

The Book of Statistical Proofs

DOI: 10.5281/zenodo.4305950

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

StatProofBook@gmail.com

2022-03-28, 03:52

Contents

I	General Theorems	1
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>	3
1.2	Random variables	3
1.2.1	<i>Random event</i>	3
1.2.2	<i>Random variable</i>	3
1.2.3	<i>Random vector</i>	4
1.2.4	<i>Random matrix</i>	4
1.2.5	<i>Constant</i>	4
1.2.6	<i>Discrete vs. continuous</i>	5
1.2.7	<i>Univariate vs. multivariate</i>	5
1.3	Probability	5
1.3.1	<i>Probability</i>	5
1.3.2	<i>Joint probability</i>	6
1.3.3	<i>Marginal probability</i>	6
1.3.4	<i>Conditional probability</i>	6
1.3.5	<i>Exceedance probability</i>	7
1.3.6	<i>Statistical independence</i>	7
1.3.7	<i>Conditional independence</i>	8
1.3.8	Probability under independence	9
1.3.9	<i>Mutual exclusivity</i>	10
1.3.10	Probability under exclusivity	10
1.4	Probability axioms	11
1.4.1	<i>Axioms of probability</i>	11
1.4.2	Monotonicity of probability	11
1.4.3	Probability of the empty set	12
1.4.4	Probability of the complement	13
1.4.5	Range of probability	13
1.4.6	Addition law of probability	14
1.4.7	Law of total probability	15
1.4.8	Probability of exhaustive events	16
1.5	Probability distributions	16
1.5.1	<i>Probability distribution</i>	16
1.5.2	<i>Joint distribution</i>	17

1.5.3	<i>Marginal distribution</i>	17
1.5.4	<i>Conditional distribution</i>	17
1.5.5	<i>Sampling distribution</i>	18
1.6	Probability functions	18
1.6.1	<i>Probability mass function</i>	18
1.6.2	Probability mass function of sum of independents	18
1.6.3	Probability mass function of strictly increasing function	19
1.6.4	Probability mass function of strictly decreasing function	20
1.6.5	Probability mass function of invertible function	20
1.6.6	<i>Probability density function</i>	21
1.6.7	Probability density function of sum of independents	21
1.6.8	Probability density function of strictly increasing function	22
1.6.9	Probability density function of strictly decreasing function	23
1.6.10	Probability density function of invertible function	25
1.6.11	Probability density function of linear transformation	27
1.6.12	Probability density function in terms of cumulative distribution function	28
1.6.13	<i>Cumulative distribution function</i>	28
1.6.14	Cumulative distribution function of sum of independents	29
1.6.15	Cumulative distribution function of strictly increasing function	30
1.6.16	Cumulative distribution function of strictly decreasing function	30
1.6.17	Cumulative distribution function of discrete random variable	31
1.6.18	Cumulative distribution function of continuous random variable	32
1.6.19	Probability integral transform	33
1.6.20	Inverse transformation method	33
1.6.21	Distributional transformation	34
1.6.22	<i>Joint cumulative distribution function</i>	35
1.6.23	<i>Quantile function</i>	35
1.6.24	Quantile function in terms of cumulative distribution function	35
1.6.25	<i>Characteristic function</i>	36
1.6.26	Characteristic function of arbitrary function	37
1.6.27	<i>Moment-generating function</i>	37
1.6.28	Moment-generating function of arbitrary function	38
1.6.29	Moment-generating function of linear transformation	38
1.6.30	Moment-generating function of linear combination	39
1.6.31	<i>Cumulant-generating function</i>	40
1.6.32	<i>Probability-generating function</i>	40
1.7	Expected value	41
1.7.1	<i>Definition</i>	41
1.7.2	<i>Sample mean</i>	41
1.7.3	Non-negative random variable	41
1.7.4	Non-negativity	42
1.7.5	Linearity	43
1.7.6	Monotonicity	45
1.7.7	(Non-)Multiplicativity	46

1.7.8	Expectation of a trace	48
1.7.9	Expectation of a quadratic form	48
1.7.10	Law of total expectation	49
1.7.11	Law of the unconscious statistician	50
1.7.12	<i>Expected value of a random vector</i>	52
1.7.13	<i>Expected value of a random matrix</i>	53
1.8	Variance	53
1.8.1	<i>Definition</i>	53
1.8.2	<i>Sample variance</i>	53
1.8.3	Partition into expected values	54
1.8.4	Non-negativity	54
1.8.5	Variance of a constant	55
1.8.6	Invariance under addition	56
1.8.7	Scaling upon multiplication	57
1.8.8	Variance of a sum	57
1.8.9	Variance of linear combination	58
1.8.10	Additivity under independence	59
1.8.11	Law of total variance	59
1.8.12	<i>Precision</i>	60
1.9	Covariance	60
1.9.1	<i>Definition</i>	60
1.9.2	<i>Sample covariance</i>	61
1.9.3	Partition into expected values	61
1.9.4	Covariance under independence	62
1.9.5	Relationship to correlation	62
1.9.6	Law of total covariance	63
1.9.7	<i>Covariance matrix</i>	63
1.9.8	<i>Sample covariance matrix</i>	64
1.9.9	Covariance matrix and expected values	64
1.9.10	Covariance matrix and correlation matrix	65
1.9.11	<i>Precision matrix</i>	66
1.9.12	Precision matrix and correlation matrix	67
1.10	Correlation	68
1.10.1	<i>Definition</i>	68
1.10.2	Range	68
1.10.3	<i>Sample correlation coefficient</i>	69
1.10.4	Relationship to standard scores	70
1.10.5	<i>Correlation matrix</i>	70
1.10.6	<i>Sample correlation matrix</i>	71
1.11	Measures of central tendency	72
1.11.1	<i>Median</i>	72
1.11.2	<i>Mode</i>	72
1.12	Measures of statistical dispersion	73
1.12.1	<i>Standard deviation</i>	73
1.12.2	<i>Full width at half maximum</i>	73
1.13	Further summary statistics	73
1.13.1	<i>Minimum</i>	73
1.13.2	<i>Maximum</i>	74

1.14	Further moments	74
1.14.1	<i>Moment</i>	74
1.14.2	Moment in terms of moment-generating function	75
1.14.3	<i>Raw moment</i>	76
1.14.4	First raw moment is mean	77
1.14.5	Second raw moment and variance	77
1.14.6	<i>Central moment</i>	78
1.14.7	First central moment is zero	78
1.14.8	Second central moment is variance	78
1.14.9	<i>Standardized moment</i>	79
2	Information theory	80
2.1	Shannon entropy	80
2.1.1	<i>Definition</i>	80
2.1.2	Non-negativity	80
2.1.3	Concavity	81
2.1.4	<i>Conditional entropy</i>	82
2.1.5	<i>Joint entropy</i>	82
2.1.6	<i>Cross-entropy</i>	83
2.1.7	Convexity of cross-entropy	83
2.1.8	Gibbs' inequality	84
2.1.9	Log sum inequality	85
2.2	Differential entropy	86
2.2.1	<i>Definition</i>	86
2.2.2	Negativity	86
2.2.3	Invariance under addition	87
2.2.4	Addition upon multiplication	88
2.2.5	Addition upon matrix multiplication	89
2.2.6	Non-invariance and transformation	91
2.2.7	<i>Conditional differential entropy</i>	92
2.2.8	<i>Joint differential entropy</i>	93
2.2.9	<i>Differential cross-entropy</i>	93
2.3	Discrete mutual information	94
2.3.1	<i>Definition</i>	94
2.3.2	Relation to marginal and conditional entropy	94
2.3.3	Relation to marginal and joint entropy	95
2.3.4	Relation to joint and conditional entropy	96
2.4	Continuous mutual information	97
2.4.1	<i>Definition</i>	97
2.4.2	Relation to marginal and conditional differential entropy	98
2.4.3	Relation to marginal and joint differential entropy	99
2.4.4	Relation to joint and conditional differential entropy	100
2.5	Kullback-Leibler divergence	101
2.5.1	<i>Definition</i>	101
2.5.2	Non-negativity	102
2.5.3	Non-negativity	102
2.5.4	Non-symmetry	103
2.5.5	Convexity	105
2.5.6	Additivity for independent distributions	105

	2.5.7	Invariance under parameter transformation	106
	2.5.8	Relation to discrete entropy	107
	2.5.9	Relation to differential entropy	108
3		Estimation theory	110
	3.1	Point estimates	110
	3.1.1	Partition of the mean squared error into bias and variance	110
	3.2	Interval estimates	111
	3.2.1	Construction of confidence intervals using Wilks' theorem	111
4		Frequentist statistics	113
	4.1	Likelihood theory	113
	4.1.1	<i>Likelihood function</i>	113
	4.1.2	<i>Log-likelihood function</i>	113
	4.1.3	<i>Maximum likelihood estimation</i>	113
	4.1.4	<i>Maximum log-likelihood</i>	114
	4.1.5	<i>Method of moments</i>	114
	4.2	Statistical hypotheses	115
	4.2.1	<i>Statistical hypothesis</i>	115
	4.2.2	<i>Simple vs. composite</i>	115
	4.2.3	<i>Point/exact vs. set/inexact</i>	115
	4.2.4	<i>One-tailed vs. two-tailed</i>	116
	4.3	Hypothesis testing	116
	4.3.1	<i>Statistical test</i>	116
	4.3.2	<i>Null hypothesis</i>	117
	4.3.3	<i>Alternative hypothesis</i>	117
	4.3.4	<i>One-tailed vs. two-tailed</i>	118
	4.3.5	<i>Test statistic</i>	118
	4.3.6	<i>Size of a test</i>	118
	4.3.7	<i>Power of a test</i>	119
	4.3.8	<i>Significance level</i>	119
	4.3.9	<i>Critical value</i>	120
	4.3.10	<i>p-value</i>	120
5		Bayesian statistics	121
	5.1	Probabilistic modeling	121
	5.1.1	<i>Generative model</i>	121
	5.1.2	<i>Likelihood function</i>	121
	5.1.3	<i>Prior distribution</i>	121
	5.1.4	<i>Full probability model</i>	122
	5.1.5	<i>Joint likelihood</i>	122
	5.1.6	Joint likelihood is product of likelihood and prior	122
	5.1.7	<i>Posterior distribution</i>	123
	5.1.8	Posterior density is proportional to joint likelihood	123
	5.1.9	<i>Marginal likelihood</i>	124
	5.1.10	Marginal likelihood is integral of joint likelihood	124
	5.2	Prior distributions	125
	5.2.1	<i>Flat vs. hard vs. soft</i>	125
	5.2.2	<i>Uniform vs. non-uniform</i>	125
	5.2.3	<i>Informative vs. non-informative</i>	125
	5.2.4	<i>Empirical vs. non-empirical</i>	126

5.2.5	<i>Conjugate vs. non-conjugate</i>	126
5.2.6	<i>Maximum entropy priors</i>	126
5.2.7	<i>Empirical Bayes priors</i>	127
5.2.8	<i>Reference priors</i>	127
5.3	Bayesian inference	128
5.3.1	Bayes' theorem	128
5.3.2	Bayes' rule	128
5.3.3	<i>Empirical Bayes</i>	129
5.3.4	<i>Variational Bayes</i>	130
II Probability Distributions		131
1	Univariate discrete distributions	132
1.1	Discrete uniform distribution	132
1.1.1	<i>Definition</i>	132
1.1.2	Probability mass function	132
1.1.3	Cumulative distribution function	133
1.1.4	Quantile function	134
1.2	Bernoulli distribution	135
1.2.1	<i>Definition</i>	135
1.2.2	Probability mass function	135
1.2.3	Mean	135
1.3	Binomial distribution	136
1.3.1	<i>Definition</i>	136
1.3.2	Probability mass function	136
1.3.3	Mean	137
1.4	Poisson distribution	138
1.4.1	<i>Definition</i>	138
1.4.2	Probability mass function	138
1.4.3	Mean	139
1.4.4	Variance	140
2	Multivariate discrete distributions	142
2.1	Categorical distribution	142
2.1.1	<i>Definition</i>	142
2.1.2	Probability mass function	142
2.1.3	Mean	142
2.2	Multinomial distribution	143
2.2.1	<i>Definition</i>	143
2.2.2	Probability mass function	143
2.2.3	Mean	144
3	Univariate continuous distributions	146
3.1	Continuous uniform distribution	146
3.1.1	<i>Definition</i>	146
3.1.2	<i>Standard uniform distribution</i>	146
3.1.3	Probability density function	146
3.1.4	Cumulative distribution function	147
3.1.5	Quantile function	148
3.1.6	Mean	149
3.1.7	Median	150

	3.1.8	Mode	151
3.2		Normal distribution	151
	3.2.1	<i>Definition</i>	151
	3.2.2	<i>Standard normal distribution</i>	152
	3.2.3	Relationship to standard normal distribution	152
	3.2.4	Relationship to standard normal distribution	153
	3.2.5	Relationship to standard normal distribution	154
	3.2.6	Relationship to chi-squared distribution	155
	3.2.7	Relationship to t-distribution	157
	3.2.8	Gaussian integral	159
	3.2.9	Probability density function	161
	3.2.10	Moment-generating function	161
	3.2.11	Cumulative distribution function	162
	3.2.12	Cumulative distribution function without error function . .	164
	3.2.13	Quantile function	166
	3.2.14	Mean	167
	3.2.15	Median	168
	3.2.16	Mode	169
	3.2.17	Variance	170
	3.2.18	Full width at half maximum	172
	3.2.19	Extreme points	173
	3.2.20	Inflection points	174
	3.2.21	Differential entropy	175
	3.2.22	Kullback-Leibler divergence	176
	3.2.23	Maximum entropy distribution	178
	3.2.24	Linear combination	179
3.3		t-distribution	181
	3.3.1	<i>Definition</i>	181
	3.3.2	<i>Non-standardized t-distribution</i>	181
	3.3.3	Relationship to non-standardized t-distribution	182
	3.3.4	Probability density function	183
3.4		Gamma distribution	185
	3.4.1	<i>Definition</i>	185
	3.4.2	<i>Standard gamma distribution</i>	185
	3.4.3	Relationship to standard gamma distribution	186
	3.4.4	Relationship to standard gamma distribution	187
	3.4.5	Probability density function	188
	3.4.6	Cumulative distribution function	188
	3.4.7	Quantile function	189
	3.4.8	Mean	190
	3.4.9	Variance	191
	3.4.10	Logarithmic expectation	192
	3.4.11	Expectation of $x \ln x$	194
	3.4.12	Differential entropy	195
	3.4.13	Kullback-Leibler divergence	196
3.5		Exponential distribution	198
	3.5.1	<i>Definition</i>	198
	3.5.2	Special case of gamma distribution	198

3.5.3	Probability density function	199
3.5.4	Cumulative distribution function	199
3.5.5	Quantile function	200
3.5.6	Mean	201
3.5.7	Median	202
3.5.8	Mode	203
3.6	Chi-squared distribution	204
3.6.1	<i>Definition</i>	204
3.6.2	Special case of gamma distribution	204
3.6.3	Probability density function	205
3.6.4	Moments	207
3.7	F-distribution	207
3.7.1	<i>Definition</i>	207
3.7.2	Probability density function	208
3.8	Beta distribution	210
3.8.1	<i>Definition</i>	210
3.8.2	Probability density function	210
3.8.3	Moment-generating function	211
3.8.4	Cumulative distribution function	212
3.8.5	Mean	213
3.8.6	Variance	214
3.9	Wald distribution	216
3.9.1	<i>Definition</i>	216
3.9.2	Probability density function	216
3.9.3	Moment-generating function	216
3.9.4	Mean	218
3.9.5	Variance	219
4	Multivariate continuous distributions	221
4.1	Multivariate normal distribution	221
4.1.1	<i>Definition</i>	221
4.1.2	Probability density function	221
4.1.3	Differential entropy	221
4.1.4	Kullback-Leibler divergence	223
4.1.5	Linear transformation	224
4.1.6	Marginal distributions	225
4.1.7	Conditional distributions	226
4.1.8	Conditions for independence	230
4.2	Multivariate t-distribution	231
4.2.1	<i>Definition</i>	231
4.2.2	Relationship to F-distribution	232
4.3	Normal-gamma distribution	233
4.3.1	<i>Definition</i>	233
4.3.2	Probability density function	234
4.3.3	Mean	235
4.3.4	Differential entropy	236
4.3.5	Kullback-Leibler divergence	238
4.3.6	Marginal distributions	239
4.3.7	Conditional distributions	242

4.4	Dirichlet distribution	244
4.4.1	<i>Definition</i>	244
4.4.2	Probability density function	244
4.4.3	Kullback-Leibler divergence	245
4.4.4	Exceedance probabilities	246
5	Matrix-variate continuous distributions	250
5.1	Matrix-normal distribution	250
5.1.1	<i>Definition</i>	250
5.1.2	Probability density function	250
5.1.3	Equivalence to multivariate normal distribution	251
5.1.4	Kullback-Leibler divergence	252
5.1.5	Linear transformation	253
5.1.6	Transposition	254
5.1.7	Drawing samples	255
5.2	Wishart distribution	256
5.2.1	<i>Definition</i>	256
5.2.2	Kullback-Leibler divergence	256
III Statistical Models		259
1	Univariate normal data	260
1.1	Univariate Gaussian	260
1.1.1	<i>Definition</i>	260
1.1.2	Maximum likelihood estimation	260
1.1.3	One-sample t-test	262
1.1.4	Two-sample t-test	264
1.1.5	Paired t-test	265
1.1.6	Conjugate prior distribution	266
1.1.7	Posterior distribution	268
1.1.8	Log model evidence	271
1.1.9	Accuracy and complexity	274
1.2	Univariate Gaussian with known variance	275
1.2.1	<i>Definition</i>	275
1.2.2	Maximum likelihood estimation	276
1.2.3	One-sample z-test	277
1.2.4	Two-sample z-test	278
1.2.5	Paired z-test	280
1.2.6	Conjugate prior distribution	280
1.2.7	Posterior distribution	282
1.2.8	Log model evidence	285
1.2.9	Accuracy and complexity	286
1.2.10	Log Bayes factor	288
1.2.11	Expectation of log Bayes factor	289
1.2.12	Cross-validated log model evidence	291
1.2.13	Cross-validated log Bayes factor	293
1.2.14	Expectation of cross-validated log Bayes factor	294
1.3	Simple linear regression	296
1.3.1	<i>Definition</i>	296
1.3.2	Special case of multiple linear regression	297

1.3.3	Ordinary least squares	298
1.3.4	Ordinary least squares	300
1.3.5	Expectation of estimates	302
1.3.6	Variance of estimates	304
1.3.7	Distribution of estimates	307
1.3.8	Effects of mean-centering	309
1.3.9	<i>Regression line</i>	311
1.3.10	Regression line includes center of mass	311
1.3.11	Projection of data point to regression line	312
1.3.12	Sums of squares	313
1.3.13	Transformation matrices	315
1.3.14	Weighted least squares	318
1.3.15	Weighted least squares	320
1.3.16	Maximum likelihood estimation	321
1.3.17	Maximum likelihood estimation	324
1.3.18	Sum of residuals is zero	325
1.3.19	Correlation with covariate is zero	326
1.3.20	Residual variance in terms of sample variance	327
1.3.21	Correlation coefficient in terms of slope estimate	329
1.3.22	Coefficient of determination in terms of correlation coefficient	330
1.4	Multiple linear regression	331
1.4.1	<i>Definition</i>	331
1.4.2	Ordinary least squares	332
1.4.3	Ordinary least squares	333
1.4.4	<i>Total sum of squares</i>	334
1.4.5	<i>Explained sum of squares</i>	334
1.4.6	<i>Residual sum of squares</i>	334
1.4.7	Total, explained and residual sum of squares	335
1.4.8	<i>Estimation matrix</i>	337
1.4.9	<i>Projection matrix</i>	337
1.4.10	<i>Residual-forming matrix</i>	337
1.4.11	Estimation, projection and residual-forming matrix	337
1.4.12	Idempotence of projection and residual-forming matrix	339
1.4.13	Weighted least squares	340
1.4.14	Weighted least squares	341
1.4.15	Maximum likelihood estimation	342
1.5	Bayesian linear regression	344
1.5.1	Conjugate prior distribution	344
1.5.2	Posterior distribution	346
1.5.3	Log model evidence	348
1.5.4	Posterior probability of alternative hypothesis	350
1.5.5	Posterior credibility region excluding null hypothesis	352
2	Multivariate normal data	354
2.1	General linear model	354
2.1.1	<i>Definition</i>	354
2.1.2	Ordinary least squares	354
2.1.3	Weighted least squares	355
2.1.4	Maximum likelihood estimation	356

2.2	Transformed general linear model	358
2.2.1	<i>Definition</i>	358
2.2.2	Derivation of the distribution	359
2.2.3	Equivalence of parameter estimates	360
2.3	Inverse general linear model	361
2.3.1	<i>Definition</i>	361
2.3.2	Derivation of the distribution	361
2.3.3	Best linear unbiased estimator	362
2.3.4	<i>Corresponding forward model</i>	364
2.3.5	Derivation of parameters	364
2.3.6	Proof of existence	365
2.4	Multivariate Bayesian linear regression	366
2.4.1	Conjugate prior distribution	366
2.4.2	Posterior distribution	368
2.4.3	Log model evidence	370
3	Poisson data	373
3.1	Poisson-distributed data	373
3.1.1	<i>Definition</i>	373
3.1.2	Maximum likelihood estimation	373
3.1.3	Conjugate prior distribution	375
3.1.4	Posterior distribution	376
3.1.5	Log model evidence	377
3.2	Poisson distribution with exposure values	379
3.2.1	<i>Definition</i>	379
3.2.2	Maximum likelihood estimation	379
3.2.3	Conjugate prior distribution	381
3.2.4	Posterior distribution	383
3.2.5	Log model evidence	384
4	Probability data	387
4.1	Beta-distributed data	387
4.1.1	<i>Definition</i>	387
4.1.2	Method of moments	387
4.2	Dirichlet-distributed data	389
4.2.1	<i>Definition</i>	389
4.2.2	Maximum likelihood estimation	389
5	Categorical data	393
5.1	Binomial observations	393
5.1.1	<i>Definition</i>	393
5.1.2	Conjugate prior distribution	393
5.1.3	Posterior distribution	394
5.1.4	Log model evidence	395
5.2	Multinomial observations	397
5.2.1	<i>Definition</i>	397
5.2.2	Conjugate prior distribution	397
5.2.3	Posterior distribution	398
5.2.4	Log model evidence	399
5.3	Logistic regression	401
5.3.1	<i>Definition</i>	401

	5.3.2	Probability and log-odds	402
	5.3.3	Log-odds and probability	403
IV		Model Selection	405
1		Goodness-of-fit measures	406
	1.1	Residual variance	406
		1.1.1 <i>Definition</i>	406
		1.1.2 Maximum likelihood estimator is biased	406
		1.1.3 Construction of unbiased estimator	408
	1.2	R-squared	409
		1.2.1 <i>Definition</i>	409
		1.2.2 Derivation of R^2 and adjusted R^2	410
		1.2.3 Relationship to maximum log-likelihood	411
	1.3	Signal-to-noise ratio	413
		1.3.1 <i>Definition</i>	413
		1.3.2 Relationship with R^2	413
2		Classical information criteria	415
	2.1	Akaike information criterion	415
		2.1.1 <i>Definition</i>	415
	2.2	Bayesian information criterion	415
		2.2.1 <i>Definition</i>	415
		2.2.2 Derivation	415
	2.3	Deviance information criterion	417
		2.3.1 <i>Definition</i>	417
3		Bayesian model selection	418
	3.1	Log model evidence	418
		3.1.1 <i>Definition</i>	418
		3.1.2 Derivation	418
		3.1.3 Partition into accuracy and complexity	419
		3.1.4 <i>Uniform-prior log model evidence</i>	420
		3.1.5 <i>Cross-validated log model evidence</i>	420
		3.1.6 <i>Empirical Bayesian log model evidence</i>	421
		3.1.7 <i>Variational Bayesian log model evidence</i>	422
	3.2	Log family evidence	422
		3.2.1 <i>Definition</i>	422
		3.2.2 Derivation	423
		3.2.3 Calculation from log model evidences	424
	3.3	Log Bayes factor	425
		3.3.1 <i>Definition</i>	425
		3.3.2 Derivation	426
		3.3.3 Calculation from log model evidences	426
	3.4	Bayes factor	427
		3.4.1 <i>Definition</i>	427
		3.4.2 Transitivity	428
		3.4.3 Computation using Savage-Dickey Density Ratio	428
		3.4.4 Computation using Encompassing Prior Method	430
		3.4.5 <i>Encompassing model</i>	431
	3.5	Posterior model probability	431

3.5.1	<i>Definition</i>	431
3.5.2	Derivation	432
3.5.3	Calculation from Bayes factors	432
3.5.4	Calculation from log Bayes factor	433
3.5.5	Calculation from log model evidences	434
3.6	Bayesian model averaging	435
3.6.1	<i>Definition</i>	435
3.6.2	Derivation	436
3.6.3	Calculation from log model evidences	436
V	Appendix	439
1	Proof by Number	440
2	Definition by Number	457
3	Proof by Topic	464
4	Definition by Topic	472

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 Definition I/1.2.2) out of a well-defined set of possible outcomes.

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

Together, sample space (\rightarrow Definition I/1.1.2), event space (\rightarrow Definition I/1.1.3) and probability function (\rightarrow Definition 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)).

Metadata: ID: D109 | shortcut: rexp | author: JoramSoch | date: 2020-11-19, 04:10.

1.1.2 Sample space

Definition: Given a random experiment (\rightarrow Definition 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 Ω 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.

Metadata: ID: D165 | shortcut: samp-spc | author: JoramSoch | date: 2021-11-26, 14:13.

1.1.3 Event space

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

Metadata: ID: D166 | shortcut: eve-spc | author: JoramSoch | date: 2021-11-26, 14:26.

1.1.4 Probability space

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

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

Metadata: ID: D167 | shortcut: prob-spc | author: JoramSoch | date: 2021-11-26, 14:30.

1.2 Random variables

1.2.1 Random event

Definition: A random event E is the outcome of a random experiment (\rightarrow Definition 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 Definition I/1.2.2) with two possible values (true and false, or 1 and 0). A random experiment (\rightarrow Definition I/1.1.1) with two possible outcomes is called a Bernoulli trial (\rightarrow Definition II/1.2.1).

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)).

Metadata: ID: D110 | shortcut: reve | author: JoramSoch | date: 2020-11-19, 04:33.

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 Definition I/1.1.1);
- formally, as a measurable function (\rightarrow Definition “meas-fct”) X defined on a probability space (\rightarrow Definition I/1.1.4) (Ω, \mathcal{E}, P) that maps from a sample space (\rightarrow Definition I/1.1.2) Ω to the real numbers \mathbb{R} using an event space (\rightarrow Definition I/1.1.3) \mathcal{E} and a probability function (\rightarrow Definition I/1.5.1) P ;
- more broadly, as any random quantity X such as a random event (\rightarrow Definition I/1.2.1), a random scalar (\rightarrow Definition I/1.2.2), a random vector (\rightarrow Definition I/1.2.3) or a random matrix (\rightarrow Definition 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.

Metadata: ID: D65 | shortcut: rvar | author: JoramSoch | date: 2020-05-27, 22:36.

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 Definition 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.

Metadata: ID: D66 | shortcut: rvec | author: JoramSoch | date: 2020-05-27, 22:44.

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 Definition I/1.2.2). Equivalently, a random matrix is an $n \times p$ matrix whose columns are n -dimensional random vectors (\rightarrow Definition 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.

Metadata: ID: D67 | shortcut: rmat | author: JoramSoch | date: 2020-05-27, 22:48.

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 Definition I/1.2.2) which is equal to its expected value (\rightarrow Definition I/1.7.1)

$$X = E(X) \tag{1}$$

or equivalently, whose variance (\rightarrow Definition I/1.8.1) is zero

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

Sources:

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

Metadata: ID: D96 | shortcut: const | author: JoramSoch | date: 2020-09-09, 01:30.

1.2.6 Discrete vs. continuous

Definition: Let X be a random variable (\rightarrow Definition 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 Definition 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 Definition I/1.6.6).

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.

Metadata: ID: D105 | shortcut: rvar-disc | author: JoramSoch | date: 2020-10-29, 04:44.

1.2.7 Univariate vs. multivariate

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

- X is called a two-valued random variable or random event (\rightarrow Definition I/1.2.1), if \mathcal{X} has exactly two elements, e.g. $\mathcal{X} = \{E, \overline{E}\}$ or $\mathcal{X} = \{\text{true}, \text{false}\}$ or $\mathcal{X} = \{1, 0\}$;
- X is called a univariate random variable or random scalar (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: D106 | shortcut: rvar-uni | author: JoramSoch | date: 2020-11-06, 03:47.

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 Definition 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>.

Metadata: ID: D48 | shortcut: prob | author: JoramSoch | date: 2020-05-10, 19:41.

1.3.2 Joint probability

Definition: Let A and B be two arbitrary statements about random variables (\rightarrow Definition 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 Definition 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.
- 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/>.

Metadata: ID: D49 | shortcut: prob-joint | author: JoramSoch | date: 2020-05-10, 19:49.

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 Definition 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 Definition 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 Definition I/1.2.6) random variable (\rightarrow Definition 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 Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with domain \mathcal{X} , is given by

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

Sources:

- Wikipedia (2020): “Marginal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: 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/>.

Metadata: ID: D50 | shortcut: prob-marg | author: JoramSoch | date: 2020-05-10, 20:01.

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 Definition 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 Definition 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 Definition 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.
- 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/>.

Metadata: ID: D51 | shortcut: prob-cond | author: JoramSoch | date: 2020-05-10, 20:06.

1.3.5 Exceedance probability

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

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

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>.

Metadata: ID: D103 | shortcut: prob-exc | author: JoramSoch | date: 2020-10-22, 04:36.

1.3.6 Statistical independence

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

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

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

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

2) A set of continuous random variables (\rightarrow Definition 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 Definition I/1.6.6) 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 Definition I/1.5.2) or marginal (\rightarrow Definition I/1.5.3) cumulative distribution functions (\rightarrow Definition I/1.6.13) and f are the respective probability density functions (\rightarrow Definition I/1.6.6).

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).

Metadata: ID: D75 | shortcut: ind | author: JoramSoch | date: 2020-06-06, 07:16.

1.3.7 Conditional independence

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

1) A set of discrete random variables (\rightarrow Definition 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 Definition I/1.3.2) of X_1, \dots, X_n given Y and $p(x_i)$ are the marginal (conditional) probabilities (\rightarrow Definition I/1.3.3) of X_i given Y .

2) A set of continuous random variables (\rightarrow Definition 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 Definition I/1.6.6) 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 Definition I/1.5.2) or marginal (conditional) (\rightarrow Definition I/1.5.3) cumulative distribution functions (\rightarrow Definition I/1.6.13) and f are the respective probability density functions (\rightarrow Definition I/1.6.6).

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.

Metadata: ID: D112 | shortcut: ind-cond | author: JoramSoch | date: 2020-11-19, 05:40.

1.3.8 Probability under independence

Theorem: Let A and B be two statements about random variables (\rightarrow Definition I/1.2.2). Then, if A and B are independent (\rightarrow Definition I/1.3.6), marginal (\rightarrow Definition I/1.3.3) and conditional (\rightarrow Definition 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 Definition I/1.3.6), then the joint probability (\rightarrow Definition I/1.3.2) is equal to the product of the marginal probabilities (\rightarrow Definition I/1.3.3):

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

The law of conditional probability (\rightarrow Definition 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).

Metadata: ID: P241 | shortcut: prob-ind | author: JoramSoch | date: 2021-07-23, 16:05.

1.3.9 Mutual exclusivity

Definition: Generally speaking, random events (\rightarrow Definition I/1.2.1) are mutually exclusive, if they cannot occur together, such that their intersection is equal to the empty set (\rightarrow Proof I/1.4.3).

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 Definition 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.

Metadata: ID: D156 | shortcut: exc | author: JoramSoch | date: 2021-07-23, 16:32.

1.3.10 Probability under exclusivity

Theorem: Let A and B be two statements about random variables (\rightarrow Definition I/1.2.2). Then, if A and B are mutually exclusive (\rightarrow Definition I/1.3.9), the probability (\rightarrow Definition I/1.3.1) of their disjunction is equal to the sum of the marginal probabilities (\rightarrow Definition I/1.3.3):

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

Proof: If A and B are mutually exclusive (\rightarrow Definition I/1.3.9), then their joint probability (\rightarrow Definition I/1.3.2) is zero:

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

The addition law of probability (\rightarrow Definition I/1.3.3) 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 Definition I/1.3.9) and the probability of the empty set is zero (\rightarrow Proof I/1.4.3), the joint probability (\rightarrow Definition 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.

Metadata: ID: P242 | shortcut: prob-exc | author: JoramSoch | date: 2021-07-23, 17:19.

1.4 Probability axioms

1.4.1 Axioms of probability

Definition: Let there be a sample space (\rightarrow Definition I/1.1.2) Ω , an event space (\rightarrow Definition I/1.1.3) \mathcal{E} and a probability measure (\rightarrow Definition “prob-meas”) P , such that $P(E)$ is the probability (\rightarrow Definition I/1.3.1) of some event (\rightarrow Definition 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 Definition I/1.3.9)) 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.

Metadata: ID: D158 | shortcut: prob-ax | author: JoramSoch | date: 2021-07-30, 11:11.

1.4.2 Monotonicity of probability

Theorem: Probability (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: P243 | shortcut: prob-mon | author: JoramSoch | date: 2021-07-30, 11:37.

1.4.3 Probability of the empty set

Theorem: The probability (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: P244 | shortcut: prob-emp | author: JoramSoch | date: 2021-07-30, 11:58.

1.4.4 Probability of the complement

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

$$P(A^c) = 1 - P(A) \quad (1)$$

where $A^c = \Omega \setminus A$ and Ω is the sample space (\rightarrow Definition I/1.1.2).

Proof: Since A and A^c are mutually exclusive (\rightarrow Definition I/1.3.9) and $A \cup A^c = \Omega$, the third axiom of probability (\rightarrow Definition I/1.4.1) implies:

$$\begin{aligned} P(A \cup A^c) &= P(A) + P(A^c) \\ P(\Omega) &= P(A) + P(A^c) \\ P(A^c) &= P(\Omega) - P(A) . \end{aligned} \quad (2)$$

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

$$P(A^c) = 1 - P(A) . \quad (3)$$

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.

Metadata: ID: P245 | shortcut: prob-comp | author: JoramSoch | date: 2021-07-30, 12:14.

1.4.5 Range of probability

Theorem: The probability (\rightarrow Definition 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 Definition I/1.4.1), we have:

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

By combining the first axiom of probability (\rightarrow Definition I/1.4.1) and the probability of the complement (\rightarrow Proof I/1.4.4), we obtain:

$$\begin{aligned} 1 - P(E) = P(E^c) &\geq 0 \\ 1 - P(E) &\geq 0 \\ P(E) &\leq 1 . \end{aligned} \quad (3)$$

Together, (2) and (3) imply that

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

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.

Metadata: ID: P246 | shortcut: prob-range | author: JoramSoch | date: 2021-07-30, 12:25.

1.4.6 Addition law of probability

Theorem: The probability (\rightarrow Definition 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 :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) . \quad (1)$$

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 Definition I/1.4.1), we have:

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

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 Definition I/1.4.1), we obtain:

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

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) . \quad (4)$$

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.

Metadata: ID: P247 | shortcut: prob-add | author: JoramSoch | date: 2021-07-30, 12:45.

1.4.7 Law of total probability

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

Metadata: ID: P248 | shortcut: prob-tot | author: JoramSoch | date: 2021-08-08, 03:56.

1.4.8 Probability of exhaustive events

Theorem: Let B_1, \dots, B_n be mutually exclusive (\rightarrow Definition I/1.3.9) and collectively exhaustive subsets of a sample space (\rightarrow Definition I/1.1.2) Ω . Then, their total probability (\rightarrow Proof I/1.4.7) 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 Definition I/1.4.1) implies that

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

and the second axiom of probability (\rightarrow Definition 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.

Metadata: ID: P249 | shortcut: prob-exh | author: JoramSoch | date: 2021-08-08, 04:10.

1.4.9 Probability of exhaustive events

Theorem: Let B_1, \dots, B_n be mutually exclusive (\rightarrow Definition I/1.3.9) and collectively exhaustive subsets of a sample space (\rightarrow Definition I/1.1.2) Ω . Then, their total probability (\rightarrow Proof I/1.4.7) is one:

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

Proof: The addition law of probability (\rightarrow Proof I/1.4.6) states that for two events (\rightarrow Definition I/1.2.1) A and B , the probability (\rightarrow Definition 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 Proof I/1.4.3), 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 Definition I/1.4.1), this means that the left-hand side of (6) becomes equal to one:

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

This proves the statement in (??).

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.

Metadata: ID: P319 | shortcut: prob-exh2 | author: JoramSoch | date: 2022-03-27, 23:14.

1.5 Probability distributions

1.5.1 Probability distribution

Definition: Let X be a random variable (\rightarrow Definition 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 Definition 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.

Metadata: ID: D55 | shortcut: dist | author: JoramSoch | date: 2020-05-17, 20:23.

1.5.2 Joint distribution

Definition: Let X and Y be random variables (\rightarrow Definition 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 Definition 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 Definition I/1.2.2) is called a bivariate distribution.
- The joint distribution of a random vector (\rightarrow Definition I/1.2.3) is called a multivariate distribution.
- The joint distribution of a random matrix (\rightarrow Definition 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.

Metadata: ID: D56 | shortcut: dist-joint | author: JoramSoch | date: 2020-05-17, 20:43.

1.5.3 Marginal distribution

Definition: Let X and Y be random variables (\rightarrow Definition 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 Definition 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 Definition I/1.5.2) of X and Y using the law of marginal probability (\rightarrow Definition 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.

Metadata: ID: D57 | shortcut: dist-marg | author: JoramSoch | date: 2020-05-17, 21:02.

1.5.4 Conditional distribution

Definition: Let X and Y be random variables (\rightarrow Definition 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 Definition 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 Definition I/1.5.2) of X and Y and the marginal distribution (\rightarrow Definition I/1.5.3) of Y using the law of conditional probability (\rightarrow Definition 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.

Metadata: ID: D58 | shortcut: dist-cond | author: JoramSoch | date: 2020-05-17, 21:25.

1.5.5 Sampling distribution

Definition: Let there be a random sample (\rightarrow Definition “samp”) with finite sample size (\rightarrow Definition “samp-size”). Then, the probability distribution (\rightarrow Definition I/1.5.1) of a given statistic (\rightarrow Definition “stat”) computed from this sample, e.g. a test statistic (\rightarrow Definition 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.

Metadata: ID: D140 | shortcut: dist-samp | author: JoramSoch | date: 2021-03-31, 09:43.

1.6 Probability functions

1.6.1 Probability mass function

Definition: Let X be a discrete (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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.

Metadata: ID: D9 | shortcut: pmf | author: JoramSoch | date: 2020-02-13, 19:09.

1.6.2 Probability mass function of sum of independents

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

Proof: Using the definition of the probability mass function (\rightarrow Definition I/1.6.1) and the expected value (\rightarrow Definition I/1.7.1), the first equation can be derived as follows:

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

Note that the third-last transition is justified by the fact that X and Y are independent (\rightarrow Definition I/1.3.6), such that conditional probabilities are equal to marginal probabilities (\rightarrow Proof I/1.3.8). 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>.

Metadata: ID: P257 | shortcut: pmf-sumind | author: JoramSoch | date: 2021-08-30, 09:14.

1.6.3 Probability mass function of strictly increasing function

Theorem: Let X be a discrete (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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 Definition 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 a strictly increasing function is invertible, the probability mass function (\rightarrow Definition 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 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>.

Metadata: ID: P184 | shortcut: pmf-sifct | author: JoramSoch | date: 2020-10-29, 05:55.

1.6.4 Probability mass function of strictly decreasing function

Theorem: Let X be a discrete (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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 Definition 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 a strictly decreasing function is invertible, the probability mass function (\rightarrow Definition 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 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>.

Metadata: ID: P187 | shortcut: pmf-sdfct | author: JoramSoch | date: 2020-11-06, 04:21.

1.6.5 Probability mass function of invertible function

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) of discrete random variables (\rightarrow Definition 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 Definition 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 Definition 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>.

Metadata: ID: P253 | shortcut: pmf-invft | author: JoramSoch | date: 2021-08-30, 05:13.

1.6.6 Probability density function

Definition: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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.

Metadata: ID: D10 | shortcut: pdf | author: JoramSoch | date: 2020-02-13, 19:26.

1.6.7 Probability density function of sum of independents

Theorem: Let X and Y be two independent (\rightarrow Definition I/1.3.6) continuous (\rightarrow Definition I/1.2.6) random variables (\rightarrow Definition I/1.2.2) with possible values \mathcal{X} and \mathcal{Y} and let $Z = X + Y$. Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition I/1.6.6) of X , Y and Z .

Proof: The cumulative distribution function of a sum of independent random variables (\rightarrow Proof I/1.6.14) 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 Proof I/1.6.12), 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>.

Metadata: ID: P258 | shortcut: pdf-sumind | author: JoramSoch | date: 2021-08-30, 09:31.

1.6.8 Probability density function of strictly increasing function

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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 Definition I/1.6.6) 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 Proof I/1.6.15) 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 Proof I/1.6.12)

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

the probability density function (\rightarrow Definition I/1.6.6) 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>.

Metadata: ID: P185 | shortcut: pdf-sifet | author: JoramSoch | date: 2020-10-29, 06:21.

1.6.9 Probability density function of strictly decreasing function

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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 Definition I/1.6.6) 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 Proof I/1.6.15) 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 Definition I/1.6.6) 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 Proof I/1.6.12)

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

the probability density function (\rightarrow Definition I/1.6.6) 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>.

Metadata: ID: P188 | shortcut: pdf-sdfct | author: JoramSoch | date: 2020-11-06, 05:30.

1.6.10 Probability density function of invertible function

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) of continuous random variables (\rightarrow Definition 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 Definition I/1.6.6) 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 Definition I/1.6.13) of $Y = g(X)$. The joint CDF (\rightarrow Definition I/1.6.22) 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

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

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 Definition I/1.6.6) of $Y = g(X)$. Because the PDF is the derivative of the CDF (\rightarrow Proof I/1.6.12), 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.
- 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.
- 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.

- 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>.

Metadata: ID: P254 | shortcut: pdf-invfc | author: JoramSoch | date: 2021-08-30, 07:05.

1.6.11 Probability density function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) of continuous random variables (\rightarrow Definition 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 Definition I/1.2.5) $n \times 1$ vector μ and constant (\rightarrow Definition I/1.2.5) $n \times n$ matrix Σ . Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Σ 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 Proof I/1.6.10) 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>.

Metadata: ID: P255 | shortcut: pdf-linfct | author: JoramSoch | date: 2021-08-30, 07:46.

1.6.12 Probability density function in terms of cumulative distribution function

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2). Then, the probability distribution function (\rightarrow Definition I/1.6.6) of X is the first derivative of the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof I/1.6.18) 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.

Metadata: ID: P191 | shortcut: pdf-cdf | author: JoramSoch | date: 2020-11-12, 07:19.

1.6.13 Cumulative distribution function

Definition: The cumulative distribution function (CDF) of a random variable (\rightarrow Definition I/1.2.2) X at a given value x is defined as the probability (\rightarrow Definition 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 Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and the probability mass function (\rightarrow Definition I/1.6.1) $f_X(x)$, then the cumulative distribution function is the function (\rightarrow Proof I/1.6.17) $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ 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 Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and the probability density function (\rightarrow Definition I/1.6.6) $f_X(x)$, then the cumulative distribution function is the function (\rightarrow Proof I/1.6.18) $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ 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.

Metadata: ID: D13 | shortcut: cdf | author: JoramSoch | date: 2020-02-17, 22:07.

1.6.14 Cumulative distribution function of sum of independents

Theorem: Let X and Y be two independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) and let $Z = X + Y$. Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Definition I/1.6.13) of X , Y and Z and $\mathbb{E}[\cdot]$ denotes the expected value (\rightarrow Definition I/1.7.1).

Proof: Using the definition of the cumulative distribution function (\rightarrow Definition I/1.6.13), 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 Definition I/1.3.6), such that conditional probabilities are equal to marginal probabilities (\rightarrow Proof I/1.3.8). 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>.

Metadata: ID: P256 | shortcut: cdf-sumind | author: JoramSoch | date: 2021-08-30, 08:53.

1.6.15 Cumulative distribution function of strictly increasing function

Theorem: Let X be a random variable (\rightarrow Definition 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 Definition I/1.6.13) 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 Definition I/1.6.13) of Y can be derived as follows:

1) If y is lower than the lowest value (\rightarrow Definition I/1.13.1) 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 Definition I/1.13.2) 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>.

Metadata: ID: P183 | shortcut: cdf-sifct | author: JoramSoch | date: 2020-10-29, 05:35.

1.6.16 Cumulative distribution function of strictly decreasing function

Theorem: Let X be a random variable (\rightarrow Definition 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 Definition I/1.6.13) 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 Definition I/1.6.13) of Y can be derived as follows:

1) If y is higher than the highest value (\rightarrow Definition I/1.13.2) 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 Definition I/1.13.1) 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>.

Metadata: ID: P186 | shortcut: cdf-sdfct | author: JoramSoch | date: 2020-11-06, 04:12.

1.6.17 Cumulative distribution function of discrete random variable

Theorem: Let X be a discrete (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with possible values \mathcal{X} and probability mass function (\rightarrow Definition I/1.6.1) $f_X(x)$. Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Definition I/1.6.13) of a random variable (\rightarrow Definition 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 Definition I/1.6.1) of a discrete (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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:

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

Sources:

- original work

Metadata: ID: P189 | shortcut: cdf-pmf | author: JoramSoch | date: 2020-11-12, 06:03.

1.6.18 Cumulative distribution function of continuous random variable

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with possible values \mathcal{X} and probability density function (\rightarrow Definition I/1.6.6) $f_X(x)$. Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) of X is

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

Proof: The cumulative distribution function (\rightarrow Definition I/1.6.13) of a random variable (\rightarrow Definition 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 Definition I/1.6.6) of a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition 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:

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

Sources:

- original work

Metadata: ID: P190 | shortcut: cdf-pdf | author: JoramSoch | date: 2020-11-12, 06:33.

1.6.19 Probability integral transform

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with invertible (\rightarrow Definition I/1.6.23) cumulative distribution function (\rightarrow Definition I/1.6.13) $F_X(x)$. Then, the random variable (\rightarrow Definition I/1.2.2)

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

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

Proof: The cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof II/3.1.4) with $a = 0$ and $b = 1$, i.e. the cumulative distribution function (\rightarrow Definition I/1.6.13) of the standard uniform distribution (\rightarrow Definition 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.

Metadata: ID: P220 | shortcut: cdf-pit | author: JoramSoch | date: 2021-04-07, 08:47.

1.6.20 Inverse transformation method

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

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

has a probability distribution (\rightarrow Definition I/1.5.1) characterized by the invertible (\rightarrow Definition I/1.6.23) cumulative distribution function (\rightarrow Definition I/1.6.13) $F_X(x)$.

Proof: The cumulative distribution function (\rightarrow Definition I/1.6.13) of the transformation $X = F_X^{-1}(U)$ can be derived as

$$\begin{aligned} \Pr(X \leq x) &= \Pr(F_X^{-1}(U) \leq x) \\ &= \Pr(U \leq F_X(x)) \\ &= F_X(x) , \end{aligned} \quad (2)$$

because the cumulative distribution function (\rightarrow Definition I/1.6.13) of the standard uniform distribution (\rightarrow Definition II/3.1.2) $\mathcal{U}(0, 1)$ is

$$U \sim \mathcal{U}(0, 1) \quad \Rightarrow \quad F_U(u) = \Pr(U \leq u) = u . \quad (3)$$

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.

Metadata: ID: P221 | shortcut: cdf-itm | author: JoramSoch | date: 2021-04-07, 08:47.

1.6.21 Distributional transformation

Theorem: Let X and Y be two continuous (\rightarrow Definition I/1.2.6) random variables (\rightarrow Definition I/1.2.2) with cumulative distribution function (\rightarrow Definition I/1.6.13) $F_X(x)$ and invertible cumulative distribution function (\rightarrow Definition I/1.6.13) $F_Y(y)$. Then, the random variable (\rightarrow Definition I/1.2.2)

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

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

Proof: The cumulative distribution function (\rightarrow Definition I/1.6.13) of the transformation $\tilde{X} = F_Y^{-1}(F_X(X))$ can be derived as

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

which shows that \tilde{X} and Y have the same cumulative distribution function (\rightarrow Definition I/1.6.13) and are thus identically distributed (\rightarrow Definition 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

Metadata: ID: P222 | shortcut: cdf-dt | author: JoramSoch | date: 2021-04-07, 09:19.

1.6.22 Joint cumulative distribution function

Definition: Let $X \in \mathbb{R}^{n \times 1}$ be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3). Then, the joint (\rightarrow Definition I/1.5.2) cumulative distribution function (\rightarrow Definition I/1.6.13) of X is defined as the probability (\rightarrow Definition I/1.3.1) that each entry X_i is smaller than a specific value x_i for $i = 1, \dots, n$:

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

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.

Metadata: ID: D141 | shortcut: cdf-joint | author: JoramSoch | date: 2020-04-07, 08:17.

1.6.23 Quantile function

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2) with the cumulative distribution function (\rightarrow Definition I/1.6.13) (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.

Metadata: ID: D14 | shortcut: qf | author: JoramSoch | date: 2020-02-17, 22:18.

1.6.24 Quantile function in terms of cumulative distribution function

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2) with the cumulative distribution function (\rightarrow Definition I/1.6.13) $F_X(x)$. If the cumulative distribution function is strictly monotonically increasing, then the quantile function (\rightarrow Definition I/1.6.23) is identical to the inverse of $F_X(x)$:

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

Proof: The quantile function (\rightarrow Definition I/1.6.23) $Q_X(p)$ 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

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

Sources:

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

Metadata: ID: P192 | shortcut: qf-cdf | author: JoramSoch | date: 2020-11-12, 07:48.

1.6.25 Characteristic function

Definition:

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

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

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

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

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

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

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>.

Metadata: ID: D159 | shortcut: cf | author: JoramSoch | date: 2021-09-22, 09:20.

1.6.26 Characteristic function of arbitrary function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) with the expected value (\rightarrow Definition I/1.7.1) function $E_X[\cdot]$. Then, the characteristic function (\rightarrow Definition I/1.6.25) of $Y = g(X)$ is equal to

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

Proof: The characteristic function (\rightarrow Definition I/1.6.25) is defined as

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

Due of the law of the unconscious statistician (\rightarrow Proof I/1.7.11)

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

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

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

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>.

Metadata: ID: P259 | shortcut: cf-fct | author: JoramSoch | date: 2021-09-22, 09:12.

1.6.27 Moment-generating function

Definition:

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

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

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

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

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.
- 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>.

Metadata: ID: D2 | shortcut: mgf | author: JoramSoch | date: 2020-01-22, 10:58.

1.6.28 Moment-generating function of arbitrary function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) with the expected value (\rightarrow Definition I/1.7.1) function $\mathbb{E}_X[\cdot]$. Then, the moment-generating function (\rightarrow Definition I/1.6.27) of $Y = g(X)$ is equal to

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

Proof: The moment-generating function (\rightarrow Definition I/1.6.27) is defined as

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

Due of the law of the unconscious statistician (\rightarrow Proof I/1.7.11)

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

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

$$M_Y(t) = \mathbb{E}_X [\exp(t g(X))] . \quad (4)$$

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>.

Metadata: ID: P260 | shortcut: mgf-fct | author: JoramSoch | date: 2021-09-22, 09:00.

1.6.29 Moment-generating function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) with the moment-generating function (\rightarrow Definition I/1.6.27) $M_X(t)$. 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 Definition I/1.6.27) X is

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

and therefore the moment-generating function of the random vector (\rightarrow Definition 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.

Metadata: ID: P154 | shortcut: mgf-ltt | author: JoramSoch | date: 2020-08-19, 08:09.

1.6.30 Moment-generating function of linear combination

Theorem: Let X_1, \dots, X_n be n independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) with moment-generating functions (\rightarrow Definition I/1.6.27) $M_{X_i}(t)$. 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 Definition I/1.6.27) X_i is

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

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

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

Because the expected value is multiplicative for independent random variables (\rightarrow Proof I/1.7.7), we have

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

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.

Metadata: ID: P155 | shortcut: mgf-lincomb | author: JoramSoch | date: 2020-08-19, 08:36.

1.6.31 Cumulant-generating function

Definition:

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

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

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

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

Sources:

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

Metadata: ID: D68 | shortcut: cgf | author: JoramSoch | date: 2020-05-31, 23:46.

1.6.32 Probability-generating function

Definition:

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

$$G_X(z) = E[z^X] = \sum_{x=0}^{\infty} p(x) z^x \quad (1)$$

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

2) If X is a discrete random vector (\rightarrow Definition 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

$$G_X(z) = E[z_1^{X_1} \cdot \dots \cdot z_n^{X_n}] = \sum_{x_1=0}^{\infty} \dots \sum_{x_n=0}^{\infty} p(x_1, \dots, x_n) z_1^{x_1} \cdot \dots \cdot z_n^{x_n} \quad (2)$$

where $z \in \mathbb{C}^n$ and $p(x)$ is the probability mass function (\rightarrow Definition 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.

Metadata: ID: D69 | shortcut: pgf | author: JoramSoch | date: 2020-05-31, 23:59.

1.7 Expected value

1.7.1 Definition

Definition:

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

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

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

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

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

where $f_X(x)$ is the probability density function (\rightarrow Definition I/1.6.6) 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.

Metadata: ID: D11 | shortcut: mean | author: JoramSoch | date: 2020-02-13, 19:38.

1.7.2 Sample mean

Definition: Let $x = \{x_1, \dots, x_n\}$ be a sample (\rightarrow Definition “samp”) from a random variable (\rightarrow Definition 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.

Metadata: ID: D142 | shortcut: mean-samp | author: JoramSoch | date: 2021-04-16, 11:53.

1.7.3 Non-negative random variable

Theorem: Let X be a non-negative random variable (\rightarrow Definition I/1.2.2). Then, the expected value (\rightarrow Definition I/1.7.1) 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 Definition I/1.6.13) of X .

Proof: Because the cumulative distribution function gives the probability of a random variable being smaller than a given value (\rightarrow Definition I/1.6.13),

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

we have

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

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 Definition I/1.6.6) of X , can be rewritten as

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

and by applying the definition of the expected value (\rightarrow Definition I/1.7.1), we see that

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

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>.

Metadata: ID: P103 | shortcut: mean-nnrvar | author: JoramSoch | date: 2020-05-18, 23:54.

1.7.4 Non-negativity

Theorem: If a random variable (\rightarrow Definition I/1.2.2) is strictly non-negative, its expected value (\rightarrow Definition I/1.7.1) is also non-negative, i.e.

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

Proof:

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

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

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 Definition I/1.6.6) is always non-negative, the integrand in

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

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.

Metadata: ID: P52 | shortcut: mean-nonneg | author: JoramSoch | date: 2020-02-13, 20:14.

1.7.5 Linearity

Theorem: The expected value (\rightarrow Definition I/1.7.1) is a linear operator, i.e.

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

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

Proof:

1) If X and Y are discrete random variables (\rightarrow Definition I/1.2.6), the expected value (\rightarrow Definition I/1.7.1) is

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

and the law of marginal probability (\rightarrow Definition 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 Definition I/1.2.6), the expected value (\rightarrow Definition I/1.7.1) is

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

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

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

Applying this, we have

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

as well as

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

Collectively, this shows that both requirements for linearity are fulfilled for the expected value (\rightarrow Definition I/1.7.1), for discrete (\rightarrow Definition I/1.2.6) as well as for continuous (\rightarrow Definition I/1.2.6) random variables.

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.
- 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/>.

Metadata: ID: P53 | shortcut: mean-lin | author: JoramSoch | date: 2020-02-13, 21:08.

1.7.6 Monotonicity

Theorem: The expected value (\rightarrow Definition I/1.7.1) is monotonic, i.e.

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

Proof: Let $Z = Y - X$. Due to the linearity of the expected value (\rightarrow Proof I/1.7.5), we have

$$E(Z) = E(Y - X) = E(Y) - E(X) . \tag{2}$$

With the non-negativity property of the expected value (\rightarrow Proof I/1.7.4), it also holds that

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

Together with (2), this yields

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

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.

Metadata: ID: P54 | shortcut: mean-mono | author: JoramSoch | date: 2020-02-17, 21:00.

1.7.7 (Non-)Multiplicativity**Theorem:**

1) If two random variables (\rightarrow Definition I/1.2.2) X and Y are independent (\rightarrow Definition I/1.3.6), the expected value (\rightarrow Definition I/1.7.1) is multiplicative, i.e.

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

2) If two random variables (\rightarrow Definition I/1.2.2) X and Y are dependent (\rightarrow Definition I/1.3.6), the expected value (\rightarrow Definition I/1.7.1) is not necessarily multiplicative, i.e. there exist X and Y such that

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

Proof:

1) If X and Y are independent (\rightarrow Definition 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 Definition I/1.7.1), we have

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

And applying it to the expected value for continuous random variables (\rightarrow Definition I/1.7.1), 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 Definition II/1.2.1) with the following joint probability (\rightarrow Definition I/1.3.2) mass function (\rightarrow Definition 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:

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

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.

Metadata: ID: P55 | shortcut: mean-mult | author: JoramSoch | date: 2020-02-17, 21:51.

1.7.8 Expectation of a trace

Theorem: Let A be an $n \times n$ random matrix (\rightarrow Definition I/1.2.4). Then, the expectation (\rightarrow Definition I/1.7.1) of the trace of A is equal to the trace of the expectation (\rightarrow Definition I/1.7.1) of A :

$$E[\text{tr}(A)] = \text{tr}(E[A]) . \quad (1)$$

Proof: The trace of an $n \times n$ matrix A is defined as:

$$\text{tr}(A) = \sum_{i=1}^n a_{ii} . \quad (2)$$

Using this definition of the trace, the linearity of the expected value (\rightarrow Proof I/1.7.5) and the expected value of a random matrix (\rightarrow Definition I/1.7.13), we have:

$$\begin{aligned} E[\text{tr}(A)] &= E\left[\sum_{i=1}^n a_{ii}\right] \\ &= \sum_{i=1}^n E[a_{ii}] \\ &= \text{tr}\left(\begin{bmatrix} E[a_{11}] & \dots & E[a_{1n}] \\ \vdots & \ddots & \vdots \\ E[a_{n1}] & \dots & E[a_{nn}] \end{bmatrix}\right) \\ &= \text{tr}(E[A]) . \end{aligned} \quad (3)$$

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>.

Metadata: ID: P298 | shortcut: mean-tr | author: JoramSoch | date: 2021-12-07, 09:03.

1.7.9 Expectation of a quadratic form

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) with mean (\rightarrow Definition I/1.7.1) μ and covariance (\rightarrow Definition I/1.9.1) Σ 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 1×1 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 Proof I/1.7.8), 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 Proof I/1.9.9)

$$\text{Cov}(X, X) = E(X X^T) - E(X)E(X)^T , \quad (5)$$

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 and covariance of X , we have

$$\begin{aligned} E [X^T A X] &= \text{tr} (A [\Sigma + \mu \mu^T]) \\ &= \text{tr} (A \Sigma + A \mu \mu^T) \\ &= \text{tr}(A \Sigma) + \text{tr}(A \mu \mu^T) \\ &= \text{tr}(A \Sigma) + \text{tr}(\mu^T A \mu) \\ &= \mu^T A \mu + \text{tr}(A \Sigma) . \end{aligned} \quad (7)$$

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.
- 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>.

Metadata: ID: P131 | shortcut: mean-qf | author: JoramSoch | date: 2020-07-13, 21:59.

1.7.10 Law of total expectation

Theorem: (law of total expectation, also called “law of iterated expectations”) Let X be a random variable (\rightarrow Definition I/1.2.2) with expected value (\rightarrow Definition I/1.7.1) $E(X)$ and let Y be any random variable (\rightarrow Definition I/1.8.1) defined on the same probability space (\rightarrow Definition I/1.1.4). Then, the expected value (\rightarrow Definition I/1.7.1) of the conditional expectation (\rightarrow Definition “mean-cond”) of X given Y is the same as the expected value (\rightarrow Definition I/1.7.1) of X :

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

Proof: Let X and Y be discrete random variables (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: P291 | shortcut: mean-tot | author: JoramSoch | date: 2021-11-26, 10:57.

1.7.11 Law of the unconscious statistician

Theorem: Let X be a random variable (\rightarrow Definition 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 Definition I/1.6.1) $f_X(x)$, the expected value (\rightarrow Definition I/1.7.1) 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 Definition I/1.6.6) $f_X(x)$, the expected value (\rightarrow Definition I/1.7.1) 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 Definition I/1.7.1) 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 Definition I/1.6.13) 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 Definition I/1.6.6) 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, substituing (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.
- 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>.

Metadata: ID: P138 | shortcut: mean-lotus | author: JoramSoch | date: 2020-07-22, 08:30.

1.7.12 Expected value of a random vector

Definition: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3). Then, the expected value (\rightarrow Definition I/1.7.1) of X is an $n \times 1$ 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.

Metadata: ID: D154 | shortcut: mean-rvec | author: JoramSoch | date: 2021-07-08, 08:34.

1.7.13 Expected value of a random matrix

Definition: Let X be an $n \times p$ random matrix (\rightarrow Definition I/1.2.4). Then, the expected value (\rightarrow Definition I/1.7.1) of X is an $n \times p$ matrix whose entries correspond to the expected values of the entries of the random matrix:

$$E(X) = E \left(\begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix} \right) = \begin{bmatrix} E(X_{11}) & \dots & E(X_{1p}) \\ \vdots & \ddots & \vdots \\ E(X_{n1}) & \dots & E(X_{np}) \end{bmatrix}. \quad (1)$$

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#hid13>.

Metadata: ID: D155 | shortcut: mean-rmat | author: JoramSoch | date: 2021-07-08, 08:42.

1.8 Variance

1.8.1 Definition

Definition: The variance of a random variable (\rightarrow Definition I/1.2.2) X is defined as the expected value (\rightarrow Definition I/1.7.1) of the squared deviation from its expected value (\rightarrow Definition I/1.7.1):

$$\text{Var}(X) = E[(X - E(X))^2]. \quad (1)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

Metadata: ID: D12 | shortcut: var | author: JoramSoch | date: 2020-02-13, 19:55.

1.8.2 Sample variance

Definition: Let $x = \{x_1, \dots, x_n\}$ be a sample (\rightarrow Definition “samp”) from a random variable (\rightarrow Definition I/1.2.2) X . Then, the sample variance of x is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1)$$

and the unbiased sample variance of x is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where \bar{x} is the sample mean (\rightarrow Definition I/1.7.2).

