_1} \end{bmatrix} \right)^{-1} \\
&\stackrel{(16)}{=} \left( (X_1^T V^{-1} X_1 - X_1^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X_1)^{-1} \right)^{-1} \\
&= X_1^T V^{-1} X_1 - X_1^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X_1.
\end{aligned} \tag{17}$$

We call this  $p_1 \times p_1$  matrix  $E$  and note that it can be written as

$$\begin{aligned}
E &\stackrel{(17)}{=} X_1^T (V^{-1} - V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1}) X_1 \\
&= X_1^T (V^{-1} - F) X_1
\end{aligned} \tag{18}$$

where the  $n \times n$  matrix  $F$  is given as follows:

$$F = V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1}. \tag{19}$$

Let  $\hat{\beta}_{0(X)}$  denote the first  $p_0$  entries of  $\hat{\beta}$ , i.e. estimates of the coefficients belonging to  $X_0$ , but estimated with  $X$  (as opposed to  $\hat{\beta}_0$  estimated with  $X_0$  given by (6)):

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix}. \tag{20}$$

Then, it obviously holds that

$$\begin{aligned}
X \hat{\beta} &\stackrel{(20)}{=} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix} \\
&= X_0 \hat{\beta}_{0(X)} + X_1 \hat{\beta}_1 \\
&\Leftrightarrow \\
X_1 \hat{\beta}_1 &= X \hat{\beta} - X_0 \hat{\beta}_{0(X)}.
\end{aligned} \tag{21}$$

Next, we focus on  $C_1^T \hat{\beta}$  which simply extracts  $\hat{\beta}_1$ :

$$\begin{aligned} C_1^T \hat{\beta} &\stackrel{(20)}{=} \begin{bmatrix} 0_{p_1, p_0} & I_{p_1} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix} \\ &= \hat{\beta}_1 . \end{aligned} \quad (22)$$

With these identities in mind, we can get back to our main quantity of interest from (14):

$$\begin{aligned} &\hat{\beta}^T C_1 (C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} \\ &\stackrel{(22)}{=} \hat{\beta}_1^T E \hat{\beta}_1 \\ &\stackrel{(18)}{=} \hat{\beta}_1^T X_1^T (V^{-1} - V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1}) X_1 \hat{\beta}_1 \\ &\stackrel{(21)}{=} (\hat{\beta}^T X^T - \hat{\beta}_{0(X)}^T X_0^T) (V^{-1} - V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1}) (X \hat{\beta} - X_0 \hat{\beta}_{0(X)}) \\ &\stackrel{(19)}{=} (\hat{\beta}^T X^T V^{-1} - \hat{\beta}^T X^T F - \hat{\beta}_{0(X)}^T X_0^T V^{-1} + \hat{\beta}_{0(X)}^T X_0^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1}) (X \hat{\beta} - X_0 \hat{\beta}_{0(X)}) \\ &= (\hat{\beta}^T X^T V^{-1} - \hat{\beta}^T X^T F - \hat{\beta}_{0(X)}^T X_0^T V^{-1} + \hat{\beta}_{0(X)}^T X_0^T V^{-1}) (X \hat{\beta} - X_0 \hat{\beta}_{0(X)}) \\ &= (\hat{\beta}^T X^T V^{-1} - \hat{\beta}^T X^T F) (X \hat{\beta} - X_0 \hat{\beta}_{0(X)}) \\ &\stackrel{(19)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X_0 \hat{\beta}_{0(X)} - \hat{\beta}^T X^T F X \hat{\beta} + \hat{\beta}^T X^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X_0 \hat{\beta}_{0(X)} \\ &= \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X_0 \hat{\beta}_{0(X)} - \hat{\beta}^T X^T F X \hat{\beta} + \hat{\beta}^T X^T V^{-1} X_0 \hat{\beta}_{0(X)} \\ &= \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T F X \hat{\beta} \\ &\stackrel{(19)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X \hat{\beta} . \end{aligned} \quad (23)$$

Let the residual vector of the full model be defined as given by (6)

$$\hat{\varepsilon} = y - X \hat{\beta} \quad \Leftrightarrow \quad y = X \hat{\beta} + \hat{\varepsilon} \quad (24)$$

and consider the term  $X^T V^{-1} \hat{\varepsilon}$ . Using the residual-forming matrix expression of the residual vector ( $\rightarrow$  III/1.5.14), we can show that this matrix product is zero:

$$\begin{aligned} X^T V^{-1} \hat{\varepsilon} &= X^T V^{-1} (I_n - X (X^T V^{-1} X)^{-1} X^T V^{-1}) y \\ &= X^T V^{-1} y - X^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ &= X^T V^{-1} y - X^T V^{-1} y \\ &= 0_p . \end{aligned} \quad (25)$$

From this, it follows that the product  $X_0^T V^{-1} \hat{\varepsilon}$  is also zero:

$$\begin{aligned}
X^T V^{-1} \hat{\varepsilon} &= 0_p \\
\begin{bmatrix} X_0^T \\ X_1^T \end{bmatrix} V^{-1} \hat{\varepsilon} &= 0_p \\
\begin{bmatrix} X_0^T V^{-1} \hat{\varepsilon} \\ X_1^T V^{-1} \hat{\varepsilon} \end{bmatrix} &= \begin{bmatrix} 0_{p_0} \\ 0_{p_1} \end{bmatrix} \\
&\Leftrightarrow \\
X_0^T V^{-1} \hat{\varepsilon} &= 0_{p_0} .
\end{aligned} \tag{26}$$

Thus, any term containing  $X_0^T V^{-1} \hat{\varepsilon} = 0_{p_0}$  can be added to a sum without changing the value of this sum. Continuing from above, we therefore write:

$$\begin{aligned}
&\hat{\beta}^T C_1 (C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} \\
&\stackrel{(23)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X \hat{\beta} \\
&\stackrel{(26)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X \hat{\beta} \\
&\quad + 2 \hat{\beta}^T X^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} \hat{\varepsilon} + \hat{\varepsilon}^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} \hat{\varepsilon} \\
&= \hat{\beta}^T X^T V^{-1} X \hat{\beta} - (X \hat{\beta} + \hat{\varepsilon})^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} (X \hat{\beta} + \hat{\varepsilon}) \\
&\stackrel{(24)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y .
\end{aligned} \tag{27}$$

In the next transformations, we will make use of the weighted least squares parameter estimates ( $\rightarrow$  III/1.5.21)

$$\begin{aligned}
\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\
\hat{\beta}_0 &= (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y
\end{aligned} \tag{28}$$

and the fact that matrices and their inverses cancel out:

$$\begin{aligned}
X^T V^{-1} X (X^T V^{-1} X)^{-1} &= (X^T V^{-1} X)^{-1} X^T V^{-1} X = I_p \\
X_0^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} &= (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X_0 = I_{p_0} .
\end{aligned} \tag{29}$$

Continuing from above, we have:

$$\begin{aligned}
& \hat{\beta}^T C_1 (C_1^T (X^T V^{-1} X)^{-1} C_1)^{-1} C_1^T \hat{\beta} \\
& \stackrel{(27)}{=} \hat{\beta}^T X^T V^{-1} X \hat{\beta} - y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y \\
& \stackrel{(28)}{=} y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y - y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y \\
& = y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y - y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y \\
& = y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y - 2y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y \\
& - y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y + 2y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y \\
& \stackrel{(29)}{=} y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y - 2y^T V^{-1} X_0 (X_0^T V^{-1} X_0)^{-1} X_0^T V^{-1} y \\
& - y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y + 2y^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} y \\
& \stackrel{(28)}{=} \hat{\beta}_0^T X_0^T V^{-1} X_0 \hat{\beta}_0 - 2y^T V^{-1} X_0 \hat{\beta}_0 - \hat{\beta}^T X^T V^{-1} X \hat{\beta} + 2y^T V^{-1} X \hat{\beta} \\
& = y^T V^{-1} y - 2y^T V^{-1} X_0 \hat{\beta}_0 + \hat{\beta}_0^T X_0^T V^{-1} X_0 \hat{\beta}_0 - y^T V^{-1} y + 2y^T V^{-1} X \hat{\beta} - \hat{\beta}^T X^T V^{-1} X \hat{\beta} \\
& = \left( y^T V^{-1} y - 2y^T V^{-1} X_0 \hat{\beta}_0 + \hat{\beta}_0^T X_0^T V^{-1} X_0 \hat{\beta}_0 \right) - \left( y^T V^{-1} y - 2y^T V^{-1} X \hat{\beta} + \hat{\beta}^T X^T V^{-1} X \hat{\beta} \right) \\
& = (y - X_0 \hat{\beta}_0)^T V^{-1} (y - X_0 \hat{\beta}_0) - (y - X \hat{\beta})^T V^{-1} (y - X \hat{\beta}) \\
& \stackrel{(6)}{=} \hat{\varepsilon}_0^T V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^T V^{-1} \hat{\varepsilon} .
\end{aligned} \tag{30}$$

With that, it is shown that (14) is true which, together with (13), finally demonstrates that the F-value in (12) is equal to the test statistic given by (3). This completes the proof. ■

### Sources:

- Ostwald, Dirk (2023): “F-Statistiken”; in: *Allgemeines Lineares Modell*, Einheit (8), Folien 20, 24;  
URL: [https://www.ipsy.ovgu.de/ipsy\\_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/8\\_F\\_Statistiken-p-9972.pdf](https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/8_F_Statistiken-p-9972.pdf).

### 1.5.32 Deviance function

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$  with correlation structure ( $\rightarrow$  I/1.17.6)  $V$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \tag{1}$$

Then, the deviance ( $\rightarrow$  IV/2.3.2) for this model is

$$D(\beta, \sigma^2) = \text{RSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] \tag{2}$$

under uncorrelated observations ( $\rightarrow$  III/1.5.1), i.e. if  $V = I_n$ , and

$$D(\beta, \sigma^2) = \text{wRSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] + \log |V| , \tag{3}$$

in the general case, i.e. if  $V \neq I_n$ , where RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9) and wRSS is the weighted residual sum of squares ( $\rightarrow$  III/1.5.22).

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for multiple linear regression is given by ( $\rightarrow$  III/1.5.23)

$$\begin{aligned}
p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\
&= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right], \tag{4}
\end{aligned}$$

such that, with  $|\sigma^2 V| = (\sigma^2)^n |V|$ , the log-likelihood function ( $\rightarrow$  I/4.1.2) for this model becomes ( $\rightarrow$  III/1.5.23)

$$\begin{aligned}
\text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta). \tag{5}
\end{aligned}$$

The last term can be expressed in terms of the (weighted) residual sum of squares ( $\rightarrow$  III/1.5.9) as

$$\begin{aligned}
-\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) &= -\frac{1}{2\sigma^2} (Wy - WX\beta)^T (Wy - WX\beta) \\
&= -\frac{1}{2\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n (W\varepsilon_i)^2 \right) = -\frac{\text{wRSS}}{2\sigma^2} \tag{6}
\end{aligned}$$

where  $W = V^{-1/2}$ . Plugging (6) into (5) and multiplying with  $-2$ , we obtain the deviance ( $\rightarrow$  IV/2.3.2) as

$$\begin{aligned}
D(\beta, \sigma^2) &= -2 \text{LL}(\beta, \sigma^2) \\
&= -2 \left( -\frac{\text{wRSS}}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |V| \right) \\
&= \text{wRSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] + \log |V| \tag{7}
\end{aligned}$$

which proves the result in (3). Assuming  $V = I_n$ , we have

$$\begin{aligned}
-\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) &= -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \\
&= -\frac{1}{2\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) = -\frac{\text{RSS}}{2\sigma^2} \tag{8}
\end{aligned}$$

and

$$\frac{1}{2} \log |V| = \frac{1}{2} \log |I_n| = \frac{1}{2} \log 1 = 0, \tag{9}$$

such that

$$D(\beta, \sigma^2) = \text{RSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] \tag{10}$$

which proves the result in (2). This completes the proof. ■

### 1.5.33 Akaike information criterion

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the Akaike information criterion ( $\rightarrow$  IV/2.1.1) for this model is

$$\text{AIC}(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + 2(p + 1) \quad (2)$$

where wRSS is the weighted residual sum of squares ( $\rightarrow$  III/1.5.9),  $p$  is the number of regressors ( $\rightarrow$  III/1.5.1) in the design matrix  $X$  and  $n$  is the number of observations ( $\rightarrow$  III/1.5.1) in the data vector  $y$ .

**Proof:** The Akaike information criterion ( $\rightarrow$  IV/2.1.1) is defined as

$$\text{AIC}(m) = -2 \text{MLL}(m) + 2k \quad (3)$$

where  $\text{MLL}(m)$  is the maximum log-likelihood ( $\rightarrow$  I/4.1.4) is  $k$  is the number of free parameters in  $m$ .

The maximum log-likelihood for multiple linear regression ( $\rightarrow$  III/1.5.24) is given by

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{wRSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] - \frac{1}{2} \log |V| \quad (4)$$

and the number of free parameters in multiple linear regression ( $\rightarrow$  III/1.5.1) is  $k = p + 1$ , i.e. one for each regressor in the design matrix ( $\rightarrow$  III/1.5.1)  $X$ , plus one for the noise variance ( $\rightarrow$  III/1.5.1)  $\sigma^2$ .

Thus, the AIC of  $m$  follows from (3) and (4) as

$$\text{AIC}(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + 2(p + 1) . \quad (5)$$

■

#### Sources:

- Claeskens G, Hjort NL (2008): “Akaike’s information criterion”; in: *Model Selection and Model Averaging*, ex. 2.2, p. 66; URL: <https://www.cambridge.org/core/books/model-selection-and-model-averaging/E6F1EC77279D1223423BB64FC3A12C37>; DOI: 10.1017/CBO9780511790485.

### 1.5.34 Bayesian information criterion

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the Bayesian information criterion ( $\rightarrow$  IV/2.2.1) for this model is

$$\text{BIC}(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \log(n) (p + 1) \quad (2)$$

where wRSS is the weighted residual sum of squares ( $\rightarrow$  III/1.5.9),  $p$  is the number of regressors ( $\rightarrow$  III/1.5.1) in the design matrix  $X$  and  $n$  is the number of observations ( $\rightarrow$  III/1.5.1) in the data vector  $y$ .

**Proof:** The Bayesian information criterion ( $\rightarrow$  IV/2.2.1) is defined as

$$\text{BIC}(m) = -2 \text{MLL}(m) + k \log(n) \quad (3)$$

where  $\text{MLL}(m)$  is the maximum log-likelihood ( $\rightarrow$  I/4.1.4),  $k$  is the number of free parameters in  $m$  and  $n$  is the number of observations.

The maximum log-likelihood for multiple linear regression ( $\rightarrow$  III/1.5.24) is given by

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{wRSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] - \frac{1}{2} \log |V| \quad (4)$$

and the number of free parameters in multiple linear regression ( $\rightarrow$  III/1.5.1) is  $k = p + 1$ , i.e. one for each regressor in the design matrix ( $\rightarrow$  III/1.5.1)  $X$ , plus one for the noise variance ( $\rightarrow$  III/1.5.1)  $\sigma^2$ .

Thus, the BIC of  $m$  follows from (3) and (4) as

$$\text{BIC}(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \log(n) (p + 1) . \quad (5)$$

■

### 1.5.35 Corrected Akaike information criterion

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) for this model is

$$\text{AIC}_c(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2n(p+1)}{n-p-2} \quad (2)$$

where wRSS is the weighted residual sum of squares ( $\rightarrow$  III/1.5.9),  $p$  is the number of regressors ( $\rightarrow$  III/1.5.1) in the design matrix  $X$  and  $n$  is the number of observations ( $\rightarrow$  III/1.5.1) in the data vector  $y$ .

**Proof:** The corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) is defined as

$$\text{AIC}_c(m) = \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} \quad (3)$$

where  $\text{AIC}(m)$  is the Akaike information criterion ( $\rightarrow$  IV/2.1.1),  $k$  is the number of free parameters in  $m$  and  $n$  is the number of observations.

The Akaike information criterion for multiple linear regression ( $\rightarrow$  III/1.5.33) is given by

$$\text{AIC}(m) = n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + 2(p + 1) \quad (4)$$

and the number of free parameters in multiple linear regression ( $\rightarrow$  III/1.5.1) is  $k = p + 1$ , i.e. one for each regressor in the design matrix ( $\rightarrow$  III/1.5.1)  $X$ , plus one for the noise variance ( $\rightarrow$  III/1.5.1)  $\sigma^2$ .

Thus, the corrected AIC of  $m$  follows from (3) and (4) as



$$\begin{aligned}
\text{AIC}_c(m) &= n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + 2k + \frac{2k^2 + 2k}{n - k - 1} \\
&= n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2nk - 2k^2 - 2k}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\
&= n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2nk}{n - k - 1} \\
&= n \log \left( \frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2n(p+1)}{n - p - 2}
\end{aligned} \tag{5}$$

### Sources:

- Claeskens G, Hjort NL (2008): “Akaike’s information criterion”; in: *Model Selection and Model Averaging*, ex. 2.5, p. 67; URL: <https://www.cambridge.org/core/books/model-selection-and-model-averaging/E6F1EC77279D1223423BB64FC3A12C37>; DOI: 10.1017/CBO9780511790485.

## 1.6 Bayesian linear regression

### 1.6.1 Conjugate prior distribution

**Theorem:** Let

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure  $V$  as well as unknown  $p \times 1$  regression coefficients  $\beta$  and unknown noise variance  $\sigma^2$ .

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for this model is a normal-gamma distribution ( $\rightarrow$  II/4.4.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \tag{2}$$

where  $\tau = 1/\sigma^2$  is the inverse noise variance or noise precision.

**Proof:** By definition, a conjugate prior ( $\rightarrow$  I/5.2.5) is a prior distribution ( $\rightarrow$  I/5.1.3) that, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right] \tag{3}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[ -\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \tag{4}$$

using the noise precision  $\tau = 1/\sigma^2$  and the  $n \times n$  precision matrix  $P = V^{-1}$ .

Separating constant and variable terms, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] . \quad (5)$$

Expanding the product in the exponent, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y^T P y - y^T P X \beta - \beta^T X^T P y + \beta^T X^T P X \beta) \right] . \quad (6)$$

Completing the square over  $\beta$ , finally gives

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} \left( (\beta - \tilde{X}y)^T X^T P X (\beta - \tilde{X}y) - y^T Q y + y^T P y \right) \right] \quad (7)$$

where  $\tilde{X} = (X^T P X)^{-1} X^T P$  and  $Q = \tilde{X}^T (X^T P X) \tilde{X}$ .

In other words, the likelihood function ( $\rightarrow$  I/5.1.2) is proportional to a power of  $\tau$ , times an exponential of  $\tau$  and an exponential of a squared form of  $\beta$ , weighted by  $\tau$ :

$$p(y|\beta, \tau) \propto \tau^{n/2} \cdot \exp \left[ -\frac{\tau}{2} (y^T P y - y^T Q y) \right] \cdot \exp \left[ -\frac{\tau}{2} (\beta - \tilde{X}y)^T X^T P X (\beta - \tilde{X}y) \right] . \quad (8)$$

The same is true for a normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $\beta$  and  $\tau$

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (9)$$

the probability density function of which ( $\rightarrow$  II/4.4.3)

$$p(\beta, \tau) = \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp \left[ -\frac{\tau}{2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \quad (10)$$

exhibits the same proportionality

$$p(\beta, \tau) \propto \tau^{a_0+p/2-1} \cdot \exp[-\tau b_0] \cdot \exp \left[ -\frac{\tau}{2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right] \quad (11)$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.112; URL: <https://www.springer.com/gp/book/9780387310732>.

### 1.6.2 Posterior distribution

**Theorem:** Let

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure  $V$  as well as unknown  $p \times 1$  regression coefficients  $\beta$  and unknown noise variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a normal-gamma distribution ( $\rightarrow$  II/4.4.1)

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (4)$$

**Proof:** According to Bayes' theorem ( $\rightarrow$  I/5.3.2), the posterior distribution ( $\rightarrow$  I/5.1.9) is given by

$$p(\beta, \tau|y) = \frac{p(y|\beta, \tau) p(\beta, \tau)}{p(y)} . \quad (5)$$

Since  $p(y)$  is just a normalization factor, the posterior is proportional ( $\rightarrow$  I/5.1.10) to the numerator:

$$p(\beta, \tau|y) \propto p(y|\beta, \tau) p(\beta, \tau) = p(y, \beta, \tau) . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[ -\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \quad (8)$$

using the noise precision  $\tau = 1/\sigma^2$  and the  $n \times n$  precision matrix ( $\rightarrow$  I/1.16.12)  $P = V^{-1}$ .

Combining the likelihood function ( $\rightarrow$  I/5.1.2) (8) with the prior distribution ( $\rightarrow$  I/5.1.3) (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned}
p(y, \beta, \tau) &= p(y|\beta, \tau) p(\beta, \tau) \\
&= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[ -\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \cdot \\
&\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp \left[ -\frac{\tau}{2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right] \cdot \\
&\quad \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] .
\end{aligned} \tag{9}$$

Collecting identical variables gives:

$$\begin{aligned}
p(y, \beta, \tau) &= \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[ -\frac{\tau}{2} \left( (y - X\beta)^T P (y - X\beta) + (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right) \right] .
\end{aligned} \tag{10}$$

Expanding the products in the exponent gives:

$$\begin{aligned}
p(y, \beta, \tau) &= \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[ -\frac{\tau}{2} \left( y^T P y - y^T P X \beta - \beta^T X^T P y + \beta^T X^T P X \beta + \right. \right. \\
&\quad \left. \left. \beta^T \Lambda_0 \beta - \beta^T \Lambda_0 \mu_0 - \mu_0^T \Lambda_0 \beta + \mu_0^T \Lambda_0 \mu_0 \right) \right] .
\end{aligned} \tag{11}$$

Completing the square over  $\beta$ , we finally have

$$\begin{aligned}
p(y, \beta, \tau) &= \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[ -\frac{\tau}{2} \left( (\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right) \right]
\end{aligned} \tag{12}$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1} (X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 .
\end{aligned} \tag{13}$$

Ergo, the joint likelihood is proportional to

$$p(y, \beta, \tau) \propto \tau^{p/2} \cdot \exp \left[ -\frac{\tau}{2} (\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) \right] \cdot \tau^{a_n-1} \cdot \exp[-b_n \tau] \tag{14}$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned}
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{15}$$

From the term in (14), we can isolate the posterior distribution over  $\beta$  given  $\tau$ :

$$p(\beta|\tau, y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) . \quad (16)$$

From the remaining term, we can isolate the posterior distribution over  $\tau$ :

$$p(\tau|y) = \text{Gam}(\tau; a_n, b_n) . \quad (17)$$

Together, (16) and (17) constitute the joint ( $\rightarrow$  I/1.3.2) posterior distribution ( $\rightarrow$  I/5.1.9) of  $\beta$  and  $\tau$ . ■

#### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.113; URL: <https://www.springer.com/gp/book/9780387310732>.

### 1.6.3 Log model evidence

**Theorem:** Let

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure  $V$  as well as unknown  $p \times 1$  regression coefficients  $\beta$  and unknown noise variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \quad (3)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (4)$$

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the model evidence ( $\rightarrow$  I/5.1.15) for this model is:

$$p(y|m) = \iint p(y|\beta, \tau) p(\beta, \tau) d\beta d\tau . \quad (5)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand is equivalent to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(y|m) = \iint p(y, \beta, \tau) d\beta d\tau . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[ -\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \quad (8)$$

using the noise precision  $\tau = 1/\sigma^2$  and the  $n \times n$  precision matrix  $P = V^{-1}$ .

When deriving the posterior distribution ( $\rightarrow$  III/1.6.2)  $p(\beta, \tau|y)$ , the joint likelihood  $p(y, \beta, \tau)$  is obtained as

$$p(y, \beta, \tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[ -\frac{\tau}{2} ((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)) \right] . \quad (9)$$

Using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), we can rewrite this as

$$p(y, \beta, \tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \sqrt{\frac{(2\pi)^p}{\tau^p |\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \exp \left[ -\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] . \quad (10)$$

Now,  $\beta$  can be integrated out easily:

$$\int p(y, \beta, \tau) d\beta = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[ -\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] . \quad (11)$$

Using the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7), we can rewrite this as

$$\int p(y, \beta, \tau) d\beta = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \text{Gam}(\tau; a_n, b_n) . \quad (12)$$

Finally,  $\tau$  can also be integrated out:

$$\iint p(y, \beta, \tau) d\beta d\tau = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} = p(y|m) . \quad (13)$$

Thus, the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is given by

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \end{aligned} \quad (14)$$

■

#### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: <https://www.springer.com/gp/book/9780387310732>.

#### 1.6.4 Accuracy and complexity

**Theorem:** Let

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure  $V$  as well as unknown  $p \times 1$  regression coefficients  $\beta$  and unknown noise variance  $\sigma^2$ . Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, accuracy and complexity ( $\rightarrow$  IV/3.1.6) of this model are

$$\begin{aligned} \text{Acc}(m) = & -\frac{1}{2} \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) - \frac{1}{2} \text{tr}(X^T P X \Lambda_n^{-1}) \\ & + \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{n}{2} (\psi(a_n) - \log(b_n)) \\ \text{Com}(m) = & \frac{1}{2} \frac{a_n}{b_n} [(\mu_0 - \mu_n)^T \Lambda_0 (\mu_0 - \mu_n) - 2(b_n - b_0)] + \frac{1}{2} \text{tr}(\Lambda_0 \Lambda_n^{-1}) - \frac{1}{2} \log \frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ & + a_0 \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \psi(a_n) . \end{aligned} \quad (3)$$

where  $\mu_n$ ,  $\Lambda_n$ ,  $a_n$  and  $b_n$  are the posterior hyperparameters for Bayesian linear regression ( $\rightarrow$  III/1.6.2) and  $P$  is the data precision matrix ( $\rightarrow$  I/1.16.12):  $P = V^{-1}$ .

**Proof:** Model accuracy and complexity are defined as ( $\rightarrow$  IV/3.1.6)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\beta, \tau, m) \rangle_{p(\beta, \tau|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\beta, \tau|y, m) || p(\beta, \tau|m)] . \end{aligned} \quad (4)$$

1) The accuracy term is the expectation ( $\rightarrow$  I/1.12.1) of the log-likelihood function ( $\rightarrow$  I/4.1.2)  $\log p(y|\beta, \tau)$  with respect to the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\beta, \tau|y)$ . This expectation can be rewritten as:

$$\begin{aligned} \text{Acc}(m) &= \iint p(\beta, \tau|y) \log p(y|\beta, \tau) d\beta d\tau \\ &= \int p(\tau|y) \int p(\beta|\tau, y) \log p(y|\beta, \tau) d\beta d\tau \\ &= \left\langle \left\langle \log p(y|\beta, \tau) \right\rangle_{p(\beta|\tau, y)} \right\rangle_{p(\tau|y)}. \end{aligned} \quad (5)$$

With the log-likelihood function for multiple linear regression ( $\rightarrow$  III/1.5.23), we have:

$$\begin{aligned} \text{Acc}(m) &= \left\langle \left\langle \log \left( \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right] \right) \right\rangle_{p(\beta|\tau, y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \log \left( \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \cdot \exp \left[ -\frac{1}{2} (y - X\beta)^T (\tau P) (y - X\beta) \right] \right) \right\rangle_{p(\beta|\tau, y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{1}{2} (y - X\beta)^T (\tau P) (y - X\beta) \right\rangle_{p(\beta|\tau, y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} [y^T P y - 2y^T P X \beta + \beta^T X^T P X \beta] \right\rangle_{p(\beta|\tau, y)} \right\rangle_{p(\tau|y)}. \end{aligned} \quad (6)$$

With the posterior distribution for Bayesian linear regression ( $\rightarrow$  III/1.6.2), this becomes:

$$\text{Acc}(m) = \left\langle \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} [y^T P y - 2y^T P X \beta + \beta^T X^T P X \beta] \right\rangle_{\mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1})} \right\rangle_{\text{Gam}(\tau; a_n, b_n)} \quad (7)$$

If  $x \sim \mathcal{N}(\mu, \Sigma)$ , then its expected value is ( $\rightarrow$  II/4.1.6)

$$\langle x \rangle = \mu \quad (8)$$

and the expectation of a quadratic form is given by ( $\rightarrow$  I/1.12.17)

$$\langle x^T A x \rangle = \mu^T A \mu + \text{tr}(A \Sigma). \quad (9)$$

Thus, the model accuracy of  $m$  evaluates to:

$$\begin{aligned} \text{Acc}(m) &= \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \right. \\ &\quad \left. \frac{\tau}{2} \left[ y^T P y - 2y^T P X \mu_n + \mu_n^T X^T P X \mu_n + \frac{1}{\tau} \text{tr}(X^T P X \Lambda_n^{-1}) \right] \right\rangle_{\text{Gam}(\tau; a_n, b_n)} \\ &= \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y - X \mu_n)^T P (y - X \mu_n) - \frac{1}{2} \text{tr}(X^T P X \Lambda_n^{-1}) \right\rangle_{\text{Gam}(\tau; a_n, b_n)}. \end{aligned} \quad (10)$$



If  $x \sim \text{Gam}(a, b)$ , then its expected value is ( $\rightarrow$  II/3.4.11)

$$\langle x \rangle = \frac{a}{b} \quad (11)$$

and its logarithmic expectation is given by ( $\rightarrow$  II/3.4.15)

$$\langle \log x \rangle = \psi(a) - \log(b) . \quad (12)$$

Thus, the model accuracy of  $m$  evaluates to

$$\begin{aligned} \text{Acc}(m) = & -\frac{1}{2} \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) - \frac{1}{2} \text{tr}(X^T P X \Lambda_n^{-1}) \\ & + \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{n}{2} (\psi(a_n) - \log(b_n)) \end{aligned} \quad (13)$$

which proofs the first part of (3).

2) The complexity penalty is the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\beta, \tau|y)$  from the prior distribution ( $\rightarrow$  I/5.1.3)  $p(\beta, \tau)$ . This can be rewritten as follows:

$$\begin{aligned} \text{Com}(m) &= \iint p(\beta, \tau|y) \log \frac{p(\beta, \tau|y)}{p(\beta, \tau)} d\beta d\tau \\ &= \iint p(\beta|\tau, y) p(\tau|y) \log \left[ \frac{p(\beta|\tau, y)}{p(\beta|\tau)} \frac{p(\tau|y)}{p(\tau)} \right] d\beta d\tau \\ &= \int p(\tau|y) \int p(\beta|\tau, y) \log \frac{p(\beta|\tau, y)}{p(\beta|\tau)} d\beta d\tau + \int p(\tau|y) \log \frac{p(\tau|y)}{p(\tau)} \int p(\beta|\tau, y) d\beta d\tau \\ &= \langle \text{KL} [p(\beta|\tau, y) || p(\beta|\tau)] \rangle_{p(\tau|y)} + \text{KL} [p(\tau|y) || p(\tau)] . \end{aligned} \quad (14)$$

With the prior distribution ( $\rightarrow$  III/1.6.1) given by (2) and the posterior distribution for Bayesian linear regression ( $\rightarrow$  III/1.6.2), this becomes:

$$\begin{aligned} \text{Com}(m) &= \langle \text{KL} [\mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) || \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1})] \rangle_{\text{Gam}(\tau; a_n, b_n)} \\ &\quad + \text{KL} [\text{Gam}(\tau; a_n, b_n) || \text{Gam}(\tau; a_0, b_0)] . \end{aligned} \quad (15)$$

With the Kullback-Leibler divergence for the multivariate normal distribution ( $\rightarrow$  II/4.1.12)

$$\text{KL}[\mathcal{N}(\mu_1, \Sigma_1) || \mathcal{N}(\mu_2, \Sigma_2)] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (16)$$

and the Kullback-Leibler divergence for the gamma distribution ( $\rightarrow$  II/3.4.18)

$$\text{KL}[\text{Gam}(a_1, b_1) || \text{Gam}(a_2, b_2)] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} , \quad (17)$$

the model complexity of  $m$  evaluates to:

$$\begin{aligned} \text{Com}(m) &= \left\langle \frac{1}{2} \left[ (\mu_0 - \mu_n)^T (\tau\Lambda_0) (\mu_0 - \mu_n) + \text{tr}((\tau\Lambda_0)(\tau\Lambda_n)^{-1}) - \log \frac{|(\tau\Lambda_n)^{-1}|}{|(\tau\Lambda_0)^{-1}|} - p \right] \right\rangle_{p(\tau|y)} \\ &\quad + a_0 \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \psi(a_n) - (b_n - b_0) \frac{a_n}{b_n} . \end{aligned} \quad (18)$$

Using  $x \sim \text{Gam}(a, b) \Rightarrow \langle x \rangle = a/b$  from (11) again, it follows that

$$\begin{aligned} \text{Com}(m) = & \frac{1}{2} \frac{a_n}{b_n} [(\mu_0 - \mu_n)^T \Lambda_0 (\mu_0 - \mu_n)] + \frac{1}{2} \text{tr}(\Lambda_0 \Lambda_n^{-1}) - \frac{1}{2} \log \frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ & + a_0 \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \psi(a_n) - (b_n - b_0) \frac{a_n}{b_n}. \end{aligned} \quad (19)$$

Thus, the model complexity of  $m$  evaluates to

$$\begin{aligned} \text{Com}(m) = & \frac{1}{2} \frac{a_n}{b_n} [(\mu_0 - \mu_n)^T \Lambda_0 (\mu_0 - \mu_n) - 2(b_n - b_0)] + \frac{1}{2} \text{tr}(\Lambda_0 \Lambda_n^{-1}) - \frac{1}{2} \log \frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ & + a_0 \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \psi(a_n) \end{aligned} \quad (20)$$

which proofs the second part of (3).

3) A control calculation confirms that

$$\text{Acc}(m) - \text{Com}(m) = \text{LME}(m) \quad (21)$$

where  $\text{LME}(m)$  is the log model evidence for Bayesian linear regression ( $\rightarrow$  III/1.6.3):

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \quad (22)$$

This requires to recognize that

$$-\frac{1}{2} \text{tr}(X^T P X \Lambda_n^{-1}) - \frac{1}{2} \text{tr}(\Lambda_0 \Lambda_n^{-1}) + \frac{p}{2} = 0 \quad (23)$$

and

$$\frac{n}{2} (\psi(a_n) - \log(b_n)) - a_0 \log \frac{b_n}{b_0} - (a_n - a_0) \psi(a_n) = a_0 \log b_0 - a_n \log b_n \quad (24)$$

thanks to the nature of the posterior hyperparameters for Bayesian linear regression ( $\rightarrow$  III/1.6.2). ■

#### Sources:

- Soch J, Allefeld A (2016): “Kullback-Leibler Divergence for the Normal-Gamma Distribution”; in: *arXiv math.ST*, 1611.01437, eqs. 23/30; URL: <https://arxiv.org/abs/1611.01437>.
- Soch J, Allefeld C, Haynes JD (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469-489, Appendix C; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage
- Soch J (2018): “cvBMS and cvBMA: filling in the gaps”; in: *arXiv stat.ME*, sect. 2.2, eqs. 8-24; URL: <https://arxiv.org/abs/1807.01585>; DOI: 10.48550/arXiv.1807.01585.

### 1.6.5 Deviance information criterion

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1)  $m$

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \sigma^2 V = (\tau P)^{-1} \quad (1)$$

with a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the deviance information criterion ( $\rightarrow$  IV/2.3.1) for this model is

$$\begin{aligned} \text{DIC}(m) = & n \cdot \log(2\pi) - n [2\psi(a_n) - \log(a_n) - \log(b_n)] - \log |P| \\ & + \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) \end{aligned} \quad (3)$$

where  $\mu_n$  and  $\Lambda_n$  as well as  $a_n$  and  $b_n$  are posterior parameters ( $\rightarrow$  I/5.1.9) describing the posterior distribution in Bayesian linear regression ( $\rightarrow$  III/1.6.2).

**Proof:** The deviance for multiple linear regression ( $\rightarrow$  III/1.5.32) is

$$D(\beta, \sigma^2) = n \cdot \log(2\pi) + n \cdot \log(\sigma^2) + \log |V| + \frac{1}{\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \quad (4)$$

which, applying the equalities  $\tau = 1/\sigma^2$  and  $P = V^{-1}$ , becomes

$$D(\beta, \tau) = n \cdot \log(2\pi) - n \cdot \log(\tau) - \log |P| + \tau \cdot (y - X\beta)^T P (y - X\beta) . \quad (5)$$

The deviance information criterion ( $\rightarrow$  IV/2.3.1) (DIC) is defined as

$$\text{DIC}(m) = -2 \log p(y | \langle \beta \rangle, \langle \tau \rangle, m) + 2 p_D \quad (6)$$

where  $\log p(y | \langle \beta \rangle, \langle \tau \rangle, m)$  is the log-likelihood function ( $\rightarrow$  III/1.5.24) at the posterior expectations ( $\rightarrow$  I/1.12.1) and the “effective number of parameters”  $p_D$  is the difference between the expectation of the deviance and the deviance at the expectation ( $\rightarrow$  IV/2.3.1):

$$p_D = \langle D(\beta, \tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle) . \quad (7)$$

With that, the DIC for multiple linear regression becomes:

$$\begin{aligned} \text{DIC}(m) &= -2 \log p(y | \langle \beta \rangle, \langle \tau \rangle, m) + 2 p_D \\ &= D(\langle \beta \rangle, \langle \tau \rangle) + 2 [\langle D(\beta, \tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle)] \\ &= 2 \langle D(\beta, \tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle) . \end{aligned} \quad (8)$$

The posterior distribution for multiple linear regression ( $\rightarrow$  III/1.6.2) is

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (9)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{10}$$

Thus, we have the following posterior expectations:

$$\langle \beta \rangle_{\beta, \tau | y} = \mu_n \tag{11}$$

$$\langle \tau \rangle_{\beta, \tau | y} = \frac{a_n}{b_n} \tag{12}$$

$$\langle \log \tau \rangle_{\beta, \tau | y} = \psi(a_n) - \log(b_n) \tag{13}$$

$$\begin{aligned}
\langle \beta^T A \beta \rangle_{\beta | \tau, y} &= \mu_n^T A \mu_n + \text{tr} (A(\tau \Lambda_n)^{-1}) \\
&= \mu_n^T A \mu_n + \frac{1}{\tau} \text{tr} (A \Lambda_n^{-1}) .
\end{aligned} \tag{14}$$

In these identities, we have used the mean of the multivariate normal distribution ( $\rightarrow$  II/4.1.6), the mean of the gamma distribution ( $\rightarrow$  II/3.4.11), the logarithmic expectation of the gamma distribution ( $\rightarrow$  II/3.4.15), the expectation of a quadratic form ( $\rightarrow$  I/1.12.17) and the covariance of the multivariate normal distribution ( $\rightarrow$  II/4.1.8).

With that, the deviance at the expectation is:

$$\begin{aligned}
D(\langle \beta \rangle, \langle \tau \rangle) &\stackrel{(5)}{=} n \cdot \log(2\pi) - n \cdot \log(\langle \tau \rangle) - \log |P| + \tau \cdot (y - X \langle \beta \rangle)^T P (y - X \langle \beta \rangle) \\
&\stackrel{(11)}{=} n \cdot \log(2\pi) - n \cdot \log(\langle \tau \rangle) - \log |P| + \tau \cdot (y - X \mu_n)^T P (y - X \mu_n) \\
&\stackrel{(12)}{=} n \cdot \log(2\pi) - n \cdot \log \left( \frac{a_n}{b_n} \right) - \log |P| + \frac{a_n}{b_n} \cdot (y - X \mu_n)^T P (y - X \mu_n) .
\end{aligned} \tag{15}$$

Moreover, the expectation of the deviance is:

$$\begin{aligned}
\langle D(\beta, \tau) \rangle &\stackrel{(5)}{=} \langle n \cdot \log(2\pi) - n \cdot \log(\tau) - \log |P| + \tau \cdot (y - X\beta)^T P (y - X\beta) \rangle \\
&= n \cdot \log(2\pi) - n \cdot \langle \log(\tau) \rangle - \log |P| + \langle \tau \cdot (y - X\beta)^T P (y - X\beta) \rangle \\
&\stackrel{(13)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \left\langle \tau \cdot \langle (y - X\beta)^T P (y - X\beta) \rangle_{\beta|\tau, y} \right\rangle_{\tau|y} \\
&= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \left\langle \tau \cdot \langle y^T P y - y^T P X \mu_n - \mu_n^T X^T P y + \mu_n^T X^T P X \mu_n \rangle_{\beta|\tau, y} \right\rangle_{\tau|y} \\
&\stackrel{(14)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \left\langle \tau \cdot \left[ y^T P y - y^T P X \mu_n - \mu_n^T X^T P y + \mu_n^T X^T P X \mu_n + \frac{1}{\tau} \text{tr} (X^T P X \Lambda_n^{-1}) \right] \right\rangle_{\tau|y} \\
&= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \langle \tau \cdot (y - X\mu_n)^T P (y - X\mu_n) \rangle_{\tau|y} + \text{tr} (X^T P X \Lambda_n^{-1}) \\
&\stackrel{(12)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \frac{a_n}{b_n} \cdot (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) .
\end{aligned} \tag{16}$$

Finally, combining the two terms, we have:

$$\begin{aligned}
\text{DIC}(m) &\stackrel{(8)}{=} 2 \langle D(\beta, \tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle) \\
&\stackrel{(16)}{=} 2 [n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\
&\quad + \frac{a_n}{b_n} \cdot (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1})] \\
&\stackrel{(15)}{=} \left[ n \cdot \log(2\pi) - n \cdot \log \left( \frac{a_n}{b_n} \right) - \log |P| + \frac{a_n}{b_n} \cdot (y - X\mu_n)^T P (y - X\mu_n) \right] \\
&= n \cdot \log(2\pi) - 2n\psi(a_n) + 2n \log(b_n) + n \log(a_n) - \log(b_n) - \log |P| \\
&\quad + \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) \\
&= n \cdot \log(2\pi) - n [2\psi(a_n) - \log(a_n) - \log(b_n)] - \log |P| \\
&\quad + \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) .
\end{aligned} \tag{17}$$

This conforms to equation (3). ■

### 1.6.6 Maximum-a-posteriori estimation

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad \sigma^2 V = (\tau P)^{-1} \tag{1}$$

and assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the maximum-a-posteriori estimates ( $\rightarrow$  I/5.1.14) of  $\beta$  and  $\tau$  are

$$\begin{aligned} \hat{\beta}_{\text{MAP}} &= (X^T P X + \Lambda_0)^{-1} (X^T P y + \Lambda_0 \mu_0) \\ \hat{\tau}_{\text{MAP}} &= (2a_0 + n - 2) \left( 2b_0 + (y - X \hat{\beta}_{\text{MAP}})^T P (y - X \hat{\beta}_{\text{MAP}}) + (\hat{\beta}_{\text{MAP}} - \mu_0)^T \Lambda_0 (\hat{\beta}_{\text{MAP}} - \mu_0) \right)^{-1} \end{aligned} \quad (3)$$

where  $n$  is the number of data points ( $\rightarrow$  III/1.5.1).

**Proof:** Given the generative model ( $\rightarrow$  I/5.1.1) in (1) and the prior distribution ( $\rightarrow$  I/5.1.3) in (2), the posterior distribution ( $\rightarrow$  I/5.1.9) for multiple linear regression ( $\rightarrow$  III/1.5.1) is also a normal-gamma distribution ( $\rightarrow$  III/1.6.2)

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (4)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are equal to

$$\begin{aligned} \mu_n &= \Lambda_n^{-1} (X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (5)$$

From this, the conditional posterior distribution over  $\beta$  follows as ( $\rightarrow$  II/4.4.1)

$$p(\beta | \tau, y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \quad (6)$$

and the marginal posterior distribution over  $\tau$  follows as ( $\rightarrow$  II/4.4.1)

$$p(\tau | y) = \text{Gam}(\tau; a_n, b_n) . \quad (7)$$

The mode of the multivariate normal distribution ( $\rightarrow$  II/4.1.7) is given by

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \text{mode}(X) = \mu \quad (8)$$

and the mode of the gamma distribution ( $\rightarrow$  II/3.4.13) is given by

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad \text{mode}(X) = \frac{a-1}{b} . \quad (9)$$

Applying (8) to (6), the maximum-a-posteriori estimate ( $\rightarrow$  I/5.1.14) of  $\beta$  follows as

$$\begin{aligned} \hat{\beta}_{\text{MAP}} &= \mu_n \\ &= \Lambda_n^{-1} (X^T P y + \Lambda_0 \mu_0) \\ &= (X^T P X + \Lambda_0)^{-1} (X^T P y + \Lambda_0 \mu_0) \end{aligned} \quad (10)$$

and applying (9) to (7), the maximum-a-posteriori estimate ( $\rightarrow$  I/5.1.14) of  $\tau$  follows as

$$\begin{aligned}
 \hat{\tau}_{\text{MAP}} &= \frac{a_n - 1}{b_n} \\
 &= \left( a_0 + \frac{n}{2} - 1 \right) \left( b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right)^{-1} \\
 &= (2a_0 + n - 2) (2b_0 + y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)^{-1} \\
 &= (2a_0 + n - 2) \left( 2b_0 + y^T P y + \mu_0^T \Lambda_0 \mu_0 - \hat{\beta}_{\text{MAP}}^T (X^T P X + \Lambda_0) \hat{\beta}_{\text{MAP}} \right)^{-1} \\
 &= (2a_0 + n - 2) \left( 2b_0 + y^T P y + \mu_0^T \Lambda_0 \mu_0 - \hat{\beta}_{\text{MAP}}^T X^T P X \hat{\beta}_{\text{MAP}} - \hat{\beta}_{\text{MAP}}^T \Lambda_0 \hat{\beta}_{\text{MAP}} \right)^{-1} \\
 &= (2a_0 + n - 2) \left( 2b_0 + (y - X \hat{\beta}_{\text{MAP}})^T P (y - X \hat{\beta}_{\text{MAP}}) + (\hat{\beta}_{\text{MAP}} - \mu_0)^T \Lambda_0 (\hat{\beta}_{\text{MAP}} - \mu_0) \right)^{-1}.
 \end{aligned} \tag{11}$$

■

### 1.6.7 Expression of posterior parameters using error terms

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \quad \sigma^2 V = (\tau P)^{-1}, \tag{1}$$

assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \tag{2}$$

and consider the Bayesian posterior distribution ( $\rightarrow$  III/1.6.2) over these model parameters:

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n). \tag{3}$$

Then, the posterior hyperparameters ( $\rightarrow$  I/5.1.9) for the noise precision ( $\rightarrow$  III/1.6.1)  $\tau$  can be expressed as

$$\begin{aligned}
 a_n &= a_0 + \frac{n}{2} \\
 b_n &= b_0 + \frac{1}{2} (\varepsilon_y^T P \varepsilon_y + \varepsilon_\beta^T \Lambda_0 \varepsilon_\beta)
 \end{aligned} \tag{4}$$

where  $\varepsilon_y$  and  $\varepsilon_\beta$  are the “prediction errors” and “parameter errors”

$$\begin{aligned}
 \varepsilon_y &= y - \hat{y} \\
 \varepsilon_\beta &= \mu_n - \mu_0
 \end{aligned} \tag{5}$$

where  $\hat{y}$  is the predicted signal ( $\rightarrow$  III/1.5.12) at the posterior mean ( $\rightarrow$  III/1.6.2) regression coefficients ( $\rightarrow$  III/1.5.1)  $\mu_n$ :

$$\hat{y} = X \mu_n. \tag{6}$$

**Proof:** The posterior hyperparameter for Bayesian linear regression ( $\rightarrow$  III/1.6.2) are:

$$\begin{aligned}\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .\end{aligned}\tag{7}$$

The shape parameter ( $\rightarrow$  II/3.4.1)  $a_n$  is given by this equation. The rate parameter ( $\rightarrow$  II/3.4.1)  $b_n$  of the posterior distribution ( $\rightarrow$  I/5.1.9) can be developed as follows:

$$\begin{aligned}b_n &\stackrel{(7)}{=} b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \\ &\stackrel{(7)}{=} b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T (X^T P X + \Lambda_0) \mu_n) \\ &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T X^T P X \mu_n - \mu_n^T \Lambda_0 \mu_n) \\ &= b_0 + \frac{1}{2}((y^T P y - \mu_n^T X^T P X \mu_n) + (\mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_0 \mu_n)) \\ &= b_0 + \frac{1}{2}((y - X \mu_n)^T P (y - X \mu_n) + (\mu_0 - \mu_n)^T \Lambda_0 (\mu_0 - \mu_n)) \\ &\stackrel{(6)}{=} b_0 + \frac{1}{2}((y - \hat{y})^T P (y - \hat{y}) + (\mu_n - \mu_0)^T \Lambda_0 (\mu_n - \mu_0)) \\ &\stackrel{(5)}{=} b_0 + \frac{1}{2}(\varepsilon_y^T P \varepsilon_y + \varepsilon_\beta^T \Lambda_0 \varepsilon_\beta) .\end{aligned}\tag{8}$$

Together with equation (??c), this completes the proof. ■

### 1.6.8 Posterior probability of alternative hypothesis

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with normally distributed ( $\rightarrow$  II/4.1.1) errors:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V)\tag{1}$$

and assume a normal-gamma ( $\rightarrow$  II/4.4.1) prior distribution ( $\rightarrow$  I/5.1.3) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) .\tag{2}$$

Then, the posterior ( $\rightarrow$  I/5.1.9) probability ( $\rightarrow$  I/1.3.1) of the alternative hypothesis ( $\rightarrow$  I/4.3.3)

$$H_1 : c^T \beta > 0\tag{3}$$

is given by

$$\Pr(H_1|y) = 1 - T\left(-\frac{c^T \mu}{\sqrt{c^T \Sigma c}}; \nu\right)\tag{4}$$



where  $c$  is a  $p \times 1$  contrast vector ( $\rightarrow$  III/1.5.26),  $T(x; \nu)$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the t-distribution ( $\rightarrow$  II/3.3.1) with  $\nu$  degrees of freedom ( $\rightarrow$  I/1.5.12) and  $\mu$ ,  $\Sigma$  and  $\nu$  can be obtained from the posterior hyperparameters ( $\rightarrow$  I/5.1.9) of Bayesian linear regression.

**Proof:** The posterior distribution for Bayesian linear regression ( $\rightarrow$  III/1.6.2) is given by a normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $\beta$  and  $\tau = 1/\sigma^2$

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (5)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (6)$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution ( $\rightarrow$  II/4.4.8), such that the marginal ( $\rightarrow$  I/1.5.3) posterior ( $\rightarrow$  I/5.1.9) distribution of  $\beta$  is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \quad (7)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left( \frac{a_n}{b_n} \Lambda_n \right)^{-1} \\ \nu &= 2 a_n . \end{aligned} \quad (8)$$

Define the quantity  $\gamma = c^T \beta$ . According to the linear transformation theorem for the multivariate t-distribution,  $\gamma$  also follows a multivariate t-distribution ( $\rightarrow$  II/4.3.1):

$$p(\gamma|y) = t(\gamma; c^T \mu, c^T \Sigma c, \nu) . \quad (9)$$

Because  $c^T$  is a  $1 \times p$  vector,  $\gamma$  is a scalar and actually has a non-standardized t-distribution ( $\rightarrow$  II/3.3.3). Therefore, the posterior probability of  $H_1$  can be calculated using a one-dimensional integral:

$$\begin{aligned} \Pr(H_1|y) &= p(\gamma > 0|y) \\ &= \int_0^{+\infty} p(\gamma|y) d\gamma \\ &= 1 - \int_{-\infty}^0 p(\gamma|y) d\gamma \\ &= 1 - T_{\text{nst}}(0; c^T \mu, c^T \Sigma c, \nu) . \end{aligned} \quad (10)$$

Using the relation between non-standardized t-distribution and standard t-distribution ( $\rightarrow$  II/3.3.4), we can finally write:

$$\begin{aligned}
\Pr(H_1|y) &= 1 - T\left(\frac{(0 - c^T\mu)}{\sqrt{c^T\Sigma c}}; \nu\right) \\
&= 1 - T\left(-\frac{c^T\mu}{\sqrt{c^T\Sigma c}}; \nu\right).
\end{aligned} \tag{11}$$

■

**Sources:**

- Koch, Karl-Rudolf (2007): “Multivariate t-distribution”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.188; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

**1.6.9 Posterior credibility region excluding null hypothesis**

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with normally distributed ( $\rightarrow$  II/4.1.1) errors:

$$y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and assume a normal-gamma ( $\rightarrow$  II/4.4.1) prior distribution ( $\rightarrow$  I/5.1.3) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0). \tag{2}$$

Then, the largest posterior ( $\rightarrow$  I/5.1.9) credibility region that does not contain the omnibus null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : C^T\beta = 0 \tag{3}$$

is given by the credibility level

$$(1 - \alpha) = F\left([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu\right) \tag{4}$$

where  $C$  is a  $p \times q$  contrast matrix ( $\rightarrow$  III/1.5.27),  $F(x; v, w)$  is the cumulative distribution function ( $\rightarrow$  I/1.8.1) of the F-distribution ( $\rightarrow$  II/3.8.1) with  $v$  numerator degrees of freedom ( $\rightarrow$  I/1.5.12),  $w$  denominator degrees of freedom ( $\rightarrow$  I/1.5.12) and  $\mu$ ,  $\Sigma$  and  $\nu$  can be obtained from the posterior hyperparameters ( $\rightarrow$  I/5.1.9) of Bayesian linear regression.

**Proof:** The posterior distribution for Bayesian linear regression ( $\rightarrow$  III/1.6.2) is given by a normal-gamma distribution ( $\rightarrow$  II/4.4.1) over  $\beta$  and  $\tau = 1/\sigma^2$

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \tag{5}$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n).
\end{aligned} \tag{6}$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution ( $\rightarrow$  II/4.4.8), such that the marginal ( $\rightarrow$  I/1.5.3) posterior ( $\rightarrow$  I/5.1.9) distribution of  $\beta$  is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \quad (7)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left( \frac{a_n}{b_n} \Lambda_n \right)^{-1} \\ \nu &= 2 a_n . \end{aligned} \quad (8)$$

Define the quantity  $\gamma = C^T \beta$ . According to the linear transformation theorem for the multivariate t-distribution,  $\gamma$  also follows a multivariate t-distribution ( $\rightarrow$  II/4.3.1):

$$p(\gamma|y) = t(\gamma; C^T \mu, C^T \Sigma C, \nu) . \quad (9)$$

Because  $C^T$  is a  $q \times p$  matrix,  $\gamma$  is a  $q \times 1$  vector. The quadratic form of a multivariate t-distributed random variable has an F-distribution ( $\rightarrow$  II/4.3.3), such that we can write:

$$QF(\gamma) = (\gamma - C^T \mu)^T (C^T \Sigma C)^{-1} (\gamma - C^T \mu) / q \sim F(q, \nu) . \quad (10)$$

Therefore, the largest posterior credibility region for  $\gamma$  which does not contain  $\gamma = 0_q$  (i.e. only touches this origin point) can be obtained by plugging  $QF(0)$  into the cumulative distribution function of the F-distribution:

$$\begin{aligned} (1 - \alpha) &= F(QF(0); q, \nu) \\ &= F([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu) . \end{aligned} \quad (11)$$

■

#### Sources:

- Koch, Karl-Rudolf (2007): “Multivariate t-distribution”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.211, 2.183; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

#### 1.6.10 Combined posterior distribution from independent data sets

**Theorem:** Let  $y = \{y_1, \dots, y_S\}$  be a set of  $S$  conditionally independent data sets ( $\rightarrow$  I/1.3.7) assumed to follow linear regression models ( $\rightarrow$  III/1.5.1) with design matrices ( $\rightarrow$  III/1.5.1)  $X_1, \dots, X_S$ , number of data points ( $\rightarrow$  III/1.5.1)  $n_1, \dots, n_S$  and precision matrices ( $\rightarrow$  III/1.6.1)  $P_1, \dots, P_n$ , governed by identical regression coefficients ( $\rightarrow$  III/1.5.1)  $\beta$  and identical noise precision ( $\rightarrow$  III/1.6.1)  $\tau$ :

$$\begin{aligned} y_1 &= X_1 \beta + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0, \sigma^2 V_1), \quad \sigma^2 V_1 = (\tau P_1)^{-1} \\ &\vdots \\ y_S &= X_S \beta + \varepsilon_S, \quad \varepsilon_S \sim \mathcal{N}(0, \sigma^2 V_S), \quad \sigma^2 V_S = (\tau P_S)^{-1} . \end{aligned} \quad (1)$$

Moreover, assume a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta$  and  $\tau = 1/\sigma^2$ :

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) . \quad (2)$$

Then, the combined posterior distribution ( $\rightarrow$  I/5.1.11) from observing these conditionally independent data sets ( $\rightarrow$  I/1.3.7) is also given by a normal-gamma distribution ( $\rightarrow$  II/4.4.1)

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (3)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1} \left( \sum_{i=1}^S X_i^T P_i y_i + \Lambda_0 \mu_0 \right) \\ \Lambda_n &= \sum_{i=1}^S X_i^T P_i X_i + \Lambda_0 \\ a_n &= a_0 + \frac{1}{2} \sum_{i=1}^S n_i \\ b_n &= b_0 + \frac{1}{2} \left( \sum_{i=1}^S y_i^T P_i y_i + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n \right) . \end{aligned} \quad (4)$$

**Proof:** This can be seen by sequentially applying Bayes' theorem ( $\rightarrow$  I/5.3.2) for calculating the posterior distribution ( $\rightarrow$  I/5.1.10), while using the posterior after one iteration as the prior for the next iteration.