Sources:

- Wikipedia (2021): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-16; URL: https://en.wikipedia.org/wiki/Variance#Sample_variance.

Metadata: ID: D143 | shortcut: var-samp | author: JoramSoch | date: 2021-04-16, 12:04.

1.8.3 Partition into expected values

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, the variance (\rightarrow Definition I/1.8.1) of X is equal to the mean (\rightarrow Definition I/1.7.1) of the square of X minus the square of the mean (\rightarrow Definition I/1.7.1) of X :

$$\text{Var}(X) = E(X^2) - E(X)^2. \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.8.1) 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 Proof I/1.7.5), 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>.

Metadata: ID: P104 | shortcut: var-mean | author: JoramSoch | date: 2020-05-19, 00:17.

1.8.4 Non-negativity

Theorem: The variance (\rightarrow Definition I/1.8.1) is always non-negative, i.e.

$$\text{Var}(X) \geq 0 . \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.8.1) of a random variable (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: P123 | shortcut: var-nonneg | author: JoramSoch | date: 2020-06-06, 07:29.

1.8.5 Variance of a constant

Theorem: The variance (\rightarrow Definition I/1.8.1) of a constant (\rightarrow Definition I/1.2.5) is zero

$$a = \text{const.} \quad \Rightarrow \quad \text{Var}(a) = 0 \quad (1)$$

and if the variance (\rightarrow Definition I/1.8.1) of X is zero, then X is a constant (\rightarrow Definition I/1.2.5)

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

Proof:

1) A constant (\rightarrow Definition I/1.2.5) is defined as a quantity that always has the same value. Thus, if understood as a random variable (\rightarrow Definition I/1.2.2), the expected value (\rightarrow Definition I/1.7.1) of a constant is equal to itself:

$$\text{E}(a) = a . \quad (3)$$

Plugged into the formula of the variance (\rightarrow Definition I/1.8.1), we have

$$\begin{aligned}
\text{Var}(a) &= E[(a - E(a))^2] \\
&= E[(a - a)^2] \\
&= E(0) .
\end{aligned} \tag{4}$$

Applied to the formula of the expected value (\rightarrow Definition I/1.7.1), this gives

$$E(0) = \sum_{x=0} x \cdot f_X(x) = 0 \cdot 1 = 0 . \tag{5}$$

Together, (4) and (5) imply (1).

2) The variance (\rightarrow Definition I/1.8.1) is defined as

$$\text{Var}(X) = E[(X - E(X))^2] . \tag{6}$$

Because $(X - E(X))^2$ is strictly non-negative (\rightarrow Proof I/1.7.4), the only way for the variance to become zero is, if the squared deviation is always zero:

$$(X - E(X))^2 = 0 . \tag{7}$$

This, in turn, requires that X is equal to its expected value (\rightarrow Definition I/1.7.1)

$$X = E(X) \tag{8}$$

which can only be the case, if X always has the same value (\rightarrow Definition I/1.2.5):

$$X = \text{const.} \tag{9}$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-27; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P124 | shortcut: var-const | author: JoramSoch | date: 2020-06-27, 06:44.

1.8.6 Invariance under addition

Theorem: The variance (\rightarrow Definition I/1.8.1) is invariant under addition of a constant (\rightarrow Definition I/1.2.5):

$$\text{Var}(X + a) = \text{Var}(X) \tag{1}$$

Proof: The variance (\rightarrow Definition I/1.8.1) is defined in terms of the expected value (\rightarrow Definition I/1.7.1) as

$$\text{Var}(X) = E[(X - E(X))^2] . \tag{2}$$

Using this and the linearity of the expected value (\rightarrow Proof I/1.7.5), 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.

Metadata: ID: P126 | shortcut: var-inv | author: JoramSoch | date: 2020-07-07, 05:23.

1.8.7 Scaling upon multiplication

Theorem: The variance (\rightarrow Definition I/1.8.1) scales upon multiplication with a constant (\rightarrow Definition I/1.2.5):

$$\text{Var}(aX) = a^2 \text{Var}(X) \tag{1}$$

Proof: The variance (\rightarrow Definition I/1.8.1) is defined in terms of the expected value (\rightarrow Definition I/1.7.1) as

$$\text{Var}(X) = \text{E} [(X - \text{E}(X))^2] . \tag{2}$$

Using this and the linearity of the expected value (\rightarrow Proof I/1.7.5), 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.

Metadata: ID: P127 | shortcut: var-scal | author: JoramSoch | date: 2020-07-07, 05:38.

1.8.8 Variance of a sum

Theorem: The variance (\rightarrow Definition I/1.8.1) of the sum of two random variables (\rightarrow Definition I/1.2.2) equals the sum of the variances of those random variables, plus two times their covariance (\rightarrow Definition I/1.9.1):

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.8.1) is defined in terms of the expected value (\rightarrow Definition I/1.7.1) as

$$\text{Var}(X) = \text{E} [(X - \text{E}(X))^2] . \quad (2)$$

Using this and the linearity of the expected value (\rightarrow Proof I/1.7.5), we can derive (1) as follows:

$$\begin{aligned} \text{Var}(X + Y) &\stackrel{(2)}{=} \text{E} [((X + Y) - \text{E}(X + Y))^2] \\ &= \text{E} [(X - \text{E}(X)) + (Y - \text{E}(Y))]^2 \\ &= \text{E} [(X - \text{E}(X))^2 + (Y - \text{E}(Y))^2 + 2(X - \text{E}(X))(Y - \text{E}(Y))] \\ &= \text{E} [(X - \text{E}(X))^2] + \text{E} [(Y - \text{E}(Y))^2] + \text{E} [2(X - \text{E}(X))(Y - \text{E}(Y))] \\ &\stackrel{(2)}{=} \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \end{aligned} \quad (3)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P128 | shortcut: var-sum | author: JoramSoch | date: 2020-07-07, 06:10.

1.8.9 Variance of linear combination

Theorem: The variance (\rightarrow Definition I/1.8.1) of the linear combination of two random variables (\rightarrow Definition I/1.2.2) is a function of the variances as well as the covariance (\rightarrow Definition I/1.9.1) 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 Definition I/1.8.1) is defined in terms of the expected value (\rightarrow Definition I/1.7.1) as

$$\text{Var}(X) = \text{E} [(X - \text{E}(X))^2] . \quad (2)$$

Using this and the linearity of the expected value (\rightarrow Proof I/1.7.5), we can derive (1) as follows:

$$\begin{aligned}
\text{Var}(aX + bY) &\stackrel{(2)}{=} \mathbb{E} [((aX + bY) - \mathbb{E}(aX + bY))^2] \\
&= \mathbb{E} [(a[X - \mathbb{E}(X)] + b[Y - \mathbb{E}(Y)])^2] \\
&= \mathbb{E} [a^2 (X - \mathbb{E}(X))^2 + b^2 (Y - \mathbb{E}(Y))^2 + 2ab (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \quad (3) \\
&= \mathbb{E} [a^2 (X - \mathbb{E}(X))^2] + \mathbb{E} [b^2 (Y - \mathbb{E}(Y))^2] + \mathbb{E} [2ab (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&\stackrel{(2)}{=} a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) .
\end{aligned}$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P129 | shortcut: var-lincomb | author: JoramSoch | date: 2020-07-07, 06:21.

1.8.10 Additivity under independence

Theorem: The variance (\rightarrow Definition I/1.8.1) is additive for independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition 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 Proof I/1.8.8) 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 Proof I/1.9.4) 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.

Metadata: ID: P130 | shortcut: var-add | author: JoramSoch | date: 2020-07-07, 06:52.

1.8.11 Law of total variance

Theorem: (law of total variance, also called “conditional variance formula”) Let X and Y be random variables (\rightarrow Definition I/1.2.2) defined on the same probability space (\rightarrow Definition I/1.1.4) and assume that the variance (\rightarrow Definition I/1.8.1) of Y is finite. Then, the sum of the expectation (\rightarrow Definition I/1.7.1) of the conditional variance and the variance (\rightarrow Definition I/1.8.1) of the conditional expectation of Y given X is equal to the variance (\rightarrow Definition I/1.8.1) of Y :

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)] . \quad (1)$$

Proof: The variance can be decomposed into expected values (\rightarrow Proof I/1.8.3) as follows:

$$\text{Var}(Y) = E(Y^2) - E(Y)^2 . \quad (2)$$

This can be rearranged into:

$$E(Y^2) = \text{Var}(Y) + E(Y)^2 . \quad (3)$$

Applying the law of total expectation (\rightarrow Proof I/1.7.10), we have:

$$E(Y^2) = E [\text{Var}(Y|X) + E(Y|X)^2] . \quad (4)$$

Now subtract the second term from (2):

$$E(Y^2) - E(Y)^2 = E [\text{Var}(Y|X) + E(Y|X)^2] - E(Y)^2 . \quad (5)$$

Again applying the law of total expectation (\rightarrow Proof I/1.7.10), 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 Proof I/1.7.5), 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 Proof I/1.8.3), 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.

Metadata: ID: P292 | shortcut: var-tot | author: JoramSoch | date: 2021-11-26, 11:20.

1.8.12 Precision

Definition: The precision of a random variable (\rightarrow Definition I/1.2.2) X is defined as the inverse of the variance (\rightarrow Definition I/1.8.1), i.e. one divided by the expected value (\rightarrow Definition I/1.7.1) of the squared deviation from its expected value (\rightarrow Definition I/1.7.1):

$$\text{Prec}(X) = \text{Var}(X)^{-1} = \frac{1}{E [(X - E(X))^2]} . \quad (1)$$

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)).

Metadata: ID: D145 | shortcut: prec | author: JoramSoch | date: 2020-04-21, 07:04.

1.9 Covariance

1.9.1 Definition

Definition: The covariance of two random variables (\rightarrow Definition I/1.2.2) X and Y is defined as the expected value (\rightarrow Definition I/1.7.1) of the product of their deviations from their individual expected values (\rightarrow Definition I/1.7.1):

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (1)$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

Metadata: ID: D70 | shortcut: cov | author: JoramSoch | date: 2020-06-02, 20:20.

1.9.2 Sample covariance

Definition: Let $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ be samples (\rightarrow Definition “samp”) from random variables (\rightarrow Definition 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 Definition I/1.7.2).

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.

Metadata: ID: D144 | shortcut: cov-samp | author: ciaranmci | date: 2021-04-21, 06:53.

1.9.3 Partition into expected values

Theorem: Let X and Y be random variables (\rightarrow Definition I/1.2.2). Then, the covariance (\rightarrow Definition I/1.9.1) of X and Y is equal to the mean (\rightarrow Definition I/1.7.1) of the product of X and Y minus the product of the means (\rightarrow Definition I/1.7.1) of X and Y :

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) . \quad (1)$$

Proof: The covariance (\rightarrow Definition I/1.9.1) 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 Proof I/1.7.5), can be rewritten as

$$\begin{aligned}
 \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
 &= E[XY - X E(Y) - E(X) Y + E(X)E(Y)] \\
 &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\
 &= E(XY) - E(X)E(Y) .
 \end{aligned} \tag{3}$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-02; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

Metadata: ID: P118 | shortcut: cov-mean | author: JoramSoch | date: 2020-06-02, 20:50.

1.9.4 Covariance under independence

Theorem: Let X and Y be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2). Then, the covariance (\rightarrow Definition I/1.9.1) 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 Proof I/1.9.3) as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) . \tag{2}$$

For independent random variables, the expected value of the product is equal to the product of the expected values (\rightarrow Proof I/1.7.7):

$$E(XY) = E(X)E(Y) . \tag{3}$$

Taking (2) and (3) together, we have

$$\begin{aligned}
 \text{Cov}(X, Y) &\stackrel{(2)}{=} E(XY) - E(X)E(Y) \\
 &\stackrel{(3)}{=} E(X)E(Y) - E(X)E(Y) \\
 &= 0 .
 \end{aligned} \tag{4}$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence.

Metadata: ID: P158 | shortcut: cov-ind | author: JoramSoch | date: 2020-09-03, 06:05.

1.9.5 Relationship to correlation

Theorem: Let X and Y be random variables (\rightarrow Definition I/1.2.2). Then, the covariance (\rightarrow Definition I/1.9.1) of X and Y is equal to the product of their correlation (\rightarrow Definition I/1.10.1) and the standard deviations (\rightarrow Definition I/1.12.1) of X and Y :

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y . \quad (1)$$

Proof: The correlation (\rightarrow Definition I/1.10.1) 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 Definition I/1.9.1) to give

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y \quad (3)$$

Sources:

- original work

Metadata: ID: P119 | shortcut: cov-corr | author: JoramSoch | date: 2020-06-02, 21:00.

1.9.6 Law of total covariance

Theorem: (law of total covariance, also called “conditional covariance formula”) Let X , Y and Z be random variables (\rightarrow Definition I/1.2.2) defined on the same probability space (\rightarrow Definition I/1.1.4) and assume that the covariance (\rightarrow Definition I/1.9.1) of X and Y is finite. Then, the sum of the expectation (\rightarrow Definition I/1.7.1) of the conditional covariance and the covariance (\rightarrow Definition I/1.9.1) of the conditional expectations of X and Y given Z is equal to the covariance (\rightarrow Definition I/1.9.1) 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 Proof I/1.9.3) 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 Proof I/1.7.10), 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 Proof I/1.9.3) 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 Proof I/1.7.5), 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 Proof I/1.9.3), 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.

Metadata: ID: P293 | shortcut: cov-tot | author: JoramSoch | date: 2021-11-26, 11:38.

1.9.7 Covariance matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition 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 Definition I/1.9.1) of X_i and X_j :

$$\Sigma_{XX} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \text{E}[(X_1 - \text{E}[X_1])(X_1 - \text{E}[X_1])] & \dots & \text{E}[(X_1 - \text{E}[X_1])(X_n - \text{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \text{E}[(X_n - \text{E}[X_n])(X_1 - \text{E}[X_1])] & \dots & \text{E}[(X_n - \text{E}[X_n])(X_n - \text{E}[X_n])] \end{bmatrix} \quad (1)$$

Sources:

- Wikipedia (2020): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Definition.

Metadata: ID: D72 | shortcut: covmat | author: JoramSoch | date: 2020-06-06, 04:24.

1.9.8 Sample covariance matrix

Definition: Let $x = \{x_1, \dots, x_n\}$ be a sample (\rightarrow Definition “samp”) from a random vector (\rightarrow Definition 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 Definition I/1.7.2).

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.

Metadata: ID: D153 | shortcut: covmat-samp | author: JoramSoch | date: 2021-05-20, 07:46.

1.9.9 Covariance matrix and expected values

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3). Then, the covariance matrix (\rightarrow Definition I/1.9.7) of X is equal to the mean (\rightarrow Definition I/1.7.1) of the outer product of X with itself minus the outer product of the mean (\rightarrow Definition I/1.7.1) of X with itself:

$$\Sigma_{XX} = E(XX^T) - E(X)E(X)^T. \quad (1)$$

Proof: The covariance matrix (\rightarrow Definition I/1.9.7) 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 Proof I/1.7.5), 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>.

Metadata: ID: P120 | shortcut: covmat-mean | author: JoramSoch | date: 2020-06-06, 05:31.

1.9.10 Covariance matrix and correlation matrix

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3). Then, the covariance matrix (\rightarrow Definition I/1.9.7) of X can be expressed in terms of its correlation matrix (\rightarrow Definition I/1.10.5) as follows

$$\Sigma_{XX} = D_X \cdot C_{XX} \cdot D_X, \quad (1)$$

where D_X is a diagonal matrix with the standard deviations (\rightarrow Definition I/1.12.1) of X_1, \dots, X_n as entries on the diagonal:

$$D_X = \text{diag}(\sigma_{X_1}, \dots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (2)$$

Proof: Reiterating (1) and applying (2), we have:

$$\Sigma_{XX} = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot C_{XX} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (3)$$

Together with the definition of the correlation matrix (\rightarrow Definition I/1.10.5), this gives

$$\begin{aligned} \Sigma_{XX} &= \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \\ &= \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix} \end{aligned} \quad (4)$$

which is nothing else than the definition of the covariance matrix (\rightarrow Definition I/1.9.7).

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.

Metadata: ID: P121 | shortcut: covmat-corrmat | author: JoramSoch | date: 2020-06-06, 06:02.

1.9.11 Precision matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition I/1.2.3). Then, the precision matrix of X is defined as the inverse of the covariance matrix (\rightarrow Definition I/1.9.7) of X :

$$\Lambda_{XX} = \Sigma_{XX}^{-1} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix}^{-1}. \quad (1)$$

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)).

Metadata: ID: D74 | shortcut: precmat | author: JoramSoch | date: 2020-06-06, 05:08.

1.9.12 Precision matrix and correlation matrix

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3). Then, the precision matrix (\rightarrow Definition I/1.9.11) of X can be expressed in terms of its correlation matrix (\rightarrow Definition I/1.10.5) as follows

$$\Lambda_{XX} = D_X^{-1} \cdot C_{XX}^{-1} \cdot D_X^{-1}, \quad (1)$$

where D_X^{-1} is a diagonal matrix with the inverse standard deviations (\rightarrow Definition I/1.12.1) of X_1, \dots, X_n as entries on the diagonal:

$$D_X^{-1} = \text{diag}(1/\sigma_{X_1}, \dots, 1/\sigma_{X_n}) = \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix}. \quad (2)$$

Proof: The precision matrix (\rightarrow Definition I/1.9.11) is defined as the inverse of the covariance matrix (\rightarrow Definition I/1.9.7)

$$\Lambda_{XX} = \Sigma_{XX}^{-1} \quad (3)$$

and the relation between covariance matrix and correlation matrix (\rightarrow Proof I/1.9.10) 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

$$\begin{aligned} \Lambda_{XX} &\stackrel{(3)}{=} \Sigma_{XX}^{-1} \\ &\stackrel{(4)}{=} (\mathbf{D}_X \cdot \mathbf{C}_{XX} \cdot \mathbf{D}_X)^{-1} \\ &\stackrel{(6)}{=} \mathbf{D}_X^{-1} \cdot \mathbf{C}_{XX}^{-1} \cdot \mathbf{D}_X^{-1} \\ &\stackrel{(7)}{=} \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \cdot \mathbf{C}_{XX}^{-1} \cdot \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \end{aligned} \quad (8)$$

which conforms to equation (1).

Sources:

- original work

Metadata: ID: P122 | shortcut: precmat-corrmat | author: JoramSoch | date: 2020-06-06, 06:28.

1.10 Correlation

1.10.1 Definition

Definition: The correlation of two random variables (\rightarrow Definition I/1.2.2) X and Y , also called Pearson product-moment correlation coefficient (PPMCC), is defined as the ratio of the covariance (\rightarrow Definition I/1.9.1) of X and Y relative to the product of their standard deviations (\rightarrow Definition I/1.12.1):

$$\text{Corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]} \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]}}. \quad (1)$$

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.

Metadata: ID: D71 | shortcut: corr | author: JoramSoch | date: 2020-06-02, 20:34.

1.10.2 Range

Theorem: Let X and Y be two random variables (\rightarrow Definition 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 Definition I/1.8.1) of X plus or minus Y , each divided by their standard deviations (\rightarrow Definition I/1.12.1):

$$\text{Var} \left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) . \quad (2)$$

Because the variance is non-negative (\rightarrow Proof I/1.8.4), 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 Proof I/1.8.9), 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 Proof I/1.9.5), 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>.

Metadata: ID: P300 | shortcut: corr-range | author: JoramSoch | date: 2021-12-14, 02:08.

1.10.3 Sample correlation coefficient

Definition: Let $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ be samples (\rightarrow Definition “samp”) from random variables (\rightarrow Definition 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 Definition I/1.7.2).

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.

Metadata: ID: D168 | shortcut: corr-samp | author: JoramSoch | date: 2021-12-14, 07:23.

1.10.4 Relationship to standard scores

Theorem: Let $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ be samples (\rightarrow Definition “samp”) from random variables (\rightarrow Definition I/1.2.2) X and Y . Then, the sample correlation coefficient (\rightarrow Definition I/1.10.3) r_{xy} can be expressed in terms of the standard scores (\rightarrow Definition “z”) 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 Definition I/1.7.2) and s_x and s_y are the sample variances (\rightarrow Definition I/1.8.2).

Proof: The sample correlation coefficient (\rightarrow Definition I/1.10.3) 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 Definition I/1.8.2) 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): “Peason correlation coefficient”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-14; URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#For_a_sample.

Metadata: ID: P299 | shortcut: corr-z | author: JoramSoch | date: 2021-12-14, 02:31.

1.10.5 Correlation matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition 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 Definition I/1.10.1) of X_i and X_j :

$$C_{XX} = \begin{bmatrix} \text{Corr}(X_1, X_1) & \dots & \text{Corr}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Corr}(X_n, X_1) & \dots & \text{Corr}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \frac{E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix}. \quad (1)$$

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.

Metadata: ID: D73 | shortcut: corrmatrix | author: JoramSoch | date: 2020-06-06, 04:56.

1.10.6 Sample correlation matrix

Definition: Let $x = \{x_1, \dots, x_n\}$ be a sample (\rightarrow Definition “samp”) from a random vector (\rightarrow Definition 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 Definition I/1.10.3) 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 Definition I/1.10.3) 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 Definition I/1.7.2)

$$\begin{aligned} \bar{x}^{(j)} &= \frac{1}{n} \sum_{i=1}^n x_{ij} \\ \bar{x}^{(k)} &= \frac{1}{n} \sum_{i=1}^n x_{ik} . \end{aligned} \quad (3)$$

Sources:

- original work

Metadata: ID: D169 | shortcut: corrmatt-samp | author: JoramSoch | date: 2021-12-14, 07:45.

1.11 Measures of central tendency

1.11.1 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 Definition “samp”) from a random variable (\rightarrow Definition 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 Definition I/1.2.2) with cumulative distribution function (\rightarrow Definition I/1.6.13) $F_X(x)$. 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>.

Metadata: ID: D101 | shortcut: med | author: JoramSoch | date: 2020-10-15, 10:53.

1.11.2 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 Definition “samp”) from a random variable (\rightarrow Definition 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 Definition I/1.2.2) with probability mass function (\rightarrow Definition I/1.6.1) or probability density function (\rightarrow Definition I/1.6.6) $f_X(x)$. Then, the mode of X is the 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)).

Metadata: ID: D102 | shortcut: mode | author: JoramSoch | date: 2020-10-15, 11:10.

1.12 Measures of statistical dispersion

1.12.1 Standard deviation

Definition: The standard deviation σ of a random variable (\rightarrow Definition I/1.2.2) X with expected value (\rightarrow Definition I/1.7.1) μ is defined as the square root of the variance (\rightarrow Definition I/1.8.1), i.e.

$$\sigma(X) = \sqrt{\mathbb{E}[(X - \mu)^2]} . \quad (1)$$

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.

Metadata: ID: D94 | shortcut: std | author: JoramSoch | date: 2020-09-03, 05:43.

1.12.2 Full width at half maximum

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.2.2) with a unimodal probability density function (\rightarrow Definition I/1.6.6) $f_X(x)$ and mode (\rightarrow Definition I/1.11.2) 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.

Metadata: ID: D91 | shortcut: fwhm | author: JoramSoch | date: 2020-08-19, 05:40.

1.13 Further summary statistics

1.13.1 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 Definition “samp”) from a random variable (\rightarrow Definition 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 Definition 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.

Metadata: ID: D107 | shortcut: min | author: JoramSoch | date: 2020-11-12, 05:25.

1.13.2 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 Definition “samp”) from a random variable (\rightarrow Definition 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 Definition 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.

Metadata: ID: D108 | shortcut: max | author: JoramSoch | date: 2020-11-12, 05:33.

1.14 Further moments

1.14.1 Moment

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2), let c be a constant (\rightarrow Definition 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 Definition I/1.7.1) 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 Definition I/1.14.3) $\mu'_n = \mu_n(0)$;
- the n -th central moment (\rightarrow Definition I/1.14.6) $\mu_n = \mu_n(\mu)$;
- the n -th standardized moment (\rightarrow Definition I/1.14.9) $\mu_n^* = \mu_n/\sigma^n$.

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).

Metadata: ID: D90 | shortcut: mom | author: JoramSoch | date: 2020-08-19, 05:24.

1.14.2 Moment in terms of moment-generating function

Theorem: Let X be a scalar random variable (\rightarrow Definition I/1.2.2) with the moment-generating function (\rightarrow Definition I/1.6.27) $M_X(t)$. Then, the n -th raw moment (\rightarrow Definition I/1.14.3) 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 Definition I/1.6.27), 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 Proof I/1.7.5), 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

$$m^{\underline{n}} = \prod_{i=0}^{n-1} (m - i) = \frac{m!}{(m - n)!} , \quad (7)$$

equation (5) becomes

$$\begin{aligned} M_X^{(n)}(t) &= \sum_{m=n}^{\infty} \frac{m^{\underline{n}} t^{m-n}}{m!} E(X^m) \\ &\stackrel{(7)}{=} \sum_{m=n}^{\infty} \frac{m! t^{m-n}}{(m - n)! m!} E(X^m) \\ &= \sum_{m=n}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= \frac{t^{n-n}}{(n - n)!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= \frac{t^0}{0!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) \\ &= E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m - n)!} E(X^m) . \end{aligned} \quad (8)$$

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.

Metadata: ID: P153 | shortcut: mom-mgf | author: JoramSoch | date: 2020-08-19, 07:51.

1.14.3 Raw moment

Definition: Let X be a random variable (\rightarrow Definition 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 Definition I/1.14.1) 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).

Metadata: ID: D97 | shortcut: mom-row | author: JoramSoch | date: 2020-10-08, 03:31.

1.14.4 First raw moment is mean

Theorem: The first raw moment (\rightarrow Definition I/1.14.3) equals the mean (\rightarrow Definition I/1.7.1), i.e.

$$\mu'_1 = \mu . \quad (1)$$

Proof: The first raw moment (\rightarrow Definition I/1.14.3) of a random variable (\rightarrow Definition I/1.2.2) X is defined as

$$\mu'_1 = E[(X - 0)^1] \quad (2)$$

which is equal to the expected value (\rightarrow Definition I/1.7.1) of X :

$$\mu'_1 = E[X] = \mu . \quad (3)$$

Sources:

- original work

Metadata: ID: P171 | shortcut: momraw-1st | author: JoramSoch | date: 2020-10-08, 04:19.

1.14.5 Second raw moment and variance

Theorem: The second raw moment (\rightarrow Definition I/1.14.3) can be expressed as

$$\mu'_2 = \text{Var}(X) + E(X)^2 \quad (1)$$

where $\text{Var}(X)$ is the variance (\rightarrow Definition I/1.8.1) of X and $E(X)$ is the expected value (\rightarrow Definition I/1.7.1) of X .

Proof: The second raw moment (\rightarrow Definition I/1.14.3) of a random variable (\rightarrow Definition 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 Proof I/1.8.3)

$$\text{Var}(X) = E(X^2) - E(X)^2 , \quad (3)$$

the second raw moment can be rearranged into:

$$\mu'_2 \stackrel{(2)}{=} E(X^2) \stackrel{(3)}{=} \text{Var}(X) + E(X)^2 . \quad (4)$$

Sources:

- original work

Metadata: ID: P172 | shortcut: momraw-2nd | author: JoramSoch | date: 2020-10-08, 05:05.

1.14.6 Central moment

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2) with expected value (\rightarrow Definition I/1.7.1) μ and let n be a positive integer. Then, the n -th central moment of X is defined as the n -th moment (\rightarrow Definition I/1.14.1) of X about the value μ :

$$\mu_n = E[(X - \mu)^n] . \quad (1)$$

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).

Metadata: ID: D98 | shortcut: mom-cent | author: JoramSoch | date: 2020-10-08, 03:37.

1.14.7 First central moment is zero

Theorem: The first central moment (\rightarrow Definition I/1.14.6) is zero, i.e.

$$\mu_1 = 0 . \quad (1)$$

Proof: The first central moment (\rightarrow Definition I/1.14.6) of a random variable (\rightarrow Definition I/1.2.2) X with mean (\rightarrow Definition I/1.7.1) μ is defined as

$$\mu_1 = E[(X - \mu)^1] . \quad (2)$$

Due to the linearity of the expected value (\rightarrow Proof I/1.7.5) 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.

Metadata: ID: P167 | shortcut: momcent-1st | author: JoramSoch | date: 2020-09-09, 07:51.

1.14.8 Second central moment is variance

Theorem: The second central moment (\rightarrow Definition I/1.14.6) equals the variance (\rightarrow Definition I/1.8.1), i.e.

$$\mu_2 = \text{Var}(X) . \quad (1)$$

Proof: The second central moment (\rightarrow Definition I/1.14.6) of a random variable (\rightarrow Definition I/1.2.2) X with mean (\rightarrow Definition I/1.7.1) μ is defined as

$$\mu_2 = \text{E} [(X - \mu)^2] \quad (2)$$

which is equivalent to the definition of the variance (\rightarrow Definition I/1.8.1):

$$\mu_2 = \text{E} [(X - \text{E}(X))^2] = \text{Var}(X) . \quad (3)$$

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).

Metadata: ID: P173 | shortcut: momcent-2nd | author: JoramSoch | date: 2020-10-08, 05:13.

1.14.9 Standardized moment

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2) with expected value (\rightarrow Definition I/1.7.1) μ and standard deviation (\rightarrow Definition I/1.12.1) σ and let n be a positive integer. Then, the n -th standardized moment of X is defined as the n -th moment (\rightarrow Definition I/1.14.1) of X about the value μ , divided by the n -th power of σ :

$$\mu_n^* = \frac{\mu_n}{\sigma^n} = \frac{\text{E}[(X - \mu)^n]}{\sigma^n} . \quad (1)$$

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).

Metadata: ID: D99 | shortcut: mom-stand | author: JoramSoch | date: 2020-10-08, 03:47.

2 Information theory

2.1 Shannon entropy

2.1.1 Definition

Definition: Let X be a discrete random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and the (observed or assumed) probability mass function (\rightarrow Definition 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.

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.

Metadata: ID: D15 | shortcut: ent | author: JoramSoch | date: 2020-02-19, 17:36.

2.1.2 Non-negativity

Theorem: The entropy of a discrete random variable (\rightarrow Definition I/1.2.2) is a non-negative number:

$$H(X) \geq 0 . \quad (1)$$

Proof: The entropy of a discrete random variable (\rightarrow Definition 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 Definition 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>.

Metadata: ID: P57 | shortcut: ent-nonneg | author: JoramSoch | date: 2020-02-19, 19:10.

2.1.3 Concavity

Theorem: The entropy (\rightarrow Definition I/2.1.1) is concave in the probability mass function (\rightarrow Definition 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 Definition I/1.2.2) with possible outcomes \mathcal{X} and let $u(x)$ be the probability mass function (\rightarrow Definition I/1.6.1) of a discrete uniform distribution (\rightarrow Definition II/1.1.1) on $X \in \mathcal{X}$. Then, the entropy (\rightarrow Definition I/2.1.1) of an arbitrary probability mass function (\rightarrow Definition I/1.6.1) $p(x)$ can be rewritten as

$$\begin{aligned} H[p] &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} u(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \cdot \log u(x) \\ &= -\text{KL}[p||u] - \log \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \\ &= \log |\mathcal{X}| - \text{KL}[p||u] \\ \log |\mathcal{X}| - H[p] &= \text{KL}[p||u] \end{aligned} \quad (2)$$

where we have applied the definition of the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1), the probability mass function of the discrete uniform distribution (\rightarrow Proof II/1.1.2) and the total sum over the probability mass function (\rightarrow Definition I/1.6.1).

Note that the KL divergence is convex (\rightarrow Proof I/2.5.5) in the pair of probability distributions (\rightarrow Definition 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

$$\begin{aligned}
\log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \lambda (\log |\mathcal{X}| - H[p_1]) + (1 - \lambda) (\log |\mathcal{X}| - H[p_2]) \\
\log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \log |\mathcal{X}| - \lambda H[p_1] - (1 - \lambda)H[p_2] \\
-H[\lambda p_1 + (1 - \lambda)p_2] &\leq -\lambda H[p_1] - (1 - \lambda)H[p_2] \\
H[\lambda p_1 + (1 - \lambda)p_2] &\geq \lambda H[p_1] + (1 - \lambda)H[p_2]
\end{aligned} \tag{5}$$

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>.

Metadata: ID: P149 | shortcut: ent-conc | author: JoramSoch | date: 2020-08-11, 08:29.

2.1.4 Conditional entropy

Definition: Let X and Y be discrete random variables (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and \mathcal{Y} and probability mass functions (\rightarrow Definition 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

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) \cdot H(Y|X = x) \tag{1}$$

where $H(Y|X = x)$ is the (marginal) entropy (\rightarrow Definition 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>.

Metadata: ID: D17 | shortcut: ent-cond | author: JoramSoch | date: 2020-02-19, 18:08.

2.1.5 Joint entropy

Definition: Let X and Y be discrete random variables (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition I/1.3.2) mass function (\rightarrow Definition I/1.6.1) $p(x, y)$. Then, the joint entropy of X and Y is defined as

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \tag{1}$$

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>.

Metadata: ID: D18 | shortcut: ent-joint | author: JoramSoch | date: 2020-02-19, 18:18.

2.1.6 Cross-entropy

Definition: Let X be a discrete random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.5.1) on X with the probability mass functions (\rightarrow Definition 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.

Metadata: ID: D85 | shortcut: ent-cross | author: JoramSoch | date: 2020-07-28, 02:51.

2.1.7 Convexity of cross-entropy

Theorem: The cross-entropy (\rightarrow Definition I/2.1.6) is convex in the probability distribution (\rightarrow Definition I/1.5.1) q , 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 distributions and $0 \leq \lambda \leq 1$.

Proof: The relationship between Kullback-Leibler divergence, entropy and cross-entropy (\rightarrow Proof I/2.5.8) is:

$$\text{KL}[P||Q] = H(P, Q) - H(P) . \quad (2)$$

Note that the KL divergence is convex (\rightarrow Proof I/2.5.5) in the pair of probability distributions (\rightarrow Definition 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.
- 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>.

Metadata: ID: P150 | shortcut: entcross-conv | author: JoramSoch | date: 2020-08-11, 09:16.

2.1.8 Gibbs’ inequality

Theorem: Let X be a discrete random variable (\rightarrow Definition I/1.2.2) and consider two probability distributions (\rightarrow Definition I/1.5.1) with probability mass functions (\rightarrow Definition I/1.6.1) $p(x)$ and $q(x)$. Then, Gibbs’ inequality states that the entropy (\rightarrow Definition I/2.1.1) of X according to P is smaller than or equal to the cross-entropy (\rightarrow Definition 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

$$\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} \geq 0 . \quad (3)$$

Because $\ln x \leq x - 1$, i.e. $-\ln x \geq 1 - x$, for all $x > 0$, with equality only if $x = 1$, we can say about the left-hand side that

$$\begin{aligned} \sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) \left(1 - \frac{p(x)}{q(x)} \right) \\ &= \sum_{x \in I} p(x) - \sum_{x \in I} q(x) . \end{aligned} \quad (4)$$

Finally, since $p(x)$ and $q(x)$ are probability mass functions (\rightarrow Definition I/1.6.1), we have

$$\begin{aligned}
0 \leq p(x) \leq 1, \quad \sum_{x \in I} p(x) &= 1 \quad \text{and} \\
0 \leq q(x) \leq 1, \quad \sum_{x \in I} q(x) &\leq 1,
\end{aligned} \tag{5}$$

such that it follows from (4) that

$$\begin{aligned}
\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \\
&= 1 - \sum_{x \in I} q(x) \geq 0.
\end{aligned} \tag{6}$$

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.

Metadata: ID: P164 | shortcut: gibbs-ineq | author: JoramSoch | date: 2020-09-09, 02:18.

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

$$\sum_{i=1}^n a_i \log_c \frac{a_i}{b_i} \geq a \log_c \frac{a}{b}. \tag{1}$$

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:

$$\log_c a = \frac{\ln a}{\ln c}. \tag{2}$$

Let $f(x) = x \ln x$. Then, the left-hand side of (1) can be rewritten as

$$\begin{aligned}
\sum_{i=1}^n a_i \ln \frac{a_i}{b_i} &= \sum_{i=1}^n b_i f\left(\frac{a_i}{b_i}\right) \\
&= b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right).
\end{aligned} \tag{3}$$

Because $f(x)$ is a convex function and

$$\begin{aligned}
\frac{b_i}{b} &\geq 0 \\
\sum_{i=1}^n \frac{b_i}{b} &= 1,
\end{aligned} \tag{4}$$

applying Jensen's inequality yields

$$\begin{aligned}
 b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) &\geq b f\left(\sum_{i=1}^n \frac{b_i}{b} \frac{a_i}{b_i}\right) \\
 &= b f\left(\frac{1}{b} \sum_{i=1}^n a_i\right) \\
 &= b f\left(\frac{a}{b}\right) \\
 &= a \ln \frac{a}{b}.
 \end{aligned} \tag{5}$$

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.
- Wikipedia (2020): “Jensen's inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Jensen%27s_inequality#Statements.

Metadata: ID: P165 | shortcut: logsum-ineq | author: JoramSoch | date: 2020-09-09, 02:46.

2.2 Differential entropy

2.2.1 Definition

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and the (estimated or assumed) probability density function (\rightarrow Definition I/1.6.6) $p(x) = f_X(x)$. Then, the differential entropy (also referred to as “continuous entropy”) of X is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx \tag{1}$$

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>.

Metadata: ID: D16 | shortcut: dent | author: JoramSoch | date: 2020-02-19, 17:53.

2.2.2 Negativity

Theorem: Unlike its discrete analogue (\rightarrow Proof I/2.1.2), the differential entropy (\rightarrow Definition I/2.2.1) can become negative.

Proof: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition 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 Proof 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 Definition 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.

Metadata: ID: P68 | shortcut: dent-neg | author: JoramSoch | date: 2020-03-02, 20:32.

2.2.3 Invariance under addition

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2). Then, the differential entropy (\rightarrow Definition 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 Definition 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 Definition I/1.6.6) 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 Proof I/1.6.8) 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.

Metadata: ID: P199 | shortcut: dent-inv | author: JoramSoch | date: 2020-12-02, 16:11.

2.2.4 Addition upon multiplication

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random variable (\rightarrow Definition I/1.2.2). Then, the differential entropy (\rightarrow Definition 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 Definition 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 Definition I/1.6.6) 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 Proof I/1.6.8) 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 Proof I/1.6.9) 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.

Metadata: ID: P200 | shortcut: dent-add | author: JoramSoch | date: 2020-12-02, 16:39.

2.2.5 Addition upon matrix multiplication

Theorem: Let X be a continuous (\rightarrow Definition I/1.2.6) random vector (\rightarrow Definition I/1.2.3). Then, the differential entropy (\rightarrow Definition 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 Definition 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 Definition I/1.6.6) 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 Proof I/1.6.11) $Y = g(X) = \Sigma X + \mu$ 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 Definition I/1.6.6) 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 Definition 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 Definition I/1.6.6) that $\int_{\mathcal{X}} f_X(x) \, dx = 1$, we can derive the differential entropy (\rightarrow Definition 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:

- 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.

Metadata: ID: P261 | shortcut: dent-addvec | author: JoramSoch | date: 2021-10-07, 09:10.

2.2.6 Non-invariance and transformation

Theorem: The differential entropy (\rightarrow Definition 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 $g : X \rightarrow Y$ 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} is the set of possible values of X .

Proof: By definition, the differential entropy (\rightarrow Definition 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 Definition I/1.6.6) of X .

The probability density function of an invertible function of a continuous random vector (\rightarrow Proof I/1.6.10) $Y = g(X)$ 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 Definition 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

$$\begin{aligned}
h(Y) &= - \int_{\mathcal{X}} [f_X(g^{-1}(g(x))) \mid J_{g^{-1}}(y)] \log [f_X(g^{-1}(g(x))) \mid J_{g^{-1}}(y)] \, d[g(x)] \\
&= - \int_{\mathcal{X}} [f_X(x) \mid J_g^{-1}(x)] \log [f_X(x) \mid J_g^{-1}(x)] \, d[g(x)] .
\end{aligned} \tag{7}$$

Using the relations $y = f(x) \Rightarrow dy = |J_f(x)| dx$ and $|A||B| = |AB|$, this becomes

$$\begin{aligned}
h(Y) &= - \int_{\mathcal{X}} [f_X(x) \mid J_g^{-1}(x) \mid J_g(x)] \log [f_X(x) \mid J_g^{-1}(x)] \, dx \\
&= - \int_{\mathcal{X}} f_X(x) \log f_X(x) \, dx - \int_{\mathcal{X}} f_X(x) \log |J_g^{-1}(x)| \, dx .
\end{aligned} \tag{8}$$

Finally, employing the fact (\rightarrow Definition I/1.6.6) that $\int_{\mathcal{X}} f_X(x) \, dx = 1$ and the determinant property $|A^{-1}| = 1/|A|$, we can derive the differential entropy (\rightarrow Definition I/2.2.1) of Y as

$$\begin{aligned}
h(Y) &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) \, dx - \int_{\mathcal{X}} f_X(x) \log \frac{1}{|J_g(x)|} \, dx \\
&\stackrel{(3)}{=} h(X) + \int_{\mathcal{X}} f_X(x) \log |J_g(x)| \, dx .
\end{aligned} \tag{9}$$

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.
- 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.
- 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.
- Wikipedia (2021): “Determinant”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Determinant#Properties_of_the_determinant.

Metadata: ID: P262 | shortcut: dent-noninv | author: JoramSoch | date: 2021-10-07, 10:39.

2.2.7 Conditional differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and \mathcal{Y} and probability density functions (\rightarrow Definition I/1.6.6) $p(x)$ and $p(y)$. Then,

the conditional differential entropy of Y given X or, differential entropy of Y conditioned on X , is defined as

$$h(Y|X) = \int_{x \in \mathcal{X}} p(x) \cdot h(Y|X = x) \quad (1)$$

where $h(Y|X = x)$ is the (marginal) differential entropy (\rightarrow Definition I/2.2.1) of Y , evaluated at x .

Sources:

- original work

Metadata: ID: D34 | shortcut: dent-cond | author: JoramSoch | date: 2020-03-21, 12:27.

2.2.8 Joint differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition I/1.3.2) density function (\rightarrow Definition I/1.6.6) $p(x, y)$. 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.

Sources:

- original work

Metadata: ID: D35 | shortcut: dent-joint | author: JoramSoch | date: 2020-03-21, 12:37.

2.2.9 Differential cross-entropy

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.5.1) on X with the probability density functions (\rightarrow Definition I/1.6.6) $p(x)$ 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.

Metadata: ID: D86 | shortcut: dent-cross | author: JoramSoch | date: 2020-07-28, 03:03.

2.3 Discrete mutual information

2.3.1 Definition

Definition:

1) The mutual information of two discrete random variables (\rightarrow Definition 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 Definition I/1.6.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.3.2) mass function of X and Y .

2) The mutual information of two continuous random variables (\rightarrow Definition 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 Definition I/1.6.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition 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>.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

2.3.2 Relation to marginal and conditional entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.1.1) of X and Y and $H(X|Y)$ and $H(Y|X)$ are the conditional entropies (\rightarrow Definition I/2.1.4).

Proof: The mutual information (\rightarrow Definition 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 Definition 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 Definition 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 Definition I/2.1.1) and conditional (\rightarrow Definition 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.

Metadata: ID: P19 | shortcut: dmi-mce | author: JoramSoch | date: 2020-01-13, 18:20.

2.3.3 Relation to marginal and joint entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.1.1) of X and Y and $H(X, Y)$ is the joint entropy (\rightarrow Definition I/2.1.5).

Proof: The mutual information (\rightarrow Definition 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 Definition 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 Definition I/2.1.1) and joint (\rightarrow Definition 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.

Metadata: ID: P20 | shortcut: dmi-mje | author: JoramSoch | date: 2020-01-13, 21:53.

2.3.4 Relation to joint and conditional entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.1.5) of X and Y and $H(X|Y)$ and $H(Y|X)$ are the conditional entropies (\rightarrow Definition I/2.1.4).

Proof: The existence of the joint probability mass function (\rightarrow Definition I/1.6.1) ensures that the mutual information (\rightarrow Definition 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 Proof 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 Proof 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.

Metadata: ID: P21 | shortcut: dmi-jce | author: JoramSoch | date: 2020-01-13, 22:17.

2.4 Continuous mutual information

2.4.1 Definition

Definition:

1) The mutual information of two discrete random variables (\rightarrow Definition 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 Definition I/1.6.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.3.2) mass function of X and Y .

2) The mutual information of two continuous random variables (\rightarrow Definition 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 Definition I/1.6.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition 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>.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

2.4.2 Relation to marginal and conditional differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.2.1) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/2.2.7).

Proof: The mutual information (\rightarrow Definition 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 Definition 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 Definition 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 Definition I/2.2.1) and conditional (\rightarrow Definition 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.

Metadata: ID: P58 | shortcut: cmi-mcde | author: JoramSoch | date: 2020-02-21, 16:53.

2.4.3 Relation to marginal and joint differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.2.1) of X and Y and $h(X, Y)$ is the joint differential entropy (\rightarrow Definition I/2.2.8).

Proof: The mutual information (\rightarrow Definition 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 Definition 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 Definition I/2.2.1) and joint (\rightarrow Definition 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.

Metadata: ID: P59 | shortcut: cmi-mjde | author: JoramSoch | date: 2020-02-21, 17:13.

2.4.4 Relation to joint and conditional differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.2.2) with the joint probability (\rightarrow Definition I/1.3.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition 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 Definition I/2.2.8) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/2.2.7).

Proof: The existence of the joint probability density function (\rightarrow Definition I/1.6.6) ensures that the mutual information (\rightarrow Definition 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 Proof 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 Proof 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.

Metadata: ID: P60 | shortcut: cmi-jcde | author: JoramSoch | date: 2020-02-21, 17:23.

2.5 Kullback-Leibler divergence

2.5.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition 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 Definition 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 Definition I/1.6.6) of 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>.

Metadata: ID: D52 | shortcut: kl | author: JoramSoch | date: 2020-05-10, 20:20.

2.5.2 Non-negativity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition 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 Definition 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 Proof I/2.1.8) states that the entropy (\rightarrow Definition I/2.1.1) of a probability distribution is always less than or equal to the cross-entropy (\rightarrow Definition 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.

Metadata: ID: P117 | shortcut: kl-nonneg | author: JoramSoch | date: 2020-05-31, 23:43.

2.5.3 Non-negativity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition 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 Definition 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 Proof 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 be non-negative real numbers and $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. Because $p(x)$ and $q(x)$ are probability mass functions (\rightarrow Definition I/1.6.1), such that

$$\begin{aligned} p(x) &\geq 0, & \sum_{x \in \mathcal{X}} p(x) &= 1 \quad \text{and} \\ q(x) &\geq 0, & \sum_{x \in \mathcal{X}} q(x) &= 1 , \end{aligned} \quad (4)$$

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.

Metadata: ID: P166 | shortcut: kl-nonneg2 | author: JoramSoch | date: 2020-09-09, 07:02.

2.5.4 Non-symmetry

Theorem: The Kullback-Leibler divergence (\rightarrow Definition 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 Definition I/1.5.1) P and Q .

Proof: Let $X \in \mathcal{X} = \{0, 1, 2\}$ be a discrete random variable (\rightarrow Definition I/1.2.2) and consider the two probability distributions (\rightarrow Definition 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 Definition II/1.3.1) and $\mathcal{U}(a, b)$ indicates a discrete uniform distribution (\rightarrow Definition II/1.1.1).

Then, the probability mass function of the binomial distribution (\rightarrow Proof II/1.3.2) 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 Proof II/1.1.2) entails that

$$q(x) = \frac{1}{3}, \quad (4)$$

such that the Kullback-Leibler divergence (\rightarrow Definition 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 Definition I/2.5.1) of Q from P is

$$\begin{aligned} \text{KL}[Q||P] &= \sum_{x \in \mathcal{X}} q(x) \cdot \log \frac{q(x)}{p(x)} \\ &= \frac{1}{3} \log \frac{4}{3} + \frac{1}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{4}{3} \\ &= \frac{1}{3} \left(\log \frac{4}{3} + \log \frac{2}{3} + \log \frac{4}{3} \right) \\ &= \frac{1}{3} \log \left(\frac{4}{3} \cdot \frac{2}{3} \cdot \frac{4}{3} \right) \\ &= \frac{1}{3} \log \frac{32}{27} = 0.0566 \end{aligned} \quad (6)$$

which provides an example for

$$\text{KL}[P||Q] \neq \text{KL}[Q||P] \quad (7)$$

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.

Metadata: ID: P147 | shortcut: kl-nonsymm | author: JoramSoch | date: 2020-08-11, 06:57.

2.5.5 Convexity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is convex in the pair of probability distributions (\rightarrow Definition I/1.5.1) (p, q) , i.e.

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \quad (1)$$

where (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions and $0 \leq \lambda \leq 1$.

Proof: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is defined as

$$\text{KL}[P || Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (2)$$

and the log sum inequality (\rightarrow Proof I/2.1.9) states that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \quad (3)$$

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

$$\begin{aligned} & \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \\ & \stackrel{(2)}{=} \sum_{x \in \mathcal{X}} \left[[\lambda p_1(x) + (1 - \lambda)p_2(x)] \cdot \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \right] \\ & \stackrel{(3)}{\leq} \sum_{x \in \mathcal{X}} \left[\lambda p_1(x) \cdot \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \cdot \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)} \right] \\ & = \lambda \sum_{x \in \mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda) \sum_{x \in \mathcal{X}} p_2(x) \cdot \log \frac{p_2(x)}{q_2(x)} \\ & \stackrel{(2)}{=} \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \end{aligned} \quad (4)$$

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.
- 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>.

Metadata: ID: P148 | shortcut: kl-conv | author: JoramSoch | date: 2020-08-11, 07:30.

2.5.6 Additivity for independent distributions

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is additive for independent distributions, i.e.

$$\text{KL}[P || Q] = \text{KL}[P_1 || Q_1] + \text{KL}[P_2 || Q_2] \quad (1)$$

where P_1 and P_2 are independent (\rightarrow Definition I/1.3.6) distributions (\rightarrow Definition I/1.5.1) with the joint distribution (\rightarrow Definition 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 Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

which, applied to the joint distributions P and Q , yields

$$\text{KL}[P||Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{q(x, y)} dy dx . \quad (3)$$

Applying $p(x, y) = p_1(x) p_2(y)$ and $q(x, y) = q_1(x) q_2(y)$, we have

$$\text{KL}[P||Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x) p_2(y)}{q_1(x) q_2(y)} dy dx . \quad (4)$$

Now we can separate the logarithm and evaluate the integrals:

$$\begin{aligned} \text{KL}[P||Q] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \left(\log \frac{p_1(x)}{q_1(x)} + \log \frac{p_2(y)}{q_2(y)} \right) dy dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x)}{q_1(x)} dy dx + \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy dx \\ &= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} \int_{\mathcal{Y}} p_2(y) dy dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \int_{\mathcal{X}} p_1(x) dx dy \\ &= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy \\ &\stackrel{(2)}{=} \text{KL}[P_1||Q_1] + \text{KL}[P_2||Q_2] . \end{aligned} \quad (5)$$

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.

Metadata: ID: P116 | shortcut: kl-add | author: JoramSoch | date: 2020-05-31, 23:26.

2.5.7 Invariance under parameter transformation

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is invariant under parameter transformation, i.e.

$$\text{KL}[p(x)||q(x)] = \text{KL}[p(y)||q(y)] \quad (1)$$

where $y(x) = mx + n$ is an affine transformation of x and $p(x)$ and $q(x)$ are the probability density functions (\rightarrow Definition I/1.6.6) of the probability distributions (\rightarrow Definition I/1.5.1) P and Q on the continuous random variable (\rightarrow Definition I/1.2.2) X .

Proof: The continuous Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) (KL divergence) is defined as

$$\text{KL}[p(x)||q(x)] = \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

where $a = \min(\mathcal{X})$ and $b = \max(\mathcal{X})$ are the lower and upper bound of the possible outcomes \mathcal{X} of X .

Due to the identity of the differentials

$$\begin{aligned} p(x) dx &= p(y) dy \\ q(x) dx &= q(y) dy \end{aligned} \quad (3)$$

which can be rearranged into

$$\begin{aligned} p(x) &= p(y) \frac{dy}{dx} \\ q(x) &= q(y) \frac{dy}{dx} , \end{aligned} \quad (4)$$

the KL divergence can be evaluated as follows:

$$\begin{aligned} \text{KL}[p(x)||q(x)] &= \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \\ &= \int_{y(a)}^{y(b)} p(y) \frac{dy}{dx} \cdot \log \left(\frac{p(y) \frac{dy}{dx}}{q(y) \frac{dy}{dx}} \right) dx \\ &= \int_{y(a)}^{y(b)} p(y) \cdot \log \frac{p(y)}{q(y)} dy \\ &= \text{KL}[p(y)||q(y)] . \end{aligned} \quad (5)$$

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#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>.

Metadata: ID: P115 | shortcut: kl-inv | author: JoramSoch | date: 2020-05-28, 00:18.

2.5.8 Relation to discrete entropy

Theorem: Let X be a discrete random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.5.1) on X . Then, the Kullback-Leibler divergence (\rightarrow Definition 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 cross-entropy (\rightarrow Definition I/2.1.6) of P and Q and $H(P)$ is the marginal entropy (\rightarrow Definition I/2.1.1) of P .

Proof: The discrete Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (2)$$

where $p(x)$ and $q(x)$ are the probability mass functions (\rightarrow Definition I/1.6.1) of P and Q . Separating the logarithm, we have:

$$\text{KL}[P||Q] = - \sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) . \quad (3)$$

Now considering the definitions of marginal entropy (\rightarrow Definition I/2.1.1) and cross-entropy (\rightarrow Definition I/2.1.6)

$$\begin{aligned} H(P) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ H(P, Q) &= - \sum_{x \in \mathcal{X}} p(x) \log q(x) , \end{aligned} \quad (4)$$

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.

Metadata: ID: P113 | shortcut: kl-ent | author: JoramSoch | date: 2020-05-27, 23:20.

2.5.9 Relation to differential entropy

Theorem: Let X be a continuous random variable (\rightarrow Definition I/1.2.2) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.5.1) on X . Then, the Kullback-Leibler divergence (\rightarrow Definition 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 Definition I/2.2.9) of P and Q and $h(P)$ is the marginal differential entropy (\rightarrow Definition I/2.2.1) of P .

Proof: The continuous Kullback-Leibler divergence (\rightarrow Definition 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 Definition I/1.6.6) 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 Definition I/2.2.1) and differential cross-entropy (\rightarrow Definition 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.

Metadata: ID: P114 | shortcut: kl-dent | author: JoramSoch | date: 2020-05-27, 23:32.

3 Estimation theory

3.1 Point estimates

3.1.1 Mean squared error

Definition: Let $\hat{\theta}$ be an estimator (\rightarrow Definition “est”) of an unknown parameter (\rightarrow Definition “para”) θ based on measured data (\rightarrow Definition “data”) y . Then, the mean squared error is defined as the expected value (\rightarrow Definition I/1.7.1) 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 Definition “samp”) 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.

Metadata: ID: D173 | shortcut: mse | author: JoramSoch | date: 2022-03-27, 23:41.

3.1.2 Partition of the mean squared error into bias and variance

Theorem: The mean squared error (\rightarrow Definition I/??) 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 Definition I/1.8.1) 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 (\rightarrow Definition “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 Definition I/??) (MSE) is defined as the expected value (\rightarrow Definition I/1.7.1) of the squared deviation of the estimated value $\hat{\theta}$ from the true value θ 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:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E_{\hat{\theta}} \left[\left(\hat{\theta} - \theta \right)^2 \right] \\
&= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) + E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right] \\
&= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 + 2 \left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right) \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right) + \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right] \\
&= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + E_{\hat{\theta}} \left[2 \left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right) \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right) \right] + E_{\hat{\theta}} \left[\left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \right].
\end{aligned} \tag{5}$$

Because $E_{\hat{\theta}}(\hat{\theta}) - \theta$ is constant as a function of $\hat{\theta}$, we have:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + 2 \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right) E_{\hat{\theta}} \left[\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right] + \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \\
&= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + 2 \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right) \left(E_{\hat{\theta}}(\hat{\theta}) - E_{\hat{\theta}}(\hat{\theta}) \right) + \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2 \\
&= E_{\hat{\theta}} \left[\left(\hat{\theta} - E_{\hat{\theta}}(\hat{\theta}) \right)^2 \right] + \left(E_{\hat{\theta}}(\hat{\theta}) - \theta \right)^2.
\end{aligned} \tag{6}$$

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.

Metadata: ID: P5 | shortcut: mse-bnv | author: JoramSoch | date: 2019-11-27, 14:26.

3.2 Interval estimates

3.2.1 Confidence interval

Definition: Let y be a random sample (\rightarrow Definition “samp”) from a probability distributions (\rightarrow Definition I/1.5.1) governed by a parameter (\rightarrow Definition “para”) of interest θ and quantities not of interest φ . A confidence interval for θ is defined as an interval $[u(y), v(y)]$ determined by the random variables (\rightarrow Definition I/1.2.2) $u(y)$ and $v(y)$ with the property

$$\Pr(u(y) < \theta < v(y) \mid \theta, \varphi) = \gamma \quad \text{for all } (\theta, \varphi). \tag{1}$$

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.

Metadata: ID: D174 | shortcut: ci | author: JoramSoch | date: 2022-03-27, 23:56.

3.2.2 Construction of confidence intervals using Wilks' theorem

Theorem: Let m be a generative model (\rightarrow Definition I/5.1.1) for measured data y with model parameters θ , consisting of a parameter of interest ϕ and nuisance parameters λ :

$$m : p(y|\theta) = \mathcal{D}(y; \theta), \quad \theta = \{\phi, \lambda\} . \quad (1)$$

Further, let $\hat{\theta}$ be an estimate of θ , obtained using maximum-likelihood-estimation (\rightarrow Definition I/4.1.3):

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta), \quad \hat{\theta} = \{\hat{\phi}, \hat{\lambda}\} . \quad (2)$$

Then, an asymptotic confidence interval (\rightarrow Definition I/??) for θ is given by

$$\text{CI}_{1-\alpha}(\hat{\phi}) = \left\{ \phi \mid \log p(y|\phi, \hat{\lambda}) \geq \log p(y|\hat{\phi}, \hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2 \right\} \quad (3)$$

where $1 - \alpha$ is the confidence level and $\chi_{1,1-\alpha}^2$ is the $(1 - \alpha)$ -quantile of the chi-squared distribution (\rightarrow Definition II/3.6.1) with 1 degree of freedom (\rightarrow Definition “dof”).

Proof: The confidence interval (\rightarrow Definition I/??) is defined as the interval that, under infinitely repeated random experiments (\rightarrow Definition I/1.1.1), contains the true parameter value with a certain probability.

Let us define the likelihood ratio (\rightarrow Definition “lr”)

$$\Lambda(\phi) = \frac{p(y|\phi, \hat{\lambda})}{p(y|\hat{\phi}, \hat{\lambda})} \quad (4)$$

and compute the log-likelihood ratio (\rightarrow Definition “llr”)

$$\log \Lambda(\phi) = \log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) . \quad (5)$$

Wilks' theorem (\rightarrow Proof “llr-wilks”) states that, when comparing two statistical models with parameter spaces Θ_1 and $\Theta_0 \subset \Theta_1$, as the sample size approaches infinity, the quantity calculated as -2 times the log-ratio of maximum likelihoods follows a chi-squared distribution (\rightarrow Definition II/3.6.1), if the null hypothesis is true:

$$H_0 : \theta \in \Theta_0 \quad \Rightarrow \quad -2 \log \frac{\max_{\theta \in \Theta_0} p(y|\theta)}{\max_{\theta \in \Theta_1} p(y|\theta)} \sim \chi_{\Delta k}^2 \quad (6)$$

where Δk is the difference in dimensionality between Θ_0 and Θ_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 Definition I/1.6.23) $\chi_{k,p}^2$ of the chi-squared distribution (\rightarrow Definition II/3.6.1), an $(1 - \alpha)$ -confidence interval is therefore given by all values ϕ that satisfy

$$-2 \log \Lambda(\phi) \leq \chi_{1,1-\alpha}^2 . \quad (8)$$

Applying (5) and rearranging, we can evaluate

$$\begin{aligned}
-2 \left[\log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) \right] &\leq \chi_{1,1-\alpha}^2 \\
\log p(y|\phi, \hat{\lambda}) - \log p(y|\hat{\phi}, \hat{\lambda}) &\geq -\frac{1}{2} \chi_{1,1-\alpha}^2 \\
\log p(y|\phi, \hat{\lambda}) &\geq \log p(y|\hat{\phi}, \hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2
\end{aligned} \tag{9}$$

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.
- 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.
- Wikipedia (2020): “Wilks’ theorem”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Wilks%27_theorem.

Metadata: ID: P56 | shortcut: ci-wilks | author: JoramSoch | date: 2020-02-19, 17:15.

4 Frequentist statistics

4.1 Likelihood theory

4.1.1 Likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the probability density function (\rightarrow Definition I/1.6.6) of the distribution of y given θ is called the likelihood function of m :

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) . \quad (1)$$

Sources:

- original work

Metadata: ID: D28 | shortcut: lf | author: JoramSoch | date: 2020-03-03, 15:50.

4.1.2 Log-likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the logarithm of the probability density function (\rightarrow Definition I/1.6.6) of the distribution of y given θ is called the log-likelihood function (\rightarrow Definition I/5.1.2) of m :

$$\text{LL}_m(\theta) = \log p(y|\theta, m) = \log \mathcal{D}(y; \theta) . \quad (1)$$

Sources:

- original work

Metadata: ID: D59 | shortcut: llf | author: JoramSoch | date: 2020-05-17, 22:52.

4.1.3 Maximum likelihood estimation

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the parameter values maximizing the likelihood function (\rightarrow Definition I/5.1.2) or log-likelihood function (\rightarrow Definition I/4.1.2) are called maximum likelihood estimates of θ :

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}_m(\theta) = \arg \max_{\theta} \text{LL}_m(\theta) . \quad (1)$$

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”.

Sources:

- original work

Metadata: ID: D60 | shortcut: mle | author: JoramSoch | date: 2020-05-15, 23:05.

4.1.4 MLE can be biased

Theorem: Maximum likelihood estimation (\rightarrow Definition I/4.1.3) can result in biased estimates (\rightarrow Definition “est-unb”) of model parameters, i.e. estimates whose long-term expected value is unequal to the quantities they estimate:

$$\mathbb{E} \left[\hat{\theta}_{\text{MLE}} \right] = \mathbb{E} \left[\arg \max_{\theta} \text{LL}_m(\theta) \right] \neq \theta . \quad (1)$$

Proof: Consider a set of independent and identical (\rightarrow Definition “iid”) normally distributed (\rightarrow Definition II/3.2.1) observations $x = \{x_1, \dots, x_n\}$ with unknown mean (\rightarrow Definition I/1.7.1) μ and variance (\rightarrow Definition I/1.8.1) σ^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 (\rightarrow Definition I/4.1.3) for the variance (\rightarrow Definition I/1.8.1) σ^2 is underestimating the true variance of the data distribution (\rightarrow Proof IV/1.1.2):

$$\mathbb{E} \left[\hat{\sigma}_{\text{MLE}}^2 \right] = \frac{n-1}{n} \sigma^2 \neq \sigma^2 . \quad (3)$$

This proves the existence of cases such as those stated by the theorem.

Sources:

- original work

Metadata: ID: P317 | shortcut: mle-bias | author: JoramSoch | date: 2022-03-18, 17:26.

4.1.5 Maximum log-likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the maximum log-likelihood (MLL) of m is the maximal value of the log-likelihood function (\rightarrow Definition 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 Definition I/4.1.3) into the log-likelihood function (\rightarrow Definition I/4.1.2).

Sources:

- original work

Metadata: ID: D61 | shortcut: mll | author: JoramSoch | date: 2020-05-15, 23:13.

4.1.6 Method of moments

Definition: Let measured data y follow a probability distribution (\rightarrow Definition I/1.5.1) with probability mass (\rightarrow Definition I/1.6.1) or probability density (\rightarrow Definition I/1.6.6) $p(y|\theta)$ 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 Definition I/1.14.1) of y in terms of θ

$$\begin{aligned}\mu_1 &= f_1(\theta_1, \dots, \theta_k) \\ &\vdots \\ \mu_k &= f_k(\theta_1, \dots, \theta_k) ,\end{aligned}\tag{1}$$

2) calculating the first k sample moments (\rightarrow Definition I/1.14.1) 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).

Metadata: ID: D151 | shortcut: mome | author: JoramSoch | date: 2021-04-29, 07:51.

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 Definition “data”).

More precisely, let m be a generative model (\rightarrow Definition 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.

Metadata: ID: D127 | shortcut: hyp | author: JoramSoch | date: 2021-03-19, 14:18.

4.2.2 Simple vs. composite

Definition: Let H be a statistical hypothesis (\rightarrow Definition 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 Definition I/1.5.5) of the test statistic (\rightarrow Definition I/4.3.5) is a function of sample size 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.

Metadata: ID: D128 | shortcut: hyp-simp | author: JoramSoch | date: 2021-03-19, 14:24.

4.2.3 Point/exact vs. set/inexact

Definition: Let H be a statistical hypothesis (\rightarrow Definition 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.

Metadata: ID: D129 | shortcut: hyp-point | author: JoramSoch | date: 2021-03-19, 14:28.

4.2.4 One-tailed vs. two-tailed

Definition: Let H_0 be a point (\rightarrow Definition I/4.2.3) null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \theta = \theta_0 \quad (1)$$

and consider a set (\rightarrow Definition I/4.2.3) alternative hypothesis (\rightarrow Definition I/4.3.3) H_1 . Then,

- H_1 is called a left-sided one-tailed hypothesis, if θ is assumed to be smaller than θ_0 :

$$H_1 : \theta < \theta_0 ; \quad (2)$$

- H_1 is called a right-sided one-tailed hypothesis, if θ is assumed to be larger than θ_0 :

$$H_1 : \theta > \theta_0 ; \quad (3)$$

- H_1 is called a two-tailed hypothesis, if θ is assumed to be unequal to θ_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.

Metadata: ID: D138 | shortcut: hyp-tail | author: JoramSoch | date: 2021-03-31, 09:21.

4.3 Hypothesis testing

4.3.1 Statistical test

Definition: Let y be a set of measured data (\rightarrow Definition “data”). Then, a statistical hypothesis test consists of the following:

- an assumption about the distribution (\rightarrow Definition I/1.5.1) of the data, often expressed in terms of a statistical model (\rightarrow Definition I/5.1.1) m ;
- a null hypothesis (\rightarrow Definition I/4.3.2) H_0 and an alternative hypothesis (\rightarrow Definition I/4.3.3) H_1 which make specific statements about the distribution of the data;
- a test statistic (\rightarrow Definition I/4.3.5) $T(Y)$ which is a function of the data and whose distribution under the null hypothesis (\rightarrow Definition I/4.3.2) is known;
- a significance level (\rightarrow Definition I/4.3.8) α which imposes an upper bound on the probability (\rightarrow Definition 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 α , a critical value (\rightarrow Definition I/4.3.9) t_{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 Definition 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.

Metadata: ID: D130 | shortcut: test | author: JoramSoch | date: 2021-03-19, 14:36.

4.3.2 Null hypothesis

Definition: The statement which is tested in a statistical hypothesis test (\rightarrow Definition 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 Definition 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 Definition 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.

Metadata: ID: D125 | shortcut: h0 | author: JoramSoch | date: 2021-03-12, 10:25.

4.3.3 Alternative hypothesis

Definition: Let H_0 be a null hypothesis (\rightarrow Definition I/4.3.2) of a statistical hypothesis test (\rightarrow Definition 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 Definition I/5.1.1) describing measured data y using model parameters $\theta \in \Theta$. Then, null and alternative hypothesis are formally specified as

$$\begin{aligned} H_0 : \theta \in \Theta_0 \quad \text{where} \quad \Theta_0 \subset \Theta \\ H_1 : \theta \in \Theta_1 \quad \text{where} \quad \Theta_1 = \Theta \setminus \Theta_0 . \end{aligned} \quad (1)$$

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.

Metadata: ID: D126 | shortcut: h1 | author: JoramSoch | date: 2021-03-12, 10:36.

4.3.4 One-tailed vs. two-tailed

Definition: Let there be a statistical test (\rightarrow Definition I/4.3.1) of an alternative hypothesis (\rightarrow Definition I/4.3.3) H_1 against a null hypothesis (\rightarrow Definition I/4.3.2) H_0 . Then,

- the test is called a one-tailed test, if H_1 is a one-tailed hypothesis (\rightarrow Definition I/4.2.4);
- the test is called a two-tailed test, if H_1 is a two-tailed hypothesis (\rightarrow Definition I/4.2.4).

The fact whether a test (\rightarrow Definition I/4.3.1) is one-tailed or two-tailed has consequences for the computation of critical value (\rightarrow Definition I/4.3.9) and p-value (\rightarrow Definition 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.

Metadata: ID: D139 | shortcut: test-tail | author: JoramSoch | date: 2021-03-31, 09:32.

4.3.5 Test statistic

Definition: In a statistical hypothesis test (\rightarrow Definition I/4.3.1), the test statistic $T(Y)$ is a scalar function of the measured data (\rightarrow Definition “data”) y whose distribution under the null hypothesis (\rightarrow Definition I/4.3.2) H_0 can be established. Together with a significance level (\rightarrow Definition I/4.3.8) α , this distribution implies a critical value (\rightarrow Definition I/4.3.9) t_{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.

Metadata: ID: D131 | shortcut: tstat | author: JoramSoch | date: 2021-03-19, 14:40.

4.3.6 Size of a test

Definition: Let there be a statistical hypothesis test (\rightarrow Definition I/4.3.1) with null hypothesis (\rightarrow Definition 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 Definition I/1.3.1) of rejecting the null hypothesis (\rightarrow Definition I/4.3.2) H_0 , given that H_0 is actually true.

For a simple null hypothesis (\rightarrow Definition I/4.2.2), the size is determined by the following conditional probability (\rightarrow Definition I/1.3.4):

$$\Pr(\text{test rejects } H_0 | H_0) . \quad (1)$$

For a composite null hypothesis (\rightarrow Definition I/4.2.2), the size is the supremum over all possible realizations of the null hypothesis (\rightarrow Definition 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)).

Metadata: ID: D132 | shortcut: size | author: JoramSoch | date: 2021-03-19, 14:46.

4.3.7 Power of a test

Definition: Let there be a statistical hypothesis test (\rightarrow Definition I/4.3.1) with null hypothesis (\rightarrow Definition I/4.3.2) H_0 and alternative hypothesis (\rightarrow Definition 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 Definition I/1.3.1) of rejecting H_0 , given that H_1 is actually true.

For given null (\rightarrow Definition I/4.3.2) and alternative (\rightarrow Definition I/4.3.3) hypothesis (\rightarrow Definition I/4.2.1), the size is determined by the following conditional probability (\rightarrow Definition 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.

Metadata: ID: D137 | shortcut: power | author: JoramSoch | date: 2021-03-31, 09:01.

4.3.8 Significance level

Definition: Let the size (\rightarrow Definition I/4.3.6) of a statistical hypothesis test (\rightarrow Definition I/4.3.1) be the probability of a false-positive result or making a type I error, i.e. the probability (\rightarrow Definition I/1.3.1) of rejecting the null hypothesis (\rightarrow Definition 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 α , if the size is less than or equal to α :

$$\Pr(\text{test rejects } H_0 | H_0) \leq \alpha . \quad (2)$$

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.

Metadata: ID: D133 | shortcut: alpha | author: JoramSoch | date: 2021-03-19, 14:50.

4.3.9 Critical value

Definition: In a statistical hypothesis test (\rightarrow Definition I/4.3.1), the critical value (\rightarrow Definition I/4.3.9) t_{crit} is that value of the test statistic (\rightarrow Definition 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 Definition I/4.3.8) α . Put differently, if the observed test statistic $t_{\text{obs}} = T(y)$ computed from actually measured data (\rightarrow Definition “data”) y is as extreme or more extreme than the critical value, the test rejects the null hypothesis (\rightarrow Definition I/4.3.2) H_0 in favor of the alternative hypothesis (\rightarrow Definition 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.

Metadata: ID: D134 | shortcut: cval | author: JoramSoch | date: 2021-03-19, 14:54.

4.3.10 p-value

Definition: Let there be a statistical test (\rightarrow Definition I/4.3.1) of the null hypothesis (\rightarrow Definition I/4.3.2) H_0 and the alternative hypothesis (\rightarrow Definition I/4.3.3) H_1 using the test statistic (\rightarrow Definition I/4.3.5) $T(Y)$. Let y be the measured data (\rightarrow Definition “data”) 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 Definition I/1.6.13) (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_{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 Definition I/4.2.4);
- $p = 1 - F_T(t_{\text{obs}})$, if H_1 is a right-sided one-tailed hypothesis (\rightarrow Definition 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 Definition 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.

Metadata: ID: D135 | shortcut: pval | author: JoramSoch | date: 2021-03-19, 14:58.

4.3.11 Distribution of p-value under null hypothesis

Theorem: Under the null hypothesis (\rightarrow Definition I/4.3.2), the p-value (\rightarrow Definition I/4.3.10) in a statistical test (\rightarrow Definition I/4.3.1) follows a continuous uniform distribution (\rightarrow Definition 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 Definition I/4.2.4). Then, the p-value is a function of the test statistic (\rightarrow Definition I/4.3.10)

$$\begin{aligned} P &= F_T(T) \\ p &= F_T(t_{\text{obs}}) \end{aligned} \quad (2)$$

where t_{obs} is the observed test statistic (\rightarrow Definition I/4.3.5) and $F_T(t)$ is the cumulative distribution function (\rightarrow Definition I/1.6.13) of the test statistic (\rightarrow Definition I/4.3.5) under the null hypothesis (\rightarrow Definition I/4.3.2).

Then, we can obtain the cumulative distribution function (\rightarrow Definition I/1.6.13) of the p-value (\rightarrow Definition I/4.3.10) as

$$\begin{aligned}
F_P(p) &= \Pr(P < p) \\
&= \Pr(F_T(T) < p) \\
&= \Pr(T < F_T^{-1}(p)) \\
&= F_T^{-1}(F_T^{-1}(p)) \\
&= p
\end{aligned} \tag{3}$$

which is the cumulative distribution function of a continuous uniform distribution (\rightarrow Proof II/3.1.4) over the interval $[0, 1]$:

$$F_X(x) = \int_{-\infty}^x \mathcal{U}(z; 0, 1) \, dz = x \quad \text{where } 0 \leq x \leq 1. \tag{4}$$

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>.

Metadata: ID: P318 | shortcut: pval-h0 | author: JoramSoch | date: 2022-03-18, 22:37.

5 Bayesian statistics

5.1 Probabilistic modeling

5.1.1 Generative model

Definition: Consider measured data (\rightarrow Definition “data”) y and some unknown latent parameters (\rightarrow Definition “para”) θ . A statement about the distribution (\rightarrow Definition I/1.5.1) of y given θ is called a generative model m

$$m : y \sim \mathcal{D}(\theta) , \quad (1)$$

where \mathcal{D} denotes an arbitrary probability distribution and θ 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.neuroimage.2007.06.048

Metadata: ID: D27 | shortcut: gm | author: JoramSoch | date: 2020-03-03, 15:50.

5.1.2 Likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the probability density function (\rightarrow Definition I/1.6.6) of the distribution of y given θ is called the likelihood function of m :

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) . \quad (1)$$

Sources:

- original work

Metadata: ID: D28 | shortcut: lf | author: JoramSoch | date: 2020-03-03, 15:50.

5.1.3 Prior distribution

Definition: Consider measured data y and some unknown latent parameters θ . A distribution of θ unconditional on y is called a prior distribution:

$$\theta \sim \mathcal{D}(\lambda) . \quad (1)$$

The parameters λ of this distribution are called the prior hyperparameters and the probability density function (\rightarrow Definition I/1.6.6) is called the prior density:

$$p(\theta|m) = \mathcal{D}(\theta; \lambda) . \quad (2)$$

Sources:

- original work

Metadata: ID: D29 | shortcut: prior | author: JoramSoch | date: 2020-03-03, 16:09.

5.1.4 Full probability model

Definition: Consider measured data y and some unknown latent parameters θ . The combination of a generative model (\rightarrow Definition I/5.1.1) for y and a prior distribution (\rightarrow Definition I/5.1.3) on θ is called a full probability model m :

$$m : y \sim \mathcal{D}(\theta), \theta \sim \mathcal{D}(\lambda) . \quad (1)$$

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/>.

Metadata: ID: D30 | shortcut: fpm | author: JoramSoch | date: 2020-03-03, 16:16.

5.1.5 Joint likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the joint probability (\rightarrow Definition I/1.3.2) density function (\rightarrow Definition I/1.6.6) of y and θ is called the joint likelihood:

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (1)$$

Sources:

- original work

Metadata: ID: D31 | shortcut: jl | author: JoramSoch | date: 2020-03-03, 16:36.

5.1.6 Joint likelihood is product of likelihood and prior

Theorem: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the joint likelihood (\rightarrow Definition I/5.1.5) is equal to the product of likelihood function (\rightarrow Definition I/5.1.2) and prior density (\rightarrow Definition I/5.1.3):

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (1)$$

Proof: The joint likelihood (\rightarrow Definition I/5.1.5) is defined as the joint probability (\rightarrow Definition I/1.3.2) density function (\rightarrow Definition I/1.6.6) of data y and parameters θ :

$$p(y, \theta|m) . \quad (2)$$

Applying the law of conditional probability (\rightarrow Definition I/1.3.4), we have:

$$\begin{aligned}
p(y|\theta, m) &= \frac{p(y, \theta|m)}{p(\theta|m)} \\
&\Leftrightarrow \\
p(y, \theta|m) &= p(y|\theta, m) p(\theta|m) .
\end{aligned} \tag{3}$$

Sources:

- original work

Metadata: ID: P89 | shortcut: jl-lfnprior | author: JoramSoch | date: 2020-05-05, 04:21.

5.1.7 Posterior distribution

Definition: Consider measured data y and some unknown latent parameters θ . The distribution of θ conditional on y is called the posterior distribution:

$$\theta|y \sim \mathcal{D}(\phi) . \tag{1}$$

The parameters ϕ of this distribution are called the posterior hyperparameters and the probability density function (\rightarrow Definition I/1.6.6) is called the posterior density:

$$p(\theta|y, m) = \mathcal{D}(\theta; \phi) . \tag{2}$$

Sources:

- original work

Metadata: ID: D32 | shortcut: post | author: JoramSoch | date: 2020-03-03, 16:43.

5.1.8 Posterior density is proportional to joint likelihood

Theorem: In a full probability model (\rightarrow Definition I/5.1.4) m describing measured data y using model parameters θ , the posterior density (\rightarrow Definition I/5.1.7) over the model parameters is proportional to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(\theta|y, m) \propto p(y, \theta|m) . \tag{1}$$

Proof: In a full probability model (\rightarrow Definition I/5.1.4), the posterior distribution (\rightarrow Definition I/5.1.7) can be expressed using Bayes' theorem (\rightarrow Proof I/5.3.1):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \tag{2}$$

Applying the law of conditional probability (\rightarrow Definition I/1.3.4) to the numerator, we have:

$$p(\theta|y, m) = \frac{p(y, \theta|m)}{p(y|m)} . \tag{3}$$

Because the denominator does not depend on θ , it is constant in θ and thus acts a proportionality factor between the posterior distribution and the joint likelihood:

$$p(\theta|y, m) \propto p(y, \theta|m) . \quad (4)$$

Sources:

- original work

Metadata: ID: P90 | shortcut: post-jl | author: JoramSoch | date: 2020-05-05, 04:46.

5.1.9 Marginal likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the marginal probability (\rightarrow Definition I/1.3.3) density function (\rightarrow Definition I/1.6.6) of y across the parameter space Θ is called the marginal likelihood:

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta . \quad (1)$$

Sources:

- original work

Metadata: ID: D33 | shortcut: ml | author: JoramSoch | date: 2020-03-03, 16:49.

5.1.10 Marginal likelihood is integral of joint likelihood

Theorem: In a full probability model (\rightarrow Definition I/5.1.4) m describing measured data y using model parameters θ , the marginal likelihood (\rightarrow Definition I/5.1.9) is the integral of the joint likelihood (\rightarrow Definition I/5.1.5) across the parameter space Θ

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta \quad (1)$$

is and related to likelihood function (\rightarrow Definition I/5.1.2) and prior distribution (\rightarrow Definition I/5.1.7) 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 Definition I/5.1.4), the marginal likelihood (\rightarrow Definition I/5.1.9) is defined as the marginal probability (\rightarrow Definition I/1.3.3) of the data y , given only the model m :

$$p(y|m) . \quad (3)$$

Using the law of marginal probability (\rightarrow Definition I/1.3.3), this can be obtained by integrating the joint likelihood (\rightarrow Definition I/5.1.5) 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 Definition I/1.3.4), the integrand can also be written as the product of likelihood function (\rightarrow Definition I/5.1.2) and prior density (\rightarrow Definition I/5.1.3):

$$p(y|m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (5)$$

Sources:

- original work

Metadata: ID: P91 | shortcut: ml-jl | author: JoramSoch | date: 2020-05-05, 04:59.

5.2 Prior distributions

5.2.1 Flat vs. hard vs. soft

Definition: Let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) for the parameter θ of a generative model (\rightarrow Definition I/5.1.1) m . Then,

- the distribution is called a “flat prior”, if its precision (\rightarrow Definition I/1.8.12) is zero or variance (\rightarrow Definition I/1.8.1) is infinite;
- the distribution is called a “hard prior”, if its precision (\rightarrow Definition I/1.8.12) is infinite or variance (\rightarrow Definition I/1.8.1) is zero;
- the distribution is called a “soft prior”, if its precision (\rightarrow Definition I/1.8.12) and variance (\rightarrow Definition I/1.8.1) 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.

Metadata: ID: D116 | shortcut: prior-flat | author: JoramSoch | date: 2020-12-02, 17:04.

5.2.2 Uniform vs. non-uniform

Definition: Let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) for the parameter θ of a generative model (\rightarrow Definition I/5.1.1) m where θ belongs to the parameter space Θ . Then,

- the distribution is called a “uniform prior”, if its density (\rightarrow Definition I/1.6.6) or mass (\rightarrow Definition I/1.6.1) is constant over Θ ;
- the distribution is called a “non-uniform prior”, if its density (\rightarrow Definition I/1.6.6) or mass (\rightarrow Definition I/1.6.1) is not constant over Θ .

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.

Metadata: ID: D117 | shortcut: prior-uni | author: JoramSoch | date: 2020-12-02, 17:21.

5.2.3 Informative vs. non-informative

Definition: Let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) for the parameter θ of a generative model (\rightarrow Definition 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 Proof I/5.1.8);
- the distribution is called a “non-informative prior”, if it does not influence (\rightarrow Proof I/5.1.8) the posterior hyperparameters (\rightarrow Definition I/5.1.7).

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.

Metadata: ID: D118 | shortcut: prior-inf | author: JoramSoch | date: 2020-12-02, 17:28.

5.2.4 Empirical vs. non-empirical

Definition: Let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) for the parameter θ of a generative model (\rightarrow Definition I/5.1.1) m . Then,

- the distribution is called an “empirical prior”, if it has been derived from empirical data (\rightarrow Proof I/5.1.8);
- 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.

Metadata: ID: D119 | shortcut: prior-emp | author: JoramSoch | date: 2020-12-02, 17:37.

5.2.5 Conjugate vs. non-conjugate

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Then,

- the prior distribution (\rightarrow Definition I/5.1.3) is called “conjugate”, if it, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition 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.

Metadata: ID: D120 | shortcut: prior-conj | author: JoramSoch | date: 2020-12-02, 17:55.

5.2.6 Maximum entropy priors

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|\lambda, m)$ using prior hyperparameters (\rightarrow Definition I/5.1.3) λ . Then, the prior distribution is called a “maximum entropy prior”, if

1) when θ is a discrete random variable (\rightarrow Definition I/1.2.6), it maximizes the entropy (\rightarrow Definition I/2.1.1) of the prior probability mass function (\rightarrow Definition I/1.6.1):

$$\lambda_{\text{maxent}} = \arg \max_{\lambda} H[p(\theta|\lambda, m)] ; \quad (1)$$

2) when θ is a continuous random variable (\rightarrow Definition I/1.2.6), it maximizes the differential entropy (\rightarrow Definition I/2.2.1) of the prior probability density function (\rightarrow Definition I/1.6.6):

$$\lambda_{\text{maxent}} = \arg \max_{\lambda} h[p(\theta|\lambda, m)] . \quad (2)$$

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.

Metadata: ID: D121 | shortcut: prior-maxent | author: JoramSoch | date: 2020-12-02, 18:13.

5.2.7 Empirical Bayes priors

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|\lambda, m)$ using prior hyperparameters (\rightarrow Definition I/5.1.3) λ . Let $p(y|\lambda, m)$ be the marginal likelihood (\rightarrow Definition I/5.1.9) when integrating the parameters out of the joint likelihood (\rightarrow Proof I/5.1.10). Then, the prior distribution is called an “Empirical Bayes (\rightarrow Definition I/5.3.3) 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.

Metadata: ID: D122 | shortcut: prior-eb | author: JoramSoch | date: 2020-12-02, 18:19.

5.2.8 Reference priors

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|\lambda, m)$ using prior hyperparameters (\rightarrow Definition I/5.1.3) λ . Let $p(\theta|y, \lambda, m)$ be the posterior distribution (\rightarrow Definition I/5.1.7) that is proportional to the the joint likelihood (\rightarrow Proof I/5.1.8). Then, the prior distribution is called a “reference prior”, if it maximizes the expected (\rightarrow Definition I/1.7.1) Kullback-Leibler divergence (\rightarrow Definition 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.

Metadata: ID: D123 | shortcut: prior-ref | author: JoramSoch | date: 2020-12-02, 18:26.

5.3 Bayesian inference

5.3.1 Bayes’ theorem

Theorem: Let A and B be two arbitrary statements about random variables (\rightarrow Definition 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 Definition I/1.3.4) is defined as the ratio of joint probability (\rightarrow Definition I/1.3.2), i.e. the probability of both statements being true, and marginal probability (\rightarrow Definition 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.

Metadata: ID: P4 | shortcut: bayes-th | author: JoramSoch | date: 2019-09-27, 16:24.

5.3.2 Bayes’ rule

Theorem: Let A_1 , A_2 and B be arbitrary statements about random variables (\rightarrow Definition I/1.2.2) where A_1 and A_2 are mutually exclusive. Then, Bayes’ rule states that the posterior odds (\rightarrow Definition “post-odd”) are equal to the Bayes factor (\rightarrow Definition IV/3.4.1) times the prior odds (\rightarrow Definition “prior-odd”), 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 Proof I/5.3.1), the conditional probabilities (\rightarrow Definition 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.

Metadata: ID: P12 | shortcut: bayes-rule | author: JoramSoch | date: 2020-01-06, 20:55.

5.3.3 Empirical Bayes

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with model parameters θ and hyper-parameters λ implying the likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, \lambda, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|\lambda, m)$. Then, an Empirical Bayes treatment of m , also referred to as “type II maximum likelihood (\rightarrow Definition I/4.1.3)” or “evidence (\rightarrow Definition IV/3.1.1) approximation”, consists in

1) evaluating the marginal likelihood (\rightarrow Definition I/5.1.9) 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 Definition IV/3.1.1) with respect to λ

$$\hat{\lambda} = \arg \max_{\lambda} \log p(y|\lambda, m) \quad (2)$$

3) and using the prior distribution (\rightarrow Definition I/5.1.3) at this maximum

$$p(\theta|m) = p(\theta|\hat{\lambda}, m) \quad (3)$$

for Bayesian inference (\rightarrow Proof I/5.3.1), i.e. obtaining the posterior distribution (\rightarrow Proof I/5.1.8) and computing the marginal likelihood (\rightarrow Proof I/5.1.10).

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.
- 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>.

Metadata: ID: D149 | shortcut: eb | author: JoramSoch | date: 2021-04-29, 06:46.

5.3.4 Variational Bayes

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with model parameters θ implying the likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition 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 Definition I/5.1.7)

$$q(\theta) \approx p(\theta|y, m), \quad (1)$$

2) evaluating the variational free energy (\rightarrow Definition IV/3.1.7)

$$F_q(m) = \int q(\theta) \log p(y|\theta, m) d\theta - \int q(\theta) \frac{q(\theta)}{p(\theta|m)} d\theta \quad (2)$$

3) and maximizing this function with respect to $q(\theta)$

$$\hat{q}(\theta) = \arg \max_q F_q(m). \quad (3)$$

for Bayesian inference, i.e. obtaining the posterior distribution (from eq. (3)) and approximating the marginal likelihood (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.
- 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.

Metadata: ID: D150 | shortcut: vb | author: JoramSoch | date: 2021-04-29, 07:15.

Chapter II

Probability Distributions

1 Univariate discrete distributions

1.1 Discrete uniform distribution

1.1.1 Definition

Definition: Let X be a discrete random variable (\rightarrow Definition 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.

Metadata: ID: D88 | shortcut: duni | author: JoramSoch | date: 2020-07-28, 04:05.

1.1.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 Definition II/1.1.1) 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 Definition I/1.6.1) is

$$\sum_{x=a}^b f_X(x) = 1 , \quad (4)$$

we have

$$f_X(x) = \frac{1}{n} = \frac{1}{b - a + 1} . \quad (5)$$

Sources:

- original work

Metadata: ID: P140 | shortcut: duni-pmf | author: JoramSoch | date: 2020-07-28, 04:57.

1.1.3 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof II/1.1.2) 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 Definition I/1.6.13) 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.

Sources:

- original work

Metadata: ID: P141 | shortcut: duni-cdf | author: JoramSoch | date: 2020-07-28, 05:34.

1.1.4 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the quantile function (\rightarrow Definition I/1.6.23) 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 Proof II/1.1.3) 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 $Q_X(p)$ is defined as (\rightarrow Definition I/1.6.23) 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 Proof II/1.1.3) multiples of $1/n$ with $n = b - a + 1$, the quantile function (\rightarrow Definition I/1.6.23) 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 Proof II/1.1.3) 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:

$$\begin{aligned} Q_X(p) &= a + p \cdot (b - a + 1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1-p) + (b+1)p - 1 . \end{aligned} \quad (7)$$

Sources:

- original work

Metadata: ID: P142 | shortcut: duni-qf | author: JoramSoch | date: 2020-07-28, 06:17.

1.2 Bernoulli distribution

1.2.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to follow a Bernoulli distribution with success probability p

$$X \sim \text{Bern}(p) , \quad (1)$$

if $X = 1$ with probability (\rightarrow Definition I/1.3.1) p and $X = 0$ with probability (\rightarrow Definition 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.

Metadata: ID: D44 | shortcut: bern | author: JoramSoch | date: 2020-03-22, 17:40.

1.2.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 Definition II/1.2.1).

Sources:

- original work

Metadata: ID: P96 | shortcut: bern-pmf | author: JoramSoch | date: 2020-05-11, 22:10.

1.2.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = p . \quad (2)$$

Proof: The expected value (\rightarrow Definition I/1.7.1) 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 Proof II/1.2.2), 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.

Metadata: ID: P22 | shortcut: bern-mean | author: JoramSoch | date: 2020-01-16, 10:58.

1.2.4 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.7.1) of X is

$$\text{Var}(X) = p(1 - p) . \quad (2)$$

Proof: The variance (\rightarrow Definition I/1.7.1) is the probability-weighted average of the squared deviation from the expected value (\rightarrow Definition I/1.7.1) 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 Proof I/1.8.3):

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (4)$$

The mean of a Bernoulli random variable (\rightarrow Proof II/1.2.3) 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 (??), (??) and (??), 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.

Metadata: ID: P301 | shortcut: bern-var | author: JoramSoch | date: 2022-01-20, 15:06.

1.2.5 Range of variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) 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 Proof II/??) 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 Definition II/1.2.1) 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 Proof I/1.8.4), 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.

Metadata: ID: P303 | shortcut: bern-varrange | author: JoramSoch | date: 2022-01-27, 09:03.

1.3 Binomial distribution

1.3.1 Definition

Definition: Let X be a random variable (\rightarrow Definition 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 Definition I/1.3.6) trials, where each trial has two possible outcomes (\rightarrow Definition II/1.2.1) (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.

Metadata: ID: D45 | shortcut: bin | author: JoramSoch | date: 2020-03-22, 17:52.

1.3.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a binomial distribution (\rightarrow Definition II/1.3.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 is defined as (\rightarrow Definition II/1.3.1) the number of successes observed in n independent (\rightarrow Definition I/1.3.6) trials, where each trial has two possible outcomes (\rightarrow Definition II/1.2.1) (success/failure) and the probability (\rightarrow Definition 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.

Sources:

- original work

Metadata: ID: P97 | shortcut: bin-pmf | author: JoramSoch | date: 2020-05-11, 22:35.

1.3.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a binomial distribution (\rightarrow Definition II/1.3.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = np . \quad (2)$$

Proof: By definition, a binomial random variable (\rightarrow Definition II/1.3.1) is the sum of n independent and identical (\rightarrow Definition “iid”) Bernoulli trials (\rightarrow Definition II/1.2.1) 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 Proof I/1.7.5), 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 Proof II/1.2.3), 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.

Metadata: ID: P23 | shortcut: bin-mean | author: JoramSoch | date: 2020-01-16, 11:06.

1.3.4 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a binomial distribution (\rightarrow Definition II/1.3.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

$$\text{Var}(X) = np(1 - p) . \quad (2)$$

Proof: By definition, a binomial random variable (\rightarrow Definition II/1.3.1) is the sum of n independent and identical (\rightarrow Definition “iid”) Bernoulli trials (\rightarrow Definition II/1.2.1) 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 Proof I/1.8.10), 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 Proof II/??), 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.

Metadata: ID: P302 | shortcut: bin-var | author: JoramSoch | date: 2022-01-20, 15:19.

1.3.5 Range of variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a binomial distribution (\rightarrow Definition II/1.3.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) 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 Definition II/1.3.1) is the sum of n independent and identical (\rightarrow Definition “iid”) Bernoulli trials (\rightarrow Definition II/1.2.1) 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 Proof I/1.8.10), 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 Proof II/??)

$$0 \leq \text{Var}(X_i) \leq \frac{1}{4} \quad \text{for all } i = 1, \dots, n , \quad (5)$$

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:

$$\text{Var}(X) \in \left[0, \frac{n}{4}\right] . \quad (8)$$

Sources:

- original work

Metadata: ID: P304 | shortcut: bin-varrange | author: JoramSoch | date: 2022-01-27, 09:20.

1.4 Poisson distribution

1.4.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to follow a Poisson distribution with rate λ

$$X \sim \text{Poiss}(\lambda) , \quad (1)$$

if and only if its probability mass function (\rightarrow Definition 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.

Metadata: ID: D62 | shortcut: poiss | author: JoramSoch | date: 2020-05-25, 23:34.

1.4.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Poisson distribution (\rightarrow Definition II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 Definition II/1.4.1).

Sources:

- original work

Metadata: ID: P102 | shortcut: poiss-pmf | author: JoramSoch | date: 2020-05-14, 20:39.

1.4.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Poisson distribution (\rightarrow Definition II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$\mathbb{E}(X) = \lambda . \quad (2)$$

Proof: The expected value of a discrete random variable (\rightarrow Definition I/1.7.1) 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 Proof 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.

Metadata: ID: P151 | shortcut: poiss-mean | author: JoramSoch | date: 2020-08-19, 06:09.

1.4.4 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Poisson distribution (\rightarrow Definition II/1.4.1):

$$X \sim \text{Poiss}(\lambda) . \tag{1}$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

$$\text{Var}(X) = \lambda . \tag{2}$$

Proof: The variance (\rightarrow Definition I/1.8.1) can be expressed in terms of expected values (\rightarrow Proof I/1.8.3) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

The expected value of a Poisson random variable (\rightarrow Proof II/1.4.3) is

$$E(X) = \lambda . \quad (4)$$

Let us now consider the expectation (\rightarrow Definition I/1.7.1) 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 Proof 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:

- jbstatistics (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.

Metadata: ID: P230 | shortcut: poiss-var | author: JoramSoch | date: 2021-04-29, 09:59.

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Definition

Definition: Let X be a random vector (\rightarrow Definition I/1.2.3). Then, X is said to follow a categorical distribution with success probability p_1, \dots, p_k

$$X \sim \text{Cat}([p_1, \dots, p_k]) , \quad (1)$$

if $X = e_i$ with probability (\rightarrow Definition 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.

Metadata: ID: D46 | shortcut: cat | author: JoramSoch | date: 2020-03-22, 18:09.

2.1.2 Probability mass function

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a categorical distribution (\rightarrow Definition II/2.1.1):

$$X \sim \text{Cat}([p_1, \dots, p_k]) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 Definition II/2.1.1).

Sources:

- original work

Metadata: ID: P98 | shortcut: cat-pmf | author: JoramSoch | date: 2020-05-11, 22:58.

2.1.3 Mean

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a categorical distribution (\rightarrow Definition II/2.1.1):

$$X \sim \text{Cat}([p_1, \dots, p_k]) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = [p_1, \dots, p_k] . \quad (2)$$

Proof: If we conceive the outcome of a categorical distribution (\rightarrow Definition 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 Definition I/1.7.1) is

$$\begin{aligned} E(X) &= \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x) \\ &= \sum_{i=1}^k e_i \cdot \Pr(X = e_i) \\ &= \sum_{i=1}^k e_i \cdot p_i \\ &= [p_1, \dots, p_k] . \end{aligned} \quad (3)$$

Sources:

- original work

Metadata: ID: P24 | shortcut: cat-mean | author: JoramSoch | date: 2020-01-16, 11:17.

2.2 Multinomial distribution

2.2.1 Definition

Definition: Let X be a random vector (\rightarrow Definition 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 Definition I/1.3.6) trials, where each trial has k possible outcomes (\rightarrow Definition 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.

Metadata: ID: D47 | shortcut: mult | author: JoramSoch | date: 2020-03-22, 17:52.

2.2.2 Probability mass function

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a multinomial distribution (\rightarrow Definition II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, the probability mass function (\rightarrow Definition 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 is defined as (\rightarrow Definition II/2.2.1) a vector of the numbers of observations belonging to k distinct categories in n independent (\rightarrow Definition I/1.3.6) trials, where each trial has k possible outcomes (\rightarrow Definition II/2.1.1) and the category probabilities (\rightarrow Definition I/1.3.1) are identical across trials.

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_k observations for category k . This number is equal to the number of possibilities in which x_1 category 1 objects, ..., x_k category k objects 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_k observations for category k , 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.

Sources:

- original work

Metadata: ID: P99 | shortcut: mult-pmf | author: JoramSoch | date: 2020-05-11, 23:30.

2.2.3 Mean

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a multinomial distribution (\rightarrow Definition II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = [np_1, \dots, np_k] . \quad (2)$$

Proof: By definition, a multinomial random variable (\rightarrow Definition II/2.2.1) is the sum of n independent and identical categorical trials (\rightarrow Definition II/2.1.1) with category probabilities p_1, \dots, p_k . 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 Proof I/1.7.5), 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 Proof II/2.1.3), we have:

$$E(X) = \sum_{i=1}^n [p_1, \dots, p_k] = n \cdot [p_1, \dots, p_k] = [np_1, \dots, np_k] . \quad (5)$$

Sources:

- original work

Metadata: ID: P25 | shortcut: mult-mean | author: JoramSoch | date: 2020-01-16, 11:26.

3 Univariate continuous distributions

3.1 Continuous uniform distribution

3.1.1 Definition

Definition: Let X be a continuous random variable (\rightarrow Definition 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)).

Metadata: ID: D3 | shortcut: cuni | author: JoramSoch | date: 2020-01-27, 14:05.

3.1.2 Standard uniform distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to be standard uniformly distributed, if X follows a continuous uniform distribution (\rightarrow Definition 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.

Metadata: ID: D157 | shortcut: suni | author: JoramSoch | date: 2021-07-23, 17:32.

3.1.3 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition 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 Definition I/1.6.6), 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:

$$\begin{aligned} \int_a^b 1 \, dx &= c(a, b) \\ [x]_a^b &= c(a, b) \\ c(a, b) &= b - a . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P37 | shortcut: cuni-pdf | author: JoramSoch | date: 2020-01-31, 15:41.

3.1.4 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof 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 Definition I/1.6.13) 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)

$$\begin{aligned} F_X(x) &= \int_{-\infty}^a \mathcal{U}(z; a, b) \, dz + \int_a^x \mathcal{U}(z; a, b) \, dz \\ &= \int_{-\infty}^a 0 \, dz + \int_a^x \frac{1}{b-a} \, dz \\ &= 0 + \frac{1}{b-a} [z]_a^x \\ &= \frac{x-a}{b-a} . \end{aligned} \quad (6)$$

Finally, if $x > b$, we have

$$\begin{aligned} F_X(x) &= \int_{-\infty}^b \mathcal{U}(z; a, b) \, dz + \int_b^x \mathcal{U}(z; a, b) \, dz \\ &= F_X(b) + \int_b^x 0 \, dz \\ &= \frac{b-a}{b-a} + 0 \\ &= 1 . \end{aligned} \quad (7)$$

This completes the proof.

Sources:

- original work

Metadata: ID: P38 | shortcut: cuni-cdf | author: JoramSoch | date: 2020-01-02, 18:05.

3.1.5 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the quantile function (\rightarrow Definition I/1.6.23) of X is

$$Q_X(p) = \begin{cases} -\infty , & \text{if } p = 0 \\ bp + a(1-p) , & \text{if } p > 0 . \end{cases} \quad (2)$$

Proof: The cumulative distribution function of the continuous uniform distribution (\rightarrow Proof II/3.1.4) is:

$$F_X(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b. \end{cases} \quad (3)$$

The quantile function $Q_X(p)$ is defined as (\rightarrow Definition I/1.6.23) 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 Proof I/1.6.24)

$$Q_X(p) = F_X^{-1}(x) . \quad (5)$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{x-a}{b-a} \\ x &= p(b-a) + a \\ x &= bp + a(1-p) . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P39 | shortcut: cuni-qf | author: JoramSoch | date: 2020-01-02, 18:27.

3.1.6 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$\mathbb{E}(X) = \frac{1}{2}(a+b) . \quad (2)$$

Proof: The expected value (\rightarrow Definition I/1.7.1) 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 continuous uniform distribution (\rightarrow Proof II/3.1.3), this becomes:

$$\begin{aligned}
E(X) &= \int_a^b x \cdot \frac{1}{b-a} dx \\
&= \left[\frac{1}{2} \frac{x^2}{b-a} \right]_a^b \\
&= \frac{1}{2} \frac{b^2 - a^2}{b-a} \\
&= \frac{1}{2} \frac{(b+a)(b-a)}{b-a} \\
&= \frac{1}{2}(a+b) .
\end{aligned} \tag{4}$$

Sources:

- original work

Metadata: ID: P82 | shortcut: cuni-mean | author: JoramSoch | date: 2020-03-16, 16:12.

3.1.7 Median

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \tag{1}$$

Then, the median (\rightarrow Definition I/1.11.1) of X is

$$\text{median}(X) = \frac{1}{2}(a+b) . \tag{2}$$

Proof: The median (\rightarrow Definition I/1.11.1) is the value at which the cumulative distribution function (\rightarrow Definition I/1.6.13) is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} . \tag{3}$$

The cumulative distribution function of the continuous uniform distribution (\rightarrow Proof II/3.1.4) is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < a \\ \frac{x-a}{b-a} , & \text{if } a \leq x \leq b \\ 1 , & \text{if } x > b . \end{cases} \tag{4}$$

Thus, the inverse CDF (\rightarrow Proof II/3.1.5) is

$$x = bp + a(1-p) . \tag{5}$$

Setting $p = 1/2$, we obtain:

$$\text{median}(X) = b \cdot \frac{1}{2} + a \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{2}(a+b) . \tag{6}$$

Sources:

- original work

Metadata: ID: P83 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:19.

3.1.8 Mode

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) . \quad (1)$$

Then, the mode (\rightarrow Definition I/1.11.2) of X is

$$\text{mode}(X) \in [a, b] . \quad (2)$$

Proof: The mode (\rightarrow Definition I/1.11.2) is the value which maximizes the probability density function (\rightarrow Definition I/1.6.6):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the continuous uniform distribution (\rightarrow Proof 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 .

Sources:

- original work

Metadata: ID: P84 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:29.

3.2 Normal distribution**3.2.1 Definition**

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to be normally distributed with mean μ and variance σ^2 (or, standard deviation σ)

$$X \sim \mathcal{N}(\mu, \sigma^2) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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.

Metadata: ID: D4 | shortcut: norm | author: JoramSoch | date: 2020-01-27, 14:15.

3.2.2 Standard normal distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to be standard normally distributed, if X follows a normal distribution (\rightarrow Definition 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.

Metadata: ID: D63 | shortcut: snorm | author: JoramSoch | date: 2020-05-26, 23:32.

3.2.3 Relationship to standard normal distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (\rightarrow Definition II/3.2.2) 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 σ is positive, $g(X)$ is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function (\rightarrow Proof I/1.6.15) 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 Proof II/3.2.11) 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 Definition I/1.6.13) of the standard normal distribution (\rightarrow Definition II/3.2.2).

Sources:

- original work

Metadata: ID: P111 | shortcut: norm-snorm | author: JoramSoch | date: 2020-05-26, 23:01.

3.2.4 Relationship to standard normal distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

$$X \sim \mathcal{N}(\mu, \sigma^2). \quad (1)$$

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (\rightarrow Definition II/3.2.2) 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 σ is positive, $g(X)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (\rightarrow Proof I/1.6.8) 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 Proof II/3.2.9), 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 Definition I/1.6.6) of the standard normal distribution (\rightarrow Definition II/3.2.2).

Sources:

- original work

Metadata: ID: P176 | shortcut: norm-snorm2 | author: JoramSoch | date: 2020-10-15, 11:42.

3.2.5 Relationship to standard normal distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (\rightarrow Definition II/3.2.2) 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 Proof II/4.1.5) states

$$x \sim \mathcal{N}(\mu, \Sigma) \Rightarrow y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \quad (3)$$

where x is an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) following a multivariate normal distribution (\rightarrow Definition II/4.1.1) with mean μ and covariance Σ , 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

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Z = \frac{X}{\sigma} - \frac{\mu}{\sigma} \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{1}{\sigma} \cdot \sigma^2 \cdot \frac{1}{\sigma}\right) \quad (5)$$

which results in the distribution:

$$Z \sim \mathcal{N}(0, 1) . \quad (6)$$

Sources:

- original work

Metadata: ID: P180 | shortcut: norm-snorm3 | author: JoramSoch | date: 2020-10-22, 06:34.

3.2.6 Relationship to chi-squared distribution

Theorem: Let X_1, \dots, X_n be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) where each of them is following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for } i = 1, \dots, n . \quad (1)$$

Define the sample mean (\rightarrow Definition I/1.7.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

and the unbiased sample variance (\rightarrow Definition I/1.8.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 . \quad (3)$$

Then, the sampling distribution (\rightarrow Definition I/1.5.5) of the sample variance is given by a chi-squared distribution (\rightarrow Definition II/3.6.1) 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 Definition I/1.2.2) U_i defined as

$$U_i = \frac{X_i - \mu}{\sigma} \quad (5)$$

which follows a standard normal distribution (\rightarrow Proof II/3.2.3)

$$U_i \sim \mathcal{N}(0, 1) . \quad (6)$$

Then, the sum of squared random variables U_i can be rewritten as

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \sum_{i=1}^n \frac{(\bar{X} - \mu)^2}{\sigma^2} + \sum_{i=1}^n \frac{(X_i - \bar{X})(\bar{X} - \mu)}{\sigma^2} \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma^2} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma^2} \right)^2 + \frac{(\bar{X} - \mu)}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}) . \end{aligned} \quad (7)$$

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.

$$\sum_{i=1}^n U_i^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma^2} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma^2} \right)^2 . \quad (9)$$

Cochran's theorem (\rightarrow Proof "snorm-cochran") states that, if a sum of squared standard normal (\rightarrow Definition II/3.2.2) random variables (\rightarrow Definition 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 Definition I/1.3.6) and each term Q_j follows a chi-squared distribution (\rightarrow Definition II/3.6.1) with r_j degrees of freedom:

$$Q_j \sim \chi^2(r_j) . \quad (11)$$

We observe that (9) can be represented as

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma^2} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma^2} \right)^2 \\ &= Q_1 + Q_2 = \sum_{i=1}^n \left(U_i - \frac{1}{n} \sum_{j=1}^n U_j \right)^2 + \frac{1}{n} \left(\sum_{i=1}^n U_i \right)^2 \end{aligned} \quad (12)$$

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 (\rightarrow Proof "snorm-cochran") are met and the squared form

$$Q_1 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma^2} \right)^2 = (n-1) \frac{1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = (n-1) \frac{s^2}{\sigma^2} \quad (14)$$

follows a chi-squared distribution (\rightarrow Definition II/3.6.1) 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.

Metadata: ID: P233 | shortcut: norm-chi2 | author: JoramSoch | date: 2021-05-20, 10:18.

3.2.7 Relationship to t-distribution

Theorem: Let X_1, \dots, X_n be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) where each of them is following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad i = 1, \dots, n . \quad (1)$$

Define the sample mean (\rightarrow Definition I/1.7.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

and the unbiased sample variance (\rightarrow Definition I/1.8.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 . \quad (3)$$

Then, subtracting μ from the sample mean (\rightarrow Definition I/1.7.1), dividing by the sample standard deviation (\rightarrow Definition I/1.12.1) and multiplying with \sqrt{n} results in a quantity that follows a t-distribution (\rightarrow Definition 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 Proof II/3.2.24), 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 Proof II/4.1.5) 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 Proof II/3.2.6) 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 Definition II/3.2.2) and the square root of a chi-squared random variable (\rightarrow Definition II/3.6.1), 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 Definition 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.
- 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.

Metadata: ID: P234 | shortcut: norm-t | author: JoramSoch | date: 2021-05-27, 08:10.

3.2.8 Gaussian integral

Theorem: The definite integral of $\exp[-x^2]$ from $-\infty$ to $+\infty$ is equal to the square root of π :

$$\int_{-\infty}^{+\infty} \exp[-x^2] \, dx = \sqrt{\pi} . \quad (1)$$

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:

$$\begin{aligned} \int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp[-r^2] r \, dr \, d\theta &\leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp[-r^2] r \, dr \, d\theta \\ \int_0^{\frac{\pi}{2}} \left[-\frac{1}{2} \exp[-r^2] \right]_0^{r_1} d\theta &\leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \left[-\frac{1}{2} \exp[-r^2] \right]_0^{r_2} d\theta \\ -\frac{1}{2} \int_0^{\frac{\pi}{2}} (\exp[-r_1^2] - 1) d\theta &\leq I_P^2 \leq -\frac{1}{2} \int_0^{\frac{\pi}{2}} (\exp[-r_2^2] - 1) d\theta \\ -\frac{1}{2} [(\exp[-r_1^2] - 1) \theta]_0^{\frac{\pi}{2}} &\leq I_P^2 \leq -\frac{1}{2} [(\exp[-r_2^2] - 1) \theta]_0^{\frac{\pi}{2}} \\ \frac{1}{2} (1 - \exp[-r_1^2]) \frac{\pi}{2} &\leq I_P^2 \leq \frac{1}{2} (1 - \exp[-r_2^2]) \frac{\pi}{2} \\ \frac{\pi}{4} (1 - \exp[-P^2]) &\leq I_P^2 \leq \frac{\pi}{4} (1 - \exp[-2P^2]) \end{aligned} \quad (9)$$

Calculating the limit for $P \rightarrow \infty$, we obtain

$$\begin{aligned} \lim_{P \rightarrow \infty} \frac{\pi}{4} (1 - \exp[-P^2]) &\leq \lim_{P \rightarrow \infty} I_P^2 \leq \lim_{P \rightarrow \infty} \frac{\pi}{4} (1 - \exp[-2P^2]) \\ \frac{\pi}{4} &\leq I^2 \leq \frac{\pi}{4} , \end{aligned} \quad (10)$$

such that we have a preliminary result for I :

$$I^2 = \frac{\pi}{4} \quad \Rightarrow \quad I = \frac{\sqrt{\pi}}{2} . \quad (11)$$

Because the integrand in (1) is an even function, we can calculate the final result as follows:

$$\begin{aligned} \int_{-\infty}^{+\infty} \exp[-x^2] \, dx &= 2 \int_0^{\infty} \exp[-x^2] \, dx \\ &\stackrel{(11)}{=} 2 \frac{\sqrt{\pi}}{2} \\ &= \sqrt{\pi} . \end{aligned} \quad (12)$$

Sources:

- ProofWiki (2020): “Gaussian Integral”; in: *ProofWiki*, retrieved on 2020-11-25; URL: 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.

Metadata: ID: P196 | shortcut: norm-gi | author: JoramSoch | date: 2020-11-25, 04:47.

3.2.9 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/3.2.1).

Sources:

- original work

Metadata: ID: P33 | shortcut: norm-pdf | author: JoramSoch | date: 2020-01-27, 15:15.

3.2.10 Moment-generating function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the moment-generating function (\rightarrow Definition I/1.6.27) 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 Proof II/3.2.9) 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 Definition I/1.6.27) is defined as

$$M_X(t) = \mathbb{E} [e^{tX}] . \quad (4)$$

Using the expected value for continuous random variables (\rightarrow Definition I/1.7.1), 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 Proof II/3.2.8)

$$\int_{-\infty}^{+\infty} \exp \left[-x^2 \right] dx = \sqrt{\pi} , \tag{8}$$

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.

Metadata: ID: P71 | shortcut: norm-mgf | author: JoramSoch | date: 2020-03-03, 11:29.

3.2.11 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distributions (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof II/3.2.9) 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 Definition I/1.6.13) 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:

$$\begin{aligned}
F_X(x) &= \frac{1}{2} \lim_{x \rightarrow \infty} \operatorname{erf}(x) + \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \\
&= \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \\
&= \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \right].
\end{aligned} \tag{7}$$

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.
- Wikipedia (2020): “Error function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Error_function.

Metadata: ID: P85 | shortcut: norm-cdf | author: JoramSoch | date: 2020-03-20, 01:33.

3.2.12 Cumulative distribution function without error function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{1}$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) of X can be expressed as

$$f_X(x) = \Phi_{\mu, \sigma}(x) = \varphi\left(\frac{x - \mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x - \mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2} \tag{2}$$

where $\varphi(x)$ is the probability density function (\rightarrow Definition I/1.6.6) of the standard normal distribution (\rightarrow Definition II/3.2.2) and $n!!$ is a double factorial.

Proof:

1) First, consider the standard normal distribution (\rightarrow Definition II/3.2.2) $\mathcal{N}(0, 1)$ which has the probability density function (\rightarrow Proof II/3.2.9)

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}. \tag{3}$$

Let $T(x)$ be the indefinite integral of this function. It can be obtained using infinitely repeated integration by parts as follows:

$$\begin{aligned}
T(x) &= \int \varphi(x) \, dx \\
&= \int \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int 1 \cdot e^{-\frac{1}{2}x^2} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \int x^2 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{3}x^4 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{15}x^5 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{15}x^6 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \right] \\
&= \dots \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^n \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \lim_{n \rightarrow \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right].
\end{aligned} \tag{4}$$

Since $(2n-1)!!$ grows faster than x^{2n} , it holds that

$$\frac{1}{\sqrt{2\pi}} \cdot \lim_{n \rightarrow \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx = \int 0 \, dx = c \tag{5}$$

for constant c , such that the indefinite integral becomes

$$\begin{aligned}
T(x) &= \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + c \\
&= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c \\
&\stackrel{(3)}{=} \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c.
\end{aligned} \tag{6}$$

2) Next, let $\Phi(x)$ be the cumulative distribution function (\rightarrow Definition I/1.6.13) of the standard normal distribution (\rightarrow Definition II/3.2.2):

$$\Phi(x) = \int_{-\infty}^x \varphi(x) \, dx. \tag{7}$$

It can be obtained by matching $T(0)$ to $\Phi(0)$ which is $1/2$, because the standard normal distribution is symmetric around zero:

$$\begin{aligned}
T(0) &= \varphi(0) \cdot \sum_{i=1}^{\infty} \frac{0^{2i-1}}{(2i-1)!!} + c = \frac{1}{2} = \Phi(0) \\
&\Leftrightarrow c = \frac{1}{2} \\
\Rightarrow \Phi(x) &= \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + \frac{1}{2}.
\end{aligned} \tag{8}$$

3) Finally, the cumulative distribution functions (\rightarrow Definition I/1.6.13) of the standard normal distribution (\rightarrow Definition II/3.2.2) and the general normal distribution (\rightarrow Definition II/3.2.1) are related to each other (\rightarrow Proof II/3.2.3) as

$$\Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{9}$$

Combining (9) with (8), we have:

$$\Phi_{\mu,\sigma}(x) = \varphi\left(\frac{x-\mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x-\mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2}. \tag{10}$$

Sources:

- Soch J (2015): “Solution for the Indefinite Integral of the Standard Normal Probability Density Function”; in: *arXiv stat.OT*, arXiv: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.

Metadata: ID: P86 | shortcut: norm-cdfwerf | author: JoramSoch | date: 2020-03-20, 04:26.

3.2.13 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distributions (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{1}$$

Then, the quantile function (\rightarrow Definition I/1.6.23) of X is

$$Q_X(p) = \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p-1) + \mu \tag{2}$$

where $\operatorname{erf}^{-1}(x)$ is the inverse error function.

Proof: The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.11) is:

$$F_X(x) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \right]. \tag{3}$$

Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF (\rightarrow Proof I/1.6.24):

$$Q_X(p) = F_X^{-1}(x) . \quad (4)$$

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.

Metadata: ID: P87 | shortcut: norm-qf | author: JoramSoch | date: 2020-03-20, 04:47.

3.2.14 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$\mathbb{E}(X) = \mu . \quad (2)$$

Proof: The expected value (\rightarrow Definition I/1.7.1) 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 normal distribution (\rightarrow Proof II/3.2.9), this reads:

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right] \, dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right] \, dx . \end{aligned} \quad (4)$$

Substituting $z = x - \mu$, we have:

$$\begin{aligned}
E(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z + \mu) \cdot \exp \left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] d(z + \mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z + \mu) \cdot \exp \left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] dz \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp \left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] dz + \mu \int_{-\infty}^{+\infty} \exp \left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] dz \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp \left[-\frac{1}{2\sigma^2} \cdot z^2 \right] dz + \mu \int_{-\infty}^{+\infty} \exp \left[-\frac{1}{2\sigma^2} \cdot z^2 \right] dz \right) .
\end{aligned} \tag{5}$$

The general antiderivatives are

$$\begin{aligned}
\int x \cdot \exp [-ax^2] dx &= -\frac{1}{2a} \cdot \exp [-ax^2] \\
\int \exp [-ax^2] dx &= \frac{1}{2} \sqrt{\frac{\pi}{a}} \cdot \operatorname{erf} [\sqrt{a}x]
\end{aligned} \tag{6}$$

where $\operatorname{erf}(x)$ is the error function. Using this, the integrals can be calculated as:

$$\begin{aligned}
E(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[-\sigma^2 \cdot \exp \left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right]_{-\infty}^{+\infty} + \mu \left[\sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[\frac{1}{\sqrt{2}\sigma} z \right] \right]_{-\infty}^{+\infty} \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[\lim_{z \rightarrow \infty} \left(-\sigma^2 \cdot \exp \left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) - \lim_{z \rightarrow -\infty} \left(-\sigma^2 \cdot \exp \left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) \right] \right. \\
&\quad \left. + \mu \left[\lim_{z \rightarrow \infty} \left(\sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[\frac{1}{\sqrt{2}\sigma} z \right] \right) - \lim_{z \rightarrow -\infty} \left(\sqrt{\frac{\pi}{2}} \sigma \cdot \operatorname{erf} \left[\frac{1}{\sqrt{2}\sigma} z \right] \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left([0 - 0] + \mu \left[\sqrt{\frac{\pi}{2}} \sigma - \left(-\sqrt{\frac{\pi}{2}} \sigma \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2\sqrt{\frac{\pi}{2}} \sigma \\
&= \mu .
\end{aligned} \tag{7}$$

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>.

Metadata: ID: P15 | shortcut: norm-mean | author: JoramSoch | date: 2020-01-09, 15:04.

3.2.15 Median

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the median (\rightarrow Definition I/1.11.1) of X is

$$\text{median}(X) = \mu . \quad (2)$$

Proof: The median (\rightarrow Definition I/1.11.1) is the value at which the cumulative distribution function (\rightarrow Definition I/1.6.13) is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.11) 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:

$$\text{median}(X) = \sqrt{2}\sigma \cdot \text{erf}^{-1}(0) + \mu = \mu . \quad (6)$$

Sources:

- original work

Metadata: ID: P16 | shortcut: norm-med | author: JoramSoch | date: 2020-01-09, 15:33.

3.2.16 Mode

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mode (\rightarrow Definition I/1.11.2) of X is

$$\text{mode}(X) = \mu . \quad (2)$$

Proof: The mode (\rightarrow Definition I/1.11.2) is the value which maximizes the probability density function (\rightarrow Definition I/1.6.6):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the normal distribution (\rightarrow Proof II/3.2.9) 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

$$\text{mode}(X) = \mu . \quad (9)$$

Sources:

- original work

Metadata: ID: P17 | shortcut: norm-mode | author: JoramSoch | date: 2020-01-09, 15:58.

3.2.17 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

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

Proof: The variance (\rightarrow Definition I/1.8.1) is the probability-weighted average of the squared deviation from the mean (\rightarrow Definition I/1.7.1):

$$\text{Var}(X) = \int_{\mathbb{R}} (x - E(X))^2 \cdot f_X(x) dx . \quad (3)$$

With the expected value (\rightarrow Proof II/3.2.14) and probability density function (\rightarrow Proof II/3.2.9) 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>.

Metadata: ID: P18 | shortcut: norm-var | author: JoramSoch | date: 2020-01-09, 22:47.

3.2.18 Full width at half maximum

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the full width at half maximum (\rightarrow Definition I/1.12.2) (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 Proof II/3.2.9) 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 Proof II/3.2.16) 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 Definition I/1.12.2)

$$f_X(x_{\text{FWHM}}) = \frac{1}{2} f_{\max} \stackrel{(5)}{=} \frac{1}{2\sqrt{2\pi}\sigma} . \quad (6)$$

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_{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 Definition I/1.12.2) 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.

Metadata: ID: P152 | shortcut: norm-fwhm | author: JoramSoch | date: 2020-08-19, 06:39.

3.2.19 Extreme points

Theorem: The probability density function (\rightarrow Definition I/1.6.6) of the normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 has a maximum at $x = \mu$ and no other extrema. Consequently, the normal distribution (\rightarrow Definition II/3.2.1) is a unimodal probability distribution (\rightarrow Definition “dist-uni”).

Proof: The probability density function of the normal distribution (\rightarrow Proof II/3.2.9) 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 Proof II/3.2.16) 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 Proof II/3.2.16).

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.

Metadata: ID: P251 | shortcut: norm-extr | author: JoramSoch | date: 2020-08-25, 21:11.

3.2.20 Inflection points

Theorem: The probability density function (\rightarrow Definition I/1.6.6) of the normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 has two inflection points at $x = \mu - \sigma$ and $x = \mu + \sigma$, i.e. exactly one standard deviation (\rightarrow Definition I/1.12.1) away from the expected value (\rightarrow Definition I/1.7.1).

Proof: The probability density function of the normal distribution (\rightarrow Proof II/3.2.9) 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 μ is the mean and σ^2 is the variance of a normal distribution (\rightarrow Definition II/3.2.1), these points are exactly one standard deviation (\rightarrow Definition I/1.12.1) 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.

Metadata: ID: P252 | shortcut: norm-infl | author: JoramSoch | date: 2020-08-26, 12:26.

3.2.21 Differential entropy

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the differential entropy (\rightarrow Definition 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 Definition 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 Definition I/1.7.1)

$$h(X) = -E[\ln p(x)] . \quad (4)$$

With the probability density function of the normal distribution (\rightarrow Proof II/3.2.9), the differential entropy of X is:

$$\begin{aligned} h(X) &= -E \left[\ln \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \right) \right] \\ &= -E \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} E \left[\left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot E[(x-\mu)^2] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \sigma^2 \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2} \ln(2\pi\sigma^2 e) . \end{aligned} \quad (5)$$

Sources:

- Wang, Peng-Hua (2012): “Differential Entropy”; in: *National Taipei University*; URL: <https://web.ntpu.edu.tw/~phwang/teaching/2012s/IT/slides/chap08.pdf>.

Metadata: ID: P101 | shortcut: norm-dent | author: JoramSoch | date: 2020-05-14, 20:09.