Let  $\mu_0^{(i)}, \Lambda_0^{(i)}, a_0^{(i)}, b_0^{(i)}$  denote the prior hyperparameters ( $\rightarrow$  I/5.1.3) before analyzing the  $i$ -th data set, such that e.g.  $\mu_0^{(1)}$  is identical to  $\mu_0$  in (2):

$$\begin{aligned} \mu_0^{(1)} &= \mu_0 \\ \Lambda_0^{(1)} &= \Lambda_0 \\ a_0^{(1)} &= a_0 \\ b_0^{(1)} &= b_0 . \end{aligned} \quad (5)$$

Moreover, let  $\mu_n^{(i)}, \Lambda_n^{(i)}, a_n^{(i)}, b_n^{(i)}$  denote the posterior hyperparameters ( $\rightarrow$  I/5.1.9) after analyzing the  $i$ -th data set, such that e.g.  $\mu_n^{(S)}$  is identical to  $\mu_n$  in (3):

$$\begin{aligned} \mu_n^{(S)} &= \mu_n \\ \Lambda_n^{(S)} &= \Lambda_n \\ a_n^{(S)} &= a_n \\ b_n^{(S)} &= b_n . \end{aligned} \quad (6)$$

The posterior ( $\rightarrow$  I/5.1.9) after seeing the  $i$ -th data set is equal to the prior ( $\rightarrow$  I/5.1.3) before seeing the  $(i+1)$ -th data set, so we have the relation:

$$\begin{aligned}
\mu_0^{(i+1)} &= \mu_n^{(i)} \\
\Lambda_0^{(i+1)} &= \Lambda_n^{(i)} \\
a_0^{(i+1)} &= a_n^{(i)} \\
b_0^{(i+1)} &= b_n^{(i)} .
\end{aligned} \tag{7}$$

The posterior distribution for Bayesian linear regression when observing a single data set is given by the following hyperparameter equations ( $\rightarrow$  III/1.6.2):

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{8}$$

We can apply (8) to calculate the posterior hyperparameters after seeing the first data set:

$$\begin{aligned}
\mu_n^{(1)} &= \Lambda_n^{(1)-1} \left( X_1^T P_1 y_1 + \Lambda_0^{(1)} \mu_0^{(1)} \right) \\
&= \Lambda_n^{(1)-1} \left( X_1^T P_1 y_1 + \Lambda_0 \mu_0 \right) \\
\Lambda_n^{(1)} &= X_1^T P_1 X_1 + \Lambda_0^{(1)} \\
&= X_1^T P_1 X_1 + \Lambda_0 \\
a_n^{(1)} &= a_0^{(1)} + \frac{1}{2} n_1 \\
&= a_0 + \frac{1}{2} n_1 \\
b_n^{(1)} &= b_0^{(1)} + \frac{1}{2} \left( y_1^T P_1 y_1 + \mu_0^{(1)T} \Lambda_0^{(1)} \mu_0^{(1)} - \mu_n^{(1)T} \Lambda_n^{(1)} \mu_n^{(1)} \right) \\
&= b_0 + \frac{1}{2} \left( y_1^T P_1 y_1 + \mu_0^T \Lambda_0 \mu_0 - \mu_n^{(1)T} \Lambda_n^{(1)} \mu_n^{(1)} \right) .
\end{aligned} \tag{9}$$

These are the prior hyperparameters before seeing the second data set:

$$\begin{aligned}
\mu_0^{(2)} &= \mu_n^{(1)} \\
\Lambda_0^{(2)} &= \Lambda_n^{(1)} \\
a_0^{(2)} &= a_n^{(1)} \\
b_0^{(2)} &= b_n^{(1)} .
\end{aligned} \tag{10}$$

Thus, we can again use (8) to calculate the posterior hyperparameters after seeing the second data set:

$$\begin{aligned}
\mu_n^{(2)} &= \Lambda_n^{(2)-1} \left( X_2^T P_2 y_2 + \Lambda_0^{(2)} \mu_0^{(2)} \right) \\
&= \Lambda_n^{(2)-1} \left( X_2^T P_2 y_2 + \Lambda_n^{(1)} \Lambda_n^{(1)-1} (X_1^T P_1 y_1 + \Lambda_0 \mu_0) \right) \\
&= \Lambda_n^{(2)-1} (X_1^T P_1 y_1 + X_2^T P_2 y_2 + \Lambda_0 \mu_0) \\
\Lambda_n^{(2)} &= X_2^T P_2 X_2 + \Lambda_0^{(2)} \\
&= X_2^T P_2 X_2 + X_1^T P_1 X_1 + \Lambda_0 \\
&= X_1^T P_1 X_1 + X_2^T P_2 X_2 + \Lambda_0 \\
a_n^{(2)} &= a_0^{(2)} + \frac{1}{2} n_2 \\
&= a_0 + \frac{1}{2} n_1 + \frac{1}{2} n_2 \\
&= a_0 + \frac{1}{2} (n_1 + n_2) \\
b_n^{(2)} &= b_0^{(2)} + \frac{1}{2} \left( y_2^T P_2 y_2 + \mu_0^{(2)T} \Lambda_0^{(2)} \mu_0^{(2)} - \mu_n^{(2)T} \Lambda_n^{(2)} \mu_n^{(2)} \right) \\
&= b_0 + \frac{1}{2} \left( y_1^T P_1 y_1 + \mu_0^T \Lambda_0 \mu_0 - \mu_n^{(1)T} \Lambda_n^{(1)} \mu_n^{(1)} \right) + \frac{1}{2} \left( y_2^T P_2 y_2 + \mu_n^{(1)T} \Lambda_n^{(1)} \mu_n^{(1)} - \mu_n^{(2)T} \Lambda_n^{(2)} \mu_n^{(2)} \right) \\
&= b_0 + \frac{1}{2} \left( y_1^T P_1 y_1 + y_2^T P_2 y_2 + \mu_0^T \Lambda_0 \mu_0 - \mu_n^{(2)T} \Lambda_n^{(2)} \mu_n^{(2)} \right) .
\end{aligned} \tag{11}$$

These are the prior hyperparameters before seeing the third data set:

$$\begin{aligned}
\mu_0^{(3)} &= \mu_n^{(2)} \\
\Lambda_0^{(3)} &= \Lambda_n^{(2)} \\
a_0^{(3)} &= a_n^{(2)} \\
b_0^{(3)} &= b_n^{(2)} .
\end{aligned} \tag{12}$$

Generalizing this, we have after observing the  $j$ -th data set:

$$\begin{aligned}
\mu_n^{(j)} &= \Lambda_n^{(j)-1} \left( \sum_{i=1}^j X_i^T P_i y_i + \Lambda_0 \mu_0 \right) \\
\Lambda_n^{(j)} &= \sum_{i=1}^j X_i^T P_i X_i + \Lambda_0 \\
a_n^{(j)} &= a_0 + \frac{1}{2} \sum_{i=1}^j n_i \\
b_n^{(j)} &= b_0 + \frac{1}{2} \left( \sum_{i=1}^j y_i^T P_i y_i + \mu_0^T \Lambda_0 \mu_0 - \mu_n^{(j)T} \Lambda_n^{(j)} \mu_n^{(j)} \right) .
\end{aligned} \tag{13}$$

Plugging in  $j = S$ , we obtain the final posterior distribution:

$$\begin{aligned}
\mu_n &= \mu_n^{(S)} = \Lambda_n^{(S)-1} \left( \sum_{i=1}^S X_i^T P_i y_i + \Lambda_0 \mu_0 \right) = \Lambda_n^{-1} \left( \sum_{i=1}^S X_i^T P_i y_i + \Lambda_0 \mu_0 \right) \\
\Lambda_n &= \Lambda_n^{(S)} = \sum_{i=1}^S X_i^T P_i X_i + \Lambda_0 \\
a_n &= a_n^{(S)} = a_0 + \frac{1}{2} \sum_{i=1}^S n_i \\
b_n &= b_n^{(S)} = b_0 + \frac{1}{2} \left( \sum_{i=1}^S y_i^T P_i y_i + \mu_0^T \Lambda_0 \mu_0 - \mu_n^{(S)T} \Lambda_n^{(S)} \mu_n^{(S)} \right) \\
&= b_0 + \frac{1}{2} \left( \sum_{i=1}^S y_i^T P_i y_i + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n \right) .
\end{aligned} \tag{14}$$

This result is also compatible with the general theorem about combined posterior distributions in terms of individual posterior distributions ( $\rightarrow$  I/5.1.11) when analyzing independent data sets. ■

### 1.6.11 Log Bayes factor for comparison of two regression models

**Theorem:** Let  $y = [y_1, \dots, y_n]^T$  be an  $n$ -dimensional vector of a measured univariate signal ( $\rightarrow$  I/1.1.5) and consider two linear regression models ( $\rightarrow$  III/1.5.1) with design matrices ( $\rightarrow$  III/1.5.1)  $X_1, X_2$  and precision matrices ( $\rightarrow$  III/1.6.1)  $P_1, P_2$ , entailing potentially different regression coefficients ( $\rightarrow$  III/1.5.1)  $\beta_1, \beta_2$  and noise precisions ( $\rightarrow$  III/1.6.1)  $\tau_1, \tau_2$ :

$$\begin{aligned}
m_1 : y &= X_1 \beta_1 + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2 V_1), \quad \sigma_1^2 V_1 = (\tau_1 P_1)^{-1} \\
m_2 : y &= X_2 \beta_2 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2 V_2), \quad \sigma_2^2 V_2 = (\tau_2 P_2)^{-1} .
\end{aligned} \tag{1}$$

Moreover, assume normal-gamma prior distributions ( $\rightarrow$  III/1.6.1) over the model parameters  $\beta_1$  and  $\tau_1 = 1/\sigma_1^2$  as well as  $\beta_2$  and  $\tau_2 = 1/\sigma_2^2$ :

$$\begin{aligned}
p(\beta_1, \tau_1) &= \mathcal{N} \left( \beta_1; \mu_0^{(1)}, \left( \tau_1 \Lambda_0^{(1)} \right)^{-1} \right) \cdot \text{Gam} \left( \tau_1; a_0^{(1)}, b_0^{(1)} \right) \\
p(\beta_2, \tau_2) &= \mathcal{N} \left( \beta_2; \mu_0^{(2)}, \left( \tau_2 \Lambda_0^{(2)} \right)^{-1} \right) \cdot \text{Gam} \left( \tau_2; a_0^{(2)}, b_0^{(2)} \right) .
\end{aligned} \tag{2}$$

Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_2$  is

$$\begin{aligned}
\text{LBF}_{12} &= \frac{1}{2} \log \frac{|P_1|}{|P_2|} + \frac{1}{2} \log \frac{|\Lambda_0^{(1)}|}{|\Lambda_0^{(2)}|} - \frac{1}{2} \log \frac{|\Lambda_n^{(1)}|}{|\Lambda_n^{(2)}|} \\
&\quad + \log \frac{\Gamma(a_n^{(1)})}{\Gamma(a_0^{(1)})} + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\
&\quad + \log \frac{\Gamma(a_n^{(2)})}{\Gamma(a_0^{(2)})} - a_0^{(2)} \log b_0^{(2)} + a_n^{(2)} \log b_n^{(2)}
\end{aligned} \tag{3}$$

where  $\mu_n^{(1)}, \Lambda_n^{(1)}, a_n^{(1)}, b_n^{(1)}$  and  $\mu_n^{(2)}, \Lambda_n^{(2)}, a_n^{(2)}, b_n^{(2)}$  are the posterior hyperparameters for Bayesian linear regression ( $\rightarrow$  III/1.6.2) for each of the two models which are functions of the design matrices, the precision matrices and the data vector.

**Proof:** For Bayesian linear regression with data vector  $y$ , design matrix  $X$ , precision matrix  $P$  and a normal-gamma prior distribution ( $\rightarrow$  III/1.6.1) over  $\beta$  and  $\tau$ , the log model evidence is given by ( $\rightarrow$  III/1.6.3)

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \quad (4)$$

where the posterior hyperparameters are equal to ( $\rightarrow$  III/1.6.2)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (5)$$

Thus, the log model evidences ( $\rightarrow$  IV/3.1.3) for  $m_1$  and  $m_2$  are given by:

$$\begin{aligned} \text{LME}(m_1) &= \frac{1}{2} \log |P_1| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0^{(1)}| - \frac{1}{2} \log |\Lambda_n^{(1)}| + \\ & \quad \log \Gamma(a_n^{(1)}) - \log \Gamma(a_0^{(1)}) + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\ \text{LME}(m_2) &= \frac{1}{2} \log |P_2| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0^{(2)}| - \frac{1}{2} \log |\Lambda_n^{(2)}| + \\ & \quad \log \Gamma(a_n^{(2)}) - \log \Gamma(a_0^{(2)}) + a_0^{(2)} \log b_0^{(2)} - a_n^{(2)} \log b_n^{(2)} . \end{aligned} \quad (6)$$

The log Bayes factor is equal to the difference of two log model evidences ( $\rightarrow$  IV/3.3.8):

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (7)$$

Plugging (6) into (7), this gives:

$$\begin{aligned} \text{LBF}_{12} &= \frac{1}{2} \log |P_1| - \frac{1}{2} \log |P_2| \\ & \quad + \frac{1}{2} \log |\Lambda_0^{(1)}| - \frac{1}{2} \log |\Lambda_0^{(2)}| \\ & \quad - \frac{1}{2} \log |\Lambda_n^{(1)}| + \frac{1}{2} \log |\Lambda_n^{(2)}| \\ & \quad + \log \Gamma(a_n^{(1)}) - \log \Gamma(a_0^{(1)}) + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\ & \quad - \log \Gamma(a_n^{(2)}) + \log \Gamma(a_0^{(2)}) - a_0^{(2)} \log b_0^{(2)} + a_n^{(2)} \log b_n^{(2)} . \end{aligned} \quad (8)$$

Applying  $\log a - \log b = \log(a/b)$ , we obtain:



$$\begin{aligned}
\text{LBF}_{12} = & \frac{1}{2} \log \frac{|P_1|}{|P_2|} + \frac{1}{2} \log \frac{|\Lambda_0^{(1)}|}{|\Lambda_0^{(2)}|} - \frac{1}{2} \log \frac{|\Lambda_n^{(1)}|}{|\Lambda_n^{(2)}|} \\
& + \log \frac{\Gamma(a_n^{(1)})}{\Gamma(a_0^{(1)})} + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\
& - \log \frac{\Gamma(a_n^{(2)})}{\Gamma(a_0^{(2)})} - a_0^{(2)} \log b_0^{(2)} + a_n^{(2)} \log b_n^{(2)} .
\end{aligned} \tag{9}$$

■

## 1.7 Bayesian linear regression with known covariance

### 1.7.1 Conjugate prior distribution

**Theorem:** Let

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{1}$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$  and known  $n \times n$  covariance matrix  $\Sigma$  as well as unknown  $p \times 1$  regression coefficients  $\beta$ .

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for this model is a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) . \tag{2}$$

**Proof:** By definition, a conjugate prior ( $\rightarrow$  I/5.2.5) is a prior distribution ( $\rightarrow$  I/5.1.3) that, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2):

$$p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right] . \tag{3}$$

Expanding the product in the exponent, we have:

$$p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (y^T \Sigma^{-1} y - y^T \Sigma^{-1} X\beta - \beta^T X^T \Sigma^{-1} y + \beta^T X^T \Sigma^{-1} X\beta) \right] . \tag{4}$$

Completing the square over  $\beta$ , one obtains

$$p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} \left( (\beta - \tilde{X}y)^T X^T \Sigma^{-1} X (\beta - \tilde{X}y) - y^T Q y + y^T \Sigma^{-1} y \right) \right] \tag{5}$$

where  $\tilde{X} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$  and  $Q = \tilde{X}^T (X^T \Sigma^{-1} X) \tilde{X}$ .

Separating constant and variable terms, we get:

$$p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (y^T Q y + y^T \Sigma^{-1} y) \right] \cdot \exp \left[ -\frac{1}{2} (\beta - \tilde{X} y)^T X^T \Sigma^{-1} X (\beta - \tilde{X} y) \right] . \quad (6)$$

In other words, the likelihood function ( $\rightarrow$  I/5.1.2) is proportional to an exponential of a squared form of  $\beta$ :

$$p(y|\beta) \propto \exp \left[ -\frac{1}{2} (\beta - \tilde{X} y)^T X^T \Sigma^{-1} X (\beta - \tilde{X} y) \right] . \quad (7)$$

The same is true for a multivariate normal distribution ( $\rightarrow$  II/4.1.1) over  $\beta$

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \quad (8)$$

the probability density function of which ( $\rightarrow$  II/4.1.4)

$$p(\beta) = \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \cdot \exp \left[ -\frac{1}{2} (\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right] \quad (9)$$

exhibits the same proportionality

$$p(\beta) \propto \exp \left[ -\frac{1}{2} (\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right] \quad (10)$$

and is therefore conjugate relative to the likelihood. ■

### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, eq. 3.48; URL: <https://www.springer.com/gp/book/9780387310732>.
- Penny WD (2012): “Comparing Dynamic Causal Models using AIC, BIC and Free Energy”; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eq. 9; URL: <https://www.sciencedirect.com/science/article/pii/S1053811911008160>; DOI: 10.1016/j.neuroimage.2011.07.039.

## 1.7.2 Posterior distribution

**Theorem:** Let

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma) \quad (1)$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$  and known  $n \times n$  covariance matrix  $\Sigma$  as well as unknown  $p \times 1$  regression coefficients  $\beta$ . Moreover, assume a multivariate normal distribution ( $\rightarrow$  III/1.7.1) over the model parameter  $\beta$ :

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a multivariate normal distribution ( $\rightarrow$  II/4.1.1)

$$p(\beta|y) = \mathcal{N}(\beta; \mu_n, \Sigma_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned}\mu_n &= \Sigma_n(X^T \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\ \Sigma_n &= (X^T \Sigma^{-1} X + \Sigma_0^{-1})^{-1}.\end{aligned}\tag{4}$$

**Proof:** According to Bayes' theorem ( $\rightarrow$  I/5.3.2), the posterior distribution ( $\rightarrow$  I/5.1.9) is given by

$$p(\beta|y) = \frac{p(y|\beta) p(\beta)}{p(y)}.\tag{5}$$

Since  $p(y)$  is just a normalization factor, the posterior is proportional ( $\rightarrow$  I/5.1.10) to the numerator:

$$p(\beta|y) \propto p(y|\beta) p(\beta) = p(y, \beta).\tag{6}$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2):

$$p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right].\tag{7}$$

Combining the likelihood function ( $\rightarrow$  I/5.1.2) (7) with the prior distribution ( $\rightarrow$  I/5.1.3) (2) using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned}p(y, \beta) &= p(y|\beta) p(\beta) \\ &= \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right] \cdot \\ &\quad \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \exp \left[ -\frac{1}{2} (\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right].\end{aligned}\tag{8}$$

Collecting identical variables gives:

$$\begin{aligned}p(y, \beta) &= \sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \cdot \\ &\quad \exp \left[ -\frac{1}{2} ((y - X\beta)^T \Sigma^{-1} (y - X\beta) + (\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0)) \right].\end{aligned}\tag{9}$$

Expanding the products in the exponent gives:

$$\begin{aligned}p(y, \beta) &= \sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \cdot \\ &\quad \exp \left[ -\frac{1}{2} (y^T \Sigma^{-1} y - y^T \Sigma^{-1} X\beta - \beta^T X^T \Sigma^{-1} y + \beta^T X^T \Sigma^{-1} X\beta + \right. \\ &\quad \left. \beta^T \Sigma_0^{-1} \beta - \beta^T \Sigma_0^{-1} \mu_0 - \mu_0^T \Sigma_0^{-1} \beta + \mu_0^T \Sigma_0^{-1} \mu_0) \right].\end{aligned}\tag{10}$$

Regrouping the terms in the exponent gives:

$$\begin{aligned}
p(y, \beta) = & \sqrt{\frac{1}{(2\pi)^{n+p}|\Sigma||\Sigma_0|}} \cdot \\
& \exp \left[ -\frac{1}{2} \left( \beta^T [X^T \Sigma^{-1} X + \Sigma_0^{-1}] \beta - 2\beta^T [X^T \Sigma^{-1} y + \Sigma_0^{-1} \mu_0] + \right. \right. \\
& \left. \left. y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 \right) \right] .
\end{aligned} \tag{11}$$

Completing the square over  $\beta$ , we finally have

$$\begin{aligned}
p(y, \beta) = & \sqrt{\frac{1}{(2\pi)^{n+p}|\Sigma||\Sigma_0|}} \cdot \\
& \exp \left[ -\frac{1}{2} \left( (\beta - \mu_n)^T \Sigma_n^{-1} (\beta - \mu_n) + (y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n) \right) \right]
\end{aligned} \tag{12}$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned}
\mu_n &= \Sigma_n (X^T \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\
\Sigma_n &= (X^T \Sigma^{-1} X + \Sigma_0^{-1})^{-1} .
\end{aligned} \tag{13}$$

Ergo, the joint likelihood is proportional to

$$p(y, \beta) \propto \exp \left[ -\frac{1}{2} (\beta - \mu_n)^T \Sigma_n^{-1} (\beta - \mu_n) \right] , \tag{14}$$

such that the posterior distribution over  $\beta$  is given by

$$p(\beta|y) = \mathcal{N}(\beta; \mu_n, \Sigma_n) \tag{15}$$

with the posterior hyperparameters given in (13). ■

### Sources:

- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, eqs. 3.49-3.51, ex. 3.7; URL: <https://www.springer.com/gp/book/9780387310732>.
- Penny WD (2012): “Comparing Dynamic Causal Models using AIC, BIC and Free Energy”; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eq. 27; URL: <https://www.sciencedirect.com/science/article/pii/S1053811911008160>; DOI: 10.1016/j.neuroimage.2011.07.039.

### 1.7.3 Log model evidence

**Theorem:** Let

$$m : y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{1}$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$  and known  $n \times n$  covariance matrix  $\Sigma$  as well as unknown  $p \times 1$  regression coefficients  $\beta$ . Moreover, assume a multivariate normal distribution ( $\rightarrow$  III/1.7.1) over the model parameter  $\beta$ :

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned} \log p(y|m) = & -\frac{1}{2} e_y^T \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) \\ & - \frac{1}{2} e_\beta^T \Sigma_0^{-1} e_\beta - \frac{1}{2} \log |\Sigma_0| + \frac{1}{2} \log |\Sigma_n| . \end{aligned} \quad (3)$$

with the “prediction error” and “parameter error” terms

$$\begin{aligned} e_y &= y - X\mu_n \\ e_\beta &= \mu_0 - \mu_n \end{aligned} \quad (4)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \Sigma_n (X^T \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\ \Sigma_n &= (X^T \Sigma^{-1} X + \Sigma_0^{-1})^{-1} . \end{aligned} \quad (5)$$

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the model evidence ( $\rightarrow$  I/5.1.15) for this model is:

$$p(y|m) = \int p(y|\beta) p(\beta) d\beta . \quad (6)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand is equivalent to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(y|m) = \int p(y, \beta) d\beta . \quad (7)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2):

$$p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right] . \quad (8)$$

When deriving the posterior distribution ( $\rightarrow$  III/1.7.2)  $p(\beta|y)$ , the joint likelihood  $p(y, \beta)$  is obtained as

$$\begin{aligned} p(y, \beta) = & \sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \\ & \exp \left[ -\frac{1}{2} ((\beta - \mu_n)^T \Sigma_n^{-1} (\beta - \mu_n) + (y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n)) \right] . \end{aligned} \quad (9)$$

Using the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), we can rewrite this as

$$p(y, \beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \sqrt{\frac{(2\pi)^p |\Sigma_n|}{1}} \cdot \mathcal{N}(\beta; \mu_n, \Sigma_n) \cdot \exp \left[ -\frac{1}{2} (y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n) \right]. \quad (10)$$

With that,  $\beta$  can be integrated out easily:

$$\int p(y, \beta) d\beta = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{|\Sigma_n|}{|\Sigma_0|}} \cdot \exp \left[ -\frac{1}{2} (y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n) \right]. \quad (11)$$

Now we turn to the intra-exponent term

$$y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n \quad (12)$$

and plug in the posterior covariance

$$\Sigma_n = (X^T \Sigma^{-1} X + \Sigma_0^{-1})^{-1}. \quad (13)$$

This gives

$$\begin{aligned} & y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T \Sigma_n^{-1} \mu_n \\ &= y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T (X^T \Sigma^{-1} X + \Sigma_0^{-1}) \mu_n \\ &= y^T \Sigma^{-1} y + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_n^T X^T \Sigma^{-1} X \mu_n - \mu_n^T \Sigma_0^{-1} \mu_n \\ &= (y - X \mu_n)^T \Sigma^{-1} (y - X \mu_n) + (\mu_0 - \mu_n)^T \Sigma_0^{-1} (\mu_0 - \mu_n) \\ &\stackrel{(4)}{=} e_y^T \Sigma^{-1} e_y + e_\beta^T \Sigma_0^{-1} e_\beta. \end{aligned} \quad (14)$$

Thus, the marginal likelihood ( $\rightarrow$  I/5.1.15) becomes

$$p(y|m) = \int p(y, \beta) d\beta \stackrel{(11)}{=} \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{|\Sigma_n|}{|\Sigma_0|}} \cdot \exp \left[ -\frac{1}{2} (e_y^T \Sigma^{-1} e_y + e_\beta^T \Sigma_0^{-1} e_\beta) \right] \quad (15)$$

and the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is given by

$$\begin{aligned} \log p(y|m) &= -\frac{1}{2} e_y^T \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) \\ &\quad - \frac{1}{2} e_\beta^T \Sigma_0^{-1} e_\beta - \frac{1}{2} \log |\Sigma_0| + \frac{1}{2} \log |\Sigma_n|. \end{aligned} \quad (16)$$

■

#### Sources:

- Penny WD (2012): “Comparing Dynamic Causal Models using AIC, BIC and Free Energy”; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eqs. 19-23; URL: <https://www.sciencedirect.com/science/article/pii/S1053811911008160>; DOI: 10.1016/j.neuroimage.2011.07.039.
- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161; URL: <https://www.springer.com/gp/book/9780387310732>.

## 1.7.4 Accuracy and complexity

**Theorem:** Let

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \Sigma) \quad (1)$$

be a linear regression model ( $\rightarrow$  III/1.5.1) with measured  $n$ -dimensional data vector  $y$ , known  $n \times p$  design matrix  $X$  and known  $n \times n$  covariance matrix  $\Sigma$  as well as unknown  $p \times 1$  regression coefficients  $\beta$ . Moreover, assume a multivariate normal distribution ( $\rightarrow$  III/1.7.1) over the model parameter  $\beta$ :

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) . \quad (2)$$

Then, accuracy and complexity ( $\rightarrow$  IV/3.1.6) of this model are

$$\begin{aligned} \text{Acc}(m) &= -\frac{1}{2}e_y^T \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \text{tr}(X^T \Sigma^{-1} X \Sigma_n) \\ \text{Com}(m) &= \frac{1}{2}e_\beta^T \Sigma_0^{-1} e_\beta + \frac{1}{2} \log |\Sigma_0| - \frac{1}{2} \log |\Sigma_n| + \frac{1}{2} \text{tr}(\Sigma_0^{-1} \Sigma_n) - \frac{p}{2} \end{aligned} \quad (3)$$

with the “prediction error” and “parameter error” terms

$$\begin{aligned} e_y &= y - X\mu_n \\ e_\beta &= \mu_0 - \mu_n \end{aligned} \quad (4)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \mu_n &= \Sigma_n (X^T \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\ \Sigma_n &= (X^T \Sigma^{-1} X + \Sigma_0^{-1})^{-1} . \end{aligned} \quad (5)$$

**Proof:** Model accuracy and complexity are defined as ( $\rightarrow$  IV/3.1.6)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\beta, m) \rangle_{p(\beta|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\beta|y, m) || p(\beta|m)] . \end{aligned} \quad (6)$$

1) The accuracy term is the expectation ( $\rightarrow$  I/1.12.1) of the log-likelihood function ( $\rightarrow$  I/4.1.2)  $\log p(y|\beta)$  with respect to the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\beta|y)$ :

$$\text{Acc}(m) = \langle \log p(y|\beta) \rangle_{p(\beta|y)} . \quad (7)$$

With the likelihood function for Bayesian linear regression with known covariance ( $\rightarrow$  III/1.7.1), we have:

$$\begin{aligned} \text{Acc}(m) &= \left\langle \log \left( \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right] \right) \right\rangle_{p(\beta|y)} \\ &= \left\langle -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right\rangle_{p(\beta|y)} \\ &= \left\langle -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} [y^T \Sigma^{-1} y - 2y^T \Sigma^{-1} X\beta + \beta^T X^T \Sigma^{-1} X\beta] \right\rangle_{p(\beta|y)} . \end{aligned} \quad (8)$$

With the posterior distribution for Bayesian linear regression with known covariance ( $\rightarrow$  III/1.7.2), this becomes:

$$\text{Acc}(m) = \left\langle -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} [y^T \Sigma^{-1} y - 2y^T \Sigma^{-1} X \beta + \beta^T X^T \Sigma^{-1} X \beta] \right\rangle_{\mathcal{N}(\beta; \mu_n, \Sigma_n)} . \quad (9)$$

If  $x \sim \mathcal{N}(\mu, \Sigma)$ , then its expected value is ( $\rightarrow$  II/4.1.6)

$$\langle x \rangle = \mu \quad (10)$$

and the expectation of a quadratic form is given by ( $\rightarrow$  I/1.12.17)

$$\langle x^T A x \rangle = \mu^T A \mu + \text{tr}(A \Sigma) . \quad (11)$$

Thus, the model accuracy of  $m$  evaluates to

$$\begin{aligned} \text{Acc}(m) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \\ &\quad \frac{1}{2} [y^T \Sigma^{-1} y - 2y^T \Sigma^{-1} X \mu_n + \mu_n^T X^T \Sigma^{-1} X \mu_n + \text{tr}(X^T \Sigma^{-1} X \Sigma_n)] \\ &= -\frac{1}{2} (y - X \mu_n)^T \Sigma^{-1} (y - X \mu_n) - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \text{tr}(X^T \Sigma^{-1} X \Sigma_n) \\ &\stackrel{(4)}{=} -\frac{1}{2} e_y^T \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \text{tr}(X^T \Sigma^{-1} X \Sigma_n) \end{aligned} \quad (12)$$

which proofs the first part of (3).

2) The complexity penalty is the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of the posterior distribution ( $\rightarrow$  I/5.1.9)  $p(\beta|y)$  from the prior distribution ( $\rightarrow$  I/5.1.3)  $p(\beta)$ :

$$\text{Com}(m) = \text{KL} [p(\beta|y) || p(\beta)] . \quad (13)$$

With the prior distribution ( $\rightarrow$  III/1.7.1) given by (2) and the posterior distribution for Bayesian linear regression with known covariance ( $\rightarrow$  III/1.7.2), this becomes:

$$\text{Com}(m) = \text{KL} [\mathcal{N}(\beta; \mu_n, \Sigma_n) || \mathcal{N}(\beta; \mu_0, \Sigma_0)] . \quad (14)$$

With the Kullback-Leibler divergence for the multivariate normal distribution ( $\rightarrow$  II/4.1.12)

$$\text{KL}[\mathcal{N}(\mu_1, \Sigma_1) || \mathcal{N}(\mu_2, \Sigma_2)] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] , \quad (15)$$

the model complexity of  $m$  evaluates to

$$\begin{aligned} \text{Com}(m) &= \frac{1}{2} \left[ (\mu_0 - \mu_n)^T \Sigma_0^{-1} (\mu_0 - \mu_n) + \text{tr}(\Sigma_0^{-1} \Sigma_n) - \log \frac{|\Sigma_n|}{|\Sigma_0|} - p \right] \\ &= \frac{1}{2} (\mu_0 - \mu_n)^T \Sigma_0^{-1} (\mu_0 - \mu_n) + \frac{1}{2} \log |\Sigma_0| - \frac{1}{2} \log |\Sigma_n| + \frac{1}{2} \text{tr}(\Sigma_0^{-1} \Sigma_n) - \frac{p}{2} \\ &\stackrel{(4)}{=} \frac{1}{2} e_\beta^T \Sigma_0^{-1} e_\beta + \frac{1}{2} \log |\Sigma_0| - \frac{1}{2} \log |\Sigma_n| + \frac{1}{2} \text{tr}(\Sigma_0^{-1} \Sigma_n) - \frac{p}{2} \end{aligned} \quad (16)$$



which proofs the second part of (3).

3) A control calculation confirms that

$$\text{Acc}(m) - \text{Com}(m) = \text{LME}(m) \quad (17)$$

where  $\text{LME}(m)$  is the log model evidence for Bayesian linear regression with known covariance ( $\rightarrow$  III/1.7.3):

$$\begin{aligned} \log p(y|m) = & -\frac{1}{2} e_y^T \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) \\ & - \frac{1}{2} e_\beta^T \Sigma_0^{-1} e_\beta - \frac{1}{2} \log |\Sigma_0| + \frac{1}{2} \log |\Sigma_n| . \end{aligned} \quad (18)$$

This requires to recognize, based on (5), that

$$\begin{aligned} & -\frac{1}{2} \text{tr}(X^T \Sigma^{-1} X \Sigma_n) - \frac{1}{2} \text{tr}(\Sigma_0^{-1} \Sigma_n) + \frac{p}{2} \\ = & -\frac{1}{2} \text{tr}([X^T \Sigma^{-1} X + \Sigma_0^{-1}] \Sigma_n) + \frac{p}{2} \\ = & -\frac{1}{2} \text{tr}(\Sigma_n^{-1} \Sigma_n) + \frac{p}{2} \\ = & -\frac{1}{2} \text{tr}(I_p) + \frac{p}{2} \\ = & -\frac{p}{2} + \frac{p}{2} \\ = & 0 . \end{aligned} \quad (19)$$

■

#### Sources:

- Penny WD (2012): “Comparing Dynamic Causal Models using AIC, BIC and Free Energy”; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eqs. 20-21; URL: <https://www.sciencedirect.com/science/article/pii/S1053811911008160>; DOI: 10.1016/j.neuroimage.2011.07.039.
- Bishop CM (2006): “Bayesian Linear Regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161; URL: <https://www.springer.com/gp/book/9780387310732>.

## 2 Multivariate normal data

### 2.1 Multivariate Gaussian

#### 2.1.1 Bivariate normally distributed data

**Definition:** Bivariate normally distributed data are defined as a set of two-dimensional vectors  $y = \{y_1, \dots, y_n\}$  with  $y_i \in \mathbb{R}^2$ ,  $i = 1, \dots, n$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a bivariate normal distribution ( $\rightarrow$  II/4.2.1) with mean vector  $\mu$  and covariance matrix  $\Sigma$

$$y_i \sim \mathcal{N}(\mu, \Sigma), \quad i = 1, \dots, n \quad (1)$$

where  $\mu$  and  $\Sigma$  can be parametrized ( $\rightarrow$  II/4.2.4) in terms of ( $\rightarrow$  II/4.2.3) means  $\mu_1, \mu_2$ , variances  $\sigma_1^2, \sigma_2^2$  and correlation  $\rho$ :

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \quad (2)$$

#### 2.1.2 Multivariate normally distributed data

**Definition:** Multivariate normally distributed data are defined as a set of  $p$ -dimensional vectors  $y = \{y_1, \dots, y_n\}$  with  $y_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a multivariate normal distribution ( $\rightarrow$  II/4.1.1) with mean vector  $\mu \in \mathbb{R}^p$  and covariance matrix  $\Sigma \in \mathbb{R}^{p \times p}$

$$y_i \sim \mathcal{N}(\mu, \Sigma), \quad i = 1, \dots, n \quad (1)$$

where  $\mu$  is a  $p$ -dimensional real vector and  $\Sigma$  is a  $p \times p$  positive-definite matrix.

#### 2.1.3 Maximum likelihood estimation ( $p = 2$ )

**Theorem:** Let there be a bivariate normally distributed data set ( $\rightarrow$  III/2.1.1)  $y = \{y_1, \dots, y_n\}$ :

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  and  $\rho$  are given by

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{n} \sum_{i=1}^n y_{i1} \\
\hat{\mu}_2 &= \frac{1}{n} \sum_{i=1}^n y_{i2} \\
\hat{\sigma}_1^2 &= \frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)^2 \\
\hat{\sigma}_2^2 &= \frac{1}{n} \sum_{i=1}^n (y_{i2} - \hat{\mu}_2)^2 \\
\hat{\rho} &= \frac{\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2)}{\sqrt{\hat{\sigma}_1^2} \sqrt{\hat{\sigma}_2^2}} .
\end{aligned} \tag{2}$$

**Proof:** This is a special case of multivariate normally distributed data ( $\rightarrow$  III/2.1.2)

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad i = 1, \dots, n \tag{3}$$

for which maximum likelihood estimates are given by ( $\rightarrow$  III/2.1.4)

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i \\
\hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})(y_i - \hat{\mu})^T .
\end{aligned} \tag{4}$$

Applying this to (1) by setting  $p = 2$ , we obtain:

$$\begin{aligned}
\hat{\mu} &= \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n y_{i1} \\ \frac{1}{n} \sum_{i=1}^n y_{i2} \end{bmatrix} \\
\hat{\Sigma} &= \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i1} - \hat{\mu}_1) & \frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2) \\ \frac{1}{n} \sum_{i=1}^n (y_{i2} - \hat{\mu}_2)(y_{i1} - \hat{\mu}_1) & \frac{1}{n} \sum_{i=1}^n (y_{i2} - \hat{\mu}_2)(y_{i2} - \hat{\mu}_2) \end{bmatrix} .
\end{aligned} \tag{5}$$

Equating this with the parametrized bivariate normal estimates ( $\rightarrow$  II/4.2.4)

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\rho} \hat{\sigma}_1 \hat{\sigma}_2 \\ \hat{\rho} \hat{\sigma}_1 \hat{\sigma}_2 & \hat{\sigma}_2^2 \end{bmatrix} , \tag{6}$$

we obtain ML estimates for the means as

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n} \sum_{i=1}^n y_{i1} \\ \hat{\mu}_2 &= \frac{1}{n} \sum_{i=1}^n y_{i2} ,\end{aligned}\tag{7}$$

we obtain ML estimates for the variances as

$$\begin{aligned}\hat{\sigma}_1^2 &= \frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)^2 \\ \hat{\sigma}_2^2 &= \frac{1}{n} \sum_{i=1}^n (y_{i2} - \hat{\mu}_2)^2\end{aligned}\tag{8}$$

and we obtain an ML estimate for the correlation as:

$$\begin{aligned}\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 &= \frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2) \\ \hat{\rho} &= \frac{\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2)}{\hat{\sigma}_1\hat{\sigma}_2} \\ \hat{\rho} &= \frac{\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2)}{\sqrt{\hat{\sigma}_1^2}\sqrt{\hat{\sigma}_2^2}} \\ \hat{\rho} &= \frac{\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)(y_{i2} - \hat{\mu}_2)}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{i2} - \hat{\mu}_2)^2}} .\end{aligned}\tag{9}$$

Together, (7), (8) and (9) constitute the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) for bivariate normally distributed data ( $\rightarrow$  III/2.1.1). ■

#### 2.1.4 Maximum likelihood estimation ( $p > 2$ )

**Theorem:** Let there be a multivariate normally distributed data set ( $\rightarrow$  III/2.1.2)  $y = \{y_1, \dots, y_n\}$ :

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad i = 1, \dots, n .\tag{1}$$

Then, the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\mu$  and  $\Sigma$  are given by

$$\begin{aligned}\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})(y_i - \hat{\mu})^T .\end{aligned}\tag{2}$$

**Proof:** With the probability density function of the multivariate normal distribution ( $\rightarrow$  II/4.1.4), the likelihood function for a single data point ( $\rightarrow$  I/1.1.5) is

$$\begin{aligned} p(y_i|\mu, \Sigma) &= \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \cdot \exp \left[ -\frac{1}{2} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right] \end{aligned} \quad (3)$$

and all data points are independent ( $\rightarrow$  I/1.3.6), the total likelihood function is equal to the product ( $\rightarrow$  I/1.3.9) of the individual likelihood functions:

$$\begin{aligned} p(y|\mu, \Sigma) &= \prod_{i=1}^n p(y_i|\mu, \Sigma) \\ &= \frac{1}{\sqrt{(2\pi)^{np} |\Sigma|^n}} \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right]. \end{aligned} \quad (4)$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$\begin{aligned} \text{LL}(\mu, \Sigma) &= \log p(y|\mu, \Sigma) \\ &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \\ &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i^T \Sigma^{-1} y_i - 2\mu^T \Sigma^{-1} y_i + \mu^T \Sigma^{-1} \mu). \end{aligned} \quad (5)$$

The derivative of the log-likelihood function (5) with respect to  $\mu$  is

$$\begin{aligned} \frac{d\text{LL}(\mu, \Sigma)}{d\mu} &= \frac{d}{d\mu} \left[ -\frac{1}{2} \sum_{i=1}^n (y_i^T \Sigma^{-1} y_i - 2\mu^T \Sigma^{-1} y_i + \mu^T \Sigma^{-1} \mu) \right] \\ &= -\frac{1}{2} \sum_{i=1}^n (-2\Sigma^{-1} y_i + 2\Sigma^{-1} \mu) \\ &= \sum_{i=1}^n (\Sigma^{-1} y_i - \Sigma^{-1} \mu) \end{aligned} \quad (6)$$

and setting this derivative to zero gives the MLE for  $\mu$ :

$$\begin{aligned}
\frac{dLL(\hat{\mu}, \Sigma)}{d\mu} &= 0_p \\
0_p &= \sum_{i=1}^n (\Sigma^{-1}y_i - \Sigma^{-1}\hat{\mu}) \\
0_p &= \Sigma^{-1} \sum_{i=1}^n (y_i - \hat{\mu}) \\
0_p &= \sum_{i=1}^n y_i - n \cdot \hat{\mu} \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i .
\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (5) at  $\hat{\mu}$  with respect to  $\Sigma$  is

$$\begin{aligned}
\frac{dLL(\hat{\mu}, \Sigma)}{d\Sigma} &= -\frac{n}{2} \frac{d}{d\Sigma} [\log |\Sigma|] - \frac{1}{2} \frac{d}{d\Sigma} \left[ \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right] \\
&= -\frac{n}{2} (\Sigma^{-1})^T - \frac{1}{2} \sum_{i=1}^n \frac{d}{d\Sigma} [\text{tr} ((y_i - \mu)^T \Sigma^{-1} (y_i - \mu))] \\
&= -\frac{n}{2} \Sigma^{-1} - \frac{1}{2} \sum_{i=1}^n \frac{d}{d\Sigma} [\text{tr} (\Sigma^{-1} (y_i - \mu) (y_i - \mu)^T)] \\
&= -\frac{n}{2} \Sigma^{-1} - \frac{1}{2} \sum_{i=1}^n \left[ -(\Sigma^{-1} (y_i - \mu) (y_i - \mu)^T \Sigma^{-1})^T \right] \\
&= -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \sum_{i=1}^n [\Sigma^{-1} (y_i - \mu) (y_i - \mu)^T \Sigma^{-1}]
\end{aligned} \tag{8}$$

and setting this derivative to zero gives the MLE for  $\Sigma$ :

$$\begin{aligned}
\frac{dLL(\hat{\mu}, \hat{\Sigma})}{d\Sigma} &= 0_{pp} \\
0_{pp} &= -\frac{n}{2} \hat{\Sigma}^{-1} + \frac{1}{2} \sum_{i=1}^n [\hat{\Sigma}^{-1} (y_i - \mu) (y_i - \mu)^T \hat{\Sigma}^{-1}] \\
\hat{\Sigma}^{-1} &= \frac{1}{n} \Sigma^{-1} \left( \sum_{i=1}^n (y_i - \mu) (y_i - \mu)^T \right) \Sigma^{-1} \\
\hat{\Sigma} \hat{\Sigma}^{-1} \hat{\Sigma} &= \frac{1}{n} \hat{\Sigma} \Sigma^{-1} \left( \sum_{i=1}^n (y_i - \mu) (y_i - \mu)^T \right) \Sigma^{-1} \hat{\Sigma} \\
I_p \hat{\Sigma} &= \frac{1}{n} I_p \left( \sum_{i=1}^n (y_i - \mu) (y_i - \mu)^T \right) I_p \\
\hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (y_i - \mu) (y_i - \mu)^T
\end{aligned} \tag{9}$$

Together, (7) and (9) constitute the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) for multivariate normally distributed data ( $\rightarrow$  III/2.1.2).

■

#### Sources:

- Alt (2014): “MLE of multivariate (bivariate) normal distribution”; in: *StackExchange Mathematics*, retrieved on 2025-07-01; URL: <https://math.stackexchange.com/a/688030>.
- Petersen, Kaare Brandt; Pedersen, Michael Syskind (2012): “Derivatives”; in: *The Matrix Cookbook*, Section 2, eqs. (100), (117), (57), (124); URL: <https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html>.
- Wikipedia (2025): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-07-01; URL: [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Parameter\\_estimation](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Parameter_estimation).

## 2.2 General linear model

### 2.2.1 Definition

**Definition:** Let  $Y$  be an  $n \times v$  matrix and let  $X$  be an  $n \times p$  matrix. Then, a statement asserting a linear mapping from  $X$  to  $Y$  with parameters  $B$  and matrix-normally distributed ( $\rightarrow$  II/5.1.1) errors  $E$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

is called a multivariate linear regression model or simply, “general linear model”.

- $Y$  is called “data matrix”, “set of dependent variables” or “measurements”;
- $X$  is called “design matrix”, “set of independent variables” or “predictors”;
- $B$  are called “regression coefficients” or “weights”;
- $E$  is called “noise matrix” or “error terms”;
- $V$  is called “covariance across rows”;
- $\Sigma$  is called “covariance across columns”;
- $n$  is the number of observations;
- $v$  is the number of measurements;
- $p$  is the number of predictors.

When rows of  $Y$  correspond to units of time, e.g. subsequent measurements,  $V$  is called “temporal covariance”. When columns of  $Y$  correspond to units of space, e.g. measurement channels,  $\Sigma$  is called “spatial covariance”.

When the covariance matrix  $V$  is a scalar multiple of the  $n \times n$  identity matrix, this is called a general linear model with independent and identically distributed ( $\rightarrow$  I/1.2.8) (i.i.d.) observations:

$$V = \lambda I_n \quad \Rightarrow \quad E \sim \mathcal{MN}(0, \lambda I_n, \Sigma) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda \Sigma) . \quad (2)$$

Otherwise, it is called a general linear model with correlated observations ( $\rightarrow$  I/1.17.1).

#### Sources:

- Wikipedia (2020): “General linear model”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: [https://en.wikipedia.org/wiki/General\\_linear\\_model](https://en.wikipedia.org/wiki/General_linear_model).

### 2.2.2 Ordinary least squares

**Theorem:** Given a general linear model ( $\rightarrow$  III/2.2.1) with independent observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, \sigma^2 I_n, \Sigma), \quad (1)$$

the ordinary least squares ( $\rightarrow$  III/1.5.3) parameters estimates are given by

$$\hat{B} = (X^T X)^{-1} X^T Y. \quad (2)$$

**Proof:** Let  $\hat{B}$  be the ordinary least squares ( $\rightarrow$  III/1.5.3) (OLS) solution and let  $\hat{E} = Y - X\hat{B}$  be the resulting matrix of residuals. According to the exogeneity assumption of OLS, the errors have conditional mean ( $\rightarrow$  I/1.12.1) zero

$$E(E|X) = 0, \quad (3)$$

a direct consequence of which is that the regressors are uncorrelated with the errors

$$E(X^T E) = 0, \quad (4)$$

which, in the finite sample, means that the residual matrix must be orthogonal to the design matrix:

$$X^T \hat{E} = 0. \quad (5)$$

From (5), the OLS formula can be directly derived:

$$\begin{aligned} X^T \hat{E} &= 0 \\ X^T (Y - X\hat{B}) &= 0 \\ X^T Y - X^T X\hat{B} &= 0 \\ X^T X\hat{B} &= X^T Y \\ \hat{B} &= (X^T X)^{-1} X^T Y. \end{aligned} \quad (6)$$

■

#### Sources:

- Wikipedia (2025): “Ordinary least squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2025-12-19; URL: [https://en.wikipedia.org/wiki/Ordinary\\_least\\_squares#Assumptions](https://en.wikipedia.org/wiki/Ordinary_least_squares#Assumptions).

### 2.2.3 Weighted least squares

**Theorem:** Given a general linear model ( $\rightarrow$  III/2.2.1) with correlated observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma), \quad (1)$$

the weighted least squares ( $\rightarrow$  III/1.5.21) parameter estimates are given by

$$\hat{B} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y. \quad (2)$$

**Proof:** Let there be an  $n \times n$  square matrix  $W$ , such that



$$WVW^T = I_n . \quad (3)$$

Since  $V$  is a covariance matrix and thus symmetric ( $\rightarrow$  I/1.15.4),  $W$  is also symmetric and can be expressed as the matrix square root of the inverse of  $V$ :

$$WW = V^{-1} \Leftrightarrow W = V^{-1/2} . \quad (4)$$

Left-multiplying the linear regression equation (1) with  $W$ , the linear transformation theorem ( $\rightarrow$  II/5.1.9) implies that

$$WY = WXB + WE, \quad WE \sim \mathcal{MN}(0, WVW^T, \Sigma) . \quad (5)$$

Applying (3), we see that (5) is actually a general linear model ( $\rightarrow$  III/2.2.1) with independent observations

$$\tilde{Y} = \tilde{X}B + \tilde{E}, \quad \tilde{E} \sim \mathcal{N}(0, I_n, \Sigma) \quad (6)$$

where  $\tilde{Y} = WY$ ,  $\tilde{X} = WX$  and  $\tilde{E} = WE$ , such that we can apply the ordinary least squares solution ( $\rightarrow$  III/2.2.2) giving

$$\begin{aligned} \hat{B} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} \\ &= ((WX)^T WX)^{-1} (WX)^T WY \\ &= (X^T W^T W X)^{-1} X^T W^T WY \\ &= (X^T W W X)^{-1} X^T W WY \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} Y \end{aligned} \quad (7)$$

which corresponds to the weighted least squares solution (2). ■

### 2.2.4 Maximum likelihood estimation

**Theorem:** Given a general linear model ( $\rightarrow$  III/2.2.1) with matrix-normally distributed ( $\rightarrow$  II/5.1.1) errors

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) , \quad (1)$$

maximum likelihood estimates ( $\rightarrow$  I/4.1.3) for the unknown parameters  $B$  and  $\Sigma$  are given by

$$\begin{aligned} \hat{B} &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) . \end{aligned} \quad (2)$$

**Proof:** In (1),  $Y$  is an  $n \times v$  matrix of measurements ( $n$  observations,  $v$  dependent variables),  $X$  is an  $n \times p$  design matrix ( $n$  observations,  $p$  independent variables) and  $V$  is an  $n \times n$  covariance matrix across observations. This multivariate GLM implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$\begin{aligned}
p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\
&= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right]
\end{aligned} \tag{3}$$

and the log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\begin{aligned}
\text{LL}(B, \Sigma) &= \log p(Y|B, \Sigma) \\
&= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{v}{2} \log |V| \\
&\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)] .
\end{aligned} \tag{4}$$

Substituting  $V^{-1}$  by the precision matrix  $P$  to ease notation, we have:

$$\begin{aligned}
\text{LL}(B, \Sigma) &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| + \frac{v}{2} \log |P| \\
&\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] .
\end{aligned} \tag{5}$$

The derivative of the log-likelihood function (5) with respect to  $B$  is

$$\begin{aligned}
\frac{d\text{LL}(B, \Sigma)}{dB} &= \frac{d}{dB} \left( -\frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] \right) \\
&= \frac{d}{dB} \left( -\frac{1}{2} \text{tr} [-2\Sigma^{-1} Y^T P X B] \right) + \frac{d}{dB} \left( -\frac{1}{2} \text{tr} [\Sigma^{-1} B^T X^T P X B] \right) \\
&= -\frac{1}{2} (-2X^T P Y \Sigma^{-1}) - \frac{1}{2} (X^T P X B \Sigma^{-1} + (X^T P X)^T B (\Sigma^{-1})^T) \\
&= X^T P Y \Sigma^{-1} - X^T P X B \Sigma^{-1}
\end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for  $B$ :

$$\begin{aligned}
\frac{d\text{LL}(\hat{B}, \Sigma)}{dB} &= 0 \\
0 &= X^T P Y \Sigma^{-1} - X^T P X \hat{B} \Sigma^{-1} \\
0 &= X^T P Y - X^T P X \hat{B} \\
X^T P X \hat{B} &= X^T P Y \\
\hat{B} &= (X^T P X)^{-1} X^T P Y .
\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at  $\hat{B}$  with respect to  $\Sigma$  is

$$\begin{aligned}
\frac{d\text{LL}(\hat{B}, \Sigma)}{d\Sigma} &= \frac{d}{d\Sigma} \left( -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B})] \right) \\
&= -\frac{n}{2} (\Sigma^{-1})^T + \frac{1}{2} \left( \Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \Sigma^{-1} \right)^T \\
&= -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \Sigma^{-1}
\end{aligned} \tag{8}$$

and setting this derivative to zero gives the MLE for  $\Sigma$ :

$$\begin{aligned}
\frac{dLL(\hat{B}, \hat{\Sigma})}{d\Sigma} &= 0 \\
0 &= -\frac{n}{2} \hat{\Sigma}^{-1} + \frac{1}{2} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\frac{n}{2} \hat{\Sigma}^{-1} &= \frac{1}{2} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) .
\end{aligned} \tag{9}$$

Together, (7) and (9) constitute the MLE for the GLM. ■

#### Sources:

- Petersen, Kaare Brandt; Pedersen, Michael Syskind (2012): “Derivatives”; in: *The Matrix Cookbook*, Section 2, eqs. (100), (117), (57), (124); URL: <https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html>.

### 2.2.5 Maximum log-likelihood

**Theorem:** Consider a general linear model ( $\rightarrow$  III/2.2.1)  $m$  with  $n \times v$  data matrix  $Y$ ,  $n \times p$  design matrix  $X$  and  $n \times n$  covariance across rows ( $\rightarrow$  III/2.2.1)  $V$

$$m : Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) . \tag{1}$$

Then, the maximum log-likelihood ( $\rightarrow$  I/4.1.4) for this model is

$$\text{MLL}(m) = -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{nv}{2} \tag{2}$$

under uncorrelated observations ( $\rightarrow$  III/2.2.1), i.e. if  $V = I_n$ , and

$$\text{MLL}(m) = -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{v}{2} \ln |V| - \frac{nv}{2} , \tag{3}$$

in the general case, i.e. if  $V \neq I_n$ , where  $\hat{\Sigma}$  is the maximum likelihood estimate ( $\rightarrow$  I/4.1.3) of the  $v \times v$  covariance across columns ( $\rightarrow$  III/2.2.1).

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for the general linear model is given by ( $\rightarrow$  III/2.2.4)

$$\begin{aligned}
p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\
&= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] ,
\end{aligned} \tag{4}$$

such that the log-likelihood function ( $\rightarrow$  I/4.1.2) for this model becomes ( $\rightarrow$  III/2.2.4)

$$\begin{aligned} \text{LL}(B, \Sigma) &= \log p(Y|B, \Sigma) \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{v}{2} \log |V| - \frac{1}{2} \text{tr} [\Sigma^{-1}(Y - XB)^T V^{-1}(Y - XB)] . \end{aligned} \quad (5)$$

The maximum likelihood estimate for the noise covariance ( $\rightarrow$  III/2.2.4) is

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^T V^{-1}(Y - X\hat{B}) \quad (6)$$

Plugging (6) into (5), we obtain the maximum log-likelihood ( $\rightarrow$  I/4.1.4) as

$$\begin{aligned} \text{MLL}(m) &= \text{LL}(\hat{B}, \hat{\Sigma}) \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\hat{\Sigma}| - \frac{v}{2} \log |V| - \frac{1}{2} \text{tr} [\hat{\Sigma}^{-1}(Y - X\hat{B})^T V^{-1}(Y - X\hat{B})] \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\hat{\Sigma}| - \frac{v}{2} \log |V| \\ &\quad - \frac{1}{2} \text{tr} \left[ \left( \frac{1}{n}(Y - X\hat{B})^T V^{-1}(Y - X\hat{B}) \right)^{-1} (Y - X\hat{B})^T V^{-1}(Y - X\hat{B}) \right] \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\hat{\Sigma}| - \frac{v}{2} \log |V| - \frac{n}{2} \text{tr} [I_v] \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\hat{\Sigma}| - \frac{v}{2} \log |V| - \frac{nv}{2} \end{aligned} \quad (7)$$

which proves the result in (3). Assuming  $V = I_n$ , we have

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^T (Y - X\hat{B}) \quad (8)$$

and

$$\frac{v}{2} \log |V| = \frac{v}{2} \log |I_n| = \frac{v}{2} \log 1 = 0 , \quad (9)$$

such that

$$\text{MLL}(m) = -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{nv}{2} \quad (10)$$

which proves the result in (2). This completes the proof. ■

### 2.2.6 Log-likelihood ratio

**Theorem:** Let  $Y = [y_1, \dots, y_v]$  be an  $n \times v$  data matrix ( $\rightarrow$  I/1.1.5) and consider two general linear models ( $\rightarrow$  III/2.2.1) with design matrices ( $\rightarrow$  III/2.2.1)  $X_1, X_2$  and row-by-row covariance matrices ( $\rightarrow$  III/2.2.1)  $V_1, V_2$ , entailing potentially different regression coefficients ( $\rightarrow$  III/2.2.1)  $B_1, B_2$  and column-by-column covariance matrices ( $\rightarrow$  III/2.2.1)  $\Sigma_1, \Sigma_2$ :

$$\begin{aligned} m_1 : Y &= X_1 B_1 + E_1, \quad E_1 \sim \mathcal{MN}(0, V_1, \Sigma_1) \\ m_2 : Y &= X_2 B_2 + E_2, \quad E_2 \sim \mathcal{MN}(0, V_2, \Sigma_2) . \end{aligned} \quad (1)$$

Then, if the models assume the same covariance matrix across observations, i.e. if  $V_1 = V_2$ , the log-likelihood ratio ( $\rightarrow$  I/4.1.7) for comparing  $m_1$  vs.  $m_2$  is given by

$$\ln \Lambda_{12} = \frac{n}{2} \ln \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} \quad (2)$$

where  $\hat{\Sigma}_1$  and  $\hat{\Sigma}_2$  are the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) of  $\Sigma_1$  and  $\Sigma_2$ .

**Proof:** The likelihood ratio ( $\rightarrow$  I/4.1.6) between two models  $m_1$  and  $m_2$  with model parameters  $\theta_1$  and  $\theta_2$  and parameter spaces  $\Theta_1$  and  $\Theta_2$  is defined as the quotient of their maximized ( $\rightarrow$  I/4.1.3) likelihood functions ( $\rightarrow$  I/5.1.2):

$$\Lambda_{12} = \frac{\max_{\theta_1 \in \Theta_1} p(y|\theta_1, m_1)}{\max_{\theta_2 \in \Theta_2} p(y|\theta_2, m_2)} . \quad (3)$$

Thus, the log-likelihood ratio ( $\rightarrow$  I/4.1.7) is equal to the difference of the maximum log-likelihoods ( $\rightarrow$  I/4.1.4) of the two models:

$$\ln \Lambda_{12} = \ln p(y|\hat{\theta}_1, m_1) - \ln p(y|\hat{\theta}_2, m_2) . \quad (4)$$

The likelihood function ( $\rightarrow$  I/5.1.2) of the general linear model ( $\rightarrow$  III/2.2.1) is a matrix-normal probability density function ( $\rightarrow$  II/5.1.3):

$$\begin{aligned} p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] . \end{aligned} \quad (5)$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is equal to a logarithmized matrix-normal ( $\rightarrow$  II/5.1.1) density ( $\rightarrow$  I/1.7.1):

$$\begin{aligned} \ln p(Y|B, \Sigma) &= \ln \mathcal{MN}(Y; XB, V, \Sigma) \\ &= -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\Sigma| - \frac{v}{2} \ln |V| - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)] . \end{aligned} \quad (6)$$

The maximum likelihood estimates for the general linear model ( $\rightarrow$  III/2.2.4) are given by

$$\begin{aligned} \hat{B} &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) . \end{aligned} \quad (7)$$

such that the last term in the maximum log-likelihood function (6) becomes

$$\begin{aligned}
& \frac{1}{2} \operatorname{tr} \left[ \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \right] \\
&= \frac{1}{2} \operatorname{tr} \left[ \left( \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \right)^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \right] \\
&= \frac{1}{2} \operatorname{tr} \left[ n \left( (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \right)^{-1} \left( (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \right) \right] \\
&= \frac{n}{2} \operatorname{tr} [I_v] \\
&= \frac{nv}{2} .
\end{aligned} \tag{8}$$

Thus, the maximum log-likelihood for the general linear model ( $\rightarrow$  III/2.2.5) is equal to

$$\ln p(Y|\hat{B}, \hat{\Sigma}) = -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{v}{2} \ln |V| - \frac{nv}{2} . \tag{9}$$

Evaluating (9) for  $m_1$  and  $m_2$  and plugging into (4), we obtain:

$$\begin{aligned}
\ln \Lambda_{12} &= \ln p(Y|\hat{B}_1, \hat{\Sigma}_1, m_1) - \ln p(Y|\hat{B}_2, \hat{\Sigma}_2, m_2) \\
&= \left( -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}_1| - \frac{v}{2} \ln |V_1| - \frac{nv}{2} \right) \\
&\quad - \left( -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln |\hat{\Sigma}_2| - \frac{v}{2} \ln |V_2| - \frac{nv}{2} \right) \\
&= -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} - \frac{v}{2} \ln \frac{|V_1|}{|V_2|} .
\end{aligned} \tag{10}$$

Thus, if  $V_1 = V_2$ , such that  $\ln(|V_2|/|V_1|) = \ln(1) = 0$ , the log-likelihood ratio is equal to

$$\ln \Lambda_{12} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} . \tag{11}$$

■

### 2.2.7 Mutual information

**Theorem:** Consider a general linear model ( $\rightarrow$  III/2.2.1)  $m_1$  with  $n \times v$  data matrix  $Y$ ,  $n \times p$  design matrix  $X$  and uncorrelated observations ( $\rightarrow$  III/2.2.1), i.e.  $V = I_n$ ,

$$m_1 : Y = XB + E_1, \quad E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1) , \tag{1}$$

as well as another model  $m_0$  in which  $X$  has no influence on  $Y$ :

$$m_0 : Y = E_0, \quad E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0) . \tag{2}$$

Then, the mutual information ( $\rightarrow$  I/2.4.1) of  $Y$  and  $X$  is equal to

$$I(X, Y) = -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} . \tag{3}$$

**Proof:** The continuous mutual information can be written in terms of marginal and conditional differential entropy ( $\rightarrow$  I/2.4.2) as follows:

$$I(X, Y) = h(Y) - h(Y|X) . \quad (4)$$

The marginal distribution of  $Y$ , unconditional on  $X$ , is given by model  $m_0$

$$Y \sim \mathcal{MN}(0, I_n, \Sigma_0) \quad (5)$$

and the conditional distribution of  $Y$  given  $X$  is given by model  $m_1$

$$Y \sim \mathcal{MN}(XB, I_n, \Sigma_1) . \quad (6)$$

Since  $X$  is constant ( $\rightarrow$  I/1.2.5) and thus only has one possible value ( $\rightarrow$  I/1.1.2), the conditional differential entropy ( $\rightarrow$  I/2.2.7) of  $Y$  given  $X$  is obtained by simply entering  $X$  into the probability distribution ( $\rightarrow$  I/1.5.1) for which the differential entropy ( $\rightarrow$  I/2.2.7) is calculated:

$$\begin{aligned} h(Y|X) &= \int_{x \in \mathcal{X}} p(x) \cdot h(Y|x) \, dx \\ &= p(X) \cdot h(Y|X) \\ &= h[p(Y|X, B, \Sigma_1)] . \end{aligned} \quad (7)$$

The differential entropy of the matrix-normal distribution ( $\rightarrow$  II/5.1.6) is

$$\begin{aligned} X &\sim \mathcal{MN}(M, U, V) \quad \text{where} \quad X \in \mathbb{R}^{n \times p} \\ \Rightarrow \quad h(X) &= \frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |V| + \frac{p}{2} \ln |U| + \frac{np}{2} , \end{aligned} \quad (8)$$

such that the mutual information of  $Y$  and  $X$  becomes

$$\begin{aligned} I(X, Y) &= h[p(Y|\Sigma_0)] - h[p(Y|X, B, \Sigma_1)] \\ &= h[\mathcal{MN}(0, I_n, \Sigma_0)] - h[\mathcal{MN}(XB, I_n, \Sigma_1)] \\ &= \left( \frac{nv}{2} \ln(2\pi) + \frac{n}{2} \ln |\Sigma_0| + \frac{v}{2} \ln |I_n| + \frac{nv}{2} \right) \\ &\quad - \left( \frac{nv}{2} \ln(2\pi) + \frac{n}{2} \ln |\Sigma_1| + \frac{v}{2} \ln |I_n| + \frac{nv}{2} \right) \\ &= \frac{n}{2} \ln |\Sigma_0| - \frac{n}{2} \ln |\Sigma_1| \\ &= \frac{n}{2} \ln \frac{|\Sigma_0|}{|\Sigma_1|} \\ &= -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} . \end{aligned} \quad (9)$$

■

### 2.2.8 Log-likelihood ratio and estimated mutual information

**Theorem:** Consider a general linear model ( $\rightarrow$  III/2.2.1)  $m_1$  with  $n \times v$  data matrix  $Y$ ,  $n \times p$  design matrix  $X$  and uncorrelated observations ( $\rightarrow$  III/2.2.1), i.e.  $V = I_n$ ,

$$m_1 : Y = XB + E_1, \quad E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1), \quad (1)$$

as well as another model  $m_0$  in which  $X$  has no influence on  $Y$ :

$$m_0 : Y = E_0, \quad E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0). \quad (2)$$

Then, the log-likelihood ratio ( $\rightarrow$  I/4.1.7) of  $m_1$  vs.  $m_0$  is equal to the estimated ( $\rightarrow$  I/3.1.1) mutual information ( $\rightarrow$  I/2.4.1) of  $X$  and  $Y$ :

$$\ln \Lambda_{10} = \hat{I}(X, Y). \quad (3)$$

**Proof:** The maximum likelihood estimates for a general linear model ( $\rightarrow$  III/2.2.4) are

$$\begin{aligned} \hat{B} &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}), \end{aligned} \quad (4)$$

such that, for the two models, the maximum likelihood estimates ( $\rightarrow$  I/4.1.3) are:

$$\begin{aligned} \hat{\Sigma}_1 &= \frac{1}{n} (Y - X \hat{B})^T (Y - X \hat{B}) \quad \text{with} \quad \hat{B} = (X^T X)^{-1} X^T Y \quad \text{and} \\ \hat{\Sigma}_0 &= \frac{1}{n} Y^T Y. \end{aligned} \quad (5)$$

The log-likelihood ratio for two general linear models ( $\rightarrow$  III/2.2.6) is

$$\ln \Lambda_{12} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|}, \quad (6)$$

such that in the present case, we have:

$$\ln \Lambda_{10} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|}. \quad (7)$$

The mutual information for the general linear model ( $\rightarrow$  III/2.2.7) is

$$I(X, Y) = -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|}, \quad (8)$$

such that with (5), the estimated mutual information is:

$$\hat{I}(X, Y) = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|}, \quad (9)$$

Together, (7) and (9) show that

$$\ln \Lambda_{10} = \hat{I}(X, Y). \quad (10)$$



**Sources:**

- Friston K, Chu C, Mourão-Miranda J, Hulme O, Rees G, Penny W, Ashburner J (2008): “Bayesian decoding of brain images”; in: *NeuroImage*, vol. 39, pp. 181-205, eq. 6; URL: <https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203>; DOI: 10.1016/j.neuroimage.2007.08.013.