3.2.22 Kullback-Leibler divergence

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2). Assume two normal distributions (\rightarrow Definition 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 Definition 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 Definition 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 Definition 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 Proof II/3.2.9), 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 Definition I/1.7.1) is a linear operator (\rightarrow Proof I/1.7.5), 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 Definition I/1.8.1)

$$\langle (X - \mu)^2 \rangle = \text{E}[(X - \text{E}(X))^2] = \text{Var}(X) \quad (7)$$

and the variance of a normally distributed random variable (\rightarrow Proof II/3.2.17) is

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \text{Var}(X) = \sigma^2 . \quad (8)$$

Additionally applying the raw moments of the normal distribution (\rightarrow Proof II/3.2.10)

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \langle x \rangle = \mu \quad \text{and} \quad \langle x^2 \rangle = \mu^2 + \sigma^2 , \quad (9)$$

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).

Sources:

- original work

Metadata: ID: P193 | shortcut: norm-kl | author: JoramSoch | date: 2020-11-19, 07:08.

3.2.23 Maximum entropy distribution

Theorem: The normal distribution (\rightarrow Definition II/3.2.1) maximizes differential entropy (\rightarrow Definition I/2.2.1) for a random variable (\rightarrow Definition I/1.2.2) with fixed variance (\rightarrow Definition I/1.8.1).

Proof: For a random variable (\rightarrow Definition I/1.2.2) X with set of possible values with probability density function (\rightarrow Definition I/1.6.6) $f(x)$, the differential entropy (\rightarrow Definition 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 Definition I/1.6.6) of a normal distribution (\rightarrow Definition II/3.2.1) with mean (\rightarrow Definition I/1.7.1) μ and variance (\rightarrow Definition I/1.8.1) σ^2 and let $f(x)$ be an arbitrary probability density function (\rightarrow Definition I/1.6.6) with the same variance (\rightarrow Definition I/1.8.1). Since differential entropy (\rightarrow Definition I/2.2.1) is translation-invariant (\rightarrow Proof I/2.2.3), we can assume that $f(x)$ has the same mean as $g(x)$.

Consider the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of distribution $f(x)$ from distribution $g(x)$ which is non-negative (\rightarrow Proof 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 Proof II/3.2.9) into the second term, we obtain:

$$\begin{aligned} \int_{\mathcal{X}} f(x) \log g(x) dx &= \int_{\mathcal{X}} f(x) \log \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \right) dx \\ &= \int_{\mathcal{X}} f(x) \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) dx + \int_{\mathcal{X}} f(x) \log \left(\exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \right) dx \quad (3) \\ &= -\frac{1}{2} \log(2\pi\sigma^2) \int_{\mathcal{X}} f(x) dx - \frac{\log(e)}{2\sigma^2} \int_{\mathcal{X}} f(x)(x-\mu)^2 dx . \end{aligned}$$

Because the entire integral over a probability density function is one (\rightarrow Definition I/1.6.6) and the second central moment is equal to the variance (\rightarrow Proof I/1.14.8), 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 Proof II/3.2.21), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) dx = -h[g(x)] . \quad (5)$$

Combining (2) with (5), we can show that

$$\begin{aligned} 0 &\leq -h[f(x)] - (-h[g(x)]) \\ h[g(x)] - h[f(x)] &\geq 0 \end{aligned} \quad (6)$$

which means that the differential entropy (\rightarrow Definition I/2.2.1) of the normal distribution (\rightarrow Definition II/3.2.1) $\mathcal{N}(\mu, \sigma^2)$ will be larger than or equal to any other distribution (\rightarrow Definition I/1.5.1) with the same variance (\rightarrow Definition I/1.8.1) σ^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.

Metadata: ID: P250 | shortcut: norm-maxent | author: JoramSoch | date: 2020-08-25, 08:31.

3.2.24 Linear combination

Theorem: Let X_1, \dots, X_n be independent (\rightarrow Definition I/1.3.6) normally distributed (\rightarrow Definition II/3.2.1) random variables (\rightarrow Definition I/1.2.2) with means (\rightarrow Definition I/1.7.1) μ_1, \dots, μ_n and variances (\rightarrow Definition I/1.8.1) $\sigma_1^2, \dots, \sigma_n^2$:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{for } i = 1, \dots, n. \quad (1)$$

Then, any linear combination of those random variables

$$Y = \sum_{i=1}^n a_i X_i \quad \text{where } a_1, \dots, a_n \in \mathbb{R} \quad (2)$$

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 Proof II/4.1.8) to an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) x following a multivariate normal distribution (\rightarrow Definition II/4.1.1) with a diagonal covariance matrix (\rightarrow Definition I/1.9.7). 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 Proof II/4.1.5)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \quad (6)$$

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 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

$$\begin{aligned}
A\mu &= [a_1, \dots, a_n] \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \sum_{i=1}^n a_i \mu_i \quad \text{and} \\
A\Sigma A^T &= [a_1, \dots, a_n] \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^n a_i^2 \sigma_i^2.
\end{aligned} \tag{9}$$

Sources:

- original work

Metadata: ID: P235 | shortcut: norm-lincomb | author: JoramSoch | date: 2021-06-02, 08:24.

3.3 t-distribution

3.3.1 Definition

Definition: Let Z and V be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) following a standard normal distribution (\rightarrow Definition II/3.2.2) and a chi-squared distribution (\rightarrow Definition II/3.6.1) with ν degrees of freedom (\rightarrow Definition “dof”), respectively:

$$\begin{aligned}
Z &\sim \mathcal{N}(0, 1) \\
V &\sim \chi^2(\nu).
\end{aligned} \tag{1}$$

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 ν :

$$Y = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu). \tag{2}$$

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.

Metadata: ID: D147 | shortcut: t | author: JoramSoch | date: 2021-04-21, 07:53.

3.3.2 Non-standardized t-distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2) following a Student’s t -distribution (\rightarrow Definition II/3.3.1) with ν degrees of freedom. Then, the random variable (\rightarrow Definition I/1.2.2)

$$Y = \sigma X + \mu \tag{1}$$

is said to follow a non-standardized t-distribution with non-centrality μ , scale σ^2 and degrees of freedom ν :

$$Y \sim \text{nst}(\mu, \sigma^2, \nu) . \quad (2)$$

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.

Metadata: ID: D152 | shortcut: nst | author: JoramSoch | date: 2021-05-20, 07:35.

3.3.3 Relationship to non-standardized t-distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a non-standardized t-distribution (\rightarrow Definition II/3.3.2) with mean μ , scale σ^2 and degrees of freedom ν :

$$X \sim \text{nst}(\mu, \sigma^2, \nu) . \quad (1)$$

Then, subtracting the mean and dividing by the square root of the scale results in a random variable (\rightarrow Definition I/1.2.2) following a t-distribution (\rightarrow Definition II/3.3.1) with degrees of freedom ν :

$$Y = \frac{X - \mu}{\sigma} \sim t(\nu) . \quad (2)$$

Proof: The non-standardized t-distribution is a special case (\rightarrow Proof “nst-mvt”) of the multivariate t-distribution (\rightarrow Definition II/4.2.1) in which the mean vector and scale matrix are scalars:

$$X \sim \text{nst}(\mu, \sigma^2, \nu) \quad \Rightarrow \quad X \sim t(\mu, \sigma^2, \nu) . \quad (3)$$

Therefore, we can apply the linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”) for an $n \times 1$ random vector x :

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad y = Ax + b \sim t(A\mu + b, A\Sigma A^T, \nu) . \quad (4)$$

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 Proof “mvt-pdf”),

$$p(x) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu + n]/2)}{\Gamma(\nu/2)} \left[1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] , \quad (6)$$

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 Proof II/3.3.4) with ν degrees of freedom.

Sources:

- original work

Metadata: ID: P232 | shortcut: nst-t | author: JoramSoch | date: 2021-05-11, 15:46.

3.3.4 Probability density function

Theorem: Let T be a random variable (\rightarrow Definition I/1.2.2) following a t-distribution (\rightarrow Definition II/3.3.1):

$$T \sim t(\nu) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/3.3.1) is defined as the ratio of a standard normal random variable (\rightarrow Definition II/3.2.2) and the square root of a chi-squared random variable (\rightarrow Definition II/3.6.1), divided by its degrees of freedom (\rightarrow Definition “dof”)

$$X \sim \mathcal{N}(0, 1), Y \sim \chi^2(\nu) \quad \Rightarrow \quad T = \frac{X}{\sqrt{Y/\nu}} \sim t(\nu) \quad (3)$$

where X and Y are independent of each other (\rightarrow Definition I/1.3.6).

The probability density function (\rightarrow Proof II/3.2.9) of the standard normal distribution (\rightarrow Definition II/3.2.2) 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 Proof II/3.6.3) 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:

$$\begin{aligned} J &= \begin{bmatrix} \frac{dX}{dT} & \frac{dX}{dW} \\ \frac{dY}{dT} & \frac{dY}{dW} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{W}{\nu}} & \frac{T}{2\sqrt{W/\nu}} \\ 0 & 1 \end{bmatrix} \\ |J| &= \sqrt{\frac{W}{\nu}}. \end{aligned} \quad (8)$$

Because X and Y are independent (\rightarrow Definition I/1.3.6), the joint density (\rightarrow Definition I/1.5.2) of X and Y is equal to the product (\rightarrow Proof I/1.3.8) of the marginal densities (\rightarrow Definition 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 Proof I/1.6.10), the joint density (\rightarrow Definition 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:

$$\begin{aligned} f_{T,W}(t, w) &= f_X\left(t \cdot \sqrt{\frac{w}{\nu}}\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(t \cdot \sqrt{\frac{w}{\nu}})^2}{2}} \cdot \frac{1}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \sqrt{\frac{w}{\nu}} \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu+1}{2}-1} \cdot e^{-\frac{w}{2}\left(\frac{t^2}{\nu}+1\right)}. \end{aligned} \quad (11)$$

The marginal density (\rightarrow Definition I/1.5.3) of T can now be obtained by integrating out (\rightarrow Definition I/1.3.3) W :

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,W}(t, w) dw \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \int_0^\infty w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)w\right] dw \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right]^{(\nu+1)/2}} \cdot \int_0^\infty \frac{\left[\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right]^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \cdot w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)w\right] dw \end{aligned} \quad (12)$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution (\rightarrow Proof II/3.4.5) with

$$a = \frac{\nu + 1}{2} \quad \text{and} \quad b = \frac{1}{2} \left(\frac{t^2}{\nu} + 1 \right), \quad (13)$$

and because a probability density function integrates to one (\rightarrow Definition I/1.6.6), we finally have:

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2} \left(\frac{t^2}{\nu} + 1\right)\right]^{(\nu+1)/2}} \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \cdot \sqrt{\nu\pi}} \cdot \left(\frac{t^2}{\nu} + 1\right)^{-\frac{\nu+1}{2}}. \end{aligned} \quad (14)$$

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>.

Metadata: ID: P263 | shortcut: t-pdf | author: JoramSoch | date: 2021-10-12, 08:15.

3.4 Gamma distribution

3.4.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to follow a gamma distribution with shape a and rate b

$$X \sim \text{Gam}(a, b), \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) is given by

$$\text{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx], \quad x > 0 \quad (2)$$

where $a > 0$ and $b > 0$, 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.

Metadata: ID: D7 | shortcut: gam | author: JoramSoch | date: 2020-02-08, 23:29.

3.4.2 Standard gamma distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to have a standard gamma distribution, if X follows a gamma distribution (\rightarrow Definition 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; 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

Metadata: ID: D64 | shortcut: sgam | author: JoramSoch | date: 2020-05-26, 23:36.

3.4.3 Relationship to standard gamma distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition 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 Definition II/3.4.2) 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 is positive, $g(X)$ is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function (\rightarrow Proof I/1.6.15) 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 Proof II/3.4.6) 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

$$\begin{aligned}
F_Y(y) &= \int_{-\infty}^{b(y/b)} \frac{b^a}{\Gamma(a)} \left(\frac{s}{b}\right)^{a-1} \exp\left[-b\left(\frac{s}{b}\right)\right] d\left(\frac{s}{b}\right) \\
&= \int_{-\infty}^y \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}b} s^{a-1} \exp[-s] ds \\
&= \int_{-\infty}^y \frac{1}{\Gamma(a)} s^{a-1} \exp[-s] ds
\end{aligned} \tag{8}$$

which is the cumulative distribution function (\rightarrow Definition I/1.6.13) of the standard gamma distribution (\rightarrow Definition II/3.4.2).

Sources:

- original work

Metadata: ID: P112 | shortcut: gam-sgam | author: JoramSoch | date: 2020-05-26, 23:14.

3.4.4 Relationship to standard gamma distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1) with shape a and rate b :

$$X \sim \text{Gam}(a, b) . \tag{1}$$

Then, the quantity $Y = bX$ will have a standard gamma distribution (\rightarrow Definition II/3.4.2) with shape a and rate 1:

$$Y = bX \sim \text{Gam}(a, 1) . \tag{2}$$

Proof: Note that Y is a function of X

$$Y = g(X) = bX \tag{3}$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{b}Y . \tag{4}$$

Because b is positive, $g(X)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (\rightarrow Proof I/1.6.8) as

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} , & \text{if } y \in \mathcal{Y} \\ 0 , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{5}$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$. With the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), we have

$$\begin{aligned}
f_Y(y) &= \frac{b^a}{\Gamma(a)} [g^{-1}(y)]^{a-1} \exp[-b g^{-1}(y)] \cdot \frac{dg^{-1}(y)}{dy} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{b}y\right)^{a-1} \exp\left[-b\left(\frac{1}{b}y\right)\right] \cdot \frac{d\left(\frac{1}{b}y\right)}{dy} \\
&= \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}} y^{a-1} \exp[-y] \cdot \frac{1}{b} \\
&= \frac{1}{\Gamma(a)} y^{a-1} \exp[-y]
\end{aligned} \tag{6}$$

which is the probability density function (\rightarrow Definition I/1.6.6) of the standard gamma distribution (\rightarrow Definition II/3.4.2).

Sources:

- original work

Metadata: ID: P177 | shortcut: gam-sgam2 | author: JoramSoch | date: 2020-10-15, 12:04.

3.4.5 Probability density function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \tag{1}$$

Then, the probability density function (\rightarrow Definition I/1.6.6) of X is

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] . \tag{2}$$

Proof: This follows directly from the definition of the gamma distribution (\rightarrow Definition II/3.4.1).

Sources:

- original work

Metadata: ID: P45 | shortcut: gam-pdf | author: JoramSoch | date: 2020-02-08, 23:41.

3.4.6 Cumulative distribution function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \tag{1}$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) of X is

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} \tag{2}$$

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 Proof II/3.4.5) is:

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] . \quad (3)$$

Thus, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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.

Metadata: ID: P178 | shortcut: gam-cdf | author: JoramSoch | date: 2020-10-15, 12:34.

3.4.7 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the quantile function (\rightarrow Definition I/1.6.23) 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 Proof II/3.4.6) 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 $Q_X(p)$ is defined as (\rightarrow Definition I/1.6.23) 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 Proof I/1.6.24)

$$Q_X(p) = F_X^{-1}(x). \quad (5)$$

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.

Metadata: ID: P194 | shortcut: gam-qf | author: JoramSoch | date: 2020-11-19, 07:31.

3.4.8 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b). \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = \frac{a}{b}. \quad (2)$$

Proof: The expected value (\rightarrow Definition I/1.7.1) 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 Proof II/3.4.5), 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 Proof II/3.4.5), 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>.

Metadata: ID: P108 | shortcut: gam-mean | author: JoramSoch | date: 2020-05-19, 06:54.

3.4.9 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

$$\text{Var}(X) = \frac{a}{b^2} . \quad (2)$$

Proof: The variance (\rightarrow Definition I/1.8.1) can be expressed in terms of expected values (\rightarrow Proof I/1.8.3) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

The expected value of a gamma random variable (\rightarrow Proof II/3.4.8) is

$$E(X) = \frac{a}{b} . \quad (4)$$

With the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), 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 Proof II/3.4.5), 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>.

Metadata: ID: P109 | shortcut: gam-var | author: JoramSoch | date: 2020-05-19, 07:20.

3.4.10 Logarithmic expectation

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the expectation (\rightarrow Definition I/1.7.1) of the natural logarithm of X is

$$E(\ln X) = \psi(a) - \ln(b) \quad (2)$$

where $\psi(x)$ is the digamma function.

Proof: Let $Y = \ln(X)$, such that $E(Y) = E(\ln X)$ and consider the special case that $b = 1$. In this case, the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) 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 $Y = \ln(X)$:

$$\begin{aligned} E(Y) &= \int_{\mathbb{R}} y f_Y(y) dy \\ &\stackrel{(7)}{=} \frac{1}{\Gamma(a)} \int_{\mathbb{R}} \frac{d}{da} \exp[ay - e^y] dy \\ &= \frac{1}{\Gamma(a)} \frac{d}{da} \int_{\mathbb{R}} \exp[ay - e^y] dy \\ &\stackrel{(6)}{=} \frac{1}{\Gamma(a)} \frac{d}{da} \Gamma(a) \\ &= \frac{\Gamma'(a)}{\Gamma(a)} . \end{aligned} \quad (8)$$

Using the derivative of a logarithmized function

$$\frac{d}{dx} \ln f(x) = \frac{f'(x)}{f(x)} \quad (9)$$

and the definition of the digamma function

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) , \quad (10)$$

we have

$$E(Y) = \psi(a) . \quad (11)$$

Finally, noting that $1/b$ acts as a scaling parameter (\rightarrow Proof II/3.4.3) on a gamma-distributed (\rightarrow Definition II/3.4.1) random variable (\rightarrow Definition I/1.2.2),

$$X \sim \text{Gam}(a, 1) \quad \Rightarrow \quad \frac{1}{b}X \sim \text{Gam}(a, b) , \quad (12)$$

and that a scaling parameter acts additively on the logarithmic expectation of a random variable,

$$E[\ln(cX)] = E[\ln(X) + \ln(c)] = E[\ln(X)] + \ln(c) , \quad (13)$$

it follows that

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad E(\ln X) = \psi(a) - \ln(b) . \quad (14)$$

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>.

Metadata: ID: P110 | shortcut: gam-logmean | author: JoramSoch | date: 2020-05-25, 21:28.

3.4.11 Expectation of $x \ln x$

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of $(X \cdot \ln X)$ is

$$E(X \ln X) = \frac{a}{b} [\psi(a) - \ln(b)] . \quad (2)$$

Proof: With the definition of the expected value (\rightarrow Definition I/1.7.1), the law of the unconscious statistician (\rightarrow Proof I/1.7.11) and the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), we have:

$$\begin{aligned}
E(X \ln X) &= \int_0^\infty x \ln x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx \\
&= \frac{1}{\Gamma(a)} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{b} x^a \exp[-bx] dx \\
&= \frac{\Gamma(a+1)}{\Gamma(a)b} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] dx
\end{aligned} \tag{3}$$

The integral now corresponds to the logarithmic expectation of a gamma distribution (\rightarrow Proof II/3.4.10) with shape $a+1$ and rate b

$$E(\ln Y) \quad \text{where} \quad Y \sim \text{Gam}(a+1, b) \tag{4}$$

which is given by (\rightarrow Proof II/3.4.10)

$$E(\ln Y) = \psi(a+1) - \ln(b) \tag{5}$$

where $\psi(x)$ is the digamma function. Additionally employing the relation

$$\Gamma(x+1) = \Gamma(x) \cdot x \quad \Leftrightarrow \quad \frac{\Gamma(x+1)}{\Gamma(x)} = x, \tag{6}$$

the expression in equation (3) develops into:

$$E(X \ln X) = \frac{a}{b} [\psi(a) - \ln(b)] . \tag{7}$$

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>.

Metadata: ID: P179 | shortcut: gam-xlogx | author: JoramSoch | date: 2020-10-15, 13:02.

3.4.12 Differential entropy

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a gamma distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Gam}(a, b) \tag{1}$$

Then, the differential entropy (\rightarrow Definition I/2.2.1) of X in nats is

$$h(X) = a + \ln \Gamma(a) + (1-a) \cdot \psi(a) + \ln b . \tag{2}$$

Proof: The differential entropy (\rightarrow Definition 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 (\rightarrow Definition I/1.7.1)

$$h(X) = -E[\ln p(x)] . \quad (4)$$

With the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), 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 Proof II/3.4.8) and logarithmic expectation (\rightarrow Proof II/3.4.10) of the gamma distribution (\rightarrow Definition 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 Definition 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.

Metadata: ID: P239 | shortcut: gam-dent | author: JoramSoch | date: 2021-07-14, 07:37.

3.4.13 Kullback-Leibler divergence

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2). Assume two gamma distributions (\rightarrow Definition 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 Definition I/2.5.1) of P from Q is given by

$$\text{KL}[P || Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} . \quad (2)$$

Proof: The KL divergence for a continuous random variable (\rightarrow Definition I/2.5.1) is given by

$$\text{KL}[P || Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the gamma distributions (\rightarrow Definition 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 Proof II/3.4.5), 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 Proof II/3.4.8) and the expected value of a logarithmized gamma variate (\rightarrow Proof II/3.4.10)

$$\begin{aligned} x \sim \text{Gam}(a, b) \quad \Rightarrow \quad \langle x \rangle &= \frac{a}{b} \quad \text{and} \\ \langle \ln x \rangle &= \psi(a) - \ln(b), \end{aligned} \quad (6)$$

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:

$$\text{KL}[P \parallel Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1}. \quad (8)$$

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>.

Metadata: ID: P93 | shortcut: gam-kl | author: JoramSoch | date: 2020-05-05, 08:41.

3.5 Exponential distribution

3.5.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to be exponentially distributed with rate (or, inverse scale) λ

$$X \sim \text{Exp}(\lambda) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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.

Metadata: ID: D8 | shortcut: exp | author: JoramSoch | date: 2020-02-08, 23:48.

3.5.2 Special case of gamma distribution

Theorem: The exponential distribution (\rightarrow Definition II/3.5.1) is a special case of the gamma distribution (\rightarrow Definition II/3.4.1) with shape $a = 1$ and rate $b = \lambda$.

Proof: The probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) 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 Proof II/3.5.3).

Sources:

- original work

Metadata: ID: P69 | shortcut: exp-gam | author: JoramSoch | date: 2020-03-02, 20:49.

3.5.3 Probability density function

Theorem: Let X be a non-negative random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) of X is

$$f_X(x) = \lambda \exp[-\lambda x] . \quad (2)$$

Proof: This follows directly from the definition of the exponential distribution (\rightarrow Definition II/3.5.1).

Sources:

- original work

Metadata: ID: P46 | shortcut: exp-pdf | author: JoramSoch | date: 2020-02-08, 23:53.

3.5.4 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) of X is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (2)$$

Proof: The probability density function of the exponential distribution (\rightarrow Proof II/3.5.3) is:

$$\text{Exp}(x; \lambda) = \begin{cases} 0 , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (3)$$

Thus, the cumulative distribution function (\rightarrow Definition I/1.6.13) is:

$$F_X(x) = \int_{-\infty}^x \text{Exp}(z; \lambda) dz . \quad (4)$$

If $x < 0$, we have:

$$F_X(x) = \int_{-\infty}^x 0 dz = 0 . \quad (5)$$

If $x \geq 0$, we have using (3):

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^0 \text{Exp}(z; \lambda) \, dz + \int_0^x \text{Exp}(z; \lambda) \, dz \\
&= \int_{-\infty}^0 0 \, dz + \int_0^x \lambda \exp[-\lambda z] \, dz \\
&= 0 + \lambda \left[-\frac{1}{\lambda} \exp[-\lambda z] \right]_0^x \\
&= \lambda \left[\left(-\frac{1}{\lambda} \exp[-\lambda x] \right) - \left(-\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\
&= 1 - \exp[-\lambda x] .
\end{aligned} \tag{6}$$

Sources:

- original work

Metadata: ID: P48 | shortcut: exp-cdf | author: JoramSoch | date: 2020-02-11, 14:48.

3.5.5 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \tag{1}$$

Then, the quantile function (\rightarrow Definition I/1.6.23) 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 Proof II/3.5.4) 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 $Q_X(p)$ is defined as (\rightarrow Definition I/1.6.23) 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 Proof I/1.6.24)

$$Q_X(p) = F_X^{-1}(x) . \tag{5}$$

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= 1 - \exp[-\lambda x] \\
\exp[-\lambda x] &= 1 - p \\
-\lambda x &= \ln(1 - p) \\
x &= -\frac{\ln(1 - p)}{\lambda} .
\end{aligned} \tag{6}$$

Sources:

- original work

Metadata: ID: P50 | shortcut: exp-qf | author: JoramSoch | date: 2020-02-12, 15:48.

3.5.6 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \tag{1}$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$\mathbb{E}(X) = \frac{1}{\lambda} . \tag{2}$$

Proof: The expected value (\rightarrow Definition I/1.7.1) is the probability-weighted average over all possible values:

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

With the probability density function of the exponential distribution (\rightarrow Proof II/3.5.3), this reads:

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^{+\infty} x \cdot \lambda \exp(-\lambda x) \, dx \\
&= \lambda \int_0^{+\infty} x \cdot \exp(-\lambda x) \, dx .
\end{aligned} \tag{4}$$

Using the following anti-derivative

$$\int x \cdot \exp(-\lambda x) \, dx = \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) , \tag{5}$$

the expected value becomes

$$\begin{aligned}
E(X) &= \lambda \left[\left(-\frac{1}{\lambda}x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \right]_0^{+\infty} \\
&= \lambda \left[\lim_{x \rightarrow \infty} \left(-\frac{1}{\lambda}x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) - \left(-\frac{1}{\lambda} \cdot 0 - \frac{1}{\lambda^2} \right) \exp(-\lambda \cdot 0) \right] \\
&= \lambda \left[0 + \frac{1}{\lambda^2} \right] \\
&= \frac{1}{\lambda} .
\end{aligned} \tag{6}$$

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.

Metadata: ID: P47 | shortcut: exp-mean | author: JoramSoch | date: 2020-02-10, 21:57.

3.5.7 Median

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \tag{1}$$

Then, the median (\rightarrow Definition I/1.11.1) of X is

$$\text{median}(X) = \frac{\ln 2}{\lambda} . \tag{2}$$

Proof: The median (\rightarrow Definition I/1.11.1) is the value at which the cumulative distribution function (\rightarrow Definition I/1.6.13) is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} . \tag{3}$$

The cumulative distribution function of the exponential distribution (\rightarrow Proof II/3.5.4) is

$$F_X(x) = 1 - \exp[-\lambda x], \quad x \geq 0 . \tag{4}$$

Thus, the inverse CDF is

$$x = -\frac{\ln(1-p)}{\lambda} \tag{5}$$

and setting $p = 1/2$, we obtain:

$$\text{median}(X) = -\frac{\ln(1-\frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda} . \tag{6}$$

Sources:

- original work

Metadata: ID: P49 | shortcut: exp-med | author: JoramSoch | date: 2020-02-11, 15:03.

3.5.8 Mode

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following an exponential distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

Then, the mode (\rightarrow Definition I/1.11.2) of X is

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

Proof: The mode (\rightarrow Definition I/1.11.2) is the value which maximizes the probability density function (\rightarrow Definition I/1.6.6):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the exponential distribution (\rightarrow Proof II/3.5.3) is:

$$f_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (4)$$

Since

$$\lim_{x \rightarrow 0} f_X(x) = \infty \quad (5)$$

and

$$f_X(x) < \infty \quad \text{for any } x \neq 0 , \quad (6)$$

it follows that

$$\text{mode}(X) = 0 . \quad (7)$$

Sources:

- original work

Metadata: ID: P51 | shortcut: exp-mode | author: JoramSoch | date: 2020-02-12, 15:53.

3.6 Log-normal distribution

3.6.1 Definition

Definition: Let $\ln X$ be a random variable (\rightarrow Definition I/1.2.2) following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 (or, standard deviation σ):

$$Y = \ln(X) \sim \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the exponential function of Y is said to have a log-normal distribution with location parameter μ and scale parameter σ

$$X = \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2) \quad (2)$$

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.

Metadata: ID: D170 | shortcut: lognorm | author: majapavlo | date: 2022-02-07, 22:33.

3.6.2 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a log-normal distribution (\rightarrow Definition II/??):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/??) is defined as the exponential function of a normal random variable (\rightarrow Definition 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 Proof II/3.2.9) is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] . \quad (4)$$

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 Proof I/1.6.8) 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 Proof II/3.2.9), we have

$$\begin{aligned} f_X(x) &= f_Y(g^{-1}(x)) \cdot \frac{dg^{-1}(x)}{dx} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{g^{-1}(x) - \mu}{\sigma} \right)^2 \right] \cdot \frac{dg^{-1}(x)}{dx} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{(\ln x) - \mu}{\sigma} \right)^2 \right] \cdot \frac{d(\ln x)}{dx} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{\ln x - \mu}{\sigma} \right)^2 \right] \cdot \frac{1}{x} \\ &= \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp \left[-\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \end{aligned} \quad (8)$$

which is the probability density function (\rightarrow Definition I/1.6.6) of the log-normal distribution (\rightarrow Definition II/??).

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>.

Metadata: ID: P310 | shortcut: lognorm-pdf | author: majapavlo | date: 2022-02-13, 10:05.

3.6.3 Median

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a log-normal distribution (\rightarrow Definition II/??):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the median (\rightarrow Definition I/1.11.1) of X is

$$\text{median}(X) = e^\mu . \quad (2)$$

Proof: The median (\rightarrow Definition I/1.11.1) 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 Proof “lognorm-cdf”) 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. Thus, the inverse CDF is

$$\begin{aligned} \ln(x) &= \sigma\sqrt{2} \cdot \text{erf}^{-1}(2p - 1) + \mu \\ x &= \exp \left[\sigma\sqrt{2} \cdot \text{erf}^{-1}(2p - 1) + \mu \right] \end{aligned} \quad (5)$$

where $\text{erf}^{-1}(x)$ is the inverse error function. Setting $p = 1/2$, we obtain:

$$\begin{aligned} \ln [\text{median}(X)] &= \sigma\sqrt{2} \cdot \text{erf}^{-1}(0) + \mu \\ \text{median}(X) &= e^\mu . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P306 | shortcut: lognorm-med | author: majapavlo | date: 2022-02-07, 22:33.

3.6.4 Mode

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a log-normal distribution (\rightarrow Definition II/??):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) . \quad (1)$$

Then, the mode (\rightarrow Definition I/1.11.2) of X is

$$\text{mode}(X) = e^{(\mu - \sigma^2)} . \quad (2)$$

Proof: The mode (\rightarrow Definition I/1.11.2) is the value which maximizes the probability density function (\rightarrow Definition I/1.6.6):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

The probability density function of the log-normal distribution (\rightarrow Proof II/??) 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:

$$f'_X(x) = -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \quad (5)$$

$$\begin{aligned} f''_X(x) &= \frac{1}{\sqrt{2\pi}\sigma^2x^3} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \cdot (\ln x - \mu) \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ &\quad + \frac{\sqrt{2}}{\sqrt{\pi}x^3} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ &\quad - \frac{1}{\sqrt{2\pi}\sigma^2x^3} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] . \end{aligned} \quad (6)$$

We now calculate the root of the first derivative (??):

$$\begin{aligned} f'_X(x) = 0 &= -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ -1 &= \frac{\ln x - \mu}{\sigma^2} \\ x &= e^{(\mu - \sigma^2)} . \end{aligned} \quad (7)$$

By plugging this value into the second derivative (??),

$$\begin{aligned} f''_X(e^{(\mu - \sigma^2)}) &= \frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \exp\left[-\frac{\sigma^2}{2}\right] \cdot (\sigma^2) \cdot (0) \\ &\quad + \frac{\sqrt{2}}{\sqrt{\pi}(e^{(\mu - \sigma^2)})^3} \exp\left[-\frac{\sigma^2}{2}\right] \cdot (0) \\ &\quad - \frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \exp\left[-\frac{\sigma^2}{2}\right] \\ &= -\frac{1}{\sqrt{2\pi}\sigma^2(e^{(\mu - \sigma^2)})^3} \exp\left[-\frac{\sigma^2}{2}\right] < 0 , \end{aligned} \quad (8)$$

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.
- 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>.

Metadata: ID: P311 | shortcut: lognorm-mode | author: majapavlo | date: 2022-02-13, 10:15.

3.7 Chi-squared distribution

3.7.1 Definition

Definition: Let X_1, \dots, X_k be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) where each of them is following a standard normal distribution (\rightarrow Definition II/3.2.2):

$$X_i \sim \mathcal{N}(0, 1) \quad \text{for } i = 1, \dots, n. \quad (1)$$

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 Proof II/3.6.3) with k degree of freedom is

$$\chi^2(x; k) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad (3)$$

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.
- 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>.

Metadata: ID: D100 | shortcut: chi2 | author: kjpetrykowski | date: 2020-10-13, 01:20.

3.7.2 Special case of gamma distribution

Theorem: The chi-squared distribution (\rightarrow Definition II/3.6.1) with k degrees of freedom is a special case of the gamma distribution (\rightarrow Definition II/3.4.1) with shape $\frac{k}{2}$ and rate $\frac{1}{2}$:

$$X \sim \text{Gam}\left(\frac{k}{2}, \frac{1}{2}\right) \Rightarrow X \sim \chi^2(k). \quad (1)$$

Proof: The probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) for $x > 0$, where α is the shape parameter and β is the rate parameter, is as follows:

$$\text{Gam}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad (2)$$

If we let $\alpha = k/2$ and $\beta = 1/2$, we obtain

$$\text{Gam}\left(x; \frac{k}{2}, \frac{1}{2}\right) = \frac{x^{k/2-1} e^{-x/2}}{\Gamma(k/2) 2^{k/2}} = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad (3)$$

which is equivalent to the probability density function of the chi-squared distribution (\rightarrow Proof II/3.6.3).

Sources:

- original work

Metadata: ID: P174 | shortcut: chi2-gam | author: kjpetrykowski | date: 2020-10-12, 22:15.

3.7.3 Probability density function

Theorem: Let Y be a random variable (\rightarrow Definition I/1.2.2) following a chi-squared distribution (\rightarrow Definition II/3.6.1):

$$Y \sim \chi^2(k) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) of Y is

$$f_Y(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} e^{-y/2} . \quad (2)$$

Proof: A chi-square-distributed random variable (\rightarrow Definition II/3.6.1) with k degrees of freedom is defined as the sum of k squared standard normal random variables (\rightarrow Definition II/3.2.2):

$$X_1, \dots, X_k \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad Y = \sum_{i=1}^k X_i^2 \sim \chi^2(k) . \quad (3)$$

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 Definition I/1.6.6) and cumulative distribution function (\rightarrow Definition I/1.6.13) of Y . Because the PDF is the first derivative of the CDF (\rightarrow Proof I/1.6.12), we can write:

$$F_Y(y) = \frac{F_Y(y)}{dy} dy = f_Y(y) dy . \quad (5)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Definition I/1.6.6) of the standard normal distribution (\rightarrow Definition II/3.2.2) 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 Proof II/3.2.9), 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

$$\frac{dr}{dy} = \frac{1}{2} y^{-1/2} \quad \Leftrightarrow \quad dr = \frac{dy}{2y^{1/2}} . \tag{10}$$

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):

$$f_Y(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} e^{-y/2} . \tag{12}$$

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.
- 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.

Metadata: ID: P197 | shortcut: chi2-pdf | author: JoramSoch | date: 2020-11-25, 05:56.

3.7.4 Moments

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a chi-squared distribution (\rightarrow Definition II/3.6.1):

$$X \sim \chi^2(k) . \quad (1)$$

If $m > -k/2$, then $E(X^m)$ exists and is equal to:

$$E(X^m) = \frac{2^m \Gamma\left(\frac{k}{2} + m\right)}{\Gamma\left(\frac{k}{2}\right)} . \quad (2)$$

Proof: Combining the definition of the m -th raw moment (\rightarrow Definition I/1.14.3) with the probability density function of the chi-squared distribution (\rightarrow Proof II/3.6.3), we have:

$$E(X^m) = \int_0^\infty \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{k/2}} x^{(k/2)+m-1} e^{-x/2} dx . \quad (3)$$

Now define a new variable $u = x/2$. As a result, we obtain:

$$E(X^m) = \int_0^\infty \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{(k/2)-1}} 2^{(k/2)+m-1} u^{(k/2)+m-1} e^{-u} du . \quad (4)$$

This leads to the desired result when $m > -k/2$. Observe that, if m is a nonnegative integer, then $m > -k/2$ is always true. Therefore, all moments (\rightarrow Definition I/1.14.1) of a chi-squared distribution (\rightarrow Definition II/3.6.1) exist and the m -th raw moment is given by the foregoing equation.

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>.

Metadata: ID: P175 | shortcut: chi2-mom | author: kjpetrykowski | date: 2020-10-13, 01:30.

3.8 F-distribution

3.8.1 Definition

Definition: Let X_1 and X_2 be independent (\rightarrow Definition I/1.3.6) random variables (\rightarrow Definition I/1.2.2) following a chi-squared distribution (\rightarrow Definition II/3.6.1) with d_1 and d_2 degrees of freedom (\rightarrow Definition “dof”), respectively:

$$\begin{aligned} X_1 &\sim \chi^2(d_1) \\ X_2 &\sim \chi^2(d_2) . \end{aligned} \quad (1)$$

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 d_1 and denominator degrees of freedom d_2 :

$$Y = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2) \quad \text{where} \quad d_1, d_2 > 0 . \quad (2)$$

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>.

Metadata: ID: D146 | shortcut: f | author: JoramSoch | date: 2020-04-21, 07:26.

3.8.2 Probability density function

Theorem: Let F be a random variable (\rightarrow Definition I/1.2.2) following an F -distribution (\rightarrow Definition II/3.7.1):

$$F \sim F(u, v) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/3.7.1) is defined as the ratio of two chi-squared random variables (\rightarrow Definition II/3.6.1), divided by their degrees of freedom (\rightarrow Definition “dof”)

$$X \sim \chi^2(u), Y \sim \chi^2(v) \quad \Rightarrow \quad F = \frac{X/u}{Y/v} \sim F(u, v) \quad (3)$$

where X and Y are independent of each other (\rightarrow Definition I/1.3.6).

The probability density function of the chi-squared distribution (\rightarrow Proof II/3.6.3) 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:

$$J = \begin{bmatrix} \frac{dX}{dF} & \frac{dX}{dW} \\ \frac{dY}{dF} & \frac{dY}{dW} \end{bmatrix} = \begin{bmatrix} \frac{u}{v}W & \frac{u}{v}F \\ 0 & 1 \end{bmatrix} \quad (7)$$

$$|J| = \frac{u}{v}W.$$

Because X and Y are independent (\rightarrow Definition I/1.3.6), the joint density (\rightarrow Definition I/1.5.2) of X and Y is equal to the product (\rightarrow Proof I/1.3.8) of the marginal densities (\rightarrow Definition 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 Proof I/1.6.10), the joint density (\rightarrow Definition I/1.5.2) of T 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 Definition I/1.5.3) of F can now be obtained by integrating out (\rightarrow Definition 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 Proof II/3.4.5) 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 Definition I/1.6.6), 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: *YouTube*, retrieved on 2021-10-11; URL: <https://www.youtube.com/watch?v=AmHiOKYmHkI>.

Metadata: ID: P264 | shortcut: f-pdf | author: JoramSoch | date: 2021-10-12, 09:00.

3.9 Beta distribution

3.9.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to follow a beta distribution with shape parameters α and β

$$X \sim \text{Bet}(\alpha, \beta) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) is given by

$$\text{Bet}(x; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (2)$$

where $\alpha > 0$ and $\beta > 0$, 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.

Metadata: ID: D53 | shortcut: beta | author: JoramSoch | date: 2020-05-10, 20:29.

3.9.2 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a beta distribution (\rightarrow Definition II/3.8.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/3.8.1).

Sources:

- original work

Metadata: ID: P94 | shortcut: beta-pdf | author: JoramSoch | date: 2020-05-05, 21:03.

3.9.3 Moment-generating function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a beta distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the moment-generating function (\rightarrow Definition I/1.6.27) 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 Proof II/3.8.2) is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (3)$$

and the moment-generating function (\rightarrow Definition I/1.6.27) is defined as

$$M_X(t) = E[e^{tX}] . \quad (4)$$

Using the expected value for continuous random variables (\rightarrow Definition I/1.7.1), the moment-generating function of X therefore is

$$\begin{aligned} M_X(t) &= \int_0^1 \exp[tx] \cdot \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 e^{tx} x^{\alpha-1} (1-x)^{\beta-1} dx . \end{aligned} \quad (5)$$

With the relationship between beta function and gamma function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (6)$$

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

$${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{a^{\overline{n}}}{b^{\overline{n}}} \frac{z^n}{n!} \quad (9)$$

where $m^{\overline{n}}$ 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.
- 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.

Metadata: ID: P198 | shortcut: beta-mgf | author: JoramSoch | date: 2020-11-25, 06:55.

3.9.4 Cumulative distribution function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a beta distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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 Proof II/3.8.2) is:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} . \quad (3)$$

Thus, the cumulative distribution function (\rightarrow Definition I/1.6.13) 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.
- 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.

Metadata: ID: P195 | shortcut: beta-cdf | author: JoramSoch | date: 2020-11-19, 08:01.

3.9.5 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a beta distribution (\rightarrow Definition II/3.8.1):

$$X \sim \text{Bet}(\alpha, \beta). \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = \frac{\alpha}{\alpha + \beta}. \quad (2)$$

Proof: The expected value (\rightarrow Definition I/1.7.1) 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 Proof II/3.8.2) 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:

$$\begin{aligned} E(X) &= \int_0^1 x \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + 1 + \beta)} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx. \end{aligned} \quad (6)$$

Employing the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$\begin{aligned} E(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \cdot \frac{\alpha \cdot \Gamma(\alpha)}{(\alpha + \beta) \cdot \Gamma(\alpha + \beta)} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx \\ &= \frac{\alpha}{\alpha + \beta} \int_0^1 \frac{\Gamma(\alpha + 1 + \beta)}{\Gamma(\alpha + 1) \cdot \Gamma(\beta)} x^{(\alpha+1)-1} (1-x)^{\beta-1} dx \end{aligned} \quad (7)$$

and again using the density of the beta distribution (\rightarrow Proof II/3.8.2), we get

$$\begin{aligned} E(X) &= \frac{\alpha}{\alpha + \beta} \int_0^1 \text{Bet}(x; \alpha + 1, \beta) dx \\ &= \frac{\alpha}{\alpha + \beta} . \end{aligned} \quad (8)$$

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>.

Metadata: ID: P228 | shortcut: beta-mean | author: JoramSoch | date: 2021-04-29, 09:12.

3.9.6 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.2.2) following a beta distribution (\rightarrow Definition II/3.8.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta + 1) \cdot (\alpha + \beta)^2} . \quad (2)$$

Proof: The variance (\rightarrow Definition I/1.8.1) can be expressed in terms of expected values (\rightarrow Proof I/1.8.3) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

The expected value of a beta random variable (\rightarrow Proof II/3.8.5) is

$$E(X) = \frac{\alpha}{\alpha + \beta} . \quad (4)$$

The probability density function of the beta distribution (\rightarrow Proof II/3.8.2) is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \quad (5)$$

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 Proof II/3.8.2), 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>.

Metadata: ID: P229 | shortcut: beta-var | author: JoramSoch | date: 2021-04-29, 09:31.

3.10 Wald distribution

3.10.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.2.2). Then, X is said to follow a Wald distribution with drift rate γ and threshold α

$$X \sim \text{Wald}(\gamma, \alpha) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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.

Metadata: ID: D95 | shortcut: wald | author: tomfaulkenberry | date: 2020-09-04, 12:00.

3.10.2 Probability density function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a Wald distribution (\rightarrow Definition II/3.9.1):

$$X \sim \text{Wald}(\gamma, \alpha) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/3.9.1).

Sources:

- original work

Metadata: ID: P162 | shortcut: wald-pdf | author: tomfaulkenberry | date: 2020-09-04, 12:00.

3.10.3 Moment-generating function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a Wald distribution (\rightarrow Definition II/3.9.1):

$$X \sim \text{Wald}(\gamma, \alpha) . \quad (1)$$

Then, the moment-generating function (\rightarrow Definition I/1.6.27) 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 Proof II/3.9.2) 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 Definition I/1.6.27) is defined as

$$M_X(t) = E \left[e^{tX} \right] . \quad (4)$$

Using the definition of expected value for continuous random variables (\rightarrow Definition I/1.7.1), 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 kind¹, 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 identities² gives an explicit solution for $K_{-1/2}$:

$$K_{-1/2}(x) = \sqrt{\frac{\pi}{2x}} e^{-x} . \quad (7)$$

The second of these identities³ 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:

$$\begin{aligned} M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \int_0^\infty x^{-3/2} \cdot \exp \left[tx - \frac{\alpha^2}{2x} + \alpha\gamma - \frac{\gamma^2 x}{2} \right] dx \\ &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp \left[\left(t - \frac{\gamma^2}{2} \right) x - \frac{\alpha^2}{2x} \right] dx \\ &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp \left[-\frac{1}{2} (\gamma^2 - 2t) x - \frac{1}{2} \cdot \frac{\alpha^2}{x} \right] dx . \end{aligned} \quad (9)$$

¹<https://dlmf.nist.gov/10.25>

²<https://dlmf.nist.gov/10.39.2>

³<https://dlmf.nist.gov/10.32.10>

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

$$\begin{aligned} M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2 \left(\frac{\gamma^2 - 2t}{\alpha^2} \right)^{1/4} \cdot \sqrt{\frac{\pi}{2\sqrt{\alpha^2(\gamma^2 - 2t)}}} \cdot \exp \left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\ &= \frac{\alpha}{\sqrt{2} \cdot \sqrt{\pi}} \cdot e^{\alpha\gamma} \cdot 2 \cdot \frac{(\gamma^2 - 2t)^{1/4}}{\sqrt{\alpha}} \cdot \frac{\sqrt{\pi}}{\sqrt{2} \cdot \sqrt{\alpha} \cdot (\gamma^2 - 2t)^{1/4}} \cdot \exp \left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\ &= e^{\alpha\gamma} \cdot \exp \left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\ &= \exp \left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right]. \end{aligned} \quad (11)$$

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>.

Metadata: ID: P168 | shortcut: wald-mgf | author: tomfaulkenberry | date: 2020-09-13, 12:00.

3.10.4 Mean

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a Wald distribution (\rightarrow Definition II/3.9.1):

$$X \sim \text{Wald}(\gamma, \alpha). \quad (1)$$

Then, the mean or expected value (\rightarrow Definition I/1.7.1) of X is

$$E(X) = \frac{\alpha}{\gamma}. \quad (2)$$

Proof: The mean or expected value $E(X)$ is the first moment (\rightarrow Definition I/1.14.1) of X , so we can use (\rightarrow Proof I/1.14.2) the moment-generating function of the Wald distribution (\rightarrow Proof II/3.9.3) 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:

$$\begin{aligned} M'_X(0) &= \exp \left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2(0))}} \\ &= \exp \left[\alpha\gamma - \sqrt{\alpha^2 \cdot \gamma^2} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2 \cdot \gamma^2}} \\ &= \exp[0] \cdot \frac{\alpha^2}{\alpha\gamma} \\ &= \frac{\alpha}{\gamma} . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P169 | shortcut: wald-mean | author: tomfaulkenberry | date: 2020-09-13, 12:00.

3.10.5 Variance

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.2.2) following a Wald distribution (\rightarrow Definition II/3.9.1):

$$X \sim \text{Wald}(\gamma, \alpha) . \quad (1)$$

Then, the variance (\rightarrow Definition I/1.8.1) of X is

$$\text{Var}(X) = \frac{\alpha}{\gamma^3} . \quad (2)$$

Proof: To compute the variance of X , we partition the variance into expected values (\rightarrow Proof I/1.8.3):

$$\text{Var}(X) = \text{E}(X^2) - \text{E}(X)^2. \quad (3)$$

We then use the moment-generating function of the Wald distribution (\rightarrow Proof II/3.9.3) to calculate

$$\text{E}(X^2) = M''_X(0) . \quad (4)$$

First we differentiate

$$M_X(t) = \exp \left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \quad (5)$$

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 Proof II/3.9.4) 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).

Sources:

- original work

Metadata: ID: P170 | shortcut: wald-var | author: tomfaulkenberry | date: 2020-09-13, 12:00.

4 Multivariate continuous distributions

4.1 Multivariate normal distribution

4.1.1 Definition

Definition: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3). Then, X is said to be multivariate normally distributed with mean μ and covariance Σ

$$X \sim \mathcal{N}(\mu, \Sigma) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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 μ is an $n \times 1$ real vector and Σ 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.

Metadata: ID: D1 | shortcut: mvn | author: JoramSoch | date: 2020-01-22, 05:20.

4.1.2 Probability density function

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a multivariate normal distribution (\rightarrow Definition II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/4.1.1).

Sources:

- original work

Metadata: ID: P34 | shortcut: mvn-pdf | author: JoramSoch | date: 2020-01-27, 15:23.

4.1.3 Differential entropy

Theorem: Let x follow a multivariate normal distribution (\rightarrow Definition II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the differential entropy (\rightarrow Definition 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 Definition 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 Definition I/1.7.1)

$$h(X) = -E [\ln p(x)] . \quad (4)$$

With the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), the differential entropy of x is:

$$\begin{aligned} h(x) &= -E \left[\ln \left(\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) \right] \\ &= -E \left[-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \\ &= \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} E [(x - \mu)^T \Sigma^{-1} (x - \mu)] . \end{aligned} \quad (5)$$

The last term can be evaluated as

$$\begin{aligned} E [(x - \mu)^T \Sigma^{-1} (x - \mu)] &= E [\text{tr} ((x - \mu)^T \Sigma^{-1} (x - \mu))] \\ &= E [\text{tr} (\Sigma^{-1} (x - \mu) (x - \mu)^T)] \\ &= \text{tr} (\Sigma^{-1} E [(x - \mu) (x - \mu)^T]) \\ &= \text{tr} (\Sigma^{-1} \Sigma) \\ &= \text{tr} (I_n) \\ &= n , \end{aligned} \quad (6)$$

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>

Metadata: ID: P100 | shortcut: mvn-dent | author: JoramSoch | date: 2020-05-14, 19:49.

4.1.4 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3). Assume two multivariate normal distributions (\rightarrow Definition II/4.1.1) P and Q specifying the probability distribution of x as

$$\begin{aligned} P : x &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ Q : x &\sim \mathcal{N}(\mu_2, \Sigma_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is given by

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] . \quad (2)$$

Proof: The KL divergence for a continuous random variable (\rightarrow Definition 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 multivariate normal distributions (\rightarrow Definition II/4.1.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{\mathbb{R}^n} \mathcal{N}(x; \mu_1, \Sigma_1) \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} dx \\ &= \left\langle \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} \right\rangle_{p(x)} . \end{aligned} \quad (4)$$

Using the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \cdot \exp \left[-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \right]}{\frac{1}{\sqrt{(2\pi)^n |\Sigma_2|}} \cdot \exp \left[-\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right]} \right\rangle_{p(x)} \\ &= \left\langle \frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)} . \end{aligned} \quad (5)$$

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:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (x - \mu_2)(x - \mu_2)^T] \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T)] \right\rangle_{p(x)} . \end{aligned} \quad (6)$$

Because trace function and expected value are both linear operators (\rightarrow Proof I/1.7.8), the expectation can be moved inside the trace:

$$\begin{aligned}
\text{KL}[P || Q] &= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} \left[\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)} \right] + \text{tr} \left[\Sigma_2^{-1} \langle xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T \rangle_{p(x)} \right] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} \left[\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)} \right] + \text{tr} \left[\Sigma_2^{-1} \left(\langle xx^T \rangle_{p(x)} - \langle 2\mu_2 x^T \rangle_{p(x)} + \langle \mu_2 \mu_2^T \rangle_{p(x)} \right) \right] \right)
\end{aligned} \tag{7}$$

Using the expectation of a linear form for the multivariate normal distribution (\rightarrow Proof II/4.1.5)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle Ax \rangle = A\mu \tag{8}$$

and the expectation of a quadratic form for the multivariate normal distribution (\rightarrow Proof I/1.7.9)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle x^T Ax \rangle = \mu^T A \mu + \text{tr}(A \Sigma), \tag{9}$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{aligned}
\text{KL}[P || Q] &= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\Sigma_1 + \mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [I_n] + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\mu_1^T \Sigma_2^{-1} \mu_1 - 2\mu_1^T \Sigma_2^{-1} \mu_2 + \mu_2^T \Sigma_2^{-1} \mu_2] \right) \\
&= \frac{1}{2} \left[\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right].
\end{aligned} \tag{10}$$

Finally, rearranging the terms, we get:

$$\text{KL}[P || Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right]. \tag{11}$$

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.

Metadata: ID: P92 | shortcut: mvn-kl | author: JoramSoch | date: 2020-05-05, 06:57.

4.1.5 Linear transformation

Theorem: Let x follow a multivariate normal distribution (\rightarrow Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma). \tag{1}$$

Then, any linear transformation of x is also multivariate normally distributed:

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T). \tag{2}$$

Proof: The moment-generating function of a random vector (\rightarrow Definition I/1.6.27) 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

$$\begin{aligned} M_y(t) &\stackrel{(2)}{=} \mathbb{E}(\exp[t^T(Ax + b)]) \\ &= \mathbb{E}(\exp[t^T Ax] \cdot \exp[t^T b]) \\ &= \exp[t^T b] \cdot \mathbb{E}(\exp[t^T Ax]) \\ &\stackrel{(3)}{=} \exp[t^T b] \cdot M_x(At) . \end{aligned} \quad (4)$$

The moment-generating function of the multivariate normal distribution (\rightarrow Proof “mvn-mgf”) 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

$$\begin{aligned} M_y(t) &\stackrel{(4)}{=} \exp[t^T b] \cdot M_x(At) \\ &\stackrel{(5)}{=} \exp[t^T b] \cdot \exp\left[t^T A\mu + \frac{1}{2} t^T A\Sigma A^T t\right] \\ &= \exp\left[t^T (A\mu + b) + \frac{1}{2} t^T A\Sigma A^T t\right] . \end{aligned} \quad (6)$$

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>.

Metadata: ID: P1 | shortcut: mvn-ltt | author: JoramSoch | date: 2019-08-27, 12:14.

4.1.6 Marginal distributions

Theorem: Let x follow a multivariate normal distribution (\rightarrow Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the marginal distribution (\rightarrow Definition 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 μ_s drops the irrelevant variables (the ones not in the subset, i.e. marginalized out) from the mean vector μ and Σ_s drops the corresponding rows and columns from the covariance matrix Σ .

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 Proof II/4.1.5) to give

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^T). \quad (4)$$

Finally, we see that $S\mu = \mu_s$ and $S\Sigma S^T = \Sigma_s$.

Sources:

- original work

Metadata: ID: P35 | shortcut: mvn-marg | author: JoramSoch | date: 2020-01-29, 15:12.

4.1.7 Conditional distributions

Theorem: Let x follow a multivariate normal distribution (\rightarrow Definition II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma). \quad (1)$$

Then, the conditional distribution (\rightarrow Definition 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 Definition I/1.7.1) and covariance (\rightarrow Definition I/1.9.1) 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 Definition 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 Definition I/1.5.3) of x_2 follows from (\rightarrow Proof II/4.1.6) (1) and (4) as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) . \quad (7)$$

According to the law of conditional probability (\rightarrow Definition 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 Proof II/4.1.2), 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 Σ 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

$$\begin{aligned}
p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
&\exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T \Sigma^{11} (x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma^{12} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma^{22} (x_2 - \mu_2) \right) \right. \\
&\quad \left. + \frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right]
\end{aligned} \tag{13}$$

where we have used the fact that $\Sigma^{21T} = \Sigma^{12}$, because Σ^{-1} is a symmetric matrix.

The inverse of a block matrix is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}, \tag{14}$$

thus the inverse of Σ in (11) is

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}. \tag{15}$$

Plugging this into (13), we have:

$$\begin{aligned}
p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
&\exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\
&\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\
&\quad (x_2 - \mu_2)^T [\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}] (x_2 - \mu_2) \\
&\quad \left. \left. + \frac{1}{2} ((x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right) \right].
\end{aligned} \tag{16}$$

Eliminating some terms, we have:

$$\begin{aligned}
p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
&\exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\
&\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\
&\quad \left. \left. (x_2 - \mu_2)^T \Sigma_{22}^{-1} \Sigma_{21} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) \right) \right].
\end{aligned} \tag{17}$$

Rearranging the terms, we have

$$\begin{aligned}
p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[-\frac{1}{2} \cdot \right. \\
&\quad \left. [(x_1 - \mu_1) - \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)]^T (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} [(x_1 - \mu_1) - \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)] \right] \\
&= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[-\frac{1}{2} \cdot \right. \\
&\quad \left. [x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2))]^T (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} [x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2))] \right]
\end{aligned} \tag{18}$$

where we have used the fact that $\Sigma_{21}^T = \Sigma_{12}$, because Σ is a covariance matrix.

The determinant of a block matrix is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C|, \tag{19}$$

such that we have for Σ that

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}|. \tag{20}$$

With this and $n - n_2 = n_1$, we finally arrive at

$$\begin{aligned}
p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n_1} |\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}|}} \cdot \exp \left[-\frac{1}{2} \cdot \right. \\
&\quad \left. [x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2))]^T (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} [x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2))] \right]
\end{aligned} \tag{21}$$

which is the probability density function of a multivariate normal distribution (\rightarrow Proof II/4.1.2)

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \tag{22}$$

with the mean $\mu_{1|2}$ and variance $\Sigma_{1|2}$ 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.

Metadata: ID: P88 | shortcut: mvn-cond | author: JoramSoch | date: 2020-03-20, 08:44.

4.1.8 Conditions for independence

Theorem: Let x be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) following a multivariate normal distribution (\rightarrow Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) . \quad (1)$$

Then, the components of x are statistically independent (\rightarrow Definition I/1.3.6), if and only if the covariance matrix (\rightarrow Definition I/1.9.7) 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 Proof II/4.1.6) where mean (\rightarrow Definition I/1.7.1) and variance (\rightarrow Definition I/1.8.1) 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 Proof II/4.1.2) 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 Proof II/3.2.9) 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, we have

$$\begin{aligned} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] &\stackrel{(4),(5)}{=} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp \left[-\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right] \\ \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] &= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \mu_i) \frac{1}{\sigma_i^2} (x_i - \mu_i) \right] \\ -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) &= -\frac{1}{2} \sum_{i=1}^n \log \sigma_i^2 - \frac{1}{2} \sum_{i=1}^n (x_i - \mu_i) \frac{1}{\sigma_i^2} (x_i - \mu_i) \end{aligned} \quad (7)$$

which, given the laws for matrix determinants and matrix inverses, is only fulfilled if

$$\Sigma = \text{diag}([\sigma_1^2, \dots, \sigma_n^2]) . \quad (8)$$

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

$$p(x) = p(x_1) \cdot \dots \cdot p(x_n) . \quad (11)$$

Sources:

- original work

Metadata: ID: P236 | shortcut: mvn-ind | author: JoramSoch | date: 2021-06-02, 09:22.

4.2 Multivariate t-distribution

4.2.1 Definition

Definition: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3). Then, X is said to follow a multivariate t -distribution with mean μ , scale matrix Σ and degrees of freedom ν

$$X \sim t(\mu, \Sigma, \nu) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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 μ is an $n \times 1$ real vector, Σ 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.

Metadata: ID: D148 | shortcut: mvt | author: JoramSoch | date: 2020-04-21, 08:16.

4.2.2 Relationship to F-distribution

Theorem: Let X be a $n \times 1$ random vector (\rightarrow Definition I/1.2.3) following a multivariate t-distribution (\rightarrow Definition II/4.2.1) with mean μ , scale matrix Σ and degrees of freedom ν :

$$X \sim t(\mu, \Sigma, \nu) . \quad (1)$$

Then, the centered, weighted and standardized quadratic form of X follows an F-distribution (\rightarrow Definition II/3.7.1) with degrees of freedom n and ν :

$$(X - \mu)^T \Sigma^{-1} (X - \mu) / n \sim F(n, \nu) . \quad (2)$$

Proof: The linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”) 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 $n \times 1$ random vector (\rightarrow Definition I/1.2.3) following a multivariate t-distribution (\rightarrow Definition II/4.2.1), 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 Σ . 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. Y is an $n \times 1$ vector of independent and identically distributed (\rightarrow Definition “iid”) random variables (\rightarrow Definition I/1.2.2) following a univariate t-distribution (\rightarrow Definition II/3.3.1) with ν 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 Definition II/3.3.1) an expression of X in terms of a standard normal (\rightarrow Definition II/3.2.2) random variable Z and a chi-squared (\rightarrow Definition II/3.6.1) 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:

$$\begin{aligned}
Z &\stackrel{(4)}{=} Y^T Y / n \\
&= \frac{1}{n} \sum_{i=1}^n Y_i^2 \\
&\stackrel{(7)}{=} \frac{1}{n} \sum_{i=1}^n \left(\frac{Z_i}{\sqrt{V/\nu}} \right)^2 \\
&= \frac{(\sum_{i=1}^n Z_i^2) / n}{V/\nu} .
\end{aligned} \tag{8}$$

Because by definition, the sum of squared standard normal random variables follows a chi-squared distribution (\rightarrow Definition II/3.6.1)

$$X_i \sim \mathcal{N}(0, 1), \ i = 1, \dots, n \quad \Rightarrow \quad \sum_{i=1}^n X_i^2 \sim \chi^2(n) , \tag{9}$$

the quantity Z becomes a ratio of the following form

$$Z = \frac{W/n}{V/\nu} \quad \text{with} \quad W \sim \chi^2(n) \quad \text{and} \quad V \sim \chi^2(\nu) , \tag{10}$$

such that Z , by definition, follows an F-distribution (\rightarrow Definition II/3.7.1):

$$Z = \frac{W/n}{V/\nu} \sim F(n, \nu) . \tag{11}$$

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.

Metadata: ID: P231 | shortcut: mvt-f | author: JoramSoch | date: 2021-05-04, 10:29.

4.3 Normal-gamma distribution

4.3.1 Definition

Definition: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) and let Y be a positive random variable (\rightarrow Definition I/1.2.2). Then, X and Y are said to follow a normal-gamma distribution

$$X, Y \sim \text{NG}(\mu, \Lambda, a, b) , \tag{1}$$

if and only if their joint probability (\rightarrow Definition I/1.3.2) density function (\rightarrow Definition I/1.6.6) is given by

$$f_{X,Y}(x, y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) \tag{2}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2) with mean μ and covariance Σ and $\text{Gam}(x; a, b)$ is the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) with shape a and rate b . The $n \times n$ matrix Λ is referred to as the precision matrix (\rightarrow Definition I/1.9.11) 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.

Metadata: ID: D5 | shortcut: ng | author: JoramSoch | date: 2020-01-27, 14:28.