**2.3 Transformed general linear model****2.3.1 Definition**

**Definition:** Let there be two general linear models ( $\rightarrow$  III/2.2.1) of measured data  $Y \in \mathbb{R}^{n \times v}$  using design matrices ( $\rightarrow$  III/2.2.1)  $X \in \mathbb{R}^{n \times p}$  and  $X_t \in \mathbb{R}^{n \times t}$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

$$Y = X_t \Gamma + E_t, \quad E_t \sim \mathcal{MN}(0, V, \Sigma_t) \quad (2)$$

and assume that  $X_t$  can be transformed into  $X$  using a transformation matrix  $T \in \mathbb{R}^{t \times p}$

$$X = X_t T \quad (3)$$

where  $p < t$  and  $X$ ,  $X_t$  and  $T$  have full ranks  $\text{rk}(X) = p$ ,  $\text{rk}(X_t) = t$  and  $\text{rk}(T) = p$ .

Then, a linear model ( $\rightarrow$  III/2.2.1) of the parameter estimates from (2), under the assumption of (1), is called a transformed general linear model.

**Sources:**

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix A; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

**2.3.2 Derivation of the distribution**

**Theorem:** Let there be two general linear models ( $\rightarrow$  III/2.2.1) of measured data  $Y$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

$$Y = X_t \Gamma + E_t, \quad E_t \sim \mathcal{MN}(0, V, \Sigma_t) \quad (2)$$

and a matrix  $T$  transforming  $X_t$  into  $X$ :

$$X = X_t T. \quad (3)$$

Then, the transformed general linear model ( $\rightarrow$  III/2.3.1) is given by

$$\hat{\Gamma} = TB + H, \quad H \sim \mathcal{MN}(0, U, \Sigma) \quad (4)$$

where the covariance across rows ( $\rightarrow$  II/5.1.1) is  $U = (X_t^T V^{-1} X_t)^{-1}$ .

**Proof:** The linear transformation theorem for the matrix-normal distribution ( $\rightarrow$  II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \Rightarrow Y = AXB + C \sim \mathcal{MN}(AMB + C, AU A^T, B^T V B) . \quad (5)$$

The weighted least squares parameter estimates ( $\rightarrow$  III/2.2.3) for (2) are given by

$$\hat{\Gamma} = (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} Y . \quad (6)$$

Using (1) and (5), the distribution of  $Y$  is

$$Y \sim \mathcal{MN}(XB, V, \Sigma) \quad (7)$$

Combining (6) with (7), the distribution of  $\hat{\Gamma}$  is

$$\begin{aligned} \hat{\Gamma} &\sim \mathcal{MN} \left( [(X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1}] XB, [(X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1}] V [V^{-1} X_t (X_t^T V^{-1} X_t)^{-1}] , \Sigma \right) \\ &\sim \mathcal{MN} \left( (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} X_t TB, (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} X_t (X_t^T V^{-1} X_t)^{-1}, \Sigma \right) \\ &\sim \mathcal{MN} \left( TB, (X_t^T V^{-1} X_t)^{-1}, \Sigma \right) . \end{aligned} \quad (8)$$

This can also be written as

$$\hat{\Gamma} = TB + H, \quad H \sim \mathcal{MN}(0, (X_t^T V^{-1} X_t)^{-1}, \Sigma) \quad (9)$$

which is equivalent to (4). ■

#### Sources:

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix A, Theorem 1; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

### 2.3.3 Equivalence of parameter estimates

**Theorem:** Let there be a general linear model ( $\rightarrow$  III/2.2.1)

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

and the transformed general linear model ( $\rightarrow$  III/2.3.1)

$$\hat{\Gamma} = TB + H, \quad H \sim \mathcal{MN}(0, U, \Sigma) \quad (2)$$

which are linked to each other ( $\rightarrow$  III/2.3.2) via

$$\hat{\Gamma} = (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} Y \quad (3)$$

and

$$X = X_t T . \quad (4)$$

Then, the parameter estimates for  $B$  from (1) and (2) are equivalent.

**Proof:** The weighted least squares parameter estimates ( $\rightarrow$  III/2.2.3) for (1) are given by

$$\hat{B} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (5)$$

and the weighted least squares parameter estimates ( $\rightarrow$  III/2.2.3) for (2) are given by

$$\hat{B} = (T^T U^{-1} T)^{-1} T^T U^{-1} \hat{\Gamma} . \quad (6)$$

The covariance across rows for the transformed general linear model ( $\rightarrow$  III/2.3.2) is equal to

$$U = (X_t^T V^{-1} X_t)^{-1} . \quad (7)$$

Applying (7), (4) and (3), the estimates in (6) can be developed into

$$\begin{aligned} \hat{B} &\stackrel{(6)}{=} (T^T U^{-1} T)^{-1} T^T U^{-1} \hat{\Gamma} \\ &\stackrel{(7)}{=} (T^T [X_t^T V^{-1} X_t] T)^{-1} T^T [X_t^T V^{-1} X_t] \hat{\Gamma} \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} T^T X_t^T V^{-1} X_t \hat{\Gamma} \\ &\stackrel{(3)}{=} (X^T V^{-1} X)^{-1} T^T X_t^T V^{-1} X_t [(X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} Y] \\ &= (X^T V^{-1} X)^{-1} T^T X_t^T V^{-1} Y \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} Y \end{aligned} \quad (8)$$

which is equivalent to the estimates in (5). ■

#### Sources:

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix A, Theorem 2; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

## 2.4 Inverse general linear model

### 2.4.1 Definition

**Definition:** Let there be a general linear model ( $\rightarrow$  III/2.2.1) of measured data  $Y \in \mathbb{R}^{n \times v}$  in terms of the design matrix ( $\rightarrow$  III/2.2.1)  $X \in \mathbb{R}^{n \times p}$ :

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) . \quad (1)$$

Then, a linear model ( $\rightarrow$  III/2.2.1) of  $X$  in terms of  $Y$ , under the assumption of (1), is called an inverse general linear model.

#### Sources:

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix C; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

### 2.4.2 Derivation of the distribution

**Theorem:** Let there be a general linear model ( $\rightarrow$  III/2.2.1) of  $Y \in \mathbb{R}^{n \times v}$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) . \quad (1)$$

Then, the inverse general linear model ( $\rightarrow$  III/2.4.1) of  $X \in \mathbb{R}^{n \times p}$  is given by

$$X = YW + N, \quad N \sim \mathcal{MN}(0, V, \Sigma_x) \quad (2)$$

where  $W \in \mathbb{R}^{v \times p}$  is a matrix, such that  $BW = I_p$ , and the covariance across columns ( $\rightarrow$  II/5.1.1) is  $\Sigma_x = W^T \Sigma W$ .

**Proof:** The linear transformation theorem for the matrix-normal distribution ( $\rightarrow$  II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AU A^T, B^T V B) . \quad (3)$$

The matrix  $W$  exists, if the rows of  $B \in \mathbb{R}^{p \times v}$  are linearly independent, such that  $\text{rk}(B) = p$ . Then, right-multiplying the model (1) with  $W$  and applying (3) yields

$$YW = XBW + EW, \quad EW \sim \mathcal{MN}(0, V, W^T \Sigma W) . \quad (4)$$

Employing  $BW = I_p$  and rearranging, we have

$$X = YW - EW, \quad EW \sim \mathcal{MN}(0, V, W^T \Sigma W) . \quad (5)$$

Substituting  $N = -EW$ , we get

$$X = YW + N, \quad N \sim \mathcal{MN}(0, V, W^T \Sigma W) \quad (6)$$

which is equivalent to (2). ■

#### Sources:

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix C, Theorem 4; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

### 2.4.3 Best linear unbiased estimator

**Theorem:** Let there be a general linear model ( $\rightarrow$  III/2.2.1) of  $Y \in \mathbb{R}^{n \times v}$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

implying the inverse general linear model ( $\rightarrow$  III/2.4.2) of  $X \in \mathbb{R}^{n \times p}$

$$X = YW + N, \quad N \sim \mathcal{MN}(0, V, \Sigma_x) . \quad (2)$$

where

$$BW = I_p \quad \text{and} \quad \Sigma_x = W^T \Sigma W . \quad (3)$$

Then, the weighted least squares solution ( $\rightarrow$  III/2.2.3) for  $W$  is the best linear unbiased estimator of  $W$ .

**Proof:** The linear transformation theorem for the matrix-normal distribution ( $\rightarrow$  II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \Rightarrow Y = AXB + C \sim \mathcal{MN}(AMB + C, AU A^T, B^T V B). \quad (4)$$

The weighted least squares parameter estimates ( $\rightarrow$  III/2.2.3) for (2) are given by

$$\hat{W} = (Y^T V^{-1} Y)^{-1} Y^T V^{-1} X. \quad (5)$$

The best linear unbiased estimator  $\hat{\theta}$  of a certain quantity  $\theta$  estimated from measured data ( $\rightarrow$  I/1.1.5)  $y$  is 1) an estimator resulting from a linear operation  $f(y)$ , 2) whose expected value is equal to  $\theta$  and 3) which has, among those satisfying 1) and 2), the minimum variance ( $\rightarrow$  I/1.13.1).

1) First,  $\hat{W}$  is a linear estimator, because it is of the form  $\tilde{W} = MX$  where  $M$  is an arbitrary  $v \times n$  matrix.

2) Second,  $\hat{W}$  is an unbiased estimator, if  $\langle \hat{W} \rangle = W$ . By applying (4) to (2), the distribution of  $\tilde{W}$  is

$$\tilde{W} = MX \sim \mathcal{MN}(MYW, MVM^T, \Sigma_x) \quad (6)$$

which requires ( $\rightarrow$  II/5.1.4) that  $MY = I_v$ . This is fulfilled by any matrix

$$M = (Y^T V^{-1} Y)^{-1} Y^T V^{-1} + D \quad (7)$$

where  $D$  is a  $v \times n$  matrix which satisfies  $DY = 0$ .

3) Third, the best linear unbiased estimator is the one with minimum variance ( $\rightarrow$  I/1.13.1), i.e. the one that minimizes the expected Frobenius norm

$$\text{Var}(\tilde{W}) = \left\langle \text{tr} \left[ (\tilde{W} - W)^T (\tilde{W} - W) \right] \right\rangle. \quad (8)$$

Using the matrix-normal distribution ( $\rightarrow$  II/5.1.1) of  $\tilde{W}$  from (6)

$$(\tilde{W} - W) \sim \mathcal{MN}(0, MVM^T, \Sigma_x) \quad (9)$$

and the property of the Wishart distribution ( $\rightarrow$  II/5.2.1)

$$X \sim \mathcal{MN}(0, U, V) \Rightarrow \langle XX^T \rangle = \text{tr}(V) U, \quad (10)$$

this variance ( $\rightarrow$  I/1.13.1) can be evaluated as a function of  $M$ :

$$\begin{aligned} \text{Var}[\tilde{W}(M)] &\stackrel{(8)}{=} \left\langle \text{tr} \left[ (\tilde{W} - W)^T (\tilde{W} - W) \right] \right\rangle \\ &= \left\langle \text{tr} \left[ (\tilde{W} - W)(\tilde{W} - W)^T \right] \right\rangle \\ &= \text{tr} \left[ \left\langle (\tilde{W} - W)(\tilde{W} - W)^T \right\rangle \right] \\ &\stackrel{(10)}{=} \text{tr} [\text{tr}(\Sigma_x) MVM^T] \\ &= \text{tr}(\Sigma_x) \text{tr}(MVM^T). \end{aligned} \quad (11)$$

As a function of  $D$  and using  $DY = 0$ , it becomes:

$$\begin{aligned}
 \text{Var} \left[ \tilde{W}(D) \right] &\stackrel{(7)}{=} \text{tr}(\Sigma_x) \text{tr} \left[ ((Y^T V^{-1} Y)^{-1} Y^T V^{-1} + D) V ((Y^T V^{-1} Y)^{-1} Y^T V^{-1} + D)^T \right] \\
 &= \text{tr}(\Sigma_x) \text{tr} \left[ (Y^T V^{-1} Y)^{-1} Y^T V^{-1} V V^{-1} Y (Y^T V^{-1} Y)^{-1} + \right. \\
 &\quad \left. (Y^T V^{-1} Y)^{-1} Y^T V^{-1} V D^T + D V V^{-1} Y (Y^T V^{-1} Y)^{-1} + D V D^T \right] \\
 &= \text{tr}(\Sigma_x) \left[ \text{tr}((Y^T V^{-1} Y)^{-1}) + \text{tr}(D V D^T) \right].
 \end{aligned} \tag{12}$$

Since  $D V D^T$  is a positive-semidefinite matrix, all its eigenvalues are non-negative. Because the trace of a square matrix is the sum of its eigenvalues, the minimum variance is achieved by  $D = 0$ , thus producing  $\hat{W}$  as in (5). ■

#### Sources:

- Soch J, Allefeld C, Haynes JD (2020): “Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding”; in: *NeuroImage*, vol. 209, art. 116449, Appendix C, Theorem 5; URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310407>; DOI: 10.1016/j.neuroimage.2019.116449.

#### 2.4.4 Equivalence of log-likelihood ratios

**Theorem:** Consider two general linear models ( $\rightarrow$  III/2.2.1)

$$\begin{aligned}
 m_1^{(Y)} : Y &= X B + E_1, \quad E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1^{(Y)}) \\
 m_0^{(Y)} : Y &= E_0, \quad E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0^{(Y)})
 \end{aligned} \tag{1}$$

and two inverse general linear models ( $\rightarrow$  III/2.4.1)

$$\begin{aligned}
 m_1^{(X)} : X &= Y W + N_1, \quad N_1 \sim \mathcal{MN}(0, I_n, \Sigma_1^{(X)}) \\
 m_0^{(X)} : X &= N_0, \quad N_0 \sim \mathcal{MN}(0, I_n, \Sigma_0^{(X)})
 \end{aligned} \tag{2}$$

where  $Y \in \mathbb{R}^{n \times v}$  and  $X \in \mathbb{R}^{n \times p}$  are data matrices, such that  $n > v$  and  $n > p$ . Then, the log-likelihood ratio ( $\rightarrow$  I/4.1.7) comparing the forward models ( $\rightarrow$  III/2.2.1) is equivalent to the log-likelihood ratio ( $\rightarrow$  I/4.1.7) comparing the backward models ( $\rightarrow$  III/2.4.1):

$$\ln \Lambda_{10}^{(Y)} = \ln \Lambda_{10}^{(X)}. \tag{3}$$

**Proof:** The maximum likelihood estimates for the general linear models ( $\rightarrow$  III/2.2.4) are

$$\begin{aligned}
 \hat{\Sigma}_1^{(Y)} &= \frac{1}{n} (Y - X \hat{B})^T (Y - X \hat{B}) \quad \text{with} \quad \hat{B} = (X^T X)^{-1} X^T Y \quad \text{and} \\
 \hat{\Sigma}_0^{(Y)} &= \frac{1}{n} Y^T Y
 \end{aligned} \tag{4}$$

as well as

$$\begin{aligned}\hat{\Sigma}_1^{(X)} &= \frac{1}{n}(X - Y\hat{W})^T(X - Y\hat{W}) \quad \text{with} \quad \hat{W} = (Y^T Y)^{-1}Y^T X \quad \text{and} \\ \hat{\Sigma}_0^{(X)} &= \frac{1}{n}X^T X.\end{aligned}\tag{5}$$

The likelihood ratio for two general linear models ( $\rightarrow$  III/2.2.6)  $m_1$  and  $m_2$  is:

$$\begin{aligned}\ln \Lambda_{12} &= -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \\ &= -\frac{n}{2} \ln \left( |\hat{\Sigma}_2^{-1}| |\hat{\Sigma}_1| \right) \\ &= -\frac{n}{2} \ln |\hat{\Sigma}_2^{-1} \hat{\Sigma}_1|.\end{aligned}\tag{6}$$

Thus, with (4), the log-likelihood ratio ( $\rightarrow$  I/4.1.7) of  $m_1^{(Y)}$  vs.  $m_0^{(Y)}$  is given as

$$\begin{aligned}\ln \Lambda_Y &= \ln \Lambda_{10}^{(Y)} \stackrel{(6)}{=} -\frac{n}{2} \ln \left| \left( \hat{\Sigma}_0^{(Y)} \right)^{-1} \hat{\Sigma}_1^{(Y)} \right| \\ &\stackrel{(4)}{=} -\frac{n}{2} \ln \left| \left( \frac{1}{n} Y^T Y \right)^{-1} \frac{1}{n} (Y - X\hat{B})^T (Y - X\hat{B}) \right| \\ &= -\frac{n}{2} \ln \left| (Y^T Y)^{-1} \left( Y^T Y - 2Y^T X\hat{B} + \hat{B}^T X^T X\hat{B} \right) \right| \\ &= -\frac{n}{2} \ln \left| \left( (Y^T Y)^{-1} Y^T Y - 2(Y^T Y)^{-1} Y^T X\hat{B} + (Y^T Y)^{-1} \hat{B}^T X^T X\hat{B} \right) \right| \\ &\stackrel{(4)}{=} -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + (Y^T Y)^{-1} Y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T Y \right| \\ &= -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + (Y^T Y)^{-1} Y^T X (X^T X)^{-1} X^T Y \right| \\ &= -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + \hat{W}\hat{B} \right| \\ &= -\frac{n}{2} \ln \left| I_v - \hat{W}\hat{B} \right|.\end{aligned}\tag{7}$$

Similarly, with (5), the log-likelihood ratio ( $\rightarrow$  I/4.1.7) of  $m_1^{(X)}$  vs.  $m_0^{(X)}$  becomes

$$\begin{aligned}
\ln \Lambda_X &= \ln \Lambda_{10}^{(X)} \stackrel{(6)}{=} -\frac{n}{2} \ln \left| \left( \hat{\Sigma}_0^{(X)} \right)^{-1} \hat{\Sigma}_1^{(X)} \right| \\
&\stackrel{(5)}{=} -\frac{n}{2} \ln \left| \left( \frac{1}{n} X^T X \right)^{-1} \frac{1}{n} (X - Y\hat{W})^T (X - Y\hat{W}) \right| \\
&= -\frac{n}{2} \ln \left| (X^T X)^{-1} \left( X^T X - 2X^T Y\hat{W} + \hat{W}^T Y^T Y \hat{W} \right) \right| \\
&= -\frac{n}{2} \ln \left| \left( (X^T X)^{-1} X^T X - 2(X^T X)^{-1} X^T Y\hat{W} + (X^T X)^{-1} \hat{W}^T Y^T Y \hat{W} \right) \right| \quad (8) \\
&\stackrel{(5)}{=} -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + (X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T Y (Y^T Y)^{-1} Y^T X \right| \\
&= -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + (X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T X \right| \\
&= -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + \hat{B}\hat{W} \right| \\
&= -\frac{n}{2} \ln \left| I_p - \hat{B}\hat{W} \right|.
\end{aligned}$$

Sylvester's determinant theorem (also known as the “Weinstein–Aronszajn identity”) states that, for two matrices  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{n \times m}$ , the following identity holds:

$$|I_m + AB| = |I_n + BA|. \quad (9)$$

Since  $\hat{B} \in \mathbb{R}^{p \times v}$  and  $(-\hat{W}) \in \mathbb{R}^{v \times p}$ , it follows that

$$|I_p - \hat{B}\hat{W}| = |I_v - \hat{W}\hat{B}| \quad (10)$$

and thus, we finally have:

$$\ln \Lambda_Y = -\frac{n}{2} \ln |I_v - \hat{W}\hat{B}| = -\frac{n}{2} \ln |I_p - \hat{B}\hat{W}| = \ln \Lambda_X. \quad (11)$$

■

#### Sources:

- Friston K, Chu C, Mourão-Miranda J, Hulme O, Rees G, Penny W, Ashburner J (2008): “Bayesian decoding of brain images”; in: *NeuroImage*, vol. 39, pp. 181-205, p. 183; URL: <https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203>; DOI: 10.1016/j.neuroimage.2007.08.013.
- Wikipedia (2024): “Weinstein–Aronszajn identity”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-28; URL: [https://en.wikipedia.org/wiki/Weinstein%E2%80%93Aronszajn\\_identity](https://en.wikipedia.org/wiki/Weinstein%E2%80%93Aronszajn_identity).

#### 2.4.5 Corresponding forward model

**Definition:** Let there be observations ( $\rightarrow$  I/1.1.5)  $Y \in \mathbb{R}^{n \times v}$  and  $X \in \mathbb{R}^{n \times p}$  and consider a weight matrix  $W = f(Y, X) \in \mathbb{R}^{v \times p}$  estimated ( $\rightarrow$  I/3.1.1) from  $Y$  and  $X$ , such that right-multiplying  $Y$  with the weight matrix gives an estimate ( $\rightarrow$  I/3.1.1) or prediction of  $X$ :

$$\hat{X} = YW. \quad (1)$$

Given that the columns of  $\hat{X}$  are linearly independent, then

$$Y = \hat{X}A^T + E \quad \text{with} \quad \hat{X}^T E = 0 \quad (2)$$



is called the corresponding forward model relative to the weight matrix  $W$ .

**Sources:**

- Haufe S, Meinecke F, Gorgen K, Dhne S, Haynes JD, Blankertz B, Biemann F (2014): “On the interpretation of weight vectors of linear models in multivariate neuroimaging”; in: *NeuroImage*, vol. 87, pp. 96–110, eq. 3; URL: <https://www.sciencedirect.com/science/article/pii/S1053811913010914>; DOI: 10.1016/j.neuroimage.2013.10.067.

### 2.4.6 Derivation of parameters

**Theorem:** Let there be observations ( $\rightarrow$  I/1.1.5)  $Y \in \mathbb{R}^{n \times v}$  and  $X \in \mathbb{R}^{n \times p}$  and consider a weight matrix  $W = f(Y, X) \in \mathbb{R}^{v \times p}$  predicting  $X$  from  $Y$ :

$$\hat{X} = YW . \quad (1)$$

Then, the parameter matrix of the corresponding forward model ( $\rightarrow$  III/2.4.5) is equal to

$$A = \Sigma_y W \Sigma_x^{-1} \quad (2)$$

with the “sample covariances ( $\rightarrow$  I/1.15.2)”

$$\begin{aligned} \Sigma_x &= \hat{X}^T \hat{X} \\ \Sigma_y &= Y^T Y . \end{aligned} \quad (3)$$

**Proof:** The corresponding forward model ( $\rightarrow$  III/2.4.5) is given by

$$Y = \hat{X} A^T + E , \quad (4)$$

subject to the constraint that predicted  $X$  and errors  $E$  are uncorrelated ( $\rightarrow$  I/1.17.1):

$$\hat{X}^T E = 0 . \quad (5)$$

With that, we can directly derive the parameter matrix  $A$ :

$$\begin{aligned} Y &\stackrel{(4)}{=} \hat{X} A^T + E \\ \hat{X} A^T &= Y - E \\ \hat{X}^T \hat{X} A^T &= \hat{X}^T (Y - E) \\ \hat{X}^T \hat{X} A^T &= \hat{X}^T Y - \hat{X}^T E \\ \hat{X}^T \hat{X} A^T &\stackrel{(5)}{=} \hat{X}^T Y \\ \hat{X}^T \hat{X} A^T &\stackrel{(1)}{=} W^T Y^T Y \\ \Sigma_x A^T &\stackrel{(3)}{=} W^T \Sigma_y \\ A^T &= \Sigma_x^{-1} W^T \Sigma_y \\ A &= \Sigma_y W \Sigma_x^{-1} . \end{aligned} \quad (6)$$

■

**Sources:**

- Haufe S, Meinecke F, Gorgen K, Dahne S, Haynes JD, Blankertz B, Biemann F (2014): “On the interpretation of weight vectors of linear models in multivariate neuroimaging”; in: *NeuroImage*, vol. 87, pp. 96–110, Theorem 1; URL: <https://www.sciencedirect.com/science/article/pii/S1053811913010914>; DOI: 10.1016/j.neuroimage.2013.10.067.

**2.4.7 Proof of existence**

**Theorem:** Let there be observations ( $\rightarrow$  I/1.1.5)  $Y \in \mathbb{R}^{n \times v}$  and  $X \in \mathbb{R}^{n \times p}$  and consider a weight matrix  $W = f(Y, X) \in \mathbb{R}^{v \times p}$  predicting  $X$  from  $Y$ :

$$\hat{X} = YW. \quad (1)$$

Then, there exists a corresponding forward model ( $\rightarrow$  III/2.4.5).

**Proof:** The corresponding forward model ( $\rightarrow$  III/2.4.5) is defined as

$$Y = \hat{X}A^T + E \quad \text{with} \quad \hat{X}^T E = 0 \quad (2)$$

and the parameters of the corresponding forward model ( $\rightarrow$  III/2.4.6) are equal to

$$A = \Sigma_y W \Sigma_x^{-1} \quad \text{where} \quad \Sigma_x = \hat{X}^T \hat{X} \quad \text{and} \quad \Sigma_y = Y^T Y. \quad (3)$$

1) Because the columns of  $\hat{X}$  are assumed to be linearly independent by definition of the corresponding forward model ( $\rightarrow$  III/2.4.5), the matrix  $\Sigma_x = \hat{X}^T \hat{X}$  is invertible, such that  $A$  in (3) is well-defined.

2) Moreover, the solution for the matrix  $A$  satisfies the constraint of the corresponding forward model ( $\rightarrow$  III/2.4.5) for predicted  $X$  and errors  $E$  to be uncorrelated ( $\rightarrow$  I/1.17.1) which can be shown as follows:

$$\begin{aligned} \hat{X}^T E &\stackrel{(2)}{=} \hat{X}^T (Y - \hat{X}A^T) \\ &\stackrel{(3)}{=} \hat{X}^T (Y - \hat{X} \Sigma_x^{-1} W^T \Sigma_y) \\ &= \hat{X}^T Y - \hat{X}^T \hat{X} \Sigma_x^{-1} W^T \Sigma_y \\ &\stackrel{(3)}{=} \hat{X}^T Y - \hat{X}^T \hat{X} (\hat{X}^T \hat{X})^{-1} W^T (Y^T Y) \\ &\stackrel{(1)}{=} (YW)^T Y - W^T (Y^T Y) \\ &= W^T Y^T Y - W^T Y^T Y \\ &= 0. \end{aligned} \quad (4)$$

This completes the proof. ■

**Sources:**

- Haufe S, Meinecke F, Gorgen K, Dahne S, Haynes JD, Blankertz B, Biemann F (2014): “On the interpretation of weight vectors of linear models in multivariate neuroimaging”; in: *NeuroImage*, vol. 87, pp. 96–110, Appendix B; URL: <https://www.sciencedirect.com/science/article/pii/S1053811913010914>; DOI: 10.1016/j.neuroimage.2013.10.067.

## 2.5 Multivariate Bayesian linear regression

### 2.5.1 Conjugate prior distribution

**Theorem:** Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model ( $\rightarrow$  III/2.2.1) with measured  $n \times v$  data matrix  $Y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure ( $\rightarrow$  II/5.1.1)  $V$  as well as unknown  $p \times v$  regression coefficients  $B$  and unknown  $v \times v$  noise covariance ( $\rightarrow$  II/5.1.1)  $\Sigma$ .

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for this model is a normal-Wishart distribution ( $\rightarrow$  II/5.3.1)

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \quad (2)$$

where  $T = \Sigma^{-1}$  is the inverse noise covariance ( $\rightarrow$  I/1.16.1) or noise precision matrix ( $\rightarrow$  I/1.16.12).

**Proof:** By definition, a conjugate prior ( $\rightarrow$  I/5.2.5) is a prior distribution ( $\rightarrow$  I/5.1.3) that, when combined with the likelihood function ( $\rightarrow$  I/5.1.2), leads to a posterior distribution ( $\rightarrow$  I/5.1.9) that belongs to the same family of probability distributions ( $\rightarrow$  I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (3)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P^{-1}, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[ -\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \quad (4)$$

using the  $v \times v$  precision matrix ( $\rightarrow$  I/1.16.12)  $T = \Sigma^{-1}$  and the  $n \times n$  precision matrix ( $\rightarrow$  I/1.16.12)  $P = V^{-1}$ .

Separating constant and variable terms, we have:

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right]. \quad (5)$$

Expanding the product in the exponent, we have:

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B]) \right]. \quad (6)$$

Completing the square over  $B$ , finally gives

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ (B - \tilde{X}Y)^T X^T P X (B - \tilde{X}Y) - Y^T Q Y + Y^T P Y \right] \right) \right] \quad (7)$$

where  $\tilde{X} = (X^T P X)^{-1} X^T P$  and  $Q = \tilde{X}^T (X^T P X) \tilde{X}$ .

In other words, the likelihood function ( $\rightarrow$  I/5.1.2) is proportional to a power of the determinant of  $T$ , times an exponential of the trace of  $T$  and an exponential of the trace of a squared form of  $B$ , weighted by  $T$ :

$$p(Y|B, T) \propto |T|^{n/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T Q Y]) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ (B - \tilde{X}Y)^T X^T P X (B - \tilde{X}Y) \right] \right) \right] . \quad (8)$$

The same is true for a normal-Wishart distribution ( $\rightarrow$  II/5.3.1) over  $B$  and  $T$

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \quad (9)$$

the probability density function of which ( $\rightarrow$  II/5.3.2)

$$p(B, T) = \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[ -\frac{1}{2} \text{tr} (T (B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \frac{1}{\Gamma_v(\frac{\nu_0}{2})} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega_0 T) \right] \quad (10)$$

exhibits the same proportionality

$$p(B, T) \propto |T|^{(\nu_0 + p - v - 1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (T \Omega_0) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [(B - M_0)^T \Lambda_0 (B - M_0)]) \right] \quad (11)$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Wikipedia (2020): “Bayesian multivariate linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: [https://en.wikipedia.org/wiki/Bayesian\\_multivariate\\_linear\\_regression#Conjugate\\_prior\\_distribution](https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Conjugate_prior_distribution).

### 2.5.2 Posterior distribution

**Theorem:** Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model ( $\rightarrow$  III/2.2.1) with measured  $n \times v$  data matrix  $Y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure ( $\rightarrow$  II/5.1.1)  $V$  as well as unknown  $p \times v$  regression coefficients  $B$  and unknown  $v \times v$  noise covariance ( $\rightarrow$  II/5.1.1)  $\Sigma$ . Moreover, assume a normal-Wishart prior distribution ( $\rightarrow$  III/2.5.1) over the model parameters  $B$  and  $T = \Sigma^{-1}$ :

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a normal-Wishart distribution ( $\rightarrow$  II/5.3.1)

$$p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1}(X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n . \end{aligned} \quad (4)$$

**Proof:** According to Bayes' theorem ( $\rightarrow$  I/5.3.2), the posterior distribution ( $\rightarrow$  I/5.1.9) is given by

$$p(B, T|Y) = \frac{p(Y|B, T) p(B, T)}{p(Y)} . \quad (5)$$

Since  $p(Y)$  is just a normalization factor, the posterior is proportional ( $\rightarrow$  I/5.1.10) to the numerator:

$$p(B, T|Y) \propto p(Y|B, T) p(B, T) = p(Y, B, T) . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[ -\frac{1}{2} \text{tr} (T(Y - XB)^T P (Y - XB)) \right] \quad (8)$$

using the  $v \times v$  precision matrix ( $\rightarrow$  I/1.16.12)  $T = \Sigma^{-1}$  and the  $n \times n$  precision matrix ( $\rightarrow$  I/1.16.12)  $P = V^{-1}$ .

Combining the likelihood function ( $\rightarrow$  I/5.1.2) (8) with the prior distribution ( $\rightarrow$  I/5.1.3) (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(Y, B, T) &= p(Y|B, T) p(B, T) \\ &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[ -\frac{1}{2} \text{tr} (T(Y - XB)^T P (Y - XB)) \right] \cdot \\ &\quad \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[ -\frac{1}{2} \text{tr} (T(B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \\ &\quad \frac{1}{\Gamma_v(\frac{\nu_0}{2})} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega_0 T) \right] . \end{aligned} \quad (9)$$

Collecting identical variables gives:

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_0 T) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ (Y - XB)^T P (Y - XB) + (B - M_0)^T \Lambda_0 (B - M_0) \right] \right) \right]. \quad (10)$$

Expanding the products in the exponent gives:

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_0 T) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B + B^T \Lambda_0 B - B^T \Lambda_0 M_0 - M_0^T \Lambda_0 B + M_0^T \Lambda_0 \mu_0 \right] \right) \right]. \quad (11)$$

Completing the square over  $B$ , we finally have

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_0 T) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ (B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n) \right] \right) \right]. \quad (12)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0. \end{aligned} \quad (13)$$

Ergo, the joint likelihood is proportional to

$$p(Y, B, T) \propto |T|^{p/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} \left( T \left[ (B - M_n)^T \Lambda_n (B - M_n) \right] \right) \right] \cdot |T|^{(\nu_n - v - 1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr}(\Omega_n T) \right] \quad (14)$$

with the posterior hyperparameters ( $\rightarrow$  I/5.1.9)

$$\begin{aligned} \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n. \end{aligned} \quad (15)$$

From the term in (14), we can isolate the posterior distribution over  $B$  given  $T$ :

$$p(B|T, Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}). \quad (16)$$

From the remaining term, we can isolate the posterior distribution over  $T$ :

$$p(T|Y) = \mathcal{W}(T; \Omega_n^{-1}, \nu_n). \quad (17)$$

Together, (16) and (17) constitute the joint ( $\rightarrow$  I/1.3.2) posterior distribution ( $\rightarrow$  I/5.1.9) of  $B$  and  $T$ .

**Sources:**

- Wikipedia (2020): “Bayesian multivariate linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: [https://en.wikipedia.org/wiki/Bayesian\\_multivariate\\_linear\\_regression#Posterior\\_distribution](https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Posterior_distribution).

**2.5.3 Log model evidence**

**Theorem:** Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model ( $\rightarrow$  III/2.2.1) with measured  $n \times v$  data matrix  $Y$ , known  $n \times p$  design matrix  $X$ , known  $n \times n$  covariance structure ( $\rightarrow$  II/5.1.1)  $V$  as well as unknown  $p \times v$  regression coefficients  $B$  and unknown  $v \times v$  noise covariance ( $\rightarrow$  II/5.1.1)  $\Sigma$ . Moreover, assume a normal-Wishart prior distribution ( $\rightarrow$  III/2.5.1) over the model parameters  $B$  and  $T = \Sigma^{-1}$ :

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned} \log p(Y|m) = & \frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ & \frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v \left( \frac{\nu_n}{2} \right) - \log \Gamma_v \left( \frac{\nu_0}{2} \right) \end{aligned} \quad (3)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n . \end{aligned} \quad (4)$$

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the model evidence ( $\rightarrow$  I/5.1.15) for this model is:

$$p(Y|m) = \iint p(Y|B, T) p(B, T) dB dT . \quad (5)$$

According to the law of conditional probability ( $\rightarrow$  I/1.3.4), the integrand is equivalent to the joint likelihood ( $\rightarrow$  I/5.1.7):

$$p(Y|m) = \iint p(Y, B, T) dB dT . \quad (6)$$

Equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp \left[ -\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[ -\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \quad (8)$$

using the  $v \times v$  precision matrix ( $\rightarrow$  I/1.16.12)  $T = \Sigma^{-1}$  and the  $n \times n$  precision matrix ( $\rightarrow$  I/1.16.12)  $P = V^{-1}$ .

When deriving the posterior distribution ( $\rightarrow$  III/2.5.2)  $p(B, T|Y)$ , the joint likelihood  $p(Y, B, T)$  is obtained as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v(\frac{\nu_0}{2})}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [(B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n)]) \right] \quad (9)$$

Using the probability density function of the matrix-normal distribution ( $\rightarrow$  II/5.1.3), we can rewrite this as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{(2\pi)^{pv}}{|T|^p |\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v(\frac{\nu_0}{2})}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]) \right] \quad (10)$$

Now,  $B$  can be integrated out easily:

$$\int p(Y, B, T) dB = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v(\frac{\nu_0}{2})}} \cdot |T|^{(\nu_0 - v - 1)/2} \cdot \exp \left[ -\frac{1}{2} \text{tr} (T [\Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]) \right] \quad (11)$$

Using the probability density function of the Wishart distribution, we can rewrite this as

$$\int p(Y, B, T) dB = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \sqrt{\frac{2^{\nu_n v}}{|\Omega_n|^{\nu_n}} \frac{\Gamma_v(\frac{\nu_n}{2})}{\Gamma_v(\frac{\nu_0}{2})}} \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \quad (12)$$

Finally,  $T$  can also be integrated out:



$$\iint p(Y, B, T) \, dB \, dT = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\frac{1}{2}\Omega_0|^{\nu_0}}{|\frac{1}{2}\Omega_n|^{\nu_n}}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} = p(y|m) . \quad (13)$$

Thus, the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is given by

$$\begin{aligned} \log p(Y|m) = & \frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ & \frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v \left( \frac{\nu_n}{2} \right) - \log \Gamma_v \left( \frac{\nu_0}{2} \right) . \end{aligned} \quad (14)$$

■

### 3 Count data

#### 3.1 Binomial observations

##### 3.1.1 Definition

**Definition:** An ordered pair  $(n, y)$  with  $n \in \mathbb{N}$  and  $y \in \mathbb{N}_0$ , where  $y$  is the number of successes in  $n$  trials, constitutes a set of binomial observations.

##### 3.1.2 Binomial test

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Then, the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : p = p_0 \quad (2)$$

is rejected ( $\rightarrow$  I/4.3.1) at significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ , if

$$y \leq c_1 \quad \text{or} \quad y \geq c_2 \quad (3)$$

where  $c_1$  is the largest integer value, such that

$$\sum_{x=0}^{c_1} \text{Bin}(x; n, p_0) \leq \frac{\alpha}{2} , \quad (4)$$

and  $c_2$  is the smallest integer value, such that

$$\sum_{x=c_2}^n \text{Bin}(x; n, p_0) \leq \frac{\alpha}{2} , \quad (5)$$

where  $\text{Bin}(x; n, p)$  is the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3):

$$\text{Bin}(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} . \quad (6)$$

**Proof:** The alternative hypothesis ( $\rightarrow$  I/4.3.3) relative to  $H_0$  for a two-sided test ( $\rightarrow$  I/4.3.4) is

$$H_1 : p \neq p_0 . \quad (7)$$

We can use  $y$  as a test statistic ( $\rightarrow$  I/4.3.5). Its sampling distribution ( $\rightarrow$  I/1.5.6) is given by (1). The cumulative distribution function ( $\rightarrow$  I/1.8.1) (CDF) of the test statistic under the null hypothesis is thus equal to the cumulative distribution function of a binomial distribution ( $\rightarrow$  II/1.2.4) with success probability ( $\rightarrow$  II/1.2.1)  $p_0$ :

$$\Pr(y \leq z | H_0) = \sum_{x=0}^z \text{Bin}(x; n, p_0) = \sum_{x=0}^z \binom{n}{x} p_0^x (1-p_0)^{n-x} . \quad (8)$$

The critical value ( $\rightarrow$  I/4.3.9) is the value of  $y$ , such that the probability of observing this or more extreme values of the test statistic is equal to or smaller than  $\alpha$ . Since  $H_0$  and  $H_1$  define a two-tailed test, we need two critical values  $y_1$  and  $y_2$  that satisfy

$$\begin{aligned}\alpha &\geq \Pr(y \in \{0, \dots, y_1\} \cup \{y_2, \dots, n\} | H_0) \\ &= \Pr(y \leq y_1 | H_0) + \Pr(y \geq y_2 | H_0) \\ &= \Pr(y \leq y_1 | H_0) + (1 - \Pr(y \leq (y_2 - 1) | H_0)) .\end{aligned}\tag{9}$$

Given the test statistic's CDF in (8), this is fulfilled by the values  $c_1$  and  $c_2$  defined in (4) and (5). Thus, the null hypothesis  $H_0$  can be rejected ( $\rightarrow$  I/4.3.9), if the observed test statistic is inside the rejection region ( $\rightarrow$  I/4.3.1):

$$y \in \{0, \dots, c_1\} \cup \{c_2, \dots, n\} .\tag{10}$$

This is equivalent to (3) and thus completes the proof. ■

#### Sources:

- Wikipedia (2023): “Binomial test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-16; URL: [https://en.wikipedia.org/wiki/Binomial\\_test#Usage](https://en.wikipedia.org/wiki/Binomial_test#Usage).
- Wikipedia (2023): “Binomialtest”; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2023-12-16; URL: [https://de.wikipedia.org/wiki/Binomialtest#Signifikanzniveau\\_und\\_kritische\\_Werte](https://de.wikipedia.org/wiki/Binomialtest#Signifikanzniveau_und_kritische_Werte).

### 3.1.3 Maximum likelihood estimation

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) .\tag{1}$$

Then, the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $p$  is

$$\hat{p} = \frac{y}{n} .\tag{2}$$

**Proof:** With the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2):

$$\begin{aligned}p(y|p) &= \text{Bin}(y; n, p) \\ &= \binom{n}{y} p^y (1 - p)^{n-y} .\end{aligned}\tag{3}$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is given by

$$\begin{aligned}\text{LL}(p) &= \log p(y|p) \\ &= \log \binom{n}{y} + y \log p + (n - y) \log(1 - p) .\end{aligned}\tag{4}$$

The derivative of the log-likelihood function (4) with respect to  $p$  is

$$\frac{dLL(p)}{dp} = \frac{y}{p} - \frac{n-y}{1-p} \quad (5)$$

and setting this derivative to zero gives the MLE for  $p$ :

$$\begin{aligned} \frac{dLL(p)}{d\hat{p}} &= 0 \\ 0 &= \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} \\ \frac{n-y}{1-\hat{p}} &= \frac{y}{\hat{p}} \\ (n-y)\hat{p} &= y(1-\hat{p}) \\ n\hat{p} - y\hat{p} &= y - y\hat{p} \\ n\hat{p} &= y \\ \hat{p} &= \frac{y}{n} . \end{aligned} \quad (6)$$

■

#### Sources:

- Wikipedia (2022): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-23; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Statistical\\_inference](https://en.wikipedia.org/wiki/Binomial_distribution#Statistical_inference).

### 3.1.4 Maximum log-likelihood

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Then, the maximum log-likelihood ( $\rightarrow$  I/4.1.4) for this model is

$$\begin{aligned} \text{MLL} &= \log \Gamma(n+1) - \log \Gamma(y+1) - \log \Gamma(n-y+1) \\ &\quad - n \log(n) + y \log(y) + (n-y) \log(n-y) . \end{aligned} \quad (2)$$

**Proof:** The log-likelihood function for binomial data ( $\rightarrow$  III/3.1.3) is given by

$$LL(p) = \log \binom{n}{y} + y \log p + (n-y) \log(1-p) \quad (3)$$

and the maximum likelihood estimate of the success probability ( $\rightarrow$  III/3.1.3)  $p$  is

$$\hat{p} = \frac{y}{n} . \quad (4)$$

Plugging (4) into (3), we obtain the maximum log-likelihood ( $\rightarrow$  I/4.1.4) of the binomial observation model in (1) as

$$\begin{aligned}
\text{MLL} &= \text{LL}(\hat{p}) \\
&= \log \binom{n}{y} + y \log \left( \frac{y}{n} \right) + (n - y) \log \left( 1 - \frac{y}{n} \right) \\
&= \log \binom{n}{y} + y \log \left( \frac{y}{n} \right) + (n - y) \log \left( \frac{n - y}{n} \right) \\
&= \log \binom{n}{y} + y \log(y) + (n - y) \log(n - y) - n \log(n) .
\end{aligned} \tag{5}$$

With the definition of the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n - k)!} \tag{6}$$

and the definition of the gamma function

$$\Gamma(n) = (n - 1)! , \tag{7}$$

the MLL finally becomes

$$\begin{aligned}
\text{MLL} &= \log \Gamma(n + 1) - \log \Gamma(y + 1) - \log \Gamma(n - y + 1) \\
&\quad - n \log(n) + y \log(y) + (n - y) \log(n - y) .
\end{aligned} \tag{8}$$

■

### 3.1.5 Maximum-a-posteriori estimation

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \tag{1}$$

Moreover, assume a beta prior distribution ( $\rightarrow$  III/3.1.6) over the model parameter  $p$ :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \tag{2}$$

Then, the maximum-a-posteriori estimate ( $\rightarrow$  I/5.1.14) of  $p$  is

$$\hat{p}_{\text{MAP}} = \frac{\alpha_0 + y - 1}{\alpha_0 + \beta_0 + n - 2} . \tag{3}$$

**Proof:** Given the prior distribution ( $\rightarrow$  I/5.1.3) in (2), the posterior distribution ( $\rightarrow$  I/5.1.9) for binomial observations ( $\rightarrow$  III/3.1.1) is also a beta distribution ( $\rightarrow$  III/3.1.7)

$$p(p|y) = \text{Bet}(p; \alpha_n, \beta_n) \tag{4}$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are equal to

$$\begin{aligned}
\alpha_n &= \alpha_0 + y \\
\beta_n &= \beta_0 + (n - y) .
\end{aligned} \tag{5}$$

The mode of the beta distribution ( $\rightarrow$  II/3.9.10) is given by:

$$X \sim \text{Bet}(\alpha, \beta) \quad \Rightarrow \quad \text{mode}(X) = \frac{\alpha - 1}{\alpha + \beta - 2} . \quad (6)$$

Applying (6) to (4) with (5), the maximum-a-posteriori estimate ( $\rightarrow$  I/5.1.14) of  $p$  follows as:

$$\begin{aligned} \hat{p}_{\text{MAP}} &= \frac{\alpha_n - 1}{\alpha_n + \beta_n - 2} \\ &\stackrel{(5)}{=} \frac{\alpha_0 + y - 1}{\alpha_0 + y + \beta_0 + (n - y) - 2} \\ &= \frac{\alpha_0 + y - 1}{\alpha_0 + \beta_0 + n - 2} . \end{aligned} \quad (7)$$

■

### 3.1.6 Conjugate prior distribution

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for the model parameter  $p$  is a beta distribution ( $\rightarrow$  II/3.9.1):

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

**Proof:** With the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1 - p)^{n-y} . \quad (3)$$

In other words, the likelihood function is proportional to a power of  $p$  times a power of  $(1 - p)$ :

$$p(y|p) \propto p^y (1 - p)^{n-y} . \quad (4)$$

The same is true for a beta distribution over  $p$

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) \quad (5)$$

the probability density function of which ( $\rightarrow$  II/3.9.5)

$$p(p) = \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1 - p)^{\beta_0-1} \quad (6)$$

exhibits the same proportionality

$$p(p) \propto p^{\alpha_0-1} (1 - p)^{\beta_0-1} \quad (7)$$

and is therefore conjugate relative to the likelihood.

■

**Sources:**

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Estimation\\_of\\_parameters](https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters).

**3.1.7 Posterior distribution**

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution ( $\rightarrow$  III/3.1.6) over the model parameter  $p$ :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a beta distribution ( $\rightarrow$  II/3.9.1)

$$p(p|y) = \text{Bet}(p; \alpha_n, \beta_n) . \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) . \end{aligned} \quad (4)$$

**Proof:** With the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1 - p)^{n-y} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y} p^y (1 - p)^{n-y} \cdot \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1 - p)^{\beta_0-1} \\ &= \frac{1}{B(\alpha_0, \beta_0)} \binom{n}{y} p^{\alpha_0+y-1} (1 - p)^{\beta_0+(n-y)-1} . \end{aligned} \quad (6)$$

Note that the posterior distribution is proportional to the joint likelihood ( $\rightarrow$  I/5.1.10):

$$p(p|y) \propto p(y, p) . \quad (7)$$

Setting  $\alpha_n = \alpha_0 + y$  and  $\beta_n = \beta_0 + (n - y)$ , the posterior distribution is therefore proportional to

$$p(p|y) \propto p^{\alpha_n-1} (1 - p)^{\beta_n-1} \quad (8)$$

which, when normalized to one, results in the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5):

$$p(p|y) = \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1 - p)^{\beta_n-1} = \text{Bet}(p; \alpha_n, \beta_n) . \quad (9)$$



**Sources:**

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: [https://en.wikipedia.org/wiki/Binomial\\_distribution#Estimation\\_of\\_parameters](https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters).

### 3.1.8 Log model evidence

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution ( $\rightarrow$  III/3.1.6) over the model parameter  $p$ :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(y+1) - \log \Gamma(n-y+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (3)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) . \end{aligned} \quad (4)$$

**Proof:** With the probability mass function of the binomial distribution ( $\rightarrow$  II/1.2.3), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1-p)^{n-y} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y} p^y (1-p)^{n-y} \cdot \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \\ &= \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0+y-1} (1-p)^{\beta_0+(n-y)-1} . \end{aligned} \quad (6)$$

Note that the model evidence is the marginal density of the joint likelihood ( $\rightarrow$  I/5.1.15):

$$p(y) = \int p(y, p) dp . \quad (7)$$

Setting  $\alpha_n = \alpha_0 + y$  and  $\beta_n = \beta_0 + (n - y)$ , the joint likelihood can also be written as



$$p(y, p) = \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} . \quad (8)$$

Using the probability density function of the beta distribution ( $\rightarrow$  II/3.9.5),  $p$  can now be integrated out easily

$$\begin{aligned} p(y) &= \int \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} dp \\ &= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \int \text{Bet}(p; \alpha_n, \beta_n) dp \\ &= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} , \end{aligned} \quad (9)$$

such that the log model evidence ( $\rightarrow$  IV/3.1.3) (LME) is shown to be

$$\log p(y|m) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \quad (10)$$

With the definition of the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (11)$$

and the definition of the gamma function

$$\Gamma(n) = (n-1)! , \quad (12)$$

the LME finally becomes

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(y+1) - \log \Gamma(n-y+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (13)$$

■

#### Sources:

- Wikipedia (2020): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-24; URL: [https://en.wikipedia.org/wiki/Beta-binomial\\_distribution#Motivation\\_and\\_derivation](https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation).

#### 3.1.9 Log Bayes factor

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $p$  is 0.5 (null model ( $\rightarrow$  I/4.3.2)), the other imposing a beta distribution ( $\rightarrow$  III/3.1.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $p$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y &\sim \text{Bin}(n, p), \quad p = 0.5 \\ m_1 : y &\sim \text{Bin}(n, p), \quad p \sim \text{Bet}(\alpha_0, \beta_0) . \end{aligned} \quad (2)$$

Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_0$  is

$$\text{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left( \frac{1}{2} \right) \quad (3)$$

where  $B(x, y)$  is the beta function and  $\alpha_n$  and  $\beta_n$  are the posterior hyperparameters for binomial observations ( $\rightarrow$  III/3.1.7) which are functions of the number of trials ( $\rightarrow$  II/1.2.1)  $n$  and the number of successes ( $\rightarrow$  II/1.2.1)  $y$ .

**Proof:** The log Bayes factor is equal to the difference of two log model evidences ( $\rightarrow$  IV/3.3.8):

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (4)$$

The LME of the alternative  $m_1$  is equal to the log model evidence for binomial observations ( $\rightarrow$  III/3.1.8):

$$\text{LME}(m_1) = \log p(y|m_1) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \quad (5)$$

Because the null model  $m_0$  has no free parameter, its log model evidence ( $\rightarrow$  IV/3.1.3) (logarithmized marginal likelihood ( $\rightarrow$  I/5.1.15)) is equal to the log-likelihood function for binomial observations ( $\rightarrow$  III/3.1.3) at the value  $p = 0.5$ :

$$\begin{aligned} \text{LME}(m_0) &= \log p(y|p = 0.5) = \log \binom{n}{y} + y \log(0.5) + (n - y) \log(1 - 0.5) \\ &= \log \binom{n}{y} + n \log \left( \frac{1}{2} \right) . \end{aligned} \quad (6)$$

Subtracting the two LMEs from each other, the LBF emerges as

$$\text{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left( \frac{1}{2} \right) \quad (7)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/3.1.7)

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) \end{aligned} \quad (8)$$

with the number of trials ( $\rightarrow$  II/1.2.1)  $n$  and the number of successes ( $\rightarrow$  II/1.2.1)  $y$ .

■

### 3.1.10 Posterior probability

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution ( $\rightarrow$  II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $p$  is 0.5 (null model ( $\rightarrow$  I/4.3.2)), the other imposing a beta distribution ( $\rightarrow$  III/3.1.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $p$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y &\sim \text{Bin}(n, p), \quad p = 0.5 \\ m_1 : y &\sim \text{Bin}(n, p), \quad p \sim \text{Bet}(\alpha_0, \beta_0) . \end{aligned} \quad (2)$$

Then, the posterior probability ( $\rightarrow$  IV/3.4.1) of the alternative model ( $\rightarrow$  I/4.3.3) is given by

$$p(m_1|y) = \frac{1}{1 + 2^{-n} [B(\alpha_0, \beta_0)/B(\alpha_n, \beta_n)]} \quad (3)$$

where  $B(x, y)$  is the beta function and  $\alpha_n$  and  $\beta_n$  are the posterior hyperparameters for binomial observations ( $\rightarrow$  III/3.1.7) which are functions of the number of trials ( $\rightarrow$  II/1.2.1)  $n$  and the number of successes ( $\rightarrow$  II/1.2.1)  $y$ .

**Proof:** The posterior probability for one of two models is a function of the log Bayes factor in favor of this model ( $\rightarrow$  IV/3.4.4):

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (4)$$

The log Bayes factor in favor of the alternative model for binomial observations ( $\rightarrow$  III/3.1.9) is given by

$$\text{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left( \frac{1}{2} \right) . \quad (5)$$

and the corresponding Bayes factor ( $\rightarrow$  IV/3.3.1), i.e. exponentiated log Bayes factor ( $\rightarrow$  IV/3.3.7), is equal to

$$\text{BF}_{10} = \exp(\text{LBF}_{10}) = 2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} . \quad (6)$$

Thus, the posterior probability of the alternative, assuming a prior distribution over the probability  $p$ , compared to the null model, assuming a fixed probability  $p = 0.5$ , follows as

$$\begin{aligned} p(m_1|y) &\stackrel{(4)}{=} \frac{\exp(\text{LBF}_{10})}{\exp(\text{LBF}_{10}) + 1} \\ &\stackrel{(6)}{=} \frac{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)}}{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} + 1} \\ &= \frac{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)}}{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \left( 1 + 2^{-n} \frac{B(\alpha_0, \beta_0)}{B(\alpha_n, \beta_n)} \right)} \\ &= \frac{1}{1 + 2^{-n} [B(\alpha_0, \beta_0)/B(\alpha_n, \beta_n)]} \end{aligned} \quad (7)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/3.1.7)

$$\begin{aligned}\alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y)\end{aligned}\tag{8}$$

with the number of trials ( $\rightarrow$  II/1.2.1)  $n$  and the number of successes ( $\rightarrow$  II/1.2.1)  $y$ . ■

### 3.1.11 Cross-validated log model evidence

**Theorem:** Let  $y$  be the number of successes resulting from  $n$  independent trials with unknown success probability  $p$ , such that  $y$  follows a binomial distribution:

$$y \sim \text{Bin}(n, p) . \tag{1}$$

Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that  $p$  is 0.5 (null model ( $\rightarrow$  I/4.3.2)), the other imposing a beta distribution ( $\rightarrow$  III/3.1.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameter  $p$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned}m_0 : y &\sim \text{Bin}(n, p), p = 0.5 \\ m_1 : y &\sim \text{Bin}(n, p), p \sim \text{Bet}(\alpha_0, \beta_0) .\end{aligned}\tag{2}$$

Then, the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are

$$\begin{aligned}\text{cvLME}(m_0) &= -n \log(2) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} \right] \\ \text{cvLME}(m_1) &= S \cdot \log B(y, n - y) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} - \log B(y_1^{(i)}, n_1 - y_1^{(i)}) \right]\end{aligned}\tag{3}$$

where  $y_1^{(i)}$  and  $y_2^{(i)}$  are the training and test data, respectively, in the  $i$ -th cross-validation fold with  $n_1$  and  $n_2$  data points, respectively,  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9) and  $B(x, y)$  is the beta function.