4.3.2 Probability density function

Theorem: Let x and y follow a normal-gamma distribution (\rightarrow Definition II/4.3.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then, the joint probability (\rightarrow Definition I/1.3.2) density function (\rightarrow Definition I/1.6.6) of x and y is

$$p(x, y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp \left[-\frac{y}{2} ((x - \mu)^T \Lambda (x - \mu) + 2b) \right] . \quad (2)$$

Proof: The probability density of the normal-gamma distribution is defined as (\rightarrow Definition II/4.3.1) as the product of a multivariate normal distribution (\rightarrow Definition II/4.1.1) over x conditional on y and a univariate gamma distribution (\rightarrow Definition II/3.4.1) over y :

$$p(x, y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) \quad (3)$$

With the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2) and the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), this becomes:

$$p(x, y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[-\frac{1}{2} (x - \mu)^T (y\Lambda) (x - \mu) \right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp [-by] . \quad (4)$$

Using the relation $|yA| = y^n |A|$ for an $n \times n$ matrix A and rearranging the terms, we have:

$$p(x, y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp \left[-\frac{y}{2} ((x - \mu)^T \Lambda (x - \mu) + 2b) \right] . \quad (5)$$

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.

Metadata: ID: P44 | shortcut: ng-pdf | author: JoramSoch | date: 2020-02-07, 20:44.

4.3.3 Mean

Theorem: Let $x \in \mathbb{R}^n$ and $y > 0$ follow a normal-gamma distribution (\rightarrow Definition II/4.3.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then, the expected value (\rightarrow Definition I/1.7.1) of x and y is

$$\mathbb{E}[(x, y)] = \left[\left(\mu, \frac{a}{b} \right) \right] . \quad (2)$$

Proof: Consider the random vector (\rightarrow Definition I/1.2.3)

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ y \end{bmatrix} . \quad (3)$$

According to the expected value of a random vector (\rightarrow Definition I/1.7.12), 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 Definition II/4.3.1) by definition x follows a multivariate normal distribution (\rightarrow Definition II/4.1.1) conditional on y and y follows a univariate gamma distribution (\rightarrow Definition II/3.4.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) \Leftrightarrow x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \quad \wedge \quad y \sim \text{Gam}(a, b) . \quad (5)$$

Thus, with the expected value of the multivariate normal distribution (\rightarrow Proof “mvn-mean”) and the law of conditional probability (\rightarrow Definition 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 Proof II/3.4.8), $E(y)$ becomes

$$\begin{aligned} E(y) &= \int y \cdot p(y) \, dy \\ &= \langle y \rangle_{\text{Gam}(a,b)} \\ &= \frac{a}{b} . \end{aligned} \tag{7}$$

Thus, the expectation of the random vector in equations (3) and (4) is

$$E \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} \mu \\ a/b \end{bmatrix} , \tag{8}$$

as indicated by equation (2).

Sources:

- original work

Metadata: ID: P237 | shortcut: ng-mean | author: JoramSoch | date: 2021-07-08, 09:40.

4.3.4 Differential entropy

Theorem: Let x be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) and let y be a positive random variable (\rightarrow Definition I/1.2.2). Assume that x and y are jointly normal-gamma distributed:

$$(x, y) \sim \text{NG}(\mu, \Lambda^{-1}, a, b) \tag{1}$$

Then, the differential entropy (\rightarrow Definition I/2.2.1) of x in nats is

$$\begin{aligned} h(x, y) &= \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n \\ &\quad + a + \ln \Gamma(a) - \frac{n-2+2a}{2} \psi(a) + \frac{n-2}{2} \ln b . \end{aligned} \tag{2}$$

Proof: The probability density function of the normal-gamma distribution (\rightarrow Proof II/4.3.2) is

$$p(x, y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) . \tag{3}$$

The differential entropy of the multivariate normal distribution (\rightarrow Proof II/4.1.3) is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2}n \tag{4}$$

and the differential entropy of the univariate gamma distribution (\rightarrow Proof II/3.4.12) is

$$h(y) = a + \ln \Gamma(a) + (1-a) \cdot \psi(a) - \ln b \tag{5}$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The differential entropy of a continuous random variable (\rightarrow Definition 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 Definition II/4.3.1) 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 Definition I/1.3.4), this can be evaluated as follows:

$$\begin{aligned} h(x, y) &= - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(x|y) p(y) dx dy \\ &= - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(x|y) dx dy - \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln p(y) dx dy \\ &= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y) \ln p(x|y) dx dy + \int_0^\infty p(y) \ln p(y) \int_{\mathbb{R}^n} p(x|y) dx dy \\ &= \langle h(x|y) \rangle_{p(y)} + h(y) . \end{aligned} \quad (8)$$

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

$$\begin{aligned} \langle h(x|y) \rangle_{p(y)} &= \left\langle \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |(y\Lambda)^{-1}| + \frac{1}{2}n \right\rangle_{p(y)} \\ &= \left\langle \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |(y\Lambda)| + \frac{1}{2}n \right\rangle_{p(y)} \\ &= \left\langle \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(y^n |\Lambda|) + \frac{1}{2}n \right\rangle_{p(y)} \\ &= \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n - \left\langle \frac{n}{2} \ln y \right\rangle_{p(y)} \end{aligned} \quad (9)$$

and using the relation (\rightarrow Proof II/3.4.10) $y \sim \text{Gam}(a, b) \Rightarrow \langle \ln y \rangle = \psi(a) - \ln(b)$, we have

$$\langle h(x|y) \rangle_{p(y)} = \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Lambda| + \frac{1}{2}n - \frac{n}{2} \psi(a) + \frac{n}{2} \ln b . \quad (10)$$

By plugging (10) and (5) into (8), one arrives at the differential entropy given by (2).

Sources:

- original work

Metadata: ID: P238 | shortcut: ng-dent | author: JoramSoch | date: 2021-07-08, 10:51.

4.3.5 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (\rightarrow Definition I/1.2.3) and let y be a positive random variable (\rightarrow Definition I/1.2.2). Assume two normal-gamma distributions (\rightarrow Definition II/4.3.1) P and Q specifying the joint distribution of x and y as

$$\begin{aligned} P : (x, y) &\sim \text{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\ Q : (x, y) &\sim \text{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is given by

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \frac{a_1}{b_1} [(\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1)] + \frac{1}{2} \text{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \\ &\quad + a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} . \end{aligned} \quad (2)$$

Proof: The probability density function of the normal-gamma distribution (\rightarrow Proof II/4.3.2) 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 Proof II/4.1.4) is

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (4)$$

and the Kullback-Leibler divergence of the univariate gamma distribution (\rightarrow Proof II/3.4.13) is

$$\text{KL}[P \parallel Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} \quad (5)$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The KL divergence for a continuous random variable (\rightarrow Definition I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)} dz \quad (6)$$

which, applied to the normal-gamma distribution (\rightarrow Definition II/4.3.1) over x and y , yields

$$\text{KL}[P \parallel Q] = \int_0^\infty \int_{\mathbb{R}^n} p(x, y) \ln \frac{p(x, y)}{q(x, y)} dx dy . \quad (7)$$

Using the law of conditional probability (\rightarrow Definition I/1.3.4), this can be evaluated as follows:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(x|y) p(y)}{q(x|y) q(y)} dx dy \\ &= \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty \int_{\mathbb{R}^n} p(x|y) p(y) \ln \frac{p(y)}{q(y)} dx dy \\ &= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty p(y) \ln \frac{p(y)}{q(y)} \int_{\mathbb{R}^n} p(x|y) dx dy \\ &= \langle \text{KL}[p(x|y) \parallel q(x|y)] \rangle_{p(y)} + \text{KL}[p(y) \parallel q(y)] . \end{aligned} \quad (8)$$

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 Proof II/3.4.8) $y \sim \text{Gam}(a, b) \Rightarrow \langle y \rangle = a/b$, 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>.

Metadata: ID: P6 | shortcut: ng-kl | author: JoramSoch | date: 2019-12-06, 09:35.

4.3.6 Marginal distributions

Theorem: Let x and y follow a normal-gamma distribution (\rightarrow Definition II/4.3.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b). \quad (1)$$

Then, the marginal distribution (\rightarrow Definition I/1.5.3) of y is a gamma distribution (\rightarrow Definition II/3.4.1)

$$y \sim \text{Gam}(a, b) \quad (2)$$

and the marginal distribution (\rightarrow Definition I/1.5.3) of x is a multivariate t-distribution (\rightarrow Definition II/4.2.1)

$$x \sim t\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right). \quad (3)$$

Proof: The probability density function of the normal-gamma distribution (\rightarrow Proof II/4.3.2) 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 Definition I/1.3.3), the marginal distribution of y can be derived as

$$\begin{aligned}
 p(y) &= \int p(x, y) \, dx \\
 &= \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \text{Gam}(y; a, b) \, dx \\
 &= \text{Gam}(y; a, b) \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \, dx \\
 &= \text{Gam}(y; a, b)
 \end{aligned} \tag{5}$$

which is the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) with shape parameter a and rate parameter b .

Using the law of marginal probability (\rightarrow Definition 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 Proof “mvt-pdf”) with mean vector μ , shape matrix $\left(\frac{a}{b}\Lambda\right)^{-1}$ and $2a$ degrees of freedom.

Sources:

- original work

Metadata: ID: P36 | shortcut: ng-marg | author: JoramSoch | date: 2020-01-29, 21:42.

4.3.7 Conditional distributions

Theorem: Let x and y follow a normal-gamma distribution (\rightarrow Definition II/4.3.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then,

1) the conditional distribution (\rightarrow Definition I/1.5.4) of x given y is a multivariate normal distribution (\rightarrow Definition II/4.1.1)

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) ; \quad (2)$$

2) the conditional distribution (\rightarrow Definition 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 Definition 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 Definition I/1.7.1) and covariance (\rightarrow Definition I/1.9.1)

$$\begin{aligned} \mu_{1|2}(y) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2}(y) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} \end{aligned} \quad (4)$$

where μ_1, μ_2 and $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}, \Sigma_{21}$ are block-wise components (\rightarrow Proof II/4.1.7) of μ and $\Sigma(y) = (y\Lambda)^{-1}$;

3) the conditional distribution (\rightarrow Definition I/1.5.4) of y given x is a gamma distribution (\rightarrow Definition 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 Definition II/4.3.1):

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) . \end{aligned} \quad (6)$$

2) This follows from (2) and the conditional distributions of the multivariate normal distribution (\rightarrow Proof II/4.1.7):

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \Sigma) \\ \Rightarrow x_1|x_2 &\sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} . \end{aligned} \quad (7)$$

3) The conditional density of y given x follows from Bayes' theorem (\rightarrow Proof I/5.3.1) as

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}. \quad (8)$$

The conditional distribution (\rightarrow Definition I/1.5.4) of x given y is a multivariate normal distribution (\rightarrow Proof II/4.3.2)

$$p(x|y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[-\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right], \quad (9)$$

the marginal distribution (\rightarrow Definition I/1.5.3) of y is a gamma distribution (\rightarrow Proof II/4.3.6)

$$p(y) = \text{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] \quad (10)$$

and the marginal distribution (\rightarrow Definition I/1.5.3) of x is a multivariate t-distribution (\rightarrow Proof II/4.3.6)

$$\begin{aligned} p(x) &= t \left(x; \mu, \left(\frac{a}{b} \Lambda \right)^{-1}, 2a \right) \\ &= \sqrt{\frac{\left| \frac{a}{b} \Lambda \right|}{(2a\pi)^n}} \cdot \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T \left(\frac{a}{b} \Lambda \right) (x - \mu) \right)^{-\frac{2a+n}{2}} \\ &= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{-(a + \frac{n}{2})}. \end{aligned} \quad (11)$$

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 Proof II/3.4.5) 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

$$p(y|x) = \text{Gam} \left(y; a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right). \quad (14)$$

Sources:

- original work

Metadata: ID: P146 | shortcut: ng-cond | author: JoramSoch | date: 2020-08-05, 06:54.

4.4 Dirichlet distribution

4.4.1 Definition

Definition: Let X be a $k \times 1$ random vector (\rightarrow Definition 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 Definition I/1.6.6) 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.

Metadata: ID: D54 | shortcut: dir | author: JoramSoch | date: 2020-05-10, 20:36.

4.4.2 Probability density function

Theorem: Let X be a random vector (\rightarrow Definition I/1.2.3) following a Dirichlet distribution (\rightarrow Definition II/4.4.1):

$$X \sim \text{Dir}(\alpha) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/4.4.1).

Sources:

- original work

Metadata: ID: P95 | shortcut: dir-pdf | author: JoramSoch | date: 2020-05-05, 21:22.

4.4.3 Kullback-Leibler divergence

Theorem: Let x be an $k \times 1$ random vector (\rightarrow Definition I/1.2.3). Assume two Dirichlet distributions (\rightarrow Definition II/4.4.1) 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 Definition 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 Definition 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 Definition 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 Proof II/4.4.2), 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 (\rightarrow Proof “dir-logmean”)

$$x \sim \text{Dir}(\alpha) \Rightarrow \langle \ln x_i \rangle = \psi(\alpha_i) - \psi\left(\sum_{i=1}^k \alpha_i\right), \quad (6)$$

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>.

Metadata: ID: P294 | shortcut: dir-kl | author: JoramSoch | date: 2021-12-02, 14:28.

4.4.4 Exceedance probabilities

Theorem: Let $r = [r_1, \dots, r_k]$ be a random vector (\rightarrow Definition I/1.2.3) following a Dirichlet distribution (\rightarrow Definition II/4.4.1) 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 Definition I/1.3.10) 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 Definition I/1.3.10) 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 Definition II/4.4.1), the exceedance probability (\rightarrow Definition I/1.3.10) 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 Proof II/4.4.2) 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

$$r_i \in [0, 1] \quad \text{for} \quad i = 1, \dots, k \quad \text{and} \quad \sum_{i=1}^k r_i = 1 , \quad (6)$$

and defined to be zero otherwise (\rightarrow Definition II/4.4.1).

1) If $k = 2$, the probability density function of the Dirichlet distribution (\rightarrow Proof II/4.4.2) 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 Proof II/3.8.2)

$$p(r_1) = \frac{r_1^{\alpha_1-1} (1 - r_1)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)} \quad (8)$$

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 Proof II/3.8.4), it evaluates to

$$\varphi_1 = 1 - \int_0^{\frac{1}{2}} p(r_1) dr_1 = 1 - \frac{B\left(\frac{1}{2}; \alpha_1, \alpha_2\right)}{B(\alpha_1, \alpha_2)} \quad (11)$$

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

$$\varphi_i = p(r_i = \max(r)) > p(r_i > 1/2) \quad \text{for} \quad i = 1, \dots, k , \quad (13)$$

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 (\rightarrow Proof “gam-dir”) which states that

$$\begin{aligned}
Y_1 &\sim \text{Gam}(\alpha_1, \beta), \dots, Y_k \sim \text{Gam}(\alpha_k, \beta), Y_s = \sum_{i=1}^k Y_i \\
\Rightarrow X &= (X_1, \dots, X_k) = \left(\frac{Y_1}{Y_s}, \dots, \frac{Y_k}{Y_s} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_k).
\end{aligned} \tag{14}$$

The probability density function of the gamma distribution (\rightarrow Proof II/3.4.5) is given by

$$\text{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \quad \text{for } x > 0. \tag{15}$$

Consider the gamma random variables (\rightarrow Definition II/3.4.1)

$$q_1 \sim \text{Gam}(\alpha_1, 1), \dots, q_k \sim \text{Gam}(\alpha_k, 1), q_s = \sum_{j=1}^k q_j \tag{16}$$

and the Dirichlet random vector (\rightarrow Definition II/4.4.1)

$$r = (r_1, \dots, r_k) = \left(\frac{q_1}{q_s}, \dots, \frac{q_k}{q_s} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_k). \tag{17}$$

Obviously, it holds that

$$r_i > r_j \Leftrightarrow q_i > q_j \quad \text{for } i, j = 1, \dots, k \quad \text{with } j \neq i. \tag{18}$$

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

$$p(q_i > q_j | q_i) = p(q_j < q_i | q_i) \tag{19}$$

which can be expressed in terms of the cumulative distribution function of the gamma distribution (\rightarrow Proof II/3.4.6) as

$$p(q_j < q_i | q_i) = \int_0^{q_i} \text{Gam}(q_j; \alpha_j, 1) dq_j = \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \tag{20}$$

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:

$$p(\forall_{j \neq i} [q_i > q_j] | q_i) = \prod_{j \neq i} p(q_i > q_j | q_i) = \prod_{j \neq i} \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)}. \tag{21}$$

In order to obtain the exceedance probability φ_i , the dependency on q_i in this probability still has to be removed. From equations (4) and (18), it follows that

$$\varphi_i = p(\forall_{j \neq i} [r_i > r_j]) = p(\forall_{j \neq i} [q_i > q_j]). \tag{22}$$

Using the law of marginal probability (\rightarrow Definition I/1.3.3), we have

$$\varphi_i = \int_0^\infty p(\forall_{j \neq i} [q_i > q_j] | q_i) p(q_i) dq_i. \tag{23}$$

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 Definition I/1.3.10) for one element from a Dirichlet-distributed (\rightarrow Definition II/4.4.1) random vector (\rightarrow Definition 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 Definition II/3.4.1) cumulative distribution functions (\rightarrow Definition I/1.6.13) and the second term is a gamma (\rightarrow Definition II/3.4.1) probability density function (\rightarrow Definition I/1.6.6).

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>.

Metadata: ID: P181 | shortcut: dir-ep | author: JoramSoch | date: 2020-10-22, 08:04.

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 Definition I/1.2.4). Then, X is said to be matrix-normally distributed with mean M , covariance (\rightarrow Definition I/1.9.7) across rows U and covariance (\rightarrow Definition I/1.9.7) across columns V

$$X \sim \mathcal{MN}(M, U, V) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.6.6) 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.

Metadata: ID: D6 | shortcut: matn | author: JoramSoch | date: 2020-01-27, 14:37.

5.1.2 Probability density function

Theorem: Let X be a random matrix (\rightarrow Definition I/1.2.4) following a matrix-normal distribution (\rightarrow Definition II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.6.6) 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 Definition II/5.1.1).

Sources:

- original work

Metadata: ID: P70 | shortcut: matn-pdf | author: JoramSoch | date: 2020-03-02, 21:03.

5.1.3 Equivalence to multivariate normal distribution

Theorem: The matrix X is matrix-normally distributed (\rightarrow Definition 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 Definition 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 Proof II/5.1.2) 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 Proof II/4.1.2) 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 Definition I/1.6.6) are identical, it is proven that the associated probability distributions (\rightarrow Definition 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.

Metadata: ID: P26 | shortcut: matn-mvn | author: JoramSoch | date: 2020-01-20, 21:09.

5.1.4 Kullback-Leibler divergence

Theorem: Let X be an $n \times p$ random matrix (\rightarrow Definition I/1.2.4). Assume two matrix-normal distributions (\rightarrow Definition 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 Definition I/2.5.1) of P from Q is given by

$$\begin{aligned} \text{KL}[P || Q] &= \frac{1}{2} \left[\text{vec}(M_2 - M_1)^T \text{vec} \left(U_2^{-1} (M_2 - M_1) V_2^{-1} \right) \right. \\ &\quad \left. + \text{tr} \left((V_2^{-1} V_1) \otimes (U_2^{-1} U_1) \right) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] . \end{aligned} \quad (2)$$

Proof: The matrix-normal distribution is equivalent to the multivariate normal distribution (\rightarrow Proof II/5.1.3),

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) , \quad (3)$$

and the Kullback-Leibler divergence for the multivariate normal distribution (\rightarrow Proof II/4.1.4) is

$$\text{KL}[P || Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (4)$$

where X is an $n \times 1$ random vector (\rightarrow Definition 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)

$$\begin{aligned} \text{KL}[P || Q] &= \frac{1}{2} \left[(\text{vec}(M_2) - \text{vec}(M_1))^T (V_2 \otimes U_2)^{-1} (\text{vec}(M_2) - \text{vec}(M_1)) \right. \\ &\quad \left. + \text{tr} \left((V_2 \otimes U_2)^{-1} (V_1 \otimes U_1) \right) - \ln \frac{|V_1 \otimes U_1|}{|V_2 \otimes U_2|} - np \right] . \end{aligned} \quad (5)$$

Using the vectorization operator and Kronecker product properties

$$\text{vec}(A) + \text{vec}(B) = \text{vec}(A + B) \quad (6)$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \quad (7)$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \quad (8)$$

$$|A \otimes B| = |A|^m |B|^n \quad \text{where } A \in \mathbb{R}^{n \times n} \quad \text{and } B \in \mathbb{R}^{m \times m}, \quad (9)$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] = & \frac{1}{2} \left[\text{vec}(M_2 - M_1)^T (V_2^{-1} \otimes U_2^{-1}) \text{vec}(M_2 - M_1) \right. \\ & \left. + \text{tr}((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right]. \end{aligned} \quad (10)$$

Using the relationship between Kronecker product and vectorization operator

$$(C^T \otimes A) \text{vec}(B) = \text{vec}(ABC), \quad (11)$$

we finally have:

$$\begin{aligned} \text{KL}[P \parallel Q] = & \frac{1}{2} \left[\text{vec}(M_2 - M_1)^T \text{vec}(U_2^{-1}(M_2 - M_1)V_2^{-1}) \right. \\ & \left. + \text{tr}((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right]. \end{aligned} \quad (12)$$

Sources:

- original work

Metadata: ID: P296 | shortcut: matn-kl | author: JoramSoch | date: 2021-12-02, 20:22.

5.1.5 Linear transformation

Theorem: Let X be an $n \times p$ random matrix (\rightarrow Definition I/1.2.4) following a matrix-normal distribution (\rightarrow Definition 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 Proof II/5.1.3),

$$X \sim \mathcal{MN}(M, U, V) \Leftrightarrow \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U), \quad (3)$$

and the linear transformation theorem for the multivariate normal distribution (\rightarrow Proof II/4.1.5) 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

$$\begin{aligned} \text{vec}(Y) &= \text{vec}(AXB + C) \\ &= \text{vec}(AXB) + \text{vec}(C) \\ &= (B^T \otimes A)\text{vec}(X) + \text{vec}(C). \end{aligned} \quad (5)$$

Using (3) and (4), we have

$$\begin{aligned} \text{vec}(Y) &\sim \mathcal{N}((B^T \otimes A)\text{vec}(M) + \text{vec}(C), (B^T \otimes A)(V \otimes U)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB) + \text{vec}(C), (B^T V \otimes AU)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB + C), B^T V B \otimes AU A^T). \end{aligned} \quad (6)$$

Using (3), we finally have:

$$Y \sim \mathcal{MN}(AMB + C, AU A^T, B^T V B). \quad (7)$$

Sources:

- original work

Metadata: ID: P145 | shortcut: matn-ltt | author: JoramSoch | date: 2020-08-03, 22:24.

5.1.6 Transposition

Theorem: Let X be a random matrix (\rightarrow Definition I/1.2.4) following a matrix-normal distribution (\rightarrow Definition 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: The probability density function of the matrix-normal distribution (\rightarrow Proof II/5.1.2) is:

$$f(X) = \mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (V^{-1}(X - M)^T U^{-1}(X - M)) \right]. \quad (3)$$

Define $Y = X^T$. Then, $X = Y^T$ and we can substitute:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (V^{-1}(Y^T - M)^T U^{-1}(Y^T - M)) \right] . \quad (4)$$

Using $(A + B)^T = (A^T + B^T)$, we have:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (V^{-1}(Y - M^T) U^{-1}(Y - M^T)^T) \right] . \quad (5)$$

Using $\text{tr}(ABC) = \text{tr}(CAB)$, we obtain

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (U^{-1}(Y - M^T)^T V^{-1}(Y - M^T)) \right] \quad (6)$$

which is the probability density function of a matrix-normal distribution (\rightarrow Proof II/5.1.2) with mean M^T , covariance across rows V and covariance across columns U .

Sources:

- original work

Metadata: ID: P144 | shortcut: matn-trans | author: JoramSoch | date: 2020-08-03, 22:21.

5.1.7 Drawing samples

Theorem: Let $X \in \mathbb{R}^{n \times p}$ be a random matrix (\rightarrow Definition I/1.2.4) with all entries independently following a standard normal distribution (\rightarrow Definition II/3.2.2). 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 Definition II/5.1.1) with mean (\rightarrow Definition I/1.7.13) M , covariance (\rightarrow Definition I/1.9.7) across rows U and covariance (\rightarrow Definition I/1.9.7) 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 Definition II/3.2.2)

$$x_{ij} \sim \mathcal{N}(0, 1) \quad \text{ind. for all } i = 1, \dots, n \quad \text{and} \quad j = 1, \dots, p , \quad (2)$$

this implies a multivariate normal distribution with diagonal covariance matrix (\rightarrow Proof II/4.1.8):

$$\begin{aligned} \text{vec}(X) &\sim \mathcal{N}(\text{vec}(0_{np}), I_{np}) \\ &\sim \mathcal{N}(\text{vec}(0_{np}), I_p \otimes I_n) . \end{aligned} \quad (3)$$

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 Proof II/5.1.3), 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 Proof II/5.1.5), it follows that

$$\begin{aligned}
Y = M + AXB &\sim \mathcal{MN}(M + A0_{np}B, AI_nA^T, B^T I_p B) \\
&\sim \mathcal{MN}(M, AA^T, B^T B) \\
&\sim \mathcal{MN}(M, U, V) .
\end{aligned} \tag{5}$$

Thus, given X defined by (2), Y defined by (1) is a sample (\rightarrow Definition “samp”) from $\mathcal{N}(M, U, V)$.

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.

Metadata: ID: P297 | shortcut: matn-samp | author: JoramSoch | date: 2021-12-07, 08:43.

5.2 Wishart distribution

5.2.1 Definition

Definition: Let X be an $n \times p$ matrix following a matrix-normal distribution (\rightarrow Definition II/5.1.1) with mean zero, independence across rows and covariance across columns V :

$$X \sim \mathcal{MN}(0, I_n, V) . \tag{1}$$

Define the scatter matrix S as the product of the transpose of X with itself:

$$S = X^T X = \sum_{i=1}^n x_i^T x_i . \tag{2}$$

Then, the matrix S is said to follow a Wishart distribution with scale matrix V and degrees of freedom n

$$S \sim \mathcal{W}(V, n) \tag{3}$$

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.

Metadata: ID: D43 | shortcut: wish | author: JoramSoch | date: 2020-03-22, 17:15.

5.2.2 Kullback-Leibler divergence

Theorem: Let S be a $p \times p$ random matrix (\rightarrow Definition I/1.2.4). Assume two Wishart distributions (\rightarrow Definition II/5.2.1) P and Q specifying the probability distribution of S as

$$\begin{aligned}
P : S &\sim \mathcal{W}(V_1, n_1) \\
Q : S &\sim \mathcal{W}(V_2, n_2) .
\end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is given by

$$\text{KL}[P \parallel Q] = \frac{1}{2} \left[n_2 (\ln |V_2| - \ln |V_1|) + n_1 \text{tr}(V_2^{-1} V_1) + 2 \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} + (n_1 - n_2) \psi_p\left(\frac{n_1}{2}\right) - n_1 p \right] \quad (2)$$

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 Definition 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 Definition 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 (\rightarrow Proof “wish-pdf”), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2^{n_1 p} |V_1|^{n_1} \Gamma_p\left(\frac{n_1}{2}\right)}} \cdot |S|^{(n_1 - p - 1)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}(V_1^{-1} S)\right]}{\frac{1}{\sqrt{2^{n_2 p} |V_2|^{n_2} \Gamma_p\left(\frac{n_2}{2}\right)}} \cdot |S|^{(n_2 - p - 1)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}(V_2^{-1} S)\right]} \right\rangle_{p(S)} \\ &= \left\langle \ln \left(\sqrt{2^{(n_2 - n_1)p}} \cdot \frac{|V_2|^{n_2}}{|V_1|^{n_1}} \cdot \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \cdot |S|^{(n_1 - n_2)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}(V_1^{-1} S) - \frac{1}{2} \text{tr}(V_2^{-1} S)\right] \right) \right\rangle_{p(S)} \\ &= \left\langle \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \right. \\ &\quad \left. + \frac{n_1 - n_2}{2} \ln |S| - \frac{1}{2} \text{tr}(V_1^{-1} S) - \frac{1}{2} \text{tr}(V_2^{-1} S) \right\rangle_{p(S)} \\ &= \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \\ &\quad + \frac{n_1 - n_2}{2} \langle \ln |S| \rangle_{p(S)} - \frac{1}{2} \langle \text{tr}(V_1^{-1} S) \rangle_{p(S)} - \frac{1}{2} \langle \text{tr}(V_2^{-1} S) \rangle_{p(S)}. \end{aligned} \quad (7)$$

Using the expected value of a Wishart random matrix (\rightarrow Proof “wish-mean”)

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle S \rangle = nV, \quad (8)$$

such that the expected value of the matrix trace (\rightarrow Proof I/1.7.8) becomes

$$\langle \text{tr}(AS) \rangle = \text{tr}(\langle AS \rangle) = \text{tr}(A \langle S \rangle) = \text{tr}(A \cdot (nV)) = n \cdot \text{tr}(AV), \quad (9)$$

and the expected value of a Wishart log-determinant (\rightarrow Proof “wish-logdetmean”)

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle \ln |S| \rangle = \psi_p\left(\frac{n}{2}\right) + p \cdot \ln 2 + \ln |V|, \quad (10)$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{aligned} \text{KL}[P || Q] &= \frac{(n_2 - n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \\ &\quad + \frac{n_1 - n_2}{2} \left[\psi_p\left(\frac{n_1}{2}\right) + p \cdot \ln 2 + \ln |V_1| \right] - \frac{n_1}{2} \text{tr}(V_1^{-1}V_1) - \frac{n_1}{2} \text{tr}(V_2^{-1}V_1) \\ &= \frac{n_2}{2} (\ln |V_2| - \ln |V_1|) + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} + \frac{n_1 - n_2}{2} \psi_p\left(\frac{n_1}{2}\right) - \frac{n_1}{2} \text{tr}(I_p) - \frac{n_1}{2} \text{tr}(V_2^{-1}V_1) \\ &= \frac{1}{2} \left[n_2 (\ln |V_2| - \ln |V_1|) + n_1 \text{tr}(V_2^{-1}V_1) + 2 \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} + (n_1 - n_2) \psi_p\left(\frac{n_1}{2}\right) - n_1 p \right]. \end{aligned} \quad (11)$$

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.

Metadata: ID: P295 | shortcut: wish-kl | author: JoramSoch | date: 2021-12-02, 15:33.

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 according to a normal distribution (\rightarrow Definition II/3.2.1) with unknown mean μ and unknown variance σ^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>.

Metadata: ID: D124 | shortcut: ug | author: JoramSoch | date: 2021-03-03, 07:21.

1.1.2 Maximum likelihood estimation

Theorem: Let there be a univariate Gaussian data set (\rightarrow Definition 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 Definition I/4.1.3) for mean μ and variance σ^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 Definition I/5.1.2) for each observation is given by the probability density function of the normal distribution (\rightarrow Proof II/3.2.9)

$$p(y_i|\mu, \sigma^2) = \mathcal{N}(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \quad (3)$$

and because observations are independent (\rightarrow Definition 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 Definition 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 μ 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 μ :

$$\begin{aligned}
\frac{d\text{LL}(\hat{\mu}, \sigma^2)}{d\mu} &= 0 \\
0 &= \frac{n}{\sigma^2} (\bar{y} - \hat{\mu}) \\
0 &= \bar{y} - \hat{\mu} \\
\hat{\mu} &= \bar{y} \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y_i .
\end{aligned} \tag{8}$$

The derivative of the log-likelihood function (6) at $\hat{\mu}$ with respect to σ^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 σ^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>.

Metadata: ID: P223 | shortcut: ug-mle | author: JoramSoch | date: 2021-04-16, 11:03.

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 Definition III/1.1.1) with unknown mean μ and unknown variance σ^2 . Then, the test statistic (\rightarrow Definition I/4.3.5)

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \tag{2}$$

with sample mean (\rightarrow Definition I/1.7.2) \bar{y} and sample variance (\rightarrow Definition I/1.8.2) s^2 follows a Student’s t-distribution (\rightarrow Definition II/3.3.1) with $n - 1$ degrees of freedom (\rightarrow Definition “dof”)

$$t \sim t(n - 1) \tag{3}$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu = \mu_0 . \tag{4}$$

Proof: The sample mean (\rightarrow Definition I/1.7.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (5)$$

and the sample variance (\rightarrow Definition I/1.8.2) is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 . \quad (6)$$

Using the linearity of the expected value (\rightarrow Proof I/1.7.5), the additivity of the variance under independence (\rightarrow Proof I/1.8.10) and scaling of the variance upon multiplication (\rightarrow Proof I/1.8.7), the sample mean follows a normal distribution (\rightarrow Definition II/3.2.1)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \sim \mathcal{N} \left(\frac{1}{n} n\mu, \left(\frac{1}{n} \right)^2 n\sigma^2 \right) = \mathcal{N} (\mu, \sigma^2/n) \quad (7)$$

and additionally using the invariance of the variance under addition (\rightarrow Proof I/1.8.6) and applying the null hypothesis from (4), the distribution of $Z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$ becomes standard normal (\rightarrow Definition II/3.2.2)

$$Z = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma} \sim \mathcal{N} \left(\frac{\sqrt{n}}{\sigma} (\mu - \mu_0), \left(\frac{\sqrt{n}}{\sigma} \right)^2 \frac{\sigma^2}{n} \right) \stackrel{H_0}{=} \mathcal{N} (0, 1) . \quad (8)$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution (\rightarrow Proof II/3.2.6), 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 Definition II/3.3.1), the distribution of the test statistic (\rightarrow Definition 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 Definition I/4.3.2) can be rejected when t is as extreme or more extreme than the critical value (\rightarrow Definition I/4.3.9) obtained from the Student's t-distribution (\rightarrow Definition II/3.3.1) with $n-1$ degrees of freedom (\rightarrow Definition “dof”) using a significance level (\rightarrow Definition I/4.3.8) α .

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.

Metadata: ID: P204 | shortcut: ug-ttest1 | author: JoramSoch | date: 2021-03-12, 08:43.

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 a univariate Gaussian data set (\rightarrow Definition III/1.1.1) representing two groups of unequal size n_1 and n_2 with unknown means μ_1 and μ_2 and equal unknown variance σ^2 . Then, the test statistic (\rightarrow Definition 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 Definition I/1.7.2) \bar{y}_1 and \bar{y}_2 and pooled standard deviation (\rightarrow Definition “std-pool”) s_p follows a Student’s t-distribution (\rightarrow Definition II/3.3.1) with $n_1 + n_2 - 1$ degrees of freedom (\rightarrow Definition “dof”)

$$t \sim t(n_1 + n_2 - 1) \quad (3)$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu_1 - \mu_2 = \mu_\Delta . \quad (4)$$

Proof: The sample means (\rightarrow Definition I/1.7.2) 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 Definition “std-pool”) 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 Definition I/1.8.2)

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 \\ s_2^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 . \end{aligned} \quad (7)$$

Using the linearity of the expected value (\rightarrow Proof I/1.7.5), the additivity of the variance under independence (\rightarrow Proof I/1.8.10) and scaling of the variance upon multiplication (\rightarrow Proof I/1.8.7), the sample means follow a normal distribution (\rightarrow Definition II/3.2.1)

$$\begin{aligned}\bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N} \left(\frac{1}{n_1} n_1 \mu_1, \left(\frac{1}{n_1} \right)^2 n_1 \sigma^2 \right) = \mathcal{N} (\mu_1, \sigma^2/n_1) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N} \left(\frac{1}{n_2} n_2 \mu_2, \left(\frac{1}{n_2} \right)^2 n_2 \sigma^2 \right) = \mathcal{N} (\mu_2, \sigma^2/n_2)\end{aligned}\tag{8}$$

and additionally using the invariance of the variance under addition (\rightarrow Proof I/1.8.6) 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 Definition II/3.2.2)

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N} \left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \left(\frac{1}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2 \left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right) \right) \stackrel{H_0}{=} \mathcal{N}(0, 1) . \tag{9}$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution (\rightarrow Proof II/3.2.6), the distribution of $V = (n_1 + n_2 - 2) s_p^2 / \sigma^2$ is

$$V = \frac{(n_1 + n_2 - 2) s_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) . \tag{10}$$

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 Definition II/3.3.1), the distribution of the test statistic (\rightarrow Definition I/4.3.5) is given by

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} \sim t(n_1 + n_2 - 2) . \tag{11}$$

This means that the null hypothesis (\rightarrow Definition I/4.3.2) can be rejected when t is as extreme or more extreme than the critical value (\rightarrow Definition I/4.3.9) obtained from the Student's t-distribution (\rightarrow Definition II/3.3.1) with $n_1 + n_2 - 2$ degrees of freedom (\rightarrow Definition “dof”) using a significance level (\rightarrow Definition I/4.3.8) α .

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.
- 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)).

Metadata: ID: P205 | shortcut: ug-ttest2 | author: JoramSoch | date: 2021-03-12, 09:20.

1.1.5 Paired t-test

Theorem: Let y_{i1} and y_{i2} with $i = 1, \dots, n$ be paired observations, such that

$$y_{i1} \sim \mathcal{N}(y_{i2} + \mu, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

is a univariate Gaussian data set (\rightarrow Definition III/1.1.1) with unknown shift μ and unknown variance σ^2 . Then, the test statistic (\rightarrow Definition 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 Definition I/1.7.2) \bar{d} and sample variance (\rightarrow Definition I/1.8.2) s_d^2 follows a Student's t-distribution (\rightarrow Definition II/3.3.1) with $n - 1$ degrees of freedom (\rightarrow Definition “dof”)

$$t \sim t(n - 1) \quad (3)$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu = \mu_0 . \quad (4)$$

Proof: Define the pair-wise difference $d_i = y_{i1} - y_{i2}$ which is, according to the linearity of the expected value (\rightarrow Proof I/1.7.5) and the invariance of the variance under addition (\rightarrow Proof I/1.8.6), 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 Proof 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.

Metadata: ID: P206 | shortcut: ug-ttestp | author: JoramSoch | date: 2021-03-12, 09:34.

1.1.6 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 Definition III/1.1.1) with unknown mean μ and unknown variance σ^2 . Then, the conjugate prior (\rightarrow Definition I/5.2.5) for this model is a normal-gamma distribution (\rightarrow Definition II/4.3.1)

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau \lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (2)$$

where $\tau = 1/\sigma^2$ is the inverse variance or precision.

Proof: By definition, a conjugate prior (\rightarrow Definition I/5.2.5) is a prior distribution (\rightarrow Definition I/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model 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 Definition I/5.1.2)

$$\begin{aligned}
p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \\
&= \frac{1}{(\sqrt{2\pi}\sigma^2)^n} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{3}$$

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 μ , 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 Definition I/5.1.2) is proportional to a power of τ times an exponential of τ and an exponential of a squared form of μ , weighted by τ :

$$p(y|\mu, \tau) \propto \tau^{n/2} \cdot \exp \left[-\frac{\tau}{2} (y^T y - n\bar{y}^2) \right] \cdot \exp \left[-\frac{\tau n}{2} (\mu - \bar{y})^2 \right]. \quad (8)$$

The same is true for a normal-gamma distribution (\rightarrow Definition II/4.3.1) over μ and τ

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau \lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (9)$$

the probability density function of which (\rightarrow Proof II/4.3.2)

$$p(\mu, \tau) = \sqrt{\frac{\tau \lambda_0}{2\pi}} \cdot \exp \left[-\frac{\tau \lambda_0}{2} (\mu - \mu_0)^2 \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \quad (10)$$

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>.

Metadata: ID: P201 | shortcut: ug-prior | author: JoramSoch | date: 2021-03-03, 08:54.

1.1.7 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 Definition III/1.1.1) with unknown mean μ and unknown variance σ^2 . Moreover, assume a normal-gamma prior distribution (\rightarrow Proof III/1.1.6) over the model parameters μ 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 Definition I/5.1.7) is also a normal-gamma distribution (\rightarrow Definition II/4.3.1)

$$p(\mu, \tau|y) = \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

$$\begin{aligned}
\mu_n &= \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) .
\end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (\rightarrow Proof I/5.3.1), the posterior distribution (\rightarrow Definition I/5.1.7) is given by

$$p(\mu, \tau|y) = \frac{p(y|\mu, \tau) p(\mu, \tau)}{p(y)} . \tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (\rightarrow Proof I/5.1.8) to the numerator:

$$p(\mu, \tau|y) \propto p(y|\mu, \tau) p(\mu, \tau) = p(y, \mu, \tau) . \tag{6}$$

Equation (1) implies the following likelihood function (\rightarrow Definition 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 Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) 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 Proof III/1.1.6) gives

$$\begin{aligned}
p(y, \mu, \tau) &= \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[-\frac{\tau}{2} ((y^T y - 2\mu n \bar{y} + n\mu^2) + \lambda_0 (\mu^2 - 2\mu\mu_0 + \mu_0^2)) \right]
\end{aligned} \tag{11}$$

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

$$\begin{aligned}
p(y, \mu, \tau) &= \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[-\frac{\tau}{2} (\mu^2 (\lambda_0 + n) - 2\mu (\lambda_0 \mu_0 + n \bar{y}) + (y^T y + \lambda_0 \mu_0^2)) \right]
\end{aligned} \tag{12}$$

Completing the square over μ , we finally have

$$\begin{aligned}
p(y, \mu, \tau) &= \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\quad \exp \left[-\frac{\tau \lambda_n}{2} (\mu - \mu_n)^2 - \frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right]
\end{aligned} \tag{13}$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned}
\mu_n &= \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n .
\end{aligned} \tag{14}$$

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 Definition I/5.1.7)

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) . \end{aligned} \quad (16)$$

From the term in (13), we can isolate the posterior distribution over μ given τ :

$$p(\mu|\tau, y) = \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) . \quad (17)$$

From the remaining term, we can isolate the posterior distribution over τ :

$$p(\tau|y) = \text{Gam}(\tau; a_n, b_n) . \quad (18)$$

Together, (17) and (18) constitute the joint (\rightarrow Definition I/1.3.2) posterior distribution (\rightarrow Definition I/5.1.7) of μ and τ .

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>.

Metadata: ID: P202 | shortcut: ug-post | author: JoramSoch | date: 2021-03-03, 09:53.

1.1.8 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 Definition III/1.1.1) with unknown mean μ and unknown variance σ^2 . Moreover, assume a normal-gamma prior distribution (\rightarrow Proof III/1.1.6) over the model parameters μ 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 Definition IV/3.1.1) 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 Definition I/5.1.7) are given by

$$\begin{aligned}
\mu_n &= \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) .
\end{aligned} \tag{4}$$

Proof: According to the law of marginal probability (\rightarrow Definition I/1.3.3), the model evidence (\rightarrow Definition I/5.1.9) for this model is:

$$p(y|m) = \iint p(y|\mu, \tau) p(\mu, \tau) d\mu d\tau . \tag{5}$$

According to the law of conditional probability (\rightarrow Definition I/1.3.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(y|m) = \iint p(y, \mu, \tau) d\mu d\tau . \tag{6}$$

Equation (1) implies the following likelihood function (\rightarrow Definition 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 Proof III/1.1.7) $p(\mu, \tau|y)$, the joint likelihood $p(y, \mu, \tau)$ is obtained as

$$p(y, \mu, \tau) = \sqrt{\frac{\tau^{n+1} \lambda_0}{(2\pi)^{n+1}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[-\frac{\tau \lambda_n}{2} (\mu - \mu_n)^2 - \frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] . \quad (9)$$

Using the probability density function of the normal distribution (\rightarrow Proof II/3.2.9), we can rewrite this as

$$p(y, \mu, \tau) = \sqrt{\frac{\tau^n}{(2\pi)^n}} \sqrt{\frac{\tau \lambda_0}{2\pi}} \sqrt{\frac{2\pi}{\tau \lambda_n}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \mathcal{N}(\mu; \mu_n, (\tau \lambda_n)^{-1}) \exp \left[-\frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] . \quad (10)$$

Now, μ can be integrated out easily:

$$\int p(y, \mu, \tau) d\mu = \sqrt{\frac{1}{(2\pi)^n}} \sqrt{\frac{\lambda_0}{\lambda_n}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0+n/2-1} \exp[-b_0 \tau] \cdot \exp \left[-\frac{\tau}{2} (y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right] . \quad (11)$$

Using the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), 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, τ 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 Definition IV/3.1.1) 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>.

Metadata: ID: P203 | shortcut: ug-lme | author: JoramSoch | date: 2021-03-03, 10:25.

1.1.9 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 Definition III/1.1.1) with unknown mean μ and unknown variance σ^2 . Moreover, assume a normal-gamma prior distribution (\rightarrow Proof III/1.1.6) over the model parameters μ 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 Proof IV/3.1.3) 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 μ_n and λ_n as well as a_n and b_n are the posterior hyperparameters for the univariate Gaussian (\rightarrow Proof III/1.1.7) and \bar{y} is the sample mean (\rightarrow Definition I/1.7.2).

Proof: Model accuracy and complexity are defined as (\rightarrow Proof IV/3.1.3)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\mu, m) \rangle_{p(\mu|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\mu|y, m) || p(\mu|m)] . \end{aligned} \quad (4)$$

The accuracy term is the expectation (\rightarrow Definition I/1.7.1) of the log-likelihood function (\rightarrow Definition I/4.1.2) $\log p(y|\mu, \tau)$ with respect to the posterior distribution (\rightarrow Definition I/5.1.7) $p(\mu, \tau|y)$. With the log-likelihood function for the univariate Gaussian (\rightarrow Proof III/1.1.2) and the posterior distribution for the univariate Gaussian (\rightarrow Proof III/1.1.7), the model accuracy of m evaluates to:

$$\begin{aligned} \text{Acc}(m) &= \langle \log p(y|\mu, \tau) \rangle_{p(\mu, \tau|y)} \\ &= \left\langle \langle \log p(y|\mu, \tau) \rangle_{p(\mu|\tau, y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \frac{n}{2} \log(\tau) - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y^T y - 2n\bar{y}\mu + n\mu^2) \right\rangle_{\mathcal{N}(\mu_n, (\tau \lambda_n)^{-1})} \right\rangle_{\text{Gam}(a_n, b_n)} \\ &= \left\langle \frac{n}{2} \log(\tau) - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} \right\rangle_{\text{Gam}(a_n, b_n)} \\ &= \frac{n}{2} (\psi(a_n) - \log(b_n)) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \frac{a_n}{b_n} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} \\ &= -\frac{1}{2} \frac{a_n}{b_n} (y^T y - 2n\bar{y}\mu_n + n\mu_n^2) - \frac{1}{2} n\lambda_n^{-1} + \frac{n}{2} (\psi(a_n) - \log(b_n)) - \frac{n}{2} \log(2\pi) \end{aligned} \quad (5)$$

The complexity penalty is the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of the posterior distribution (\rightarrow Definition I/5.1.7) $p(\mu, \tau|y)$ from the prior distribution (\rightarrow Definition I/5.1.3) $p(\mu, \tau)$. With the prior distribution (\rightarrow Proof III/1.1.6) given by (2), the posterior distribution for the univariate Gaussian (\rightarrow Proof III/1.1.7) and the Kullback-Leibler divergence of the normal-gamma distribution (\rightarrow Proof II/4.3.5), the model complexity of m evaluates to:

$$\begin{aligned}
\text{Com}(m) &= \text{KL} [p(\mu, \tau|y) || p(\mu, \tau)] \\
&= \text{KL} [\text{NG}(\mu_n, \lambda_n^{-1}, a_n, b_n) || \text{NG}(\mu_0, \lambda_0^{-1}, a_0, b_0)] \\
&= \frac{1}{2} \frac{a_n}{b_n} [\lambda_0(\mu_0 - \mu_n)^2] + \frac{1}{2} \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&\quad + a_0 \cdot \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \cdot \psi(a_n) - (b_n - b_0) \cdot \frac{a_n}{b_n} \\
&= \frac{1}{2} \frac{a_n}{b_n} [\lambda_0(\mu_0 - \mu_n)^2 - 2(b_n - b_0)] + \frac{1}{2} \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \log \frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&\quad + a_0 \cdot \log \frac{b_n}{b_0} - \log \frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0) \cdot \psi(a_n) .
\end{aligned} \tag{6}$$

A control calculation confirms that

$$\text{Acc}(m) - \text{Com}(m) = \text{LME}(m) \tag{7}$$

where $\text{LME}(m)$ is the log model evidence for the univariate Gaussian (\rightarrow Proof III/1.1.8).

Sources:

- original work

Metadata: ID: P240 | shortcut: ug-anc | author: JoramSoch | date: 2021-07-14, 08:26.

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 according to a normal distribution (\rightarrow Definition II/3.2.1) with unknown mean μ and known variance σ^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>.

Metadata: ID: D136 | shortcut: ugkv | author: JoramSoch | date: 2021-03-23, 16:12.

1.2.2 Maximum likelihood estimation

Theorem: Let there be univariate Gaussian data with known variance (\rightarrow Definition III/1.2.1) $y = \{y_1, \dots, y_n\}$:

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate (\rightarrow Definition I/4.1.3) for the mean μ is given by

$$\hat{\mu} = \bar{y} \quad (2)$$

where \bar{y} is the sample mean (\rightarrow Definition I/1.7.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (3)$$

Proof: The likelihood function (\rightarrow Definition I/5.1.2) for each observation is given by the probability density function of the normal distribution (\rightarrow Proof II/3.2.9)

$$p(y_i|\mu) = \mathcal{N}(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \quad (4)$$

and because observations are independent (\rightarrow Definition 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 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 Definition 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 μ 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:

$$\begin{aligned}
\frac{dLL(\hat{\mu})}{d\mu} &= 0 \\
0 &= \frac{n}{\sigma^2}(\bar{y} - \hat{\mu}) \\
0 &= \bar{y} - \hat{\mu} \\
\hat{\mu} &= \bar{y}
\end{aligned} \tag{9}$$

Plugging this value into the second derivative, we confirm:

$$\frac{d^2LL(\hat{\mu})}{d\mu^2} = -\frac{n}{\sigma^2} < 0 . \tag{10}$$

This demonstrates that the estimate $\hat{\mu} = \bar{y}$ maximizes the likelihood $p(y|\mu)$.

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>.

Metadata: ID: P207 | shortcut: ugkv-mle | author: JoramSoch | date: 2021-03-24, 03:48.

1.2.3 One-sample z-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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Then, the test statistic (\rightarrow Definition I/4.3.5)

$$z = \sqrt{n} \frac{\bar{y} - \mu_0}{\sigma} \tag{2}$$

with sample mean (\rightarrow Definition I/1.7.2) \bar{y} follows a standard normal distribution (\rightarrow Definition II/3.2.2)

$$z \sim \mathcal{N}(0, 1) \tag{3}$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu = \mu_0 . \tag{4}$$

Proof: The sample mean (\rightarrow Definition I/1.7.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \tag{5}$$

Using the linearity of the expected value (\rightarrow Proof I/1.7.5), the additivity of the variance under independence (\rightarrow Proof I/1.8.10) and scaling of the variance upon multiplication (\rightarrow Proof I/1.8.7), the sample mean follows a normal distribution (\rightarrow Definition II/3.2.1)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \sim \mathcal{N} \left(\frac{1}{n} n \mu, \left(\frac{1}{n} \right)^2 n \sigma^2 \right) = \mathcal{N} (\mu, \sigma^2/n) \quad (6)$$

and additionally using the invariance of the variance under addition (\rightarrow Proof I/1.8.6), 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 Definition I/4.3.2) can be rejected when z is as extreme or more extreme than the critical value (\rightarrow Definition I/4.3.9) obtained from the standard normal distribution (\rightarrow Definition II/3.2.2) using a significance level (\rightarrow Definition I/4.3.8) α .

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.
- 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>.

Metadata: ID: P208 | shortcut: ugkv-ztest1 | author: JoramSoch | date: 2021-03-24, 04:23.

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 a univariate Gaussian data set (\rightarrow Definition III/1.1.1) representing two groups of unequal size n_1 and n_2 with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2 . Then, the test statistic (\rightarrow Definition 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 Definition I/1.7.2) \bar{y}_1 and \bar{y}_2 follows a standard normal distribution (\rightarrow Definition II/3.2.2)

$$z \sim \mathcal{N}(0, 1) \quad (3)$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu_1 - \mu_2 = \mu_\Delta. \quad (4)$$

Proof: The sample means (\rightarrow Definition I/1.7.2) are given by

$$\begin{aligned}\bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} .\end{aligned}\tag{5}$$

Using the linearity of the expected value (\rightarrow Proof I/1.7.5), the additivity of the variance under independence (\rightarrow Proof I/1.8.10) and scaling of the variance upon multiplication (\rightarrow Proof I/1.8.7), the sample means follow a normal distribution (\rightarrow Definition II/3.2.1)

$$\begin{aligned}\bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N} \left(\frac{1}{n_1} n_1 \mu_1, \left(\frac{1}{n_1} \right)^2 n_1 \sigma^2 \right) = \mathcal{N} (\mu_1, \sigma_1^2/n_1) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N} \left(\frac{1}{n_2} n_2 \mu_2, \left(\frac{1}{n_2} \right)^2 n_2 \sigma^2 \right) = \mathcal{N} (\mu_2, \sigma_2^2/n_2)\end{aligned}\tag{6}$$

and additionally using the invariance of the variance under addition (\rightarrow Proof I/1.8.6), the distribution of $z = [(\bar{y}_1 - \bar{y}_2) - \mu_\Delta]/\sigma_\Delta$ becomes

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma_\Delta} \sim \mathcal{N} \left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, \left(\frac{1}{\sigma_\Delta} \right)^2 \sigma_\Delta^2 \right) = \mathcal{N} \left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, 1 \right)\tag{7}$$

where σ_Δ is the pooled standard deviation (\rightarrow Definition “std-pool”)

$$\sigma_\Delta = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\tag{8}$$

such that, under the null hypothesis in (4), we have:

$$z \sim \mathcal{N}(0, 1), \quad \text{if } \mu_\Delta = \mu_1 - \mu_2 .\tag{9}$$

This means that the null hypothesis (\rightarrow Definition I/4.3.2) can be rejected when z is as extreme or more extreme than the critical value (\rightarrow Definition I/4.3.9) obtained from the standard normal distribution (\rightarrow Definition II/3.2.2) using a significance level (\rightarrow Definition I/4.3.8) α .

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.
- 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.

Metadata: ID: P209 | shortcut: ugkv-ztest2 | author: JoramSoch | date: 2021-03-24, 04:38.

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 Definition III/1.2.1) with unknown shift μ and known variance σ^2 . Then, the test statistic (\rightarrow Definition 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 Definition I/1.7.2) \bar{d} follows a standard normal distribution (\rightarrow Definition II/3.2.2)

$$z \sim \mathcal{N}(0, 1) \quad (3)$$

under the null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : \mu = \mu_0 . \quad (4)$$

Proof: Define the pair-wise difference $d_i = y_{i1} - y_{i2}$ which is, according to the linearity of the expected value (\rightarrow Proof I/1.7.5) and the invariance of the variance under addition (\rightarrow Proof I/1.8.6), 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 Proof 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.
- 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).

Metadata: ID: P210 | shortcut: ugkv-ztestp | author: JoramSoch | date: 2021-03-24, 05:10.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Then, the conjugate prior (\rightarrow Definition I/5.2.5) for this model is a normal distribution (\rightarrow Definition II/3.2.1)

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \quad (2)$$

with prior (\rightarrow Definition I/5.1.3) mean (\rightarrow Definition I/1.7.1) μ_0 and prior (\rightarrow Definition I/5.1.3) precision (\rightarrow Definition I/1.8.12) λ_0 .

Proof: By definition, a conjugate prior (\rightarrow Definition I/5.2.5) is a prior distribution (\rightarrow Definition I/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition 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 Definition 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

$$\begin{aligned} p(y|\mu) &= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2) \right] \\ &= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp \left[-\frac{\tau}{2} \left(\sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2 \right) \right] \\ &= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp \left[-\frac{\tau}{2} (y^T y - 2\mu n\bar{y} + n\mu^2) \right] \\ &= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp \left[-\frac{\tau n}{2} \left(\frac{1}{n} y^T y - 2\mu\bar{y} + \mu^2 \right) \right] \end{aligned} \quad (5)$$

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 μ , finally gives

$$p(y|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \exp \left[-\frac{\tau n}{2} \left((\mu - \bar{y})^2 - \bar{y}^2 + \frac{1}{n} y^T y \right) \right] \quad (6)$$

In other words, the likelihood function (\rightarrow Definition I/5.1.2) is proportional to an exponential of a squared form of μ , weighted by some constant:

$$p(y|\mu) \propto \exp \left[-\frac{\tau n}{2} (\mu - \bar{y})^2 \right]. \quad (7)$$

The same is true for a normal distribution (\rightarrow Definition II/3.2.1) over μ

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \quad (8)$$

the probability density function of which (\rightarrow Proof II/3.2.9)

$$p(\mu) = \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[-\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] \quad (9)$$

exhibits the same proportionality

$$p(\mu) \propto \exp \left[-\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right] \quad (10)$$

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>.

Metadata: ID: P211 | shortcut: ugkv-prior | author: JoramSoch | date: 2021-03-24, 05:57.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume a normal distribution (\rightarrow Proof III/1.2.6) over the model parameter μ :

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}). \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a normal distribution (\rightarrow Definition II/3.2.1)

$$p(\mu|y) = \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

$$\begin{aligned}\mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n\end{aligned}\tag{4}$$

with the sample mean (\rightarrow Definition I/1.7.2) \bar{y} and the inverse variance or precision (\rightarrow Definition I/1.8.12) $\tau = 1/\sigma^2$.

Proof: According to Bayes' theorem (\rightarrow Proof I/5.3.1), the posterior distribution (\rightarrow Definition I/5.1.7) is given by

$$p(\mu|y) = \frac{p(y|\mu)p(\mu)}{p(y)}.\tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (\rightarrow Proof I/5.1.8) to the numerator:

$$p(\mu|y) \propto p(y|\mu)p(\mu) = p(y, \mu).\tag{6}$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$\begin{aligned}p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \left(\sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right]\end{aligned}\tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}p(y|\mu) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\ &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[-\frac{\tau}{2} (y_i - \mu)^2 \right] \\ &= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]\end{aligned}\tag{8}$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

Combining the likelihood function (\rightarrow Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) of the model is given by

$$\begin{aligned}p(y, \mu) &= p(y|\mu)p(\mu) \\ &= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[-\frac{\lambda_0}{2} (\mu - \mu_0)^2 \right].\end{aligned}\tag{9}$$

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 Proof 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 μ 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 Definition I/5.1.7)

$$\begin{aligned} \mu_n &= \frac{\lambda_0\mu_0 + \tau n\bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \quad (13)$$

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 μ 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>.

Metadata: ID: P212 | shortcut: ugkv-post | author: JoramSoch | date: 2021-03-24, 06:10.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume a normal distribution (\rightarrow Proof III/1.2.6) over the model parameter μ :

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) 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 Definition I/5.1.7) 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 Definition I/1.7.2) \bar{y} and the inverse variance or precision (\rightarrow Definition I/1.8.12) $\tau = 1/\sigma^2$.

Proof: According to the law of marginal probability (\rightarrow Definition I/1.3.3), the model evidence (\rightarrow Definition I/5.1.9) 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 Definition I/1.3.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(y|m) = \int p(y, \mu) d\mu . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$\begin{aligned} p(y|\mu, \sigma^2) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \left(\sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right] \end{aligned} \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned}
p(y|\mu, \tau) &= \prod_{i=1}^n \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \cdot \exp \left[-\frac{\tau}{2} (y_i - \mu)^2 \right] \\
&= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]
\end{aligned} \tag{8}$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

When deriving the posterior distribution (\rightarrow Proof III/1.2.7) $p(\mu|y)$, the joint likelihood $p(y, \mu)$ is obtained as

$$p(y, \mu) = \left(\frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp \left[-\frac{\lambda_n}{2} (\mu - \mu_n)^2 - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right]. \tag{9}$$

Using the probability density function of the normal distribution (\rightarrow Proof II/3.2.9), we can rewrite this as

$$p(y, \mu) = \left(\frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \sqrt{\frac{2\pi}{\lambda_n}} \cdot \mathcal{N}(\mu; \lambda_n^{-1}) \cdot \exp \left[-\frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right]. \tag{10}$$

Now, μ can be integrated out using the properties of the probability density function (\rightarrow Definition I/1.6.6):

$$p(y|m) = \int p(y, \mu) d\mu = \left(\frac{\tau}{2\pi} \right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{\lambda_n}} \cdot \exp \left[-\frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2) \right]. \tag{11}$$

Thus, the log model evidence (\rightarrow Definition IV/3.1.1) of this model is given by

$$\log p(y|m) = \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2). \tag{12}$$

Sources:

- original work

Metadata: ID: P213 | shortcut: ugkv-lme | author: JoramSoch | date: 2021-03-24, 06:45.

1.2.9 Accuracy and complexity

Theorem: Let

$$y = \{y_1, \dots, y_n\}, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

be a univariate Gaussian data set (\rightarrow Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume a statistical model (\rightarrow Definition I/5.1.4) imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ :

$$m : y_i \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}). \tag{2}$$

Then, accuracy and complexity (\rightarrow Proof IV/3.1.3) 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 μ_n and λ_n are the posterior hyperparameters for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.7), $\tau = 1/\sigma^2$ is the inverse variance or precision (\rightarrow Definition I/1.8.12) and \bar{y} is the sample mean (\rightarrow Definition I/1.7.2).

Proof: Model accuracy and complexity are defined as (\rightarrow Proof IV/3.1.3)

$$\begin{aligned} \text{LME}(m) &= \text{Acc}(m) - \text{Com}(m) \\ \text{Acc}(m) &= \langle \log p(y|\mu, m) \rangle_{p(\mu|y, m)} \\ \text{Com}(m) &= \text{KL} [p(\mu|y, m) || p(\mu|m)] . \end{aligned} \quad (4)$$

The accuracy term is the expectation (\rightarrow Definition I/1.7.1) of the log-likelihood function (\rightarrow Definition I/4.1.2) $\log p(y|\mu)$ with respect to the posterior distribution (\rightarrow Definition I/5.1.7) $p(\mu|y)$. With the log-likelihood function for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.2) and the posterior distribution for the univariate Gaussian with known variance (\rightarrow Proof 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 Definition I/2.5.1) of the posterior distribution (\rightarrow Definition I/5.1.7) $p(\mu|y)$ from the prior distribution (\rightarrow Definition I/5.1.3) $p(\mu)$. With the prior distribution (\rightarrow Proof III/1.2.6) given by (2), the posterior distribution for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.7) and the Kullback-Leibler divergence of the normal distribution (\rightarrow Proof II/3.2.22), 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 Proof III/1.2.8).

Sources:

- original work

Metadata: ID: P214 | shortcut: ugkv-anc | author: JoramSoch | date: 2021-03-24, 07:49.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume two statistical models (\rightarrow Definition I/5.1.4), one assuming that μ is zero (null model (\rightarrow Definition I/4.3.2)), the other imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ (alternative (\rightarrow Definition 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 Definition IV/3.3.1) 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 μ_n and λ_n are the posterior hyperparameters for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.7) which are functions of the inverse variance or precision (\rightarrow Definition I/1.8.12) $\tau = 1/\sigma^2$ and the sample mean (\rightarrow Definition I/1.7.2) \bar{y} .