**Proof:** For evaluation of the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME), we assume that  $n$  data points are divided into  $S \mid n$  data subsets without remainder. Then, the number of training data points  $n_1$  and test data points  $n_2$  are given by

$$\begin{aligned}n &= n_1 + n_2 \\ n_1 &= \frac{S-1}{S}n \\ n_2 &= \frac{1}{S}n ,\end{aligned}\tag{4}$$

such that  $y_1^{(i)}$  is the number of successes ( $\rightarrow$  II/1.2.1) from the  $n_1$  trials of the training set ( $\rightarrow$  III/3.1.1) and  $y_2^{(i)}$  is the number of successes ( $\rightarrow$  II/1.2.1) from the  $n_2$  trials of the test set ( $\rightarrow$  III/3.1.1) in the  $i$ -th cross-validation fold and it holds that

$$y = y_1^{(i)} + y_2^{(i)} . \quad (5)$$

First, we consider the null model  $m_0$  assuming  $p = 0.5$ . Because this model has no free parameter, nothing is estimated from the training data and the assumed parameter value is applied to the test data. Consequently, the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) is equal to the log-likelihood function ( $\rightarrow$  III/3.1.3) of the test data at  $p = 0.5$ :

$$\begin{aligned} \text{oosLME}_i(m_0) &= \log p \left( y_2^{(i)} \middle| p = 0.5 \right) \\ &= \log \binom{n_2}{y_2^{(i)}} + y_2^{(i)} \log 0.5 + (n_2 - y_2^{(i)}) \log(1 - 0.5) \\ &= \log \binom{n_2}{y_2^{(i)}} + n_2 \log \frac{1}{2} . \end{aligned} \quad (6)$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, such that the cvLME of  $m_0$  is:

$$\begin{aligned} \text{cvLME}(m_0) &= \sum_{i=1}^S \text{oosLME}_i(m_0) \\ &= \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} + n_2 \log \left( \frac{1}{2} \right) \right] \\ &= S \cdot n_2 \log \left( \frac{1}{2} \right) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} \right] \\ &= -n \log(2) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} \right] . \end{aligned} \quad (7)$$

Next, we have a look at the alternative  $m_1$  assuming  $p \neq 0.5$ . First, the training data  $y_1^{(i)}$  are analyzed using a non-informative prior distribution ( $\rightarrow$  I/5.2.3) and applying the posterior distribution for binomial observations ( $\rightarrow$  III/3.1.7):

$$\begin{aligned} \alpha_0^{(1)} = 0 &\Rightarrow \alpha_n^{(1)} = \alpha_0^{(1)} + y_1^{(i)} = y_1^{(i)} \\ \beta_0^{(1)} = 0 &\Rightarrow \beta_n^{(1)} = \beta_0^{(1)} + (n_1 - y_1^{(i)}) = n_1 - y_1^{(i)} . \end{aligned} \quad (8)$$

This results in a posterior characterized by  $\alpha_n^{(1)}$  and  $\beta_n^{(1)}$ . Then, the test data  $y_2^{(i)}$  are analyzed using this posterior as an informative prior distribution ( $\rightarrow$  I/5.2.3), again applying the posterior distribution for binomial observations ( $\rightarrow$  III/3.1.7):

$$\begin{aligned} \alpha_0^{(2)} = \alpha_n^{(1)} = y_1^{(i)} &\Rightarrow \alpha_n^{(2)} = \alpha_0^{(2)} + y_2^{(i)} = y \\ \beta_0^{(2)} = \beta_n^{(1)} = n_1 - y_1^{(i)} &\Rightarrow \beta_n^{(2)} = \beta_0^{(2)} + (n_2 - y_2^{(i)}) = n - y . \end{aligned} \quad (9)$$

In the test data, we now have a prior characterized by  $\alpha_0^{(2)}$  and  $\beta_0^{(2)}$  and a posterior characterized  $\alpha_n^{(2)}$  and  $\beta_n^{(2)}$ . Applying the log model evidence for binomial observations ( $\rightarrow$  III/3.1.8), the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) therefore follows as

$$\begin{aligned} \text{oosLME}_i(m_1) &= \log \binom{n_2}{y_2^{(i)}} + \log B(\alpha_n^{(2)}, \beta_n^{(2)}) - \log B(\alpha_0^{(2)}, \beta_0^{(2)}) \\ &= \log \binom{n_2}{y_2^{(i)}} + \log B(y, n - y) - \log B(y_1^{(i)}, n_1 - y_1^{(i)}) . \end{aligned} \quad (10)$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, the cvLME of  $m_1$  becomes:

$$\begin{aligned} \text{cvLME}(m_1) &= \sum_{i=1}^S \text{oosLME}_i(m_1) \\ &= \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} + \log B(y, n - y) - \log B(y_1^{(i)}, n_1 - y_1^{(i)}) \right] \\ &= S \cdot \log B(y, n - y) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_2^{(i)}} - \log B(y_1^{(i)}, n_1 - y_1^{(i)}) \right] \end{aligned} \quad (11)$$

Together, (7) and (11) conform to the results given in (3). ■

## 3.2 Multinomial observations

### 3.2.1 Definition

**Definition:** An ordered pair  $(n, y)$  with  $n \in \mathbb{N}$  and  $y = [y_1, \dots, y_k] \in \mathbb{N}_0^{1 \times k}$ , where  $y_i$  is the number of observations for the  $i$ -th out of  $k$  categories obtained in  $n$  trials,  $i = 1, \dots, k$ , constitutes a set of multinomial observations.

### 3.2.2 Multinomial test

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the null hypothesis ( $\rightarrow$  I/4.3.2)

$$H_0 : p = p_0 = [p_{01}, \dots, p_{0k}] \quad (2)$$

is rejected ( $\rightarrow$  I/4.3.1) at significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ , if

$$\text{Pr}_{\text{sig}} = \sum_{x: \text{Pr}_0(x) \leq \text{Pr}_0(y)} \text{Pr}_0(x) < \alpha \quad (3)$$

where  $\Pr_0(x)$  is the probability of observing the numbers of occurrences  $x = [x_1, \dots, x_k]$  under the null hypothesis:

$$\Pr_0(x) = n! \prod_{j=1}^k \frac{p_{0j}^{x_j}}{x_j!} . \quad (4)$$

**Proof:** The alternative hypothesis ( $\rightarrow$  I/4.3.3) relative to  $H_0$  is

$$H_1 : p_j \neq p_{0j} \quad \text{for at least one } j = 1, \dots, k . \quad (5)$$

We can use  $y$  as a test statistic ( $\rightarrow$  I/4.3.5). Its sampling distribution ( $\rightarrow$  I/1.5.6) is given by (1). The probability mass function ( $\rightarrow$  I/1.6.1) (PMF) of the test statistic under the null hypothesis is thus equal to the probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2) with category probabilities ( $\rightarrow$  II/2.2.1)  $p_0$ :

$$\Pr(y = x | H_0) = \text{Mult}(x; n, p_0) = \binom{n}{x_1, \dots, x_k} \prod_{j=1}^k p_j^{x_j} . \quad (6)$$

The multinomial coefficient in this equation is equal to

$$\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \cdot \dots \cdot k_m!} , \quad (7)$$

such that the probability of observing the counts  $y$ , given  $H_0$ , is

$$\Pr(y | H_0) = n! \prod_{j=1}^k \frac{p_{0j}^{y_j}}{y_j!} . \quad (8)$$

The probability of observing any other set of counts  $x$ , given  $H_0$ , is

$$\Pr(x | H_0) = n! \prod_{j=1}^k \frac{p_{0j}^{x_j}}{x_j!} . \quad (9)$$

The p-value ( $\rightarrow$  I/4.3.10) is the probability of observing a value of the test statistic ( $\rightarrow$  I/4.3.5) that is as extreme or more extreme than the actually observed test statistic. Any set of counts  $x$  might be considered as extreme or more extreme than the actually observed counts  $y$ , if the former is equally probable or less probable than the latter according to the PMF:

$$\Pr_0(x) \leq \Pr_0(y) . \quad (10)$$

Thus, the p-value ( $\rightarrow$  I/4.3.10) for the data in (1) is equal to

$$p = \sum_{x: \Pr_0(x) \leq \Pr_0(y)} \Pr_0(x) \quad (11)$$

and the null hypothesis in (2) is rejected ( $\rightarrow$  I/4.3.1), if

$$p < \alpha . \quad (12)$$

■

**Sources:**

- Wikipedia (2023): “Multinomial test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-23; URL: [https://en.wikipedia.org/wiki/Multinomial\\_test](https://en.wikipedia.org/wiki/Multinomial_test).

### 3.2.3 Maximum likelihood estimation

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $p$  is

$$\hat{p} = \frac{1}{n}y, \quad \text{i.e.} \quad \hat{p}_j = \frac{y_j}{n} \quad \text{for all } j = 1, \dots, k . \quad (2)$$

**Proof:** Note that the marginal distribution of each element in a multinomial random vector is a binomial distribution ( $\rightarrow$  II/2.2.7)

$$X \sim \text{Mult}(n, p) \quad \Rightarrow \quad X_j \sim \text{Bin}(n, p_j) \quad \text{for all } j = 1, \dots, k . \quad (3)$$

Thus, combining (1) with (3), we have

$$y_j \sim \text{Bin}(n, p_j) \quad (4)$$

which implies the likelihood function ( $\rightarrow$  III/3.1.3)

$$p(y|p_j) = \text{Bin}(y_j; n, p_j) = \binom{n}{y_j} p_j^{y_j} (1 - p_j)^{n-y_j} . \quad (5)$$

To this, we can apply maximum likelihood estimation for binomial observations ( $\rightarrow$  III/3.1.3), such that the MLE for each  $p_j$  is

$$\hat{p}_j = \frac{y_j}{n} . \quad (6)$$

■

### 3.2.4 Maximum log-likelihood

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the maximum log-likelihood ( $\rightarrow$  I/4.1.4) for this model is

$$\text{MLL} = \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) - n \log(n) + \sum_{j=1}^k y_j \log(y_j) . \quad (2)$$



**Proof:** With the probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2), equation (1) implies the following likelihood function ( $\rightarrow$  I/5.1.2):

$$\begin{aligned} p(y|p) &= \text{Mult}(y; n, p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \end{aligned} \quad (3)$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is given by

$$\begin{aligned} \text{LL}(p) &= \log p(y|p) \\ &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k y_j \log(p_j) . \end{aligned} \quad (4)$$

The maximum likelihood estimates of the category probabilities ( $\rightarrow$  III/3.2.3)  $p$  are

$$\hat{p} = [\hat{p}_1, \dots, \hat{p}_k] \quad \text{with} \quad \hat{p}_j = \frac{y_j}{n} \quad \text{for all} \quad j = 1, \dots, k . \quad (5)$$

Plugging (5) into (4), we obtain the maximum log-likelihood ( $\rightarrow$  I/4.1.4) of the multinomial observation model in (1) as

$$\begin{aligned} \text{MLL} &= \text{LL}(\hat{p}) \\ &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k y_j \log \left( \frac{y_j}{n} \right) \\ &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k [y_j \log(y_j) - y_j \log(n)] \\ &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k y_j \log(y_j) - \sum_{j=1}^k y_j \log(n) \\ &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k y_j \log(y_j) - n \log(n) . \end{aligned} \quad (6)$$

With the definition of the multinomial coefficient

$$\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \cdot \dots \cdot k_m!} \quad (7)$$

and the definition of the gamma function

$$\Gamma(n) = (n-1)! , \quad (8)$$

the MLL finally becomes

$$\text{MLL} = \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) - n \log(n) + \sum_{j=1}^k y_j \log(y_j) . \quad (9)$$

■

### 3.2.5 Maximum-a-posteriori estimation

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution ( $\rightarrow$  III/3.2.6) over the model parameter  $p$ :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the maximum-a-posteriori estimates ( $\rightarrow$  I/5.1.14) of  $p$  are

$$\hat{p}_{\text{MAP}} = \frac{\alpha_0 + y - 1}{\sum_{j=1}^k \alpha_{0j} + n - k} . \quad (3)$$

**Proof:** Given the prior distribution ( $\rightarrow$  I/5.1.3) in (2), the posterior distribution ( $\rightarrow$  I/5.1.9) for multinomial observations ( $\rightarrow$  III/3.2.1) is also a Dirichlet distribution ( $\rightarrow$  III/3.2.7)

$$p(p|y) = \text{Dir}(p; \alpha_n) \quad (4)$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are equal to

$$\alpha_{nj} = \alpha_{0j} + y_j, \quad j = 1, \dots, k . \quad (5)$$

The mode of the Dirichlet distribution is given by:

$$X \sim \text{Dir}(\alpha) \quad \Rightarrow \quad \text{mode}(X_i) = \frac{\alpha_i - 1}{\sum_j \alpha_j - k} . \quad (6)$$

Applying (6) to (4) with (5), the maximum-a-posteriori estimates ( $\rightarrow$  I/5.1.14) of  $p$  follow as

$$\begin{aligned} \hat{p}_{i,\text{MAP}} &= \frac{\alpha_{ni} - 1}{\sum_j \alpha_{nj} - k} \\ &\stackrel{(5)}{=} \frac{\alpha_{0i} + y_i - 1}{\sum_j (\alpha_{0j} + y_j) - k} \\ &= \frac{\alpha_{0i} + y_i - 1}{\sum_j \alpha_{0j} + \sum_j y_j - k} . \end{aligned} \quad (7)$$

Since  $y_1 + \dots + y_k = n$  by definition ( $\rightarrow$  III/3.2.1), this becomes

$$\hat{p}_{i,\text{MAP}} = \frac{\alpha_{0i} + y_i - 1}{\sum_j \alpha_{0j} + n - k} \quad (8)$$

which, using the  $1 \times k$  vectors ( $\rightarrow$  III/3.2.1)  $y$ ,  $p$  and  $\alpha_0$ , can be written as:

$$\hat{p}_{\text{MAP}} = \frac{\alpha_0 + y - 1}{\sum_{j=1}^k \alpha_{0j} + n - k} . \quad (9)$$

■

### 3.2.6 Conjugate prior distribution

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for the model parameter  $p$  is a Dirichlet distribution ( $\rightarrow$  II/4.5.1):

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

**Proof:** With the probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (3)$$

In other words, the likelihood function is proportional to a product of powers of the entries of the vector  $p$ :

$$p(y|p) \propto \prod_{j=1}^k p_j^{y_j} . \quad (4)$$

The same is true for a Dirichlet distribution over  $p$

$$p(p) = \text{Dir}(p; \alpha_0) \quad (5)$$

the probability density function of which ( $\rightarrow$  II/4.5.2)

$$p(p) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (6)$$

exhibits the same proportionality

$$p(p) \propto \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (7)$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: [https://en.wikipedia.org/wiki/Dirichlet\\_distribution#Conjugate\\_to\\_categorical/multinomial](https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomial)

### 3.2.7 Posterior distribution

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution ( $\rightarrow$  III/3.2.6) over the model parameter  $p$ :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a Dirichlet distribution ( $\rightarrow$  II/4.5.1)

$$p(p|y) = \text{Dir}(p; \alpha_n) . \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\alpha_{nj} = \alpha_{0j} + y_j, \quad j = 1, \dots, k . \quad (4)$$

**Proof:** With the probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \\ &= \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{\alpha_{0j}+y_j-1} . \end{aligned} \quad (6)$$

Note that the posterior distribution is proportional to the joint likelihood ( $\rightarrow$  I/5.1.10):

$$p(p|y) \propto p(y, p) . \quad (7)$$

Setting  $\alpha_{nj} = \alpha_{0j} + y_j$ , the posterior distribution is therefore proportional to

$$p(p|y) \propto \prod_{j=1}^k p_j^{\alpha_{nj}-1} \quad (8)$$

which, when normalized to one, results in the probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2):

$$p(p|y) = \frac{\Gamma(\sum_{j=1}^k \alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} = \text{Dir}(p; \alpha_n) . \quad (9)$$

**Sources:**

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: [https://en.wikipedia.org/wiki/Dirichlet\\_distribution#Conjugate\\_to\\_categorical/multinomial](https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomial)

**3.2.8 Log model evidence**

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution ( $\rightarrow$  III/3.2.6) over the model parameter  $p$ :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) \\ &\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) . \end{aligned} \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\alpha_{nj} = \alpha_{0j} + y_j, \quad j = 1, \dots, k . \quad (4)$$

**Proof:** With the probability mass function of the multinomial distribution ( $\rightarrow$  II/2.2.2), the likelihood function ( $\rightarrow$  I/5.1.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}+y_j-1} . \end{aligned} \quad (6)$$

Note that the model evidence is the marginal density of the joint likelihood:

$$p(y) = \int p(y, p) \, dp. \quad (7)$$

Setting  $\alpha_{nj} = \alpha_{0j} + y_j$ , the joint likelihood can also be written as

$$p(y, p) = \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1}. \quad (8)$$

Using the probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2),  $p$  can now be integrated out easily

$$\begin{aligned} p(y) &= \int \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} \, dp \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \int \text{Dir}(p; \alpha_n) \, dp \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})}, \end{aligned} \quad (9)$$

such that the log model evidence ( $\rightarrow$  IV/3.1.3) (LME) is shown to be

$$\begin{aligned} \log p(y|m) &= \log \binom{n}{y_1, \dots, y_k} + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}). \end{aligned} \quad (10)$$

With the definition of the multinomial coefficient

$$\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \cdot \dots \cdot k_m!} \quad (11)$$

and the definition of the gamma function

$$\Gamma(n) = (n-1)!, \quad (12)$$

the LME finally becomes

$$\begin{aligned}
\log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) \\
&\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
&\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) .
\end{aligned} \tag{13}$$

■

### 3.2.9 Log Bayes factor

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \tag{1}$$

Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that each  $p_j$  is  $1/k$  (null model ( $\rightarrow$  I/4.3.2)), the other imposing a Dirichlet distribution ( $\rightarrow$  III/3.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameters  $p_1, \dots, p_k$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned}
m_0 : y &\sim \text{Mult}(n, p), \quad p = [1/k, \dots, 1/k] \\
m_1 : y &\sim \text{Mult}(n, p), \quad p \sim \text{Dir}(\alpha_0) .
\end{aligned} \tag{2}$$

Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of  $m_1$  against  $m_0$  is

$$\begin{aligned}
\text{LBF}_{10} &= \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
&\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) - n \log\left(\frac{1}{k}\right)
\end{aligned} \tag{3}$$

where  $\Gamma(x)$  is the gamma function and  $\alpha_n$  are the posterior hyperparameters for multinomial observations ( $\rightarrow$  III/3.2.7) which are functions of the numbers of observations ( $\rightarrow$  II/2.2.1)  $y_1, \dots, y_k$ .

**Proof:** The log Bayes factor is equal to the difference of two log model evidences ( $\rightarrow$  IV/3.3.8):

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \tag{4}$$

The LME of the alternative  $m_1$  is equal to the log model evidence for multinomial observations ( $\rightarrow$  III/3.2.8):

$$\begin{aligned}
\text{LME}(m_1) = \log p(y|m_1) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) \\
&+ \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
&+ \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) .
\end{aligned} \tag{5}$$

Because the null model  $m_0$  has no free parameter, its log model evidence ( $\rightarrow$  IV/3.1.3) (logarithmized marginal likelihood ( $\rightarrow$  I/5.1.15)) is equal to the log-likelihood function for multinomial observations ( $\rightarrow$  III/3.2.3) at the value  $p_0 = [1/k, \dots, 1/k]$ :

$$\begin{aligned}
\text{LME}(m_0) = \log p(y|p=p_0) &= \log \binom{n}{y_1, \dots, y_k} + \sum_{j=1}^k y_j \log \left(\frac{1}{k}\right) \\
&= \log \binom{n}{y_1, \dots, y_k} + n \log \left(\frac{1}{k}\right) .
\end{aligned} \tag{6}$$

Subtracting the two LMEs from each other, the LBF emerges as

$$\begin{aligned}
\text{LBF}_{10} &= \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
&+ \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) - n \log \left(\frac{1}{k}\right)
\end{aligned} \tag{7}$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/3.2.7)

$$\begin{aligned}
\alpha_n &= \alpha_0 + y \\
&= [\alpha_{01}, \dots, \alpha_{0k}] + [y_1, \dots, y_k] \\
&= [\alpha_{01} + y_1, \dots, \alpha_{0k} + y_k] \\
\text{i.e. } \alpha_{nj} &= \alpha_{0j} + y_j \quad \text{for all } j = 1, \dots, k
\end{aligned} \tag{8}$$

with the numbers of observations ( $\rightarrow$  II/2.2.1)  $y_1, \dots, y_k$ .

■

### 3.2.10 Posterior probability

**Theorem:** Let  $y = [y_1, \dots, y_k]$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \tag{1}$$



Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that each  $p_j$  is  $1/k$  (null model ( $\rightarrow$  I/4.3.2)), the other imposing a Dirichlet distribution ( $\rightarrow$  III/3.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameters  $p_1, \dots, p_k$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned} m_0 : y &\sim \text{Mult}(n, p), \quad p = [1/k, \dots, 1/k] \\ m_1 : y &\sim \text{Mult}(n, p), \quad p \sim \text{Dir}(\alpha_0) . \end{aligned} \quad (2)$$

Then, the posterior probability ( $\rightarrow$  IV/3.4.1) of the alternative model ( $\rightarrow$  I/4.3.3) is given by

$$p(m_1|y) = \frac{1}{1 + k^{-n} \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{nj})}{\Gamma(\sum_{j=1}^k \alpha_{0j})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{nj})}} \quad (3)$$

where  $\Gamma(x)$  is the gamma function and  $\alpha_n$  are the posterior hyperparameters for multinomial observations ( $\rightarrow$  III/3.2.7) which are functions of the numbers of observations ( $\rightarrow$  II/2.2.1)  $y_1, \dots, y_k$ .

**Proof:** The posterior probability for one of two models is a function of the log Bayes factor in favor of this model ( $\rightarrow$  IV/3.4.4):

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (4)$$

The log Bayes factor in favor of the alternative model for multinomial observations ( $\rightarrow$  III/3.2.9) is given by

$$\begin{aligned} \text{LBF}_{10} &= \log \Gamma \left( \sum_{j=1}^k \alpha_{0j} \right) - \log \Gamma \left( \sum_{j=1}^k \alpha_{nj} \right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) - n \log \left( \frac{1}{k} \right) \end{aligned} \quad (5)$$

and the corresponding Bayes factor ( $\rightarrow$  IV/3.3.1), i.e. exponentiated log Bayes factor ( $\rightarrow$  IV/3.3.7), is equal to

$$\text{BF}_{10} = \exp(\text{LBF}_{10}) = k^n \cdot \frac{\Gamma \left( \sum_{j=1}^k \alpha_{0j} \right)}{\Gamma \left( \sum_{j=1}^k \alpha_{nj} \right)} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} . \quad (6)$$

Thus, the posterior probability of the alternative, assuming a prior distribution over the probabilities  $p_1, \dots, p_k$ , compared to the null model, assuming fixed probabilities  $p = [1/k, \dots, 1/k]$ , follows as

$$\begin{aligned}
p(m_1|y) &\stackrel{(4)}{=} \frac{\exp(\text{LBF}_{10})}{\exp(\text{LBF}_{10}) + 1} \\
&\stackrel{(6)}{=} \frac{k^n \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\Gamma(\sum_{j=1}^k \alpha_{nj})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})}}{k^n \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\Gamma(\sum_{j=1}^k \alpha_{nj})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} + 1} \\
&= \frac{k^n \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\Gamma(\sum_{j=1}^k \alpha_{nj})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})}}{1 + k^{-n} \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{nj})}{\Gamma(\sum_{j=1}^k \alpha_{0j})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{nj})}} \\
&= \frac{1}{1 + k^{-n} \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{nj})}{\Gamma(\sum_{j=1}^k \alpha_{0j})} \cdot \frac{\prod_{j=1}^k \Gamma(\alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{nj})}}
\end{aligned} \tag{7}$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by ( $\rightarrow$  III/3.2.7)

$$\alpha_n = \alpha_0 + y, \quad \text{i.e.} \quad \alpha_{nj} = \alpha_{0j} + y_j \tag{8}$$

with the numbers of observations ( $\rightarrow$  II/2.2.1)  $y_1, \dots, y_k$ .

■

### 3.2.11 Cross-validated log model evidence

**Theorem:** Let  $y = [y_1, \dots, y_k] \in \mathbb{N}_0^{1 \times k}$  be the number of observations in  $k$  categories resulting from  $n$  independent trials with unknown category probabilities  $p = [p_1, \dots, p_k]$ , such that  $y$  follows a multinomial distribution ( $\rightarrow$  II/2.2.1):

$$y \sim \text{Mult}(n, p) . \tag{1}$$

Moreover, assume two statistical models ( $\rightarrow$  I/5.1.6), one assuming that each  $p_j$  is  $1/k$  (null model ( $\rightarrow$  I/4.3.2)), the other imposing a Dirichlet distribution ( $\rightarrow$  III/3.2.6) as the prior distribution ( $\rightarrow$  I/5.1.3) on the model parameters  $p_1, \dots, p_k$  (alternative ( $\rightarrow$  I/4.3.3)):

$$\begin{aligned}
m_0 : y &\sim \text{Mult}(n, p), \quad p = [1/k, \dots, 1/k] \\
m_1 : y &\sim \text{Mult}(n, p), \quad p \sim \text{Dir}(\alpha_0) .
\end{aligned} \tag{2}$$

Then, the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) of  $m_0$  and  $m_1$  are

$$\begin{aligned}
\text{cvLME}(m_0) &= -n \log(k) + \sum_{i=1}^S \left[ \log \left( \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} \right) \right] \\
\text{cvLME}(m_1) &= S \cdot \log \frac{\Gamma(n_1)}{\Gamma(y)} + \sum_{i=1}^S \left[ \log \left( \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} \right) - \sum_{j=1}^k \log \left( \frac{y_{1j}^{(i)}}{y_j} \right) \right] .
\end{aligned} \tag{3}$$

where  $y_1^{(i)}$  and  $y_2^{(i)}$  are the training and test data, respectively, in the  $i$ -th cross-validation fold with  $n_1$  and  $n_2$  data points, respectively,  $S$  is the number of data subsets ( $\rightarrow$  IV/3.1.9) and  $\Gamma(x)$  is the gamma function.

**Proof:** For evaluation of the cross-validated log model evidences ( $\rightarrow$  IV/3.1.9) (cvLME), we assume that  $n$  data points are divided into  $S \mid n$  data subsets without remainder. Then, the number of training data points  $n_1$  and test data points  $n_2$  are given by

$$\begin{aligned} n &= n_1 + n_2 \\ n_1 &= \frac{S-1}{S}n \\ n_2 &= \frac{1}{S}n, \end{aligned} \quad (4)$$

such that  $y_1^{(i)}$  are the number of category observations ( $\rightarrow$  II/2.2.1) from the  $n_1$  trials of the training set ( $\rightarrow$  III/3.2.1) and  $y_2^{(i)}$  is the number of category observations ( $\rightarrow$  II/2.2.1) from the  $n_2$  trials of the test set ( $\rightarrow$  III/3.2.1) in the  $i$ -th cross-validation fold and it holds that

$$y = y_1^{(i)} + y_2^{(i)} \quad (5)$$

as well as

$$\begin{aligned} n &= \sum_{j=1}^k y_j \\ n_1 &= \sum_{j=1}^k y_{1j}^{(i)} \\ n_2 &= \sum_{j=1}^k y_{2j}^{(i)}. \end{aligned} \quad (6)$$

First, we consider the null model  $m_0$  assuming  $p_j = 1/k$  for each  $j = 1, \dots, k$ . Because this model has no free parameter, nothing is estimated from the training data and the assumed parameter value is applied to the test data. Consequently, the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) is equal to the log-likelihood function ( $\rightarrow$  III/3.2.3) of the test data at  $p = [1/k, \dots, 1/k]$ :

$$\begin{aligned} \text{oosLME}_i(m_0) &= \log p \left( y_2^{(i)} \mid p = [1/k, \dots, 1/k] \right) \\ &= \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + \sum_{j=1}^k y_{2j}^{(i)} \log(1/k) \\ &= \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + n_2 \log \frac{1}{k}. \end{aligned} \quad (7)$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, such that the cvLME of  $m_0$  is:

$$\begin{aligned}
\text{cvLME}(m_0) &= \sum_{i=1}^S \text{oosLME}_i(m_0) \\
&= \sum_{i=1}^S \left[ \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + n_2 \log \left( \frac{1}{k} \right) \right] \\
&= S \cdot n_2 \log \left( \frac{1}{k} \right) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} \right] \\
&= -n \log(k) + \sum_{i=1}^S \left[ \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} \right].
\end{aligned} \tag{8}$$

Next, we have a look at the alternative  $m_1$  assuming  $p_j = 1/k$  for at least one  $j = 1, \dots, k$ . First, the training data  $y_1^{(i)}$  are analyzed using a non-informative prior distribution ( $\rightarrow$  I/5.2.3) and applying the posterior distribution for multinomial observations ( $\rightarrow$  III/3.2.7):

$$\alpha_0^{(1)} = 0_{1k} \quad \Rightarrow \quad \alpha_n^{(1)} = \alpha_0^{(1)} + y_1^{(i)} = y_1^{(i)}. \tag{9}$$

This results in a posterior characterized by  $\alpha_n^{(1)}$ . Then, the test data  $y_2^{(i)}$  are analyzed using this posterior as an informative prior distribution ( $\rightarrow$  I/5.2.3), again applying the posterior distribution for multinomial observations ( $\rightarrow$  III/3.2.7):

$$\alpha_0^{(2)} = \alpha_n^{(1)} = y_1^{(i)} \quad \Rightarrow \quad \alpha_n^{(2)} = \alpha_0^{(2)} + y_2^{(i)} = y. \tag{10}$$

In the test data, we now have a prior characterized by  $\alpha_0^{(2)}$  and a posterior characterized  $\alpha_n^{(2)}$ . Applying the log model evidence for multinomial observations ( $\rightarrow$  III/3.2.8), the out-of-sample log model evidence ( $\rightarrow$  IV/3.1.9) (oosLME) therefore follows as

$$\begin{aligned}
\text{oosLME}_i(m_1) &= \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + \log \Gamma \left( \sum_{j=1}^k \alpha_{0j}^{(2)} \right) - \log \Gamma \left( \sum_{j=1}^k \alpha_{nj}^{(2)} \right) \\
&\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}^{(2)}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}^{(2)}) \\
&= \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + \log \Gamma \left( \sum_{j=1}^k y_{1j}^{(i)} \right) - \log \Gamma \left( \sum_{j=1}^k y_j \right) \\
&\quad + \sum_{j=1}^k \log \Gamma(y_j) - \sum_{j=1}^k \log \Gamma(y_{1j}^{(i)}) \\
&= \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + \log \frac{\Gamma(n_1)}{\Gamma(y)} - \sum_{j=1}^k \log \left( \frac{y_{1j}^{(i)}}{y_j} \right).
\end{aligned} \tag{11}$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences ( $\rightarrow$  IV/3.1.9) over cross-validation folds, the cvLME of  $m_1$  becomes:

$$\begin{aligned}
\text{cvLME}(m_1) &= \sum_{i=1}^S \text{oosLME}_i(m_1) \\
&= \sum_{i=1}^S \left[ \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} + \log \frac{\Gamma(n_1)}{\Gamma(y)} - \sum_{j=1}^k \log \left( \frac{y_{1j}^{(i)}}{y_j} \right) \right] \\
&= S \cdot \log \frac{\Gamma(n_1)}{\Gamma(y)} + \sum_{i=1}^S \left[ \log \binom{n_2}{y_{21}^{(i)}, \dots, y_{2k}^{(i)}} - \sum_{j=1}^k \log \left( \frac{y_{1j}^{(i)}}{y_j} \right) \right].
\end{aligned} \tag{12}$$

Together, (8) and (12) conform to the results given in (3). ■

### 3.3 Poisson-distributed data

#### 3.3.1 Definition

**Definition:** Poisson-distributed data are defined as a set of observed counts  $y = \{y_1, \dots, y_n\}$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a Poisson distribution ( $\rightarrow$  II/1.4.1) with rate  $\lambda$ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \tag{1}$$

#### Sources:

- Wikipedia (2020): “Poisson distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: [https://en.wikipedia.org/wiki/Poisson\\_distribution#Parameter\\_estimation](https://en.wikipedia.org/wiki/Poisson_distribution#Parameter_estimation).

#### 3.3.2 Maximum likelihood estimation

**Theorem:** Let there be a Poisson-distributed data ( $\rightarrow$  III/3.3.1) set  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \tag{1}$$

Then, the maximum likelihood estimate ( $\rightarrow$  I/4.1.3) for the rate parameter  $\lambda$  is given by

$$\hat{\lambda} = \bar{y} \tag{2}$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \tag{3}$$

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for each observation is given by the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2)

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \tag{4}$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} . \quad (5)$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$LL(\lambda) = \log p(y|\lambda) = \log \left[ \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \quad (6)$$

which can be developed into

$$\begin{aligned} LL(\lambda) &= \sum_{i=1}^n \log \left[ \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\ &= \sum_{i=1}^n [y_i \cdot \log(\lambda) - \lambda - \log(y_i!)] \\ &= - \sum_{i=1}^n \lambda + \sum_{i=1}^n y_i \cdot \log(\lambda) - \sum_{i=1}^n \log(y_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!) \end{aligned} \quad (7)$$

The derivatives of the log-likelihood with respect to  $\lambda$  are

$$\begin{aligned} \frac{dLL(\lambda)}{d\lambda} &= \frac{1}{\lambda} \sum_{i=1}^n y_i - n \\ \frac{d^2LL(\lambda)}{d\lambda^2} &= -\frac{1}{\lambda^2} \sum_{i=1}^n y_i . \end{aligned} \quad (8)$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned} \frac{dLL(\hat{\lambda})}{d\lambda} &= 0 \\ 0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^n y_i - n \\ \hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} . \end{aligned} \quad (9)$$

Plugging this value into the second derivative, we confirm:

$$\begin{aligned} \frac{d^2LL(\hat{\lambda})}{d\lambda^2} &= -\frac{1}{\bar{y}^2} \sum_{i=1}^n y_i \\ &= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\ &= -\frac{n}{\bar{y}} < 0 . \end{aligned} \quad (10)$$

This demonstrates that the estimate  $\hat{\lambda} = \bar{y}$  maximizes the likelihood  $p(y|\lambda)$ .

■

### 3.3.3 Conjugate prior distribution

**Theorem:** Let there be a Poisson-distributed data ( $\rightarrow$  III/3.3.1) set  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for the model parameter  $\lambda$  is a gamma distribution ( $\rightarrow$  II/3.4.1):

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0). \quad (2)$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \quad (3)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!}. \quad (4)$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^n \frac{1}{y_i!} \cdot \prod_{i=1}^n \lambda^{y_i} \cdot \prod_{i=1}^n \exp[-\lambda] \\ &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \cdot \lambda^{n\bar{y}} \cdot \exp[-n\lambda] \end{aligned} \quad (5)$$

where  $\bar{y}$  is the mean ( $\rightarrow$  I/1.12.2) of  $y$ :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (6)$$

In other words, the likelihood function is proportional to a power of  $\lambda$  times an exponential of  $\lambda$ :

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp[-n\lambda]. \quad (7)$$

The same is true for a gamma distribution over  $\lambda$

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) \quad (8)$$

the probability density function of which ( $\rightarrow$  II/3.4.7)

$$p(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \quad (9)$$

exhibits the same proportionality

$$p(\lambda) \propto \lambda^{a_0-1} \cdot \exp[-b_0\lambda] \quad (10)$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Other standard single-parameter models”; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: <http://www.stat.columbia.edu/~gelman/book/>.

### 3.3.4 Posterior distribution

**Theorem:** Let there be a Poisson-distributed data ( $\rightarrow$  III/3.3.1) set  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Moreover, assume a gamma prior distribution ( $\rightarrow$  III/3.3.3) over the model parameter  $\lambda$ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0). \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a gamma distribution ( $\rightarrow$  II/3.4.1)

$$p(\lambda|y) = \text{Gam}(\lambda; a_n, b_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n. \end{aligned} \quad (4)$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \quad (5)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!}. \quad (6)$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda]. \end{aligned} \quad (7)$$



Resolving the product in the joint likelihood, we have

$$\begin{aligned}
 p(y, \lambda) &= \prod_{i=1}^n \frac{1}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
 &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
 &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\lambda)]
 \end{aligned} \tag{8}$$

where  $\bar{y}$  is the mean ( $\rightarrow$  I/1.12.2) of  $y$ :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \tag{9}$$

Note that the posterior distribution is proportional to the joint likelihood ( $\rightarrow$  I/5.1.10):

$$p(\lambda|y) \propto p(y, \lambda) . \tag{10}$$

Setting  $a_n = a_0 + n\bar{y}$  and  $b_n = b_0 + n$ , the posterior distribution is therefore proportional to

$$p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp[-b_n \lambda] \tag{11}$$

which, when normalized to one, results in the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] = \text{Gam}(\lambda; a_n, b_n) . \tag{12}$$

■

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Other standard single-parameter models”; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: <http://www.stat.columbia.edu/~gelman/book/>.

### 3.3.5 Log model evidence

**Theorem:** Let there be a Poisson-distributed data ( $\rightarrow$  III/3.3.1) set  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \text{Pois}(\lambda), \quad i = 1, \dots, n . \tag{1}$$

Moreover, assume a gamma prior distribution ( $\rightarrow$  III/3.3.3) over the model parameter  $\lambda$ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) . \tag{2}$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\log p(y|m) = - \sum_{i=1}^n \log y_i! + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \tag{3}$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n . \end{aligned} \quad (4)$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \quad (5)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} . \quad (6)$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] . \end{aligned} \quad (7)$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^n \frac{1}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\ &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\ &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\lambda)] \end{aligned} \quad (8)$$

where  $\bar{y}$  is the mean ( $\rightarrow$  I/1.12.2) of  $y$ :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \quad (9)$$

Note that the model evidence is the marginal density of the joint likelihood ( $\rightarrow$  I/5.1.15):

$$p(y) = \int p(y, \lambda) d\lambda . \quad (10)$$

Setting  $a_n = a_0 + n\bar{y}$  and  $b_n = b_0 + n$ , the joint likelihood can also be written as

$$p(y, \lambda) = \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n\lambda] . \quad (11)$$

Using the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7),  $\lambda$  can now be integrated out easily

$$\begin{aligned} p(y) &= \int \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] d\lambda \\ &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} \int \text{Gam}(\lambda; a_n, b_n) d\lambda \\ &= \prod_{i=1}^n \left( \frac{1}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}}, \end{aligned} \quad (12)$$

such that the log model evidence ( $\rightarrow$  IV/3.1.3) is shown to be

$$\log p(y|m) = - \sum_{i=1}^n \log y_i! + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n. \quad (13)$$

■

### 3.4 Poisson distribution with exposure values

#### 3.4.1 Definition

**Definition:** A Poisson distribution with exposure values is defined as a set of observed counts  $y = \{y_1, \dots, y_n\}$ , independently ( $\rightarrow$  I/1.3.6) distributed ( $\rightarrow$  I/1.5.1) according to a Poisson distribution ( $\rightarrow$  II/1.4.1) with common rate ( $\rightarrow$  I/1.5.10)  $\lambda$  and a set of concurrent exposures  $x = \{x_1, \dots, x_n\}$ :

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Other standard single-parameter models”; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14; URL: <http://www.stat.columbia.edu/~gelman/book/>.

#### 3.4.2 Maximum likelihood estimation

**Theorem:** Consider data  $y = \{y_1, \dots, y_n\}$  following a Poisson distribution with exposure values ( $\rightarrow$  III/3.4.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate ( $\rightarrow$  I/4.1.3) for the rate parameter  $\lambda$  is given by

$$\hat{\lambda} = \frac{\bar{y}}{\bar{x}} \quad (2)$$

where  $\bar{y}$  and  $\bar{x}$  are the sample means ( $\rightarrow$  I/1.12.2)

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i.\end{aligned}\tag{3}$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}\tag{4}$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}.\tag{5}$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$\text{LL}(\lambda) = \log p(y|\lambda) = \log \left[ \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right]\tag{6}$$

which can be developed into

$$\begin{aligned}\text{LL}(\lambda) &= \sum_{i=1}^n \log \left[ \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right] \\ &= \sum_{i=1}^n [y_i \cdot \log(\lambda x_i) - \lambda x_i - \log(y_i!)] \\ &= -\sum_{i=1}^n \lambda x_i + \sum_{i=1}^n y_i \cdot [\log(\lambda) + \log(x_i)] - \sum_{i=1}^n \log(y_i!) \\ &= -\lambda \sum_{i=1}^n x_i + \log(\lambda) \sum_{i=1}^n y_i + \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log(y_i!) \\ &= -n\bar{x}\lambda + n\bar{y} \log(\lambda) + \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log(y_i!)\end{aligned}\tag{7}$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means from equation (3).

The derivatives of the log-likelihood with respect to  $\lambda$  are

$$\begin{aligned}\frac{d\text{LL}(\lambda)}{d\lambda} &= -n\bar{x} + \frac{n\bar{y}}{\lambda} \\ \frac{d^2\text{LL}(\lambda)}{d\lambda^2} &= -\frac{n\bar{y}}{\lambda^2}.\end{aligned}\tag{8}$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned}\frac{dLL(\hat{\lambda})}{d\lambda} &= 0 \\ 0 &= -n\bar{x} + \frac{n\bar{y}}{\hat{\lambda}} \\ \hat{\lambda} &= \frac{n\bar{y}}{n\bar{x}} = \frac{\bar{y}}{\bar{x}}.\end{aligned}\tag{9}$$

Plugging this value into the second derivative, we confirm:

$$\begin{aligned}\frac{d^2LL(\hat{\lambda})}{d\lambda^2} &= -\frac{n\bar{y}}{\hat{\lambda}^2} \\ &= -\frac{n \cdot \bar{y}}{(\bar{y}/\bar{x})^2} \\ &= -\frac{n \cdot \bar{x}^2}{\bar{y}} < 0.\end{aligned}\tag{10}$$

This demonstrates that the estimate  $\hat{\lambda} = \bar{y}/\bar{x}$  maximizes the likelihood  $p(y|\lambda)$ . ■

### 3.4.3 Conjugate prior distribution

**Theorem:** Consider data  $y = \{y_1, \dots, y_n\}$  following a Poisson distribution with exposure values ( $\rightarrow$  III/3.4.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n.\tag{1}$$

Then, the conjugate prior ( $\rightarrow$  I/5.2.5) for the model parameter  $\lambda$  is a gamma distribution ( $\rightarrow$  II/3.4.1):

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0).\tag{2}$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}\tag{3}$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}.\tag{4}$$

Resolving the product in the likelihood function, we have

$$\begin{aligned}
p(y|\lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \cdot \prod_{i=1}^n \lambda^{y_i} \cdot \prod_{i=1}^n \exp[-\lambda x_i] \\
&= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{\sum_{i=1}^n y_i} \cdot \exp \left[ -\lambda \sum_{i=1}^n x_i \right] \\
&= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda]
\end{aligned} \tag{5}$$

where  $\bar{y}$  and  $\bar{x}$  are the means ( $\rightarrow$  I/1.12.2) of  $y$  and  $x$  respectively:

$$\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i .
\end{aligned} \tag{6}$$

In other words, the likelihood function is proportional to a power of  $\lambda$  times an exponential of  $\lambda$ :

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda] . \tag{7}$$

The same is true for a gamma distribution over  $\lambda$

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) \tag{8}$$

the probability density function of which ( $\rightarrow$  II/3.4.7)

$$p(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \tag{9}$$

exhibits the same proportionality

$$p(\lambda) \propto \lambda^{a_0-1} \cdot \exp[-b_0\lambda] \tag{10}$$

and is therefore conjugate relative to the likelihood. ■

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Other standard single-parameter models”; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: <http://www.stat.columbia.edu/~gelman/book/>.

### 3.4.4 Posterior distribution

**Theorem:** Consider data  $y = \{y_1, \dots, y_n\}$  following a Poisson distribution with exposure values ( $\rightarrow$  III/3.4.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n . \tag{1}$$

Moreover, assume a gamma prior distribution ( $\rightarrow$  III/3.4.3) over the model parameter  $\lambda$ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) . \quad (2)$$

Then, the posterior distribution ( $\rightarrow$  I/5.1.9) is also a gamma distribution ( $\rightarrow$  II/3.4.1)

$$p(\lambda|y) = \text{Gam}(\lambda; a_n, b_n) \quad (3)$$

and the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n\bar{x} . \end{aligned} \quad (4)$$

**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \quad (5)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} . \quad (6)$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] . \end{aligned} \quad (7)$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda x_i] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^n y_i} \exp \left[ -\lambda \sum_{i=1}^n x_i \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\bar{x}\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\bar{x})\lambda] \end{aligned} \quad (8)$$

where  $\bar{y}$  and  $\bar{x}$  are the means ( $\rightarrow$  I/1.12.2) of  $y$  and  $x$  respectively:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i .\end{aligned}\tag{9}$$

Note that the posterior distribution is proportional to the joint likelihood ( $\rightarrow$  I/5.1.10):

$$p(\lambda|y) \propto p(y, \lambda) .\tag{10}$$

Setting  $a_n = a_0 + n\bar{y}$  and  $b_n = b_0 + n\bar{x}$ , the posterior distribution is therefore proportional to

$$p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp[-b_n \lambda]\tag{11}$$

which, when normalized to one, results in the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] = \text{Gam}(\lambda; a_n, b_n) .\tag{12}$$

■

#### Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): “Other standard single-parameter models”; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: <http://www.stat.columbia.edu/~gelman/book/>.

### 3.4.5 Log model evidence

**Theorem:** Consider data  $y = \{y_1, \dots, y_n\}$  following a Poisson distribution with exposure values ( $\rightarrow$  III/3.4.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n .\tag{1}$$

Moreover, assume a gamma prior distribution ( $\rightarrow$  III/3.4.3) over the model parameter  $\lambda$ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) .\tag{2}$$

Then, the log model evidence ( $\rightarrow$  IV/3.1.3) for this model is

$$\begin{aligned}\log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\ &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n .\end{aligned}\tag{3}$$

where the posterior hyperparameters ( $\rightarrow$  I/5.1.9) are given by

$$\begin{aligned}a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n\bar{x} .\end{aligned}\tag{4}$$



**Proof:** With the probability mass function of the Poisson distribution ( $\rightarrow$  II/1.4.2), the likelihood function ( $\rightarrow$  I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \quad (5)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} . \quad (6)$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood ( $\rightarrow$  I/5.1.7) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] . \end{aligned} \quad (7)$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda x_i] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^n y_i} \exp \left[ -\lambda \sum_{i=1}^n x_i \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\bar{x}\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\bar{x})\lambda] \end{aligned} \quad (8)$$

where  $\bar{y}$  and  $\bar{x}$  are the means ( $\rightarrow$  I/1.12.2) of  $y$  and  $x$  respectively:

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i . \end{aligned} \quad (9)$$

Note that the model evidence is the marginal density of the joint likelihood ( $\rightarrow$  I/5.1.15):

$$p(y) = \int p(y, \lambda) d\lambda . \quad (10)$$

Setting  $a_n = a_0 + n\bar{y}$  and  $b_n = b_0 + n\bar{x}$ , the joint likelihood can also be written as

$$p(y, \lambda) = \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] . \quad (11)$$

Using the probability density function of the gamma distribution ( $\rightarrow$  II/3.4.7),  $\lambda$  can now be integrated out easily

$$\begin{aligned}
 p(y) &= \int \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] d\lambda \\
 &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} \int \text{Gam}(\lambda; a_n, b_n) d\lambda \\
 &= \prod_{i=1}^n \left( \frac{x_i^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}},
 \end{aligned} \tag{12}$$

such that the log model evidence ( $\rightarrow$  IV/3.1.3) is shown to be

$$\begin{aligned}
 \log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\
 &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n.
 \end{aligned} \tag{13}$$

■

## 4 Frequency data

### 4.1 Beta-distributed data

#### 4.1.1 Definition

**Definition:** Beta-distributed data are defined as a set of proportions  $y = \{y_1, \dots, y_n\}$  with  $y_i \in [0, 1]$ ,  $i = 1, \dots, n$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a beta distribution ( $\rightarrow$  II/3.9.1) with shapes  $\alpha$  and  $\beta$ :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

#### 4.1.2 Method of moments

**Theorem:** Let  $y = \{y_1, \dots, y_n\}$  be a set of observed counts independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a beta distribution ( $\rightarrow$  II/3.9.1) with shapes  $\alpha$  and  $\beta$ :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

Then, the method-of-moments estimates ( $\rightarrow$  I/4.1.8) for the shape parameters  $\alpha$  and  $\beta$  are given by

$$\begin{aligned} \hat{\alpha} &= \bar{y} \left( \frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) \\ \hat{\beta} &= (1 - \bar{y}) \left( \frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) \end{aligned} \quad (2)$$

where  $\bar{y}$  is the sample mean ( $\rightarrow$  I/1.12.2) and  $\bar{v}$  is the unbiased sample variance ( $\rightarrow$  I/1.13.2):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (3)$$

**Proof:** Mean ( $\rightarrow$  II/3.9.8) and variance ( $\rightarrow$  II/3.9.11) of the beta distribution ( $\rightarrow$  II/3.9.1) in terms of the parameters  $\alpha$  and  $\beta$  are given by

$$\begin{aligned} E(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (4)$$

Thus, matching the moments ( $\rightarrow$  I/4.1.8) requires us to solve the following equation system for  $\alpha$  and  $\beta$ :

$$\begin{aligned} \bar{y} &= \frac{\alpha}{\alpha + \beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (5)$$

From the first equation, we can deduce:

$$\begin{aligned}
 \bar{y}(\alpha + \beta) &= \alpha \\
 \alpha\bar{y} + \beta\bar{y} &= \alpha \\
 \beta\bar{y} &= \alpha - \alpha\bar{y} \\
 \beta &= \frac{\alpha}{\bar{y}} - \alpha \\
 \beta &= \alpha \left( \frac{1}{\bar{y}} - 1 \right) .
 \end{aligned} \tag{6}$$

If we define  $q = 1/\bar{y} - 1$  and plug (6) into the second equation, we have:

$$\begin{aligned}
 \bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2 (\alpha + \alpha q + 1)} \\
 &= \frac{\alpha^2 q}{(\alpha(1 + q))^2 (\alpha(1 + q) + 1)} \\
 &= \frac{q}{(1 + q)^2 (\alpha(1 + q) + 1)} \\
 &= \frac{q}{\alpha(1 + q)^3 + (1 + q)^2} \\
 q &= \bar{v} [\alpha(1 + q)^3 + (1 + q)^2] .
 \end{aligned} \tag{7}$$

Noting that  $1 + q = 1/\bar{y}$  and  $q = (1 - \bar{y})/\bar{y}$ , one obtains for  $\alpha$ :

$$\begin{aligned}
 \frac{1 - \bar{y}}{\bar{y}} &= \bar{v} \left[ \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \right] \\
 \frac{1 - \bar{y}}{\bar{y} \bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\
 \frac{\bar{y}^3(1 - \bar{y})}{\bar{y} \bar{v}} &= \alpha + \bar{y} \\
 \alpha &= \frac{\bar{y}^2(1 - \bar{y})}{\bar{v}} - \bar{y} \\
 &= \bar{y} \left( \frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) .
 \end{aligned} \tag{8}$$

Plugging this into equation (6), one obtains for  $\beta$ :

$$\begin{aligned}
 \beta &= \bar{y} \left( \frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) \cdot \left( \frac{1 - \bar{y}}{\bar{y}} \right) \\
 &= (1 - \bar{y}) \left( \frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) .
 \end{aligned} \tag{9}$$

Together, (8) and (9) constitute the method-of-moment estimates of  $\alpha$  and  $\beta$ . ■

#### Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: [https://en.wikipedia.org/wiki/Beta\\_distribution#Method\\_of\\_moments](https://en.wikipedia.org/wiki/Beta_distribution#Method_of_moments).

## 4.2 Dirichlet-distributed data

### 4.2.1 Definition

**Definition:** Dirichlet-distributed data are defined as a set of vectors of proportions  $y = \{y_1, \dots, y_n\}$  where

$$\begin{aligned} y_i &= [y_{i1}, \dots, y_{ik}], \\ y_{ij} &\in [0, 1] \quad \text{and} \\ \sum_{j=1}^k y_{ij} &= 1 \end{aligned} \quad (1)$$

for all  $i = 1, \dots, n$  (and  $j = 1, \dots, k$ ) and each  $y_i$  is independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a Dirichlet distribution ( $\rightarrow$  II/4.5.1) with concentration parameters  $\alpha = [\alpha_1, \dots, \alpha_k]$ :

$$y_i \sim \text{Dir}(\alpha), \quad i = 1, \dots, n. \quad (2)$$

### 4.2.2 Maximum likelihood estimation

**Theorem:** Let there be a Dirichlet-distributed data ( $\rightarrow$  III/4.2.1) set  $y = \{y_1, \dots, y_n\}$ :

$$y_i \sim \text{Dir}(\alpha), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate ( $\rightarrow$  I/4.1.3) for the concentration parameters  $\alpha$  can be obtained by iteratively computing

$$\alpha_j^{(\text{new})} = \psi^{-1} \left[ \psi \left( \sum_{j=1}^k \alpha_j^{(\text{old})} \right) + \log \bar{y}_j \right] \quad (2)$$

where  $\psi(x)$  is the digamma function and  $\log \bar{y}_j$  is given by:

$$\log \bar{y}_j = \frac{1}{n} \sum_{i=1}^n \log y_{ij}. \quad (3)$$

**Proof:** The likelihood function ( $\rightarrow$  I/5.1.2) for each observation is given by the probability density function of the Dirichlet distribution ( $\rightarrow$  II/4.5.2)

$$p(y_i|\alpha) = \frac{\Gamma \left( \sum_{j=1}^k \alpha_j \right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k y_{ij}^{\alpha_j-1} \quad (4)$$

and because observations are independent ( $\rightarrow$  I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\alpha) = \prod_{i=1}^n p(y_i|\alpha) = \prod_{i=1}^n \left[ \frac{\Gamma \left( \sum_{j=1}^k \alpha_j \right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k y_{ij}^{\alpha_j-1} \right]. \quad (5)$$

Thus, the log-likelihood function ( $\rightarrow$  I/4.1.2) is

$$\text{LL}(\alpha) = \log p(y|\alpha) = \log \prod_{i=1}^n \left[ \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k y_{ij}^{\alpha_j-1} \right] \quad (6)$$

which can be developed into

$$\begin{aligned} \text{LL}(\alpha) &= \sum_{i=1}^n \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - \sum_{i=1}^n \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{i=1}^n \sum_{j=1}^k (\alpha_j - 1) \log y_{ij} \\ &= n \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - n \sum_{j=1}^k \log \Gamma(\alpha_j) + n \sum_{j=1}^k (\alpha_j - 1) \frac{1}{n} \sum_{i=1}^n \log y_{ij} \\ &= n \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - n \sum_{j=1}^k \log \Gamma(\alpha_j) + n \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \end{aligned} \quad (7)$$

where we have specified

$$\log \bar{y}_j = \frac{1}{n} \sum_{i=1}^n \log y_{ij} . \quad (8)$$

The derivative of the log-likelihood with respect to a particular parameter  $\alpha_j$  is

$$\begin{aligned} \frac{d\text{LL}(\alpha)}{d\alpha_j} &= \frac{d}{d\alpha_j} \left[ n \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - n \sum_{j=1}^k \log \Gamma(\alpha_j) + n \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \right] \\ &= \frac{d}{d\alpha_j} \left[ n \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) \right] - \frac{d}{d\alpha_j} [n \log \Gamma(\alpha_j)] + \frac{d}{d\alpha_j} [n(\alpha_j - 1) \log \bar{y}_j] \\ &= n\psi\left(\sum_{j=1}^k \alpha_j\right) - n\psi(\alpha_j) + n \log \bar{y}_j \end{aligned} \quad (9)$$

where we have used the digamma function

$$\psi(x) = \frac{d \log \Gamma(x)}{dx} . \quad (10)$$

Setting this derivative to zero, we obtain:

$$\begin{aligned}
\frac{dLL(\alpha)}{d\alpha_j} &= 0 \\
0 &= n\psi\left(\sum_{j=1}^k \alpha_j\right) - n\psi(\alpha_j) + n \log \bar{y}_j \\
0 &= \psi\left(\sum_{j=1}^k \alpha_j\right) - \psi(\alpha_j) + \log \bar{y}_j \\
\psi(\alpha_j) &= \psi\left(\sum_{j=1}^k \alpha_j\right) + \log \bar{y}_j \\
\alpha_j &= \psi^{-1}\left[\psi\left(\sum_{j=1}^k \alpha_j\right) + \log \bar{y}_j\right].
\end{aligned} \tag{11}$$

In the following, we will use a fixed-point iteration to maximize  $LL(\alpha)$ . Given an initial guess for  $\alpha$ , we construct a lower bound on the likelihood function (7) which is tight at  $\alpha$ . The maximum of this bound is computed and it becomes the new guess. Because the Dirichlet distribution ( $\rightarrow$  II/4.5.1) belongs to the exponential family, the log-likelihood function is convex in  $\alpha$  and the maximum is the only stationary point, such that the procedure is guaranteed to converge to the maximum.

In our case, we use a bound on the gamma function

$$\begin{aligned}
\Gamma(x) &\geq \Gamma(\hat{x}) \cdot \exp[(x - \hat{x})\psi(\hat{x})] \\
\log \Gamma(x) &\geq \log \Gamma(\hat{x}) + (x - \hat{x})\psi(\hat{x})
\end{aligned} \tag{12}$$

and apply it to  $\Gamma\left(\sum_{j=1}^k \alpha_j\right)$  in (7) to yield

$$\begin{aligned}
\frac{1}{n}LL(\alpha) &= \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \\
\frac{1}{n}LL(\alpha) &\geq \log \Gamma\left(\sum_{j=1}^k \hat{\alpha}_j\right) + \left(\sum_{j=1}^k \alpha_j - \sum_{j=1}^k \hat{\alpha}_j\right) \psi\left(\sum_{j=1}^k \hat{\alpha}_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \\
\frac{1}{n}LL(\alpha) &\geq \left(\sum_{j=1}^k \alpha_j\right) \psi\left(\sum_{j=1}^k \hat{\alpha}_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j + \text{const.}
\end{aligned} \tag{13}$$

which leads to the following fixed-point iteration using (11):

$$\alpha_j^{(\text{new})} = \psi^{-1}\left[\psi\left(\sum_{j=1}^k \alpha_j^{(\text{old})}\right) + \log \bar{y}_j\right]. \tag{14}$$

■

#### Sources:

- Minka TP (2012): “Estimating a Dirichlet distribution”; in: *Papers by Tom Minka*, retrieved on 2020-10-22; URL: <https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf>.

### 4.3 Beta-binomial data

#### 4.3.1 Definition

**Definition:** Beta-binomial data are defined as a set of counts  $y = \{y_1, \dots, y_N\}$  with  $y_i \in \mathbb{N}$ ,  $i = 1, \dots, N$ , independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a beta-binomial distribution ( $\rightarrow$  II/1.3.1) with number of trials  $n$  as well as shapes  $\alpha$  and  $\beta$ :

$$y_i \sim \text{BetBin}(n, \alpha, \beta), \quad i = 1, \dots, N. \quad (1)$$

#### 4.3.2 Method of moments

**Theorem:** Let  $y = \{y_1, \dots, y_N\}$  be a set of observed counts, independent and identically distributed ( $\rightarrow$  I/1.2.8) according to a beta-binomial distribution ( $\rightarrow$  II/1.3.1) with number of trials  $n$  as well as parameters  $\alpha$  and  $\beta$ :

$$y_i \sim \text{BetBin}(n, \alpha, \beta), \quad i = 1, \dots, N. \quad (1)$$

Then, the method-of-moments estimates ( $\rightarrow$  I/4.1.8) for the parameters  $\alpha$  and  $\beta$  are given by

$$\begin{aligned} \hat{\alpha} &= \frac{nm_1 - m_2}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \\ \hat{\beta} &= \frac{(n - m_1) \left( n - \frac{m_2}{m_1} \right)}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \end{aligned} \quad (2)$$

where  $m_1$  and  $m_2$  are the first two raw sample moments ( $\rightarrow$  I/1.21.3):

$$\begin{aligned} m_1 &= \frac{1}{N} \sum_{i=1}^N y_i \\ m_2 &= \frac{1}{N} \sum_{i=1}^N y_i^2. \end{aligned} \quad (3)$$

**Proof:** The first two raw moments of the beta-binomial distribution in terms of the parameters  $\alpha$  and  $\beta$  are given by

$$\begin{aligned} \mu_1 &= \frac{n\alpha}{\alpha + \beta} \\ \mu_2 &= \frac{n\alpha(n\alpha + \beta + n)}{(\alpha + \beta)(n\alpha + \beta + 1)} \end{aligned} \quad (4)$$

Thus, matching the moments ( $\rightarrow$  I/4.1.8) requires us to solve the following equation system for  $\alpha$  and  $\beta$ :

$$\begin{aligned} m_1 &= \frac{n\alpha}{\alpha + \beta} \\ m_2 &= \frac{n\alpha(n\alpha + \beta + n)}{(\alpha + \beta)(n\alpha + \beta + 1)}. \end{aligned} \quad (5)$$



From the first equation, we can deduce:

$$\begin{aligned}
 m_1(\alpha + \beta) &= n\alpha \\
 m_1\alpha + m_1\beta &= n\alpha \\
 m_1\beta &= n\alpha - m_1\alpha \\
 \beta &= \frac{n\alpha}{m_1} - \alpha \\
 \beta &= \alpha \left( \frac{n}{m_1} - 1 \right) .
 \end{aligned} \tag{6}$$

If we define  $q = n/m_1 - 1$  and plug (6) into the second equation, we have:

$$\begin{aligned}
 m_2 &= \frac{n\alpha(n\alpha + \alpha q + n)}{(\alpha + \alpha q)(\alpha + \alpha q + 1)} \\
 &= \frac{n\alpha(\alpha(n + q) + n)}{\alpha(1 + q)(\alpha(1 + q) + 1)} \\
 &= \frac{n(\alpha(n + q) + n)}{(1 + q)(\alpha(1 + q) + 1)} \\
 &= \frac{n(\alpha(n + q) + n)}{\alpha(1 + q)^2 + (1 + q)} .
 \end{aligned} \tag{7}$$

Noting that  $1 + q = n/m_1$  and expanding the fraction with  $m_1$ , one obtains:

$$\begin{aligned}
 m_2 &= \frac{n \left( \alpha \left( \frac{n}{m_1} + n - 1 \right) + n \right)}{n \left( \alpha \frac{n}{m_1^2} + \frac{1}{m_1} \right)} \\
 m_2 &= \frac{\alpha (n + nm_1 - m_1) + nm_1}{\alpha \frac{n}{m_1} + 1} \\
 m_2 \left( \frac{\alpha n}{m_1} + 1 \right) &= \alpha (n + nm_1 - m_1) + nm_1 \\
 \alpha \left( n \frac{m_2}{m_1} - (n + nm_1 - m_1) \right) &= nm_1 - m_2 \\
 \alpha \left( n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1 \right) &= nm_1 - m_2 \\
 \hat{\alpha} &= \frac{nm_1 - m_2}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} .
 \end{aligned} \tag{8}$$

Plugging this into equation (6), one obtains for  $\beta$ :

$$\begin{aligned}
\beta &= \hat{\alpha} \left( \frac{n}{m_1} - 1 \right) \\
\beta &= \left( \frac{nm_1 - m_2}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \right) \left( \frac{n}{m_1} - 1 \right) \\
\beta &= \frac{n^2 - nm_1 - n \frac{m_2}{m_1} + m_2}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \\
\hat{\beta} &= \frac{(n - m_1) \left( n - \frac{m_2}{m_1} \right)}{n \left( \frac{m_2}{m_1} - m_1 - 1 \right) + m_1} .
\end{aligned} \tag{9}$$

Together, (8) and (9) constitute the method-of-moment estimates of  $\alpha$  and  $\beta$ .