Proof: The log Bayes factor is equal to the difference of two log model evidences (\rightarrow Proof IV/3.3.3):

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2). \quad (4)$$

The LME of the alternative m_1 is equal to the log model evidence for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.8):

$$\text{LME}(m_1) = \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} (\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2). \quad (5)$$

Because the null model m_0 has no free parameter, its log model evidence (\rightarrow Definition IV/3.1.1) (logarithmized marginal likelihood (\rightarrow Definition I/5.1.9)) is equal to the log-likelihood function for the univariate Gaussian with known variance (\rightarrow Proof 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 Definition I/5.1.7) are given by (\rightarrow Proof 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 Definition I/1.7.2) \bar{y} and the inverse variance or precision (\rightarrow Definition I/1.8.12) $\tau = 1/\sigma^2$.

Sources:

- original work

Metadata: ID: P215 | shortcut: ugkv-lbf | author: JoramSoch | date: 2021-03-24, 09:05.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume two statistical models (\rightarrow Definition I/5.1.4), one assuming that μ is zero (null model (\rightarrow Definition I/4.3.2)), the other imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ (alternative (\rightarrow Definition 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 Definition I/4.3.2) that m_0 generated the data, the expectation (\rightarrow Definition I/1.7.1) of the log Bayes factor (\rightarrow Definition IV/3.3.1) 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 λ_n is the posterior precision for the univariate Gaussian with known variance (\rightarrow Proof III/1.2.7).

Proof: The log Bayes factor for the univariate Gaussian with known variance (\rightarrow Proof 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 Definition I/5.1.7) are given by (\rightarrow Proof 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 Definition I/1.7.2) \bar{y} and the inverse variance or precision (\rightarrow Definition I/1.8.12) $\tau = 1/\sigma^2$. Plugging μ_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 Definition I/5.1.3) with prior mean (\rightarrow Definition I/1.7.1) $\mu_0 = 0$ 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 Definition I/1.7.1) 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 Definition I/1.7.1) 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 Definition 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 λ_n from (5) into (10), we obtain:

$$\langle \text{LBF}_{10} \rangle = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left(\frac{\lambda_n - \lambda_0}{\lambda_n} \right).\tag{11}$$

Sources:

- original work

Metadata: ID: P216 | shortcut: ugkv-lbfmean | author: JoramSoch | date: 2021-03-24, 10:03.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume two statistical models (\rightarrow Definition I/5.1.4), one assuming that μ is zero (null model (\rightarrow Definition I/4.3.2)), the other imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ (alternative (\rightarrow Definition 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 Definition IV/3.1.5) 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 Definition I/1.7.2), $\tau = 1/\sigma^2$ is the inverse variance or precision (\rightarrow Definition I/1.8.12), $y_1^{(i)}$ are the training data in the i -th cross-validation fold and S is the number of data subsets (\rightarrow Definition IV/3.1.5).

Proof: For evaluation of the cross-validated log model evidences (\rightarrow Definition IV/3.1.5) (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 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 Definition “ooslme”) (oosLME) is equal to the log-likelihood function (\rightarrow Proof III/1.2.2) of the test data at $\mu = 0$:

$$\text{oosLME}_i(m_0) = \log p\left(y_2^{(i)} \middle| \mu = 0\right) = \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} \left[\tau y_2^{(i)\top} y_2^{(i)} \right]. \quad (6)$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences (\rightarrow Definition IV/3.1.5) over cross-validation folds, such that the cvLME of m_0 is:

$$\begin{aligned} \text{cvLME}(m_0) &= \sum_{i=1}^S \text{oosLME}_i(m_0) \\ &= \sum_{i=1}^S \left(\frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} \left[\tau y_2^{(i)\top} y_2^{(i)} \right] \right) \\ &= \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2} [\tau y^\top y]. \end{aligned} \quad (7)$$

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 Definition I/5.2.3) and applying the posterior distribution for the univariate Gaussian with known variance (\rightarrow Proof 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 Definition I/5.2.3), again applying the posterior distribution for the univariate Gaussian with known variance (\rightarrow Proof 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 Proof III/1.2.8), the out-of-sample log model evidence (\rightarrow Definition “ooslme”) (oosLME) therefore follows as

$$\begin{aligned} \text{oosLME}_i(m_1) &= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{\lambda_0^{(2)}}{\lambda_n^{(2)}}\right) - \frac{1}{2} \left[\tau y_2^{(i)\top} y_2^{(i)} + \lambda_0^{(2)} \mu_0^{(2)2} - \lambda_n^{(2)} \mu_n^{(2)2} \right] \\ &= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{n_1}{n}\right) - \frac{1}{2} \left[\tau y_2^{(i)\top} y_2^{(i)} + \frac{\tau}{n_1} \left(n_1 \bar{y}_1^{(i)} \right)^2 - \frac{\tau}{n} (n \bar{y})^2 \right]. \end{aligned} \quad (10)$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences (\rightarrow Definition IV/3.1.5) over cross-validation folds, the cvLME of m_1 becomes:

$$\begin{aligned}
 \text{cvLME}(m_1) &= \sum_{i=1}^S \text{oosLME}_i(m_1) \\
 &= \sum_{i=1}^S \left(\frac{n_2}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left(\frac{n_1}{n} \right) - \frac{1}{2} \left[\tau y_2^{(i)\top} y_2^{(i)} + \frac{\tau}{n_1} \left(n_1 \bar{y}_1^{(i)} \right)^2 - \frac{\tau}{n} (n\bar{y})^2 \right] \right) \\
 &= \frac{S \cdot n_2}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left(\frac{n_1}{n} \right) - \frac{\tau}{2} \sum_{i=1}^S \left[y_2^{(i)\top} y_2^{(i)} + \frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right] \\
 &= \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \left[y^\top y + \sum_{i=1}^S \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right].
 \end{aligned} \tag{11}$$

Together, (7) and (11) conform to the results given in (3).

Sources:

- original work

Metadata: ID: P217 | shortcut: ugkv-cvlme | author: JoramSoch | date: 2021-03-24, 10:57.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume two statistical models (\rightarrow Definition I/5.1.4), one assuming that μ is zero (null model (\rightarrow Definition I/4.3.2)), the other imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ (alternative (\rightarrow Definition 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 Definition IV/3.1.5) log Bayes factor (\rightarrow Definition IV/3.3.1) 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 Definition I/1.7.2), $\tau = 1/\sigma^2$ is the inverse variance or precision (\rightarrow Definition I/1.8.12), $y_1^{(i)}$ are the training data in the i -th cross-validation fold and S is the number of data subsets (\rightarrow Definition IV/3.1.5).

Proof: The relationship between log Bayes factor and log model evidences (\rightarrow Proof IV/3.3.3) also holds for cross-validated log bayes factor (\rightarrow Definition IV/3.3.1) (cvLBF) and cross-validated log model evidences (\rightarrow Definition IV/3.1.5) (cvLME):

$$\text{cvLBF}_{12} = \text{cvLME}(m_1) - \text{cvLME}(m_2) . \quad (4)$$

The cross-validated log model evidences (\rightarrow Definition IV/3.1.5) of m_0 and m_1 are given by (\rightarrow Proof 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

$$\begin{aligned} \text{cvLBF}_{10} &= \text{cvLME}(m_1) - \text{LME}(m_0) \\ &= \left(\frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \left[y^T y + \sum_{i=1}^S \left(\frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] \right) \\ &\quad - \left(\frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) - \frac{1}{2} (\tau y^T y) \right) \\ &= \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S \left(\frac{(n_1 \bar{y}_1^{(i)})^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P218 | shortcut: ugkv-cvlf | author: JoramSoch | date: 2021-03-24, 11:13.

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 Definition III/1.2.1) with unknown mean μ and known variance σ^2 . Moreover, assume two statistical models (\rightarrow Definition I/5.1.4), one assuming that μ is zero (null model (\rightarrow Definition I/4.3.2)), the other imposing a normal distribution (\rightarrow Proof III/1.2.6) as the prior distribution (\rightarrow Definition I/5.1.3) on the model parameter μ (alternative (\rightarrow Definition 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 Definition I/1.7.1) of the cross-validated (\rightarrow Definition IV/3.1.5) log Bayes factor (\rightarrow Definition IV/3.3.1) (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 Definition I/1.8.12) and S is the number of data subsets (\rightarrow Definition IV/3.1.5).

Proof: The cross-validated log Bayes factor for the univariate Gaussian with known variance (\rightarrow Proof 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 Definition I/1.7.1) 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 Definition I/1.7.1) 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 Proof III/1.2.12) $n_1 + n_2 = n$ and $n_2 = n/S$, we finally have:

$$\begin{aligned}
\langle \text{cvLBF}_{10} \rangle &= \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^S (-n_2) \mu^2 \\
&= \frac{S}{2} \log \left(\frac{S-1}{S} \right) + \frac{1}{2} [\tau n \mu^2] .
\end{aligned} \tag{7}$$

Sources:

- original work

Metadata: ID: P219 | shortcut: ugkv-cvlfmean | author: JoramSoch | date: 2021-03-24, 12:27.

1.3 Simple linear regression

1.3.1 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 Definition II/4.1.1) for ε

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{2}$$

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”;
- β_1 is called “slope of the regression line (\rightarrow Definition III/1.3.9)”;
- β_0 is called “intercept of the regression line (\rightarrow Definition III/1.3.9)”;
- ε is called “noise”, “errors” or “error terms”;
- σ^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 (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) . \tag{3}$$

In this case, the linear regression model can also be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) . \tag{4}$$

Otherwise, it is called simple linear regression with correlated observations.

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.

Metadata: ID: D163 | shortcut: slr | author: JoramSoch | date: 2021-10-27, 07:07.

1.3.2 Special case of multiple linear regression

Theorem: Simple linear regression (\rightarrow Definition III/1.3.1) is a special case of multiple linear regression (\rightarrow Definition III/1.4.1) with design matrix X and regression coefficients β

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (1)$$

where 1_n is an $n \times 1$ vector of ones, x is the $n \times 1$ single predictor variable, β_0 is the intercept and β_1 is the slope of the regression line (\rightarrow Definition III/1.3.9).

Proof: Without loss of generality, consider the simple linear regression case with uncorrelated errors (\rightarrow Definition III/1.3.1):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n. \quad (2)$$

In matrix notation and using the multivariate normal distribution (\rightarrow Definition 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 Definition III/1.4.1), 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 Definition III/1.3.1), the error distribution changes to (\rightarrow Definition III/1.4.1):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V). \quad (5)$$

Sources:

- original work

Metadata: ID: P281 | shortcut: slr-mlr | author: JoramSoch | date: 2021-11-09, 07:57.

1.3.3 Ordinary least squares

Theorem: Given a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Definition III/1.4.6) 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 Definition I/1.7.2), s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x and s_{xy} is the sample covariance (\rightarrow Definition I/1.9.2) between x and y .

Proof: The residual sum of squares (\rightarrow Definition III/1.4.6) 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 β_0 and β_1 are

$$\begin{aligned} \frac{d\text{RSS}(\beta_0, \beta_1)}{d\beta_0} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) \\ &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{d\text{RSS}(\beta_0, \beta_1)}{d\beta_1} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \\ &= -2 \sum_{i=1}^n (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \end{aligned} \quad (4)$$

and setting these derivatives to zero

$$\begin{aligned} 0 &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ 0 &= -2 \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \end{aligned} \quad (5)$$

yields the following equations:

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^n x_i + \hat{\beta}_0 \cdot n &= \sum_{i=1}^n y_i \\ \hat{\beta}_1 \sum_{i=1}^n x_i^2 + \hat{\beta}_0 \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i. \end{aligned} \quad (6)$$

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.
- 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.

Metadata: ID: P271 | shortcut: slr-ols | author: JoramSoch | date: 2021-10-27, 08:56.

1.3.4 Ordinary least squares

Theorem: Given a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Definition III/1.4.6) 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 Definition I/1.7.2), s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x and s_{xy} is the sample covariance (\rightarrow Definition I/1.9.2) between x and y .

Proof: Simple linear regression is a special case of multiple linear regression (\rightarrow Proof III/1.3.2) 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 Proof III/1.4.2) 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 Proof III/1.3.3):

$$\begin{aligned}\hat{\beta}_1 &= \frac{n x^T y - (n\bar{x})(n\bar{y})}{nx^T x - (n\bar{x})^2} \\ &= \frac{x^T y - n\bar{x}\bar{y}}{x^T x - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{s_{xy}}{s_x^2}.\end{aligned}\tag{6}$$

Moreover, the first entry of $\hat{\beta}$ is equal to:

$$\begin{aligned}
 \hat{\beta}_0 &= \frac{n\bar{y}x^T x - n\bar{x}x^T y}{nx^T x - (n\bar{x})^2} \\
 &= \frac{\bar{y}x^T x - \bar{x}x^T y}{x^T x - n\bar{x}^2} \\
 &= \frac{\bar{y}x^T x - \bar{x}x^T y + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y}}{x^T x - n\bar{x}^2} \\
 &= \frac{\bar{y}(x^T x - n\bar{x}^2) - \bar{x}(x^T y - n\bar{x}\bar{y})}{x^T x - n\bar{x}^2} \\
 &= \frac{\bar{y}(x^T x - n\bar{x}^2)}{x^T x - n\bar{x}^2} - \frac{\bar{x}(x^T y - n\bar{x}\bar{y})}{x^T x - n\bar{x}^2} \\
 &= \bar{y} - \bar{x} \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
 &= \bar{y} - \hat{\beta}_1 \bar{x} .
 \end{aligned} \tag{7}$$

Sources:

- original work

Metadata: ID: P288 | shortcut: slr-ols2 | author: JoramSoch | date: 2021-11-16, 09:36.

1.3.5 Expectation of estimates

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3). Then, the expected values (\rightarrow Definition I/1.7.1) 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 Proof III/1.3.3) produces unbiased estimators (\rightarrow Definition “est-unb”).

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 Proof III/1.3.3) 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.
- 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.

Metadata: ID: P272 | shortcut: slr-olsmean | author: JoramSoch | date: 2021-10-27, 09:54.

1.3.6 Variance of estimates

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3). Then, the variances (\rightarrow Definition I/1.8.1) 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 Definition I/1.8.2) 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 Proof III/1.3.3) 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

$$\begin{aligned}\sum_{i=1}^n c_i^2 &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \\ &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}\tag{6}$$

With (5), the estimate for the slope from (4) becomes

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum_{i=1}^n c_i (y_i - \bar{y}) \\ &= \sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i\end{aligned}\tag{7}$$

and with (3) and (6) as well as invariance (\rightarrow Proof I/1.8.6), scaling (\rightarrow Proof I/1.8.7) and additivity (\rightarrow Proof I/1.8.10) of the variance, the variance of $\hat{\beta}_1$ is:

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \text{Var} \left(\sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i \right) \\
&= \text{Var} \left(\sum_{i=1}^n c_i y_i \right) \\
&= \sum_{i=1}^n c_i^2 \text{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^n c_i^2 \\
&= \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{(n-1) s_x^2} .
\end{aligned} \tag{8}$$

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 Definition I/1.8.2) s_x^2 , we finally get:

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \bar{x}^2) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + 2\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \frac{x^T x}{n} \cdot \frac{\sigma^2}{(n-1)s_x^2}.
\end{aligned} \tag{10}$$

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.
- 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.

Metadata: ID: P273 | shortcut: slr-olsvar | author: JoramSoch | date: 2021-10-27, 11:53.

1.3.7 Distribution of estimates

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

and consider estimation using ordinary least squares (\rightarrow Proof III/1.3.3). Then, the estimated parameters are normally distributed (\rightarrow Definition II/4.1.1) as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \frac{\sigma^2}{(n-1)s_x^2} \cdot \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \tag{2}$$

where \bar{x} is the sample mean (\rightarrow Definition I/1.7.2) and s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x .

Proof: Simple linear regression is a special case of multiple linear regression (\rightarrow Proof III/1.3.2) 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 Proof III/1.4.2) 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 Proof II/4.1.5), 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 Proof II/4.1.5), 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 Definition II/4.1.1) can be further developed as follows:

$$\begin{aligned} \sigma^2 (X^T X)^{-1} &= \sigma^2 \left(\begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \\ &= \sigma^2 \left(\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^T x \end{bmatrix} \right)^{-1} \\ &= \frac{\sigma^2}{nx^T x - (n\bar{x})^2} \begin{bmatrix} x^T x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \frac{\sigma^2}{x^T x - n\bar{x}^2} \begin{bmatrix} x^T x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}. \end{aligned} \quad (8)$$

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.

Metadata: ID: P282 | shortcut: slr-olsdist | author: JoramSoch | date: 2021-11-09, 09:09.

1.3.8 Effects of mean-centering

Theorem: In simple linear regression (\rightarrow Definition III/1.3.1), when the independent variable y and/or the dependent variable x are mean-centered (\rightarrow Definition I/1.7.1), the ordinary least squares (\rightarrow Proof III/1.3.3) 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 Proof III/1.3.3) are

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{1}$$

with sample means (\rightarrow Definition I/1.7.2) \bar{x} and \bar{y} , sample variance (\rightarrow Definition I/1.8.2) s_x^2 and sample covariance (\rightarrow Definition I/1.9.2) s_{xy} , such that β_0 estimates “the mean y at $x = 0$ ”.

2) Let \tilde{x} be the mean-centered covariate vector (\rightarrow Definition III/1.3.1):

$$\tilde{x}_i = x_i - \bar{x} \quad \Rightarrow \quad \bar{\tilde{x}} = 0. \quad (2)$$

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 β_0 now estimates “the mean y at the mean x ”.

3) Let \tilde{y} be the mean-centered data vector (\rightarrow Definition III/1.3.1):

$$\tilde{y}_i = y_i - \bar{y} \quad \Rightarrow \quad \bar{\tilde{y}} = 0. \quad (4)$$

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 β_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 β_0 is now forced to become zero.

Sources:

- original work

Metadata: ID: P274 | shortcut: slr-meancent | author: JoramSoch | date: 2021-10-27, 12:38.

1.3.9 Regression line

Definition: Let there be a simple linear regression with independent observations (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3).

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.

Metadata: ID: D164 | shortcut: regline | author: JoramSoch | date: 2021-10-27, 07:30.

1.3.10 Regression line includes center of mass

Theorem: In simple linear regression (\rightarrow Definition III/1.3.1), the regression line (\rightarrow Definition III/1.3.9) estimated using ordinary least squares (\rightarrow Proof III/1.3.3) includes the point $M(\bar{x}, \bar{y})$.

Proof: The fitted regression line (\rightarrow Definition III/1.3.9) 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 Proof III/1.3.3), 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 Definition III/1.3.9) goes through the center of mass point (\bar{x}, \bar{y}) , if the model (\rightarrow Definition III/1.3.1) includes an intercept term β_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.

Metadata: ID: P275 | shortcut: slr-comp | author: JoramSoch | date: 2021-10-27, 12:52.

1.3.11 Projection of data point to regression line

Theorem: Consider simple linear regression (\rightarrow Definition III/1.3.1) and an estimated regression line (\rightarrow Definition III/1.3.9) 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:

$$P(w \mid \hat{\beta}_0 + \hat{\beta}_1 w) \quad \text{with} \quad w = \frac{x_o + (y_o - \hat{\beta}_0)\hat{\beta}_1}{1 + \hat{\beta}_1^2} . \quad (2)$$

Proof: The intersection point of the regression line (\rightarrow Definition III/1.3.9) 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

$$\begin{aligned}
a^T(b - p) &= 0 \\
a^T(b - w \cdot a) &= 0 \\
a^T b - w \cdot a^T a &= 0 \\
w \cdot a^T a &= a^T b \\
w &= \frac{a^T b}{a^T a} .
\end{aligned} \tag{8}$$

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.

Metadata: ID: P283 | shortcut: slr-proj | author: JoramSoch | date: 2021-11-09, 10:16.

1.3.12 Sums of squares

Theorem: Under ordinary least squares (\rightarrow Proof III/1.3.3) for simple linear regression (\rightarrow Definition III/1.3.1), total (\rightarrow Definition III/1.4.4), explained (\rightarrow Definition III/1.4.5) and residual (\rightarrow Definition III/1.4.6) 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 Definition I/1.8.2) of x and y and s_{xy} is the sample covariance (\rightarrow Definition I/1.9.2) between x and y .

Proof: The ordinary least squares parameter estimates (\rightarrow Proof III/1.3.3) are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \quad (2)$$

1) The total sum of squares (\rightarrow Definition III/1.4.4) is defined as

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

which can be reformulated as follows:

$$\begin{aligned} \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= (n-1) \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= (n-1) s_y^2. \end{aligned} \quad (4)$$

2) The explained sum of squares (\rightarrow Definition III/1.4.5) is defined as

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (5)$$

which, with the OLS parameter estimates, becomes:

$$\begin{aligned} \text{ESS} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &\stackrel{(2)}{=} \sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^n \left(\hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\ &\stackrel{(2)}{=} \sum_{i=1}^n \left(\frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 \\ &= \left(\frac{s_{xy}}{s_x^2} \right)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \left(\frac{s_{xy}}{s_x^2} \right)^2 (n-1) s_x^2 \\ &= (n-1) \frac{s_{xy}^2}{s_x^2}. \end{aligned} \quad (6)$$

3) The residual sum of squares (\rightarrow Definition III/1.4.6) is defined as

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (7)$$

which, with the OLS parameter estimates, becomes:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &\stackrel{(2)}{=} \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \\ &= \sum_{i=1}^n \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\ &= \sum_{i=1}^n \left((y_i - \bar{y})^2 - 2\hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 (x_i - \bar{x})^2 \right) \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= (n-1) s_y^2 - 2(n-1) \hat{\beta}_1 s_{xy} + (n-1) \hat{\beta}_1^2 s_x^2 \\ &\stackrel{(2)}{=} (n-1) s_y^2 - 2(n-1) \left(\frac{s_{xy}}{s_x^2} \right) s_{xy} + (n-1) \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 \\ &= (n-1) s_y^2 - (n-1) \frac{s_{xy}^2}{s_x^2} \\ &= (n-1) \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) . \end{aligned} \quad (8)$$

Sources:

- original work

Metadata: ID: P284 | shortcut: slr-sss | author: JoramSoch | date: 2021-11-09, 11:34.

1.3.13 Transformation matrices

Theorem: Under ordinary least squares (\rightarrow Proof III/1.3.3) for simple linear regression (\rightarrow Definition III/1.3.1), estimation (\rightarrow Definition III/1.4.8), projection (\rightarrow Definition III/1.4.9) and residual-forming (\rightarrow Definition III/1.4.10) matrices are given by

$$\begin{aligned}
E &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) 1_n^T - \bar{x} x^T \\ -\bar{x} 1_n^T + x^T \end{bmatrix} \\
P &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n & \cdots & (x^T x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\
R &= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (n-1)(x^T x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n \\ \vdots & \ddots & \vdots \\ -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n & \cdots & (n-1)(x^T x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix}
\end{aligned} \tag{1}$$

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 Definition I/1.7.2) of x and s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x .

Proof: Simple linear regression is a special case of multiple linear regression (\rightarrow Proof III/1.3.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \tag{2}$$

such that the simple linear regression model can also be written as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{3}$$

Moreover, we note the following equality (\rightarrow Proof III/1.3.7):

$$x^T x - n\bar{x}^2 = (n-1)s_x^2. \tag{4}$$

1) The estimation matrix is given by (\rightarrow Proof III/1.4.11)

$$E = (X^T X)^{-1} X^T \tag{5}$$

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 Proof III/1.4.11)

$$P = X(X^T X)^{-1} X^T = X E \tag{7}$$

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 Proof III/1.4.11)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P \tag{9}$$

which also is an $n \times n$ matrix and can be reformulated as follows:

$$\begin{aligned}
R = I_n - P &= \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \\
&\stackrel{(4)}{=} \frac{1}{(n-1)s_x^2} \begin{bmatrix} x^T x - n\bar{x}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x^T x - n\bar{x}^2 \end{bmatrix} \\
&\quad - \frac{1}{(n-1)s_x^2} \begin{bmatrix} (x^T x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^T x/n) - \bar{x}(x_1 + x_n) + x_1 x_n & \cdots & (x^T x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\
&= \frac{1}{(n-1)s_x^2} \begin{bmatrix} (n-1)(x^T x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n \\ \vdots & \ddots & \vdots \\ -(x^T x/n) + \bar{x}(x_1 + x_n) - x_1 x_n & \cdots & (n-1)(x^T x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix}.
\end{aligned} \tag{10}$$

Sources:

- original work

Metadata: ID: P285 | shortcut: slr-mat | author: JoramSoch | date: 2021-11-09, 15:19.

1.3.14 Weighted least squares

Theorem: Given a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Definition III/1.4.6) 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, W is also symmetric and can be expressed as the matrix square root of the inverse of V :

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2}. \quad (4)$$

Because β_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 Proof II/4.1.5) implies that

$$Wy = \beta_0 W\mathbf{1}_n + \beta_1 Wx + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T). \quad (6)$$

Applying (3), we see that (6) is actually a linear regression model (\rightarrow Definition III/1.4.1) 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 Proof III/1.4.2) 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 2×2 matrix, this reformulates to:

$$\begin{aligned} \hat{\beta} &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} & -\tilde{x}_0^T \tilde{x} \\ -\tilde{x}^T \tilde{x}_0 & \tilde{x}_0^T \tilde{x}_0 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}_0^T \\ \tilde{x}^T \end{bmatrix} \tilde{y} \\ &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} \tilde{x}_0^T - \tilde{x}_0^T \tilde{x} \tilde{x}^T \\ \tilde{x}_0^T \tilde{x}_0 \tilde{x}^T - \tilde{x}^T \tilde{x}_0 \tilde{x}_0^T \end{bmatrix} \tilde{y} \\ &= \frac{1}{\tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{x} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{x}_0} \begin{bmatrix} \tilde{x}^T \tilde{x} \tilde{x}_0^T \tilde{y} - \tilde{x}_0^T \tilde{x} \tilde{x}^T \tilde{y} \\ \tilde{x}_0^T \tilde{x}_0 \tilde{x}^T \tilde{y} - \tilde{x}^T \tilde{x}_0 \tilde{x}_0^T \tilde{y} \end{bmatrix}. \end{aligned} \quad (9)$$

Applying $\tilde{x}_0 = W\mathbf{1}_n$, $\tilde{x} = Wx$ and $W^T W = WW = V^{-1}$, we finally have

$$\begin{aligned}
\hat{\beta} &= \frac{1}{1_n^T W^T W 1_n x^T W^T W x - 1_n^T W^T W x x^T W^T W 1_n} \begin{bmatrix} x^T W^T W x 1_n^T W^T W y - 1_n^T W^T W x x^T W^T W y \\ 1_n^T W^T W 1_n x^T W^T W y - x^T W^T W 1_n 1_n^T W^T W y \end{bmatrix} \\
&= \frac{1}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n} \begin{bmatrix} x^T V^{-1} x 1_n^T V^{-1} y - 1_n^T V^{-1} x x^T V^{-1} y \\ 1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y \end{bmatrix} \\
&= \begin{bmatrix} \frac{x^T V^{-1} x 1_n^T V^{-1} y - 1_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n} \\ \frac{1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y}{1_n^T V^{-1} 1_n x^T V^{-1} x - x^T V^{-1} 1_n 1_n^T V^{-1} x} \end{bmatrix}
\end{aligned} \tag{10}$$

which corresponds to the weighted least squares solution (2).

Sources:

- original work

Metadata: ID: P286 | shortcut: slr-wls | author: JoramSoch | date: 2021-11-16, 07:16.

1.3.15 Weighted least squares

Theorem: Given a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Definition III/1.4.6) are given by

$$\begin{aligned}
\hat{\beta}_0 &= \frac{x^T V^{-1} x 1_n^T V^{-1} y - 1_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n} \\
\hat{\beta}_1 &= \frac{1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y}{1_n^T V^{-1} 1_n x^T V^{-1} x - x^T V^{-1} 1_n 1_n^T V^{-1} x}
\end{aligned} \tag{2}$$

where 1_n is an $n \times 1$ vector of ones.

Proof: Simple linear regression is a special case of multiple linear regression (\rightarrow Proof III/1.3.2) 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 weighted least squares estimates (\rightarrow Proof III/1.4.13) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \tag{4}$$

Writing out equation (4), we have

$$\begin{aligned}
\hat{\beta} &= \left(\begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} y \\
&= \begin{bmatrix} 1_n^T V^{-1} 1_n & 1_n^T V^{-1} x \\ x^T V^{-1} 1_n & x^T V^{-1} x \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T V^{-1} y \\ x^T V^{-1} y \end{bmatrix} \\
&= \frac{1}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n} \begin{bmatrix} x^T V^{-1} x & -1_n^T V^{-1} x \\ -x^T V^{-1} 1_n & 1_n^T V^{-1} 1_n \end{bmatrix} \begin{bmatrix} 1_n^T V^{-1} y \\ x^T V^{-1} y \end{bmatrix} \\
&= \frac{1}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n} \begin{bmatrix} x^T V^{-1} x 1_n^T V^{-1} y - 1_n^T V^{-1} x x^T V^{-1} y \\ 1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y \end{bmatrix}.
\end{aligned} \tag{5}$$

Thus, the first entry of $\hat{\beta}$ is equal to:

$$\hat{\beta}_0 = \frac{x^T V^{-1} x 1_n^T V^{-1} y - 1_n^T V^{-1} x x^T V^{-1} y}{x^T V^{-1} x 1_n^T V^{-1} 1_n - 1_n^T V^{-1} x x^T V^{-1} 1_n}. \tag{6}$$

Moreover, the second entry of $\hat{\beta}$ is equal to (\rightarrow Proof III/1.3.14):

$$\hat{\beta}_1 = \frac{1_n^T V^{-1} 1_n x^T V^{-1} y - x^T V^{-1} 1_n 1_n^T V^{-1} y}{1_n^T V^{-1} 1_n x^T V^{-1} x - x^T V^{-1} 1_n 1_n^T V^{-1} x}. \tag{7}$$

Sources:

- original work

Metadata: ID: P289 | shortcut: slr-wls2 | author: JoramSoch | date: 2021-11-16, 11:20.

1.3.16 Maximum likelihood estimation

Theorem: Given a simple linear regression model (\rightarrow Definition 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 maximum likelihood estimates (\rightarrow Definition I/4.1.3) of β_0 , β_1 and σ^2 are given by

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned} \tag{2}$$

where \bar{x} and \bar{y} are the sample means (\rightarrow Definition I/1.7.2), s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x and s_{xy} is the sample covariance (\rightarrow Definition I/1.9.2) between x and y .

Proof: With the probability density function of the normal distribution (\rightarrow Proof II/3.2.9) and probability under independence (\rightarrow Definition I/1.3.6), the linear regression equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$\begin{aligned}
 p(y|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n p(y_i|\beta_0, \beta_1, \sigma^2) \\
 &= \prod_{i=1}^n \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right] \\
 &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right]
 \end{aligned} \tag{3}$$

and the log-likelihood function (\rightarrow Definition I/4.1.2)

$$\begin{aligned}
 \text{LL}(\beta_0, \beta_1, \sigma^2) &= \log p(y|\beta_0, \beta_1, \sigma^2) \\
 &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.
 \end{aligned} \tag{4}$$

The derivative of the log-likelihood function (4) with respect to β_0 is

$$\frac{d\text{LL}(\beta_0, \beta_1, \sigma^2)}{d\beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \tag{5}$$

and setting this derivative to zero gives the MLE for β_0 :

$$\begin{aligned}
 \frac{d\text{LL}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\beta_0} &= 0 \\
 0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
 0 &= \sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i \\
 \hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i \\
 \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}.
 \end{aligned} \tag{6}$$

The derivative of the log-likelihood function (4) at $\hat{\beta}_0$ with respect to β_1 is

$$\frac{d\text{LL}(\hat{\beta}_0, \beta_1, \sigma^2)}{d\beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \beta_1 x_i^2) \tag{7}$$

and setting this derivative to zero gives the MLE for β_1 :

$$\begin{aligned}
 \frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\beta_1} &= 0 \\
 0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \\
 0 &= \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
 0 &\stackrel{(6)}{=} \sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
 0 &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
 0 &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
 \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
 \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} .
 \end{aligned} \tag{8}$$

The derivative of the log-likelihood function (4) at $(\hat{\beta}_0, \hat{\beta}_1)$ with respect to σ^2 is

$$\frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{9}$$

and setting this derivative to zero gives the MLE for σ^2 :

$$\begin{aligned}
 \frac{dLL(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{d\sigma^2} &= 0 \\
 0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
 \frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
 \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 .
 \end{aligned} \tag{10}$$

Together, (6), (8) and (10) constitute the MLE for simple linear regression.

Sources:

- original work

Metadata: ID: P287 | shortcut: slr-mle | author: JoramSoch | date: 2021-11-16, 08:34.

1.3.17 Maximum likelihood estimation

Theorem: Given a simple linear regression model (\rightarrow Definition III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \quad (1)$$

the maximum likelihood estimates (\rightarrow Definition I/4.1.3) of β_0 , β_1 and σ^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 Definition I/1.7.2), s_x^2 is the sample variance (\rightarrow Definition I/1.8.2) of x and s_{xy} is the sample covariance (\rightarrow Definition I/1.9.2) between x and y .

Proof: Simple linear regression is a special case of multiple linear regression (\rightarrow Proof III/1.3.2) 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 weighted least squares estimates (\rightarrow Proof III/1.4.15) 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 β

$$\begin{aligned} \hat{\beta} &= \left(\begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} V^{-1} y \\ &= \left(\begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^T \\ x^T \end{bmatrix} y \end{aligned} \quad (6)$$

which is equal to the ordinary least squares solution for simple linear regression (\rightarrow Proof III/1.3.4):

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} .\end{aligned}\tag{7}$$

Additionally, we can write out the estimate of σ^2 :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= \frac{1}{n} \left(y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right)^T \left(y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right) \\ &= \frac{1}{n} (y - \hat{\beta}_0 - \hat{\beta}_1 x)^T (y - \hat{\beta}_0 - \hat{\beta}_1 x) \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 .\end{aligned}\tag{8}$$

Sources:

- original work

Metadata: ID: P290 | shortcut: slr-mle2 | author: JoramSoch | date: 2021-11-16, 11:53.

1.3.18 Sum of residuals is zero

Theorem: In simple linear regression (\rightarrow Definition III/1.3.1), the sum of the residuals (\rightarrow Definition III/1.4.6) is zero when estimated using ordinary least squares (\rightarrow Proof III/1.3.3).

Proof: The residuals are defined as the estimated error terms (\rightarrow Definition III/1.3.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{1}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameter estimates obtained using ordinary least squares (\rightarrow Proof III/1.3.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} . \tag{2}$$

With that, we can calculate the sum of the residuals:

$$\begin{aligned}\sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n y_i - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i \\ &= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} \\ &= 0 .\end{aligned}\tag{3}$$

Thus, the sum of the residuals (\rightarrow Definition III/1.4.6) is zero under ordinary least squares (\rightarrow Proof III/1.3.3), if the model (\rightarrow Definition III/1.3.1) includes an intercept term β_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.

Metadata: ID: P276 | shortcut: slr-ressum | author: JoramSoch | date: 2021-10-27, 13:07.

1.3.19 Correlation with covariate is zero

Theorem: In simple linear regression (\rightarrow Definition III/1.3.1), the residuals (\rightarrow Definition III/1.4.6) and the covariate (\rightarrow Definition III/1.3.1) are uncorrelated (\rightarrow Definition I/1.10.1) when estimated using ordinary least squares (\rightarrow Proof III/1.3.3).

Proof: The residuals are defined as the estimated error terms (\rightarrow Definition III/1.3.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (1)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameter estimates obtained using ordinary least squares (\rightarrow Proof III/1.3.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} . \quad (2)$$

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 Definition I/1.10.1), this demonstrates that residuals (\rightarrow Definition III/1.4.6) and covariate (\rightarrow Definition III/1.3.1) values are uncorrelated under ordinary least squares (\rightarrow Proof III/1.3.3).

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.

Metadata: ID: P277 | shortcut: slr-rescorr | author: JoramSoch | date: 2021-10-27, 13:07.

1.3.20 Residual variance in terms of sample variance

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3). Then, residual variance (\rightarrow Definition IV/1.1.1) and sample variance (\rightarrow Definition I/1.8.2) are related to each other via the correlation coefficient (\rightarrow Definition I/1.10.1):

$$\hat{\sigma}^2 = (1 - r_{xy}^2) s_y^2 . \quad (2)$$

Proof: The residual variance (\rightarrow Definition IV/1.1.1) can be expressed in terms of the residual sum of squares (\rightarrow Definition III/1.4.6):

$$\hat{\sigma}^2 = \frac{1}{n-1} \text{RSS}(\hat{\beta}_0, \hat{\beta}_1) \quad (3)$$

and the residual sum of squares for simple linear regression (\rightarrow Proof III/1.3.12) 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 Definition I/1.10.1)

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \quad (6)$$

which also holds for sample correlation, sample covariance (\rightarrow Definition I/1.9.2) and sample standard deviation (\rightarrow Definition I/1.12.1)

$$r_{xy} = \frac{s_{xy}}{s_x s_y} , \quad (7)$$

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.
- 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.

Metadata: ID: P278 | shortcut: slr-resvar | author: JoramSoch | date: 2021-10-27, 14:37.

1.3.21 Correlation coefficient in terms of slope estimate

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3). Then, correlation coefficient (\rightarrow Definition I/1.10.1) and the estimated value of the slope parameter (\rightarrow Definition III/1.3.1) are related to each other via the sample standard deviations (\rightarrow Definition I/1.12.1):

$$r_{xy} = \frac{s_x}{s_y} \hat{\beta}_1 . \quad (2)$$

Proof: The ordinary least squares estimate of the slope (\rightarrow Proof III/1.3.3) is given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} . \quad (3)$$

Using the relationship between covariance and correlation (\rightarrow Proof I/1.9.5)

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y \quad (4)$$

which also holds for sample correlation (\rightarrow Definition I/1.10.1) and sample covariance (\rightarrow Definition I/1.9.2)

$$s_{xy} = s_x r_{xy} s_y , \quad (5)$$

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.
- 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.

Metadata: ID: P279 | shortcut: slr-corr | author: JoramSoch | date: 2021-10-27, 14:58.

1.3.22 Coefficient of determination in terms of correlation coefficient

Theorem: Assume a simple linear regression model (\rightarrow Definition III/1.3.1) 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 Proof III/1.3.3). Then, the coefficient of determination (\rightarrow Definition IV/1.2.1) is equal to the squared correlation coefficient (\rightarrow Definition I/1.10.1) between x and y :

$$R^2 = r_{xy}^2. \quad (2)$$

Proof: The ordinary least squares estimates for simple linear regression (\rightarrow Proof III/1.3.3) 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 Definition IV/1.2.1) 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 Definition III/1.4.5) to total sum of squares (\rightarrow Definition III/1.4.4):

$$R^2 = \frac{\text{ESS}}{\text{TSS}}. \quad (4)$$

Using the explained and total sum of squares for simple linear regression (\rightarrow Proof III/1.3.12), 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:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{\beta}_1 (x_i - \bar{x}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{s_x^2}{s_y^2} \\ &= \left(\frac{s_x}{s_y} \hat{\beta}_1 \right)^2. \end{aligned} \quad (6)$$

Using the relationship between correlation coefficient and slope estimate (\rightarrow Proof III/1.3.21), 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.
- 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.
- Wikipedia (2021): “Correlation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Correlation#Sample_correlation_coefficient.

Metadata: ID: P280 | shortcut: slr-rsq | author: JoramSoch | date: 2021-10-27, 15:31.

1.4 Multiple linear regression

1.4.1 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 Definition II/4.1.1) for ε

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (2)$$

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”;
- β are called “regression coefficients” or “weights”;
- ε is called “noise”, “errors” or “error terms”;
- σ^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 (i.i.d.) observations:

$$V = I_n \Rightarrow \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \Rightarrow \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) . \quad (5)$$

Otherwise, it is called multiple linear regression with correlated observations.

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.

Metadata: ID: D36 | shortcut: mlr | author: JoramSoch | date: 2020-03-21, 20:09.

1.4.2 Ordinary least squares

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) with independent observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) , \quad (1)$$

the parameters minimizing the residual sum of squares (\rightarrow Definition III/1.4.6) 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>.

Metadata: ID: P2 | shortcut: mlr-ols | author: JoramSoch | date: 2019-09-27, 07:18.

1.4.3 Ordinary least squares

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) 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 Definition III/1.4.6) are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad (2)$$

Proof: The residual sum of squares (\rightarrow Definition III/1.4.6) 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 β 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:

$$\begin{aligned} \frac{d\text{RSS}(\hat{\beta})}{d\beta} &= 0 \\ 0 &= -2X^T y + 2X^T X\hat{\beta} \\ X^T X\hat{\beta} &= X^T y \\ \hat{\beta} &= (X^T X)^{-1} X^T y. \end{aligned} \quad (6)$$

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.
- 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/.

Metadata: ID: P40 | shortcut: mlr-ols2 | author: JoramSoch | date: 2020-02-03, 18:43.

1.4.4 Total sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.4.1) 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 total sum of squares (TSS) is defined as the sum of squared deviations of the measured signal from the average signal:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (2)$$

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.

Metadata: ID: D37 | shortcut: tss | author: JoramSoch | date: 2020-03-21, 21:44.

1.4.5 Explained sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.4.1) 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 explained sum of squares (ESS) is defined as the sum of squared deviations of the fitted signal from the average signal:

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2)$$

with estimated regression coefficients $\hat{\beta}$, e.g. obtained via ordinary least squares (\rightarrow Proof III/1.4.2).

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.

Metadata: ID: D38 | shortcut: ess | author: JoramSoch | date: 2020-03-21, 21:57.

1.4.6 Residual sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.4.1) 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 Proof III/1.4.2).

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.

Metadata: ID: D39 | shortcut: rss | author: JoramSoch | date: 2020-03-21, 22:03.

1.4.7 Total, explained and residual sum of squares

Theorem: Assume a linear regression model (\rightarrow Definition III/1.4.1) 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 Definition III/1.4.4), ESS is the explained sum of squares (\rightarrow Definition III/1.4.5) and RSS is the residual sum of squares (\rightarrow Definition III/1.4.6).

Proof: The total sum of squares (\rightarrow Definition III/1.4.4) 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 Proof III/1.4.2), 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 Definition III/1.4.5) and residual sum of squares (\rightarrow Definition III/1.4.6), 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.

Metadata: ID: P76 | shortcut: mlr-pss | author: JoramSoch | date: 2020-03-09, 22:18.

1.4.8 Estimation matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.4.1), the estimation matrix is the matrix E that results in ordinary least squares (\rightarrow Proof III/1.4.2) or weighted least squares (\rightarrow Proof III/1.4.13) parameter estimates when right-multiplied with the measured data:

$$Ey = \hat{\beta} . \quad (1)$$

Sources:

- original work

Metadata: ID: D81 | shortcut: emat | author: JoramSoch | date: 2020-07-22, 05:17.

1.4.9 Projection matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.4.1), the projection matrix is the matrix P that results in the fitted signal explained by estimated parameters (\rightarrow Definition III/1.4.8) 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.

Metadata: ID: D82 | shortcut: pmat | author: JoramSoch | date: 2020-07-22, 05:25.

1.4.10 Residual-forming matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.4.1), the residual-forming matrix is the matrix R that results in the vector of residuals left over by estimated parameters (\rightarrow Definition III/1.4.8) 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.

Metadata: ID: D83 | shortcut: rformat | author: JoramSoch | date: 2020-07-22, 05:35.

1.4.11 Estimation, projection and residual-forming matrix

Theorem: Assume a linear regression model (\rightarrow Definition III/1.4.1) 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 Proof III/1.4.2). Then, the estimated parameters, fitted signal and residuals are given by

$$\begin{aligned}\hat{\beta} &= Ey \\ \hat{y} &= Py \\ \hat{\varepsilon} &= Ry\end{aligned}\tag{2}$$

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 Definition III/1.4.8), projection matrix (\rightarrow Definition III/1.4.9) and residual-forming matrix (\rightarrow Definition III/1.4.10) and n is the number of observations.

Proof:

1) Ordinary least squares parameter estimates of β are defined as minimizing the residual sum of squares (\rightarrow Definition III/1.4.6)

$$\hat{\beta} = \arg \min_{\beta} [(y - X\beta)^T (y - X\beta)]\tag{4}$$

and the solution to this (\rightarrow Proof III/1.4.2) 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>.

Metadata: ID: P75 | shortcut: mlr-mat | author: JoramSoch | date: 2020-03-09, 21:18.

1.4.12 Idempotence of projection and residual-forming matrix

Theorem: The projection matrix (\rightarrow Definition III/1.4.9) and the residual-forming matrix (\rightarrow Definition III/1.4.10) 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 Proof III/1.4.11)

$$P = X(X^T X)^{-1} X^T , \tag{2}$$

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 Proof III/1.4.11)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P , \tag{4}$$

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.

Metadata: ID: P135 | shortcut: mlr-idem | author: JoramSoch | date: 2020-07-22, 06:28.

1.4.13 Weighted least squares

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) 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 Definition III/1.4.6) 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, 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 Proof II/4.1.5) 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 Definition III/1.4.1) 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 Proof III/1.4.2) 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.

Metadata: ID: P77 | shortcut: mlr-wls | author: JoramSoch | date: 2020-03-11, 11:22.

1.4.14 Weighted least squares

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) 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 Definition III/1.4.6) 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, 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 Proof II/4.1.5) 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 Definition III/1.4.1) 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 Definition III/1.4.6) 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 β 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:

$$\begin{aligned}
\frac{dwRSS(\hat{\beta})}{d\beta} &= 0 \\
0 &= -2X^T V^{-1} y + 2X^T V^{-1} X \hat{\beta} \\
X^T V^{-1} X \hat{\beta} &= X^T V^{-1} y \\
\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y .
\end{aligned} \tag{10}$$

Since the quadratic form $y^T V^{-1} y$ in (8) is positive, $\hat{\beta}$ minimizes $wRSS(\beta)$.

Sources:

- original work

Metadata: ID: P136 | shortcut: mlr-wls2 | author: JoramSoch | date: 2020-07-22, 06:48.

1.4.15 Maximum likelihood estimation

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) with correlated observations

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) , \tag{1}$$

the maximum likelihood estimates (\rightarrow Definition I/4.1.3) of β and σ^2 are given by

$$\begin{aligned}
\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\
\hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) .
\end{aligned} \tag{2}$$

Proof: With the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), the linear regression equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$\begin{aligned}
p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\
&= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[-\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right]
\end{aligned} \tag{3}$$

and, using $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (\rightarrow Definition I/4.1.2)

$$\begin{aligned}
LL(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| \\
&\quad - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) .
\end{aligned} \tag{4}$$

Substituting the precision matrix $P = V^{-1}$ into (4) to ease notation, we have:

$$\begin{aligned} \text{LL}(\beta, \sigma^2) = & -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|V|) \\ & - \frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) . \end{aligned} \quad (5)$$

The derivative of the log-likelihood function (5) with respect to β 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 β :

$$\begin{aligned} \frac{d\text{LL}(\hat{\beta}, \sigma^2)}{d\beta} &= 0 \\ 0 &= \frac{1}{\sigma^2} (X^T P y - X^T P X \hat{\beta}) \\ 0 &= X^T P y - X^T P X \hat{\beta} \\ X^T P X \hat{\beta} &= X^T P y \\ \hat{\beta} &= (X^T P X)^{-1} X^T P y \end{aligned} \quad (7)$$

The derivative of the log-likelihood function (4) at $\hat{\beta}$ with respect to σ^2 is

$$\begin{aligned} \frac{d\text{LL}(\hat{\beta}, \sigma^2)}{d\sigma^2} &= \frac{d}{d\sigma^2} \left(-\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \right) \\ &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \end{aligned} \quad (8)$$

and setting this derivative to zero gives the MLE for σ^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.

Sources:

- original work

Metadata: ID: P78 | shortcut: mlr-mle | author: JoramSoch | date: 2020-03-11, 12:27.

1.4.16 Maximum log-likelihood

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) m with correlation structure (\rightarrow Definition I/1.10.5) V

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \tag{1}$$

Then, the maximum log-likelihood (\rightarrow Definition I/4.1.4) for this model is

$$MLL(m) = -\frac{n}{2} \log \left(\frac{RSS}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] \tag{2}$$

under uncorrelated observations (\rightarrow Definition III/1.4.1), i.e. if $V = I_n$, and

$$MLL(m) = -\frac{n}{2} \log \left(\frac{wRSS}{n} \right) - \frac{n}{2} [1 + \log(2\pi)] - \frac{1}{2} \log |V| , \tag{3}$$

in the general case, i.e. if $V \neq I_n$, where RSS is the residual sum of squares (\rightarrow Definition III/1.4.6) and wRSS is the weighted residual sum of squares (\rightarrow Proof III/1.4.14).

Proof: The likelihood function (\rightarrow Definition I/5.1.2) for multiple linear regression is given by (\rightarrow Proof III/1.4.15)

$$\begin{aligned}
p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\
&= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[-\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right] ,
\end{aligned} \tag{4}$$

such that, with $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (\rightarrow Definition I/4.1.2) for this model becomes (\rightarrow Proof III/1.4.15)

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) . \end{aligned} \quad (5)$$

The maximum likelihood estimate for the noise variance (\rightarrow Proof III/1.4.15) 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 Definition III/1.4.6) 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 (??) into (??), we obtain the maximum log-likelihood (\rightarrow Definition 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 (??). 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 (??). 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.

Metadata: ID: P305 | shortcut: mlr-mll | author: JoramSoch | date: 2022-02-04, 07:27.

1.4.17 Deviance function

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) m with correlation structure (\rightarrow Definition I/1.10.5) V

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the deviance (\rightarrow Definition IV/??) for this model is

$$D(\beta, \sigma^2) = \text{RSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] \quad (2)$$

under uncorrelated observations (\rightarrow Definition III/1.4.1), i.e. if $V = I_n$, and

$$D(\beta, \sigma^2) = \text{wRSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] + \log |V| , \quad (3)$$

in the general case, i.e. if $V \neq I_n$, where RSS is the residual sum of squares (\rightarrow Definition III/1.4.6) and wRSS is the weighted residual sum of squares (\rightarrow Proof III/1.4.14).

Proof: The likelihood function (\rightarrow Definition I/5.1.2) for multiple linear regression is given by (\rightarrow Proof III/1.4.15)

$$\begin{aligned} p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp \left[-\frac{1}{2} (y - X\beta)^T (\sigma^2 V)^{-1} (y - X\beta) \right] , \end{aligned} \quad (4)$$

such that, with $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (\rightarrow Definition I/4.1.2) for this model becomes (\rightarrow Proof III/1.4.15)

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) . \end{aligned} \quad (5)$$

The last term can be expressed in terms of the (weighted) residual sum of squares (\rightarrow Definition III/1.4.6) as

$$\begin{aligned} -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) &= -\frac{1}{2\sigma^2} (Wy - WX\beta)^T (Wy - WX\beta) \\ &= -\frac{1}{2\sigma^2} \left(\frac{1}{n} \sum_{i=1}^n (W\varepsilon)_i^2 \right) = -\frac{\text{wRSS}}{2\sigma^2} \end{aligned} \quad (6)$$

where $W = V^{-1/2}$. Plugging (??) into (??) and multiplying with -2 , we obtain the deviance (\rightarrow Definition IV/??) as

$$\begin{aligned} D(\beta, \sigma^2) &= -2 \text{LL}(\beta, \sigma^2) \\ &= -2 \left(-\frac{\text{wRSS}}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |V| \right) \\ &= \text{wRSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] + \log |V| \end{aligned} \quad (7)$$

which proves the result in (??). Assuming $V = I_n$, we have

$$\begin{aligned} -\frac{1}{2\sigma^2}(y - X\beta)^T V^{-1}(y - X\beta) &= -\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta) \\ &= -\frac{1}{2\sigma^2} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) = -\frac{\text{RSS}}{2\sigma^2} \end{aligned} \quad (8)$$

and

$$\frac{1}{2} \log |V| = \frac{1}{2} \log |I_n| = \frac{1}{2} \log 1 = 0, \quad (9)$$

such that

$$D(\beta, \sigma^2) = \text{RSS}/\sigma^2 + n \cdot [\log(\sigma^2) + \log(2\pi)] \quad (10)$$

which proves the result in (??). This completes the proof.

Sources:

- original work

Metadata: ID: P312 | shortcut: mlr-dev | author: JoramSoch | date: 2022-03-01, 08:42.

1.4.18 Akaike information criterion

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) m

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V). \quad (1)$$

Then, the Akaike information criterion (\rightarrow Definition 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 Definition III/1.4.6), p is the number of regressors (\rightarrow Definition III/1.4.1) in the design matrix X and n is the number of observations (\rightarrow Definition III/1.4.1) in the data vector y .

Proof: The Akaike information criterion (\rightarrow Definition 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 Definition I/4.1.4) is k is the number of free parameters in m .

The maximum log-likelihood for multiple linear regression (\rightarrow Proof III/??) 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 paramters in multiple linear regression (\rightarrow Definition III/1.4.1) is $k = p + 1$, i.e. one for each regressor in the design matrix (\rightarrow Definition III/1.4.1) X , plus one for the noise variance (\rightarrow Definition III/1.4.1) σ^2 .

Thus, the AIC of m follows from (??) and (??) 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.

Metadata: ID: P307 | shortcut: mlr-aic | author: JoramSoch | date: 2022-02-11, 06:26.

1.4.19 Bayesian information criterion

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) m

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the Bayesian information criterion (\rightarrow Definition 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 Definition III/1.4.6), p is the number of regressors (\rightarrow Definition III/1.4.1) in the design matrix X and n is the number of observations (\rightarrow Definition III/1.4.1) in the data vector y .

Proof: The Bayesian information criterion (\rightarrow Definition 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 Definition 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 Proof III/??) 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 paramters in multiple linear regression (\rightarrow Definition III/1.4.1) is $k = p+1$, i.e. one for each regressor in the design matrix (\rightarrow Definition III/1.4.1) X , plus one for the noise variance (\rightarrow Definition III/1.4.1) σ^2 .

Thus, the BIC of m follows from (??) and (??) as

$$\text{BIC}(m) = n \log \left(\frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \log(n) (p+1) . \quad (5)$$

Sources:

- original work

Metadata: ID: P308 | shortcut: mlr-bic | author: JoramSoch | date: 2022-02-11, 06:43.

1.4.20 Corrected Akaike information criterion

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) m

$$m : y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) . \quad (1)$$

Then, the corrected Akaike information criterion (\rightarrow Definition IV/??) 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 Definition III/1.4.6), p is the number of regressors (\rightarrow Definition III/1.4.1) in the design matrix X and n is the number of observations (\rightarrow Definition III/1.4.1) in the data vector y .

Proof: The corrected Akaike information criterion (\rightarrow Definition IV/??) 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 Definition 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 Proof III/??) 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 Definition III/1.4.1) is $k = p + 1$, i.e. one for each regressor in the design matrix (\rightarrow Definition III/1.4.1) X , plus one for the noise variance (\rightarrow Definition III/1.4.1) σ^2 .

Thus, the corrected AIC of m follows from (??) and (??) as

$$\begin{aligned} \text{AIC}_c(m) &= n \log \left(\frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + 2k + \frac{2k^2 + 2k}{n - k - 1} \\ &= n \log \left(\frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2nk - 2k^2 - 2k}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\ &= n \log \left(\frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2nk}{n - k - 1} \\ &= n \log \left(\frac{\text{wRSS}}{n} \right) + n [1 + \log(2\pi)] + \log |V| + \frac{2n(p+1)}{n - p - 2} \end{aligned} \quad (5)$$

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.

Metadata: ID: P309 | shortcut: mlr-aicc | author: JoramSoch | date: 2022-02-11, 07:07.

1.5 Bayesian linear regression

1.5.1 Conjugate prior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

be a linear regression model (\rightarrow Definition III/1.4.1) with measured $n \times 1$ 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 β and unknown noise variance σ^2 .

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for this model is a normal-gamma distribution (\rightarrow Definition II/4.3.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (2)$$

where $\tau = 1/\sigma^2$ is the inverse noise variance or noise precision.

Proof: By definition, a conjugate prior (\rightarrow Definition I/5.2.5) is a prior distribution (\rightarrow Definition I/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition 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 Definition I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp \left[-\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right] \quad (3)$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[-\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \quad (4)$$

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 β , 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 Definition I/5.1.2) is proportional to a power of τ , times an exponential of τ and an exponential of a squared form of β , weighted by τ :

$$p(y|\beta, \tau) \propto \tau^{n/2} \cdot \exp \left[-\frac{\tau}{2} (y^T P y - y^T Q y) \right] \cdot \exp \left[-\frac{\tau}{2} (\beta - \tilde{X} y)^T X^T P X (\beta - \tilde{X} y) \right] . \quad (8)$$

The same is true for a normal-gamma distribution (\rightarrow Definition II/4.3.1) over β and τ

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0) \quad (9)$$

the probability density function of which (\rightarrow Proof II/4.3.2)

$$p(\beta, \tau) = \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp \left[-\frac{\tau}{2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \quad (10)$$

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>.

Metadata: ID: P9 | shortcut: blr-prior | author: JoramSoch | date: 2020-01-03, 15:26.

1.5.2 Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

be a linear regression model (\rightarrow Definition III/1.4.1) with measured $n \times 1$ 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 β and unknown noise variance σ^2 . Moreover, assume a normal-gamma prior distribution (\rightarrow Proof III/1.5.1) over the model parameters β 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 Definition I/5.1.7) is also a normal-gamma distribution (\rightarrow Definition II/4.3.1)

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof I/5.3.1), the posterior distribution (\rightarrow Definition I/5.1.7) 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 Proof I/5.1.8) 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 Definition 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 Definition I/1.9.11) $P = V^{-1}$.

Combining the likelihood function (\rightarrow Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) of the model is given by

$$\begin{aligned} p(y, \beta, \tau) &= p(y|\beta, \tau) p(\beta, \tau) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp \left[-\frac{\tau}{2} (y - X\beta)^T P (y - X\beta) \right] \cdot \\ &\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp \left[-\frac{\tau}{2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) \right] \cdot \\ &\quad \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau]. \end{aligned} \quad (9)$$

Collecting identical variables gives:

$$\begin{aligned} p(y, \beta, \tau) &= \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\ &\quad \exp \left[-\frac{\tau}{2} ((y - X\beta)^T P (y - X\beta) + (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0)) \right]. \end{aligned} \quad (10)$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(y, \beta, \tau) &= \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\ &\quad \exp \left[-\frac{\tau}{2} (y^T P y - y^T P X \beta - \beta^T X^T P y + \beta^T X^T P X \beta + \right. \\ &\quad \left. \beta^T \Lambda_0 \beta - \beta^T \Lambda_0 \mu_0 - \mu_0^T \Lambda_0 \beta + \mu_0^T \Lambda_0 \mu_0) \right]. \end{aligned} \quad (11)$$

Completing the square over β , we finally have

$$p(y, \beta, \tau) = \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}}} |P| |\Lambda_0| \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[-\frac{\tau}{2} ((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)) \right] \quad (12)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1} (X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 . \end{aligned} \quad (13)$$

Ergo, the joint likelihood is proportional to

$$p(y, \beta, \tau) \propto \tau^{p/2} \cdot \exp \left[-\frac{\tau}{2} (\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) \right] \cdot \tau^{a_n-1} \cdot \exp [-b_n \tau] \quad (14)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (15)$$

From the term in (14), we can isolate the posterior distribution over β given τ :

$$p(\beta | \tau, y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) . \quad (16)$$

From the remaining term, we can isolate the posterior distribution over τ :

$$p(\tau | y) = \text{Gam}(\tau; a_n, b_n) . \quad (17)$$

Together, (16) and (17) constitute the joint (\rightarrow Definition I/1.3.2) posterior distribution (\rightarrow Definition I/5.1.7) of β and τ .

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>.

Metadata: ID: P10 | shortcut: blr-post | author: JoramSoch | date: 2020-01-03, 17:53.

1.5.3 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 Definition III/1.4.1) with measured $n \times 1$ 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

β and unknown noise variance σ^2 . Moreover, assume a normal-gamma prior distribution (\rightarrow Proof III/1.5.1) over the model parameters β 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 Definition IV/3.1.1) 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 Definition I/5.1.7) 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 Definition I/1.3.3), the model evidence (\rightarrow Definition I/5.1.9) 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 Definition I/1.3.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(y|m) = \iint p(y, \beta, \tau) d\beta d\tau . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition 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 Proof III/1.5.2) $p(\beta, \tau|y)$, the joint likelihood $p(y, \beta, \tau)$ is obtained as

$$\begin{aligned} p(y, \beta, \tau) = & \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\ & \exp \left[-\frac{\tau}{2} ((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)) \right] . \end{aligned} \quad (9)$$

Using the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), 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, β 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 Proof II/3.4.5), 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, τ 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 Definition IV/3.1.1) of this model is given by

$$\log p(y|m) = \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n. \quad (14)$$

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>.

Metadata: ID: P11 | shortcut: blr-lme | author: JoramSoch | date: 2020-01-03, 22:05.

1.5.4 Deviance information criterion

Theorem: Consider a linear regression model (\rightarrow Definition III/1.4.1) 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 Proof III/1.5.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \text{Gam}(\tau; a_0, b_0). \quad (2)$$

Then, the deviance information criterion (\rightarrow Definition 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 μ_n and Λ_n as well as a_n and b_n are posterior parameters (\rightarrow Definition I/5.1.7) describing the posterior distribution in Bayesian linear regression (\rightarrow Proof III/1.5.2).

Proof: The deviance for multiple linear regression (\rightarrow Proof III/??) 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 Definition 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 Definition “mlr-mll”) at the posterior expectations (\rightarrow Definition I/1.7.1) and the “effective number of parameters” p_D is the difference between the expectation of the deviance and the deviance at the expectation (\rightarrow Definition 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 Proof III/1.5.2) 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 Definition I/5.1.7) are given by

$$\begin{aligned} \mu_n &= \Lambda_n^{-1} (X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (10)$$

Thus, we have the following posterior expectations:

$$\langle \beta \rangle_{\beta, \tau | y} = \mu_n \quad (11)$$

$$\langle \tau \rangle_{\beta, \tau|y} = \frac{a_n}{b_n} \quad (12)$$

$$\langle \log \tau \rangle_{\beta, \tau|y} = \psi(a_n) - \log(b_n) \quad (13)$$

$$\begin{aligned} \langle \beta^T A \beta \rangle_{\beta|\tau, y} &= \mu_n^T A \mu_n + \text{tr} (A(\tau \Lambda_n)^{-1}) \\ &= \mu_n^T A \mu_n + \frac{1}{\tau} \text{tr} (A \Lambda_n^{-1}) . \end{aligned} \quad (14)$$

In these identities, we have used the mean of the multivariate normal distribution (\rightarrow Proof “mvn-mean”), the mean of the gamma distribution (\rightarrow Proof II/3.4.8), the logarithmic expectation of the gamma distribution (\rightarrow Proof II/3.4.10), the expectation of a quadratic form (\rightarrow Proof I/1.7.9) and the covariance of the multivariate normal distribution (\rightarrow Proof “mvn-cov”).

With that, the deviance at the expectation is:

$$\begin{aligned} D(\langle \beta \rangle, \langle \tau \rangle) &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot \log(\langle \tau \rangle) - \log |P| + \tau \cdot (y - X \langle \beta \rangle)^T P (y - X \langle \beta \rangle) \\ &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot \log(\langle \tau \rangle) - \log |P| + \tau \cdot (y - X \mu_n)^T P (y - X \mu_n) \\ &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot \log \left(\frac{a_n}{b_n} \right) - \log |P| + \frac{a_n}{b_n} \cdot (y - X \mu_n)^T P (y - X \mu_n) . \end{aligned} \quad (15)$$

Moreover, the expectation of the deviance is:

$$\begin{aligned} \langle D(\beta, \tau) \rangle &\stackrel{(\text{??})}{=} \langle n \cdot \log(2\pi) - n \cdot \log(\tau) - \log |P| + \tau \cdot (y - X \beta)^T P (y - X \beta) \rangle \\ &= n \cdot \log(2\pi) - n \cdot \langle \log(\tau) \rangle - \log |P| + \langle \tau \cdot (y - X \beta)^T P (y - X \beta) \rangle \\ &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\ &\quad + \left\langle \tau \cdot \langle (y - X \beta)^T P (y - X \beta) \rangle_{\beta|\tau, y} \right\rangle_{\tau|y} \\ &= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\ &\quad + \left\langle \tau \cdot \langle y^T P y - y^T P X \beta - \beta^T X^T P y + \beta^T X^T P X \beta \rangle_{\beta|\tau, y} \right\rangle_{\tau|y} \\ &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\ &\quad + \left\langle \tau \cdot \left[y^T P y - y^T P X \mu_n - \mu_n^T X^T P y + \mu_n^T X^T P X \mu_n + \frac{1}{\tau} \text{tr} (X^T P X \Lambda_n^{-1}) \right] \right\rangle_{\tau|y} \\ &= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\ &\quad + \langle \tau \cdot (y - X \mu_n)^T P (y - X \mu_n) \rangle_{\tau|y} + \text{tr} (X^T P X \Lambda_n^{-1}) \\ &\stackrel{(\text{??})}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \\ &\quad + \frac{a_n}{b_n} \cdot (y - X \mu_n)^T P (y - X \mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) . \end{aligned} \quad (16)$$

Finally, combining the two terms, we have:

$$\begin{aligned}
\text{DIC}(m) &\stackrel{(\text{??})}{=} 2 \langle D(\beta, \tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle) \\
&\stackrel{(\text{??})}{=} 2 \left[n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log |P| \right. \\
&\quad \left. + \frac{a_n}{b_n} \cdot (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) \right] \\
&\stackrel{(\text{??})}{=} \left[n \cdot \log(2\pi) - n \cdot \log \left(\frac{a_n}{b_n} \right) - \log |P| + \frac{a_n}{b_n} \cdot (y - X\mu_n)^T P (y - X\mu_n) \right] \quad (17) \\
&= n \cdot \log(2\pi) - 2n\psi(a_n) + 2n \log(b_n) + n \log(a_n) - \log(b_n) - \log |P| \\
&\quad + \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) \\
&= n \cdot \log(2\pi) - n [2\psi(a_n) - \log(a_n) - \log(b_n)] - \log |P| \\
&\quad + \frac{a_n}{b_n} (y - X\mu_n)^T P (y - X\mu_n) + \text{tr} (X^T P X \Lambda_n^{-1}) .
\end{aligned}$$

This conforms to equation (??).

Sources:

- original work

Metadata: ID: P313 | shortcut: blr-dic | author: JoramSoch | date: 2022-03-01, 12:10.

1.5.5 Posterior probability of alternative hypothesis

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.4.1) with normally distributed (\rightarrow Definition II/4.1.1) errors:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

and assume a normal-gamma (\rightarrow Definition II/4.3.1) prior distribution (\rightarrow Definition I/5.1.3) over the model parameters β 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 (\rightarrow Definition I/5.1.7) probability (\rightarrow Definition I/1.3.1) of the alternative hypothesis (\rightarrow Definition I/4.3.3)

$$H_1 : c^T \beta > 0 \quad (3)$$

is given by

$$\Pr(H_1 | y) = 1 - T \left(-\frac{c^T \mu}{\sqrt{c^T \Sigma c}}; \nu \right) \quad (4)$$

where c is a $p \times 1$ contrast vector (\rightarrow Definition “con”), $T(x; \nu)$ is the cumulative distribution function (\rightarrow Definition I/1.6.13) of the t-distribution (\rightarrow Definition II/3.3.1) with ν degrees of freedom (\rightarrow Definition “dof”) and μ , Σ and ν can be obtained from the posterior hyperparameters (\rightarrow Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (\rightarrow Proof III/1.5.2) is given by a normal-gamma distribution (\rightarrow Definition II/4.3.1) over β 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 Definition I/5.1.7)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (6)$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (\rightarrow Proof II/4.3.6), such that the marginal (\rightarrow Definition I/1.5.3) posterior (\rightarrow Definition I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \quad (7)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left(\frac{a_n}{b_n} \Lambda_n \right)^{-1} \\ \nu &= 2 a_n . \end{aligned} \quad (8)$$

Define the quantity $\gamma = c^T \beta$. According to the linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”), γ also follows a multivariate t-distribution (\rightarrow Definition II/4.2.1):

$$p(\gamma|y) = t(\gamma; c^T \mu, c^T \Sigma c, \nu) . \quad (9)$$

Because c^T is a $1 \times p$ vector, γ is a scalar and actually has a non-standardized t-distribution (\rightarrow Definition II/3.3.2). 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 Proof II/3.3.3), 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.

Metadata: ID: P133 | shortcut: blr-pp | author: JoramSoch | date: 2020-07-17, 17:03.

1.5.6 Posterior credibility region excluding null hypothesis

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.4.1) with normally distributed (\rightarrow Definition 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 Definition II/4.3.1) prior distribution (\rightarrow Definition I/5.1.3) over the model parameters β 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 Definition I/5.1.7) credibility region (\rightarrow Definition “cr”) that does not contain the omnibus null hypothesis (\rightarrow Definition I/4.3.2)

$$H_0 : C^T\beta = 0 \tag{3}$$

is given by the credibility level (\rightarrow Definition “cr”)

$$(1 - \alpha) = F([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu) \tag{4}$$

where C is a $p \times q$ contrast matrix (\rightarrow Definition “con”), $F(x; v, w)$ is the cumulative distribution function (\rightarrow Definition I/1.6.13) of the F-distribution (\rightarrow Definition II/3.7.1) with v numerator degrees of freedom (\rightarrow Definition “dof”), w denominator degrees of freedom (\rightarrow Definition “dof”) and μ , Σ and ν can be obtained from the posterior hyperparameters (\rightarrow Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (\rightarrow Proof III/1.5.2) is given by a normal-gamma distribution (\rightarrow Definition II/4.3.1) over β 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 Definition I/5.1.7)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{6}$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (\rightarrow Proof II/4.3.6), such that the marginal (\rightarrow Definition I/1.5.3) posterior (\rightarrow Definition I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \tag{7}$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned}
\mu &= \mu_n \\
\Sigma &= \left(\frac{a_n}{b_n} \Lambda_n \right)^{-1} \\
\nu &= 2 a_n .
\end{aligned} \tag{8}$$

Define the quantity $\gamma = C^T \beta$. According to the linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”), γ also follows a multivariate t-distribution (\rightarrow Definition II/4.2.1):

$$p(\gamma|y) = t(\gamma; C^T \mu, C^T \Sigma C, \nu) . \tag{9}$$

Because C^T is a $q \times p$ matrix, γ is a $q \times 1$ vector. The quadratic form of a multivariate t-distributed random variable has an F-distribution (\rightarrow Proof II/4.2.2), such that we can write:

$$\text{QF}(\gamma) = (\gamma - C^T \mu)^T (C^T \Sigma C)^{-1} (\gamma - C^T \mu) / q \sim \text{F}(q, \nu) . \tag{10}$$

Therefore, the largest posterior credibility region for γ which does not contain $\gamma = 0_q$ (i.e. only touches this origin point) can be obtained by plugging $\text{QF}(0)$ into the cumulative distribution function of the F-distribution:

$$\begin{aligned}
(1 - \alpha) &= \text{F}(\text{QF}(0); q, \nu) \\
&= \text{F}([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu) .
\end{aligned} \tag{11}$$

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, 2.211, 2.183; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P134 | shortcut: blr-pcr | author: JoramSoch | date: 2020-07-17, 17:41.

2 Multivariate normal data

2.1 General linear model

2.1.1 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 Definition 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”;
- Σ 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, Σ 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 (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.

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.

Metadata: ID: D40 | shortcut: glm | author: JoramSoch | date: 2020-03-21, 22:24.

2.1.2 Ordinary least squares

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) with independent observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, \sigma^2 I_n, \Sigma) , \quad (1)$$

the ordinary least squares (\rightarrow Proof III/1.4.2) 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 Proof III/1.4.2) (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 Definition I/1.7.1) 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:

- original work

Metadata: ID: P106 | shortcut: glm-ols | author: JoramSoch | date: 2020-05-19, 06:02.

2.1.3 Weighted least squares

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) with correlated observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) , \quad (1)$$

the weighted least squares (\rightarrow Proof III/1.4.13) 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, W is also symmetric and can be expressed as the matrix square root of the inverse of V :

$$WW = V^{-1} \quad \Leftrightarrow \quad W = V^{-1/2} . \quad (4)$$

Left-multiplying the linear regression equation (1) with W , the linear transformation theorem (\rightarrow Proof II/5.1.5) 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 Definition III/2.1.1) 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 Proof III/2.1.2) 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).

Sources:

- original work

Metadata: ID: P107 | shortcut: glm-wls | author: JoramSoch | date: 2020-05-19, 06:27.

2.1.4 Maximum likelihood estimation

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) with matrix-normally distributed (\rightarrow Definition II/5.1.1) errors

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma), \quad (1)$$

maximum likelihood estimates (\rightarrow Definition I/4.1.3) for the unknown parameters B and Σ 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 Definition I/5.1.2)

$$\begin{aligned} p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \end{aligned} \quad (3)$$

and the log-likelihood function (\rightarrow Definition 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:

$$\begin{aligned}
\text{LL}(B, \Sigma) &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{v}{2} \log(|V|) \\
&\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] .
\end{aligned} \tag{5}$$

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 Σ 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 Σ :

$$\begin{aligned}
\frac{dLL(\hat{B}, \hat{\Sigma})}{d\Sigma} &= 0 \\
0 &= -\frac{n}{2} \hat{\Sigma}^{-1} + \frac{1}{2} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\frac{n}{2} \hat{\Sigma}^{-1} &= \frac{1}{2} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n} \hat{\Sigma}^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B})
\end{aligned} \tag{9}$$

Together, (7) and (9) constitute the MLE for the GLM.

Sources:

- original work

Metadata: ID: P7 | shortcut: glm-mle | author: JoramSoch | date: 2019-12-06, 10:40.

2.2 Transformed general linear model

2.2.1 Definition

Definition: Let there be two general linear models (\rightarrow Definition III/2.1.1) of measured data $Y \in \mathbb{R}^{n \times v}$ using design matrices (\rightarrow Definition III/2.1.1) $X \in \mathbb{R}^{n \times p}$ and $X_t \in \mathbb{R}^{n \times t}$

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

$$Y = X_t \Gamma + E_t, \quad E_t \sim \mathcal{MN}(0, V, \Sigma_t) \tag{2}$$

and assume that X_t can be transformed into X using a transformation matrix $T \in \mathbb{R}^{t \times p}$

$$X = X_t T \tag{3}$$

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 Definition III/2.1.1) 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.

Metadata: ID: D160 | shortcut: tglm | author: JoramSoch | date: 2021-10-21, 14:43.

2.2.2 Derivation of the distribution

Theorem: Let there be two general linear models (\rightarrow Definition III/2.1.1) 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 Definition III/2.2.1) is given by

$$\hat{\Gamma} = TB + H, \quad H \sim \mathcal{MN}(0, U, \Sigma) \quad (4)$$

where the covariance across rows (\rightarrow Definition 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 Proof II/5.1.5) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AU A^T, B^T V B). \quad (5)$$

The weighted least squares parameter estimates (\rightarrow Proof III/2.1.3) 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

$$\begin{aligned} \hat{\Gamma} &\sim \mathcal{MN} \left([(X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1}] XB, [(X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1}] V [V^{-1} X_t (X_t^T V^{-1} X_t)^{-1}], \Sigma \right) \\ &\sim \mathcal{MN} \left((X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} X_t TB, (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} X_t (X_t^T V^{-1} X_t)^{-1}, \Sigma \right) \\ &\sim \mathcal{MN} (TB, (X_t^T V^{-1} X_t)^{-1}, \Sigma). \end{aligned} \quad (8)$$

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.

Metadata: ID: P265 | shortcut: tglm-dist | author: JoramSoch | date: 2021-10-21, 15:03.

2.2.3 Equivalence of parameter estimates

Theorem: Let there be a general linear model (\rightarrow Definition III/2.1.1)

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

and the transformed general linear model (\rightarrow Definition III/2.2.1)

$$\hat{\Gamma} = TB + H, \quad H \sim \mathcal{MN}(0, U, \Sigma) \quad (2)$$

which are linked to each other (\rightarrow Proof III/2.2.2) 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 Proof III/2.1.3) 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 Proof III/2.1.3) 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 Proof III/2.2.2) 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.

Metadata: ID: P266 | shortcut: tglm-para | author: JoramSoch | date: 2021-10-21, 15:25.

2.3 Inverse general linear model

2.3.1 Definition

Definition: Let there be a general linear model (\rightarrow Definition III/2.1.1) of measured data $Y \in \mathbb{R}^{n \times v}$ in terms of the design matrix (\rightarrow Definition III/2.1.1) $X \in \mathbb{R}^{n \times p}$:

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) . \quad (1)$$

Then, a linear model (\rightarrow Definition III/2.1.1) 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.

Metadata: ID: D161 | shortcut: iglm | author: JoramSoch | date: 2021-10-21, 15:31.

2.3.2 Derivation of the distribution

Theorem: Let there be a general linear model (\rightarrow Definition III/2.1.1) 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 Definition III/2.3.1) 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 Definition II/5.1.1) is $\Sigma_x = W^T \Sigma W$.

Proof: The linear transformation theorem for the matrix-normal distribution (\rightarrow Proof II/5.1.5) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^T, B^T V B) . \quad (3)$$

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.

Metadata: ID: P267 | shortcut: iglm-dist | author: JoramSoch | date: 2021-10-21, 16:03.

2.3.3 Best linear unbiased estimator

Theorem: Let there be a general linear model (\rightarrow Definition III/2.1.1) 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 Proof III/2.3.2) 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 Proof III/2.1.3) for W is the best linear unbiased estimator (\rightarrow Definition “blue”) of W .

Proof: The linear transformation theorem for the matrix-normal distribution (\rightarrow Proof II/5.1.5) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^T, B^T V B) . \quad (4)$$

The weighted least squares parameter estimates (\rightarrow Proof III/2.1.3) 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 (\rightarrow Definition “blue”) $\hat{\theta}$ of a certain quantity θ estimated from measured data (\rightarrow Definition “data”) y is 1) an estimator resulting from a linear operation $f(y)$, 2) whose expected value is equal to θ and 3) which has, among those satisfying 1) and 2), the minimum variance (\rightarrow Definition I/1.8.1).

1) First, \hat{W} is a linear estimator, because it is of the form $\tilde{W} = M\hat{X}$ 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

$$\tilde{W} = MX \sim \mathcal{MN}(MYW, MV M^T, \Sigma_x) \quad (6)$$

which requires (\rightarrow Proof “matn-mean”) 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 (\rightarrow Definition “blue”) is the one with minimum variance (\rightarrow Definition I/1.8.1), 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 Definition 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 Definition II/5.2.1)

$$X \sim \mathcal{MN}(0, U, V) \quad \Rightarrow \quad \langle XX^T \rangle = \text{tr}(V) U , \quad (10)$$

this variance (\rightarrow Definition I/1.8.1) 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:

$$\begin{aligned} \text{Var}[\tilde{W}(D)] &\stackrel{(7)}{=} \text{tr}(\Sigma_x) \text{tr} \left[((Y^T V^{-1} Y)^{-1} Y^T V^{-1} + D) V ((Y^T V^{-1} Y)^{-1} Y^T V^{-1} + D)^T \right] \\ &= \text{tr}(\Sigma_x) \text{tr} \left[(Y^T V^{-1} Y)^{-1} Y^T V^{-1} V V^{-1} Y (Y^T V^{-1} Y)^{-1} + \right. \\ &\quad \left. (Y^T V^{-1} Y)^{-1} Y^T V^{-1} V D^T + D V V^{-1} Y (Y^T V^{-1} Y)^{-1} + D V D^T \right] \\ &= \text{tr}(\Sigma_x) [\text{tr}((Y^T V^{-1} Y)^{-1}) + \text{tr}(D V D^T)] . \end{aligned} \quad (12)$$

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.

Metadata: ID: P268 | shortcut: iglm-blue | author: JoramSoch | date: 2021-10-21, 16:46.

2.3.4 Corresponding forward model

Definition: Let there be observations $Y \in \mathbb{R}^{n \times v}$ and $X \in \mathbb{R}^{n \times p}$ and consider a weight matrix $W = f(Y, X) \in \mathbb{R}^{v \times p}$ estimated from Y and X , such that right-multiplying Y with the weight matrix gives an estimate 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, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann 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.

Metadata: ID: D162 | shortcut: cfm | author: JoramSoch | date: 2021-10-21, 17:01.

2.3.5 Derivation of parameters

Theorem: Let there be observations $Y \in \mathbb{R}^{n \times v}$ 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 Definition III/2.3.4) is equal to

$$A = \Sigma_y W \Sigma_x^{-1} \quad (2)$$

with the “sample covariances (\rightarrow Definition I/1.9.2)”

$$\begin{aligned} \Sigma_x &= \hat{X}^T \hat{X} \\ \Sigma_y &= Y^T Y . \end{aligned} \quad (3)$$

Proof: The corresponding forward model (\rightarrow Definition III/2.3.4) is given by

$$Y = \hat{X}A^T + E , \quad (4)$$

subject to the constraint that predicted X and errors E are uncorrelated:

$$\hat{X}^T E = 0 . \quad (5)$$

With that, we can directly derive the parameter matrix A :

$$\begin{aligned}
Y &\stackrel{(4)}{=} \hat{X}A^T + E \\
\hat{X}A^T &= Y - E \\
\hat{X}^T \hat{X}A^T &= \hat{X}^T(Y - E) \\
\hat{X}^T \hat{X}A^T &= \hat{X}^T Y - \hat{X}^T E \\
\hat{X}^T \hat{X}A^T &\stackrel{(5)}{=} \hat{X}^T Y \\
\hat{X}^T \hat{X}A^T &\stackrel{(1)}{=} W^T Y^T Y \\
\Sigma_x A^T &\stackrel{(3)}{=} W^T \Sigma_y \\
A^T &= \Sigma_x^{-1} W^T \Sigma_y \\
A &= \Sigma_y W \Sigma_x^{-1} .
\end{aligned} \tag{6}$$

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, Theorem 1; URL: <https://www.sciencedirect.com/science/article/pii/S1053811913010914>; DOI: 10.1016/j.neuroimage.2013.10.067.

Metadata: ID: P269 | shortcut: cfm-para | author: JoramSoch | date: 2021-10-21, 17:20.

2.3.6 Proof of existence

Theorem: Let there be observations $Y \in \mathbb{R}^{n \times v}$ 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 :

$$\hat{X} = YW . \tag{1}$$

Then, there exists a corresponding forward model (\rightarrow Definition III/2.3.4).

Proof: The corresponding forward model (\rightarrow Definition III/2.3.4) is defined as

$$Y = \hat{X}A^T + E \quad \text{with} \quad \hat{X}^T E = 0 \tag{2}$$

and the parameters of the corresponding forward model (\rightarrow Proof III/2.3.5) are equal to

$$A = \Sigma_y W \Sigma_x^{-1} \quad \text{where} \quad \Sigma_x = \hat{X}^T \hat{X} \quad \text{and} \quad \Sigma_y = Y^T Y . \tag{3}$$

1) Because the columns of \hat{X} are assumed to be linearly independent by definition of the corresponding forward model (\rightarrow Definition III/2.3.4), 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 Definition III/2.3.4) for predicted X and errors E to be uncorrelated which can be shown as follows:

$$\begin{aligned}
\hat{X}^T E &\stackrel{(2)}{=} \hat{X}^T (Y - \hat{X} A^T) \\
&\stackrel{(3)}{=} \hat{X}^T (Y - \hat{X} \Sigma_x^{-1} W^T \Sigma_y) \\
&= \hat{X}^T Y - \hat{X}^T \hat{X} \Sigma_x^{-1} W^T \Sigma_y \\
&\stackrel{(3)}{=} \hat{X}^T Y - \hat{X}^T \hat{X} (\hat{X}^T \hat{X})^{-1} W^T (Y^T Y) \\
&\stackrel{(1)}{=} (Y W)^T Y - W^T (Y^T Y) \\
&= W^T Y^T Y - W^T Y^T Y \\
&= 0.
\end{aligned} \tag{4}$$

This completes the proof.

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, Appendix B; URL: <https://www.sciencedirect.com/science/article/pii/S1053811913010914>; DOI: 10.1016/j.neuroimage.2013.10.067.

Metadata: ID: P270 | shortcut: cfm-exist | author: JoramSoch | date: 2021-10-21, 17:43.

2.4 Multivariate Bayesian linear regression

2.4.1 Conjugate prior distribution

Theorem: Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ .

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for this model is a normal-Wishart distribution (\rightarrow Definition “nw”)

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \tag{2}$$

where $T = \Sigma^{-1}$ is the inverse noise covariance (\rightarrow Definition I/1.9.7) or noise precision matrix (\rightarrow Definition I/1.9.11).

Proof: By definition, a conjugate prior (\rightarrow Definition I/5.2.5) is a prior distribution (\rightarrow Definition I/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition 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 Definition 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

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \quad (4)$$

using the $v \times v$ precision matrix (\rightarrow Definition I/1.9.11) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.9.11) $P = V^{-1}$.

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 β , 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 Definition 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 Definition “nw”) 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 Proof “nw-pdf”)

$$p(B, T) = \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[-\frac{1}{2} \text{tr} (T(B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \quad (10)$$

exhibits the same proportionality

$$p(B, T) \propto |T|^{(\nu_0 + p - v - 1)/2} \cdot \exp \left[-\frac{1}{2} \text{tr}(T \Omega_0) \right] \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T [(B - M_0)^T \Lambda_0 (B - M_0)] \right) \right] \quad (11)$$

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.

Metadata: ID: P159 | shortcut: mblr-prior | author: JoramSoch | date: 2020-09-03, 07:33.

2.4.2 Posterior distribution

Theorem: Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ . Moreover, assume a normal-Wishart prior distribution (\rightarrow Proof III/2.4.1) 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 Definition I/5.1.7) is also a normal-Wishart distribution (\rightarrow Definition “nw”)

$$p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof I/5.3.1), the posterior distribution (\rightarrow Definition I/5.1.7) 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 Proof I/5.1.8) 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 Definition 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 Definition I/1.9.11) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.9.11) $P = V^{-1}$.

Combining the likelihood function (\rightarrow Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) of the model is given by

$$\begin{aligned} p(Y, B, T) &= p(Y|B, T) p(B, T) \\ &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \cdot \\ &\quad \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[-\frac{1}{2} \text{tr} (T(B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \\ &\quad \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] . \end{aligned} \quad (9)$$

Collecting identical variables gives:

$$\begin{aligned} p(Y, B, T) &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \\ &\quad \exp \left[-\frac{1}{2} \text{tr} (T [(Y - XB)^T P(Y - XB) + (B - M_0)^T \Lambda_0 (B - M_0)]) \right] . \end{aligned} \quad (10)$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(Y, B, T) &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \\ &\quad \exp \left[-\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B + \right. \\ &\quad \left. B^T \Lambda_0 B - B^T \Lambda_0 M_0 - M_0^T \Lambda_0 B + M_0^T \Lambda_0 M_0]) \right] . \end{aligned} \quad (11)$$

Completing the square over B , we finally have

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v(\frac{\nu_0}{2})}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr}(\Omega_0 T) \right] \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T \left[(B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n) \right] \right) \right]. \quad (12)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0. \end{aligned} \quad (13)$$

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 Definition I/5.1.7)

$$\begin{aligned} \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n. \end{aligned} \quad (15)$$

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 Definition I/1.3.2) posterior distribution (\rightarrow Definition I/5.1.7) 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.

Metadata: ID: P160 | shortcut: mblr-post | author: JoramSoch | date: 2020-09-03, 08:37.

2.4.3 Log model evidence

Theorem: Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$

regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ . Moreover, assume a normal-Wishart prior distribution (\rightarrow Proof III/2.4.1) 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 Definition IV/3.1.1) 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 Definition I/5.1.7) 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 Definition I/1.3.3), the model evidence (\rightarrow Definition I/5.1.9) 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 Definition I/1.3.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(Y|m) = \iint p(Y, B, T) dB dT . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma| |V|^v}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (7)$$

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 Definition I/1.9.11) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.9.11) $P = V^{-1}$.

When deriving the posterior distribution (\rightarrow Proof III/2.4.2) $p(B, T|Y)$, the joint likelihood $p(Y, B, T)$ is obtained as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \text{tr}(\Omega_0 T)\right] \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[(B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n)]\right)\right]. \quad (9)$$

Using the probability density function of the matrix-normal distribution (\rightarrow Proof II/5.1.2), we can rewrite this as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{(2\pi)^{pv}}{|T|^p |\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \text{tr}(\Omega_0 T)\right] \cdot \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]\right)\right]. \quad (10)$$

Now, B can be integrated out easily:

$$\int p(Y, B, T) dB = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[\Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]\right)\right]. \quad (11)$$

Using the probability density function of the Wishart distribution (\rightarrow Proof “wish-pdf”), we can rewrite this as

$$\int p(Y, B, T) dB = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \sqrt{\frac{2^{\nu_n v}}{|\Omega_n|^{\nu_n}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n). \quad (12)$$

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 Definition IV/3.1.1) of this model is given by

$$\begin{aligned} \log p(y|m) &= \frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ &\quad \frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v\left(\frac{\nu_n}{2}\right) - \log \Gamma_v\left(\frac{\nu_0}{2}\right). \end{aligned} \quad (14)$$

Sources:

- original work

Metadata: ID: P161 | shortcut: mblr-lme | author: JoramSoch | date: 2020-09-03, 09:23.

3 Poisson data

3.1 Poisson-distributed data

3.1.1 Definition

Definition: Poisson-distributed data are defined as a set of observed counts $y = \{y_1, \dots, y_n\}$, independent and identically distributed according to a Poisson distribution (\rightarrow Definition II/1.4.1) with rate λ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

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.

Metadata: ID: D41 | shortcut: poiss-data | author: JoramSoch | date: 2020-03-22, 22:50.

3.1.2 Maximum likelihood estimation

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/3.1.1) set $y = \{y_1, \dots, y_n\}$:

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate (\rightarrow Definition I/4.1.3) for the rate parameter λ is given by

$$\hat{\lambda} = \bar{y} \quad (2)$$

where \bar{y} is the sample mean (\rightarrow Definition I/1.7.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (3)$$

Proof: The likelihood function (\rightarrow Definition I/5.1.2) for each observation is given by the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2)

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \quad (4)$$

and because observations are independent (\rightarrow Definition 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 Definition I/4.1.2) is

$$\text{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \quad (6)$$

which can be developed into

$$\begin{aligned}
\text{LL}(\lambda) &= \sum_{i=1}^n \log \left[\frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\
&= \sum_{i=1}^n [y_i \cdot \log(\lambda) - \lambda - \log(y_i!)] \\
&= -\sum_{i=1}^n \lambda + \sum_{i=1}^n y_i \cdot \log(\lambda) - \sum_{i=1}^n \log(y_i!) \\
&= -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!)
\end{aligned} \tag{7}$$

The derivatives of the log-likelihood with respect to λ are

$$\begin{aligned}
\frac{d\text{LL}(\lambda)}{d\lambda} &= \frac{1}{\lambda} \sum_{i=1}^n y_i - n \\
\frac{d^2\text{LL}(\lambda)}{d\lambda^2} &= -\frac{1}{\lambda^2} \sum_{i=1}^n y_i .
\end{aligned} \tag{8}$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned}
\frac{d\text{LL}(\hat{\lambda})}{d\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^n y_i - n \\
\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} .
\end{aligned} \tag{9}$$

Plugging this value into the second derivative, we confirm:

$$\begin{aligned}
\frac{d^2\text{LL}(\hat{\lambda})}{d\lambda^2} &= -\frac{1}{\bar{y}^2} \sum_{i=1}^n y_i \\
&= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 .
\end{aligned} \tag{10}$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}$ maximizes the likelihood $p(y|\lambda)$.

Sources:

- original work

Metadata: ID: P27 | shortcut: poiss-mle | author: JoramSoch | date: 2020-01-20, 21:53.

3.1.3 Conjugate prior distribution

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/3.1.1) set $y = \{y_1, \dots, y_n\}$:

$$y_i \sim \text{Pois}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for the model parameter λ is a gamma distribution (\rightarrow Definition 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 Proof II/1.4.2), the likelihood function (\rightarrow Definition I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Pois}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp[-\lambda]}{y_i!} \quad (3)$$

and because observations are independent (\rightarrow Definition 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 Definition I/1.7.2) 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 λ times an exponential of λ :

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp[-n\lambda]. \quad (7)$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) \quad (8)$$

the probability density function of which (\rightarrow Proof II/3.4.5)

$$p(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \quad (9)$$

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/>.

Metadata: ID: P225 | shortcut: poiss-prior | author: JoramSoch | date: 2020-04-21, 08:31.

3.1.4 Posterior distribution

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/3.1.1) 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 Proof III/3.1.3) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0). \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a gamma distribution (\rightarrow Definition II/3.4.1)

$$p(\lambda|y) = \text{Gam}(\lambda; a_n, b_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/1.4.2), the likelihood function (\rightarrow Definition 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 Definition 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 Definition I/5.1.5) 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 Definition I/1.7.2) 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 Proof I/5.1.8):

$$p(\lambda|y) \propto p(y, \lambda) . \tag{10}$$

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 Proof II/3.4.5):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] = \text{Gam}(\lambda; a_n, b_n) . \tag{12}$$

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/>.

Metadata: ID: P226 | shortcut: poiss-post | author: JoramSoch | date: 2020-04-21, 08:48.

3.1.5 Log model evidence

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/3.1.1) 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 Proof III/3.1.3) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) . \tag{2}$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\log p(y|m) = - \sum_{i=1}^n \log y_i! + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \tag{3}$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/1.4.2), the likelihood function (\rightarrow Definition 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 Definition 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 Definition I/5.1.5) 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 Definition I/1.7.2) 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 Definition I/5.1.9):

$$p(y) = \int p(y, \lambda) d\lambda . \quad (10)$$

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 Proof II/3.4.5), λ 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 Definition IV/3.1.1) is shown to be

$$\log p(y|m) = - \sum_{i=1}^n \log y_i! + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n. \quad (13)$$

Sources:

- original work

Metadata: ID: P227 | shortcut: poiss-lme | author: JoramSoch | date: 2020-04-21, 09:09.

3.2 Poisson distribution with exposure values

3.2.1 Definition

Definition: A Poisson distribution with exposure values is defined as a set of observed counts $y = \{y_1, \dots, y_n\}$, independently distributed according to a Poisson distribution (\rightarrow Definition II/1.4.1) with common rate λ 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/>.

Metadata: ID: D42 | shortcut: poissexp | author: JoramSoch | date: 2020-03-22, 22:57.

3.2.2 Maximum likelihood estimation

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate (\rightarrow Definition I/4.1.3) for the rate parameter λ is given by

$$\hat{\lambda} = \frac{\bar{y}}{\bar{x}} \quad (2)$$

where \bar{y} and \bar{x} are the sample means (\rightarrow Definition “mean-sample”)

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i . \end{aligned} \quad (3)$$

Proof: With the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2), the likelihood function (\rightarrow Definition I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \quad (4)$$

and because observations are independent (\rightarrow Definition I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} . \quad (5)$$

Thus, the log-likelihood function (\rightarrow Definition I/4.1.2) is

$$\text{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right] \quad (6)$$

which can be developed into

$$\begin{aligned} \text{LL}(\lambda) &= \sum_{i=1}^n \log \left[\frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right] \\ &= \sum_{i=1}^n [y_i \cdot \log(\lambda x_i) - \lambda x_i - \log(y_i!)] \\ &= - \sum_{i=1}^n \lambda x_i + \sum_{i=1}^n y_i \cdot [\log(\lambda) + \log(x_i)] - \sum_{i=1}^n \log(y_i!) \\ &= -\lambda \sum_{i=1}^n x_i + \log(\lambda) \sum_{i=1}^n y_i + \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log(y_i!) \\ &= -n\bar{x}\lambda + n\bar{y} \log(\lambda) + \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log(y_i!) \end{aligned} \quad (7)$$

where \bar{x} and \bar{y} are the sample means from equation (3).

The derivatives of the log-likelihood with respect to λ are

$$\begin{aligned}\frac{dLL(\lambda)}{d\lambda} &= -n\bar{x} + \frac{n\bar{y}}{\lambda} \\ \frac{d^2LL(\lambda)}{d\lambda^2} &= -\frac{n\bar{y}}{\lambda^2} .\end{aligned}\tag{8}$$

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 $p(y|\lambda)$.

Sources:

- original work

Metadata: ID: P224 | shortcut: poissexp-mle | author: JoramSoch | date: 2021-04-16, 11:42.

3.2.3 Conjugate prior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n .\tag{1}$$

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for the model parameter λ is a gamma distribution (\rightarrow Definition 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 Proof II/1.4.2), the likelihood function (\rightarrow Definition 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 Definition I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}. \quad (4)$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \cdot \prod_{i=1}^n \lambda^{y_i} \cdot \prod_{i=1}^n \exp[-\lambda x_i] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{\sum_{i=1}^n y_i} \cdot \exp \left[-\lambda \sum_{i=1}^n x_i \right] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda] \end{aligned} \quad (5)$$

where \bar{y} and \bar{x} are the means (\rightarrow Definition I/1.7.2) of y and x respectively:

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i. \end{aligned} \quad (6)$$

In other words, the likelihood function is proportional to a power of λ times an exponential of λ :

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda]. \quad (7)$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) \quad (8)$$

the probability density function of which (\rightarrow Proof II/3.4.5)

$$p(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \quad (9)$$

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/>.

Metadata: ID: P41 | shortcut: poissexp-prior | author: JoramSoch | date: 2020-02-04, 14:11.

3.2.4 Posterior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/3.2.3) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0). \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a gamma distribution (\rightarrow Definition II/3.4.1)

$$p(\lambda|y) = \text{Gam}(\lambda; a_n, b_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/1.4.2), the likelihood function (\rightarrow Definition 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 Definition 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 Definition I/5.1.5) 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

$$\begin{aligned}
p(y, \lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda x_i] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^n y_i} \exp \left[-\lambda \sum_{i=1}^n x_i \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\bar{x}\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\bar{x})\lambda]
\end{aligned} \tag{8}$$

where \bar{y} and \bar{x} are the means (\rightarrow Definition I/1.7.2) 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 Proof I/5.1.8):

$$p(\lambda|y) \propto p(y, \lambda) . \tag{10}$$

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 Proof II/3.4.5):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] = \text{Gam}(\lambda; a_n, b_n) . \tag{12}$$

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/>.

Metadata: ID: P42 | shortcut: poissexp-post | author: JoramSoch | date: 2020-02-04, 14:42.

3.2.5 Log model evidence

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n . \tag{1}$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/3.2.3) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\ &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} . \end{aligned} \quad (4)$$

Proof: With the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2), the likelihood function (\rightarrow Definition 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 Definition 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 Definition I/5.1.5) 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 Definition I/1.7.2) 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 model evidence is the marginal density of the joint likelihood (\rightarrow Definition I/5.1.9):

$$p(y) = \int p(y, \lambda) d\lambda .\tag{10}$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the joint likelihood can also be written as

$$p(y, \lambda) = \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] .\tag{11}$$

Using the probability density function of the gamma distribution (\rightarrow Proof II/3.4.5), λ 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 Definition IV/3.1.1) is shown to be

$$\begin{aligned}\log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\ &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n .\end{aligned}\tag{13}$$

Sources:

- original work

Metadata: ID: P43 | shortcut: poissexp-lme | author: JoramSoch | date: 2020-02-04, 15:12.

4 Probability 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 according to a Beta distribution (\rightarrow Definition II/3.8.1) with shapes α and β :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

Sources:

- original work

Metadata: ID: D77 | shortcut: beta-data | author: JoramSoch | date: 2020-06-28, 21:16.

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 Definition “iid”) according to a beta distribution (\rightarrow Definition II/3.8.1) with shapes α and β :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

Then, the method-of-moments estimates (\rightarrow Definition I/4.1.5) for the shape parameters α and β 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 Definition I/1.7.2) and \bar{v} is the unbiased sample variance (\rightarrow Definition I/1.8.2):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (3)$$

Proof: Mean (\rightarrow Proof II/3.8.5) and variance (\rightarrow Proof II/3.8.6) of the beta distribution (\rightarrow Definition II/3.8.1) in terms of the parameters α and β are given by

$$\begin{aligned} \text{E}(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (4)$$

Thus, matching the moments (\rightarrow Definition I/4.1.5) requires us to solve the following equation system for α and β :

$$\begin{aligned}\bar{y} &= \frac{\alpha}{\alpha + \beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.\end{aligned}\tag{5}$$

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 α :

$$\begin{aligned}\frac{1 - \bar{y}}{\bar{y}} &= \bar{v} \left[\frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \right] \\ \frac{1 - \bar{y}}{\bar{y} \bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\ \frac{\bar{y}^3(1 - \bar{y})}{\bar{y} \bar{v}} &= \alpha + \bar{y} \\ \alpha &= \frac{\bar{y}^2(1 - \bar{y})}{\bar{v}} - \bar{y} \\ &= \bar{y} \left(\frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right).\end{aligned}\tag{8}$$

Plugging this into equation (6), one obtains for β :

$$\begin{aligned}
\beta &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \cdot \left(\frac{1-\bar{y}}{\bar{y}} \right) \\
&= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) .
\end{aligned} \tag{9}$$

Together, (8) and (9) constitute the method-of-moment estimates of α and β .

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.

Metadata: ID: P28 | shortcut: beta-mome | author: JoramSoch | date: 2020-01-22, 02:53.

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

$$\begin{aligned}
y_i &= [y_{i1}, \dots, y_{ik}], \\
y_{ij} &\in [0, 1] \quad \text{and} \\
\sum_{j=1}^k y_{ij} &= 1
\end{aligned} \tag{1}$$

for all $i = 1, \dots, n$ (and $j = 1, \dots, k$) and each y_i is independent and identically distributed according to a Dirichlet distribution (\rightarrow Definition II/4.4.1) with concentration parameters $\alpha = [\alpha_1, \dots, \alpha_k]$:

$$y_i \sim \text{Dir}(\alpha), \quad i = 1, \dots, n. \tag{2}$$

Sources:

- original work

Metadata: ID: D104 | shortcut: dir-data | author: JoramSoch | date: 2020-10-22, 05:06.

4.2.2 Maximum likelihood estimation

Theorem: Let there be a Dirichlet-distributed data (\rightarrow Definition III/4.2.1) set $y = \{y_1, \dots, y_n\}$:

$$y_i \sim \text{Dir}(\alpha), \quad i = 1, \dots, n. \tag{1}$$

Then, the maximum likelihood estimate (\rightarrow Definition I/4.1.3) for the concentration parameters α can be obtained by iteratively computing

$$\alpha_j^{(\text{new})} = \psi^{-1} \left[\psi \left(\sum_{j=1}^k \alpha_j^{(\text{old})} \right) + \log \bar{y}_j \right] \tag{2}$$

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 Definition I/5.1.2) for each observation is given by the probability density function of the Dirichlet distribution (\rightarrow Proof II/4.4.2)

$$p(y_i|\alpha) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k y_{ij}^{\alpha_j-1} \quad (4)$$

and because observations are independent (\rightarrow Definition 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 Definition 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 α_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:

$$\begin{aligned} \frac{dLL(\alpha)}{d\alpha_j} &= 0 \\ 0 &= n\psi\left(\sum_{j=1}^k \alpha_j\right) - n\psi(\alpha_j) + n \log \bar{y}_j \\ 0 &= \psi\left(\sum_{j=1}^k \alpha_j\right) - \psi(\alpha_j) + \log \bar{y}_j \\ \psi(\alpha_j) &= \psi\left(\sum_{j=1}^k \alpha_j\right) + \log \bar{y}_j \\ \alpha_j &= \psi^{-1}\left[\psi\left(\sum_{j=1}^k \alpha_j\right) + \log \bar{y}_j\right] . \end{aligned} \quad (11)$$

In the following, we will use a fixed-point iteration to maximize $LL(\alpha)$. Given an initial guess for α , we construct a lower bound on the likelihood function (7) which is tight at α . The maximum of this bound is computed and it becomes the new guess. Because the Dirichlet distribution (\rightarrow Definition II/4.4.1) belongs to the exponential family (\rightarrow Definition “dist-expfam”), the log-likelihood function is convex in α 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

$$\begin{aligned} \Gamma(x) &\geq \Gamma(\hat{x}) \cdot \exp[(x - \hat{x}) \psi(\hat{x})] \\ \log \Gamma(x) &\geq \log \Gamma(\hat{x}) + (x - \hat{x}) \psi(\hat{x}) \end{aligned} \quad (12)$$

and apply it to $\Gamma\left(\sum_{j=1}^k \alpha_j\right)$ in (7) to yield

$$\begin{aligned} \frac{1}{n}LL(\alpha) &= \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \\ \frac{1}{n}LL(\alpha) &\geq \log \Gamma\left(\sum_{j=1}^k \hat{\alpha}_j\right) + \left(\sum_{j=1}^k \alpha_j - \sum_{j=1}^k \hat{\alpha}_j\right) \psi\left(\sum_{j=1}^k \hat{\alpha}_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j \\ \frac{1}{n}LL(\alpha) &\geq \left(\sum_{j=1}^k \alpha_j\right) \psi\left(\sum_{j=1}^k \hat{\alpha}_j\right) - \sum_{j=1}^k \log \Gamma(\alpha_j) + \sum_{j=1}^k (\alpha_j - 1) \log \bar{y}_j + \text{const.} \end{aligned} \quad (13)$$

which leads to the following fixed-point iteration using (11):

$$\alpha_j^{(\text{new})} = \psi^{-1} \left[\psi \left(\sum_{j=1}^k \alpha_j^{(\text{old})} \right) + \log \bar{y}_j \right] . \quad (14)$$

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>.

Metadata: ID: P182 | shortcut: dir-mle | author: JoramSoch | date: 2020-10-22, 09:31.

5 Categorical data

5.1 Binomial observations

5.1.1 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.

Sources:

- original work

Metadata: ID: D78 | shortcut: bin-data | author: JoramSoch | date: 2020-07-07, 07:04.

5.1.2 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 Definition II/1.3.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for the model parameter p is a beta distribution (\rightarrow Definition II/3.8.1):

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Proof: With the probability mass function of the binomial distribution (\rightarrow Proof II/1.3.2), the likelihood function (\rightarrow Definition 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 Proof II/3.8.2)

$$p(p) = \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \quad (6)$$

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.

Metadata: ID: P29 | shortcut: bin-prior | author: JoramSoch | date: 2020-01-23, 23:38.

5.1.3 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 Definition II/1.3.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution (\rightarrow Proof III/5.1.2) over the model parameter p :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a beta distribution (\rightarrow Definition II/3.8.1)

$$p(p|y) = \text{Bet}(p; \alpha_n, \beta_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/1.3.2), the likelihood function (\rightarrow Definition 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 Definition I/5.1.5) 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 Proof I/5.1.8):

$$p(p|y) \propto p(y, p) . \quad (7)$$

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 Proof II/3.8.2):

$$p(p|y) = \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} = \text{Bet}(p; \alpha_n, \beta_n) . \quad (9)$$

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.

Metadata: ID: P30 | shortcut: bin-post | author: JoramSoch | date: 2020-01-24, 00:20.

5.1.4 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 Definition II/1.3.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution (\rightarrow Proof III/5.1.2) over the model parameter p :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/1.3.2), the likelihood function (\rightarrow Definition 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 Definition I/5.1.5) 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 Definition I/5.1.9):

$$p(y) = \int p(y, p) \, dp . \quad (7)$$

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 Proof II/3.8.2), 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 Definition IV/3.1.1) (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

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (13)$$

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.

Metadata: ID: P31 | shortcut: bin-lme | author: JoramSoch | date: 2020-01-24, 00:44.

5.2 Multinomial observations

5.2.1 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.

Sources:

- original work

Metadata: ID: D79 | shortcut: mult-data | author: JoramSoch | date: 2020-07-07, 07:12.

5.2.2 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 Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the conjugate prior (\rightarrow Definition I/5.2.5) for the model parameter p is a Dirichlet distribution (\rightarrow Definition II/4.4.1):

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Proof: With the probability mass function of the multinomial distribution (\rightarrow Proof II/2.2.2), the likelihood function (\rightarrow Definition 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 Proof II/4.4.2)

$$p(p) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (6)$$

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

Metadata: ID: P79 | shortcut: mult-prior | author: JoramSoch | date: 2020-03-11, 14:15.

5.2.3 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 Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution (\rightarrow Proof III/5.2.2) over the model parameter p :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a Dirichlet distribution (\rightarrow Definition II/4.4.1)

$$p(p|y) = \text{Dir}(p; \alpha_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/2.2.2), the likelihood function (\rightarrow Definition 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 Definition I/5.1.5) 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 Proof I/5.1.8):

$$p(p|y) \propto p(y, p) . \quad (7)$$

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 Proof II/4.4.2):

$$p(p|y) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} = \text{Dir}(p; \alpha_n) . \quad (9)$$

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

Metadata: ID: P80 | shortcut: mult-post | author: JoramSoch | date: 2020-03-11, 14:40.

5.2.4 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 Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution (\rightarrow Proof III/5.2.2) over the model parameter p :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(k_j+1) \\ &\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) . \end{aligned} \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Proof II/2.2.2), the likelihood function (\rightarrow Definition 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 Definition I/5.1.5) 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 Proof II/4.4.2), 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 Definition IV/3.1.1) (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

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(k_j+1) \\ &\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) . \end{aligned} \quad (13)$$

Sources:

- original work

Metadata: ID: P81 | shortcut: mult-lme | author: JoramSoch | date: 2020-03-11, 15:17.

5.3 Logistic regression

5.3.1 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”;
- β are called “regression coefficients” or “weights”;
- ε_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.

Metadata: ID: D76 | shortcut: logreg | author: JoramSoch | date: 2020-06-28, 20:51.

5.3.2 Probability and log-odds

Theorem: Assume a logistic regression model (\rightarrow Definition III/5.3.1)

$$l_i = x_i \beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

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 Definition I/5.1.2) consistent with (1), $p(y_i)$ are prior probabilities (\rightarrow Definition 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 Proof I/5.3.1) and the law of marginal probability (\rightarrow Definition I/1.3.3), the posterior probabilities (\rightarrow Definition I/5.1.7) for $y_i = 1$ and $y_i = 0$ are given by

$$\begin{aligned} p(y_i = 1|x_i) &= \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)} \\ p(y_i = 0|x_i) &= \frac{p(x_i|y_i = 0) p(y_i = 0)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)}. \end{aligned} \quad (3)$$

Calculating the log-odds from the posterior probabilities, we have

$$\begin{aligned} l_i &= \log_b \frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} \\ &= \log_b \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 0) p(y_i = 0)}. \end{aligned} \quad (4)$$

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%202006.pdf>.

Metadata: ID: P105 | shortcut: logreg-pnlo | author: JoramSoch | date: 2020-05-19, 05:08.

5.3.3 Log-odds and probability

Theorem: Assume a logistic regression model (\rightarrow Definition III/5.3.1)

$$l_i = x_i\beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

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 probability that $y_i = 1$ is given by

$$\Pr(y_i = 1) = \frac{1}{1 + b^{-(x_i\beta + \varepsilon_i)}} \quad (2)$$

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

$$l_i = \log_b \frac{p_i}{1 - p_i} \quad (3)$$

and using (1), we have

$$\begin{aligned} \log_b \frac{p_i}{1 - p_i} &= x_i\beta + \varepsilon_i \\ \frac{p_i}{1 - p_i} &= b^{x_i\beta + \varepsilon_i} \\ p_i &= (b^{x_i\beta + \varepsilon_i}) (1 - p_i) \\ p_i (1 + b^{x_i\beta + \varepsilon_i}) &= b^{x_i\beta + \varepsilon_i} \\ p_i &= \frac{b^{x_i\beta + \varepsilon_i}}{1 + b^{x_i\beta + \varepsilon_i}} \\ p_i &= \frac{b^{x_i\beta + \varepsilon_i}}{b^{x_i\beta + \varepsilon_i} (1 + b^{-(x_i\beta + \varepsilon_i)})} \\ p_i &= \frac{1}{1 + b^{-(x_i\beta + \varepsilon_i)}} \end{aligned} \quad (4)$$

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.

Metadata: ID: P72 | shortcut: logreg-lonp | author: JoramSoch | date: 2020-03-03, 12:01.

Chapter IV

Model Selection

1 Goodness-of-fit measures

1.1 Residual variance

1.1.1 Definition

Definition: Let there be a linear regression model (\rightarrow Definition III/1.4.1)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (1)$$

with measured data y , known design matrix X and covariance structure V as well as unknown regression coefficients β and noise variance σ^2 .

Then, an estimate of the noise variance σ^2 is called the “residual variance” $\hat{\sigma}^2$, e.g. obtained via maximum likelihood estimation (\rightarrow Definition I/4.1.3).

Sources:

- original work

Metadata: ID: D20 | shortcut: resvar | author: JoramSoch | date: 2020-02-25, 11:21.

1.1.2 Maximum likelihood estimator is biased

Theorem: Let $x = \{x_1, \dots, x_n\}$ be a set of independent normally distributed (\rightarrow Definition II/3.2.1) observations with unknown mean (\rightarrow Definition I/1.7.1) μ and variance (\rightarrow Definition I/1.8.1) σ^2 :

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

Then,

1) the maximum likelihood estimator (\rightarrow Definition I/4.1.3) of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

2) and $\hat{\sigma}^2$ is a biased estimator (\rightarrow Definition “est-unb”) of σ^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 Proof III/1.1.2) and a special case of the maximum likelihood estimator for multiple linear regression (\rightarrow Proof III/1.4.15) in which $y = x$, $X = 1_n$ and $\hat{\beta} = \bar{x}$:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}) \\
&= \frac{1}{n}(x - 1_n\bar{x})^T(x - 1_n\bar{x}) \\
&= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 .
\end{aligned} \tag{6}$$

2) The expectation (\rightarrow Definition I/1.7.1) of the maximum likelihood estimator (\rightarrow Definition I/4.1.3) can be developed as follows:

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i\bar{x} + \sum_{i=1}^n \bar{x}^2\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n} \left(\sum_{i=1}^n E[x_i^2] - nE[\bar{x}^2]\right) \\
&= \frac{1}{n} \sum_{i=1}^n E[x_i^2] - E[\bar{x}^2]
\end{aligned} \tag{7}$$

Due to the partition of variance into expected values (\rightarrow Proof I/1.8.3)

$$\text{Var}(X) = E(X^2) - E(X)^2 , \tag{8}$$

we have

$$\begin{aligned}
\text{Var}(x_i) &= E(x_i^2) - E(x_i)^2 \\
\text{Var}(\bar{x}) &= E(\bar{x}^2) - E(\bar{x})^2 ,
\end{aligned} \tag{9}$$

such that (7) becomes

$$E[\hat{\sigma}^2] = \frac{1}{n} \sum_{i=1}^n (\text{Var}(x_i) + E(x_i)^2) - (\text{Var}(\bar{x}) + E(\bar{x})^2) . \tag{10}$$

From (1), it follows that

$$\mathbb{E}(x_i) = \mu \quad \text{and} \quad \text{Var}(x_i) = \sigma^2. \quad (11)$$

The expectation (\rightarrow Definition I/1.7.1) of \bar{x} given by (3) is

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i] \\ &\stackrel{(11)}{=} \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu \\ &= \mu. \end{aligned} \quad (12)$$

The variance of \bar{x} given by (3) is

$$\begin{aligned} \text{Var}[\bar{x}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[x_i] \\ &\stackrel{(11)}{=} \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \frac{1}{n} \sigma^2. \end{aligned} \quad (13)$$

Plugging (11), (12) and (13) into (10), we have

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\ \mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \cdot n \cdot (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\ \mathbb{E}[\hat{\sigma}^2] &= \sigma^2 + \mu^2 - \frac{1}{n} \sigma^2 - \mu^2 \\ \mathbb{E}[\hat{\sigma}^2] &= \frac{n-1}{n} \sigma^2 \end{aligned} \quad (14)$$

which proves the bias (\rightarrow Definition “est-unb”) 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.

Metadata: ID: P61 | shortcut: resvar-bias | author: JoramSoch | date: 2020-02-24, 23:44.

1.1.3 Construction of unbiased estimator

Theorem: Let $x = \{x_1, \dots, x_n\}$ be a set of independent normally distributed (\rightarrow Definition II/3.2.1) observations with unknown mean (\rightarrow Definition I/1.7.1) μ and variance (\rightarrow Definition I/1.8.1) σ^2 :

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

An unbiased estimator (\rightarrow Definition “est-unb”) of σ^2 is given by

$$\hat{\sigma}_{\text{unb}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2)$$

Proof: It can be shown that (\rightarrow Proof IV/1.1.2) the maximum likelihood estimator (\rightarrow Definition I/4.1.3) of σ^2

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3)$$

is a biased estimator (\rightarrow Definition “est-unb”) 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 Definition “est-unb”) can be constructed as

$$\begin{aligned} \hat{\sigma}_{\text{unb}}^2 &= \frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \\ &\stackrel{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (6)$$

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.

Metadata: ID: P62 | shortcut: resvar-unb | author: JoramSoch | date: 2020-02-25, 15:38.

1.2 R-squared

1.2.1 Definition

Definition: Let there be a linear regression model (\rightarrow Definition III/1.4.1) with independent (\rightarrow Definition 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 β and noise variance σ^2 .

Then, the proportion of the variance of the dependent variable y (“total variance (\rightarrow Definition III/1.4.4)”) that can be predicted from the independent variables X (“explained variance (\rightarrow Definition III/1.4.5)”) 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.

Metadata: ID: D21 | shortcut: rsq | author: JoramSoch | date: 2020-02-25, 11:41.

1.2.2 Derivation of R^2 and adjusted R^2

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1)

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1)$$

with n independent observations and p independent variables,

1) the coefficient of determination (\rightarrow Definition IV/1.2.1) 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 Definition III/1.4.6) and total sum of squares (\rightarrow Definition III/1.4.4) 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 Proof III/1.4.2) estimates.

Proof: The coefficient of determination R^2 is defined as (\rightarrow Definition IV/1.2.1) 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 Definition III/1.4.5) 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 (\rightarrow Proof III/1.4.7) $\text{TSS} = \text{ESS} + \text{RSS}$.

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} . \quad (8)$$

If we replace the variance estimates by their unbiased estimators (\rightarrow Proof IV/1.1.3), we obtain

$$R_{\text{adj}}^2 = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{RSS}/\text{df}_r}{\text{TSS}/\text{df}_t} \quad (9)$$

where $\text{df}_r = n - p$ and $\text{df}_t = n - 1$ are the residual and total degrees of freedom (\rightarrow Definition “dof”).

This gives the adjusted R^2 which adjusts R^2 for the number of explanatory variables.

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.

Metadata: ID: P8 | shortcut: rsq-der | author: JoramSoch | date: 2019-12-06, 11:19.

1.2.3 Relationship to maximum log-likelihood

Theorem: Given a linear regression model (\rightarrow Definition III/1.4.1) with independent observations

$$y = X\beta + \varepsilon, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) , \quad (1)$$

the coefficient of determination (\rightarrow Definition IV/1.2.1) can be expressed in terms of the maximum log-likelihood (\rightarrow Definition I/4.1.4) as

$$R^2 = 1 - (\exp[\Delta\text{MLL}])^{-2/n} \quad (2)$$

where n is the number of observations and ΔMLL 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 Definition I/4.1.4) (MLL) of a linear regression model in terms of its residual sum of squares (\rightarrow Definition III/1.4.6) (RSS). The model in (1) implies the following log-likelihood function (\rightarrow Definition 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 Proof III/1.4.15)

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (5)$$

and the residual sum of squares (\rightarrow Definition III/1.4.6) 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 Definition 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 Definition III/1.4.4):

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 . \quad (10)$$

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:

$$\begin{aligned} \exp[\Delta\text{MLL}] &= \exp\left[-\frac{n}{2} \log(\text{RSS}) + \frac{n}{2} \log(\text{TSS})\right] \\ &= (\exp[\log(\text{RSS}) - \log(\text{TSS})])^{-n/2} \\ &= \left(\frac{\exp[\log(\text{RSS})]}{\exp[\log(\text{TSS})]}\right)^{-n/2} \\ &= \left(\frac{\text{RSS}}{\text{TSS}}\right)^{-n/2} . \end{aligned} \quad (12)$$

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.

Sources:

- original work

Metadata: ID: P14 | shortcut: rsq-mll | author: JoramSoch | date: 2020-01-08, 04:46.

1.3 Signal-to-noise ratio

1.3.1 Definition

Definition: Let there be a linear regression model (\rightarrow Definition III/1.4.1) with independent (\rightarrow Definition I/1.3.6) observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with measured data y , known design matrix X as well as unknown regression coefficients β and noise variance σ^2 .

Given estimated regression coefficients (\rightarrow Proof III/1.4.15) $\hat{\beta}$ and residual variance (\rightarrow Definition IV/1.1.1) $\hat{\sigma}^2$, 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.

Metadata: ID: D22 | shortcut: snr | author: JoramSoch | date: 2020-02-25, 12:01.

1.3.2 Relationship with R^2

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.4.1) with independent (\rightarrow Definition I/1.3.6) observations

$$y = X\beta + \varepsilon, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and parameter estimates (\rightarrow Definition “est”) obtained with ordinary least squares (\rightarrow Proof III/1.4.2)

$$\hat{\beta} = (X^T X)^{-1} X^T y . \tag{2}$$

Then, the signal-to noise ratio (\rightarrow Definition IV/1.3.1) can be expressed in terms of the coefficient of determination (\rightarrow Definition IV/1.2.1)

$$\text{SNR} = \frac{R^2}{1 - R^2} \quad (3)$$

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 (SNR) is defined as (\rightarrow Definition IV/1.3.1)

$$\text{SNR} = \frac{\text{Var}(X\hat{\beta})}{\hat{\sigma}^2} = \frac{\text{Var}(\hat{y})}{\hat{\sigma}^2} . \quad (5)$$

Writing out the variances, 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 Proof IV/1.1.2) (dividing by n) or the unbiased estimator for the variance (\rightarrow Proof IV/1.1.3) (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 Definition III/1.4.5), residual (\rightarrow Definition III/1.4.6) and total sum of squares (\rightarrow Definition III/1.4.4):

$$\text{SNR} = \frac{\text{ESS}}{\text{RSS}} = \frac{\text{ESS/TSS}}{\text{RSS/TSS}} . \quad (8)$$

With the derivation of the coefficient of determination (\rightarrow Proof IV/1.2.2), this becomes

$$\text{SNR} = \frac{R^2}{1 - R^2} . \quad (9)$$

Rearranging this equation for the coefficient of determination (\rightarrow Definition IV/1.2.1), we have

$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}} , \quad (10)$$

Sources:

- original work

Metadata: ID: P63 | shortcut: snr-rsq | author: JoramSoch | date: 2020-02-26, 10:37.

2 Classical information criteria

2.1 Akaike information criterion

2.1.1 Definition

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition 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.

Metadata: ID: D23 | shortcut: aic | author: JoramSoch | date: 2020-02-25, 12:31.

2.1.2 Corrected AIC

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the corrected Akaike information criterion (\rightarrow Definition IV/??) (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 Definition IV/2.1.1) and k is the number of free parameters estimated via (??).

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.

Metadata: ID: D171 | shortcut: aicc | author: JoramSoch | date: 2022-02-11, 06:49.

2.1.3 Corrected AIC and uncorrected AIC

Theorem: In the infinite data limit, the corrected Akaike information criterion (\rightarrow Definition IV/??) converges to the uncorrected Akaike information criterion (\rightarrow Definition IV/2.1.1)

$$\lim_{n \rightarrow \infty} \text{AIC}_c(m) = \text{AIC}(m) . \quad (1)$$

Proof: The corrected Akaike information criterion (\rightarrow Definition IV/??) 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.

Metadata: ID: P316 | shortcut: aicc-aic | author: JoramSoch | date: 2022-03-18, 17:00.

2.1.4 Corrected AIC and maximum log-likelihood

Theorem: The corrected Akaike information criterion (\rightarrow Definition IV/??) of a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition 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 Definition I/4.1.4), k is the number of free parameters and n is the number of observations.

Proof: The Akaike information criterion (\rightarrow Definition 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 Definition IV/??) is defined as

$$\text{AIC}_c(m) = \text{AIC}(m) + \frac{2k^2 + 2k}{n - k - 1} . \quad (3)$$

Plugging (??) into (??), we obtain:

$$\begin{aligned}
\text{AIC}_c(m) &= -2 \log p(y|\hat{\theta}, m) + 2k + \frac{2k^2 + 2k}{n - k - 1} \\
&= -2 \log p(y|\hat{\theta}, m) + \frac{2k(n - k - 1)}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\
&= -2 \log p(y|\hat{\theta}, m) + \frac{2nk - 2k^2 - 2k}{n - k - 1} + \frac{2k^2 + 2k}{n - k - 1} \\
&= -2 \log p(y|\hat{\theta}, m) + \frac{2nk}{n - k - 1} .
\end{aligned} \tag{4}$$

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.