■

#### Sources:

- statisticsmatt (2022): “Method of Moments Estimation Beta Binomial Distribution”; in: *YouTube*, retrieved on 2022-10-07; URL: <https://www.youtube.com/watch?v=18PWnWJsPnA>.
- Wikipedia (2022): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-07; URL: [https://en.wikipedia.org/wiki/Beta-binomial\\_distribution#Method\\_of\\_moments](https://en.wikipedia.org/wiki/Beta-binomial_distribution#Method_of_moments).

## 5 Categorical data

### 5.1 Logistic regression

#### 5.1.1 Definition

**Definition:** A logistic regression model is given by a set of binary observations  $y_i \in \{0, 1\}$ ,  $i = 1, \dots, n$ , a set of predictors  $x_j \in \mathbb{R}^n$ ,  $j = 1, \dots, p$ , a base  $b$  and the assumption that the log-odds are a linear combination of the predictors:

$$l_i = x_i \beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where  $l_i$  are the log-odds that  $y_i = 1$

$$l_i = \log_b \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} \quad (2)$$

and  $x_i$  is the  $i$ -th row of the  $n \times p$  matrix

$$X = [x_1, \dots, x_p] \quad (3)$$

Within this model,

- $y$  are called “categorical observations” or “dependent variable”;
- $X$  is called “design matrix” or “set of independent variables”;
- $\beta$  are called “regression coefficients” or “weights”;
- $\varepsilon_i$  is called “noise” or “error term”;
- $n$  is the number of observations;
- $p$  is the number of predictors.

#### Sources:

- Wikipedia (2020): “Logistic regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-28; URL: [https://en.wikipedia.org/wiki/Logistic\\_regression#Logistic\\_model](https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model).

#### 5.1.2 Probability and log-odds

**Theorem:** Assume a logistic regression model ( $\rightarrow$  III/5.1.1)

$$l_i = x_i \beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where  $x_i$  are the predictors corresponding to the  $i$ -th observation  $y_i$  and  $l_i$  are the log-odds that  $y_i = 1$ .

Then, the log-odds in favor of  $y_i = 1$  against  $y_i = 0$  can also be expressed as

$$l_i = \log_b \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 0) p(y_i = 0)} \quad (2)$$

where  $p(x_i|y_i)$  is a likelihood function ( $\rightarrow$  I/5.1.2) consistent with (1),  $p(y_i)$  are prior probabilities ( $\rightarrow$  I/5.1.3) for  $y_i = 1$  and  $y_i = 0$  and where  $b$  is the base used to form the log-odds  $l_i$ .

**Proof:** Using Bayes’ theorem ( $\rightarrow$  I/5.3.2) and the law of marginal probability ( $\rightarrow$  I/1.3.3), the posterior probabilities ( $\rightarrow$  I/5.1.9) for  $y_i = 1$  and  $y_i = 0$  are given by

$$\begin{aligned}
p(y_i = 1|x_i) &= \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)} \\
p(y_i = 0|x_i) &= \frac{p(x_i|y_i = 0) p(y_i = 0)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)} .
\end{aligned} \tag{3}$$

Calculating the log-odds from the posterior probabilities, we have

$$\begin{aligned}
l_i &= \log_b \frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} \\
&= \log_b \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 0) p(y_i = 0)} .
\end{aligned} \tag{4}$$

■

#### Sources:

- Bishop, Christopher M. (2006): “Linear Models for Classification”; in: *Pattern Recognition for Machine Learning*, ch. 4, p. 197, eq. 4.58; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

### 5.1.3 Log-odds and probability

**Theorem:** Assume a logistic regression model ( $\rightarrow$  III/5.1.1)

$$l_i = x_i \beta + \varepsilon_i, \quad i = 1, \dots, n \tag{1}$$

where  $x_i$  are the predictors corresponding to the  $i$ -th observation  $y_i$  and  $l_i$  are the log-odds that  $y_i = 1$ .

Then, the probability that  $y_i = 1$  is given by

$$\Pr(y_i = 1) = \frac{1}{1 + b^{-(x_i \beta + \varepsilon_i)}} \tag{2}$$

where  $b$  is the base used to form the log-odds  $l_i$ .

**Proof:** Let us denote  $\Pr(y_i = 1)$  as  $p_i$ . Then, the log-odds are

$$l_i = \log_b \frac{p_i}{1 - p_i} \tag{3}$$

and using (1), we have

$$\begin{aligned}
\log_b \frac{p_i}{1-p_i} &= x_i \beta + \varepsilon_i \\
\frac{p_i}{1-p_i} &= b^{x_i \beta + \varepsilon_i} \\
p_i &= (b^{x_i \beta + \varepsilon_i}) (1-p_i) \\
p_i (1 + b^{x_i \beta + \varepsilon_i}) &= b^{x_i \beta + \varepsilon_i} \\
p_i &= \frac{b^{x_i \beta + \varepsilon_i}}{1 + b^{x_i \beta + \varepsilon_i}} \\
p_i &= \frac{b^{x_i \beta + \varepsilon_i}}{b^{x_i \beta + \varepsilon_i} (1 + b^{-(x_i \beta + \varepsilon_i)})} \\
p_i &= \frac{1}{1 + b^{-(x_i \beta + \varepsilon_i)}}
\end{aligned} \tag{4}$$

which proves the identity given by (2).

■

#### Sources:

- Wikipedia (2020): “Logistic regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-03; URL: [https://en.wikipedia.org/wiki/Logistic\\_regression#Logistic\\_model](https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model).



# Chapter IV

## Model Selection

# 1 Goodness-of-fit measures

## 1.0.1 Definition

**Definition:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

with measured data  $y$ , known design matrix  $X$  and covariance structure  $V$  as well as unknown regression coefficients  $\beta$  and noise variance  $\sigma^2$ .

Then, an estimate of the noise variance  $\sigma^2$  is called the “residual variance”  $\hat{\sigma}^2$ , e.g. obtained via maximum likelihood estimation ( $\rightarrow$  I/4.1.3).

## 1.0.2 Maximum likelihood estimator is biased ( $p = 1$ )

**Theorem:** Let  $y = \{y_1, \dots, y_n\}$  be a set of independent normally distributed ( $\rightarrow$  II/3.2.1) observations with unknown mean ( $\rightarrow$  I/1.12.1)  $\mu$  and variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ :

$$y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

Then,

1) the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2)$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (3)$$

2) and  $\hat{\sigma}^2$  is a biased estimator ( $\rightarrow$  I/3.1.2) of  $\sigma^2$

$$\mathbb{E} [\hat{\sigma}^2] \neq \sigma^2, \quad (4)$$

more precisely:

$$\mathbb{E} [\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2. \quad (5)$$

**Proof:**

1) This is equivalent to the maximum likelihood estimator for the univariate Gaussian with unknown variance ( $\rightarrow$  III/1.1.2) and a special case of the maximum likelihood estimator for multiple linear regression ( $\rightarrow$  III/1.5.23) in which  $X = 1_n$  and  $\hat{\beta} = \bar{y}$ :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= \frac{1}{n} (y - 1_n \bar{y})^T (y - 1_n \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (6)$$



2) The expectation ( $\rightarrow$  I/1.12.1) of the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) can be developed as follows:

$$\begin{aligned}
 E[\hat{\sigma}^2] &= E\left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2\right] \\
 &= \frac{1}{n} E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] \\
 &= \frac{1}{n} E\left[\sum_{i=1}^n (y_i^2 - 2y_i\bar{y} + \bar{y}^2)\right] \\
 &= \frac{1}{n} E\left[\sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n y_i\bar{y} + \sum_{i=1}^n \bar{y}^2\right] \\
 &= \frac{1}{n} E\left[\sum_{i=1}^n y_i^2 - 2n\bar{y}^2 + n\bar{y}^2\right] \\
 &= \frac{1}{n} E\left[\sum_{i=1}^n y_i^2 - n\bar{y}^2\right] \\
 &= \frac{1}{n} \left(\sum_{i=1}^n E[y_i^2] - nE[\bar{y}^2]\right) \\
 &= \frac{1}{n} \sum_{i=1}^n E[y_i^2] - E[\bar{y}^2]
 \end{aligned} \tag{7}$$

Due to the partition of variance into expected values ( $\rightarrow$  I/1.13.4)

$$\text{Var}(X) = E(X^2) - E(X)^2, \tag{8}$$

we have

$$\begin{aligned}
 \text{Var}(y_i) &= E(y_i^2) - E(y_i)^2 \\
 \text{Var}(\bar{y}) &= E(\bar{y}^2) - E(\bar{y})^2,
 \end{aligned} \tag{9}$$

such that (7) becomes

$$E[\hat{\sigma}^2] = \frac{1}{n} \sum_{i=1}^n (\text{Var}(y_i) + E(y_i)^2) - (\text{Var}(\bar{y}) + E(\bar{y})^2). \tag{10}$$

From (1), it follows that

$$E(y_i) = \mu \quad \text{and} \quad \text{Var}(y_i) = \sigma^2. \tag{11}$$

The expectation ( $\rightarrow$  I/1.12.1) of  $\bar{y}$  given by (3) is

$$\begin{aligned}
\mathbb{E}[\bar{y}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n y_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[y_i] \\
&\stackrel{(11)}{=} \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu \\
&= \mu .
\end{aligned} \tag{12}$$

The variance of  $\bar{y}$  given by (3) is

$$\begin{aligned}
\text{Var}[\bar{y}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n y_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[y_i] \\
&\stackrel{(11)}{=} \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\
&= \frac{1}{n} \sigma^2 .
\end{aligned} \tag{13}$$

Plugging (11), (12) and (13) into (10), we have

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\
\mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \cdot n \cdot (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\
\mathbb{E}[\hat{\sigma}^2] &= \sigma^2 + \mu^2 - \frac{1}{n} \sigma^2 - \mu^2 \\
\mathbb{E}[\hat{\sigma}^2] &= \frac{n-1}{n} \sigma^2
\end{aligned} \tag{14}$$

which proves the bias ( $\rightarrow$  I/3.1.2) given by (5). ■

#### Sources:

- Liang, Dawen (????): “Maximum Likelihood Estimator for Variance is Biased: Proof”, retrieved on 2020-02-24; URL: [https://dawenl.github.io/files/mle\\_biased.pdf](https://dawenl.github.io/files/mle_biased.pdf).

### 1.0.3 Maximum likelihood estimator is biased ( $p > 1$ )

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \tag{1}$$

Then,

1) the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) \quad (2)$$

where

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \quad (3)$$

2) and  $\hat{\sigma}^2$  is a biased estimator ( $\rightarrow$  I/3.1.2) of  $\sigma^2$

$$E[\hat{\sigma}^2] \neq \sigma^2, \quad (4)$$

more precisely:

$$E[\hat{\sigma}^2] = \frac{n-p}{n} \sigma^2. \quad (5)$$

**Proof:**

1) This follows from maximum likelihood estimation for multiple linear regression ( $\rightarrow$  III/1.5.23) and is a special case ( $\rightarrow$  III/1.5.2) of maximum likelihood estimation for the general linear model ( $\rightarrow$  III/2.2.4) in which  $Y = y$ ,  $B = \beta$  and  $\Sigma = \sigma^2$ :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}(Y - X\hat{B})^T V^{-1}(Y - X\hat{B}) \\ &= \frac{1}{n}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}). \end{aligned} \quad (6)$$

2) We know that the residual sum of squares, divided by the true noise variance, is following a chi-squared distribution ( $\rightarrow$  III/1.5.20):

$$\begin{aligned} \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} &\sim \chi^2(n-p) \\ \text{where } \hat{\varepsilon}^T \hat{\varepsilon} &= (y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}). \end{aligned} \quad (7)$$

Thus, combining (7) and (6), we have:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p). \quad (8)$$

Using the relationship between chi-squared distribution and gamma distribution ( $\rightarrow$  II/3.7.2)

$$X \sim \chi^2(k) \quad \Rightarrow \quad cX \sim \text{Gam}\left(\frac{k}{2}, \frac{1}{2c}\right), \quad (9)$$

we can deduce from (8) that

$$\hat{\sigma}^2 = \frac{\sigma^2}{n} \cdot \frac{n\hat{\sigma}^2}{\sigma^2} \sim \text{Gam}\left(\frac{n-p}{2}, \frac{n}{2\sigma^2}\right). \quad (10)$$

Using the expected value of the gamma distribution ( $\rightarrow$  II/3.4.11)

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad E(X) = \frac{a}{b}, \quad (11)$$

we can deduce from (10) that

$$\mathbb{E} [\hat{\sigma}^2] = \frac{\frac{n-p}{2}}{\frac{n}{2\sigma^2}} = \frac{n-p}{n} \sigma^2 \quad (12)$$

which proves the relationship given by (5). ■

#### Sources:

- ocram (2022): “Why is RSS distributed chi square times n-p?”; in: *StackExchange Cross Validated*, retrieved on 2022-12-21; URL: <https://stats.stackexchange.com/a/20230>.

### 1.0.4 Construction of unbiased estimator ( $p = 1$ )

**Theorem:** Let  $y = \{y_1, \dots, y_n\}$  be a set of independent normally distributed ( $\rightarrow$  II/3.2.1) observations with unknown mean ( $\rightarrow$  I/1.12.1)  $\mu$  and variance ( $\rightarrow$  I/1.13.1)  $\sigma^2$ :

$$y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

An unbiased estimator ( $\rightarrow$  I/3.1.2) of  $\sigma^2$  is given by

$$\hat{\sigma}_{\text{unb}}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (2)$$

**Proof:** It can be shown that ( $\rightarrow$  IV/1.0.2) the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $\sigma^2$

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

is a biased estimator ( $\rightarrow$  I/3.1.2) in the sense that

$$\mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] = \frac{n-1}{n} \sigma^2. \quad (4)$$

From (4), it follows that

$$\begin{aligned} \mathbb{E} \left[ \frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \right] &= \frac{n}{n-1} \mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] \\ &\stackrel{(4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 \\ &= \sigma^2, \end{aligned} \quad (5)$$

such that an unbiased estimator ( $\rightarrow$  I/3.1.2) can be constructed as

$$\begin{aligned} \hat{\sigma}_{\text{unb}}^2 &= \frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \\ &\stackrel{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (6)$$

**Sources:**

- Liang, Dawen (????): “Maximum Likelihood Estimator for Variance is Biased: Proof”, retrieved on 2020-02-25; URL: [https://dawenl.github.io/files/mle\\_biased.pdf](https://dawenl.github.io/files/mle_biased.pdf).

**1.0.5 Construction of unbiased estimator ( $p > 1$ )**

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V). \quad (1)$$

An unbiased estimator ( $\rightarrow$  I/3.1.2) of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \quad (2)$$

where

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \quad (3)$$

**Proof:** It can be shown that ( $\rightarrow$  IV/1.0.3) the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) of  $\sigma^2$

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \quad (4)$$

is a biased estimator ( $\rightarrow$  I/3.1.2) in the sense that

$$\mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] = \frac{n-p}{n} \sigma^2. \quad (5)$$

From (5), it follows that

$$\begin{aligned} \mathbb{E} \left[ \frac{n}{n-p} \hat{\sigma}_{\text{MLE}}^2 \right] &= \frac{n}{n-p} \mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] \\ &\stackrel{(5)}{=} \frac{n}{n-p} \cdot \frac{n-p}{n} \sigma^2 \\ &= \sigma^2, \end{aligned} \quad (6)$$

such that an unbiased estimator ( $\rightarrow$  I/3.1.2) can be constructed as

$$\begin{aligned} \hat{\sigma}_{\text{unb}}^2 &= \frac{n}{n-p} \hat{\sigma}_{\text{MLE}}^2 \\ &\stackrel{(4)}{=} \frac{n}{n-p} \cdot \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= \frac{1}{n-p} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}). \end{aligned} \quad (7)$$



## 1.1 R-squared

### 1.1.1 Definition

**Definition:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

with measured data  $y$ , known design matrix  $X$  as well as unknown regression coefficients  $\beta$  and noise variance  $\sigma^2$ .

Then, the proportion of the variance of the dependent variable  $y$  (“total variance ( $\rightarrow$  III/1.5.7)”) that can be predicted from the independent variables  $X$  (“explained variance ( $\rightarrow$  III/1.5.8)”) is called “coefficient of determination”, “R-squared” or  $R^2$ .

**Sources:**

- Wikipedia (2020): “Coefficient of determination”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-25; URL: [https://en.wikipedia.org/wiki/Mean\\_squared\\_error#Proof\\_of\\_variance\\_and\\_bias\\_relationship](https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship).

### 1.1.2 Derivation of $R^2$ and adjusted $R^2$

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

with  $n$  independent observations and  $p$  independent variables,

1) the coefficient of determination ( $\rightarrow$  IV/1.1.1) is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (2)$$

2) the adjusted coefficient of determination is

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \quad (3)$$

where the residual ( $\rightarrow$  III/1.5.9) and total sum of squares ( $\rightarrow$  III/1.5.7) are

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \end{aligned} \quad (4)$$

where  $X$  is the  $n \times p$  design matrix and  $\hat{\beta}$  are the ordinary least squares ( $\rightarrow$  III/1.5.3) estimates.

**Proof:** The coefficient of determination ( $\rightarrow$  IV/1.1.1)  $R^2$  is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data.

1) If we define the explained sum of squares ( $\rightarrow$  III/1.5.8) as

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (5)$$

then  $R^2$  is given by

$$R^2 = \frac{\text{ESS}}{\text{TSS}}. \quad (6)$$

which is equal to

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \quad (7)$$

because it holds that ( $\rightarrow$  III/1.5.10)

$$\text{TSS} = \text{ESS} + \text{RSS}. \quad (8)$$

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}. \quad (9)$$

If we replace the variance estimates by their unbiased estimators ( $\rightarrow$  IV/1.0.5), we obtain

$$R_{\text{adj}}^2 = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{RSS}/\text{df}_r}{\text{TSS}/\text{df}_t} \quad (10)$$

where  $\text{df}_r = n - p$  and  $\text{df}_t = n - 1$  are the residual and total degrees of freedom ( $\rightarrow$  I/1.5.12). This gives the adjusted  $R^2$  which adjusts  $R^2$  for the number of explanatory variables ( $\rightarrow$  III/1.5.1). ■

#### Sources:

- Wikipedia (2019): “Coefficient of determination”; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-12-06; URL: [https://en.wikipedia.org/wiki/Coefficient\\_of\\_determination#Adjusted\\_R2](https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2).

### 1.1.3 Relationship to residual variance

**Theorem:** Given a linear regression model with independent observations ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (1)$$

the coefficient of determination ( $\rightarrow$  IV/1.1.1) can be expressed in terms of residual variances ( $\rightarrow$  IV/1.0.1) as

$$R^2 = 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot s^2} \quad (2)$$

where  $n$  is the number of observations,  $p$  is the number of predictors,  $\hat{\sigma}^2$  is an unbiased estimate of the noise variance ( $\rightarrow$  IV/1.0.5)  $\sigma^2$  and  $s^2$  is the unbiased sample variance ( $\rightarrow$  I/1.13.2) of  $y$ .

**Proof:** The coefficient of determination ( $\rightarrow$  IV/1.1.2) is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (3)$$

where RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9)

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad (4)$$

and TSS is the total sum of squares ( $\rightarrow$  III/1.5.7)

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \quad (5)$$

Note that the residual sum of squares can be written as:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (X\hat{\beta})_i)^2 = (y - X\hat{\beta})^T (y - X\hat{\beta}) . \quad (6)$$

The unbiased estimate of the noise variance ( $\rightarrow$  IV/1.0.5) is

$$\hat{\sigma}^2 = \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (7)$$

and the unbiased sample variance of the dependent variable ( $\rightarrow$  I/1.13.2) is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 , \quad (8)$$

Combining (3), (4) and (5), the coefficient of determination can be rewritten as follows:

$$\begin{aligned} R^2 &\stackrel{(3)}{=} 1 - \frac{\text{RSS}}{\text{TSS}} \\ &\stackrel{(5)}{=} 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &\stackrel{(6)}{=} 1 - \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{(n-p) \cdot \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta})}{(n-1) \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &\stackrel{(7)}{=} 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &\stackrel{(8)}{=} 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot s^2} . \end{aligned} \quad (9)$$

This completes the proof. ■

#### 1.1.4 Relationship to maximum log-likelihood

**Theorem:** Given a linear regression model with independent observations ( $\rightarrow$  III/1.5.1)

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) , \quad (1)$$



the coefficient of determination ( $\rightarrow$  IV/1.1.1) can be expressed in terms of the maximum log-likelihood ( $\rightarrow$  I/4.1.4) as

$$R^2 = 1 - (\exp[\Delta\text{MLL}])^{-2/n} \quad (2)$$

where  $n$  is the number of observations and  $\Delta\text{MLL}$  is the difference in maximum log-likelihood between the model given by (1) and a linear regression model with only a constant regressor.

**Proof:** First, we express the maximum log-likelihood ( $\rightarrow$  I/4.1.4) (MLL) of a linear regression model in terms of its residual sum of squares ( $\rightarrow$  III/1.5.9) (RSS). The model in (1) implies the following log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\text{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta), \quad (3)$$

such that maximum likelihood estimates are ( $\rightarrow$  III/1.5.23)

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (5)$$

and the residual sum of squares ( $\rightarrow$  III/1.5.9) is

$$\text{RSS} = \sum_{i=1}^n \hat{\varepsilon}_i = \hat{\varepsilon}^T \hat{\varepsilon} = (y - X\hat{\beta})^T (y - X\hat{\beta}) = n \cdot \hat{\sigma}^2. \quad (6)$$

Since  $\hat{\beta}$  and  $\hat{\sigma}^2$  are maximum likelihood estimates ( $\rightarrow$  I/4.1.3), plugging them into the log-likelihood function gives the maximum log-likelihood:

$$\text{MLL} = \text{LL}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})^T (y - X\hat{\beta}). \quad (7)$$

With (6) for the first  $\hat{\sigma}^2$  and (5) for the second  $\hat{\sigma}^2$ , the MLL becomes

$$\text{MLL} = -\frac{n}{2} \log(\text{RSS}) - \frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2}. \quad (8)$$

Second, we establish the relationship between maximum log-likelihood (MLL) and coefficient of determination ( $R^2$ ). Consider the two models

$$\begin{aligned} m_0 : X_0 &= 1_n \\ m_1 : X_1 &= X \end{aligned} \quad (9)$$

For  $m_1$ , the residual sum of squares is given by (6); and for  $m_0$ , the residual sum of squares is equal to the total sum of squares ( $\rightarrow$  III/1.5.7):

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (10)$$

Using (8), we can therefore write

$$\Delta\text{MLL} = \text{MLL}(m_1) - \text{MLL}(m_0) = -\frac{n}{2} \log(\text{RSS}) + \frac{n}{2} \log(\text{TSS}). \quad (11)$$

Exponentiating both sides of the equation, we have:

$$\begin{aligned}
 \exp[\Delta\text{MLL}] &= \exp\left[-\frac{n}{2}\log(\text{RSS}) + \frac{n}{2}\log(\text{TSS})\right] \\
 &= (\exp[\log(\text{RSS}) - \log(\text{TSS})])^{-n/2} \\
 &= \left(\frac{\exp[\log(\text{RSS})]}{\exp[\log(\text{TSS})]}\right)^{-n/2} \\
 &= \left(\frac{\text{RSS}}{\text{TSS}}\right)^{-n/2}.
 \end{aligned} \tag{12}$$

Taking both sides to the power of  $-2/n$  and subtracting from 1, we have

$$\begin{aligned}
 (\exp[\Delta\text{MLL}])^{-2/n} &= \frac{\text{RSS}}{\text{TSS}} \\
 1 - (\exp[\Delta\text{MLL}])^{-2/n} &= 1 - \frac{\text{RSS}}{\text{TSS}} = R^2
 \end{aligned} \tag{13}$$

which proves the identity given above. ■

### 1.1.5 Statistical significance test for $R^2$

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V). \tag{1}$$

Further assume that  $X$  contains a constant regressor ( $\rightarrow$  III/1.5.1). Then, the coefficient of determination ( $\rightarrow$  IV/1.1.1) can be used to calculate a test statistic ( $\rightarrow$  I/4.3.5)

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \tag{2}$$

where  $n$  and  $p$  are the dimensions of the design matrix ( $\rightarrow$  III/1.5.1)  $X$ , and this test statistic follows an F-distribution ( $\rightarrow$  II/3.8.1)

$$F \sim F(p-1, n-p) \tag{3}$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the true coefficient of determination ( $\rightarrow$  IV/1.1.1) is zero:

$$H_0 : R^2 = 0. \tag{4}$$

**Proof:** Consider two linear regression models ( $\rightarrow$  III/1.5.1) for the same measured data  $y$ , with design matrices  $X = X_0 \in \mathbb{R}^{n \times p_0}$  and  $X = [X_0, X_1] \in \mathbb{R}^{n \times p}$  as well as regression coefficients  $\beta = \beta_0 \in \mathbb{R}^{p_0 \times 1}$  and  $\beta = [\beta_0^T, \beta_1^T]^T \in \mathbb{R}^{p \times 1}$ .

Then, under the null hypothesis that all regression coefficients ( $\rightarrow$  III/1.5.1)  $\beta_1$  associated with  $X_1$  are zero

$$H_0 : \beta_1 = 0_{p-p_0} \quad \Leftrightarrow \quad \beta_i = 0 \quad \text{for all } j = p_0 + 1, \dots, p, \tag{5}$$

the omnibus F-statistic follows an F-distribution ( $\rightarrow$  III/1.5.31)

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)} \sim F(p - p_0, n - p) \quad (6)$$

where  $\text{RSS}_0$  and  $\text{RSS}$  are the residual sums of squares ( $\rightarrow$  III/1.5.9) of the null model with  $X_0$  and the full model with  $X_0$  nested in  $X$ , after regression coefficients have been estimated with weighted least squares ( $\rightarrow$  III/1.5.21) or maximum likelihood ( $\rightarrow$  III/1.5.23).

Since by the requirements of our theorem,  $X$  contains a constant regressor, we can assume the following design matrices without loss of generality:

$$X_0 = 1_n \in \mathbb{R}^{n \times 1} \quad \text{and} \quad X = [1_n, X_1] \in \mathbb{R}^{n \times p}. \quad (7)$$

Thus, since a single constant regressor estimates the mean and considering the definition of the total sum of squares ( $\rightarrow$  III/1.5.7) TSS, we in our case have:

$$\text{RSS}_0 = \text{TSS} \quad \text{and} \quad p_0 = 1. \quad (8)$$

The coefficient of determination is given by ( $\rightarrow$  IV/1.1.2)

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (9)$$

which can also be written as ( $\rightarrow$  III/1.5.10)

$$R^2 = \frac{\text{ESS}}{\text{TSS}}. \quad (10)$$

If all regression coefficients  $\beta_1$  associated with  $X_1$  are zero, then the true  $R^2$  is zero, because there is no variance explained beyond the constant regressor, the explained sum of squares ( $\rightarrow$  III/1.5.8) ESS is zero and the residual sum of squares ( $\rightarrow$  III/1.5.9) RSS is equal to the total sum of squares ( $\rightarrow$  III/1.5.7) TSS.

Then, by virtue of (6), we get the following F-statistic:

$$\begin{aligned} F &\stackrel{(6)}{=} \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)} \\ &\stackrel{(8)}{=} \frac{(\text{TSS} - \text{RSS})/(p - 1)}{\text{RSS}/(n - p)} \\ &= \frac{\frac{\text{TSS} - \text{RSS}}{\text{TSS}}/(p - 1)}{\frac{\text{RSS}}{\text{TSS}}/(n - p)} \\ &= \frac{\left(1 - \frac{\text{RSS}}{\text{TSS}}\right)/(p - 1)}{\left(1 - \left(1 - \frac{\text{RSS}}{\text{TSS}}\right)\right)/(n - p)} \\ &\stackrel{(9)}{=} \frac{(R^2)/(p - 1)}{(1 - R^2)/(n - p)}. \end{aligned} \quad (11)$$

This means that the null hypothesis ( $\rightarrow$  I/4.3.2) can be rejected when  $F$  as a function of  $R^2$  is as extreme or more extreme than the critical value ( $\rightarrow$  I/4.3.9) obtained from the F-distribution ( $\rightarrow$  II/3.8.1) with  $p - 1$  denominator and  $n - p$  numerator degrees of freedom ( $\rightarrow$  I/1.5.12) using a significance level ( $\rightarrow$  I/4.3.8)  $\alpha$ .



### Sources:

- Alecos Papadopoulos (2014): “What is the distribution of  $R^2$  in linear regression under the null hypothesis?”; in: *StackExchange CrossValidated*, retrieved on 2024-03-15; URL: <https://stats.stackexchange.com/a/130082>.

### 1.1.6 Distribution under null hypothesis

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Further assume that  $X$  contains a constant regressor ( $\rightarrow$  III/1.5.1), i.e. it is of the following form:

$$X = [1_n, X_1] \in \mathbb{R}^{n \times p} . \quad (2)$$

Then, the coefficient of determination ( $\rightarrow$  IV/1.1.1) follows a beta distribution ( $\rightarrow$  II/3.9.1)

$$R^2 \sim \text{Bet} \left( \frac{p-1}{2}, \frac{n-p}{2} \right) \quad (3)$$

under the null hypothesis ( $\rightarrow$  I/4.3.2) that the true coefficient of determination ( $\rightarrow$  IV/1.1.1) is zero:

$$H_0 : R^2 = 0 . \quad (4)$$

**Proof:** We know that the F-statistic is related to R-squared as ( $\rightarrow$  IV/1.2.2)

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \quad (5)$$

and that the F-statistic follows an F-distribution ( $\rightarrow$  IV/1.1.5) under  $H_0$ :

$$F \sim F(p-1, n-p), \quad \text{if } R^2 = 0 . \quad (6)$$

Rearranging equation (5) gives R-squared as a function of the F-statistic ( $\rightarrow$  IV/1.2.2)

$$R^2 = \frac{F \cdot \frac{p-1}{n-p}}{1 + F \cdot \frac{p-1}{n-p}} \quad (7)$$

which, when expanding with  $(n-p)$ , can also be written as:

$$R^2 = \frac{(p-1)F}{(n-p) + (p-1)F} . \quad (8)$$

Using the relationship between F-distribution and beta distribution ( $\rightarrow$  II/3.9.4)

$$X \sim F(d_1, d_2) \quad \Rightarrow \quad Y = \frac{d_1 X}{d_2 + d_1 X} \sim \text{Bet} \left( \frac{d_1}{2}, \frac{d_2}{2} \right) , \quad (9)$$

i.e. combining (6) and (8), it follows that

$$R^2 \sim \text{Bet} \left( \frac{p-1}{2}, \frac{n-p}{2} \right) \quad (10)$$

under the null hypothesis that  $R^2 = 0$ .

■

#### Sources:

- Alecos Papadopoulos (2014): “What is the distribution of  $R^2$  in linear regression under the null hypothesis?”; in: *StackExchange CrossValidated*, retrieved on 2024-03-15; URL: <https://stats.stackexchange.com/a/130082>.

#### 1.1.7 Mean/mode/median under null hypothesis

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X = [1_n, X_1] \in \mathbb{R}^{n \times p}$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, under the null hypothesis ( $\rightarrow$  I/4.3.2) that the true coefficient of determination ( $\rightarrow$  IV/1.1.1) is zero, i.e.  $H_0 : R^2 = 0$ , the expected value ( $\rightarrow$  I/1.12.1), the median ( $\rightarrow$  I/1.18.1) and the mode ( $\rightarrow$  I/1.18.3) of  $R^2$  are

$$\begin{aligned} E(R^2) &= \frac{p-1}{n-1} \\ \text{median}(R^2) &= I^{-1}\left(\frac{1}{2}; \frac{p-1}{2}, \frac{n-p}{2}\right) \\ \text{mode}(R^2) &= \frac{p-3}{n-5} \end{aligned} \quad (2)$$

where  $I^{-1}(p; a, b)$  is the inverse function of the regularized incomplete beta function ( $\rightarrow$  II/3.9.7).

**Proof:** We know that R-squared follows a beta distribution ( $\rightarrow$  IV/1.1.6) under  $H_0$ :

$$R^2 \sim \text{Bet}\left(\frac{p-1}{2}, \frac{n-p}{2}\right) . \quad (3)$$

Using mean ( $\rightarrow$  II/3.9.8), median ( $\rightarrow$  II/3.9.9) and mode ( $\rightarrow$  II/3.9.10) of the beta distribution ( $\rightarrow$  II/3.9.1)

$$\begin{aligned} X &\sim \text{Bet}(\alpha, \beta) \\ \Rightarrow E(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{median}(X) &= I^{-1}(1/2; \alpha, \beta) \\ \text{mode}(X) &= \frac{\alpha - 1}{\alpha + \beta - 2} , \end{aligned} \quad (4)$$

we have:

$$\begin{aligned}
E(R^2) &= \frac{(p-1)/2}{(p-1)/2 + (n-p)/2} \\
&= \frac{p-1}{n-1} \\
\text{median}(R^2) &= I^{-1}(1/2; (p-1)/2, (n-p)/2) \\
&= I^{-1}\left(\frac{1}{2}, \frac{p-1}{2}, \frac{n-p}{2}\right) \\
\text{mode}(R^2) &= \frac{(p-1)/2 - 1}{(p-1)/2 + (n-p)/2 - 2} \\
&= \frac{p-3}{n-5}.
\end{aligned} \tag{5}$$

This completes the proof. ■

### 1.1.8 Variance under null hypothesis

**Theorem:** Consider a linear regression model ( $\rightarrow$  III/1.5.1) with known design matrix  $X = [1_n, X_1] \in \mathbb{R}^{n \times p}$ , known covariance structure  $V$ , unknown regression parameters  $\beta$  and unknown noise variance  $\sigma^2$ :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V). \tag{1}$$

Then, under the null hypothesis ( $\rightarrow$  I/4.3.2) that the true coefficient of determination ( $\rightarrow$  IV/1.1.1) is zero, i.e.  $H_0 : R^2 = 0$ , the variance ( $\rightarrow$  I/1.13.1) of the coefficient of determination is

$$\text{Var}(R^2) = 2 \cdot \frac{(p-1) \cdot (n-p)}{(n+1) \cdot (n-1)^2}. \tag{2}$$

**Proof:** We know that R-squared follows a beta distribution ( $\rightarrow$  IV/1.1.6) under  $H_0$ :

$$R^2 \sim \text{Bet}\left(\frac{p-1}{2}, \frac{n-p}{2}\right). \tag{3}$$

Using the variance of the beta distribution ( $\rightarrow$  II/3.9.11)

$$X \sim \text{Bet}(\alpha, \beta) \quad \Rightarrow \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2}, \tag{4}$$

we have:

$$\begin{aligned}
\text{Var}(R^2) &= \frac{\frac{p-1}{2} \cdot \frac{n-p}{2}}{\left(\frac{p-1}{2} + \frac{n-p}{2} + 1\right) \cdot \left(\frac{p-1}{2} + \frac{n-p}{2}\right)^2} \\
&= \frac{\frac{(p-1) \cdot (n-p)}{4}}{\frac{1}{2} \cdot (n+1) \cdot \frac{(n-1)^2}{4}} \\
&= 2 \cdot \frac{(p-1) \cdot (n-p)}{(n+1) \cdot (n-1)^2}.
\end{aligned} \tag{5}$$

This completes the proof.



## 1.2 F-statistic

### 1.2.1 Definition

**Definition:** Consider two linear regression models ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$\begin{aligned} m_1 : y &= X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\ m_0 : y &= X_0\beta_0 + \varepsilon_0, \varepsilon_{0i} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2) \end{aligned} \quad (1)$$

operating on identical measured data  $y$ , but with different design matrices  $X \in \mathbb{R}^{n \times p}$  and  $X_0 \in \mathbb{R}^{n \times p_0}$  and thus different regression coefficients  $\beta \in \mathbb{R}^{p \times 1}$  and  $\beta_0 \in \mathbb{R}^{p_0 \times 1}$ . Furthermore, let the design matrix of the null model be fully contained in the design matrix of the full model:

$$X = \begin{bmatrix} X_0 & X_1 \end{bmatrix}. \quad (2)$$

Then, the F-statistic for model comparison is defined as the ratio of the difference in residual sum of squares ( $\rightarrow$  III/1.5.9) between the two models, divided by the difference in number of parameters ( $\rightarrow$  III/1.5.1), to the residual sum of squares ( $\rightarrow$  III/1.5.9) of the full model, divided by the number of degrees of freedom ( $\rightarrow$  I/1.5.12):

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)}. \quad (3)$$

#### Sources:

- Wikipedia (2024): “F-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-03-15; URL: [https://en.wikipedia.org/wiki/F-test#Regression\\_problems](https://en.wikipedia.org/wiki/F-test#Regression_problems).

### 1.2.2 Relationship to coefficient of determination

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \quad (1)$$

Then, the F-statistic ( $\rightarrow$  IV/1.2.1) for comparing this model against a null model containing only a constant regressor ( $\rightarrow$  III/1.5.1)  $x_0 = 1_n$  can be expressed in terms of the coefficient of determination ( $\rightarrow$  IV/1.1.1)

$$F = \frac{R^2/(p - 1)}{(1 - R^2)/(n - p)} \quad (2)$$

and vice versa

$$R^2 = 1 - \frac{1}{F \cdot \frac{p-1}{n-p} + 1} \quad (3)$$

where  $n$  and  $p$  are the dimensions of the design matrix  $X \in \mathbb{R}^{n \times p}$ .

**Proof:** Consider two linear regression models ( $\rightarrow$  III/1.5.1) for the same measured data  $y$ , one using design matrix  $X$  from (1) and the other with design matrix  $X_0 = \mathbf{1}_n \in \mathbb{R}^{n \times 1}$ . Then, RSS is the residual sum of squares ( $\rightarrow$  III/1.5.9) of the model in (1) and the residual sum of squares for the model using  $X_0$  is equal to the total sum of squares ( $\rightarrow$  III/1.5.7).

1) Thus, the F-statistic ( $\rightarrow$  IV/1.2.1)

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)} \quad (4)$$

becomes

$$F = \frac{(\text{TSS} - \text{RSS})/(p - 1)}{\text{RSS}/(n - p)} . \quad (5)$$

From this, we can derive  $F$  in terms of  $R^2$ :

$$\begin{aligned} F &= \frac{(\text{TSS} - \text{RSS})/(p - 1)}{\text{RSS}/(n - p)} \\ &= \frac{\frac{\text{TSS} - \text{RSS}}{\text{TSS}}/(p - 1)}{\frac{\text{RSS}}{\text{TSS}}/(n - p)} \\ &= \frac{\left(1 - \frac{\text{RSS}}{\text{TSS}}\right)/(p - 1)}{\left(1 - \left(1 - \frac{\text{RSS}}{\text{TSS}}\right)\right)/(n - p)} \\ &= \frac{(R^2)/(p - 1)}{(1 - R^2)/(n - p)} . \end{aligned} \quad (6)$$

2) Rearranging this equation, we can derive  $R^2$  in terms of  $F$ :



$$\begin{aligned}
F &= \frac{(R^2)/(p-1)}{(1-R^2)/(n-p)} \\
F \cdot \frac{p-1}{n-p} &= \frac{R^2}{(1-R^2)} \\
F \cdot \frac{p-1}{n-p} \cdot (1-R^2) &= R^2 \\
F \cdot \frac{p-1}{n-p} - F \cdot \frac{p-1}{n-p} \cdot R^2 &= R^2 \\
F \cdot \frac{p-1}{n-p} \cdot R^2 + R^2 &= F \cdot \frac{p-1}{n-p} \\
R^2 \left( F \cdot \frac{p-1}{n-p} + 1 \right) &= F \cdot \frac{p-1}{n-p} \\
R^2 &= \frac{F \cdot \frac{p-1}{n-p}}{F \cdot \frac{p-1}{n-p} + 1} \\
R^2 &= \frac{F \cdot \frac{p-1}{n-p} + 1 - 1}{F \cdot \frac{p-1}{n-p} + 1} \\
R^2 &= \frac{F \cdot \frac{p-1}{n-p} + 1}{F \cdot \frac{p-1}{n-p} + 1} - \frac{1}{F \cdot \frac{p-1}{n-p} + 1} \\
R^2 &= 1 - \frac{1}{F \cdot \frac{p-1}{n-p} + 1}
\end{aligned} \tag{7}$$

This completes the proof. ■

#### Sources:

- Alecos Papadopoulos (2014): “What is the distribution of  $R^2$  in linear regression under the null hypothesis?”; in: *StackExchange CrossValidated*, retrieved on 2024-03-15; URL: <https://stats.stackexchange.com/a/130082>.

### 1.2.3 Relationship to maximum log-likelihood

**Theorem:** Consider two linear regression models ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$\begin{aligned}
m_1 : y &= X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\
m_0 : y &= X_0\beta_0 + \varepsilon_0, \quad \varepsilon_{0i} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2) .
\end{aligned} \tag{1}$$

Then, the F-statistic ( $\rightarrow$  IV/1.2.1) can be expressed in terms of the maximum log-likelihood ( $\rightarrow$  I/4.1.4) as

$$F = \left[ (\exp[\Delta\text{MLL}])^{2/n} - 1 \right] \cdot \frac{n-p}{p-p_0} \tag{2}$$

where  $n$ ,  $p$  and  $p_0$  are the dimensions of the design matrices  $X = [X_0, X_1] \in \mathbb{R}^{n \times p}$  and  $X_0 \in \mathbb{R}^{n \times p_0}$  and  $\Delta\text{MLL}$  is the difference in maximum log-likelihood between the two models given by (1)

**Proof:** Under the conditions mentioned in the theorem, the F-statistic is defined in terms of the residual sum of squares ( $\rightarrow$  IV/1.2.1) as

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)} . \quad (3)$$

We also know that the maximum log-likelihood can be expressed in terms of residual sum of squares ( $\rightarrow$  III/1.5.24):

$$\text{MLL}(m) = -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] . \quad (4)$$

Based on this, we see that the difference of the maximum log-likelihoods develops into

$$\begin{aligned} \Delta\text{MLL} &= \text{MLL}(m_1) - \text{MLL}(m_0) \\ &= \left( -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] \right) \\ &\quad - \left( -\frac{n}{2} \log \left( \frac{\text{RSS}_0}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] \right) \\ &= -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right) + \frac{n}{2} \log \left( \frac{\text{RSS}_0}{n} \right) . \end{aligned} \quad (5)$$

Finally, we simply perform algebraic operations on both sides ( $\rightarrow$  IV/1.1.4) until we reach the F-statistic on the right side. We start by exponentiating the MLL difference:

$$\begin{aligned}
\exp[\Delta\text{MLL}] &= \exp \left[ -\frac{n}{2} \log(\text{RSS}/n) + \frac{n}{2} \log(\text{RSS}_0/n) \right] \\
\exp[\Delta\text{MLL}] &= (\exp [\log(\text{RSS}/n) - \log(\text{RSS}_0/n)])^{-n/2} \\
\exp[\Delta\text{MLL}] &= \left( \frac{\exp[\log(\text{RSS}/n)]}{\exp[\log(\text{RSS}_0/n)]} \right)^{-n/2} \\
\exp[\Delta\text{MLL}] &= \left( \frac{\text{RSS}/n}{\text{RSS}_0/n} \right)^{-n/2} \\
\exp[\Delta\text{MLL}] &= \left( \frac{\text{RSS}_0}{\text{RSS}} \right)^{n/2} \\
(\exp[\Delta\text{MLL}])^{2/n} &= \frac{\text{RSS}_0}{\text{RSS}} \\
(\exp[\Delta\text{MLL}])^{2/n} - 1 &= \frac{\text{RSS}_0}{\text{RSS}} - 1 \\
(\exp[\Delta\text{MLL}])^{2/n} - 1 &= \frac{\text{RSS}_0}{\text{RSS}} - \frac{\text{RSS}}{\text{RSS}} \\
\left[ (\exp[\Delta\text{MLL}])^{2/n} - 1 \right] \cdot \frac{n-p}{p-p_0} &= \left[ \frac{\text{RSS}_0 - \text{RSS}}{\text{RSS}} \right] \cdot \frac{n-p}{p-p_0} \\
\left[ (\exp[\Delta\text{MLL}])^{2/n} - 1 \right] \cdot \frac{n-p}{p-p_0} &= \frac{(\text{RSS}_0 - \text{RSS})/(p-p_0)}{\text{RSS}/(n-p)} \\
\left[ (\exp[\Delta\text{MLL}])^{2/n} - 1 \right] \cdot \frac{n-p}{p-p_0} &= F .
\end{aligned} \tag{6}$$

This completes the proof. ■

### 1.3 Signal-to-noise ratio

#### 1.3.1 Definition

**Definition:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with measured data  $y$ , known design matrix  $X$  as well as unknown regression coefficients  $\beta$  and noise variance  $\sigma^2$ .

Given estimated regression coefficients ( $\rightarrow$  III/1.5.23)  $\hat{\beta}$  and residual variance ( $\rightarrow$  IV/1.0.1)  $\hat{\sigma}^2$ , the signal-to-noise ratio (SNR) is defined as the ratio of estimated signal variance to estimated noise variance:

$$\text{SNR} = \frac{\text{Var}(X\hat{\beta})}{\hat{\sigma}^2} . \tag{2}$$

**Sources:**

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 6; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 1.3.2 Relationship to coefficient of determination

**Theorem:** Let there be a linear regression model ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

and parameter estimates ( $\rightarrow$  I/3.1.1) obtained with ordinary least squares ( $\rightarrow$  III/1.5.3)

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad (2)$$

Then, the signal-to noise ratio ( $\rightarrow$  IV/1.3.1) can be expressed in terms of the coefficient of determination ( $\rightarrow$  IV/1.1.1)

$$\text{SNR} = \frac{R^2}{1 - R^2} \quad (3)$$

and vice versa

$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}}, \quad (4)$$

if the predicted signal mean is equal to the actual signal mean.

**Proof:** The signal-to-noise ratio ( $\rightarrow$  IV/1.3.1) (SNR) is defined as

$$\text{SNR} = \frac{\text{Var}(X\hat{\beta})}{\hat{\sigma}^2} = \frac{\text{Var}(\hat{y})}{\hat{\sigma}^2}. \quad (5)$$

Writing out the sample variances ( $\rightarrow$  I/1.13.2), we have

$$\text{SNR} = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (6)$$

Note that it is irrelevant whether we use the biased estimator of the variance ( $\rightarrow$  IV/1.0.2) (dividing by  $n$ ) or the unbiased estimator for the variance ( $\rightarrow$  IV/1.0.4) (dividing by  $n - 1$ ), because the relevant terms cancel out.

If the predicted signal mean is equal to the actual signal mean – which is the case when variable regressors in  $X$  have mean zero, such that they are orthogonal to a constant regressor in  $X$  –, this means that  $\bar{\hat{y}} = \bar{y}$ , such that

$$\text{SNR} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (7)$$

Then, the SNR can be written in terms of the explained ( $\rightarrow$  III/1.5.8), residual ( $\rightarrow$  III/1.5.9) and total sum of squares ( $\rightarrow$  III/1.5.7):

$$\text{SNR} = \frac{\text{ESS}}{\text{RSS}} = \frac{\text{ESS/TSS}}{\text{RSS/TSS}}. \quad (8)$$

With the derivation of the coefficient of determination ( $\rightarrow$  IV/1.1.2), this becomes

$$\text{SNR} = \frac{R^2}{1 - R^2} . \quad (9)$$

Rearranging this equation for the coefficient of determination ( $\rightarrow$  IV/1.1.1), we have

$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}} . \quad (10)$$

■

### 1.3.3 Relationship to maximum log-likelihood

**Theorem:** Given a linear regression model ( $\rightarrow$  III/1.5.1) with independent ( $\rightarrow$  I/1.3.6) observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) ; \quad (1)$$

the signal-to-noise ratio ( $\rightarrow$  IV/1.3.1) can be expressed in terms of the maximum log-likelihood ( $\rightarrow$  I/4.1.4) as

$$\text{SNR} = (\exp[\Delta\text{MLL}])^{2/n} - 1 , \quad (2)$$

where  $n$  is the number of observations and  $\Delta\text{MLL}$  is the difference in maximum log-likelihood between the model given by (1) and a linear regression model with only a constant regressor.

This holds, if the predicted signal mean is equal to the actual signal mean

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n (X\hat{\beta})_i = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} \quad (3)$$

where  $X$  is the  $n \times p$  design matrix and  $\hat{\beta}$  are the ordinary least squares ( $\rightarrow$  III/1.5.3) estimates.

**Proof:** Under the conditions mentioned in the theorem, the signal-to-noise ratio can be expressed in terms of the coefficient of determination ( $\rightarrow$  IV/1.3.2) as

$$\text{SNR} = \frac{R^2}{1 - R^2} \quad (4)$$

and R-squared can be expressed in terms of maximum likelihood ( $\rightarrow$  IV/1.1.4) as

$$R^2 = 1 - (\exp[\Delta\text{MLL}])^{-2/n} . \quad (5)$$

Plugging (5) into (4), we obtain:

$$\begin{aligned} \text{SNR} &= \frac{1 - (\exp[\Delta\text{MLL}])^{-2/n}}{(\exp[\Delta\text{MLL}])^{-2/n}} \\ &= \frac{1}{(\exp[\Delta\text{MLL}])^{-2/n}} - \frac{(\exp[\Delta\text{MLL}])^{-2/n}}{(\exp[\Delta\text{MLL}])^{-2/n}} \\ &= (\exp[\Delta\text{MLL}])^{2/n} - 1 . \end{aligned} \quad (6)$$

■

## 2 Classical information criteria

### 2.1 Akaike information criterion

#### 2.1.1 Definition

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and maximum likelihood estimates ( $\rightarrow$  I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the Akaike information criterion (AIC) of this model is defined as

$$\text{AIC}(m) = -2 \log p(y|\hat{\theta}, m) + 2k \quad (2)$$

where  $k$  is the number of free parameters estimated via (1).

**Sources:**

- Akaike H (1974): “A New Look at the Statistical Model Identification”; in: *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716-723; URL: <https://ieeexplore.ieee.org/document/1100705>; DOI: 10.1109/TAC.1974.1100705.

#### 2.1.2 Corrected AIC

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and maximum likelihood estimates ( $\rightarrow$  I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) ( $\text{AIC}_c$ ) of this model is defined as

$$\text{AIC}_c(m) = \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} \quad (2)$$

where  $\text{AIC}(m)$  is the Akaike information criterion ( $\rightarrow$  IV/2.1.1) and  $k$  is the number of free parameters estimated via (1).

**Sources:**

- Hurvich CM, Tsai CL (1989): “Regression and time series model selection in small samples”; in: *Biometrika*, vol. 76, no. 2, pp. 297-307; URL: <https://academic.oup.com/biomet/article-abstract/76/2/297/265326>; DOI: 10.1093/biomet/76.2.297.

#### 2.1.3 Corrected AIC and uncorrected AIC

**Theorem:** In the infinite data limit, the corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) converges to the uncorrected Akaike information criterion ( $\rightarrow$  IV/2.1.1)

$$\lim_{n \rightarrow \infty} \text{AIC}_c(m) = \text{AIC}(m) . \quad (1)$$

**Proof:** The corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) is defined as

$$\text{AIC}_c(m) = \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} . \quad (2)$$

Note that the number of free model parameters  $k$  is finite. Thus, we have:

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{AIC}_c(m) &= \lim_{n \rightarrow \infty} \left[ \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} \right] \\ &= \lim_{n \rightarrow \infty} \text{AIC}(m) + \lim_{n \rightarrow \infty} \frac{2k^2 + 2k}{n - k - 1} \\ &= \text{AIC}(m) + 0 \\ &= \text{AIC}(m) . \end{aligned} \quad (3)$$

■

#### Sources:

- Wikipedia (2022): “Akaike information criterion”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-18; URL: [https://en.wikipedia.org/wiki/Akaike\\_information\\_criterion#Modification\\_for\\_small\\_sample\\_size](https://en.wikipedia.org/wiki/Akaike_information_criterion#Modification_for_small_sample_size).

#### 2.1.4 Corrected AIC and maximum log-likelihood

**Theorem:** The corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) of a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  is equal to

$$\text{AIC}_c(m) = -2 \log p(y|\hat{\theta}, m) + \frac{2nk}{n - k - 1} \quad (1)$$

where  $\log p(y|\hat{\theta}, m)$  is the maximum log-likelihood ( $\rightarrow$  I/4.1.4),  $k$  is the number of free parameters and  $n$  is the number of observations.

**Proof:** The Akaike information criterion ( $\rightarrow$  IV/2.1.1) (AIC) is defined as

$$\text{AIC}(m) = -2 \log p(y|\hat{\theta}, m) + 2k \quad (2)$$

and the corrected Akaike information criterion ( $\rightarrow$  IV/2.1.2) is defined as

$$\text{AIC}_c(m) = \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} . \quad (3)$$

Plugging (2) into (3), we obtain:

$$\begin{aligned} \text{AIC}_c(m) &= -2 \log p(y|\hat{\theta}, m) + 2k + \frac{2k^2 + 2k}{n - k - 1} \\ &= -2 \log p(y|\hat{\theta}, m) + \frac{2k(n - k - 1)}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\ &= -2 \log p(y|\hat{\theta}, m) + \frac{2nk - 2k^2 - 2k}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\ &= -2 \log p(y|\hat{\theta}, m) + \frac{2nk}{n - k - 1} . \end{aligned} \quad (4)$$

**Sources:**

- Wikipedia (2022): “Akaike information criterion”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-11; URL: [https://en.wikipedia.org/wiki/Akaike\\_information\\_criterion#Modification\\_for\\_small\\_sample\\_size](https://en.wikipedia.org/wiki/Akaike_information_criterion#Modification_for_small_sample_size).

**2.2 Bayesian information criterion****2.2.1 Definition**

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and maximum likelihood estimates ( $\rightarrow$  I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the Bayesian information criterion (BIC) of this model is defined as

$$\text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + k \log n \quad (2)$$

where  $n$  is the number of data points and  $k$  is the number of free parameters estimated via (1).

**Sources:**

- Schwarz G (1978): “Estimating the Dimension of a Model”; in: *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464; URL: <https://www.jstor.org/stable/2958889>.

**2.2.2 Derivation**

**Theorem:** Let  $p(y|\theta, m)$  be the likelihood function ( $\rightarrow$  I/5.1.2) of a generative model ( $\rightarrow$  I/5.1.1)  $m \in \mathcal{M}$  with model parameters ( $\rightarrow$  I/1.5.7)  $\theta \in \Theta$  describing measured data ( $\rightarrow$  I/1.1.5)  $y \in \mathbb{R}^n$ . Let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) on the model parameters. Assume that likelihood function and prior density are twice differentiable.

Then, as the number of data points ( $\rightarrow$  I/1.1.7) goes to infinity, an approximation to the log-marginal likelihood ( $\rightarrow$  I/5.1.15)  $\log p(y|m)$ , up to constant terms not depending on the model, is given by the Bayesian information criterion ( $\rightarrow$  IV/2.2.1) (BIC) as

$$-2 \log p(y|m) \approx \text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \quad (1)$$

where  $\hat{\theta}$  is the maximum likelihood estimator ( $\rightarrow$  I/4.1.3) (MLE) of  $\theta$ ,  $n$  is the number of data points and  $p$  is the number of model parameters.

**Proof:** Let  $\text{LL}(\theta)$  be the log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\text{LL}(\theta) = \log p(y|\theta, m) \quad (2)$$

and define the functions  $g$  and  $h$  as follows:

$$\begin{aligned} g(\theta) &= p(\theta|m) \\ h(\theta) &= \frac{1}{n} \text{LL}(\theta) . \end{aligned} \quad (3)$$



Then, the marginal likelihood ( $\rightarrow$  I/5.1.15) can be written as follows:

$$\begin{aligned} p(y|m) &= \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta \\ &= \int_{\Theta} \exp[n h(\theta)] g(\theta) d\theta . \end{aligned} \quad (4)$$

This is an integral suitable for Laplace approximation which states that

$$\int_{\Theta} \exp[n h(\theta)] g(\theta) d\theta = \left( \sqrt{\frac{2\pi}{n}} \right)^p \exp[n h(\theta_0)] \left( g(\theta_0) |J(\theta_0)|^{-1/2} + O(1/n) \right) \quad (5)$$

where  $\theta_0$  is the value that maximizes  $h(\theta)$  and  $J(\theta_0)$  is the Hessian matrix evaluated at  $\theta_0$ . In our case, we have  $h(\theta) = 1/n \text{LL}(\theta)$  such that  $\theta_0$  is the maximum likelihood estimator  $\hat{\theta}$ :

$$\hat{\theta} = \arg \max_{\theta} \text{LL}(\theta) . \quad (6)$$

With this, (5) can be applied to (4) using (3) to give:

$$p(y|m) \approx \left( \sqrt{\frac{2\pi}{n}} \right)^p p(y|\hat{\theta}, m) p(\hat{\theta}|m) |J(\hat{\theta})|^{-1/2} . \quad (7)$$

Logarithmizing and multiplying with  $-2$ , we have:

$$-2 \log p(y|m) \approx -2 \text{LL}(\hat{\theta}) + p \log n - p \log(2\pi) - 2 \log p(\hat{\theta}|m) + \log |J(\hat{\theta})| . \quad (8)$$

As  $n \rightarrow \infty$ , the last three terms are  $O_p(1)$  and can therefore be ignored when comparing between models  $\mathcal{M} = \{m_1, \dots, m_M\}$  and using  $p(y|m_j)$  to compute posterior model probabilities ( $\rightarrow$  IV/3.4.1)  $p(m_j|y)$ . With that, the BIC is given as

$$\text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n . \quad (9)$$

■

#### Sources:

- Claeskens G, Hjort NL (2008): “The Bayesian information criterion”; in: *Model Selection and Model Averaging*, ch. 3.2, pp. 78-81; URL: <https://www.cambridge.org/core/books/model-selection-and-model-averaging>; E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

## 2.3 Deviance information criterion

### 2.3.1 Definition

**Definition:** Let  $m$  be a full probability model ( $\rightarrow$  I/5.1.6) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Together, likelihood function and prior distribution imply a posterior distribution ( $\rightarrow$  I/5.1.10)  $p(\theta|y, m)$ . Consider the deviance ( $\rightarrow$  IV/2.3.2) which is minus two times the log-likelihood function ( $\rightarrow$  I/4.1.2):

$$D(\theta) = -2 \log p(y|\theta, m) . \quad (1)$$

Then, the deviance information criterion (DIC) of the model  $m$  is defined as

$$\text{DIC}(m) = -2 \log p(y | \langle \theta \rangle, m) + 2 p_D \quad (2)$$

where  $\log p(y | \langle \theta \rangle, m)$  is the log-likelihood function ( $\rightarrow$  I/4.1.2) at the posterior ( $\rightarrow$  I/5.1.9) expectation ( $\rightarrow$  I/1.12.1) and the “effective number of parameters”  $p_D$  is the difference between the expectation of the deviance and the deviance at the expectation ( $\rightarrow$  IV/2.3.1):

$$p_D = \langle D(\theta) \rangle - D(\langle \theta \rangle) . \quad (3)$$

In these equations,  $\langle \cdot \rangle$  denotes expected values ( $\rightarrow$  I/1.12.1) across the posterior distribution ( $\rightarrow$  I/5.1.9).

#### Sources:

- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002): “Bayesian measures of model complexity and fit”; in: *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 64, iss. 4, pp. 583-639; URL: <https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353>; DOI: 10.1111/1467-9868.00353.
- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eqs. 10-12; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 2.3.2 Deviance

**Definition:** Let there be a generative model ( $\rightarrow$  I/5.1.1)  $m$  describing measured data  $y$  using model parameters  $\theta$ . Then, the deviance of  $m$  is a function of  $\theta$  which multiplies the log-likelihood function ( $\rightarrow$  I/4.1.2) with  $-2$ :

$$D(\theta) = -2 \log p(y | \theta, m) . \quad (1)$$

The deviance function serves the definition of the deviance information criterion ( $\rightarrow$  IV/2.3.1).

#### Sources:

- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002): “Bayesian measures of model complexity and fit”; in: *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 64, iss. 4, pp. 583-639; URL: <https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353>; DOI: 10.1111/1467-9868.00353.
- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eqs. 10-12; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.
- Wikipedia (2022): “Deviance information criterion”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-01; URL: [https://en.wikipedia.org/wiki/Deviance\\_information\\_criterion#Definition](https://en.wikipedia.org/wiki/Deviance_information_criterion#Definition).

### 3 Bayesian model selection

#### 3.1 Model evidence

##### 3.1.1 Definition

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Then, the model evidence (ME) of  $m$  is defined as the marginal likelihood ( $\rightarrow$  I/5.1.15) of this model:

$$\text{ME}(m) = p(y|m) . \quad (1)$$

**Sources:**

- Penny WD (2012): “Comparing Dynamic Causal Models using AIC, BIC and Free Energy”; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eq. 15; URL: <https://www.sciencedirect.com/science/article/pii/S1053811911008160>; DOI: 10.1016/j.neuroimage.2011.07.039.

##### 3.1.2 Derivation

**Theorem:** Let  $p(y|\theta, m)$  be a likelihood function ( $\rightarrow$  I/5.1.2) of a generative model ( $\rightarrow$  I/5.1.1)  $m$  for making inferences on model parameters  $\theta$  given measured data  $y$ . Moreover, let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) on model parameters  $\theta$  in the parameter space  $\Theta$ . Then, the model evidence ( $\rightarrow$  IV/3.1.1) (ME) can be expressed in terms of likelihood ( $\rightarrow$  I/5.1.2) and prior ( $\rightarrow$  I/5.1.3) as

$$\text{ME}(m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (1)$$

**Proof:** This is a consequence of the law of marginal probability ( $\rightarrow$  I/1.3.3) for continuous variables ( $\rightarrow$  I/1.2.6)

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta \quad (2)$$

and the law of conditional probability ( $\rightarrow$  I/1.3.4) according to which

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (3)$$

Plugging (3) into (2), we obtain:

$$\text{ME}(m) = p(y|m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (4)$$

■

##### 3.1.3 Log model evidence

**Definition:** Let  $m$  be a full probability model ( $\rightarrow$  I/5.1.6) with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Then, the log model evidence (LME) of  $m$  is defined as the logarithm of the marginal likelihood ( $\rightarrow$  I/5.1.15) of this model:

$$\text{LME}(m) = \log p(y|m) . \quad (1)$$

**Sources:**

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 13; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

**3.1.4 Derivation of the log model evidence**

**Theorem:** Let  $p(y|\theta, m)$  be a likelihood function ( $\rightarrow$  I/5.1.2) of a generative model ( $\rightarrow$  I/5.1.1)  $m$  for making inferences on model parameters  $\theta$  given measured data  $y$ . Moreover, let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) on model parameters  $\theta$ . Then, the log model evidence ( $\rightarrow$  IV/3.1.3) (LME), also called marginal log-likelihood,

$$\text{LME}(m) = \log p(y|m) , \quad (1)$$

can be expressed in terms of likelihood ( $\rightarrow$  I/5.1.2) and prior ( $\rightarrow$  I/5.1.3) as

$$\text{LME}(m) = \log \int p(y|\theta, m) p(\theta|m) d\theta . \quad (2)$$

**Proof:** This is a consequence of the law of marginal probability ( $\rightarrow$  I/1.3.3) for continuous variables ( $\rightarrow$  I/1.2.6)

$$p(y|m) = \int p(y, \theta|m) d\theta \quad (3)$$

and the law of conditional probability ( $\rightarrow$  I/1.3.4) according to which

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (4)$$

Combining (3) with (4) and logarithmizing, we have:

$$\text{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m) p(\theta|m) d\theta . \quad (5)$$

■

**3.1.5 Expression using prior and posterior**

**Theorem:** Let  $p(y|\theta, m)$  be a likelihood function ( $\rightarrow$  I/5.1.2) of a generative model ( $\rightarrow$  I/5.1.1)  $m$  for making inferences on model parameters  $\theta$  given measured data  $y$ . Moreover, let  $p(\theta|m)$  be a prior distribution ( $\rightarrow$  I/5.1.3) on model parameters  $\theta$ . Then, the log model evidence ( $\rightarrow$  IV/3.1.3) (LME), also called marginal log-likelihood,

$$\text{LME}(m) = \log p(y|m) , \quad (1)$$

can be expressed in terms of prior ( $\rightarrow$  I/5.1.3) and posterior ( $\rightarrow$  I/5.1.9) as

$$\text{LME}(m) = \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) . \quad (2)$$

**Proof:** For a full probability model ( $\rightarrow$  I/5.1.6), Bayes' theorem ( $\rightarrow$  I/5.3.2) makes a statement about the posterior distribution ( $\rightarrow$  I/5.1.9):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (3)$$

Rearranging for  $p(y|m)$  and logarithmizing, we have:

$$\begin{aligned} \text{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta, m) p(\theta|m)}{p(\theta|y, m)} \\ &= \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) . \end{aligned} \quad (4)$$

■

### 3.1.6 Partition into accuracy and complexity

**Theorem:** The log model evidence ( $\rightarrow$  IV/3.1.3) can be partitioned into accuracy and complexity

$$\text{LME}(m) = \text{Acc}(m) - \text{Com}(m) \quad (1)$$

where the accuracy term is the posterior ( $\rightarrow$  I/5.1.9) expectation ( $\rightarrow$  I/1.12.9) of the log-likelihood function ( $\rightarrow$  I/4.1.2)

$$\text{Acc}(m) = \langle \log p(y|\theta, m) \rangle_{p(\theta|y, m)} \quad (2)$$

and the complexity penalty is the Kullback-Leibler divergence ( $\rightarrow$  I/2.5.1) of posterior ( $\rightarrow$  I/5.1.9) from prior ( $\rightarrow$  I/5.1.3)

$$\text{Com}(m) = \text{KL} [p(\theta|y, m) || p(\theta|m)] . \quad (3)$$

**Proof:** We consider Bayesian inference on data ( $\rightarrow$  I/1.1.5)  $y$  using model ( $\rightarrow$  I/5.1.1)  $m$  with parameters  $\theta$ . Then, Bayes' theorem ( $\rightarrow$  I/5.3.2) makes a statement about the posterior distribution ( $\rightarrow$  I/5.1.9), i.e. the probability of parameters, given the data and the model:

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (4)$$

Rearranging this for the model evidence ( $\rightarrow$  IV/3.1.5), we have:

$$p(y|m) = \frac{p(y|\theta, m) p(\theta|m)}{p(\theta|y, m)} . \quad (5)$$

Logarithmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta, m) - \log \frac{p(\theta|y, m)}{p(\theta|m)} . \quad (6)$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y, m) \log p(y|\theta, m) d\theta - \int p(\theta|y, m) \log \frac{p(\theta|y, m)}{p(\theta|m)} d\theta . \quad (7)$$

By definition, the left-hand side is the log model evidence and the terms on the right-hand side correspond to the posterior expectation of the log-likelihood function and the Kullback-Leibler divergence of posterior from prior

$$\text{LME}(m) = \langle \log p(y|\theta, m) \rangle_{p(\theta|y, m)} - \text{KL} [p(\theta|y, m) || p(\theta|m)] \quad (8)$$

which proofs the partition given by (1). ■

#### Sources:

- Beal & Ghahramani (2003): “The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures”; in: *Bayesian Statistics*, vol. 7; URL: <https://mlg.eng.cam.ac.uk/zoubin/papers/valencia02.pdf>.
- Penny et al. (2007): “Bayesian Comparison of Spatially Regularised General Linear Models”; in: *Human Brain Mapping*, vol. 28, pp. 275–293; URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327>; DOI: 10.1002/hbm.20327.
- Ostwald et al. (2014): “A tutorial on variational Bayes for latent linear stochastic time-series models”; in: *Journal of Mathematical Psychology*, vol. 60, pp. 1-19; URL: <https://www.sciencedirect.com/science/article/abs/pii/S0022249614000352>; DOI: 10.1016/j.jmp.2014.04.003.
- Soch et al. (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469–489; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage.2016.07.047.

#### 3.1.7 Subtraction of mean from LMEs

**Theorem:** Subtracting the arithmetic mean from a set of log model evidences ( $\rightarrow$  IV/3.1.3) is equivalent to dividing the corresponding model evidences ( $\rightarrow$  IV/3.1.1) by their geometric mean.

**Proof:** Consider a model space  $\mathcal{M} = \{m_1, \dots, m_M\}$  consisting of  $M$  models ( $\rightarrow$  I/5.1.1). Then, the normalized log model evidence ( $\rightarrow$  IV/3.1.3) of any model  $m_i$ , denoted as  $\text{LME}^*(m_i)$ , may be calculated by subtracting the mean across model space:

$$\text{LME}^*(m_i) = \log p(y|m_i) - \frac{1}{M} \sum_{j=1}^M \log p(y|m_j) . \quad (1)$$

To prove the theorem, we will now rewrite the right-hand side until we arrive at an expression for the normalized model evidence ( $\rightarrow$  IV/3.1.3). First, applying  $c \log_b a = \log_b a^c$ , we obtain

$$\text{LME}^*(m_i) = \log p(y|m_i) - \sum_{j=1}^M \left[ \log p(y|m_j)^{1/M} \right] . \quad (2)$$

Then, exponentiating both sides, we have:

$$\begin{aligned}
\exp [\text{LME}^*(m_i)] &= \frac{\exp [\log p(y|m_i)]}{\exp \left[ \sum_{j=1}^M [\log p(y|m_j)^{1/M}] \right]} \\
&= \frac{p(y|m_i)}{\prod_{j=1}^M \exp [\log p(y|m_j)^{1/M}]} \\
&= \frac{p(y|m_i)}{\prod_{j=1}^M p(y|m_j)^{1/M}} \\
&= \frac{p(y|m_i)}{\left( \prod_{j=1}^M p(y|m_j) \right)^{1/M}} \\
&= \frac{p(y|m_i)}{\sqrt[M]{\prod_{j=1}^M p(y|m_j)}} .
\end{aligned} \tag{3}$$

Finally, the right-hand side is equal to ratio of  $m_i$ 's model evidence to the geometric mean of all model evidences.

■

#### Sources:

- Penny, Will (2015): “Bayesian model selection for group studies using Gibbs sampling”; in: *SPM12*, retrieved on 2023-09-08; URL: [https://github.com/spm/spm12/blob/master/spm\\_BMS\\_gibbs.m](https://github.com/spm/spm12/blob/master/spm_BMS_gibbs.m).
- Soch, Joram (2018): “Random Effects Bayesian Model Selection using Variational Bayes”; in: *MACS – a new SPM toolbox for model assessment, comparison and selection*, retrieved on 2023-09-08; URL: [https://github.com/JoramSoch/MACS/blob/master/ME\\_BMS\\_RFX\\_VB.m](https://github.com/JoramSoch/MACS/blob/master/ME_BMS_RFX_VB.m); DOI: 10.5281/zenodo.845404.

### 3.1.8 Uniform-prior log model evidence

**Definition:** Assume a generative model ( $\rightarrow$  I/5.1.1)  $m$  with likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and a uniform ( $\rightarrow$  I/5.2.2) prior distribution ( $\rightarrow$  I/5.1.3)  $p_{\text{uni}}(\theta|m)$ . Then, the log model evidence ( $\rightarrow$  IV/3.1.3) of this model is called “log model evidence with uniform prior” or “uniform-prior log model evidence” (upLME):

$$\text{upLME}(m) = \log \int p(y|\theta, m) p_{\text{uni}}(\theta|m) d\theta . \tag{1}$$

#### Sources:

- Wikipedia (2020): “Lindley’s paradox”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Lindley%27s\\_paradox#Bayesian\\_approach](https://en.wikipedia.org/wiki/Lindley%27s_paradox#Bayesian_approach).

### 3.1.9 Cross-validated log model evidence

**Definition:** Let there be a data set ( $\rightarrow$  I/1.1.5)  $y$  with mutually exclusive and collectively exhaustive subsets  $y_1, \dots, y_S$ . Assume a generative model ( $\rightarrow$  I/5.1.1)  $m$  with model parameters  $\theta$  implying a

likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and a non-informative ( $\rightarrow$  I/5.2.3) prior density ( $\rightarrow$  I/5.1.3)  $p_{\text{ni}}(\theta|m)$ .

Then, the cross-validated log model evidence (cvLME) of  $m$  is given by

$$\text{cvLME}(m) = \sum_{i=1}^S \log \int p(y_i|\theta, m) p(\theta|y_{-i}, m) d\theta \quad (1)$$

where  $y_{-i} = \bigcup_{j \neq i} y_j$  is the union of all data subsets except  $y_i$  and  $p(\theta|y_{-i}, m)$  is the posterior distribution ( $\rightarrow$  I/5.1.9) obtained from  $y_{-i}$  when using the prior distribution ( $\rightarrow$  I/5.1.3)  $p_{\text{ni}}(\theta|m)$ :

$$p(\theta|y_{-i}, m) = \frac{p(y_{-i}|\theta, m) p_{\text{ni}}(\theta|m)}{p(y_{-i}|m)}. \quad (2)$$

One addend of the cvLME is referred to as the out-of-sample log model evidence (oosLME) of  $m$  for the  $i$ -th data subset:

$$\text{oosLME}_i(m) = \log \int p(y_i|\theta, m) p(\theta|y_{-i}, m) d\theta. \quad (3)$$

#### Sources:

- Soch J, Allefeld C, Haynes JD (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469-489, eqs. 13-15; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage.2016.06.056.
- Soch J, Meyer AP, Allefeld C, Haynes JD (2017): “How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging”; in: *NeuroImage*, vol. 158, pp. 186-195, eq. 6; URL: <https://www.sciencedirect.com/science/article/pii/S105381191730527X>; DOI: 10.1016/j.neuroimage.2017.06.056.
- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eqs. 14-15; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.
- Soch J (2018): “cvBMS and cvBMA: filling in the gaps”; in: *arXiv stat.ME*, 1807.01585, eq. 1; URL: <https://arxiv.org/abs/1807.01585>.

#### 3.1.10 Empirical Bayesian log model evidence

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with model parameters  $\theta$  and hyper-parameters  $\lambda$  implying the likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, \lambda, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|\lambda, m)$ . Then, the Empirical Bayesian ( $\rightarrow$  I/5.3.4) log model evidence ( $\rightarrow$  IV/3.1.3) is the logarithm of the marginal likelihood ( $\rightarrow$  I/5.1.15), maximized with respect to the hyper-parameters:

$$\text{ebLME}(m) = \log p(y|\hat{\lambda}, m) \quad (1)$$

where

$$p(y|\lambda, m) = \int p(y|\theta, \lambda, m) p(\theta|\lambda, m) d\theta \quad (2)$$

and ( $\rightarrow$  I/5.2.7)

$$\hat{\lambda} = \arg \max_{\lambda} \log p(y|\lambda, m). \quad (3)$$



**Sources:**

- Wikipedia (2020): “Empirical Bayes method”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Empirical\\_Bayes\\_method#Introduction](https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction).
- Penny, W.D. and Ridgway, G.R. (2013): “Efficient Posterior Probability Mapping Using Savage-Dickey Ratios”; in: *PLoS ONE*, vol. 8, iss. 3, art. e59655, eqs. 7/11; URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059655>; DOI: 10.1371/journal.pone.0059655.

**3.1.11 Variational Bayesian log model evidence**

**Definition:** Let  $m$  be a generative model ( $\rightarrow$  I/5.1.1) with model parameters ( $\rightarrow$  I/1.5.7)  $\theta \in \Theta$  implying the likelihood function ( $\rightarrow$  I/5.1.2)  $p(y|\theta, m)$  and prior distribution ( $\rightarrow$  I/5.1.3)  $p(\theta|m)$ . Moreover, assume an approximate ( $\rightarrow$  I/5.3.5) posterior distribution ( $\rightarrow$  I/5.1.9)  $q(\theta)$ . Then, the Variational Bayesian ( $\rightarrow$  I/5.3.5) log model evidence ( $\rightarrow$  IV/3.1.3), also referred to as the “variational free energy”, is defined as the expected logarithm of the likelihood function, divided by the approximate posterior:

$$\text{vbLME}(m) = F_m[q(\theta)] = \int_{\Theta} q(\theta) \log \frac{p(\theta|y, m)}{q(\theta)} d\theta . \quad (1)$$

The variational free energy can be decomposed into the difference between log model evidence and KL divergence of approximate from true posterior ( $\rightarrow$  I/5.3.6) or, alternatively, as the difference of expected log-likelihood and KL divergence of approximate posterior from prior ( $\rightarrow$  I/5.3.6).

**Sources:**

- Wikipedia (2020): “Variational Bayesian methods”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: [https://en.wikipedia.org/wiki/Variational\\_Bayesian\\_methods#Evidence\\_lower\\_bound](https://en.wikipedia.org/wiki/Variational_Bayesian_methods#Evidence_lower_bound).
- Penny W, Flandin G, Trujillo-Barreto N (2007): “Bayesian Comparison of Spatially Regularised General Linear Models”; in: *Human Brain Mapping*, vol. 28, pp. 275–293, eqs. 2-9; URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327>; DOI: 10.1002/hbm.20327.

**3.2 Family evidence****3.2.1 Definition**

**Definition:** Let  $f$  be a family of  $M$  generative models ( $\rightarrow$  I/5.1.1)  $m_1, \dots, m_M$ , such that the following statement holds true:

$$f \Leftrightarrow m_1 \vee \dots \vee m_M . \quad (1)$$

Then, the family evidence (FE) of  $f$  is defined as the marginal probability ( $\rightarrow$  I/1.3.3) relative to the model evidences ( $\rightarrow$  IV/3.1.1)  $p(y|m_i)$ , conditional only on  $f$ :

$$\text{FE}(f) = p(y|f) . \quad (2)$$

**Sources:**

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 16; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 3.2.2 Derivation

**Theorem:** Let  $f$  be a family of  $M$  generative models ( $\rightarrow$  I/5.1.1)  $m_1, \dots, m_M$  with model evidences ( $\rightarrow$  IV/3.1.1)  $p(y|m_1), \dots, p(y|m_M)$ . Then, the family evidence ( $\rightarrow$  IV/3.2.1) can be expressed in terms of the model evidences as

$$\text{FE}(f) = \sum_{i=1}^M p(y|m_i) p(m_i|f) \quad (1)$$

where  $p(m_i|f)$  are the within-family ( $\rightarrow$  IV/3.2.3) prior ( $\rightarrow$  I/5.1.3) model ( $\rightarrow$  I/5.1.1) probabilities ( $\rightarrow$  I/1.3.1).

**Proof:** This is a consequence of the law of marginal probability ( $\rightarrow$  I/1.3.3) for discrete variables ( $\rightarrow$  I/1.2.6)

$$p(y|f) = \sum_{i=1}^M p(y, m_i|f) \quad (2)$$

and the law of conditional probability ( $\rightarrow$  I/1.3.4) according to which

$$p(y, m_i|f) = p(y|m_i, f) p(m_i|f) . \quad (3)$$

Since models are nested within model families ( $\rightarrow$  IV/3.2.1), such that  $m_i \wedge f \leftrightarrow m_i$ , we have the following equality of probabilities:

$$p(y|m_i, f) = p(y|m_i \wedge f) = p(y|m_i) . \quad (4)$$

Plugging (3) into (2) and applying (4), we obtain:

$$\text{FE}(f) = p(y|f) = \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (5)$$

■

### 3.2.3 Log family evidence

**Definition:** Let  $f$  be a family of  $M$  generative models ( $\rightarrow$  I/5.1.1)  $m_1, \dots, m_M$ , such that the following statement holds true:

$$f \Leftrightarrow m_1 \vee \dots \vee m_M . \quad (1)$$

Then, the log family evidence is given by the logarithm of the family evidence ( $\rightarrow$  IV/3.2.1):

$$\text{LFE}(f) = \log p(y|f) = \log \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (2)$$

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 16; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 3.2.4 Derivation of the log family evidence

**Theorem:** Let  $f$  be a family of  $M$  generative models ( $\rightarrow$  I/5.1.1)  $m_1, \dots, m_M$  with model evidences ( $\rightarrow$  I/5.1.15)  $p(y|m_1), \dots, p(y|m_M)$ . Then, the log family evidence ( $\rightarrow$  IV/3.2.3)

$$\text{LFE}(f) = \log p(y|f) \quad (1)$$

can be expressed as

$$\text{LFE}(f) = \log \sum_{i=1}^M p(y|m_i) p(m_i|f) \quad (2)$$

where  $p(m_i|f)$  are the within-family ( $\rightarrow$  IV/3.2.3) prior ( $\rightarrow$  I/5.1.3) model ( $\rightarrow$  I/5.1.1) probabilities ( $\rightarrow$  I/1.3.1).

**Proof:** We will assume “prior additivity”

$$p(f) = \sum_{i=1}^M p(m_i) \quad (3)$$

and “posterior additivity” for family probabilities:

$$p(f|y) = \sum_{i=1}^M p(m_i|y) \quad (4)$$

Bayes’ theorem ( $\rightarrow$  I/5.3.2) for the family evidence ( $\rightarrow$  IV/3.2.3) gives

$$p(y|f) = \frac{p(f|y) p(y)}{p(f)} . \quad (5)$$

Applying (3) and (4), we have

$$p(y|f) = \frac{\sum_{i=1}^M p(m_i|y) p(y)}{\sum_{i=1}^M p(m_i)} . \quad (6)$$

Bayes’ theorem ( $\rightarrow$  I/5.3.2) for the model evidence ( $\rightarrow$  IV/3.2.3) gives

$$p(y|m_i) = \frac{p(m_i|y) p(y)}{p(m_i)} \quad (7)$$

which can be rearranged into

$$p(m_i|y) p(y) = p(y|m_i) p(m_i) . \quad (8)$$

Plugging (8) into (6), we have

$$\begin{aligned}
p(y|f) &= \frac{\sum_{i=1}^M p(y|m_i) p(m_i)}{\sum_{i=1}^M p(m_i)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i)}{\sum_{i=1}^M p(m_i)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i, f)}{p(f)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot p(m_i|f) .
\end{aligned} \tag{9}$$

Equation (2) follows by logarithmizing both sides of (9). ■

### 3.2.5 Calculation from log model evidences

**Theorem:** Let  $m_1, \dots, m_M$  be  $M$  statistical models with log model evidences ( $\rightarrow$  IV/3.1.3)  $\text{LME}(m_1), \dots, \text{LME}(m_M)$  and belonging to  $F$  mutually exclusive model families ( $\rightarrow$  IV/3.2.1)  $f_1, \dots, f_F$ . Then, the log family evidences ( $\rightarrow$  IV/3.2.3) are given by:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)], \quad j = 1, \dots, F, \tag{1}$$

where  $p(m_i|f_j)$  are within-family ( $\rightarrow$  IV/3.2.1) prior ( $\rightarrow$  I/5.1.3) model ( $\rightarrow$  I/5.1.1) probabilities ( $\rightarrow$  I/1.3.1).

**Proof:** Let us consider the (unlogarithmized) family evidence ( $\rightarrow$  IV/3.2.1)  $p(y|f_j)$ . According to the law of marginal probability ( $\rightarrow$  I/1.3.3), this conditional probability is given by

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i, f_j) \cdot p(m_i|f_j)] . \tag{2}$$

Because model families are mutually exclusive, it holds that  $p(y|m_i, f_j) = p(y|m_i)$ , such that

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \tag{3}$$

Logarithmizing transforms the family evidence  $p(y|f_j)$  into the log family evidence ( $\rightarrow$  IV/3.2.3)  $\text{LFE}(f_j)$ :

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \tag{4}$$

The definition of the log model evidence ( $\rightarrow$  IV/3.1.3)

$$\text{LME}(m) = \log p(y|m) \tag{5}$$

can be exponentiated to then read

$$\exp[\text{LME}(m)] = p(y|m) \quad (6)$$

and applying (6) to (4), we finally have:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)] . \quad (7)$$

■

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 16; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 3.2.6 Approximation of log family evidences

**Theorem:** Let  $m_1, \dots, m_M$  be  $M$  statistical models with log model evidences ( $\rightarrow$  IV/3.1.3)  $\text{LME}(m_1), \dots, \text{LME}(m_M)$  and belonging to  $F$  mutually exclusive ( $\rightarrow$  I/1.3.10) model families ( $\rightarrow$  IV/3.2.1)  $f_1, \dots, f_F$ .

1) Then, the log family evidences ( $\rightarrow$  IV/3.2.3) can be approximated as

$$\text{LFE}(f_j) = L^*(f_j) + \log \left[ \sum_{m_i \in f_j} \exp[L'(m_i)] \cdot p(m_i|f_j) \right] \quad (1)$$

where  $L^*(f_j)$  is the maximum log model evidence in family  $f_j$ ,  $L'(m_i)$  is the difference of each log model evidence to each family's maximum and  $p(m_i|f_j)$  are within-family prior model probabilities ( $\rightarrow$  IV/3.2.2).

2) Under the condition that prior model probabilities are equal within model families, the approximation simplifies to

$$\text{LFE}(f_j) = L^*(f_j) + \log \sum_{m_i \in f_j} \exp[L'(m_i)] - \log M_j \quad (2)$$

where  $M_j$  is the number of models within family  $f_j$ .

**Proof:** The log family evidence is given in terms of log model evidences ( $\rightarrow$  IV/3.2.5) as

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)] . \quad (3)$$

Often, especially for complex models or many observations, log model evidences ( $\rightarrow$  IV/3.1.3) are highly negative, such that calculation of the term  $\exp[\text{LME}(m_i)]$  in modern computers will give model evidences ( $\rightarrow$  IV/3.1.1) as zero, making calculation of LFEs impossible.

1) As a solution, we select the maximum LME within each family

$$L^*(f_j) = \max_{m_i \in f_j} [\text{LME}(m_i)] \quad (4)$$

and define differences between LMEs and maximum LME as

$$L'(m_i) = \text{LME}(m_i) - L^*(f_j) . \quad (5)$$

In this way, only the differences  $L'(m_i)$  need to be exponentiated. If such a difference is highly negative, this model's contribution to the LFE will be zero – making this an approximation. However, the model is also much less evident that the family's best model in this case – making the approximation acceptable.

Using the relation (5), equation (3) can be reworked into

$$\begin{aligned}
 \text{LFE}(f_j) &= \log \sum_{m_i \in f_j} \exp[L'(m_i) + L^*(f_j)] \cdot p(m_i|f_j) \\
 &= \log \sum_{m_i \in f_j} \exp[L'(m_i)] \cdot \exp[L^*(f_j)] \cdot p(m_i|f_j) \\
 &= \log \left[ \exp[L^*(f_j)] \cdot \sum_{m_i \in f_j} \exp[L'(m_i)] \cdot p(m_i|f_j) \right] \\
 &= L^*(f_j) + \log \left[ \sum_{m_i \in f_j} \exp[L'(m_i)] \cdot p(m_i|f_j) \right].
 \end{aligned} \tag{6}$$

2) Under uniform within-family prior model probabilities ( $\rightarrow$  IV/3.2.2), we have

$$p(m_i|f_j) = \frac{1}{M_j} \quad \text{for all } m_i \in f_j, \tag{7}$$

such that the approximated log family evidences ( $\rightarrow$  IV/3.2.3) becomes

$$\begin{aligned}
 \text{LFE}(f_j) &= L^*(f_j) + \log \left[ \sum_{m_i \in f_j} \exp[L'(m_i)] \cdot \frac{1}{M_j} \right] \\
 &= L^*(f_j) + \log \left[ \frac{1}{M_j} \cdot \sum_{m_i \in f_j} \exp[L'(m_i)] \right] \\
 &= L^*(f_j) + \log \sum_{m_i \in f_j} \exp[L'(m_i)] - \log M_j.
 \end{aligned} \tag{8}$$

■

#### Sources:

- Soch J (2018): “cvBMS and cvBMA: filling in the gaps”; in: *arXiv stat.ME*, 1807.01585, sect. 2.3, eq. 32; URL: <https://arxiv.org/abs/1807.01585>.

## 3.3 Bayes factor

### 3.3.1 Definition

**Definition:** Consider two competing generative models ( $\rightarrow$  I/5.1.1)  $m_1$  and  $m_2$  for observed data  $y$ . Then the Bayes factor in favor  $m_1$  over  $m_2$  is the ratio of marginal likelihoods ( $\rightarrow$  I/5.1.15) of  $m_1$  and  $m_2$ :

$$\text{BF}_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}. \tag{1}$$

Note that by Bayes' theorem ( $\rightarrow$  I/5.3.2), the ratio of posterior model probabilities ( $\rightarrow$  IV/3.4.1) (i.e., the posterior model odds) can be written as

$$\frac{p(m_1 | y)}{p(m_2 | y)} = \frac{p(m_1)}{p(m_2)} \cdot \frac{p(y | m_1)}{p(y | m_2)}, \quad (2)$$

or equivalently by (1),

$$\frac{p(m_1 | y)}{p(m_2 | y)} = \frac{p(m_1)}{p(m_2)} \cdot \text{BF}_{12}. \quad (3)$$

In other words, the Bayes factor can be viewed as the factor by which the prior model odds are updated (after observing data  $y$ ) to posterior model odds – which is also expressed by Bayes' rule ( $\rightarrow$  I/5.3.3).

#### Sources:

- Kass, Robert E. and Raftery, Adrian E. (1995): “Bayes Factors”; in: *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773-795; URL: <https://dx.doi.org/10.1080/01621459.1995.10476572>; DOI: 10.1080/01621459.1995.10476572.

#### 3.3.2 Transitivity

**Theorem:** Consider three competing models ( $\rightarrow$  I/5.1.1)  $m_1$ ,  $m_2$ , and  $m_3$  for observed data  $y$ . Then, the Bayes factor ( $\rightarrow$  IV/3.3.1) in favor of  $m_1$  against  $m_3$  can be written as:

$$\text{BF}_{13} = \text{BF}_{12} \cdot \text{BF}_{23}. \quad (1)$$

**Proof:** By definition ( $\rightarrow$  IV/3.3.1), the Bayes factor  $\text{BF}_{13}$  is the ratio of marginal likelihoods of data  $y$  under  $m_1$  and  $m_3$ , respectively. That is,

$$\text{BF}_{13} = \frac{p(y | m_1)}{p(y | m_3)}. \quad (2)$$

We can equivalently write

$$\begin{aligned} \text{BF}_{13} &\stackrel{(2)}{=} \frac{p(y | m_1)}{p(y | m_3)} \\ &= \frac{p(y | m_1)}{p(y | m_3)} \cdot \frac{p(y | m_2)}{p(y | m_2)} \\ &= \frac{p(y | m_1)}{p(y | m_2)} \cdot \frac{p(y | m_2)}{p(y | m_3)} \\ &\stackrel{(2)}{=} \text{BF}_{12} \cdot \text{BF}_{23}, \end{aligned} \quad (3)$$

which completes the proof of (1). ■

### 3.3.3 Computation using Savage-Dickey density ratio

**Theorem:** Consider two competing models ( $\rightarrow$  I/5.1.1) on data  $y$  containing parameters  $\delta$  and  $\varphi$ , namely  $m_0 : \delta = \delta_0, \varphi$  and  $m_1 : \delta, \varphi$ . In this context, we say that  $\delta$  is a parameter of interest,  $\varphi$  is a nuisance parameter (i.e., common to both models), and  $m_0$  is a sharp point hypothesis nested within  $m_1$ . Suppose further that the prior for the nuisance parameter  $\varphi$  in  $m_0$  is equal to the prior for  $\varphi$  in  $m_1$  after conditioning on the restriction – that is,  $p(\varphi | m_0) = p(\varphi | \delta = \delta_0, m_1)$ . Then the Bayes factor ( $\rightarrow$  IV/3.3.1) for  $m_0$  over  $m_1$  can be computed as:

$$\text{BF}_{01} = \frac{p(\delta = \delta_0 | y, m_1)}{p(\delta = \delta_0 | m_1)}. \quad (1)$$

**Proof:** By definition ( $\rightarrow$  IV/3.3.1), the Bayes factor  $\text{BF}_{01}$  is the ratio of marginal likelihoods of data  $y$  over  $m_0$  and  $m_1$ , respectively. That is,

$$\text{BF}_{01} = \frac{p(y | m_0)}{p(y | m_1)}. \quad (2)$$

The key idea in the proof is that we can use a “change of variables” technique to express  $\text{BF}_{01}$  entirely in terms of the “encompassing” model  $m_1$ . This proceeds by first unpacking the marginal likelihood ( $\rightarrow$  I/5.1.15) for  $m_0$  over the nuisance parameter  $\varphi$  and then using the fact that  $m_0$  is a sharp hypothesis nested within  $m_1$  to rewrite everything in terms of  $m_1$ . Specifically,

$$\begin{aligned} p(y | m_0) &= \int p(y | \varphi, m_0) p(\varphi | m_0) d\varphi \\ &= \int p(y | \varphi, \delta = \delta_0, m_1) p(\varphi | \delta = \delta_0, m_1) d\varphi \\ &= p(y | \delta = \delta_0, m_1). \end{aligned} \quad (3)$$

By Bayes’ theorem ( $\rightarrow$  I/5.3.2), we can rewrite this last line as

$$p(y | \delta = \delta_0, m_1) = \frac{p(\delta = \delta_0 | y, m_1) p(y | m_1)}{p(\delta = \delta_0 | m_1)}. \quad (4)$$

Thus we have

$$\begin{aligned} \text{BF}_{01} &\stackrel{(2)}{=} \frac{p(y | m_0)}{p(y | m_1)} \\ &= p(y | m_0) \cdot \frac{1}{p(y | m_1)} \\ &\stackrel{(3)}{=} p(y | \delta = \delta_0, m_1) \cdot \frac{1}{p(y | m_1)} \\ &\stackrel{(4)}{=} \frac{p(\delta = \delta_0 | y, m_1) p(y | m_1)}{p(\delta = \delta_0 | m_1)} \cdot \frac{1}{p(y | m_1)} \\ &= \frac{p(\delta = \delta_0 | y, m_1)}{p(\delta = \delta_0 | m_1)}, \end{aligned} \quad (5)$$

which completes the proof of (1).



**Sources:**

- Faulkenberry, Thomas J. (2019): “A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors”; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: <https://dx.doi.org/10.29220/CSAM.2019.26.2.217>; DOI: 10.29220/CSAM.2019.26.2.217.
- Penny, W.D. and Ridgway, G.R. (2013): “Efficient Posterior Probability Mapping Using Savage-Dickey Ratios”; in: *PLoS ONE*, vol. 8, iss. 3, art. e59655, eq. 16; URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059655>; DOI: 10.1371/journal.pone.0059655.

**3.3.4 Computation using encompassing prior method**

**Theorem:** Consider two models  $m_1$  and  $m_e$ , where  $m_1$  is nested within an encompassing model ( $\rightarrow$  IV/3.3.5)  $m_e$  via an inequality constraint on some parameter  $\theta$ , and  $\theta$  is unconstrained under  $m_e$ . Then, the Bayes factor ( $\rightarrow$  IV/3.3.1) is

$$\text{BF}_{1e} = \frac{c}{d} = \frac{1/d}{1/c} \quad (1)$$

where  $1/d$  and  $1/c$  represent the proportions of the posterior and prior of the encompassing model, respectively, that are in agreement with the inequality constraint imposed by the nested model  $m_1$ .

**Proof:** Consider first that for any model  $m_1$  on data  $y$  with parameter  $\theta$ , Bayes’ theorem ( $\rightarrow$  I/5.3.2) implies

$$p(\theta \mid y, m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(y \mid m_1)}. \quad (2)$$

Rearranging equation (2) allows us to write the marginal likelihood ( $\rightarrow$  I/5.1.15) for  $y$  under  $m_1$  as

$$p(y \mid m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(\theta \mid y, m_1)}. \quad (3)$$

Taking the ratio of the marginal likelihoods for  $m_1$  and the encompassing model ( $\rightarrow$  IV/3.3.5)  $m_e$  yields the following Bayes factor ( $\rightarrow$  IV/3.3.1):

$$\text{BF}_{1e} = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1) / p(\theta \mid y, m_1)}{p(y \mid \theta, m_e) \cdot p(\theta \mid m_e) / p(\theta \mid y, m_e)}. \quad (4)$$

Now, both the constrained model  $m_1$  and the encompassing model ( $\rightarrow$  IV/3.3.5)  $m_e$  contain the same parameter vector  $\theta$ . Choose a specific value of  $\theta$ , say  $\theta'$ , that exists in the support of both models  $m_1$  and  $m_e$  (we can do this, because  $m_1$  is nested within  $m_e$ ). Then, for this parameter value  $\theta'$ , we have  $p(y \mid \theta', m_1) = p(y \mid \theta', m_e)$ , so the expression for the Bayes factor in equation (4) reduces to an expression involving only the priors and posteriors for  $\theta'$  under  $m_1$  and  $m_e$ :

$$\text{BF}_{1e} = \frac{p(\theta' \mid m_1) / p(\theta' \mid y, m_1)}{p(\theta' \mid m_e) / p(\theta' \mid y, m_e)}. \quad (5)$$

Because  $m_1$  is nested within  $m_e$  via an inequality constraint, the prior  $p(\theta' \mid m_1)$  is simply a truncation of the encompassing prior  $p(\theta' \mid m_e)$ . Thus, we can express  $p(\theta' \mid m_1)$  in terms of the encompassing prior  $p(\theta' \mid m_e)$  by multiplying the encompassing prior by an indicator function over  $m_1$  and then normalizing the resulting product. That is,

$$\begin{aligned}
p(\theta' | m_1) &= \frac{p(\theta' | m_e) \cdot I_{\theta' \in m_1}}{\int p(\theta' | m_e) \cdot I_{\theta' \in m_1} d\theta'} \\
&= \left( \frac{I_{\theta' \in m_1}}{\int p(\theta' | m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' | m_e),
\end{aligned} \tag{6}$$

where  $I_{\theta' \in m_1}$  is an indicator function. For parameters  $\theta' \in m_1$ , this indicator function is identically equal to 1, so the expression in parentheses reduces to a constant, say  $c$ , allowing us to write the prior as

$$p(\theta' | m_1) = c \cdot p(\theta' | m_e). \tag{7}$$

By similar reasoning, we can write the posterior as

$$p(\theta' | y, m_1) = \left( \frac{I_{\theta' \in m_1}}{\int p(\theta' | y, m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' | y, m_e) = d \cdot p(\theta' | y, m_e). \tag{8}$$

Plugging (7) and (8) into (5), this gives us

$$\text{BF}_{1e} = \frac{c \cdot p(\theta' | m_e) / d \cdot p(\theta' | y, m_e)}{p(\theta' | m_e) / p(\theta' | y, m_e)} = \frac{c}{d} = \frac{1/d}{1/c}, \tag{9}$$

which completes the proof. Note that by definition,  $1/d$  represents the proportion of the posterior distribution for  $\theta$  under the encompassing model ( $\rightarrow$  IV/3.3.5)  $m_e$  that agrees with the constraints imposed by  $m_1$ . Similarly,  $1/c$  represents the proportion of the prior distribution for  $\theta$  under the encompassing model ( $\rightarrow$  IV/3.3.5)  $m_e$  that agrees with the constraints imposed by  $m_1$ . ■

#### Sources:

- Klugkist, I., Kato, B., and Hoijsink, H. (2005): “Bayesian model selection using encompassing priors”; in: *Statistica Neerlandica*, vol. 59, no. 1, pp. 57-69; URL: <https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x>; DOI: 10.1111/j.1467-9574.2005.00279.x.
- Faulkenberry, Thomas J. (2019): “A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors”; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: <https://dx.doi.org/10.29220/CSAM.2019.26.2.217>; DOI: 10.29220/CSAM.2019.26.2.217.

### 3.3.5 Encompassing model

**Definition:** Consider a family  $f$  of generative models ( $\rightarrow$  I/5.1.1)  $m$  on data  $y$ , where each  $m \in f$  is defined by placing an inequality constraint on model parameter(s)  $\theta$  (e.g.,  $m : \theta > 0$ ). Then the encompassing model  $m_e$  is constructed such that each  $m$  is nested within  $m_e$  and all inequality constraints on the parameter(s)  $\theta$  are removed.

#### Sources:

- Klugkist, I., Kato, B., and Hoijsink, H. (2005): “Bayesian model selection using encompassing priors”; in: *Statistica Neerlandica*, vol. 59, no. 1, pp. 57-69; URL: <https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x>; DOI: 10.1111/j.1467-9574.2005.00279.x.

### 3.3.6 Log Bayes factor

**Definition:** Let there be two generative models ( $\rightarrow$  I/5.1.1)  $m_1$  and  $m_2$  which are mutually exclusive, but not necessarily collectively exhaustive:

$$\neg(m_1 \wedge m_2) \quad (1)$$

Then, the Bayes factor in favor of  $m_1$  and against  $m_2$  is the ratio of the model evidences ( $\rightarrow$  I/5.1.15) of  $m_1$  and  $m_2$ :

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} . \quad (2)$$

The log Bayes factor is given by the logarithm of the Bayes factor:

$$\text{LBF}_{12} = \log \text{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (3)$$

**Sources:**

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 18; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 3.3.7 Derivation of the log Bayes factor

**Theorem:** Let there be two generative models ( $\rightarrow$  I/5.1.1)  $m_1$  and  $m_2$  with model evidences ( $\rightarrow$  I/5.1.15)  $p(y|m_1)$  and  $p(y|m_2)$ . Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6)

$$\text{LBF}_{12} = \log \text{BF}_{12} \quad (1)$$

can be expressed as

$$\text{LBF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (2)$$

**Proof:** The Bayes factor ( $\rightarrow$  IV/3.3.1) is defined as the posterior ( $\rightarrow$  I/5.1.9) odds ratio ( $\rightarrow$  I/5.3.1) when both models ( $\rightarrow$  I/5.1.1) are equally likely apriori ( $\rightarrow$  I/5.1.3):

$$\text{BF}_{12} = \frac{p(m_1|y)}{p(m_2|y)} \quad (3)$$

Plugging in the posterior odds ratio according to Bayes’ rule ( $\rightarrow$  I/5.3.3), we have

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} . \quad (4)$$

When both models are equally likely apriori, the prior ( $\rightarrow$  I/5.1.3) odds ratio ( $\rightarrow$  I/5.3.1) is one, such that

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} . \quad (5)$$

Equation (2) follows by logarithmizing both sides of (5). ■

### 3.3.8 Calculation from log model evidences

**Theorem:** Let  $m_1$  and  $m_2$  be two statistical models with log model evidences ( $\rightarrow$  IV/3.1.3)  $\text{LME}(m_1)$  and  $\text{LME}(m_2)$ . Then, the log Bayes factor ( $\rightarrow$  IV/3.3.6) in favor of model  $m_1$  and against model  $m_2$  is the difference of the log model evidences:

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (1)$$

**Proof:** The Bayes factor ( $\rightarrow$  IV/3.3.1) is defined as the ratio of the model evidences ( $\rightarrow$  I/5.1.15) of  $m_1$  and  $m_2$

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \quad (2)$$

and the log Bayes factor ( $\rightarrow$  IV/3.3.6) is defined as the logarithm of the Bayes factor

$$\text{LBF}_{12} = \log \text{BF}_{12} . \quad (3)$$

Thus, the log Bayes factor can be expressed as

$$\text{LBF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (4)$$

and, with the definition of the log model evidence ( $\rightarrow$  IV/3.1.3)

$$\text{LME}(m) = \log p(y|m) \quad (5)$$

and resolving the logarithm, we finally have:

$$\begin{aligned} \text{LBF}_{12} &= \log p(y|m_1) - \log p(y|m_2) \\ &= \text{LME}(m_1) - \text{LME}(m_2) . \end{aligned} \quad (6)$$

■

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 18; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

## 3.4 Posterior model probability

### 3.4.1 Definition

**Definition:** Let  $m_1, \dots, m_M$  be  $M$  statistical models ( $\rightarrow$  I/5.1.6) with model evidences ( $\rightarrow$  I/5.1.15)  $p(y|m_1), \dots, p(y|m_M)$  and prior probabilities ( $\rightarrow$  I/5.1.3)  $p(m_1), \dots, p(m_M)$ . Then, the conditional probability ( $\rightarrow$  I/1.3.4) of model  $m_i$ , given the data  $y$ , is called the posterior probability ( $\rightarrow$  I/5.1.9) of model  $m_i$ :

$$\text{PP}(m_i) = p(m_i|y) . \quad (1)$$

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 23; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

### 3.4.2 Derivation

**Theorem:** Let there be a set of generative models ( $\rightarrow$  I/5.1.1)  $m_1, \dots, m_M$  with model evidences ( $\rightarrow$  I/5.1.15)  $p(y|m_1), \dots, p(y|m_M)$  and prior probabilities ( $\rightarrow$  I/5.1.3)  $p(m_1), \dots, p(m_M)$ . Then, the posterior probability ( $\rightarrow$  IV/3.4.1) of model  $m_i$  is given by

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)}, \quad i = 1, \dots, M. \quad (1)$$

**Proof:** From Bayes’ theorem ( $\rightarrow$  I/5.3.2), the posterior model probability ( $\rightarrow$  IV/3.4.1) of the  $i$ -th model can be derived as

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{p(y)}. \quad (2)$$

Using the law of marginal probability ( $\rightarrow$  I/1.3.3), the denominator can be rewritten, such that

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y, m_j)}. \quad (3)$$

Finally, using the law of conditional probability ( $\rightarrow$  I/1.3.4), we have

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)}. \quad (4)$$

■

### 3.4.3 Calculation from Bayes factors

**Theorem:** Let  $m_0, m_1, \dots, m_M$  be  $M + 1$  statistical models ( $\rightarrow$  I/5.1.1) with model evidences ( $\rightarrow$  IV/3.1.3)  $p(y|m_0), p(y|m_1), \dots, p(y|m_M)$ . Then, the posterior model probabilities ( $\rightarrow$  IV/3.4.1) of the models  $m_1, \dots, m_M$  are given by

$$p(m_i|y) = \frac{\text{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j}, \quad i = 1, \dots, M, \quad (1)$$

where  $\text{BF}_{i,0}$  is the Bayes factor ( $\rightarrow$  IV/3.3.1) comparing model  $m_i$  with  $m_0$  and  $\alpha_i$  is the prior ( $\rightarrow$  I/5.1.3) odds ratio ( $\rightarrow$  I/5.3.1) of model  $m_i$  against  $m_0$ .

**Proof:** Define the Bayes factor ( $\rightarrow$  IV/3.3.1) for  $m_i$

$$\text{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \quad (2)$$

and prior odds ratio ( $\rightarrow$  I/5.3.1) of  $m_i$  against  $m_0$

$$\alpha_i = \frac{p(m_i)}{p(m_0)}. \quad (3)$$

The posterior model probability ( $\rightarrow$  IV/3.4.2) of  $m_i$  is given by

$$p(m_i|y) = \frac{p(y|m_i) \cdot p(m_i)}{\sum_{j=1}^M p(y|m_j) \cdot p(m_j)} . \quad (4)$$

Now applying (2) and (3) to (4), we have

$$\begin{aligned} p(m_i|y) &= \frac{\text{BF}_{i,0} p(y|m_0) \cdot \alpha_i p(m_0)}{\sum_{j=1}^M \text{BF}_{j,0} p(y|m_0) \cdot \alpha_j p(m_0)} \\ &= \frac{[p(y|m_0) p(m_0)] \text{BF}_{i,0} \cdot \alpha_i}{[p(y|m_0) p(m_0)] \sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j} , \end{aligned} \quad (5)$$

such that

$$p(m_i|y) = \frac{\text{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j} . \quad (6)$$

■

#### Sources:

- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999): “Bayesian Model Averaging: A Tutorial”; in: *Statistical Science*, vol. 14, no. 4, pp. 382–417, eq. 9; URL: <https://projecteuclid.org/euclid.ss/1009212519>; DOI: 10.1214/ss/1009212519.

### 3.4.4 Calculation from log Bayes factor

**Theorem:** Let  $m_1$  and  $m_2$  be two statistical models with the log Bayes factor ( $\rightarrow$  IV/3.3.6)  $\text{LBF}_{12}$  in favor of model  $m_1$  and against model  $m_2$ . Then, if both models are equally likely apriori ( $\rightarrow$  I/5.1.3), the posterior model probability ( $\rightarrow$  IV/3.4.1) of  $m_1$  is

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (1)$$

**Proof:** From Bayes’ rule ( $\rightarrow$  I/5.3.3), the posterior odds ratio ( $\rightarrow$  I/5.3.1) is

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} . \quad (2)$$

When both models are equally likely apriori ( $\rightarrow$  I/5.1.3), the prior odds ratio ( $\rightarrow$  I/5.3.1) is one, such that

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} . \quad (3)$$

Now the right-hand side corresponds to the Bayes factor ( $\rightarrow$  IV/3.3.1), therefore

$$\frac{p(m_1|y)}{p(m_2|y)} = \text{BF}_{12} . \quad (4)$$

Because the two posterior model probabilities ( $\rightarrow$  IV/3.4.1) add up to 1, we have

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \text{BF}_{12} . \quad (5)$$

Now rearranging for the posterior probability ( $\rightarrow$  IV/3.4.1), this gives

$$p(m_1|y) = \frac{\text{BF}_{12}}{\text{BF}_{12} + 1} . \quad (6)$$

Because the log Bayes factor is the logarithm of the Bayes factor ( $\rightarrow$  IV/3.3.6), we finally have

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (7)$$

■

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 21; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.
- Zeidman P, Silson EH, Schwarzkopf DS, Baker CI, Penny W (2018): “Bayesian population receptive field modelling”; in: *NeuroImage*, vol. 180, pp. 173-187, eq. 11; URL: <https://www.sciencedirect.com/science/article/pii/S1053811917307462>; DOI: 10.1016/j.neuroimage.2017.09.008.
- Zeidman P, Friston K, Parr T (2023): “A primer on Variational Laplace (VL)”; in: *NeuroImage*, vol. 279, art. 120310, eq. 56; URL: <https://www.sciencedirect.com/science/article/pii/S1053811923004615>; DOI: 10.1016/j.neuroimage.2023.120310.

#### 3.4.5 Calculation from log model evidences

**Theorem:** Let  $m_1, \dots, m_M$  be  $M$  statistical models with log model evidences ( $\rightarrow$  IV/3.1.3)  $\text{LME}(m_1), \dots, \text{LME}(m_M)$ . Then, the posterior model probabilities ( $\rightarrow$  IV/3.4.1) are given by:

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)}, \quad i = 1, \dots, M , \quad (1)$$

where  $p(m_i)$  are prior ( $\rightarrow$  I/5.1.3) model probabilities.

**Proof:** The posterior model probability ( $\rightarrow$  IV/3.4.2) can be derived as

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)} . \quad (2)$$

The definition of the log model evidence ( $\rightarrow$  IV/3.1.3)

$$\text{LME}(m) = \log p(y|m) \quad (3)$$

can be exponentiated to then read

$$\exp[\text{LME}(m)] = p(y|m) \quad (4)$$

and applying (4) to (2), we finally have:

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} . \quad (5)$$

**Sources:**

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 23; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

**3.5 Bayesian model averaging****3.5.1 Definition**

**Definition:** Let  $m_1, \dots, m_M$  be  $M$  statistical models ( $\rightarrow$  I/5.1.6) with posterior model probabilities ( $\rightarrow$  IV/3.4.1)  $p(m_1|y), \dots, p(m_M|y)$  and posterior distributions ( $\rightarrow$  I/5.1.9)  $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$ . Then, Bayesian model averaging (BMA) consists in finding the marginal ( $\rightarrow$  I/1.5.3) posterior ( $\rightarrow$  I/5.1.9) density ( $\rightarrow$  I/1.7.1), conditional ( $\rightarrow$  I/1.3.4) on the measured data  $y$ , but unconditional ( $\rightarrow$  I/1.3.3) on the modelling approach  $m$ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) . \quad (1)$$

**Sources:**

- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999): “Bayesian Model Averaging: A Tutorial”; in: *Statistical Science*, vol. 14, no. 4, pp. 382–417, eq. 1; URL: <https://projecteuclid.org/euclid.ss/1009212519>; DOI: 10.1214/ss/1009212519.

**3.5.2 Derivation**

**Theorem:** Let  $m_1, \dots, m_M$  be  $M$  statistical models ( $\rightarrow$  I/5.1.6) with posterior model probabilities ( $\rightarrow$  IV/3.4.1)  $p(m_1|y), \dots, p(m_M|y)$  and posterior distributions ( $\rightarrow$  I/5.1.9)  $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$ . Then, the marginal ( $\rightarrow$  I/1.5.3) posterior ( $\rightarrow$  I/5.1.9) density ( $\rightarrow$  I/1.7.1), conditional ( $\rightarrow$  I/1.3.4) on the measured data  $y$ , but unconditional ( $\rightarrow$  I/1.3.3) on the modelling approach  $m$ , is given by:

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) . \quad (1)$$

**Proof:** Using the law of marginal probability ( $\rightarrow$  I/1.3.3), the probability distribution of the shared parameters  $\theta$  conditional ( $\rightarrow$  I/1.3.4) on the measured data  $y$  can be obtained by marginalizing ( $\rightarrow$  I/1.3.3) over the discrete random variable ( $\rightarrow$  I/1.2.2) model  $m$ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta, m_i|y) . \quad (2)$$

Using the law of the conditional probability ( $\rightarrow$  I/1.3.4), the summand can be expanded to give

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) \quad (3)$$

where  $p(\theta|y, m_i)$  is the posterior distribution ( $\rightarrow$  I/5.1.9) of the  $i$ -th model and  $p(m_i|y)$  happens to be the posterior probability ( $\rightarrow$  IV/3.4.1) of the  $i$ -th model.



■

### 3.5.3 Calculation from log model evidences

**Theorem:** Let  $m_1, \dots, m_M$  be  $M$  statistical models ( $\rightarrow$  I/5.1.6) describing the same measured data  $y$  with log model evidences ( $\rightarrow$  IV/3.1.3)  $\text{LME}(m_1), \dots, \text{LME}(m_M)$  and shared model parameters  $\theta$ . Then, Bayesian model averaging ( $\rightarrow$  IV/3.5.1) determines the following posterior distribution over  $\theta$ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|m_i, y) \cdot \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} , \quad (1)$$

where  $p(\theta|m_i, y)$  is the posterior distributions over  $\theta$  obtained using  $m_i$ .

**Proof:** According to the law of marginal probability ( $\rightarrow$  I/1.3.3), the probability of the shared parameters  $\theta$  conditional on the measured data  $y$  can be obtained ( $\rightarrow$  IV/3.5.2) by marginalizing over the discrete variable model  $m$ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|m_i, y) \cdot p(m_i|y) , \quad (2)$$

where  $p(m_i|y)$  is the posterior probability ( $\rightarrow$  IV/3.4.1) of the  $i$ -th model. One can express posterior model probabilities in terms of log model evidences ( $\rightarrow$  IV/3.4.5) as

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} \quad (3)$$

and by plugging (3) into (2), one arrives at (1).

■

#### Sources:

- Soch J, Allefeld C (2018): “MACS – a new SPM toolbox for model assessment, comparison and selection”; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 25; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.



Chapter V

Appendix

# 1 Proof by Number

| ID  | Shortcut   | Theorem  | Author    | Date       | Page |
|-----|------------|--|-----------|------------|------|
| P1  | mvn-ltt    | Linear transformation theorem for the multivariate normal distribution | JoramSoch | 2019-08-27 | 347  |
| P2  | mlr-ols    | Ordinary least squares for multiple linear regression                  | JoramSoch | 2019-09-27 | 524  |
| P3  | lme-anc    | Partition of the log model evidence into accuracy and complexity       | JoramSoch | 2019-09-27 | 713  |
| P4  | bayes-th   | Bayes' theorem   | JoramSoch | 2019-09-27 | 156  |
| P5  | mse-bnv    | Partition of the mean squared error into bias and variance             | JoramSoch | 2019-11-27 | 134  |
| P6  | ng-kl      | Kullback-Leibler divergence for the normal-gamma distribution          | JoramSoch | 2019-12-06 | 373  |
| P7  | glm-mle    | Maximum likelihood estimation for the general linear model             | JoramSoch | 2019-12-06 | 605  |
| P8  | rsq-der    | Derivation of $R^2$ and adjusted $R^2$                                 | JoramSoch | 2019-12-06 | 690  |
| P9  | blr-prior  | Conjugate prior distribution for Bayesian linear regression            | JoramSoch | 2020-01-03 | 565  |
| P10 | blr-post   | Posterior distribution for Bayesian linear regression                  | JoramSoch | 2020-01-03 | 567  |
| P11 | blr-lme    | Log model evidence for Bayesian linear regression                      | JoramSoch | 2020-01-03 | 569  |
| P12 | bayes-rule | Bayes' rule  | JoramSoch | 2020-01-06 | 157  |
| P13 | lme-der    | Derivation of the log model evidence                                   | JoramSoch | 2020-01-06 | 712  |
| P14 | rsq-mll    | Relationship between $R^2$ and maximum log-likelihood                  | JoramSoch | 2020-01-08 | 692  |
| P15 | norm-mean  | Mean of the normal distribution  | JoramSoch | 2020-01-09 | 235  |
| P16 | norm-med   | Median of the normal distribution                                      | JoramSoch | 2020-01-09 | 237  |
| P17 | norm-mode  | Mode of the normal distribution  | JoramSoch | 2020-01-09 | 237  |
| P18 | norm-var   | Variance of the normal distribution                                    | JoramSoch | 2020-01-09 | 238  |
| P19 | dmi-mce    | Relation of mutual information to marginal and conditional entropy     | JoramSoch | 2020-01-13 | 117  |

|     |           |  |           |            |     |
|-----|-----------|--|-----------|------------|-----|
| P20 | dmi-mje   | Relation of mutual information to marginal and joint entropy                   | JoramSoch | 2020-01-13 | 118 |
| P21 | dmi-jce   | Relation of mutual information to joint and conditional entropy                | JoramSoch | 2020-01-13 | 119 |
| P22 | bern-mean | Mean of the Bernoulli distribution   | JoramSoch | 2020-01-16 | 179 |
| P23 | bin-mean  | Mean of the binomial distribution  | JoramSoch | 2020-01-16 | 186 |
| P24 | cat-mean  | Mean of the categorical distribution   | JoramSoch | 2020-01-16 | 201 |
| P25 | mult-mean | Mean of the multinomial distribution   | JoramSoch | 2020-01-16 | 205 |
| P26 | matn-mvn  | Equivalence of matrix-normal distribution and multivariate normal distribution | JoramSoch | 2020-01-20 | 386 |
| P27 | poiss-mle | Maximum likelihood estimation for Poisson-distributed data                     | JoramSoch | 2020-01-20 | 657 |
| P28 | beta-mome | Method of moments for beta-distributed data                                    | JoramSoch | 2020-01-22 | 671 |
| P29 | bin-prior | Conjugate prior distribution for binomial observations                         | JoramSoch | 2020-01-23 | 634 |
| P30 | bin-post  | Posterior distribution for binomial observations                               | JoramSoch | 2020-01-24 | 635 |
| P31 | bin-lme   | Log model evidence for binomial observations                                   | JoramSoch | 2020-01-24 | 636 |
| P32 | bic-der   | Derivation of the Bayesian information criterion                               | JoramSoch | 2020-01-26 | 708 |
| P33 | norm-pdf  | Probability density function of the normal distribution                        | JoramSoch | 2020-01-27 | 228 |
| P34 | mvn-pdf   | Probability density function of the multivariate normal distribution           | JoramSoch | 2020-01-27 | 336 |
| P35 | mvn-marg  | Marginal distributions of the multivariate normal distribution                 | JoramSoch | 2020-01-29 | 348 |
| P36 | ng-marg   | Marginal distributions of the normal-gamma distribution                        | JoramSoch | 2020-01-29 | 375 |
| P37 | cuni-pdf  | Probability density function of the continuous uniform distribution            | JoramSoch | 2020-01-31 | 210 |
| P38 | cuni-cdf  | Cumulative distribution function of the continuous uniform distribution        | JoramSoch | 2020-01-02 | 211 |

|     |                |  |                 |            |     |
|-----|----------------|--|-----------------|------------|-----|
| P39 | cuni-qf        | Quantile function of the continuous uniform distribution                       | JoramSoch       | 2020-01-02 | 212 |
| P40 | mlr-ols2       | Ordinary least squares for multiple linear regression                          | JoramSoch       | 2020-02-03 | 525 |
| P41 | poissexp-prior | Conjugate prior distribution for the Poisson distribution with exposure values | JoramSoch       | 2020-02-04 | 665 |
| P42 | poissexp-post  | Posterior distribution for the Poisson distribution with exposure values       | JoramSoch       | 2020-02-04 | 666 |
| P43 | poissexp-lme   | Log model evidence for the Poisson distribution with exposure values           | JoramSoch       | 2020-02-04 | 668 |
| P44 | ng-pdf         | Probability density function of the normal-gamma distribution                  | JoramSoch       | 2020-02-07 | 368 |
| P45 | gam-pdf        | Probability density function of the gamma distribution                         | JoramSoch       | 2020-02-08 | 260 |
| P46 | exp-pdf        | Probability density function of the exponential distribution                   | JoramSoch       | 2020-02-08 | 274 |
| P47 | exp-mean       | Mean of the exponential distribution   | JoramSoch       | 2020-02-10 | 277 |
| P48 | exp-cdf        | Cumulative distribution function of the exponential distribution               | JoramSoch       | 2020-02-11 | 275 |
| P49 | exp-med        | Median of the exponential distribution   | JoramSoch       | 2020-02-11 | 278 |
| P50 | exp-qf         | Quantile function of the exponential distribution                              | JoramSoch       | 2020-02-12 | 276 |
| P51 | exp-mode       | Mode of the exponential distribution   | kantundpeterpan | 2020-02-12 | 278 |
| P52 | mean-nonneg    | Non-negativity of the expected value   | JoramSoch       | 2020-02-13 | 52  |
| P53 | mean-lin       | Linearity of the expected value  | JoramSoch       | 2020-02-13 | 53  |
| P54 | mean-mono      | Monotonicity of the expected value   | JoramSoch       | 2020-02-17 | 54  |
| P55 | mean-mult      | (Non-)Multiplicativity of the expected value                                   | JoramSoch       | 2020-02-17 | 55  |
| P56 | ci-wilks       | Construction of confidence intervals using Wilks' theorem                      | JoramSoch       | 2020-02-19 | 135 |
| P57 | ent-nonneg     | Non-negativity of the Shannon entropy  | JoramSoch       | 2020-02-19 | 104 |

|     |             |  |           |            |     |
|-----|-------------|--|-----------|------------|-----|
| P58 | cmi-mcde    | Relation of continuous mutual information to marginal and conditional differential entropy | JoramSoch | 2020-02-21 | 121 |
| P59 | cmi-mjde    | Relation of continuous mutual information to marginal and joint differential entropy       | JoramSoch | 2020-02-21 | 122 |
| P60 | cmi-jcde    | Relation of continuous mutual information to joint and conditional differential entropy    | JoramSoch | 2020-02-21 | 123 |
| P61 | resvar-bias | Maximum likelihood estimator of variance is biased   | JoramSoch | 2020-02-24 | 684 |
| P62 | resvar-unb  | Construction of unbiased estimator for variance  | JoramSoch | 2020-02-25 | 688 |
| P63 | snr-rsq     | Relationship between signal-to-noise ratio and $R^2$                                       | JoramSoch | 2020-02-26 | 704 |
| P64 | lbf-lme     | Log Bayes factor in terms of log model evidences   | JoramSoch | 2020-02-27 | 728 |
| P65 | lfe-lme     | Log family evidences in terms of log model evidences                                       | JoramSoch | 2020-02-27 | 720 |
| P66 | pmp-lme     | Posterior model probabilities in terms of log model evidences                              | JoramSoch | 2020-02-27 | 731 |
| P67 | bma-lme     | Bayesian model averaging in terms of log model evidences                                   | JoramSoch | 2020-02-27 | 733 |
| P68 | dent-neg    | Differential entropy can be negative   | JoramSoch | 2020-03-02 | 110 |
| P69 | exp-gam     | Exponential distribution is a special case of gamma distribution                           | JoramSoch | 2020-03-02 | 273 |
| P70 | matn-pdf    | Probability density function of the matrix-normal distribution                             | JoramSoch | 2020-03-02 | 387 |
| P71 | norm-mgf    | Moment-generating function of the normal distribution                                      | JoramSoch | 2020-03-03 | 229 |
| P72 | logreg-lonp | Log-odds and probability in logistic regression  | JoramSoch | 2020-03-03 | 680 |
| P73 | pmp-lbf     | Posterior model probability in terms of log Bayes factor                                   | JoramSoch | 2020-03-03 | 730 |
| P74 | pmp-bf      | Posterior model probabilities in terms of Bayes factors                                    | JoramSoch | 2020-03-03 | 729 |

|     |              |  |           |            |     |
|-----|--------------|--|-----------|------------|-----|
| P75 | mlr-mat      | Transformation matrices for ordinary least squares   | JoramSoch | 2020-03-09 | 532 |
| P76 | mlr-pss      | Partition of sums of squares for multiple linear regression  | JoramSoch | 2020-03-09 | 530 |
| P77 | mlr-wls      | Weighted least squares for multiple linear regression  | JoramSoch | 2020-03-11 | 542 |
| P78 | mlr-mle      | Maximum likelihood estimation for multiple linear regression   | JoramSoch | 2020-03-11 | 544 |
| P79 | mult-prior   | Conjugate prior distribution for multinomial observations  | JoramSoch | 2020-03-11 | 647 |
| P80 | mult-post    | Posterior distribution for multinomial observations  | JoramSoch | 2020-03-11 | 648 |
| P81 | mult-lme     | Log model evidence for multinomial observations  | JoramSoch | 2020-03-11 | 649 |
| P82 | cuni-mean    | Mean of the continuous uniform distribution  | JoramSoch | 2020-03-16 | 213 |
| P83 | cuni-med     | Median of the continuous uniform distribution  | JoramSoch | 2020-03-16 | 214 |
| P84 | cuni-mode    | Mode of the continuous uniform distribution  | JoramSoch | 2020-03-16 | 214 |
| P85 | norm-cdf     | Cumulative distribution function of the normal distribution  | JoramSoch | 2020-03-20 | 230 |
| P86 | norm-cdfwerf | Expression of the cumulative distribution function of the normal distribution without the error function | JoramSoch | 2020-03-20 | 232 |
| P87 | norm-qf      | Quantile function of the normal distribution   | JoramSoch | 2020-03-20 | 235 |
| P88 | mvn-cond     | Conditional distributions of the multivariate normal distribution  | JoramSoch | 2020-03-20 | 349 |
| P89 | jl-lfnprior  | Joint likelihood is the product of likelihood function and prior density                                 | JoramSoch | 2020-05-05 | 149 |
| P90 | post-jl      | Posterior density is proportional to joint likelihood  | JoramSoch | 2020-05-05 | 149 |
| P91 | ml-jl        | Marginal likelihood is a definite integral of the joint likelihood                                       | JoramSoch | 2020-05-05 | 152 |
| P92 | mvn-kl       | Kullback-Leibler divergence for the multivariate normal distribution                                     | JoramSoch | 2020-05-05 | 344 |



|      |             |   |           |            |     |
|------|-------------|---|-----------|------------|-----|
| P93  | gam-kl      | Kullback-Leibler divergence for the gamma distribution                    | JoramSoch | 2020-05-05 | 272 |
| P94  | beta-pdf    | Probability density function of the beta distribution                     | JoramSoch | 2020-05-05 | 305 |
| P95  | dir-pdf     | Probability density function of the Dirichlet distribution                | JoramSoch | 2020-05-05 | 380 |
| P96  | bern-pmf    | Probability mass function of the Bernoulli distribution                   | JoramSoch | 2020-05-11 | 178 |
| P97  | bin-pmf     | Probability mass function of the binomial distribution                    | JoramSoch | 2020-05-11 | 183 |
| P98  | cat-pmf     | Probability mass function of the categorical distribution                 | JoramSoch | 2020-05-11 | 201 |
| P99  | mult-pmf    | Probability mass function of the multinomial distribution                 | JoramSoch | 2020-05-11 | 203 |
| P100 | mvn-dent    | Differential entropy of the multivariate normal distribution              | JoramSoch | 2020-05-14 | 342 |
| P101 | norm-dent   | Differential entropy of the normal distribution                           | JoramSoch | 2020-05-14 | 243 |
| P102 | poiss-pmf   | Probability mass function of the Poisson distribution                     | JoramSoch | 2020-05-14 | 196 |
| P103 | mean-nnrvar | Expected value of a non-negative random variable                          | JoramSoch | 2020-05-18 | 51  |
| P104 | var-mean    | Partition of variance into expected values                                | JoramSoch | 2020-05-19 | 70  |
| P105 | logreg-pnlo | Probability and log-odds in logistic regression                           | JoramSoch | 2020-05-19 | 679 |
| P106 | glm-ols     | Ordinary least squares for the general linear model                       | JoramSoch | 2020-05-19 | 604 |
| P107 | glm-wls     | Weighted least squares for the general linear model                       | JoramSoch | 2020-05-19 | 604 |
| P108 | gam-mean    | Mean of the gamma distribution  | JoramSoch | 2020-05-19 | 263 |
| P109 | gam-var     | Variance of the gamma distribution  | JoramSoch | 2020-05-19 | 266 |
| P110 | gam-logmean | Logarithmic expectation of the gamma distribution                         | JoramSoch | 2020-05-25 | 268 |
| P111 | norm-snorm  | Relationship between normal distribution and standard normal distribution | JoramSoch | 2020-05-26 | 220 |

|      |                 |  |           |            |     |
|------|-----------------|--|-----------|------------|-----|
| P112 | gam-sgam        | Relationship between gamma distribution and standard gamma distribution      | JoramSoch | 2020-05-26 | 258 |
| P113 | kl-ent          | Relation of discrete Kullback-Leibler divergence to Shannon entropy          | JoramSoch | 2020-05-27 | 131 |
| P114 | kl-dent         | Relation of continuous Kullback-Leibler divergence to differential entropy   | JoramSoch | 2020-05-27 | 132 |
| P115 | kl-inv          | Invariance of the Kullback-Leibler divergence under parameter transformation | JoramSoch | 2020-05-28 | 130 |
| P116 | kl-add          | Additivity of the Kullback-Leibler divergence for independent distributions  | JoramSoch | 2020-05-31 | 129 |
| P117 | kl-nonneg       | Non-negativity of the Kullback-Leibler divergence                            | JoramSoch | 2020-05-31 | 124 |
| P118 | cov-mean        | Partition of covariance into expected values                                 | JoramSoch | 2020-06-02 | 78  |
| P119 | cov-corr        | Relationship between covariance and correlation                              | JoramSoch | 2020-06-02 | 81  |
| P120 | covmat-mean     | Partition of a covariance matrix into expected values                        | JoramSoch | 2020-06-06 | 82  |
| P121 | covmat-corrmat  | Relationship between covariance matrix and correlation matrix                | JoramSoch | 2020-06-06 | 86  |
| P122 | precmat-corrmat | Relationship between precision matrix and correlation matrix                 | JoramSoch | 2020-06-06 | 89  |
| P123 | var-nonneg      | Non-negativity of the variance   | JoramSoch | 2020-06-06 | 71  |
| P124 | var-const       | Variance of constant is zero   | JoramSoch | 2020-06-27 | 71  |
| P126 | var-inv         | Invariance of the variance under addition of a constant                      | JoramSoch | 2020-07-07 | 72  |
| P127 | var-scal        | Scaling of the variance upon multiplication with a constant                  | JoramSoch | 2020-07-07 | 73  |
| P128 | var-sum         | Variance of the sum of two random variables                                  | JoramSoch | 2020-07-07 | 73  |
| P129 | var-lincomb     | Variance of the linear combination of two random variables                   | JoramSoch | 2020-07-07 | 74  |

|      |            |   |           |            |     |
|------|------------|---|-----------|------------|-----|
| P130 | var-add    | Additivity of the variance for independent random variables                                     | JoramSoch | 2020-07-07 | 75  |
| P131 | mean-qf    | Expectation of a quadratic form   | JoramSoch | 2020-07-13 | 66  |
| P132 | lfe-der    | Derivation of the log family evidence   | JoramSoch | 2020-07-13 | 719 |
| P133 | blr-pp     | Posterior probability of the alternative hypothesis for Bayesian linear regression              | JoramSoch | 2020-07-17 | 580 |
| P134 | blr-pcr    | Posterior credibility region against the omnibus null hypothesis for Bayesian linear regression | JoramSoch | 2020-07-17 | 582 |
| P135 | mlr-idem   | Projection matrix and residual-forming matrix are idempotent                                    | JoramSoch | 2020-07-22 | 535 |
| P136 | mlr-wls2   | Weighted least squares for multiple linear regression   | JoramSoch | 2020-07-22 | 543 |
| P137 | lbf-der    | Derivation of the log Bayes factor  | JoramSoch | 2020-07-22 | 727 |
| P138 | mean-lotus | Law of the unconscious statistician   | JoramSoch | 2020-07-22 | 58  |
| P139 | pmp-der    | Derivation of the posterior model probability   | JoramSoch | 2020-07-28 | 729 |
| P140 | duni-pmf   | Probability mass function of the discrete uniform distribution                                  | JoramSoch | 2020-07-28 | 172 |
| P141 | duni-cdf   | Cumulative distribution function of the discrete uniform distribution                           | JoramSoch | 2020-07-28 | 172 |
| P142 | duni-qf    | Quantile function of the discrete uniform distribution  | JoramSoch | 2020-07-28 | 173 |
| P143 | bma-der    | Derivation of Bayesian model averaging  | JoramSoch | 2020-08-03 | 732 |
| P144 | matn-trans | Transposition of a matrix-normal random variable  | JoramSoch | 2020-08-03 | 391 |
| P145 | matn-ltt   | Linear transformation theorem for the matrix-normal distribution                                | JoramSoch | 2020-08-03 | 393 |
| P146 | ng-cond    | Conditional distributions of the normal-gamma distribution                                      | JoramSoch | 2020-08-05 | 377 |
| P147 | kl-nonsymm | Non-symmetry of the Kullback-Leibler divergence   | JoramSoch | 2020-08-11 | 127 |
| P148 | kl-conv    | Convexity of the Kullback-Leibler divergence  | JoramSoch | 2020-08-11 | 128 |

|      |               |  |                 |            |     |
|------|---------------|--|-----------------|------------|-----|
| P149 | ent-conc      | Concavity of the Shannon entropy   | JoramSoch       | 2020-08-11 | 105 |
| P150 | entcross-conv | Convexity of the cross-entropy   | JoramSoch       | 2020-08-11 | 107 |
| P151 | poiss-mean    | Mean of the Poisson distribution   | JoramSoch       | 2020-08-19 | 196 |
| P152 | norm-fwhm     | Full width at half maximum for the normal distribution                           | JoramSoch       | 2020-08-19 | 240 |
| P153 | mom-mgf       | Moment in terms of moment-generating function                                    | JoramSoch       | 2020-08-19 | 99  |
| P154 | mgf-ltt       | Linear transformation theorem for the moment-generating function                 | JoramSoch       | 2020-08-19 | 44  |
| P155 | mgf-lincomb   | Moment-generating function of linear combination of independent random variables | JoramSoch       | 2020-08-19 | 45  |
| P156 | bf-sddr       | Savage-Dickey density ratio for computing Bayes factors                          | tomfaulkenberry | 2020-08-26 | 724 |
| P157 | bf-ep         | Encompassing prior method for computing Bayes factors                            | tomfaulkenberry | 2020-09-02 | 725 |
| P158 | cov-ind       | Covariance of independent random variables                                       | JoramSoch       | 2020-09-03 | 80  |
| P159 | mblr-prior    | Conjugate prior distribution for multivariate Bayesian linear regression         | JoramSoch       | 2020-09-03 | 623 |
| P160 | mblr-post     | Posterior distribution for multivariate Bayesian linear regression               | JoramSoch       | 2020-09-03 | 624 |
| P161 | mblr-lme      | Log model evidence for multivariate Bayesian linear regression                   | JoramSoch       | 2020-09-03 | 627 |
| P162 | wald-pdf      | Probability density function of the Wald distribution                            | tomfaulkenberry | 2020-09-04 | 315 |
| P163 | bf-trans      | Transitivity of Bayes Factors  | tomfaulkenberry | 2020-09-07 | 723 |
| P164 | gibbs-ineq    | Gibbs' inequality  | JoramSoch       | 2020-09-09 | 108 |
| P165 | logsum-ineq   | Log sum inequality   | JoramSoch       | 2020-09-09 | 109 |
| P166 | kl-nonneg2    | Non-negativity of the Kullback-Leibler divergence                                | JoramSoch       | 2020-09-09 | 125 |
| P167 | momcent-1st   | First central moment is zero   | JoramSoch       | 2020-09-09 | 102 |

|      |             |   |                 |            |     |
|------|-------------|---|-----------------|------------|-----|
| P168 | wald-mgf    | Moment-generating function of the Wald distribution                                       | tomfaulkenberry | 2020-09-13 | 315 |
| P169 | wald-mean   | Mean of the Wald distribution   | tomfaulkenberry | 2020-09-13 | 317 |
| P170 | wald-var    | Variance of the Wald distribution   | tomfaulkenberry | 2020-09-13 | 318 |
| P171 | momraw-1st  | First raw moment is mean  | JoramSoch       | 2020-10-08 | 101 |
| P172 | momraw-2nd  | Relationship between second raw moment, variance and mean                                 | JoramSoch       | 2020-10-08 | 101 |
| P173 | momcent-2nd | Second central moment is variance   | JoramSoch       | 2020-10-08 | 102 |
| P174 | chi2-gam    | Chi-squared distribution is a special case of gamma distribution                          | kjpetrykowski   | 2020-10-12 | 294 |
| P175 | chi2-momraw | Raw moments of the chi-squared distribution   | kjpetrykowski   | 2020-10-13 | 297 |
| P176 | norm-snorm2 | Relationship between normal distribution and standard normal distribution                 | JoramSoch       | 2020-10-15 | 222 |
| P177 | gam-sgam2   | Relationship between gamma distribution and standard gamma distribution                   | JoramSoch       | 2020-10-15 | 259 |
| P178 | gam-cdf     | Cumulative distribution function of the gamma distribution                                | JoramSoch       | 2020-10-15 | 262 |
| P179 | gam-xlogx   | Expected value of $x \ln(x)$ for a gamma distribution                                     | JoramSoch       | 2020-10-15 | 270 |
| P180 | norm-snorm3 | Relationship between normal distribution and standard normal distribution                 | JoramSoch       | 2020-10-22 | 222 |
| P181 | dir-ep      | Exceedance probabilities for the Dirichlet distribution                                   | JoramSoch       | 2020-10-22 | 382 |
| P182 | dir-mle     | Maximum likelihood estimation for Dirichlet-distributed data                              | JoramSoch       | 2020-10-22 | 673 |
| P183 | cdf-sifct   | Cumulative distribution function of a strictly increasing function of a random variable   | JoramSoch       | 2020-10-29 | 37  |
| P184 | pmf-sifct   | Probability mass function of a strictly increasing function of a discrete random variable | JoramSoch       | 2020-10-29 | 26  |

|      |           |   |           |            |     |
|------|-----------|---|-----------|------------|-----|
| P185 | pdf-sifct | Probability density function of a strictly increasing function of a continuous random variable            | JoramSoch | 2020-10-29 | 29  |
| P186 | cdf-sdfct | Cumulative distribution function of a strictly decreasing function of a random variable                   | JoramSoch | 2020-11-06 | 38  |
| P187 | pmf-sdfct | Probability mass function of a strictly decreasing function of a discrete random variable                 | JoramSoch | 2020-11-06 | 27  |
| P188 | pdf-sdfct | Probability density function of a strictly decreasing function of a continuous random variable            | JoramSoch | 2020-11-06 | 30  |
| P189 | cdf-pmf   | Cumulative distribution function in terms of probability mass function of a discrete random variable      | JoramSoch | 2020-11-12 | 39  |
| P190 | cdf-pdf   | Cumulative distribution function in terms of probability density function of a continuous random variable | JoramSoch | 2020-11-12 | 39  |
| P191 | pdf-cdf   | Probability density function is first derivative of cumulative distribution function                      | JoramSoch | 2020-11-12 | 34  |
| P192 | qf-cdf    | Quantile function is inverse of strictly monotonically increasing cumulative distribution function        | JoramSoch | 2020-11-12 | 48  |
| P193 | norm-kl   | Kullback-Leibler divergence for the normal distribution   | JoramSoch | 2020-11-19 | 244 |
| P194 | gam-qf    | Quantile function of the gamma distribution   | JoramSoch | 2020-11-19 | 263 |
| P195 | beta-cdf  | Cumulative distribution function of the beta distribution   | JoramSoch | 2020-11-19 | 307 |
| P196 | norm-gi   | Gaussian integral   | JoramSoch | 2020-11-25 | 227 |
| P197 | chi2-pdf  | Probability density function of the chi-squared distribution  | JoramSoch | 2020-11-25 | 295 |
| P198 | beta-mgf  | Moment-generating function of the beta distribution   | JoramSoch | 2020-11-25 | 306 |
| P199 | dent-inv  | Invariance of the differential entropy under addition of a constant                                       | JoramSoch | 2020-12-02 | 111 |

|      |             |   |           |            |     |
|------|-------------|---|-----------|------------|-----|
| P200 | dent-add    | Addition of the differential entropy upon multiplication with a constant      | JoramSoch | 2020-12-02 | 112 |
| P201 | ug-prior    | Conjugate prior distribution for the univariate Gaussian                      | JoramSoch | 2021-03-03 | 416 |
| P202 | ug-post     | Posterior distribution for the univariate Gaussian                            | JoramSoch | 2021-03-03 | 418 |
| P203 | ug-lme      | Log model evidence for the univariate Gaussian                                | JoramSoch | 2021-03-03 | 421 |
| P204 | ug-ttest1   | One-sample t-test for independent observations                                | JoramSoch | 2021-03-12 | 408 |
| P205 | ug-ttest2   | Two-sample t-test for independent observations                                | JoramSoch | 2021-03-12 | 409 |
| P206 | ug-ttestp   | Paired t-test for dependent observations                                      | JoramSoch | 2021-03-12 | 411 |
| P207 | ugkv-mle    | Maximum likelihood estimation for the univariate Gaussian with known variance | JoramSoch | 2021-03-24 | 430 |
| P208 | ugkv-ztest1 | One-sample z-test for independent observations                                | JoramSoch | 2021-03-24 | 432 |
| P209 | ugkv-ztest2 | Two-sample z-test for independent observations                                | JoramSoch | 2021-03-24 | 433 |
| P210 | ugkv-ztestp | Paired z-test for dependent observations                                      | JoramSoch | 2021-03-24 | 434 |
| P211 | ugkv-prior  | Conjugate prior distribution for the univariate Gaussian with known variance  | JoramSoch | 2021-03-24 | 435 |
| P212 | ugkv-post   | Posterior distribution for the univariate Gaussian with known variance        | JoramSoch | 2021-03-24 | 437 |
| P213 | ugkv-lme    | Log model evidence for the univariate Gaussian with known variance            | JoramSoch | 2021-03-24 | 439 |
| P214 | ugkv-anc    | Accuracy and complexity for the univariate Gaussian with known variance       | JoramSoch | 2021-03-24 | 441 |
| P215 | ugkv-lbf    | Log Bayes factor for the univariate Gaussian with known variance              | JoramSoch | 2021-03-24 | 442 |

|      |                |   |           |            |     |
|------|----------------|---|-----------|------------|-----|
| P216 | ugkv-lbfmean   | Expectation of the log Bayes factor for the univariate Gaussian with known variance                 | JoramSoch | 2021-03-24 | 443 |
| P217 | ugkv-cvlme     | Cross-validated log model evidence for the univariate Gaussian with known variance                  | JoramSoch | 2021-03-24 | 445 |
| P218 | ugkv-cvlbf     | Cross-validated log Bayes factor for the univariate Gaussian with known variance                    | JoramSoch | 2021-03-24 | 447 |
| P219 | ugkv-cvlbfmean | Expectation of the cross-validated log Bayes factor for the univariate Gaussian with known variance | JoramSoch | 2021-03-24 | 448 |
| P220 | cdf-pit        | Probability integral transform using cumulative distribution function                               | JoramSoch | 2021-04-07 | 41  |
| P221 | cdf-itm        | Inverse transformation method using cumulative distribution function                                | JoramSoch | 2021-04-07 | 41  |
| P222 | cdf-dt         | Distributional transformation using cumulative distribution function                                | JoramSoch | 2021-04-07 | 42  |
| P223 | ug-mle         | Maximum likelihood estimation for the univariate Gaussian   | JoramSoch | 2021-04-16 | 406 |
| P224 | poissexp-mle   | Maximum likelihood estimation for the Poisson distribution with exposure values                     | JoramSoch | 2021-04-16 | 663 |
| P225 | poiss-prior    | Conjugate prior distribution for Poisson-distributed data   | JoramSoch | 2020-04-21 | 659 |
| P226 | poiss-post     | Posterior distribution for Poisson-distributed data   | JoramSoch | 2020-04-21 | 660 |
| P227 | poiss-lme      | Log model evidence for Poisson-distributed data   | JoramSoch | 2020-04-21 | 661 |
| P228 | beta-mean      | Mean of the beta distribution   | JoramSoch | 2021-04-29 | 308 |
| P229 | beta-var       | Variance of the beta distribution   | JoramSoch | 2021-04-29 | 313 |
| P230 | poiss-var      | Variance of the Poisson distribution  | JoramSoch | 2021-04-29 | 197 |
| P231 | mvt-f          | Relationship between multivariate t-distribution and F-distribution                                 | JoramSoch | 2021-05-04 | 365 |
| P232 | nst-t          | Relationship between non-standardized t-distribution and t-distribution                             | JoramSoch | 2021-05-11 | 253 |



|      |              |   |           |            |     |
|------|--------------|---|-----------|------------|-----|
| P233 | norm-chi2    | Relationship between normal distribution and chi-squared distribution                       | JoramSoch | 2021-05-20 | 223 |
| P234 | norm-t       | Relationship between normal distribution and t-distribution                                 | JoramSoch | 2021-05-27 | 225 |
| P235 | norm-lincomb | Linear combination of independent normal random variables                                   | JoramSoch | 2021-06-02 | 247 |
| P236 | mvn-ind      | Necessary and sufficient condition for independence of multivariate normal random variables | JoramSoch | 2021-06-02 | 352 |
| P237 | ng-mean      | Mean of the normal-gamma distribution   | JoramSoch | 2021-07-08 | 369 |
| P238 | ng-dent      | Differential entropy of the normal-gamma distribution                                       | JoramSoch | 2021-07-08 | 372 |
| P239 | gam-dent     | Differential entropy of the gamma distribution  | JoramSoch | 2021-07-14 | 271 |
| P240 | ug-anc       | Accuracy and complexity for the univariate Gaussian   | JoramSoch | 2021-07-14 | 423 |
| P241 | prob-ind     | Probability under statistical independence  | JoramSoch | 2021-07-23 | 10  |
| P242 | prob-exc     | Probability under mutual exclusivity  | JoramSoch | 2021-07-23 | 11  |
| P243 | prob-mon     | Monotonicity of probability   | JoramSoch | 2021-07-30 | 12  |
| P244 | prob-emp     | Probability of the empty set  | JoramSoch | 2021-07-30 | 14  |
| P245 | prob-comp    | Probability of the complement   | JoramSoch | 2021-07-30 | 15  |
| P246 | prob-range   | Range of probability  | JoramSoch | 2021-07-30 | 15  |
| P247 | prob-add     | Addition law of probability   | JoramSoch | 2021-07-30 | 16  |
| P248 | prob-tot     | Law of total probability  | JoramSoch | 2021-08-08 | 19  |
| P249 | prob-exh     | Probability of exhaustive events  | JoramSoch | 2021-08-08 | 19  |
| P250 | norm-maxent  | Normal distribution maximizes differential entropy for fixed variance                       | JoramSoch | 2020-08-25 | 246 |
| P251 | norm-extr    | Extreme points of the probability density function of the normal distribution               | JoramSoch | 2020-08-25 | 241 |
| P252 | norm-infl    | Inflection points of the probability density function of the normal distribution            | JoramSoch | 2020-08-26 | 242 |

|      |             |  |           |            |     |
|------|-------------|--|-----------|------------|-----|
| P253 | pmf-invft   | Probability mass function of an invertible function of a random vector               | JoramSoch | 2021-08-30 | 28  |
| P254 | pdf-invft   | Probability density function of an invertible function of a continuous random vector | JoramSoch | 2021-08-30 | 32  |
| P255 | pdf-linft   | Probability density function of a linear function of a continuous random vector      | JoramSoch | 2021-08-30 | 34  |
| P256 | cdf-sumind  | Cumulative distribution function of a sum of independent random variables            | JoramSoch | 2021-08-30 | 36  |
| P257 | pmf-sumind  | Probability mass function of a sum of independent discrete random variables          | JoramSoch | 2021-08-30 | 25  |
| P258 | pdf-sumind  | Probability density function of a sum of independent continuous random variables     | JoramSoch | 2021-08-30 | 29  |
| P259 | cf-fct      | Characteristic function of a function of a random variable                           | JoramSoch | 2021-09-22 | 49  |
| P260 | mgf-fct     | Moment-generating function of a function of a random variable                        | JoramSoch | 2021-09-22 | 44  |
| P261 | dent-addvec | Addition of the differential entropy upon multiplication with invertible matrix      | JoramSoch | 2021-10-07 | 113 |
| P262 | dent-noninv | Non-invariance of the differential entropy under change of variables                 | JoramSoch | 2021-10-07 | 114 |
| P263 | t-pdf       | Probability density function of the t-distribution                                   | JoramSoch | 2021-10-12 | 254 |
| P264 | f-pdf       | Probability density function of the F-distribution                                   | JoramSoch | 2021-10-12 | 298 |
| P265 | tglm-dist   | Distribution of the transformed general linear model                                 | JoramSoch | 2021-10-21 | 613 |
| P266 | tglm-para   | Equivalence of parameter estimates from the transformed general linear model         | JoramSoch | 2021-10-21 | 614 |
| P267 | iglm-dist   | Distribution of the inverse general linear model                                     | JoramSoch | 2021-10-21 | 616 |

|      |              |   |           |            |     |
|------|--------------|---|-----------|------------|-----|
| P268 | iglm-blue    | Best linear unbiased estimator for the inverse general linear model                                       | JoramSoch | 2021-10-21 | 616 |
| P269 | cfm-para     | Parameters of the corresponding forward model   | JoramSoch | 2021-10-21 | 621 |
| P270 | cfm-exist    | Existence of a corresponding forward model  | JoramSoch | 2021-10-21 | 622 |
| P271 | slr-ols      | Ordinary least squares for simple linear regression   | JoramSoch | 2021-10-27 | 483 |
| P272 | slr-olsmean  | Expectation of parameter estimates for simple linear regression   | JoramSoch | 2021-10-27 | 487 |
| P273 | slr-olsvar   | Variance of parameter estimates for simple linear regression  | JoramSoch | 2021-10-27 | 489 |
| P274 | slr-meancent | Effects of mean-centering on parameter estimates for simple linear regression                             | JoramSoch | 2021-10-27 | 494 |
| P275 | slr-comp     | The regression line goes through the center of mass point   | JoramSoch | 2021-10-27 | 496 |
| P276 | slr-ressum   | The sum of residuals is zero in simple linear regression  | JoramSoch | 2021-10-27 | 517 |
| P277 | slr-rescorr  | The residuals and the covariate are uncorrelated in simple linear regression                              | JoramSoch | 2021-10-27 | 518 |
| P278 | slr-resvar   | Relationship between residual variance and sample variance in simple linear regression                    | JoramSoch | 2021-10-27 | 519 |
| P279 | slr-corr     | Relationship between correlation coefficient and slope estimate in simple linear regression               | JoramSoch | 2021-10-27 | 521 |
| P280 | slr-rsq      | Relationship between coefficient of determination and correlation coefficient in simple linear regression | JoramSoch | 2021-10-27 | 521 |
| P281 | slr-mlr      | Simple linear regression is a special case of multiple linear regression                                  | JoramSoch | 2021-11-09 | 482 |
| P282 | slr-olsdist  | Distribution of parameter estimates for simple linear regression  | JoramSoch | 2021-11-09 | 491 |
| P283 | slr-proj     | Projection of a data point to the regression line   | JoramSoch | 2021-11-09 | 497 |

|      |            |  |           |            |     |
|------|------------|--|-----------|------------|-----|
| P284 | slr-sss    | Sums of squares for simple linear regression                   | JoramSoch | 2021-11-09 | 498 |
| P285 | slr-mat    | Transformation matrices for simple linear regression           | JoramSoch | 2021-11-09 | 503 |
| P286 | slr-wls    | Weighted least squares for simple linear regression            | JoramSoch | 2021-11-16 | 505 |
| P287 | slr-mle    | Maximum likelihood estimation for simple linear regression     | JoramSoch | 2021-11-16 | 508 |
| P288 | slr-ols2   | Ordinary least squares for simple linear regression            | JoramSoch | 2021-11-16 | 485 |
| P289 | slr-wls2   | Weighted least squares for simple linear regression            | JoramSoch | 2021-11-16 | 507 |
| P290 | slr-mle2   | Maximum likelihood estimation for simple linear regression     | JoramSoch | 2021-11-16 | 510 |
| P291 | mean-tot   | Law of total expectation                                       | JoramSoch | 2021-11-26 | 57  |
| P292 | var-tot    | Law of total variance  | JoramSoch | 2021-11-26 | 75  |
| P293 | cov-tot    | Law of total covariance  | JoramSoch | 2021-11-26 | 80  |
| P294 | dir-kl     | Kullback-Leibler divergence for the Dirichlet distribution     | JoramSoch | 2021-12-02 | 381 |
| P295 | wish-kl    | Kullback-Leibler divergence for the Wishart distribution       | JoramSoch | 2021-12-02 | 399 |
| P296 | matn-kl    | Kullback-Leibler divergence for the matrix-normal distribution | JoramSoch | 2021-12-02 | 390 |
| P297 | matn-samp  | Sampling from the matrix-normal distribution                   | JoramSoch | 2021-12-07 | 397 |
| P298 | mean-tr    | Expected value of the trace of a matrix                        | JoramSoch | 2021-12-07 | 68  |
| P299 | corr-z     | Correlation coefficient in terms of standard scores            | JoramSoch | 2021-12-14 | 92  |
| P300 | corr-range | Correlation always falls between -1 and +1                     | JoramSoch | 2021-12-14 | 91  |
| P301 | bern-var   | Variance of the Bernoulli distribution                         | JoramSoch | 2022-01-20 | 179 |
| P302 | bin-var    | Variance of the binomial distribution                          | JoramSoch | 2022-01-20 | 186 |

|      |               |   |           |            |     |
|------|---------------|---|-----------|------------|-----|
| P303 | bern-varrange | Range of the variance of the Bernoulli distribution   | JoramSoch | 2022-01-27 | 180 |
| P304 | bin-varrange  | Range of the variance of the binomial distribution  | JoramSoch | 2022-01-27 | 187 |
| P305 | mlr-mll       | Maximum log-likelihood for multiple linear regression   | JoramSoch | 2022-02-04 | 545 |
| P306 | lognorm-med   | Median of the log-normal distribution   | majapavlo | 2022-02-07 | 288 |
| P307 | mlr-aic       | Akaike information criterion for multiple linear regression   | JoramSoch | 2022-02-11 | 563 |
| P308 | mlr-bic       | Bayesian information criterion for multiple linear regression   | JoramSoch | 2022-02-11 | 563 |
| P309 | mlr-aicc      | Corrected Akaike information criterion for multiple linear regression   | JoramSoch | 2022-02-11 | 564 |
| P310 | lognorm-pdf   | Probability density function of the log-normal distribution   | majapavlo | 2022-02-13 | 283 |
| P311 | lognorm-mode  | Mode of the log-normal distribution   | majapavlo | 2022-02-13 | 288 |
| P312 | mlr-dev       | Deviance for multiple linear regression   | JoramSoch | 2022-03-01 | 561 |
| P313 | blr-dic       | Deviance information criterion for multiple linear regression   | JoramSoch | 2022-03-01 | 575 |
| P314 | lme-pnp       | Log model evidence in terms of prior and posterior distribution   | JoramSoch | 2022-03-11 | 712 |
| P315 | aicc-mll      | Corrected Akaike information criterion in terms of maximum log-likelihood   | JoramSoch | 2022-03-11 | 707 |
| P316 | aicc-aic      | Corrected Akaike information criterion converges to uncorrected Akaike information criterion when infinite data are available | JoramSoch | 2022-03-18 | 706 |
| P317 | mle-bias      | Maximum likelihood estimation can result in biased estimates  | JoramSoch | 2022-03-18 | 137 |
| P318 | pval-h0       | The p-value follows a uniform distribution under the null hypothesis  | JoramSoch | 2022-03-18 | 144 |
| P319 | prob-exh2     | Probability of exhaustive events  | JoramSoch | 2022-03-27 | 20  |

|      |              |  |           |            |     |
|------|--------------|--|-----------|------------|-----|
| P320 | slr-olscorr  | Parameter estimates for simple linear regression are uncorrelated after mean-centering | JoramSoch | 2022-04-14 | 494 |
| P321 | norm-probstd | Probability of normal random variable being within standard deviations from its mean   | JoramSoch | 2022-05-08 | 234 |
| P322 | mult-cov     | Covariance matrix of the multinomial distribution                                      | adkipnis  | 2022-05-11 | 206 |
| P323 | nw-pdf       | Probability density function of the normal-Wishart distribution                        | JoramSoch | 2022-05-14 | 401 |
| P324 | ng-nw        | Normal-gamma distribution is a special case of normal-Wishart distribution             | JoramSoch | 2022-05-20 | 367 |
| P325 | lognorm-cdf  | Cumulative distribution function of the log-normal distribution                        | majapavlo | 2022-06-29 | 284 |
| P326 | lognorm-qf   | Quantile function of the log-normal distribution                                       | majapavlo | 2022-07-09 | 285 |
| P327 | nw-mean      | Mean of the normal-Wishart distribution  | JoramSoch | 2022-07-14 | 402 |
| P328 | gam-wish     | Gamma distribution is a special case of Wishart distribution                           | JoramSoch | 2022-07-14 | 257 |
| P329 | mlr-glm      | Multiple linear regression is a special case of the general linear model               | JoramSoch | 2022-07-21 | 524 |
| P330 | mvn-matn     | Multivariate normal distribution is a special case of matrix-normal distribution       | JoramSoch | 2022-07-31 | 334 |
| P331 | norm-mvn     | Normal distribution is a special case of multivariate normal distribution              | JoramSoch | 2022-08-19 | 220 |
| P332 | t-mvt        | t-distribution is a special case of multivariate t-distribution                        | JoramSoch | 2022-08-25 | 252 |
| P333 | mvt-pdf      | Probability density function of the multivariate t-distribution                        | JoramSoch | 2022-09-02 | 365 |
| P334 | bern-ent     | Entropy of the Bernoulli distribution  | JoramSoch | 2022-09-02 | 181 |
| P335 | bin-ent      | Entropy of the binomial distribution   | JoramSoch | 2022-09-02 | 188 |
| P336 | cat-ent      | Entropy of the categorical distribution  | JoramSoch | 2022-09-09 | 202 |

|      |              |   |           |            |     |
|------|--------------|---|-----------|------------|-----|
| P337 | mult-ent     | Entropy of the multinomial distribution                                   | JoramSoch | 2022-09-09 | 207 |
| P338 | cat-cov      | Covariance matrix of the categorical distribution                         | JoramSoch | 2022-09-09 | 202 |
| P339 | mvn-mean     | Mean of the multivariate normal distribution                              | JoramSoch | 2022-09-15 | 337 |
| P340 | mvn-cov      | Covariance matrix of the multivariate normal distribution                 | JoramSoch | 2022-09-15 | 340 |
| P341 | matn-mean    | Mean of the matrix-normal distribution                                    | JoramSoch | 2022-09-15 | 388 |
| P342 | matn-cov     | Covariance matrices of the matrix-normal distribution                     | JoramSoch | 2022-09-15 | 388 |
| P343 | matn-marg    | Marginal distributions for the matrix-normal distribution                 | JoramSoch | 2022-09-15 | 394 |
| P344 | matn-dent    | Differential entropy for the matrix-normal distribution                   | JoramSoch | 2022-09-22 | 389 |
| P345 | ng-cov       | Covariance and variance of the normal-gamma distribution                  | JoramSoch | 2022-09-22 | 371 |
| P346 | ng-samp      | Sampling from the normal-gamma distribution                               | JoramSoch | 2022-09-22 | 379 |
| P347 | covmat-inv   | Invariance of the covariance matrix under addition of constant vector     | JoramSoch | 2022-09-22 | 85  |
| P348 | covmat-scal  | Scaling of the covariance matrix upon multiplication with constant matrix | JoramSoch | 2022-09-22 | 85  |
| P349 | covmat-sum   | Covariance matrix of the sum of two random vectors                        | JoramSoch | 2022-09-26 | 88  |
| P350 | covmat-symm  | Symmetry of the covariance matrix   | JoramSoch | 2022-09-26 | 83  |
| P351 | covmat-psd   | Positive semi-definiteness of the covariance matrix                       | JoramSoch | 2022-09-26 | 84  |
| P352 | cov-var      | Self-covariance equals variance   | JoramSoch | 2022-09-26 | 79  |
| P353 | cov-symm     | Symmetry of the covariance  | JoramSoch | 2022-09-26 | 79  |
| P354 | lognorm-mean | Mean of the log-normal distribution                                       | majapavlo | 2022-10-02 | 286 |
| P355 | lognorm-var  | Variance of the log-normal distribution                                   | majapavlo | 2022-10-02 | 290 |

|      |                  |   |           |            |     |
|------|------------------|---|-----------|------------|-----|
| P356 | beta-chi2        | Relationship between chi-squared distribution and beta distribution   | JoramSoch | 2022-10-07 | 301 |
| P357 | betabin-mome     | Method of moments for beta-binomial data  | JoramSoch | 2022-10-07 | 676 |
| P358 | bin-margcond     | Marginal distribution of a conditional binomial distribution  | JoramSoch | 2022-10-07 | 190 |
| P359 | mean-prodsqr     | Square of expectation of product is less than or equal to product of expectation of squares                 | JoramSoch | 2022-10-11 | 59  |
| P360 | pgf-mean         | Probability-generating function is expectation of function of random variable                               | JoramSoch | 2022-10-11 | 46  |
| P361 | pgf-zero         | Value of the probability-generating function for argument zero  | JoramSoch | 2022-10-11 | 47  |
| P362 | pgf-one          | Value of the probability-generating function for argument one   | JoramSoch | 2022-10-11 | 47  |
| P363 | bin-pgf          | Probability-generating function of the binomial distribution  | JoramSoch | 2022-10-11 | 185 |
| P364 | betabin-pmf      | Probability mass function of the beta-binomial distribution   | JoramSoch | 2022-10-20 | 192 |
| P365 | betabin-pmfitogf | Expression of the probability mass function of the beta-binomial distribution using only the gamma function | JoramSoch | 2022-10-20 | 194 |
| P366 | betabin-cdf      | Cumulative distribution function of the beta-binomial distribution  | JoramSoch | 2022-10-22 | 194 |
| P367 | me-der           | Derivation of the model evidence  | JoramSoch | 2022-10-20 | 711 |
| P368 | fe-der           | Derivation of the family evidence   | JoramSoch | 2022-10-20 | 718 |
| P369 | anova1-ols       | Ordinary least squares for one-way analysis of variance   | JoramSoch | 2022-11-06 | 451 |
| P370 | anova1-f         | F-test for main effect in one-way analysis of variance  | JoramSoch | 2022-11-06 | 453 |
| P371 | anova2-ols       | Ordinary least squares for two-way analysis of variance   | JoramSoch | 2022-11-06 | 462 |
| P372 | anova2-fme       | F-test for main effect in two-way analysis of variance  | JoramSoch | 2022-11-10 | 474 |



|      |                |   |           |            |     |
|------|----------------|---|-----------|------------|-----|
| P373 | anova2-fia     | F-test for interaction in two-way analysis of variance                                    | JoramSoch | 2022-11-11 | 477 |
| P374 | anova2-fgm     | F-test for grand mean in two-way analysis of variance                                     | JoramSoch | 2022-11-11 | 478 |
| P375 | anova1-repara  | Reparametrization for one-way analysis of variance  | JoramSoch | 2022-11-15 | 458 |
| P376 | anova1-pss     | Partition of sums of squares in one-way analysis of variance                              | JoramSoch | 2022-11-15 | 452 |
| P377 | anova1-fols    | F-statistic in terms of ordinary least squares estimates in one-way analysis of variance  | JoramSoch | 2022-11-15 | 457 |
| P378 | anova2-cochran | Application of Cochran's theorem to two-way analysis of variance                          | JoramSoch | 2022-11-16 | 468 |
| P379 | anova2-pss     | Partition of sums of squares in two-way analysis of variance                              | JoramSoch | 2022-11-16 | 466 |
| P380 | anova2-fols    | F-statistics in terms of ordinary least squares estimates in two-way analysis of variance | JoramSoch | 2022-11-16 | 480 |
| P381 | bin-mle        | Maximum likelihood estimation for binomial observations                                   | JoramSoch | 2022-11-23 | 631 |
| P382 | bin-mll        | Maximum log-likelihood for binomial observations  | JoramSoch | 2022-11-24 | 632 |
| P383 | bin-lbf        | Log Bayes factor for binomial observations  | JoramSoch | 2022-11-25 | 637 |
| P384 | bin-pp         | Posterior probability of the alternative model for binomial observations                  | JoramSoch | 2022-11-26 | 639 |
| P385 | mult-mle       | Maximum likelihood estimation for multinomial observations                                | JoramSoch | 2022-12-02 | 644 |
| P386 | mult-mll       | Maximum log-likelihood for multinomial observations                                       | JoramSoch | 2022-12-02 | 644 |
| P387 | mult-lbf       | Log Bayes factor for multinomial observations   | JoramSoch | 2022-12-02 | 651 |
| P388 | mult-pp        | Posterior probability of the alternative model for multinomial observations               | JoramSoch | 2022-12-02 | 652 |

|      |              |  |                 |            |     |
|------|--------------|--|-----------------|------------|-----|
| P389 | mlr-wlsdist  | Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon weighted least squares | JoramSoch       | 2022-12-13 | 538 |
| P390 | mlr-rssdist  | Distribution of residual sum of squares in multiple linear regression with weighted least squares                            | JoramSoch       | 2022-12-13 | 540 |
| P391 | mlr-t        | t-test for multiple linear regression using contrast-based inference   | JoramSoch       | 2022-12-13 | 549 |
| P392 | mlr-f        | F-test for multiple linear regression using contrast-based inference   | JoramSoch       | 2022-12-13 | 552 |
| P393 | mlr-ind      | Independence of estimated parameters and residuals in multiple linear regression   | JoramSoch       | 2022-12-13 | 535 |
| P394 | mvn-indprod  | Independence of products of multivariate normal random vector  | JoramSoch       | 2022-12-13 | 354 |
| P395 | mvn-chi2     | Relationship between multivariate normal distribution and chi-squared distribution   | JoramSoch       | 2022-12-20 | 335 |
| P396 | cuni-var     | Variance of the continuous uniform distribution  | JoramSoch       | 2022-12-20 | 215 |
| P397 | cuni-dent    | Differential entropy of the continuous uniform distribution  | JoramSoch       | 2022-12-20 | 216 |
| P398 | resvar-biasp | Maximum likelihood estimator of variance in multiple linear regression is biased   | JoramSoch       | 2022-12-21 | 686 |
| P399 | mlr-symm     | Projection matrix and residual-forming matrix are symmetric  | JoramSoch       | 2022-12-22 | 534 |
| P400 | mlr-olsdist  | Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon ordinary least squares | JoramSoch       | 2022-12-23 | 537 |
| P401 | exp-var      | Variance of the exponential distribution   | majapavlo       | 2023-01-23 | 279 |
| P402 | exg-pdf      | Probability density function of the ex-Gaussian distribution   | tomfaulkenberry | 2023-04-18 | 324 |
| P403 | exp-mgf      | Moment-generating function of the exponential distribution   | tomfaulkenberry | 2023-04-19 | 274 |

|      |             |  |                 |            |     |
|------|-------------|--|-----------------|------------|-----|
| P404 | exg-mgf     | Moment-generating function of the ex-Gaussian distribution   | tomfaulkenberry | 2023-04-19 | 326 |
| P405 | exg-mean    | Mean of the ex-Gaussian distribution   | tomfaulkenberry | 2023-04-19 | 327 |
| P406 | exg-var     | Variance of the ex-Gaussian distribution   | tomfaulkenberry | 2023-04-19 | 328 |
| P407 | skew-mean   | Partition of skewness into expected values   | tomfaulkenberry | 2023-04-20 | 77  |
| P408 | exg-skew    | Skewness of the ex-Gaussian distribution   | tomfaulkenberry | 2023-04-21 | 329 |
| P409 | exp-skew    | Skewness of the exponential distribution   | tomfaulkenberry | 2023-04-24 | 280 |
| P410 | duni-ent    | Entropy of the discrete uniform distribution   | JoramSoch       | 2023-08-11 | 174 |
| P411 | duni-maxent | Discrete uniform distribution maximizes entropy for finite support   | JoramSoch       | 2023-08-18 | 176 |
| P412 | cuni-maxent | Continuous uniform distribution maximizes differential entropy for fixed range   | JoramSoch       | 2023-08-25 | 218 |
| P413 | post-ind    | Combined posterior distributions in terms of individual posterior distributions obtained from conditionally independent data | JoramSoch       | 2023-09-01 | 150 |
| P414 | lme-mean    | Equivalence of operations for model evidence and log model evidence  | JoramSoch       | 2023-09-08 | 714 |
| P415 | lfe-approx  | Approximation of log family evidences based on log model evidences   | JoramSoch       | 2023-09-15 | 721 |
| P416 | bvn-pdf     | Probability density function of the bivariate normal distribution  | JoramSoch       | 2023-09-22 | 356 |
| P417 | bvn-pdfcorr | Probability density function of the bivariate normal distribution in terms of correlation coefficient                        | JoramSoch       | 2023-09-29 | 357 |
| P418 | mlr-olstr   | Ordinary least squares for multiple linear regression with two regressors  | JoramSoch       | 2023-10-06 | 527 |
| P419 | bern-kl     | Kullback-Leibler divergence for the Bernoulli distribution   | JoramSoch       | 2023-10-13 | 181 |
| P420 | bin-kl      | Kullback-Leibler divergence for the binomial distribution  | JoramSoch       | 2023-10-20 | 189 |

|      |             |   |                 |            |     |
|------|-------------|---|-----------------|------------|-----|
| P421 | wald-skew   | Skewness of the Wald distribution   | tomfaulkenberry | 2023-10-24 | 319 |
| P422 | cuni-kl     | Kullback-Leibler divergence for the continuous uniform distribution               | JoramSoch       | 2023-10-27 | 217 |
| P423 | wald-mome   | Method of moments for Wald-distributed data                                       | tomfaulkenberry | 2023-10-30 | 322 |
| P424 | exg-mome    | Method of moments for ex-Gaussian-distributed data                                | tomfaulkenberry | 2023-10-30 | 331 |
| P425 | duni-kl     | Kullback-Leibler divergence for the discrete uniform distribution                 | JoramSoch       | 2023-11-17 | 175 |
| P426 | gam-scal    | Scaling of a random variable following the gamma distribution                     | JoramSoch       | 2023-11-24 | 259 |
| P427 | bin-map     | Maximum-a-posteriori estimation for binomial observations                         | JoramSoch       | 2023-12-01 | 633 |
| P428 | mult-map    | Maximum-a-posteriori estimation for multinomial observations                      | JoramSoch       | 2023-12-08 | 646 |
| P429 | bin-test    | Binomial test   | JoramSoch       | 2023-12-16 | 630 |
| P430 | mult-test   | Multinomial test  | JoramSoch       | 2023-12-23 | 642 |
| P431 | blr-anc     | Accuracy and complexity for Bayesian linear regression                            | JoramSoch       | 2024-01-12 | 571 |
| P432 | blrkc-prior | Conjugate prior distribution for Bayesian linear regression with known covariance | JoramSoch       | 2024-01-19 | 589 |
| P433 | blrkc-post  | Posterior distribution for Bayesian linear regression with known covariance       | JoramSoch       | 2024-01-19 | 590 |
| P434 | blrkc-lme   | Log model evidence for Bayesian linear regression with known covariance           | JoramSoch       | 2024-01-19 | 592 |
| P435 | blrkc-anc   | Accuracy and complexity for Bayesian linear regression with known covariance      | JoramSoch       | 2024-01-19 | 595 |
| P436 | mvn-mgf     | Moment-generating function of the multivariate normal distribution                | JoramSoch       | 2024-02-16 | 336 |
| P437 | gam-mgf     | Moment-generating function of the gamma distribution                              | JoramSoch       | 2024-02-23 | 261 |
| P438 | lpsr-spsr   | The log probability scoring rule is a strictly proper scoring rule                | KarahanS        | 2024-02-28 | 163 |

|      |             |   |           |            |     |
|------|-------------|---|-----------|------------|-----|
| P439 | resvar-unbp | Construction of unbiased estimator for variance in multiple linear regression                                     | JoramSoch | 2024-03-08 | 689 |
| P440 | rsq-resvar  | Expression of $R^2$ in terms of residual variances  | JoramSoch | 2024-03-08 | 691 |
| P441 | rsq-test    | Statistical significance test for the coefficient of determination based on an omnibus F-test                     | JoramSoch | 2024-03-08 | 694 |
| P442 | fstat-rsq   | Relationship between F-statistic and $R^2$  | JoramSoch | 2024-03-15 | 699 |
| P443 | fstat-mll   | Relationship between F-statistic and maximum log-likelihood   | JoramSoch | 2024-03-28 | 701 |
| P444 | snr-mll     | Relationship between signal-to-noise ratio and maximum log-likelihood   | JoramSoch | 2024-03-28 | 705 |
| P445 | bsr-spsr    | Brier scoring rule is strictly proper scoring rule  | KarahanS  | 2024-03-28 | 167 |
| P446 | blr-posterr | Expression of the noise precision posterior for Bayesian linear regression using prediction and parameter errors  | JoramSoch | 2024-04-05 | 579 |
| P447 | blr-postind | Combined posterior distribution for Bayesian linear regression when analyzing conditionally independent data sets | JoramSoch | 2024-04-12 | 583 |
| P448 | blr-map     | Maximum-a-posteriori estimation for Bayesian linear regression  | JoramSoch | 2024-04-19 | 577 |
| P449 | blr-lbf     | Log Bayes factor for Bayesian linear regression   | JoramSoch | 2024-04-26 | 587 |
| P450 | mlr-tsingle | Specific t-test for single regressor in multiple linear regression  | JoramSoch | 2024-05-03 | 554 |
| P451 | slr-tint    | Statistical test for intercept parameter in simple linear regression model  | JoramSoch | 2024-05-10 | 511 |
| P452 | slr-tslo    | Statistical test for slope parameter in simple linear regression model  | JoramSoch | 2024-05-17 | 513 |
| P453 | slr-fcomp   | Statistical test for comparing simple linear regression models with and without slope parameter                   | JoramSoch | 2024-05-24 | 515 |

|      |              |   |            |            |     |
|------|--------------|---|------------|------------|-----|
| P454 | mlr-fomnibus | Omnibus F-test for multiple regressors in multiple linear regression                      | JoramSoch  | 2024-05-31 | 556 |
| P455 | glm-llr      | Log-likelihood ratio for the general linear model   | JoramSoch  | 2024-06-07 | 608 |
| P456 | glm-mll      | Maximum log-likelihood for the general linear model                                       | JoramSoch  | 2024-06-14 | 607 |
| P457 | glm-mi       | Mutual information of dependent and independent variables in the general linear model     | JoramSoch  | 2024-06-21 | 610 |
| P458 | glm-llrmi    | Equivalence of log-likelihood ratio and mutual information for the general linear model   | JoramSoch  | 2024-06-21 | 612 |
| P459 | iglm-llrs    | Equivalence of log-likelihood ratios for regular and inverse general linear model         | JoramSoch  | 2024-06-28 | 618 |
| P460 | ug-fev       | F-test for equality of variances in two independent samples                               | JoramSoch  | 2024-07-05 | 412 |
| P461 | slr-pss      | Partition of sums of squares for simple linear regression                                 | JoramSoch  | 2024-07-12 | 500 |
| P462 | mlr-ols3     | Ordinary least squares for multiple linear regression                                     | JoramSoch  | 2024-07-18 | 526 |
| P463 | mlr-llr      | Log-likelihood ratio for multiple linear regression                                       | JoramSoch  | 2024-07-25 | 547 |
| P464 | prob-emp2    | Probability of the empty set  | JoramSoch  | 2024-08-08 | 14  |
| P465 | prob-mon2    | Monotonicity of probability   | JoramSoch  | 2024-08-08 | 13  |
| P466 | cdf-probexc  | Exceedance probability for a random variable in terms of cumulative distribution function | JoramSoch  | 2024-09-06 | 40  |
| P467 | postpred-jl  | Posterior predictive distribution is a marginal distribution of the joint likelihood      | alocavodia | 2024-09-11 | 151 |
| P468 | mean-wlln    | Weak law of large numbers   | JoramSoch  | 2024-09-13 | 63  |
| P469 | mean-mse     | The expected value minimizes the mean squared error                                       | salbalkus  | 2024-09-13 | 65  |
| P470 | ind-self     | Self-independence of random event   | JoramSoch  | 2024-09-20 | 10  |
| P471 | med-mae      | The median minimizes the mean absolute error  | salbalkus  | 2024-09-23 | 95  |

|      |                 |   |           |            |     |
|------|-----------------|---|-----------|------------|-----|
| P472 | corr-ind        | Independent random variables are uncorrelated   | JoramSoch | 2024-09-27 | 92  |
| P473 | norm-corrind    | Normally distributed and uncorrelated does not imply independent                              | JoramSoch | 2024-10-04 | 249 |
| P474 | norm-margjoint  | Marginally normal does not imply jointly normal   | JoramSoch | 2024-10-11 | 251 |
| P475 | bvn-lincomb     | Linear combination of bivariate normal random variables                                       | JoramSoch | 2024-10-25 | 363 |
| P476 | bvn-mi          | Mutual information of the bivariate normal distribution                                       | JoramSoch | 2024-11-01 | 361 |
| P477 | mvn-mi          | Mutual information of the multivariate normal distribution                                    | JoramSoch | 2024-11-01 | 343 |
| P478 | mgf-sumind      | Moment-generating function of a sum of independent random variables                           | JoramSoch | 2024-11-08 | 43  |
| P479 | lognorm-prodind | The product of independent log-normal random variables is a log-normal random variable        | JoramSoch | 2024-11-15 | 292 |
| P480 | lognorm-geomind | The geometric mean of independent log-normal random variables is a log-normal random variable | JoramSoch | 2024-11-22 | 293 |
| P481 | mark-ineq       | Markov's inequality   | JoramSoch | 2025-01-10 | 62  |
| P482 | cheb-ineq       | Chebyshev's inequality  | JoramSoch | 2025-01-10 | 62  |
| P483 | bool-ineq       | Boole's inequality  | JoramSoch | 2025-01-10 | 17  |
| P484 | bonf-ineq       | Bonferroni's inequality   | JoramSoch | 2025-01-10 | 17  |
| P485 | mult-marg       | Marginal distributions of the multinomial distribution  | JoramSoch | 2025-02-06 | 209 |
| P486 | poiss-ent       | Entropy of the Poisson distribution   | JoramSoch | 2025-02-06 | 199 |
| P487 | gam-mode        | Mode of the gamma distribution  | JoramSoch | 2025-02-06 | 265 |
| P488 | bin-cdf         | Cumulative distribution function of the binomial distribution                                 | JoramSoch | 2025-02-06 | 184 |
| P489 | mult-cdf        | Cumulative distribution function of the multinomial distribution                              | JoramSoch | 2025-02-06 | 204 |
| P490 | ug-cvlme        | Cross-validated log model evidence for the univariate Gaussian                                | JoramSoch | 2025-03-06 | 424 |

|      |            |  |           |            |     |
|------|------------|--|-----------|------------|-----|
| P491 | ug-cvlbf   | Cross-validated log Bayes factor for the univariate Gaussian                                     | JoramSoch | 2025-03-07 | 429 |
| P492 | bin-cvlme  | Cross-validated log model evidence for binomial observations                                     | JoramSoch | 2025-03-07 | 640 |
| P493 | mult-cvlme | Cross-validated log model evidence for multinomial observations                                  | JoramSoch | 2025-03-07 | 654 |
| P494 | bern-cat   | Bernoulli distribution is a special case of categorical distribution                             | JoramSoch | 2025-04-04 | 178 |
| P495 | bin-mult   | Binomial distribution is a special case of multinomial distribution                              | JoramSoch | 2025-04-04 | 183 |
| P496 | beta-dir   | Beta distribution is a special case of Dirichlet distribution                                    | JoramSoch | 2025-04-04 | 300 |
| P497 | beta-f     | Relationship between F-distribution and beta distribution  | JoramSoch | 2025-04-04 | 303 |
| P498 | mvn-mode   | Mode of the multivariate normal distribution   | JoramSoch | 2025-05-21 | 338 |
| P499 | mvn-meanqf | Expectation of a quadratic form for the multivariate normal distribution                         | JoramSoch | 2025-05-21 | 341 |
| P500 | mvn-maxent | Multivariate normal distribution maximizes differential entropy for fixed covariance             | JoramSoch | 2025-05-21 | 346 |
| P501 | mvn-samp   | Sampling from the multivariate normal distribution   | JoramSoch | 2025-05-21 | 355 |
| P502 | bvn-snorm  | Construction of a bivariate normal distribution from univariate standard normal random variables | JoramSoch | 2025-05-22 | 359 |
| P503 | bvn-mle    | Maximum likelihood estimation for bivariate normally distributed data                            | JoramSoch | 2025-06-20 | 598 |
| P504 | mvn-mle    | Maximum likelihood estimation for multivariate normally distributed data                         | JoramSoch | 2025-06-20 | 600 |
| P505 | matn-red   | Redundancy of parameters describing the matrix-normal distribution                               | JoramSoch | 2025-06-24 | 396 |
| P506 | matn-red2  | Redundancy of parameters describing the matrix-normal distribution                               | JoramSoch | 2025-06-24 | 397 |
| P507 | rsq-dist   | Distribution of the coefficient of determination under the null hypothesis                       | JoramSoch | 2025-07-04 | 696 |



|      |                  |  |                  |            |     |
|------|------------------|--|------------------|------------|-----|
| P508 | rsq-mmm          | Mean, mode and median of the coefficient of determination under the null hypothesis                | JoramSoch        | 2025-07-04 | 697 |
| P509 | rsq-var          | Variance of the coefficient of determination under the null hypothesis                             | JoramSoch        | 2025-07-04 | 698 |
| P510 | priorpred-jl     | Prior predictive distribution is a marginal distribution of the joint likelihood                   | JoramSoch        | 2025-07-04 | 148 |
| P511 | ug-ttest1power   | Power analysis, minimum detectable effect and minimum required sample size for a one-sample t-test | alexanderdbolton | 2025-07-23 | 413 |
| P512 | covmatcross-mean | Partition of a cross-covariance matrix into expected values  | JoramSoch        | 2025-08-29 | 88  |
| P513 | mean-blf         | Expectation of a bilinear form   | JoramSoch        | 2025-08-29 | 67  |
| P514 | jens-ineq        | Jensen's inequality  | JoramSoch        | 2025-09-25 | 61  |
| P515 | kl-nonneg3       | Non-negativity of the Kullback-Leibler divergence  | JoramSoch        | 2025-09-25 | 126 |
| P516 | fren-dec         | Decompositions of the variational free energy  | JoramSoch        | 2025-09-25 | 159 |
| P517 | fren-lme         | The variational free energy is a lower bound on the log model evidence                             | JoramSoch        | 2025-09-25 | 161 |
| P518 | gam-med          | Median of the gamma distribution   | JoramSoch        | 2025-10-24 | 264 |
| P519 | beta-med         | Median of the beta distribution  | JoramSoch        | 2025-10-24 | 309 |
| P520 | beta-mode        | Mode of the beta distribution  | JoramSoch        | 2025-10-24 | 310 |

## 2 Definition by Number

| ID  | Shortcut  | Definition                       | Author    | Date       | Page |
|-----|-----------|----------------------------------|-----------|------------|------|
| D1  | mvn       | Multivariate normal distribution | JoramSoch | 2020-01-22 | 334  |
| D2  | mgf       | Moment-generating function       | JoramSoch | 2020-01-22 | 43   |
| D3  | cuni      | Continuous uniform distribution  | JoramSoch | 2020-01-27 | 210  |
| D4  | norm      | Normal distribution              | JoramSoch | 2020-01-27 | 219  |
| D5  | ng        | Normal-gamma distribution        | JoramSoch | 2020-01-27 | 366  |
| D6  | matn      | Matrix-normal distribution       | JoramSoch | 2020-01-27 | 386  |
| D7  | gam       | Gamma distribution               | JoramSoch | 2020-02-08 | 256  |
| D8  | exp       | Exponential distribution         | JoramSoch | 2020-02-08 | 273  |
| D9  | pmf       | Probability mass function        | JoramSoch | 2020-02-13 | 25   |
| D10 | pdf       | Probability density function     | JoramSoch | 2020-02-13 | 28   |
| D11 | mean      | Expected value                   | JoramSoch | 2020-02-13 | 50   |
| D12 | var       | Variance                         | JoramSoch | 2020-02-13 | 69   |
| D13 | cdf       | Cumulative distribution function | JoramSoch | 2020-02-17 | 36   |
| D14 | qf        | Quantile function                | JoramSoch | 2020-02-17 | 48   |
| D15 | ent       | Shannon entropy                  | JoramSoch | 2020-02-19 | 104  |
| D16 | dent      | Differential entropy             | JoramSoch | 2020-02-19 | 110  |
| D17 | ent-cond  | Conditional entropy              | JoramSoch | 2020-02-19 | 106  |
| D18 | ent-joint | Joint entropy                    | JoramSoch | 2020-02-19 | 106  |
| D19 | mi        | Mutual information               | JoramSoch | 2020-02-19 | 120  |
| D19 | mi        | Mutual information               | JoramSoch | 2020-02-19 | 120  |
| D20 | resvar    | Residual variance                | JoramSoch | 2020-02-25 | 684  |
| D21 | rsq       | Coefficient of determination     | JoramSoch | 2020-02-25 | 690  |
| D22 | snr       | Signal-to-noise ratio            | JoramSoch | 2020-02-25 | 703  |
| D23 | aic       | Akaike information criterion     | JoramSoch | 2020-02-25 | 706  |
| D24 | bic       | Bayesian information criterion   | JoramSoch | 2020-02-25 | 708  |
| D25 | dic       | Deviance information criterion   | JoramSoch | 2020-02-25 | 709  |
| D26 | lme       | Log model evidence               | JoramSoch | 2020-02-25 | 711  |
| D27 | gm        | Generative model                 | JoramSoch | 2020-03-03 | 147  |

|     |            |   |           |            |     |
|-----|------------|---|-----------|------------|-----|
| D28 | lf         | Likelihood function                       | JoramSoch | 2020-03-03 | 147 |
| D28 | lf         | Likelihood function                       | JoramSoch | 2020-03-03 | 147 |
| D29 | prior      | Prior distribution                        | JoramSoch | 2020-03-03 | 147 |
| D30 | fpm        | Full probability model                    | JoramSoch | 2020-03-03 | 148 |
| D31 | jl         | Joint likelihood                          | JoramSoch | 2020-03-03 | 148 |
| D32 | post       | Posterior distribution                    | JoramSoch | 2020-03-03 | 149 |
| D33 | ml         | Marginal likelihood                       | JoramSoch | 2020-03-03 | 152 |
| D34 | dent-cond  | Conditional differential entropy          | JoramSoch | 2020-03-21 | 116 |
| D35 | dent-joint | Joint differential entropy                | JoramSoch | 2020-03-21 | 116 |
| D36 | mlr        | Multiple linear regression                | JoramSoch | 2020-03-21 | 523 |
| D37 | tss        | Total sum of squares                      | JoramSoch | 2020-03-21 | 529 |
| D38 | ess        | Explained sum of squares                  | JoramSoch | 2020-03-21 | 529 |
| D39 | rss        | Residual sum of squares                   | JoramSoch | 2020-03-21 | 530 |
| D40 | glm        | General linear model                      | JoramSoch | 2020-03-21 | 603 |
| D41 | poiss-data | Poisson-distributed data                  | JoramSoch | 2020-03-22 | 657 |
| D42 | poissexp   | Poisson distribution with exposure values | JoramSoch | 2020-03-22 | 663 |
| D43 | wish       | Wishart distribution                      | JoramSoch | 2020-03-22 | 398 |
| D44 | bern       | Bernoulli distribution                    | JoramSoch | 2020-03-22 | 177 |
| D45 | bin        | Binomial distribution                     | JoramSoch | 2020-03-22 | 182 |
| D46 | cat        | Categorical distribution                  | JoramSoch | 2020-03-22 | 201 |
| D47 | mult       | Multinomial distribution                  | JoramSoch | 2020-03-22 | 203 |
| D48 | prob       | Probability                               | JoramSoch | 2020-05-10 | 6   |
| D49 | prob-joint | Joint probability                         | JoramSoch | 2020-05-10 | 7   |
| D50 | prob-marg  | Law of marginal probability               | JoramSoch | 2020-05-10 | 7   |
| D51 | prob-cond  | Law of conditional probability            | JoramSoch | 2020-05-10 | 7   |
| D52 | kl         | Kullback-Leibler divergence               | JoramSoch | 2020-05-10 | 124 |
| D53 | beta       | Beta distribution                         | JoramSoch | 2020-05-10 | 300 |
| D54 | dir        | Dirichlet distribution                    | JoramSoch | 2020-05-10 | 380 |
| D55 | dist       | Probability distribution                  | JoramSoch | 2020-05-17 | 21  |
| D56 | dist-joint | Joint probability distribution            | JoramSoch | 2020-05-17 | 21  |

|     |            |                                      |           |            |     |
|-----|------------|--------------------------------------|-----------|------------|-----|
| D57 | dist-marg  | Marginal probability distribution    | JoramSoch | 2020-05-17 | 22  |
| D58 | dist-cond  | Conditional probability distribution | JoramSoch | 2020-05-17 | 22  |
| D59 | llf        | Log-likelihood function              | JoramSoch | 2020-05-17 | 137 |
| D60 | mle        | Maximum likelihood estimation        | JoramSoch | 2020-05-15 | 137 |
| D61 | mll        | Maximum log-likelihood               | JoramSoch | 2020-05-15 | 137 |
| D62 | poiss      | Poisson distribution                 | JoramSoch | 2020-05-25 | 196 |
| D63 | snorm      | Standard normal distribution         | JoramSoch | 2020-05-26 | 220 |
| D64 | sgam       | Standard gamma distribution          | JoramSoch | 2020-05-26 | 257 |
| D65 | rvar       | Random variable                      | JoramSoch | 2020-05-27 | 4   |
| D66 | rvec       | Random vector                        | JoramSoch | 2020-05-27 | 5   |
| D67 | rmat       | Random matrix                        | JoramSoch | 2020-05-27 | 5   |
| D68 | cgf        | Cumulant-generating function         | JoramSoch | 2020-05-31 | 50  |
| D69 | pgf        | Probability-generating function      | JoramSoch | 2020-05-31 | 46  |
| D70 | cov        | Covariance                           | JoramSoch | 2020-06-02 | 78  |
| D71 | corr       | Correlation                          | JoramSoch | 2020-06-02 | 90  |
| D72 | covmat     | Covariance matrix                    | JoramSoch | 2020-06-06 | 82  |
| D73 | corrmat    | Correlation matrix                   | JoramSoch | 2020-06-06 | 93  |
| D74 | precmat    | Precision matrix                     | JoramSoch | 2020-06-06 | 89  |
| D75 | ind        | Statistical independence             | JoramSoch | 2020-06-06 | 8   |
| D76 | logreg     | Logistic regression                  | JoramSoch | 2020-06-28 | 679 |
| D77 | beta-data  | Beta-distributed data                | JoramSoch | 2020-06-28 | 671 |
| D78 | bin-data   | Binomial observations                | JoramSoch | 2020-07-07 | 630 |
| D79 | mult-data  | Multinomial observations             | JoramSoch | 2020-07-07 | 642 |
| D80 | lfe        | Log family evidence                  | JoramSoch | 2020-07-13 | 718 |
| D81 | emat       | Estimation matrix                    | JoramSoch | 2020-07-22 | 532 |
| D82 | pmat       | Projection matrix                    | JoramSoch | 2020-07-22 | 532 |
| D83 | rformat    | Residual-forming matrix              | JoramSoch | 2020-07-22 | 532 |
| D84 | lbf        | Log Bayes factor                     | JoramSoch | 2020-07-22 | 727 |
| D85 | ent-cross  | Cross-entropy                        | JoramSoch | 2020-07-28 | 107 |
| D86 | dent-cross | Differential cross-entropy           | JoramSoch | 2020-07-28 | 116 |

|      |           |   |                 |            |     |
|------|-----------|---|-----------------|------------|-----|
| D87  | pmp       | Posterior model probability                 | JoramSoch       | 2020-07-28 | 728 |
| D88  | duni      | Discrete uniform distribution               | JoramSoch       | 2020-07-28 | 172 |
| D89  | bma       | Bayesian model averaging                    | JoramSoch       | 2020-08-03 | 732 |
| D90  | mom       | Moment                                      | JoramSoch       | 2020-08-19 | 99  |
| D91  | fwhm      | Full width at half maximum                  | JoramSoch       | 2020-08-19 | 97  |
| D92  | bf        | Bayes factor                                | tomfaulkenberry | 2020-08-26 | 722 |
| D93  | encm      | Encompassing model                          | tomfaulkenberry | 2020-09-02 | 726 |
| D94  | std       | Standard deviation                          | JoramSoch       | 2020-09-03 | 96  |
| D95  | wald      | Wald distribution                           | tomfaulkenberry | 2020-09-04 | 314 |
| D96  | const     | Constant                                    | JoramSoch       | 2020-09-09 | 5   |
| D97  | mom-raw   | Raw moment                                  | JoramSoch       | 2020-10-08 | 100 |
| D98  | mom-cent  | Central moment                              | JoramSoch       | 2020-10-08 | 102 |
| D99  | mom-stand | Standardized moment                         | JoramSoch       | 2020-10-08 | 103 |
| D100 | chi2      | Chi-squared distribution                    | kjpetrykowski   | 2020-10-13 | 294 |
| D101 | med       | Median                                      | JoramSoch       | 2020-10-15 | 94  |
| D102 | mode      | Mode  | JoramSoch       | 2020-10-15 | 96  |
| D103 | prob-exc  | Exceedance probability                      | JoramSoch       | 2020-10-22 | 11  |
| D104 | dir-data  | Dirichlet-distributed data                  | JoramSoch       | 2020-10-22 | 673 |
| D105 | rvar-disc | Discrete and continuous random variable     | JoramSoch       | 2020-10-29 | 5   |
| D106 | rvar-uni  | Univariate and multivariate random variable | JoramSoch       | 2020-11-06 | 6   |
| D107 | min       | Minimum                                     | JoramSoch       | 2020-11-12 | 98  |
| D108 | max       | Maximum                                     | JoramSoch       | 2020-11-12 | 98  |
| D109 | rexp      | Random experiment                           | JoramSoch       | 2020-11-19 | 2   |
| D110 | reve      | Random event                                | JoramSoch       | 2020-11-19 | 4   |
| D111 | cvlme     | Cross-validated log model evidence          | JoramSoch       | 2020-11-19 | 715 |
| D112 | ind-cond  | Conditional independence                    | JoramSoch       | 2020-11-19 | 9   |
| D113 | uplme     | Uniform-prior log model evidence            | JoramSoch       | 2020-11-25 | 715 |
| D114 | ebhme     | Empirical Bayesian log model evidence       | JoramSoch       | 2020-11-25 | 716 |

|      |              |  |           |            |     |
|------|--------------|--|-----------|------------|-----|
| D115 | vblme        | Variational Bayesian log model evidence            | JoramSoch | 2020-11-25 | 717 |
| D116 | prior-flat   | Flat, hard and soft prior distribution             | JoramSoch | 2020-12-02 | 153 |
| D117 | prior-uni    | Uniform and non-uniform prior distribution         | JoramSoch | 2020-12-02 | 153 |
| D118 | prior-inf    | Informative and non-informative prior distribution | JoramSoch | 2020-12-02 | 153 |
| D119 | prior-emp    | Empirical and theoretical prior distribution       | JoramSoch | 2020-12-02 | 154 |
| D120 | prior-conj   | Conjugate and non-conjugate prior distribution     | JoramSoch | 2020-12-02 | 154 |
| D121 | prior-maxent | Maximum entropy prior distribution                 | JoramSoch | 2020-12-02 | 154 |
| D122 | prior-eb     | Empirical Bayes prior distribution                 | JoramSoch | 2020-12-02 | 155 |
| D123 | prior-ref    | Reference prior distribution                       | JoramSoch | 2020-12-02 | 155 |
| D124 | ug           | Univariate Gaussian                                | JoramSoch | 2021-03-03 | 406 |
| D125 | h0           | Null hypothesis                                    | JoramSoch | 2021-03-12 | 141 |
| D126 | h1           | Alternative hypothesis                             | JoramSoch | 2021-03-12 | 142 |
| D127 | hyp          | Statistical hypothesis                             | JoramSoch | 2021-03-19 | 139 |
| D128 | hyp-simp     | Simple and composite hypothesis                    | JoramSoch | 2021-03-19 | 139 |
| D129 | hyp-point    | Point and set hypothesis                           | JoramSoch | 2021-03-19 | 140 |
| D130 | test         | Statistical hypothesis test                        | JoramSoch | 2021-03-19 | 141 |
| D131 | tstat        | Test statistic                                     | JoramSoch | 2021-03-19 | 142 |
| D132 | size         | Size of a statistical test                         | JoramSoch | 2021-03-19 | 143 |
| D133 | alpha        | Significance level                                 | JoramSoch | 2021-03-19 | 143 |
| D134 | cval         | Critical value                                     | JoramSoch | 2021-03-19 | 144 |
| D135 | pval         | p-value  | JoramSoch | 2021-03-19 | 144 |
| D136 | ugkv         | Univariate Gaussian with known variance            | JoramSoch | 2021-03-23 | 430 |
| D137 | power        | Power of a statistical test                        | JoramSoch | 2021-03-31 | 143 |
| D138 | hyp-tail     | One-tailed and two-tailed hypothesis               | JoramSoch | 2021-03-31 | 140 |
| D139 | test-tail    | One-tailed and two-tailed test                     | JoramSoch | 2021-03-31 | 142 |

|      |             |  |           |            |     |
|------|-------------|--|-----------|------------|-----|
| D140 | dist-samp   | Sampling distribution                  | JoramSoch | 2021-03-31 | 22  |
| D141 | cdf-joint   | Joint cumulative distribution function | JoramSoch | 2020-04-07 | 42  |
| D142 | mean-samp   | Sample mean                            | JoramSoch | 2021-04-16 | 51  |
| D143 | var-samp    | Sample variance                        | JoramSoch | 2021-04-16 | 69  |
| D144 | cov-samp    | Sample covariance                      | ciaranmci | 2021-04-21 | 78  |
| D145 | prec        | Precision                              | JoramSoch | 2020-04-21 | 76  |
| D146 | f           | F-distribution                         | JoramSoch | 2020-04-21 | 298 |
| D147 | t           | t-distribution                         | JoramSoch | 2021-04-21 | 252 |
| D148 | mvt         | Multivariate t-distribution            | JoramSoch | 2020-04-21 | 364 |
| D149 | eb          | Empirical Bayes                        | JoramSoch | 2021-04-29 | 157 |
| D150 | vb          | Variational Bayes                      | JoramSoch | 2021-04-29 | 158 |
| D151 | mome        | Method-of-moments estimation           | JoramSoch | 2021-04-29 | 138 |
| D152 | nst         | Non-standardized t-distribution        | JoramSoch | 2021-05-20 | 253 |
| D153 | covmat-samp | Sample covariance matrix               | JoramSoch | 2021-05-20 | 82  |
| D154 | mean-rvec   | Expected value of a random vector      | JoramSoch | 2021-07-08 | 66  |
| D155 | mean-rmat   | Expected value of a random matrix      | JoramSoch | 2021-07-08 | 68  |
| D156 | exc         | Mutual exclusivity                     | JoramSoch | 2021-07-23 | 11  |
| D157 | sun         | Standard uniform distribution          | JoramSoch | 2021-07-23 | 210 |
| D158 | prob-ax     | Kolmogorov axioms of probability       | JoramSoch | 2021-07-30 | 12  |
| D159 | cf          | Characteristic function                | JoramSoch | 2021-09-22 | 49  |
| D160 | tglm        | Transformed general linear model       | JoramSoch | 2021-10-21 | 613 |
| D161 | iglm        | Inverse general linear model           | JoramSoch | 2021-10-21 | 615 |
| D162 | cfm         | Corresponding forward model            | JoramSoch | 2021-10-21 | 620 |
| D163 | slr         | Simple linear regression               | JoramSoch | 2021-10-27 | 481 |
| D164 | regline     | Regression line                        | JoramSoch | 2021-10-27 | 496 |
| D165 | samp-spc    | Sample space                           | JoramSoch | 2021-11-26 | 2   |
| D166 | eve-spc     | Event space                            | JoramSoch | 2021-11-26 | 2   |
| D167 | prob-spc    | Probability space                      | JoramSoch | 2021-11-26 | 2   |

|      |              |   |                 |            |     |
|------|--------------|---|-----------------|------------|-----|
| D168 | corr-samp    | Sample correlation coefficient  | JoramSoch       | 2021-12-14 | 92  |
| D169 | corrmat-samp | Sample correlation matrix   | JoramSoch       | 2021-12-14 | 93  |
| D170 | lognorm      | Log-normal distribution   | majapavlo       | 2022-02-07 | 282 |
| D171 | aicc         | Corrected Akaike information criterion                                | JoramSoch       | 2022-02-11 | 706 |
| D172 | dev          | Deviance  | JoramSoch       | 2022-03-01 | 710 |
| D173 | mse          | Mean squared error  | JoramSoch       | 2022-03-27 | 133 |
| D174 | ci           | Confidence interval   | JoramSoch       | 2022-03-27 | 135 |
| D175 | nw           | Normal-Wishart distribution   | JoramSoch       | 2022-05-14 | 401 |
| D176 | covmat-cross | Cross-covariance matrix   | JoramSoch       | 2022-09-26 | 87  |
| D177 | betabin      | Beta-binomial distribution  | JoramSoch       | 2022-10-20 | 192 |
| D178 | betabin-data | Beta-binomial data  | JoramSoch       | 2022-10-20 | 676 |
| D179 | me           | Model evidence  | JoramSoch       | 2022-10-20 | 711 |
| D180 | fe           | Family evidence   | JoramSoch       | 2022-10-20 | 717 |
| D181 | anova1       | One-way analysis of variance  | JoramSoch       | 2022-11-06 | 450 |
| D182 | anova2       | Two-way analysis of variance  | JoramSoch       | 2022-11-06 | 461 |
| D183 | trss         | Treatment sum of squares  | JoramSoch       | 2022-12-14 | 450 |
| D184 | iass         | Interaction sum of squares  | JoramSoch       | 2022-12-14 | 462 |
| D185 | tcon         | t-contrast for contrast-based inference in multiple linear regression | JoramSoch       | 2022-12-16 | 548 |
| D186 | fcon         | F-contrast for contrast-based inference in multiple linear regression | JoramSoch       | 2022-12-16 | 549 |
| D187 | exg          | ex-Gaussian distribution  | tomfaulkenberry | 2023-04-18 | 324 |
| D188 | skew         | Skewness  | tomfaulkenberry | 2023-04-20 | 76  |
| D189 | bvn          | Bivariate normal distribution   | JoramSoch       | 2023-09-22 | 355 |
| D190 | skew-samp    | Sample skewness   | tomfaulkenberry | 2023-10-30 | 77  |
| D191 | map          | Maximum-a-posteriori estimation                                       | JoramSoch       | 2023-12-01 | 152 |
| D192 | sr           | Scoring rule  | KarahanS        | 2024-02-28 | 162 |
| D193 | psr          | Proper scoring rule   | KarahanS        | 2024-02-28 | 162 |
| D194 | spsr         | Strictly proper scoring rule  | KarahanS        | 2024-02-28 | 162 |



|      |            |  |             |            |     |
|------|------------|--|-------------|------------|-----|
| D195 | lpsr       | Log probability scoring rule                     | KarahanS    | 2024-02-28 | 163 |
| D196 | fstat      | F-statistic                                      | JoramSoch   | 2024-03-15 | 699 |
| D197 | bsr        | Brier scoring rule                               | KarahanS    | 2024-03-23 | 166 |
| D198 | lr         | Likelihood ratio                                 | JoramSoch   | 2024-06-14 | 138 |
| D199 | llr        | Log-likelihood ratio                             | JoramSoch   | 2024-06-14 | 138 |
| D200 | iid        | independent and identically distributed          | JoramSoch   | 2024-08-08 | 6   |
| D201 | post-pred  | Posterior predictive distribution                | aloctavodia | 2024-08-18 | 151 |
| D202 | prior-pred | Prior predictive distribution                    | aloctavodia | 2024-08-19 | 147 |
| D203 | data       | Data   | JoramSoch   | 2024-09-20 | 3   |
| D204 | para       | Parameter  | JoramSoch   | 2024-09-27 | 23  |
| D205 | stat       | Statistic  | JoramSoch   | 2024-10-04 | 3   |
| D206 | dof        | Degrees of freedom                               | JoramSoch   | 2024-10-11 | 25  |
| D207 | dist-uni   | Unimodal and multimodal probability distribution | JoramSoch   | 2024-10-25 | 22  |
| D208 | est        | Estimand, estimator and estimate                 | JoramSoch   | 2024-11-01 | 133 |
| D209 | est-bias   | Biased vs. unbiased estimator                    | JoramSoch   | 2024-11-08 | 133 |
| D210 | samp       | Sample   | JoramSoch   | 2024-11-15 | 3   |
| D211 | samp-size  | Sample size                                      | JoramSoch   | 2024-11-22 | 3   |
| D212 | var-pool   | Pooled sample variance                           | JoramSoch   | 2025-01-10 | 70  |
| D213 | std-pool   | Pooled sample standard deviation                 | JoramSoch   | 2025-01-10 | 97  |
| D214 | std-samp   | Sample standard deviation                        | JoramSoch   | 2025-01-10 | 96  |
| D215 | odds       | Odds ratio, prior and posterior odds             | JoramSoch   | 2025-02-06 | 155 |
| D216 | pdf-joint  | Joint probability density function               | JoramSoch   | 2025-02-06 | 35  |
| D217 | para-loc   | Location parameter                               | JoramSoch   | 2025-03-14 | 23  |
| D218 | para-scal  | Scale parameter                                  | JoramSoch   | 2025-03-14 | 23  |
| D219 | para-rate  | Rate parameter                                   | JoramSoch   | 2025-03-14 | 24  |
| D220 | para-shape | Shape parameter                                  | JoramSoch   | 2025-03-14 | 24  |
| D221 | stat-desc  | Descriptive and inferential statistics           | JoramSoch   | 2025-04-04 | 4   |
| D222 | bvn-data   | Bivariate normally distributed data              | JoramSoch   | 2025-06-20 | 598 |

|      |          |  |                  |            |     |
|------|----------|--|------------------|------------|-----|
| D223 | mvn-data | Multivariate normally distributed data | JoramSoch        | 2025-06-20 | 598 |
| D224 | mde      | Minimum detectable effect              | alexanderdbolton | 2025-07-15 | 145 |
| D225 | mrss     | Minimum required sample size           | alexanderdbolton | 2025-07-15 | 146 |

## 3 Proof by Topic

### A

- Accuracy and complexity for Bayesian linear regression, 571
- Accuracy and complexity for Bayesian linear regression with known covariance, 595
- Accuracy and complexity for the univariate Gaussian, 423
- Accuracy and complexity for the univariate Gaussian with known variance, 441
- Addition law of probability, 16
- Addition of the differential entropy upon multiplication with a constant, 112
- Addition of the differential entropy upon multiplication with invertible matrix, 113
- Additivity of the Kullback-Leibler divergence for independent distributions, 129
- Additivity of the variance for independent random variables, 75
- Akaike information criterion for multiple linear regression, 563
- Application of Cochran's theorem to two-way analysis of variance, 468
- Approximation of log family evidences based on log model evidences, 721

### B

- Bayes' rule, 157
- Bayes' theorem, 156
- Bayesian information criterion for multiple linear regression, 563
- Bayesian model averaging in terms of log model evidences, 733
- Bernoulli distribution is a special case of categorical distribution, 178
- Best linear unbiased estimator for the inverse general linear model, 616
- Beta distribution is a special case of Dirichlet distribution, 300
- Binomial distribution is a special case of multinomial distribution, 183
- Binomial test, 630
- Bonferroni's inequality, 17
- Boole's inequality, 17
- Brier scoring rule is strictly proper scoring rule, 167

### C

- Characteristic function of a function of a random variable, 49
- Chebyshev's inequality, 62
- Chi-squared distribution is a special case of gamma distribution, 294
- Combined posterior distribution for Bayesian linear regression when analyzing conditionally independent data sets, 583
- Combined posterior distributions in terms of individual posterior distributions obtained from conditionally independent data, 150
- Concavity of the Shannon entropy, 105
- Conditional distributions of the multivariate normal distribution, 349
- Conditional distributions of the normal-gamma distribution, 377
- Conjugate prior distribution for Bayesian linear regression, 565
- Conjugate prior distribution for Bayesian linear regression with known covariance, 589
- Conjugate prior distribution for binomial observations, 634
- Conjugate prior distribution for multinomial observations, 647
- Conjugate prior distribution for multivariate Bayesian linear regression, 623
- Conjugate prior distribution for Poisson-distributed data, 659
- Conjugate prior distribution for the Poisson distribution with exposure values, 665

- Conjugate prior distribution for the univariate Gaussian, 416
- Conjugate prior distribution for the univariate Gaussian with known variance, 435
- Construction of a bivariate normal distribution from univariate standard normal random variables, 359
- Construction of confidence intervals using Wilks' theorem, 135
- Construction of unbiased estimator for variance, 688
- Construction of unbiased estimator for variance in multiple linear regression, 689
- Continuous uniform distribution maximizes differential entropy for fixed range, 218
- Convexity of the cross-entropy, 107
- Convexity of the Kullback-Leibler divergence, 128
- Corrected Akaike information criterion converges to uncorrected Akaike information criterion when infinite data are available, 706
- Corrected Akaike information criterion for multiple linear regression, 564
- Corrected Akaike information criterion in terms of maximum log-likelihood, 707
- Correlation always falls between -1 and +1, 91
- Correlation coefficient in terms of standard scores, 92
- Covariance and variance of the normal-gamma distribution, 371
- Covariance matrices of the matrix-normal distribution, 388
- Covariance matrix of the categorical distribution, 202
- Covariance matrix of the multinomial distribution, 206
- Covariance matrix of the multivariate normal distribution, 340
- Covariance matrix of the sum of two random vectors, 88
- Covariance of independent random variables, 80
- Cross-validated log Bayes factor for the univariate Gaussian, 429
- Cross-validated log Bayes factor for the univariate Gaussian with known variance, 447
- Cross-validated log model evidence for binomial observations, 640
- Cross-validated log model evidence for multinomial observations, 654
- Cross-validated log model evidence for the univariate Gaussian, 424
- Cross-validated log model evidence for the univariate Gaussian with known variance, 445
- Cumulative distribution function in terms of probability density function of a continuous random variable, 39
- Cumulative distribution function in terms of probability mass function of a discrete random variable, 39
- Cumulative distribution function of a strictly decreasing function of a random variable, 38
- Cumulative distribution function of a strictly increasing function of a random variable, 37
- Cumulative distribution function of a sum of independent random variables, 36
- Cumulative distribution function of the beta distribution, 307
- Cumulative distribution function of the beta-binomial distribution, 194
- Cumulative distribution function of the binomial distribution, 184
- Cumulative distribution function of the continuous uniform distribution, 211
- Cumulative distribution function of the discrete uniform distribution, 172
- Cumulative distribution function of the exponential distribution, 275
- Cumulative distribution function of the gamma distribution, 262
- Cumulative distribution function of the log-normal distribution, 284
- Cumulative distribution function of the multinomial distribution, 204
- Cumulative distribution function of the normal distribution, 230

- Decompositions of the variational free energy, 159
- Derivation of Bayesian model averaging, 732
- Derivation of  $R^2$  and adjusted  $R^2$ , 690
- Derivation of the Bayesian information criterion, 708
- Derivation of the family evidence, 718
- Derivation of the log Bayes factor, 727
- Derivation of the log family evidence, 719
- Derivation of the log model evidence, 712
- Derivation of the model evidence, 711
- Derivation of the posterior model probability, 729
- Deviance for multiple linear regression, 561
- Deviance information criterion for multiple linear regression, 575
- Differential entropy can be negative, 110
- Differential entropy for the matrix-normal distribution, 389
- Differential entropy of the continuous uniform distribution, 216
- Differential entropy of the gamma distribution, 271
- Differential entropy of the multivariate normal distribution, 342
- Differential entropy of the normal distribution, 243
- Differential entropy of the normal-gamma distribution, 372
- Discrete uniform distribution maximizes entropy for finite support, 176
- Distribution of parameter estimates for simple linear regression, 491
- Distribution of residual sum of squares in multiple linear regression with weighted least squares, 540
- Distribution of the coefficient of determination under the null hypothesis, 696
- Distribution of the inverse general linear model, 616
- Distribution of the transformed general linear model, 613
- Distributional transformation using cumulative distribution function, 42
- Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon ordinary least squares, 537
- Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon weighted least squares, 538

## E

- Effects of mean-centering on parameter estimates for simple linear regression, 494
- Encompassing prior method for computing Bayes factors, 725
- Entropy of the Bernoulli distribution, 181
- Entropy of the binomial distribution, 188
- Entropy of the categorical distribution, 202
- Entropy of the discrete uniform distribution, 174
- Entropy of the multinomial distribution, 207
- Entropy of the Poisson distribution, 199
- Equivalence of log-likelihood ratio and mutual information for the general linear model, 612
- Equivalence of log-likelihood ratios for regular and inverse general linear model, 618
- Equivalence of matrix-normal distribution and multivariate normal distribution, 386
- Equivalence of operations for model evidence and log model evidence, 714
- Equivalence of parameter estimates from the transformed general linear model, 614
- Exceedance probabilities for the Dirichlet distribution, 382
- Exceedance probability for a random variable in terms of cumulative distribution function, 40

- Existence of a corresponding forward model, 622
- Expectation of a bilinear form, 67
- Expectation of a quadratic form, 66
- Expectation of a quadratic form for the multivariate normal distribution, 341
- Expectation of parameter estimates for simple linear regression, 487
- Expectation of the cross-validated log Bayes factor for the univariate Gaussian with known variance, 448
- Expectation of the log Bayes factor for the univariate Gaussian with known variance, 443
- Expected value of a non-negative random variable, 51
- Expected value of the trace of a matrix, 68
- Expected value of  $x$  times  $\ln(x)$  for a gamma distribution, 270
- Exponential distribution is a special case of gamma distribution, 273
- Expression of  $R^2$  in terms of residual variances, 691
- Expression of the cumulative distribution function of the normal distribution without the error function, 232
- Expression of the noise precision posterior for Bayesian linear regression using prediction and parameter errors, 579
- Expression of the probability mass function of the beta-binomial distribution using only the gamma function, 194
- Extreme points of the probability density function of the normal distribution, 241

## F

- F-statistic in terms of ordinary least squares estimates in one-way analysis of variance, 457
- F-statistics in terms of ordinary least squares estimates in two-way analysis of variance, 480
- F-test for equality of variances in two independent samples, 412
- F-test for grand mean in two-way analysis of variance, 478
- F-test for interaction in two-way analysis of variance, 477
- F-test for main effect in one-way analysis of variance, 453
- F-test for main effect in two-way analysis of variance, 474
- F-test for multiple linear regression using contrast-based inference, 552
- First central moment is zero, 102
- First raw moment is mean, 101
- Full width at half maximum for the normal distribution, 240

## G

- Gamma distribution is a special case of Wishart distribution, 257
- Gaussian integral, 227
- Gibbs' inequality, 108

## I

- Independence of estimated parameters and residuals in multiple linear regression, 535
- Independence of products of multivariate normal random vector, 354
- Independent random variables are uncorrelated, 92
- Inflection points of the probability density function of the normal distribution, 242
- Invariance of the covariance matrix under addition of constant vector, 85
- Invariance of the differential entropy under addition of a constant, 111
- Invariance of the Kullback-Leibler divergence under parameter transformation, 130
- Invariance of the variance under addition of a constant, 72

- Inverse transformation method using cumulative distribution function, 41

## J

- Jensen's inequality, 61
- Joint likelihood is the product of likelihood function and prior density, 149

## K

- Kullback-Leibler divergence for the Bernoulli distribution, 181
- Kullback-Leibler divergence for the binomial distribution, 189
- Kullback-Leibler divergence for the continuous uniform distribution, 217
- Kullback-Leibler divergence for the Dirichlet distribution, 381
- Kullback-Leibler divergence for the discrete uniform distribution, 175
- Kullback-Leibler divergence for the gamma distribution, 272
- Kullback-Leibler divergence for the matrix-normal distribution, 390
- Kullback-Leibler divergence for the multivariate normal distribution, 344
- Kullback-Leibler divergence for the normal distribution, 244
- Kullback-Leibler divergence for the normal-gamma distribution, 373
- Kullback-Leibler divergence for the Wishart distribution, 399

## L

- Law of the unconscious statistician, 58
- Law of total covariance, 80
- Law of total expectation, 57
- Law of total probability, 19
- Law of total variance, 75
- Linear combination of bivariate normal random variables, 363
- Linear combination of independent normal random variables, 247
- Linear transformation theorem for the matrix-normal distribution, 393
- Linear transformation theorem for the moment-generating function, 44
- Linear transformation theorem for the multivariate normal distribution, 347
- Linearity of the expected value, 53
- Log Bayes factor for Bayesian linear regression, 587
- Log Bayes factor for binomial observations, 637
- Log Bayes factor for multinomial observations, 651
- Log Bayes factor for the univariate Gaussian with known variance, 442
- Log Bayes factor in terms of log model evidences, 728
- Log family evidences in terms of log model evidences, 720
- Log model evidence for Bayesian linear regression, 569
- Log model evidence for Bayesian linear regression with known covariance, 592
- Log model evidence for binomial observations, 636
- Log model evidence for multinomial observations, 649
- Log model evidence for multivariate Bayesian linear regression, 627
- Log model evidence for Poisson-distributed data, 661
- Log model evidence for the Poisson distribution with exposure values, 668
- Log model evidence for the univariate Gaussian, 421
- Log model evidence for the univariate Gaussian with known variance, 439
- Log model evidence in terms of prior and posterior distribution, 712
- Log sum inequality, 109

- Log-likelihood ratio for multiple linear regression, 547
- Log-likelihood ratio for the general linear model, 608
- Log-odds and probability in logistic regression, 680
- Logarithmic expectation of the gamma distribution, 268

## M

- Marginal distribution of a conditional binomial distribution, 190
- Marginal distributions for the matrix-normal distribution, 394
- Marginal distributions of the multinomial distribution, 209
- Marginal distributions of the multivariate normal distribution, 348
- Marginal distributions of the normal-gamma distribution, 375
- Marginal likelihood is a definite integral of the joint likelihood, 152
- Marginally normal does not imply jointly normal, 251
- Markov's inequality, 62
- Maximum likelihood estimation can result in biased estimates, 137
- Maximum likelihood estimation for binomial observations, 631
- Maximum likelihood estimation for bivariate normally distributed data, 598
- Maximum likelihood estimation for Dirichlet-distributed data, 673
- Maximum likelihood estimation for multinomial observations, 644
- Maximum likelihood estimation for multiple linear regression, 544
- Maximum likelihood estimation for multivariate normally distributed data, 600
- Maximum likelihood estimation for Poisson-distributed data, 657
- Maximum likelihood estimation for simple linear regression, 508
- Maximum likelihood estimation for simple linear regression, 510
- Maximum likelihood estimation for the general linear model, 605
- Maximum likelihood estimation for the Poisson distribution with exposure values, 663
- Maximum likelihood estimation for the univariate Gaussian, 406
- Maximum likelihood estimation for the univariate Gaussian with known variance, 430
- Maximum likelihood estimator of variance in multiple linear regression is biased, 686
- Maximum likelihood estimator of variance is biased, 684
- Maximum log-likelihood for binomial observations, 632
- Maximum log-likelihood for multinomial observations, 644
- Maximum log-likelihood for multiple linear regression, 545
- Maximum log-likelihood for the general linear model, 607
- Maximum-a-posteriori estimation for Bayesian linear regression, 577
- Maximum-a-posteriori estimation for binomial observations, 633
- Maximum-a-posteriori estimation for multinomial observations, 646
- Mean, mode and median of the coefficient of determination under the null hypothesis, 697
- Mean of the Bernoulli distribution, 179
- Mean of the beta distribution, 308
- Mean of the binomial distribution, 186
- Mean of the categorical distribution, 201
- Mean of the continuous uniform distribution, 213
- Mean of the ex-Gaussian distribution, 327
- Mean of the exponential distribution, 277
- Mean of the gamma distribution, 263
- Mean of the log-normal distribution, 286
- Mean of the matrix-normal distribution, 388



- Mean of the multinomial distribution, 205
- Mean of the multivariate normal distribution, 337
- Mean of the normal distribution, 235
- Mean of the normal-gamma distribution, 369
- Mean of the normal-Wishart distribution, 402
- Mean of the Poisson distribution, 196
- Mean of the Wald distribution, 317
- Median of the beta distribution, 309
- Median of the continuous uniform distribution, 214
- Median of the exponential distribution, 278
- Median of the gamma distribution, 264
- Median of the log-normal distribution, 288
- Median of the normal distribution, 237
- Method of moments for beta-binomial data, 676
- Method of moments for beta-distributed data, 671
- Method of moments for ex-Gaussian-distributed data, 331
- Method of moments for Wald-distributed data, 322
- Mode of the beta distribution, 310
- Mode of the continuous uniform distribution, 214
- Mode of the exponential distribution, 278
- Mode of the gamma distribution, 265
- Mode of the log-normal distribution, 288
- Mode of the multivariate normal distribution, 338
- Mode of the normal distribution, 237
- Moment in terms of moment-generating function, 99
- Moment-generating function of a function of a random variable, 44
- Moment-generating function of a sum of independent random variables, 43
- Moment-generating function of linear combination of independent random variables, 45
- Moment-generating function of the beta distribution, 306
- Moment-generating function of the ex-Gaussian distribution, 326
- Moment-generating function of the exponential distribution, 274
- Moment-generating function of the gamma distribution, 261
- Moment-generating function of the multivariate normal distribution, 336
- Moment-generating function of the normal distribution, 229
- Moment-generating function of the Wald distribution, 315
- Monotonicity of probability, 12
- Monotonicity of probability, 13
- Monotonicity of the expected value, 54
- Multinomial test, 642
- Multiple linear regression is a special case of the general linear model, 524
- Multivariate normal distribution is a special case of matrix-normal distribution, 334
- Multivariate normal distribution maximizes differential entropy for fixed covariance, 346
- Mutual information of dependent and independent variables in the general linear model, 610
- Mutual information of the bivariate normal distribution, 361
- Mutual information of the multivariate normal distribution, 343

## N

- Necessary and sufficient condition for independence of multivariate normal random variables, 352

- Non-invariance of the differential entropy under change of variables, 114
- (Non-)Multiplicativity of the expected value, 55
- Non-negativity of the expected value, 52
- Non-negativity of the Kullback-Leibler divergence, 124
- Non-negativity of the Kullback-Leibler divergence, 125
- Non-negativity of the Kullback-Leibler divergence, 126
- Non-negativity of the Shannon entropy, 104
- Non-negativity of the variance, 71
- Non-symmetry of the Kullback-Leibler divergence, 127
- Normal distribution is a special case of multivariate normal distribution, 220
- Normal distribution maximizes differential entropy for fixed variance, 246
- Normal-gamma distribution is a special case of normal-Wishart distribution, 367
- Normally distributed and uncorrelated does not imply independent, 249

## O

- Omnibus F-test for multiple regressors in multiple linear regression, 556
- One-sample t-test for independent observations, 408
- One-sample z-test for independent observations, 432
- Ordinary least squares for multiple linear regression, 524
- Ordinary least squares for multiple linear regression, 525
- Ordinary least squares for multiple linear regression, 526
- Ordinary least squares for multiple linear regression with two regressors, 527
- Ordinary least squares for one-way analysis of variance, 451
- Ordinary least squares for simple linear regression, 483
- Ordinary least squares for simple linear regression, 485
- Ordinary least squares for the general linear model, 604
- Ordinary least squares for two-way analysis of variance, 462

## P

- Paired t-test for dependent observations, 411
- Paired z-test for dependent observations, 434
- Parameter estimates for simple linear regression are uncorrelated after mean-centering, 494
- Parameters of the corresponding forward model, 621
- Partition of a covariance matrix into expected values, 82
- Partition of a cross-covariance matrix into expected values, 88
- Partition of covariance into expected values, 78
- Partition of skewness into expected values, 77
- Partition of sums of squares for multiple linear regression, 530
- Partition of sums of squares for simple linear regression, 500
- Partition of sums of squares in one-way analysis of variance, 452
- Partition of sums of squares in two-way analysis of variance, 466
- Partition of the log model evidence into accuracy and complexity, 713
- Partition of the mean squared error into bias and variance, 134
- Partition of variance into expected values, 70
- Positive semi-definiteness of the covariance matrix, 84
- Posterior credibility region against the omnibus null hypothesis for Bayesian linear regression, 582
- Posterior density is proportional to joint likelihood, 149
- Posterior distribution for Bayesian linear regression, 567

- Posterior distribution for Bayesian linear regression with known covariance, 590
- Posterior distribution for binomial observations, 635
- Posterior distribution for multinomial observations, 648
- Posterior distribution for multivariate Bayesian linear regression, 624
- Posterior distribution for Poisson-distributed data, 660
- Posterior distribution for the Poisson distribution with exposure values, 666
- Posterior distribution for the univariate Gaussian, 418
- Posterior distribution for the univariate Gaussian with known variance, 437
- Posterior model probabilities in terms of Bayes factors, 729
- Posterior model probabilities in terms of log model evidences, 731
- Posterior model probability in terms of log Bayes factor, 730
- Posterior predictive distribution is a marginal distribution of the joint likelihood, 151
- Posterior probability of the alternative hypothesis for Bayesian linear regression, 580
- Posterior probability of the alternative model for binomial observations, 639
- Posterior probability of the alternative model for multinomial observations, 652
- Power analysis, minimum detectable effect and minimum required sample size for a one-sample t-test, 413
- Prior predictive distribution is a marginal distribution of the joint likelihood, 148
- Probability and log-odds in logistic regression, 679
- Probability density function is first derivative of cumulative distribution function, 34
- Probability density function of a linear function of a continuous random vector, 34
- Probability density function of a strictly decreasing function of a continuous random variable, 30
- Probability density function of a strictly increasing function of a continuous random variable, 29
- Probability density function of a sum of independent continuous random variables, 29
- Probability density function of an invertible function of a continuous random vector, 32
- Probability density function of the beta distribution, 305
- Probability density function of the bivariate normal distribution, 356
- Probability density function of the bivariate normal distribution in terms of correlation coefficient, 357
- Probability density function of the chi-squared distribution, 295
- Probability density function of the continuous uniform distribution, 210
- Probability density function of the Dirichlet distribution, 380
- Probability density function of the ex-Gaussian distribution, 324
- Probability density function of the exponential distribution, 274
- Probability density function of the F-distribution, 298
- Probability density function of the gamma distribution, 260
- Probability density function of the log-normal distribution, 283
- Probability density function of the matrix-normal distribution, 387
- Probability density function of the multivariate normal distribution, 336
- Probability density function of the multivariate t-distribution, 365
- Probability density function of the normal distribution, 228
- Probability density function of the normal-gamma distribution, 368
- Probability density function of the normal-Wishart distribution, 401
- Probability density function of the t-distribution, 254
- Probability density function of the Wald distribution, 315
- Probability integral transform using cumulative distribution function, 41
- Probability mass function of a strictly decreasing function of a discrete random variable, 27
- Probability mass function of a strictly increasing function of a discrete random variable, 26

- Probability mass function of a sum of independent discrete random variables, 25
- Probability mass function of an invertible function of a random vector, 28
- Probability mass function of the Bernoulli distribution, 178
- Probability mass function of the beta-binomial distribution, 192
- Probability mass function of the binomial distribution, 183
- Probability mass function of the categorical distribution, 201
- Probability mass function of the discrete uniform distribution, 172
- Probability mass function of the multinomial distribution, 203
- Probability mass function of the Poisson distribution, 196
- Probability of exhaustive events, 19
- Probability of exhaustive events, 20
- Probability of normal random variable being within standard deviations from its mean, 234
- Probability of the complement, 15
- Probability of the empty set, 14
- Probability of the empty set, 14
- Probability under mutual exclusivity, 11
- Probability under statistical independence, 10
- Probability-generating function is expectation of function of random variable, 46
- Probability-generating function of the binomial distribution, 185
- Projection matrix and residual-forming matrix are idempotent, 535
- Projection matrix and residual-forming matrix are symmetric, 534
- Projection of a data point to the regression line, 497

## Q

- Quantile function is inverse of strictly monotonically increasing cumulative distribution function, 48
- Quantile function of the continuous uniform distribution, 212
- Quantile function of the discrete uniform distribution, 173
- Quantile function of the exponential distribution, 276
- Quantile function of the gamma distribution, 263
- Quantile function of the log-normal distribution, 285
- Quantile function of the normal distribution, 235

## R

- Range of probability, 15
- Range of the variance of the Bernoulli distribution, 180
- Range of the variance of the binomial distribution, 187
- Raw moments of the chi-squared distribution, 297
- Redundancy of parameters describing the matrix-normal distribution, 396
- Redundancy of parameters describing the matrix-normal distribution, 397
- Relation of continuous Kullback-Leibler divergence to differential entropy, 132
- Relation of continuous mutual information to joint and conditional differential entropy, 123
- Relation of continuous mutual information to marginal and conditional differential entropy, 121
- Relation of continuous mutual information to marginal and joint differential entropy, 122
- Relation of discrete Kullback-Leibler divergence to Shannon entropy, 131
- Relation of mutual information to joint and conditional entropy, 119
- Relation of mutual information to marginal and conditional entropy, 117
- Relation of mutual information to marginal and joint entropy, 118

- Relationship between chi-squared distribution and beta distribution, 301
- Relationship between coefficient of determination and correlation coefficient in simple linear regression, 521
- Relationship between correlation coefficient and slope estimate in simple linear regression, 521
- Relationship between covariance and correlation, 81
- Relationship between covariance matrix and correlation matrix, 86
- Relationship between F-distribution and beta distribution, 303
- Relationship between F-statistic and maximum log-likelihood, 701
- Relationship between F-statistic and  $R^2$ , 699
- Relationship between gamma distribution and standard gamma distribution, 258
- Relationship between gamma distribution and standard gamma distribution, 259
- Relationship between multivariate normal distribution and chi-squared distribution, 335
- Relationship between multivariate t-distribution and F-distribution, 365
- Relationship between non-standardized t-distribution and t-distribution, 253
- Relationship between normal distribution and chi-squared distribution, 223
- Relationship between normal distribution and standard normal distribution, 220
- Relationship between normal distribution and standard normal distribution, 222
- Relationship between normal distribution and standard normal distribution, 222
- Relationship between normal distribution and t-distribution, 225
- Relationship between precision matrix and correlation matrix, 89
- Relationship between  $R^2$  and maximum log-likelihood, 692
- Relationship between residual variance and sample variance in simple linear regression, 519
- Relationship between second raw moment, variance and mean, 101
- Relationship between signal-to-noise ratio and maximum log-likelihood, 705
- Relationship between signal-to-noise ratio and  $R^2$ , 704
- Reparametrization for one-way analysis of variance, 458

## S

- Sampling from the matrix-normal distribution, 397
- Sampling from the multivariate normal distribution, 355
- Sampling from the normal-gamma distribution, 379
- Savage-Dickey density ratio for computing Bayes factors, 724
- Scaling of a random variable following the gamma distribution, 259
- Scaling of the covariance matrix upon multiplication with constant matrix, 85
- Scaling of the variance upon multiplication with a constant, 73
- Second central moment is variance, 102
- Self-covariance equals variance, 79
- Self-independence of random event, 10
- Simple linear regression is a special case of multiple linear regression, 482
- Skewness of the ex-Gaussian distribution, 329
- Skewness of the exponential distribution, 280
- Skewness of the Wald distribution, 319
- Specific t-test for single regressor in multiple linear regression, 554
- Square of expectation of product is less than or equal to product of expectation of squares, 59
- Statistical significance test for the coefficient of determinant based on an omnibus F-test, 694
- Statistical test for comparing simple linear regression models with and without slope parameter, 515
- Statistical test for intercept parameter in simple linear regression model, 511

- Statistical test for slope parameter in simple linear regression model, 513
- Sums of squares for simple linear regression, 498
- Symmetry of the covariance, 79
- Symmetry of the covariance matrix, 83

## T

- t-distribution is a special case of multivariate t-distribution, 252
- t-test for multiple linear regression using contrast-based inference, 549
- The expected value minimizes the mean squared error, 65
- The geometric mean of independent log-normal random variables is a log-normal random variable, 293
- The log probability scoring rule is a strictly proper scoring rule, 163
- The median minimizes the mean absolute error, 95
- The p-value follows a uniform distribution under the null hypothesis, 144
- The product of independent log-normal random variables is a log-normal random variable, 292
- The regression line goes through the center of mass point, 496
- The residuals and the covariate are uncorrelated in simple linear regression, 518
- The sum of residuals is zero in simple linear regression, 517
- The variational free energy is a lower bound on the log model evidence, 161
- Transformation matrices for ordinary least squares, 532
- Transformation matrices for simple linear regression, 503
- Transitivity of Bayes Factors, 723
- Transposition of a matrix-normal random variable, 391
- Two-sample t-test for independent observations, 409
- Two-sample z-test for independent observations, 433

## V

- Value of the probability-generating function for argument one, 47
- Value of the probability-generating function for argument zero, 47
- Variance of constant is zero, 71
- Variance of parameter estimates for simple linear regression, 489
- Variance of the Bernoulli distribution, 179
- Variance of the beta distribution, 313
- Variance of the binomial distribution, 186
- Variance of the coefficient of determination under the null hypothesis, 698
- Variance of the continuous uniform distribution, 215
- Variance of the ex-Gaussian distribution, 328
- Variance of the exponential distribution, 279
- Variance of the gamma distribution, 266
- Variance of the linear combination of two random variables, 74
- Variance of the log-normal distribution, 290
- Variance of the normal distribution, 238
- Variance of the Poisson distribution, 197
- Variance of the sum of two random variables, 73
- Variance of the Wald distribution, 318

## W

- Weak law of large numbers, 63

- Weighted least squares for multiple linear regression, 542
- Weighted least squares for multiple linear regression, 543
- Weighted least squares for simple linear regression, 505
- Weighted least squares for simple linear regression, 507
- Weighted least squares for the general linear model, 604

## 4 Definition by Topic

### A

- Akaike information criterion, 706
- Alternative hypothesis, 142

### B

- Bayes factor, 722
- Bayesian information criterion, 708
- Bayesian model averaging, 732
- Bernoulli distribution, 177
- Beta distribution, 300
- Beta-binomial data, 676
- Beta-binomial distribution, 192
- Beta-distributed data, 671
- Biased vs. unbiased estimator, 133
- Binomial distribution, 182
- Binomial observations, 630
- Bivariate normal distribution, 355
- Bivariate normally distributed data, 598
- Brier scoring rule, 166

### C

- Categorical distribution, 201
- Central moment, 102
- Characteristic function, 49
- Chi-squared distribution, 294
- Coefficient of determination, 690
- Conditional differential entropy, 116
- Conditional entropy, 106
- Conditional independence, 9
- Conditional probability distribution, 22
- Confidence interval, 135
- Conjugate and non-conjugate prior distribution, 154
- Constant, 5
- Continuous uniform distribution, 210
- Corrected Akaike information criterion, 706
- Correlation, 90
- Correlation matrix, 93
- Corresponding forward model, 620
- Covariance, 78
- Covariance matrix, 82
- Critical value, 144
- Cross-covariance matrix, 87
- Cross-entropy, 107
- Cross-validated log model evidence, 715
- Cumulant-generating function, 50
- Cumulative distribution function, 36



**D**

- Data, 3
- Degrees of freedom, 25
- Descriptive and inferential statistics, 4
- Deviance, 710
- Deviance information criterion, 709
- Differential cross-entropy, 116
- Differential entropy, 110
- Dirichlet distribution, 380
- Dirichlet-distributed data, 673
- Discrete and continuous random variable, 5
- Discrete uniform distribution, 172

**E**

- Empirical and theoretical prior distribution, 154
- Empirical Bayes, 157
- Empirical Bayes prior distribution, 155
- Empirical Bayesian log model evidence, 716
- Encompassing model, 726
- Estimand, estimator and estimate, 133
- Estimation matrix, 532
- Event space, 2
- ex-Gaussian distribution, 324
- Exceedance probability, 11
- Expected value, 50
- Expected value of a random matrix, 68
- Expected value of a random vector, 66
- Explained sum of squares, 529
- Exponential distribution, 273

**F**

- F-contrast for contrast-based inference in multiple linear regression, 549
- F-distribution, 298
- F-statistic, 699
- Family evidence, 717
- Flat, hard and soft prior distribution, 153
- Full probability model, 148
- Full width at half maximum, 97

**G**

- Gamma distribution, 256
- General linear model, 603
- Generative model, 147

**I**

- independent and identically distributed, 6
- Informative and non-informative prior distribution, 153
- Interaction sum of squares, 462

- Inverse general linear model, 615

**J**

- Joint cumulative distribution function, 42
- Joint differential entropy, 116
- Joint entropy, 106
- Joint likelihood, 148
- Joint probability, 7
- Joint probability density function, 35
- Joint probability distribution, 21

**K**

- Kolmogorov axioms of probability, 12
- Kullback-Leibler divergence, 124

**L**

- Law of conditional probability, 7
- Law of marginal probability, 7
- Likelihood function, 147
- Likelihood function, 147
- Likelihood ratio, 138
- Location parameter, 23
- Log Bayes factor, 727
- Log family evidence, 718
- Log model evidence, 711
- Log probability scoring rule, 163
- Log-likelihood function, 137
- Log-likelihood ratio, 138
- Log-normal distribution, 282
- Logistic regression, 679

**M**

- Marginal likelihood, 152
- Marginal probability distribution, 22
- Matrix-normal distribution, 386
- Maximum, 98
- Maximum entropy prior distribution, 154
- Maximum likelihood estimation, 137
- Maximum log-likelihood, 137
- Maximum-a-posteriori estimation, 152
- Mean squared error, 133
- Median, 94
- Method-of-moments estimation, 138
- Minimum, 98
- Minimum detectable effect, 145
- Minimum required sample size, 146
- Mode, 96
- Model evidence, 711

- Moment, 99
- Moment-generating function, 43
- Multinomial distribution, 203
- Multinomial observations, 642
- Multiple linear regression, 523
- Multivariate normal distribution, 334
- Multivariate normally distributed data, 598
- Multivariate t-distribution, 364
- Mutual exclusivity, 11
- Mutual information, 120
- Mutual information, 120

## N

- Non-standardized t-distribution, 253
- Normal distribution, 219
- Normal-gamma distribution, 366
- Normal-Wishart distribution, 401
- Null hypothesis, 141

## O

- Odds ratio, prior and posterior odds, 155
- One-tailed and two-tailed hypothesis, 140
- One-tailed and two-tailed test, 142
- One-way analysis of variance, 450

## P

- p-value, 144
- Parameter, 23
- Point and set hypothesis, 140
- Poisson distribution, 196
- Poisson distribution with exposure values, 663
- Poisson-distributed data, 657
- Pooled sample standard deviation, 97
- Pooled sample variance, 70
- Posterior distribution, 149
- Posterior model probability, 728
- Posterior predictive distribution, 151
- Power of a statistical test, 143
- Precision, 76
- Precision matrix, 89
- Prior distribution, 147
- Prior predictive distribution, 147
- Probability, 6
- Probability density function, 28
- Probability distribution, 21
- Probability mass function, 25
- Probability space, 2
- Probability-generating function, 46

- Projection matrix, 532
- Proper scoring rule, 162

**Q**

- Quantile function, 48

**R**

- Random event, 4
- Random experiment, 2
- Random matrix, 5
- Random variable, 4
- Random vector, 5
- Rate parameter, 24
- Raw moment, 100
- Reference prior distribution, 155
- Regression line, 496
- Residual sum of squares, 530
- Residual variance, 684
- Residual-forming matrix, 532

**S**

- Sample, 3
- Sample correlation coefficient, 92
- Sample correlation matrix, 93
- Sample covariance, 78
- Sample covariance matrix, 82
- Sample mean, 51
- Sample size, 3
- Sample skewness, 77
- Sample space, 2
- Sample standard deviation, 96
- Sample variance, 69
- Sampling distribution, 22
- Scale parameter, 23
- Scoring rule, 162
- Shannon entropy, 104
- Shape parameter, 24
- Signal-to-noise ratio, 703
- Significance level, 143
- Simple and composite hypothesis, 139
- Simple linear regression, 481
- Size of a statistical test, 143
- Skewness, 76
- Standard deviation, 96
- Standard gamma distribution, 257
- Standard normal distribution, 220
- Standard uniform distribution, 210
- Standardized moment, 103

- Statistic, 3
- Statistical hypothesis, 139
- Statistical hypothesis test, 141
- Statistical independence, 8
- Strictly proper scoring rule, 162

**T**

- t-contrast for contrast-based inference in multiple linear regression, 548
- t-distribution, 252
- Test statistic, 142
- Total sum of squares, 529
- Transformed general linear model, 613
- Treatment sum of squares, 450
- Two-way analysis of variance, 461

**U**

- Uniform and non-uniform prior distribution, 153
- Uniform-prior log model evidence, 715
- Unimodal and multimodal probability distribution, 22
- Univariate and multivariate random variable, 6
- Univariate Gaussian, 406
- Univariate Gaussian with known variance, 430

**V**

- Variance, 69
- Variational Bayes, 158
- Variational Bayesian log model evidence, 717

**W**

- Wald distribution, 314
- Wishart distribution, 398