Metadata: ID: P315 | shortcut: aicc-mll | author: JoramSoch | date: 2022-03-11, 16:53.

2.2 Bayesian information criterion

2.2.1 Definition

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \tag{1}$$

Then, the Bayesian information criterion (BIC) of this model is defined as

$$\text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + k \log n \tag{2}$$

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>.

Metadata: ID: D24 | shortcut: bic | author: JoramSoch | date: 2020-02-25, 12:21.

2.2.2 Derivation

Theorem: Let $p(y|\theta, m)$ be the likelihood function (\rightarrow Definition I/5.1.2) of a generative model (\rightarrow Definition I/5.1.1) $m \in \mathcal{M}$ with model parameters $\theta \in \Theta$ describing measured data $y \in \mathbb{R}^n$. Let $p(\theta|m)$ be a prior distribution (\rightarrow Definition 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 goes to infinity, an approximation to the log-marginal likelihood (\rightarrow Definition I/5.1.9) $\log p(y|m)$, up to constant terms not depending on the model, is given by the Bayesian information criterion (\rightarrow Definition 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 Definition I/4.1.3) (MLE) of θ , 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 Definition 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 Definition I/5.1.9) 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 θ_0 is the value that maximizes $h(\theta)$ and $J(\theta_0)$ is the Hessian matrix evaluated at θ_0 . In our case, we have $h(\theta) = 1/n \text{LL}(\theta)$ such that θ_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 Definition IV/3.5.1) $p(m_j|y)$. 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>; DOI: 10.1017/CBO9780511790485.

Metadata: ID: P32 | shortcut: bic-der | author: JoramSoch | date: 2020-01-26, 23:36.

2.3 Deviance information criterion

2.3.1 Definition

Definition: Let m be a full probability model (\rightarrow Definition I/5.1.4) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Together, likelihood function and prior distribution imply a posterior distribution (\rightarrow Proof I/5.1.8) $p(\theta|y, m)$. Consider the deviance (\rightarrow Definition IV/??) which is minus two times the log-likelihood function (\rightarrow Definition 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 Definition I/4.1.2) at the posterior (\rightarrow Definition I/5.1.7) expectation (\rightarrow Definition I/1.7.1) and the “effective number of parameters” p_D is the difference between the expectation of the deviance and the deviance at the expectation (\rightarrow Definition 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 Definition I/1.7.1) across the posterior distribution (\rightarrow Definition I/5.1.7).

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.

Metadata: ID: D25 | shortcut: dic | author: JoramSoch | date: 2020-02-25, 12:46.

2.3.2 Deviance

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the deviance of m is a function of θ which multiplies the log-likelihood function (\rightarrow Definition 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 Definition 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.

Metadata: ID: D172 | shortcut: dev | author: JoramSoch | date: 2022-03-01, 07:48.

3 Bayesian model selection

3.1 Log model evidence

3.1.1 Definition

Definition: Let m be a full probability model (\rightarrow Definition I/5.1.4) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Then, the log model evidence (LME) of this model is defined as the logarithm of the marginal likelihood (\rightarrow Definition I/5.1.9):

$$\text{LME}(m) = \log p(y|m) . \quad (1)$$

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.

Metadata: ID: D26 | shortcut: lme | author: JoramSoch | date: 2020-02-25, 12:56.

3.1.2 Derivation

Theorem: Let $p(y|\theta, m)$ be a likelihood function (\rightarrow Definition I/5.1.2) of a generative model (\rightarrow Definition I/5.1.1) m for making inferences on model parameters θ given measured data y . Moreover, let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) on model parameters θ . Then, the log model evidence (\rightarrow Definition IV/3.1.1) (LME), also called marginal log-likelihood,

$$\text{LME}(m) = \log p(y|m) , \quad (1)$$

can be expressed in terms of likelihood (\rightarrow Definition I/5.1.2) and prior (\rightarrow Definition I/5.1.3) as

$$\text{LME}(m) = \log \int p(y|\theta, m) p(\theta|m) d\theta . \quad (2)$$

Proof: This a consequence of the law of marginal probability (\rightarrow Definition I/1.3.3) for continuous variables

$$p(y|m) = \int p(y, \theta|m) d\theta \quad (3)$$

and the law of conditional probability (\rightarrow Definition I/1.3.4) according to which

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (4)$$

Combining (??) with (??) and logarithmizing, we have:

$$\text{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m) p(\theta|m) d\theta . \quad (5)$$

Sources:

- original work

Metadata: ID: P13 | shortcut: lme-der | author: JoramSoch | date: 2020-01-06, 21:27.

3.1.3 Expression using prior and posterior

Theorem: Let $p(y|\theta, m)$ be a likelihood function (\rightarrow Definition I/5.1.2) of a generative model (\rightarrow Definition I/5.1.1) m for making inferences on model parameters θ given measured data y . Moreover, let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) on model parameters θ . Then, the log model evidence (\rightarrow Definition IV/3.1.1) (LME), also called marginal log-likelihood,

$$\text{LME}(m) = \log p(y|m) , \quad (1)$$

can be expressed in terms of prior (\rightarrow Definition I/5.1.3) and posterior (\rightarrow Definition I/5.1.7) 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 Definition I/5.1.4), Bayes' theorem (\rightarrow Proof I/5.3.1) makes a statement about the posterior distribution (\rightarrow Definition I/5.1.7):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (3)$$

Rearranging for $p(y|m)$ and logarithmizing, we have:

$$\begin{aligned} \text{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta, m) p(\theta|m)}{p(\theta|y, m)} \\ &= \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) . \end{aligned} \quad (4)$$

Sources:

- original work

Metadata: ID: P314 | shortcut: lme-pnp | author: JoramSoch | date: 2022-03-11, 16:25.

3.1.4 Partition into accuracy and complexity

Theorem: The log model evidence (\rightarrow Definition IV/3.1.1) 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 Definition I/5.1.7) expectation (\rightarrow Proof I/1.7.11) of the log-likelihood function (\rightarrow Definition 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 Definition I/2.5.1) of posterior (\rightarrow Definition I/5.1.7) from prior (\rightarrow Definition 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 Definition “data”) y using model (\rightarrow Definition I/5.1.1) m with parameters θ . Then, Bayes’ theorem (\rightarrow Proof I/5.3.1) makes a statement about the posterior distribution (\rightarrow Definition I/5.1.7), 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 Proof IV/??), 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:

- 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.
- 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.

Metadata: ID: P3 | shortcut: lme-anc | author: JoramSoch | date: 2019-09-27, 16:13.

3.1.5 Uniform-prior log model evidence

Definition: Assume a generative model (\rightarrow Definition I/5.1.1) m with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and a uniform (\rightarrow Definition I/5.2.2) prior distribution (\rightarrow Definition I/5.1.3) $p_{\text{uni}}(\theta|m)$. Then, the log model evidence (\rightarrow Definition IV/3.1.1) of this model is called “log model evidence with uniform prior” or “uniform-prior log model evidence” (upLME):

$$\text{upLME}(m) = \log \int p(y|\theta, m) p_{\text{uni}}(\theta|m) d\theta . \quad (1)$$

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.

Metadata: ID: D113 | shortcut: uplme | author: JoramSoch | date: 2020-11-25, 07:28.

3.1.6 Cross-validated log model evidence

Definition: Let there be a data set (\rightarrow Definition “data”) y with mutually exclusive and collectively exhaustive subsets y_1, \dots, y_S . Assume a generative model (\rightarrow Definition I/5.1.1) m with model parameters θ implying a likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and a non-informative (\rightarrow Definition I/5.2.3) prior density (\rightarrow Definition I/5.1.3) $p_{\text{ni}}(\theta|m)$.

Then, the cross-validated log model evidence 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 Definition I/5.1.7) obtained from y_{-i} when using the prior distribution (\rightarrow Definition I/5.1.3) $p_{\text{ni}}(\theta|m)$:

$$p(\theta|y_{-i}, m) = \frac{p(y_{-i}|\theta, m) p_{\text{ni}}(\theta|m)}{p(y_{-i}|m)}. \quad (2)$$

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*, arXiv:1807.01585; URL: <https://arxiv.org/abs/1807.01585>.

Metadata: ID: D111 | shortcut: cvlme | author: JoramSoch | date: 2020-11-19, 04:55.

3.1.7 Empirical Bayesian log model evidence

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with model parameters θ and hyper-parameters λ implying the likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, \lambda, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|\lambda, m)$. Then, the Empirical Bayesian (\rightarrow Definition I/5.3.3) log model evidence (\rightarrow Definition IV/3.1.1) is the logarithm of the marginal likelihood (\rightarrow Definition I/5.1.9), maximized with respect to the hyper-parameters:

$$\text{ebLME}(m) = \log p(y|\hat{\lambda}, m) \quad (1)$$

where

$$p(y|\lambda, m) = \int p(y|\theta, \lambda, m) (\theta|\lambda, m) d\theta \quad (2)$$

and (\rightarrow Definition 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.
- 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.

Metadata: ID: D114 | shortcut: eblme | author: JoramSoch | date: 2020-11-25, 07:43.

3.1.8 Variational Bayesian log model evidence

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with model parameters θ implying the likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Moreover, assume an approximate (\rightarrow Definition I/5.3.4) posterior distribution (\rightarrow Definition I/5.1.7) $q(\theta)$. Then, the Variational Bayesian (\rightarrow Definition I/5.3.4) log model evidence (\rightarrow Definition IV/3.1.1), also referred to as the “negative free energy”, is the expectation of the log-likelihood function (\rightarrow Definition I/4.1.2) with respect to the approximate posterior, minus the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) between approximate posterior and the prior distribution:

$$\text{vbLME}(m) = \langle \log p(y|\theta, m) \rangle_{q(\theta)} - \text{KL} [q(\theta)||p(\theta|m)] \quad (1)$$

where

$$\langle \log p(y|\theta, m) \rangle_{q(\theta)} = \int q(\theta) \log p(y|\theta, m) d\theta \quad (2)$$

and

$$\text{KL} [q(\theta)||p(\theta|m)] = \int q(\theta) \log \frac{q(\theta)}{p(\theta|m)} d\theta . \quad (3)$$

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.
- 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.

Metadata: ID: D115 | shortcut: vblme | author: JoramSoch | date: 2020-11-25, 08:10.

3.2 Log family evidence

3.2.1 Definition

Definition: Let f be a family of M generative models (\rightarrow Definition 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 of f is the weighted average of the model evidences (\rightarrow Definition I/5.1.9) of m_1, \dots, m_M where the weights are the within-family prior model probabilities (\rightarrow Definition I/5.1.3)

$$p(y|f) = \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (2)$$

The log family evidence is given by the logarithm of the family evidence:

$$\text{LFE}(f) = \log p(y|f) = \log \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (3)$$

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. 16; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D80 | shortcut: lfe | author: JoramSoch | date: 2020-07-13, 22:31.

3.2.2 Derivation

Theorem: Let f be a family of M generative models (\rightarrow Definition I/5.1.1) m_1, \dots, m_M with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$. Then, the log family evidence (\rightarrow Definition IV/3.2.1)

$$\text{LFE}(f) = \log p(y|f) \quad (1)$$

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 Definition IV/3.2.1) prior (\rightarrow Definition I/5.1.3) model (\rightarrow Definition I/5.1.1) probabilities (\rightarrow Definition 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 Proof I/5.3.1) for the family evidence (\rightarrow Definition IV/3.2.1) 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 Proof I/5.3.1) for the model evidence (\rightarrow Definition IV/3.2.1) 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

$$\begin{aligned} p(y|f) &= \frac{\sum_{i=1}^M p(y|m_i) p(m_i)}{\sum_{i=1}^M p(m_i)} \\ &= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i)}{\sum_{i=1}^M p(m_i)} \\ &= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i, f)}{p(f)} \\ &= \sum_{i=1}^M p(y|m_i) \cdot p(m_i|f) . \end{aligned} \quad (9)$$

Equation (2) follows by logarithmizing both sides of (9).

Sources:

- original work

Metadata: ID: P132 | shortcut: lfe-der | author: JoramSoch | date: 2020-07-13, 22:58.

3.2.3 Calculation from log model evidences

Theorem: Let m_1, \dots, m_M be M statistical models with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$ and belonging to F mutually exclusive model families f_1, \dots, f_F . Then, the log family evidences (\rightarrow Definition IV/3.2.1) are given by:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)] , \quad j = 1, \dots, F, \quad (1)$$

where $p(m_i|f_j)$ are within-family (\rightarrow Definition IV/3.2.1) prior (\rightarrow Definition I/5.1.3) model (\rightarrow Definition I/5.1.1) probabilities (\rightarrow Definition I/1.3.1).

Proof: Let us consider the (unlogarithmized) family evidence $p(y|f_j)$. According to the law of marginal probability (\rightarrow Definition I/1.3.3), this conditional probability is given by

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i, f_j) \cdot p(m_i|f_j)] . \quad (2)$$

Because model families are mutually exclusive, it holds that $p(y|m_i, f_j) = p(y|m_i)$, such that

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \quad (3)$$

Logarithmizing transforms the family evidence $p(y|f_j)$ into the log family evidence $\text{LFE}(f_j)$:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \quad (4)$$

The definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (5)$$

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.

Metadata: ID: P65 | shortcut: lfe-lme | author: JoramSoch | date: 2020-02-27, 21:16.

3.3 Log Bayes factor

3.3.1 Definition

Definition: Let there be two generative models (\rightarrow Definition 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 Definition I/5.1.9) 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.

Metadata: ID: D84 | shortcut: lbf | author: JoramSoch | date: 2020-07-22, 07:02.

3.3.2 Derivation

Theorem: Let there be two generative models (\rightarrow Definition I/5.1.1) m_1 and m_2 with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1)$ and $p(y|m_2)$. Then, the log Bayes factor (\rightarrow Definition IV/3.3.1)

$$\text{LBF}_{12} = \log \text{BF}_{12} \quad (1)$$

can be expressed as

$$\text{LBF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)}. \quad (2)$$

Proof: The Bayes factor (\rightarrow Definition IV/3.4.1) is defined as the posterior (\rightarrow Definition I/5.1.7) odds ratio (\rightarrow Definition “odds”) when both models (\rightarrow Definition I/5.1.1) are equally likely apriori (\rightarrow Definition 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 Proof I/5.3.2), 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 Definition I/5.1.3) odds ratio (\rightarrow Definition “odds”) 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).

Sources:

- original work

Metadata: ID: P137 | shortcut: lbf-der | author: JoramSoch | date: 2020-07-22, 07:27.

3.3.3 Calculation from log model evidences

Theorem: Let m_1 and m_2 be two statistical models with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1)$ and $\text{LME}(m_2)$. Then, the log Bayes factor (\rightarrow Definition IV/3.3.1) 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 Definition IV/3.4.1) is defined as the ratio of the model evidences (\rightarrow Definition I/5.1.9) 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 Definition IV/3.3.1) is defined as the logarithm of the Bayes factor

$$\text{LBF}_{12} = \log \text{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (3)$$

With the definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (4)$$

the log Bayes factor can be expressed as:

Resolving the logarithm and applying the definition of the log model evidence (\rightarrow Definition IV/3.1.1), we finally have:

$$\begin{aligned} \text{LBF}_{12} &= \log p(y|m_1) - \log p(y|m_2) \\ &= \text{LME}(m_1) - \text{LME}(m_2) . \end{aligned} \quad (5)$$

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. 18; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P64 | shortcut: lbf-lme | author: JoramSoch | date: 2020-02-27, 20:51.

3.4 Bayes factor

3.4.1 Definition

Definition: Consider two competing generative models (\rightarrow Definition 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 Definition I/5.1.9) of m_1 and m_2 :

$$\text{BF}_{12} = \frac{p(y | m_1)}{p(y | m_2)} . \quad (1)$$

Note that by Bayes' theorem (\rightarrow Proof I/5.3.1), the ratio of posterior model probabilities (\rightarrow Definition IV/3.5.1) (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 Proof I/5.3.2).

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.

Metadata: ID: D92 | shortcut: bf | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.4.2 Transitivity

Theorem: Consider three competing models (\rightarrow Definition I/5.1.1) m_1 , m_2 , and m_3 for observed data y . Then the Bayes factor (\rightarrow Definition IV/3.4.1) for m_1 over m_3 can be written as:

$$\text{BF}_{13} = \text{BF}_{12} \cdot \text{BF}_{23}. \quad (1)$$

Proof: By definition (\rightarrow Definition IV/3.4.1), the Bayes factor BF_{13} is the ratio of marginal likelihoods of data y over 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).

Sources:

- original work

Metadata: ID: P163 | shortcut: bf-trans | author: tomfaulkenberry | date: 2020-09-07, 12:00.

3.4.3 Computation using Savage-Dickey Density Ratio

Theorem: Consider two competing models (\rightarrow Definition I/5.1.1) on data y containing parameters δ and φ , namely $m_0 : \delta = \delta_0, \varphi$ and $m_1 : \delta, \varphi$. In this context, we say that δ is a parameter of interest, φ 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 φ in m_0 is equal to the prior for φ 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 Definition IV/3.4.1) 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 Definition IV/3.4.1), the Bayes factor 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 BF_{01} entirely in terms of the “encompassing” model m_1 . This proceeds by first unpacking the marginal likelihood (\rightarrow Definition I/5.1.9) for m_0 over the nuisance parameter φ 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 Proof I/5.3.1), 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.

Metadata: ID: P156 | shortcut: bf-sddr | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.4.4 Computation using Encompassing Prior Method

Theorem: Consider two models m_1 and m_e , where m_1 is nested within an encompassing model (\rightarrow Definition IV/3.4.5) m_e via an inequality constraint on some parameter θ , and θ is unconstrained under m_e . Then, the Bayes factor (\rightarrow Definition IV/3.4.1) 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 θ , Bayes’ theorem (\rightarrow Proof I/5.3.1) 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 Definition I/5.1.9) 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 Definition IV/3.4.5) m_e yields the following Bayes factor (\rightarrow Definition IV/3.4.1):

$$\text{BF}_{1e} = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1) / p(\theta \mid y, m_1)}{p(y \mid \theta, m_e) \cdot p(\theta \mid m_e) / p(\theta \mid y, m_e)}. \quad (4)$$

Now, both the constrained model m_1 and the encompassing model (\rightarrow Definition IV/3.4.5) m_e contain the same parameter vector θ . Choose a specific value of θ , say θ' , 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 θ' , 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 θ' 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 $p(\theta' | m_1)$ is simply a truncation of the encompassing prior $p(\theta' | m_e)$. Thus, we can express $p(\theta' | m_1)$ in terms of the encompassing prior $p(\theta' | m_e)$ by multiplying the encompassing prior by an indicator function over m_1 and then normalizing the resulting product. That is,

$$\begin{aligned} p(\theta' | m_1) &= \frac{p(\theta' | m_e) \cdot I_{\theta' \in m_1}}{\int p(\theta' | m_e) \cdot I_{\theta' \in m_1} d\theta'} \\ &= \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' | m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' | m_e), \end{aligned} \quad (6)$$

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

$$p(\theta' | m_1) = c \cdot p(\theta' | m_e). \quad (7)$$

By similar reasoning, we can write the posterior as

$$p(\theta' | y, m_1) = \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' | y, m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' | y, m_e) = d \cdot p(\theta' | y, m_e). \quad (8)$$

Plugging (7) and (8) into (5), this gives us

$$\text{BF}_{1e} = \frac{c \cdot p(\theta' | m_e) / d \cdot p(\theta' | y, m_e)}{p(\theta' | m_e) / p(\theta' | y, m_e)} = \frac{c}{d} = \frac{1/d}{1/c}, \quad (9)$$

which completes the proof. Note that by definition, $1/d$ represents the proportion of the posterior distribution for θ under the encompassing model (\rightarrow Definition IV/3.4.5) m_e that agrees with the constraints imposed by m_1 . Similarly, $1/c$ represents the proportion of the prior distribution for θ under the encompassing model (\rightarrow Definition IV/3.4.5) 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.

Metadata: ID: P157 | shortcut: bf-ep | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.4.5 Encompassing model

Definition: Consider a family f of generative models (\rightarrow Definition I/5.1.1) m on data y , where each $m \in f$ is defined by placing an inequality constraint on model parameter(s) θ (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) θ 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.

Metadata: ID: D93 | shortcut: encm | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.5 Posterior model probability

3.5.1 Definition

Definition: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$ and prior probabilities (\rightarrow Definition I/5.1.3) $p(m_1), \dots, p(m_M)$. Then, the conditional probability (\rightarrow Definition I/1.3.4) of model m_i , given the data y , is called the posterior probability (\rightarrow Definition I/5.1.7) 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.

Metadata: ID: D87 | shortcut: pmp | author: JoramSoch | date: 2020-07-28, 03:30.

3.5.2 Derivation

Theorem: Let there be a set of generative models (\rightarrow Definition I/5.1.1) m_1, \dots, m_M with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$ and prior probabilities (\rightarrow Definition I/5.1.3) $p(m_1), \dots, p(m_M)$. Then, the posterior probability (\rightarrow Definition IV/3.5.1) 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 Proof I/5.3.1), the posterior model probability (\rightarrow Definition IV/3.5.1) 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 Definition 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 Definition I/1.3.4), we have

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)} . \quad (4)$$

Sources:

- original work

Metadata: ID: P139 | shortcut: pmp-der | author: JoramSoch | date: 2020-07-28, 03:58.

3.5.3 Calculation from Bayes factors

Theorem: Let m_0, m_1, \dots, m_M be $M + 1$ statistical models with model evidences (\rightarrow Definition IV/3.1.1) $p(y|m_0), p(y|m_1), \dots, p(y|m_M)$. Then, the posterior model probabilities (\rightarrow Definition IV/3.5.1) 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 Definition IV/3.4.1) comparing model m_i with m_0 and α_i is the prior (\rightarrow Definition I/5.1.3) odds ratio (\rightarrow Definition “odds”) of model m_i against m_0 .

Proof: Define the Bayes factor (\rightarrow Definition IV/3.4.1) for m_i

$$\text{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \quad (2)$$

and prior odds ratio of m_i against m_0

$$\alpha_i = \frac{p(m_i)}{p(m_0)}. \quad (3)$$

The posterior model probability (\rightarrow Proof IV/3.5.2) 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.

Metadata: ID: P74 | shortcut: pmp-bf | author: JoramSoch | date: 2020-03-03, 13:13.

3.5.4 Calculation from log Bayes factor

Theorem: Let m_1 and m_2 be two statistical models with the log Bayes factor (\rightarrow Definition IV/3.3.1) LBF_{12} in favor of model m_1 and against model m_2 . Then, if both models are equally likely apriori (\rightarrow Definition I/5.1.3), the posterior model probability (\rightarrow Definition IV/3.5.1) 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 Proof I/5.3.2), the posterior odds ratio (\rightarrow Definition “odds”) 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 Definition I/5.1.3), the prior odds ratio (\rightarrow Definition “odds”) 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 Definition IV/3.4.1), therefore

$$\frac{p(m_1|y)}{p(m_2|y)} = \text{BF}_{12} . \quad (4)$$

Because the two posterior model probabilities (\rightarrow Definition IV/3.5.1) 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 Definition IV/3.5.1), 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 Definition IV/3.3.1), 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.

Metadata: ID: P73 | shortcut: pmp-lbf | author: JoramSoch | date: 2020-03-03, 12:27.

3.5.5 Calculation from log model evidences

Theorem: Let m_1, \dots, m_M be M statistical models with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$. Then, the posterior model probabilities (\rightarrow Definition IV/3.5.1) 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 Definition I/5.1.3) model probabilities.

Proof: The posterior model probability (\rightarrow Proof IV/3.5.2) 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 Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (3)$$

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.

Metadata: ID: P66 | shortcut: pmp-lme | author: JoramSoch | date: 2020-02-27, 21:33.

3.6 Bayesian model averaging

3.6.1 Definition

Definition: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with posterior model probabilities (\rightarrow Definition IV/3.5.1) $p(m_1|y), \dots, p(m_M|y)$ and posterior distributions (\rightarrow Definition I/5.1.7) $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$. Then, Bayesian model averaging (BMA) consists in finding the marginal (\rightarrow Definition I/1.5.3) posterior (\rightarrow Definition I/5.1.7) density (\rightarrow Definition I/1.6.6), conditional (\rightarrow Definition I/1.3.4) on the measured data y , but unconditional (\rightarrow Definition 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.

Metadata: ID: D89 | shortcut: bma | author: JoramSoch | date: 2020-08-03, 21:34.

3.6.2 Derivation

Theorem: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with posterior model probabilities (\rightarrow Definition IV/3.5.1) $p(m_1|y), \dots, p(m_M|y)$ and posterior distributions (\rightarrow Definition I/5.1.7) $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$. Then, the marginal (\rightarrow Definition I/1.5.3) posterior (\rightarrow Definition I/5.1.7) density (\rightarrow Definition I/1.6.6), conditional (\rightarrow Definition I/1.3.4) on the measured data y , but unconditional (\rightarrow Definition 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 Definition I/1.3.3), the probability distribution of the shared parameters θ conditional (\rightarrow Definition I/1.3.4) on the measured data y can be obtained by marginalizing (\rightarrow Definition I/1.3.3) over the discrete random variable (\rightarrow Definition 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 Definition 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 Definition I/5.1.7) of the i -th model and $p(m_i|y)$ happens to be the posterior probability (\rightarrow Definition IV/3.5.1) of the i -th model.

Sources:

- original work

Metadata: ID: P143 | shortcut: bma-der | author: JoramSoch | date: 2020-08-03, 22:05.

3.6.3 Calculation from log model evidences

Theorem: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) describing the same measured data y with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$ and shared model parameters θ . Then, Bayesian model averaging (\rightarrow Definition IV/3.6.1) determines the following posterior distribution over θ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|m_i, y) \cdot \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} , \quad (1)$$

where $p(\theta|m_i, y)$ is the posterior distributions over θ obtained using m_i .

Proof: According to the law of marginal probability (\rightarrow Definition I/1.3.3), the probability of the shared parameters θ conditional on the measured data y can be obtained (\rightarrow Proof IV/3.6.2) 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 Definition IV/3.5.1) of the i -th model. One can express posterior model probabilities in terms of log model evidences (\rightarrow Proof IV/3.5.5) 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.

Metadata: ID: P67 | shortcut: bma-lme | author: JoramSoch | date: 2020-02-27, 21:58.

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	224
P2	mlr-ols	Ordinary least squares for multiple linear regression	JoramSoch	2019-09-27	332
P3	lme-anc	Partition of the log model evidence into accuracy and complexity	JoramSoch	2019-09-27	419
P4	bayes-th	Bayes' theorem	JoramSoch	2019-09-27	128
P5	mse-bnv	Partition of the mean squared error into bias and variance	JoramSoch	2019-11-27	110
P6	ng-kl	Kullback-Leibler divergence for the normal-gamma distribution	JoramSoch	2019-12-06	238
P7	glm-mle	Maximum likelihood estimation for the general linear model	JoramSoch	2019-12-06	356
P8	rsq-der	Derivation of R^2 and adjusted R^2	JoramSoch	2019-12-06	410
P9	blr-prior	Conjugate prior distribution for Bayesian linear regression	JoramSoch	2020-01-03	344
P10	blr-post	Posterior distribution for Bayesian linear regression	JoramSoch	2020-01-03	346
P11	blr-lme	Log model evidence for Bayesian linear regression	JoramSoch	2020-01-03	348
P12	bayes-rule	Bayes' rule	JoramSoch	2020-01-06	128
P13	lme-der	Derivation of the log model evidence	JoramSoch	2020-01-06	418
P14	rsq-mll	Relationship between R^2 and maximum log-likelihood	JoramSoch	2020-01-08	411
P15	norm-mean	Mean of the normal distribution	JoramSoch	2020-01-09	167
P16	norm-med	Median of the normal distribution	JoramSoch	2020-01-09	168
P17	norm-mode	Mode of the normal distribution	JoramSoch	2020-01-09	169
P18	norm-var	Variance of the normal distribution	JoramSoch	2020-01-09	170
P19	dmi-mce	Relation of mutual information to marginal and conditional entropy	JoramSoch	2020-01-13	94

P20	dmi-mje	Relation of mutual information to marginal and joint entropy	JoramSoch	2020-01-13	95
P21	dmi-jce	Relation of mutual information to joint and conditional entropy	JoramSoch	2020-01-13	96
P22	bern-mean	Mean of the Bernoulli distribution	JoramSoch	2020-01-16	135
P23	bin-mean	Mean of the binomial distribution	JoramSoch	2020-01-16	137
P24	cat-mean	Mean of the categorical distribution	JoramSoch	2020-01-16	142
P25	mult-mean	Mean of the multinomial distribution	JoramSoch	2020-01-16	144
P26	matn-mvn	Equivalence of matrix-normal distribution and multivariate normal distribution	JoramSoch	2020-01-20	251
P27	poiss-mle	Maximum likelihood estimation for Poisson-distributed data	JoramSoch	2020-01-20	373
P28	beta-mome	Method of moments for beta-distributed data	JoramSoch	2020-01-22	387
P29	bin-prior	Conjugate prior distribution for binomial observations	JoramSoch	2020-01-23	393
P30	bin-post	Posterior distribution for binomial observations	JoramSoch	2020-01-24	394
P31	bin-lme	Log model evidence for binomial observations	JoramSoch	2020-01-24	395
P32	bic-der	Derivation of the Bayesian information criterion	JoramSoch	2020-01-26	415
P33	norm-pdf	Probability density function of the normal distribution	JoramSoch	2020-01-27	161
P34	mvn-pdf	Probability density function of the multivariate normal distribution	JoramSoch	2020-01-27	221
P35	mvn-marg	Marginal distributions of the multivariate normal distribution	JoramSoch	2020-01-29	225
P36	ng-marg	Marginal distributions of the normal-gamma distribution	JoramSoch	2020-01-29	239
P37	cuni-pdf	Probability density function of the continuous uniform distribution	JoramSoch	2020-01-31	146
P38	cuni-cdf	Cumulative distribution function of the continuous uniform distribution	JoramSoch	2020-01-02	147

P39	cuni-qf	Quantile function of the continuous uniform distribution	JoramSoch	2020-01-02	148
P40	mlr-ols2	Ordinary least squares for multiple linear regression	JoramSoch	2020-02-03	333
P41	poissexp-prior	Conjugate prior distribution for the Poisson distribution with exposure values	JoramSoch	2020-02-04	381
P42	poissexp-post	Posterior distribution for the Poisson distribution with exposure values	JoramSoch	2020-02-04	383
P43	poissexp-lme	Log model evidence for the Poisson distribution with exposure values	JoramSoch	2020-02-04	384
P44	ng-pdf	Probability density function of the normal-gamma distribution	JoramSoch	2020-02-07	234
P45	gam-pdf	Probability density function of the gamma distribution	JoramSoch	2020-02-08	188
P46	exp-pdf	Probability density function of the exponential distribution	JoramSoch	2020-02-08	199
P47	exp-mean	Mean of the exponential distribution	JoramSoch	2020-02-10	201
P48	exp-cdf	Cumulative distribution function of the exponential distribution	JoramSoch	2020-02-11	199
P49	exp-med	Median of the exponential distribution	JoramSoch	2020-02-11	202
P50	exp-qf	Quantile function of the exponential distribution	JoramSoch	2020-02-12	200
P51	exp-mode	Mode of the exponential distribution	JoramSoch	2020-02-12	203
P52	mean-nonneg	Non-negativity of the expected value	JoramSoch	2020-02-13	42
P53	mean-lin	Linearity of the expected value	JoramSoch	2020-02-13	43
P54	mean-mono	Monotonicity of the expected value	JoramSoch	2020-02-17	45
P55	mean-mult	(Non-)Multiplicativity of the expected value	JoramSoch	2020-02-17	46
P56	ci-wilks	Construction of confidence intervals using Wilks' theorem	JoramSoch	2020-02-19	111
P57	ent-nonneg	Non-negativity of the Shannon entropy	JoramSoch	2020-02-19	80

P58	cmi-mcde	Relation of continuous mutual information to marginal and conditional differential entropy	JoramSoch	2020-02-21	98
P59	cmi-mjde	Relation of continuous mutual information to marginal and joint differential entropy	JoramSoch	2020-02-21	99
P60	cmi-jcde	Relation of continuous mutual information to joint and conditional differential entropy	JoramSoch	2020-02-21	100
P61	resvar-bias	Maximum likelihood estimator of variance is biased	JoramSoch	2020-02-24	406
P62	resvar-unb	Construction of unbiased estimator for variance	JoramSoch	2020-02-25	408
P63	snr-rsq	Relationship between signal-to-noise ratio and R^2	JoramSoch	2020-02-26	413
P64	lbf-lme	Log Bayes factor in terms of log model evidences	JoramSoch	2020-02-27	426
P65	lfe-lme	Log family evidences in terms of log model evidences	JoramSoch	2020-02-27	424
P66	pmp-lme	Posterior model probabilities in terms of log model evidences	JoramSoch	2020-02-27	434
P67	bma-lme	Bayesian model averaging in terms of log model evidences	JoramSoch	2020-02-27	436
P68	dent-neg	Differential entropy can be negative	JoramSoch	2020-03-02	86
P69	exp-gam	Exponential distribution is a special case of gamma distribution	JoramSoch	2020-03-02	198
P70	matn-pdf	Probability density function of the matrix-normal distribution	JoramSoch	2020-03-02	250
P71	norm-mgf	Moment-generating function of the normal distribution	JoramSoch	2020-03-03	161
P72	logreg-lonp	Log-odds and probability in logistic regression	JoramSoch	2020-03-03	403
P73	pmp-lbf	Posterior model probability in terms of log Bayes factor	JoramSoch	2020-03-03	433
P74	pmp-bf	Posterior model probabilities in terms of Bayes factors	JoramSoch	2020-03-03	432

P75	mlr-mat	Transformation matrices for ordinary least squares	JoramSoch	2020-03-09	337
P76	mlr-pss	Partition of sums of squares in ordinary least squares	JoramSoch	2020-03-09	335
P77	mlr-wls	Weighted least squares for multiple linear regression	JoramSoch	2020-03-11	340
P78	mlr-mle	Maximum likelihood estimation for multiple linear regression	JoramSoch	2020-03-11	342
P79	mult-prior	Conjugate prior distribution for multinomial observations	JoramSoch	2020-03-11	397
P80	mult-post	Posterior distribution for multinomial observations	JoramSoch	2020-03-11	398
P81	mult-lme	Log model evidence for multinomial observations	JoramSoch	2020-03-11	399
P82	cuni-mean	Mean of the continuous uniform distribution	JoramSoch	2020-03-16	149
P83	cuni-med	Median of the continuous uniform distribution	JoramSoch	2020-03-16	151
P84	cuni-med	Mode of the continuous uniform distribution	JoramSoch	2020-03-16	151
P85	norm-cdf	Cumulative distribution function of the normal distribution	JoramSoch	2020-03-20	162
P86	norm-cdfwerf	Expression of the cumulative distribution function of the normal distribution without the error function	JoramSoch	2020-03-20	164
P87	norm-qf	Quantile function of the normal distribution	JoramSoch	2020-03-20	166
P88	mvn-cond	Conditional distributions of the multivariate normal distribution	JoramSoch	2020-03-20	226
P89	jl-lfnprior	Joint likelihood is the product of likelihood function and prior density	JoramSoch	2020-05-05	122
P90	post-jl	Posterior density is proportional to joint likelihood	JoramSoch	2020-05-05	123
P91	ml-jl	Marginal likelihood is a definite integral of joint likelihood	JoramSoch	2020-05-05	124
P92	mvn-kl	Kullback-Leibler divergence for the multivariate normal distribution	JoramSoch	2020-05-05	223

P93	gam-kl	Kullback-Leibler divergence for the gamma distribution	JoramSoch	2020-05-05	196
P94	beta-pdf	Probability density function of the beta distribution	JoramSoch	2020-05-05	210
P95	dir-pdf	Probability density function of the Dirichlet distribution	JoramSoch	2020-05-05	244
P96	bern-pmf	Probability mass function of the Bernoulli distribution	JoramSoch	2020-05-11	135
P97	bin-pmf	Probability mass function of the binomial distribution	JoramSoch	2020-05-11	136
P98	cat-pmf	Probability mass function of the categorical distribution	JoramSoch	2020-05-11	142
P99	mult-pmf	Probability mass function of the multinomial distribution	JoramSoch	2020-05-11	143
P100	mvn-dent	Differential entropy of the multivariate normal distribution	JoramSoch	2020-05-14	221
P101	norm-dent	Differential entropy of the normal distribution	JoramSoch	2020-05-14	175
P102	poiss-pmf	Probability mass function of the Poisson distribution	JoramSoch	2020-05-14	138
P103	mean-nnrvar	Expected value of a non-negative random variable	JoramSoch	2020-05-18	41
P104	var-mean	Partition of variance into expected values	JoramSoch	2020-05-19	54
P105	logreg-pnlo	Probability and log-odds in logistic regression	JoramSoch	2020-05-19	402
P106	glm-ols	Ordinary least squares for the general linear model	JoramSoch	2020-05-19	354
P107	glm-wls	Weighted least squares for the general linear model	JoramSoch	2020-05-19	355
P108	gam-mean	Mean of the gamma distribution	JoramSoch	2020-05-19	190
P109	gam-var	Variance of the gamma distribution	JoramSoch	2020-05-19	191
P110	gam-logmean	Logarithmic expectation of the gamma distribution	JoramSoch	2020-05-25	192
P111	norm-snorm	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-05-26	152

P112	gam-sgam	Relationship between gamma distribution and standard gamma distribution	JoramSoch	2020-05-26	186
P113	kl-ent	Relation of Kullback-Leibler divergence to entropy	JoramSoch	2020-05-27	107
P114	kl-dent	Relation of continuous Kullback-Leibler divergence to differential entropy	JoramSoch	2020-05-27	108
P115	kl-inv	Invariance of the Kullback-Leibler divergence under parameter transformation	JoramSoch	2020-05-28	106
P116	kl-add	Additivity of the Kullback-Leibler divergence for independent distributions	JoramSoch	2020-05-31	105
P117	kl-nonneg	Non-negativity of the Kullback-Leibler divergence	JoramSoch	2020-05-31	102
P118	cov-mean	Partition of covariance into expected values	JoramSoch	2020-06-02	61
P119	cov-corr	Relationship between covariance and correlation	JoramSoch	2020-06-02	62
P120	covmat-mean	Partition of a covariance matrix into expected values	JoramSoch	2020-06-06	64
P121	covmat-corrmat	Relationship between covariance matrix and correlation matrix	JoramSoch	2020-06-06	65
P122	precmat-corrmat	Relationship between precision matrix and correlation matrix	JoramSoch	2020-06-06	67
P123	var-nonneg	Non-negativity of the variance	JoramSoch	2020-06-06	54
P124	var-const	Variance of constant is zero	JoramSoch	2020-06-27	55
P126	var-inv	Invariance of the variance under addition of a constant	JoramSoch	2020-07-07	56
P127	var-scal	Scaling of the variance upon multiplication with a constant	JoramSoch	2020-07-07	57
P128	var-sum	Variance of the sum of two random variables	JoramSoch	2020-07-07	57
P129	var-lincomb	Variance of the linear combination of two random variables	JoramSoch	2020-07-07	58

P130	var-add	Additivity of the variance for independent random variables	JoramSoch	2020-07-07	59
P131	mean-qf	Expectation of a quadratic form	JoramSoch	2020-07-13	48
P132	lfe-der	Derivation of the log family evidence	JoramSoch	2020-07-13	423
P133	blr-pp	Posterior probability of the alternative hypothesis for Bayesian linear regression	JoramSoch	2020-07-17	350
P134	blr-pcr	Posterior credibility region against the omnibus null hypothesis for Bayesian linear regression	JoramSoch	2020-07-17	352
P135	mlr-idem	Projection matrix and residual-forming matrix are idempotent	JoramSoch	2020-07-22	339
P136	mlr-wls2	Weighted least squares for multiple linear regression	JoramSoch	2020-07-22	341
P137	lbf-der	Derivation of the log Bayes factor	JoramSoch	2020-07-22	426
P138	mean-lotus	Law of the unconscious statistician	JoramSoch	2020-07-22	50
P139	pmp-der	Derivation of the posterior model probability	JoramSoch	2020-07-28	432
P140	duni-pmf	Probability mass function of the discrete uniform distribution	JoramSoch	2020-07-28	132
P141	duni-cdf	Cumulative distribution function of the discrete uniform distribution	JoramSoch	2020-07-28	133
P142	duni-qf	Quantile function of the discrete uniform distribution	JoramSoch	2020-07-28	134
P143	bma-der	Derivation of Bayesian model averaging	JoramSoch	2020-08-03	436
P144	matn-trans	Transposition of a matrix-normal random variable	JoramSoch	2020-08-03	254
P145	matn-ltt	Linear transformation theorem for the matrix-normal distribution	JoramSoch	2020-08-03	253
P146	ng-cond	Conditional distributions of the normal-gamma distribution	JoramSoch	2020-08-05	242
P147	kl-nonsymm	Non-symmetry of the Kullback-Leibler divergence	JoramSoch	2020-08-11	103
P148	kl-conv	Convexity of the Kullback-Leibler divergence	JoramSoch	2020-08-11	105

P149	ent-conc	Concavity of the Shannon entropy	JoramSoch	2020-08-11	81
P150	entcross-conv	Convexity of the cross-entropy	JoramSoch	2020-08-11	83
P151	poiss-mean	Mean of the Poisson distribution	JoramSoch	2020-08-19	139
P152	norm-fwhm	Full width at half maximum for the normal distribution	JoramSoch	2020-08-19	172
P153	mom-mgf	Moment in terms of moment-generating function	JoramSoch	2020-08-19	75
P154	mgf-ltt	Linear transformation theorem for the moment-generating function	JoramSoch	2020-08-19	38
P155	mgf-lincomb	Moment-generating function of linear combination of independent random variables	JoramSoch	2020-08-19	39
P156	bf-sddr	Savage-Dickey Density Ratio for computing Bayes Factors	tomfaulkenberry	2020-08-26	428
P157	bf-ep	Encompassing Prior Method for computing Bayes Factors	tomfaulkenberry	2020-09-02	430
P158	cov-ind	Covariance of independent random variables	JoramSoch	2020-09-03	62
P159	mblr-prior	Conjugate prior distribution for multivariate Bayesian linear regression	JoramSoch	2020-09-03	366
P160	mblr-post	Posterior distribution for multivariate Bayesian linear regression	JoramSoch	2020-09-03	368
P161	mblr-lme	Log model evidence for multivariate Bayesian linear regression	JoramSoch	2020-09-03	370
P162	wald-pdf	Probability density function of the Wald distribution	tomfaulkenberry	2020-09-04	216
P163	bf-trans	Transitivity of Bayes Factors	tomfaulkenberry	2020-09-07	428
P164	gibbs-ineq	Gibbs' inequality	JoramSoch	2020-09-09	84
P165	logsum-ineq	Log sum inequality	JoramSoch	2020-09-09	85
P166	kl-nonneg2	Non-negativity of the Kullback-Leibler divergence	JoramSoch	2020-09-09	102
P167	momcent-1st	First central moment is zero	JoramSoch	2020-09-09	78

P168	wald-mgf	Moment-generating function of the Wald distribution	tomfaulkenberry	2020-09-13	216
P169	wald-mean	Mean of the Wald distribution	tomfaulkenberry	2020-09-13	218
P170	wald-var	Variance of the Wald distribution	tomfaulkenberry	2020-09-13	219
P171	momraw-1st	First raw moment is mean	JoramSoch	2020-10-08	77
P172	momraw-2nd	Relationship between second raw moment, variance and mean	JoramSoch	2020-10-08	77
P173	momcent-2nd	Second central moment is variance	JoramSoch	2020-10-08	78
P174	chi2-gam	Chi-squared distribution is a special case of gamma distribution	kjpetrykowski	2020-10-12	204
P175	chi2-mom	Moments of the chi-squared distribution	kjpetrykowski	2020-10-13	207
P176	norm-snorm2	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-10-15	153
P177	gam-sgam2	Relationship between gamma distribution and standard gamma distribution	JoramSoch	2020-10-15	187
P178	gam-cdf	Cumulative distribution function of the gamma distribution	JoramSoch	2020-10-15	188
P179	gam-xlogx	Expected value of $x \ln(x)$ for a gamma distribution	JoramSoch	2020-10-15	194
P180	norm-snorm3	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-10-22	154
P181	dir-ep	Exceedance probabilities for the Dirichlet distribution	JoramSoch	2020-10-22	246
P182	dir-mle	Maximum likelihood estimation for Dirichlet-distributed data	JoramSoch	2020-10-22	389
P183	cdf-sifct	Cumulative distribution function of a strictly increasing function of a random variable	JoramSoch	2020-10-29	30
P184	pmf-sifct	Probability mass function of a strictly increasing function of a discrete random variable	JoramSoch	2020-10-29	19

P185	pdf-sifct	Probability density function of a strictly increasing function of a continuous random variable	JoramSoch	2020-10-29	22
P186	cdf-sdfct	Cumulative distribution function of a strictly decreasing function of a random variable	JoramSoch	2020-11-06	30
P187	pmf-sdfct	Probability mass function of a strictly decreasing function of a discrete random variable	JoramSoch	2020-11-06	20
P188	pdf-sdfct	Probability density function of a strictly decreasing function of a continuous random variable	JoramSoch	2020-11-06	23
P189	cdf-pmf	Cumulative distribution function in terms of probability mass function of a discrete random variable	JoramSoch	2020-11-12	31
P190	cdf-pdf	Cumulative distribution function in terms of probability density function of a continuous random variable	JoramSoch	2020-11-12	32
P191	pdf-cdf	Probability density function is first derivative of cumulative distribution function	JoramSoch	2020-11-12	28
P192	qf-cdf	Quantile function is inverse of strictly monotonically increasing cumulative distribution function	JoramSoch	2020-11-12	35
P193	norm-kl	Kullback-Leibler divergence for the normal distribution	JoramSoch	2020-11-19	176
P194	gam-qf	Quantile function of the gamma distribution	JoramSoch	2020-11-19	189
P195	beta-cdf	Cumulative distribution function of the beta distribution	JoramSoch	2020-11-19	212
P196	norm-gi	Gaussian integral	JoramSoch	2020-11-25	159
P197	chi2-pdf	Probability density function of the chi-squared distribution	JoramSoch	2020-11-25	205
P198	beta-mgf	Moment-generating function of the beta distribution	JoramSoch	2020-11-25	211
P199	dent-inv	Invariance of the differential entropy under addition of a constant	JoramSoch	2020-12-02	87

P200	dent-add	Addition of the differential entropy upon multiplication with a constant	JoramSoch	2020-12-02	88
P201	ug-prior	Conjugate prior distribution for the univariate Gaussian	JoramSoch	2021-03-03	266
P202	ug-post	Posterior distribution for the univariate Gaussian	JoramSoch	2021-03-03	268
P203	ug-lme	Log model evidence for the univariate Gaussian	JoramSoch	2021-03-03	271
P204	ug-ttest1	One-sample t-test for independent observations	JoramSoch	2021-03-12	262
P205	ug-ttest2	Two-sample t-test for independent observations	JoramSoch	2021-03-12	264
P206	ug-ttestp	Paired t-test for dependent observations	JoramSoch	2021-03-12	265
P207	ugkv-mle	Maximum likelihood estimation for the univariate Gaussian with known variance	JoramSoch	2021-03-24	276
P208	ugkv-ztest1	One-sample z-test for independent observations	JoramSoch	2021-03-24	277
P209	ugkv-ztest2	Two-sample z-test for independent observations	JoramSoch	2021-03-24	278
P210	ugkv-ztestp	Paired z-test for dependent observations	JoramSoch	2021-03-24	280
P211	ugkv-prior	Conjugate prior distribution for the univariate Gaussian with known variance	JoramSoch	2021-03-24	280
P212	ugkv-post	Posterior distribution for the univariate Gaussian with known variance	JoramSoch	2021-03-24	282
P213	ugkv-lme	Log model evidence for the univariate Gaussian with known variance	JoramSoch	2021-03-24	285
P214	ugkv-anc	Accuracy and complexity for the univariate Gaussian with known variance	JoramSoch	2021-03-24	286
P215	ugkv-lbf	Log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	288

P216	ugkv-lbfmean	Expectation of the log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	289
P217	ugkv-cvlme	Cross-validated log model evidence for the univariate Gaussian with known variance	JoramSoch	2021-03-24	291
P218	ugkv-cvlbf	Cross-validated log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	293
P219	ugkv-cvlbfmean	Expectation of the cross-validated log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	294
P220	cdf-pit	Probability integral transform using cumulative distribution function	JoramSoch	2021-04-07	33
P221	cdf-itm	Inverse transformation method using cumulative distribution function	JoramSoch	2021-04-07	33
P222	cdf-dt	Distributional transformation using cumulative distribution function	JoramSoch	2021-04-07	34
P223	ug-mle	Maximum likelihood estimation for the univariate Gaussian	JoramSoch	2021-04-16	260
P224	poissexp-mle	Maximum likelihood estimation for the Poisson distribution with exposure values	JoramSoch	2021-04-16	379
P225	poiss-prior	Conjugate prior distribution for Poisson-distributed data	JoramSoch	2020-04-21	375
P226	poiss-post	Posterior distribution for Poisson-distributed data	JoramSoch	2020-04-21	376
P227	poiss-lme	Log model evidence for Poisson-distributed data	JoramSoch	2020-04-21	377
P228	beta-mean	Mean of the beta distribution	JoramSoch	2021-04-29	213
P229	beta-var	Variance of the beta distribution	JoramSoch	2021-04-29	214
P230	poiss-var	Variance of the Poisson distribution	JoramSoch	2021-04-29	140
P231	mvt-f	Relationship between multivariate t-distribution and F-distribution	JoramSoch	2021-05-04	232
P232	nst-t	Relationship between non-standardized t-distribution and t-distribution	JoramSoch	2021-05-11	182

P233	norm-chi2	Relationship between normal distribution and chi-squared distribution	JoramSoch	2021-05-20	155
P234	norm-t	Relationship between normal distribution and t-distribution	JoramSoch	2021-05-27	157
P235	norm-lincomb	Linear combination of independent normal random variables	JoramSoch	2021-06-02	179
P236	mvn-ind	Necessary and sufficient condition for independence of multivariate normal random variables	JoramSoch	2021-06-02	230
P237	ng-mean	Mean of the normal-gamma distribution	JoramSoch	2021-07-08	235
P238	ng-dent	Differential entropy of the normal-gamma distribution	JoramSoch	2021-07-08	236
P239	gam-dent	Differential entropy of the gamma distribution	JoramSoch	2021-07-14	195
P240	ug-anc	Accuracy and complexity for the univariate Gaussian	JoramSoch	2021-07-14	274
P241	prob-ind	Probability under statistical independence	JoramSoch	2021-07-23	9
P242	prob-exc	Probability under mutual exclusivity	JoramSoch	2021-07-23	10
P243	prob-mon	Monotonicity of probability	JoramSoch	2021-07-30	11
P244	prob-emp	Probability of the empty set	JoramSoch	2021-07-30	12
P245	prob-comp	Probability of the complement	JoramSoch	2021-07-30	13
P246	prob-range	Range of probability	JoramSoch	2021-07-30	13
P247	prob-add	Addition law of probability	JoramSoch	2021-07-30	14
P248	prob-tot	Law of total probability	JoramSoch	2021-08-08	15
P249	prob-exh	Probability of exhaustive events	JoramSoch	2021-08-08	16
P250	norm-maxent	Normal distribution maximizes differential entropy for fixed variance	JoramSoch	2020-08-25	178
P251	norm-extr	Extreme points of the probability density function of the normal distribution	JoramSoch	2020-08-25	173
P252	norm-infl	Inflection points of the probability density function of the normal distribution	JoramSoch	2020-08-26	174

P253	pmf-invft	Probability mass function of an invertible function of a random vector	JoramSoch	2021-08-30	20
P254	pdf-invft	Probability density function of an invertible function of a continuous random vector	JoramSoch	2021-08-30	25
P255	pdf-linfct	Probability density function of a linear function of a continuous random vector	JoramSoch	2021-08-30	27
P256	cdf-sumind	Cumulative distribution function of a sum of independent random variables	JoramSoch	2021-08-30	29
P257	pmf-sumind	Probability mass function of a sum of independent discrete random variables	JoramSoch	2021-08-30	18
P258	pdf-sumind	Probability density function of a sum of independent discrete random variables	JoramSoch	2021-08-30	21
P259	cf-fct	Characteristic function of a function of a random variable	JoramSoch	2021-09-22	37
P260	mgf-fct	Moment-generating function of a function of a random variable	JoramSoch	2021-09-22	38
P261	dent-addvec	Addition of the differential entropy upon multiplication with invertible matrix	JoramSoch	2021-10-07	89
P262	dent-noninv	Non-invariance of the differential entropy under change of variables	JoramSoch	2021-10-07	91
P263	t-pdf	Probability density function of the t-distribution	JoramSoch	2021-10-12	183
P264	f-pdf	Probability density function of the F-distribution	JoramSoch	2021-10-12	208
P265	tglm-dist	Distribution of the transformed general linear model	JoramSoch	2021-10-21	359
P266	tglm-para	Equivalence of parameter estimates from the transformed general linear model	JoramSoch	2021-10-21	360
P267	iglm-dist	Distribution of the inverse general linear model	JoramSoch	2021-10-21	361

P268	iglm-blue	Best linear unbiased estimator for the inverse general linear model	JoramSoch	2021-10-21	362
P269	cfm-para	Parameters of the corresponding forward model	JoramSoch	2021-10-21	364
P270	cfm-exist	Existence of a corresponding forward model	JoramSoch	2021-10-21	365
P271	slr-ols	Ordinary least squares for simple linear regression	JoramSoch	2021-10-27	298
P272	slr-olsmean	Expectation of parameter estimates for simple linear regression	JoramSoch	2021-10-27	302
P273	slr-olsvar	Variance of parameter estimates for simple linear regression	JoramSoch	2021-10-27	304
P274	slr-meancent	Effects of mean-centering on parameter estimates for simple linear regression	JoramSoch	2021-10-27	309
P275	slr-comp	The regression line goes through the center of mass point	JoramSoch	2021-10-27	311
P276	slr-ressum	The sum of residuals is zero in simple linear regression	JoramSoch	2021-10-27	325
P277	slr-rescorr	The residuals and the covariate are uncorrelated in simple linear regression	JoramSoch	2021-10-27	326
P278	slr-resvar	Relationship between residual variance and sample variance in simple linear regression	JoramSoch	2021-10-27	327
P279	slr-corr	Relationship between correlation coefficient and slope estimate in simple linear regression	JoramSoch	2021-10-27	329
P280	slr-rsq	Relationship between coefficient of determination and correlation coefficient in simple linear regression	JoramSoch	2021-10-27	330
P281	slr-mlr	Simple linear regression is a special case of multiple linear regression	JoramSoch	2021-11-09	297
P282	slr-olsdist	Distribution of parameter estimates for simple linear regression	JoramSoch	2021-11-09	307
P283	slr-proj	Projection of a data point to the regression line	JoramSoch	2021-11-09	312

P284	slr-sss	Sums of squares for simple linear regression	JoramSoch	2021-11-09	313
P285	slr-mat	Transformation matrices for simple linear regression	JoramSoch	2021-11-09	315
P286	slr-wls	Weighted least squares for simple linear regression	JoramSoch	2021-11-16	318
P287	slr-mle	Maximum likelihood estimation for simple linear regression	JoramSoch	2021-11-16	321
P288	slr-ols2	Ordinary least squares for simple linear regression	JoramSoch	2021-11-16	300
P289	slr-wls2	Weighted least squares for simple linear regression	JoramSoch	2021-11-16	320
P290	slr-mle2	Maximum likelihood estimation for simple linear regression	JoramSoch	2021-11-16	324
P291	mean-tot	Law of total expectation	JoramSoch	2021-11-26	49
P292	var-tot	Law of total variance	JoramSoch	2021-11-26	59
P293	cov-tot	Law of total covariance	JoramSoch	2021-11-26	63
P294	dir-kl	Kullback-Leibler divergence for the Dirichlet distribution	JoramSoch	2021-12-02	245
P295	wish-kl	Kullback-Leibler divergence for the Wishart distribution	JoramSoch	2021-12-02	256
P296	matn-kl	Kullback-Leibler divergence for the matrix-normal distribution	JoramSoch	2021-12-02	252
P297	matn-samp	Sampling from the matrix-normal distribution	JoramSoch	2021-12-07	255
P298	mean-tr	Expected value of the trace of a matrix	JoramSoch	2021-12-07	48
P299	corr-z	Correlation coefficient in terms of standard scores	JoramSoch	2021-12-14	70
P300	corr-range	Correlation always falls between -1 and +1	JoramSoch	2021-12-14	68
P301	bern-var	Variance of the Bernoulli distribution	JoramSoch	2022-01-20	??
P302	bin-var	Variance of the binomial distribution	JoramSoch	2022-01-20	??

P303	bern-varrange	Range of the variance of the Bernoulli distribution	JoramSoch	2022-01-27	??
P304	bin-varrange	Range of the variance of the binomial distribution	JoramSoch	2022-01-27	??
P305	mlr-mll	Maximum log-likelihood for multiple linear regression	JoramSoch	2022-02-04	??
P306	lognorm-med	Median of the log-normal distribution	majapavlo	2022-02-07	??
P307	mlr-aic	Akaike information criterion for multiple linear regression	JoramSoch	2022-02-11	??
P308	mlr-bic	Bayesian information criterion for multiple linear regression	JoramSoch	2022-02-11	??
P309	mlr-aicc	Corrected Akaike information criterion for multiple linear regression	JoramSoch	2022-02-11	??
P310	lognorm-pdf	Probability density function of the log-normal distribution	majapavlo	2022-02-13	??
P311	lognorm-mode	Mode of the log-normal distribution	majapavlo	2022-02-13	??
P312	mlr-dev	Deviance for multiple linear regression	JoramSoch	2022-03-01	??
P313	blr-dic	Deviance information criterion for multiple linear regression	JoramSoch	2022-03-01	??
P314	lme-pnp	Log model evidence in terms of prior and posterior distribution	JoramSoch	2022-03-11	??
P315	aicc-mll	Corrected Akaike information criterion in terms of maximum log-likelihood	JoramSoch	2022-03-11	??
P316	aicc-aic	Corrected Akaike information criterion converges to uncorrected Akaike information criterion when infinite data are available	JoramSoch	2022-03-18	??
P317	mle-bias	Maximum likelihood estimation can result in biased estimates	JoramSoch	2022-03-18	??
P318	pval-h0	The p-value follows a uniform distribution under the null hypothesis	JoramSoch	2022-03-18	??
P319	prob-exh2	Probability of exhaustive events	JoramSoch	2022-03-27	??

2 Definition by Number

ID	Shortcut	Definition	Author	Date	Page
D1	mvn	Multivariate normal distribution	JoramSoch	2020-01-22	221
D2	mgf	Moment-generating function	JoramSoch	2020-01-22	37
D3	cuni	Continuous uniform distribution	JoramSoch	2020-01-27	146
D4	norm	Normal distribution	JoramSoch	2020-01-27	151
D5	ng	Normal-gamma distribution	JoramSoch	2020-01-27	233
D6	matn	Matrix-normal distribution	JoramSoch	2020-01-27	250
D7	gam	Gamma distribution	JoramSoch	2020-02-08	185
D8	exp	Exponential distribution	JoramSoch	2020-02-08	198
D9	pmf	Probability mass function	JoramSoch	2020-02-13	18
D10	pdf	Probability density function	JoramSoch	2020-02-13	21
D11	mean	Expected value	JoramSoch	2020-02-13	41
D12	var	Variance	JoramSoch	2020-02-13	53
D13	cdf	Cumulative distribution function	JoramSoch	2020-02-17	28
D14	qf	Quantile function	JoramSoch	2020-02-17	35
D15	ent	Shannon entropy	JoramSoch	2020-02-19	80
D16	dent	Differential entropy	JoramSoch	2020-02-19	86
D17	ent-cond	Conditional entropy	JoramSoch	2020-02-19	82
D18	ent-joint	Joint entropy	JoramSoch	2020-02-19	82
D19	mi	Mutual information	JoramSoch	2020-02-19	97
D19	mi	Mutual information	JoramSoch	2020-02-19	97
D20	resvar	Residual variance	JoramSoch	2020-02-25	406
D21	rsq	Coefficient of determination	JoramSoch	2020-02-25	409
D22	snr	Signal-to-noise ratio	JoramSoch	2020-02-25	413
D23	aic	Akaike information criterion	JoramSoch	2020-02-25	415
D24	bic	Bayesian information criterion	JoramSoch	2020-02-25	415
D25	dic	Deviance information criterion	JoramSoch	2020-02-25	417
D26	lme	Log model evidence	JoramSoch	2020-02-25	418
D27	gm	Generative model	JoramSoch	2020-03-03	121

D28	lf	Likelihood function	JoramSoch	2020-03-03	121
D28	lf	Likelihood function	JoramSoch	2020-03-03	121
D29	prior	Prior distribution	JoramSoch	2020-03-03	121
D30	fpm	Full probability model	JoramSoch	2020-03-03	122
D31	jl	Joint likelihood	JoramSoch	2020-03-03	122
D32	post	Posterior distribution	JoramSoch	2020-03-03	123
D33	ml	Marginal likelihood	JoramSoch	2020-03-03	124
D34	dent-cond	Conditional differential entropy	JoramSoch	2020-03-21	92
D35	dent-joint	Joint differential entropy	JoramSoch	2020-03-21	93
D36	mlr	Multiple linear regression	JoramSoch	2020-03-21	331
D37	tss	Total sum of squares	JoramSoch	2020-03-21	334
D38	ess	Explained sum of squares	JoramSoch	2020-03-21	334
D39	rss	Residual sum of squares	JoramSoch	2020-03-21	334
D40	glm	General linear model	JoramSoch	2020-03-21	354
D41	poiss-data	Poisson-distributed data	JoramSoch	2020-03-22	373
D42	poissexp	Poisson distribution with exposure values	JoramSoch	2020-03-22	379
D43	wish	Wishart distribution	JoramSoch	2020-03-22	256
D44	bern	Bernoulli distribution	JoramSoch	2020-03-22	135
D45	bin	Binomial distribution	JoramSoch	2020-03-22	136
D46	cat	Categorical distribution	JoramSoch	2020-03-22	142
D47	mult	Multinomial distribution	JoramSoch	2020-03-22	143
D48	prob	Probability	JoramSoch	2020-05-10	5
D49	prob-joint	Joint probability	JoramSoch	2020-05-10	6
D50	prob-marg	Law of marginal probability	JoramSoch	2020-05-10	6
D51	prob-cond	Law of conditional probability	JoramSoch	2020-05-10	6
D52	kl	Kullback-Leibler divergence	JoramSoch	2020-05-10	101
D53	beta	Beta distribution	JoramSoch	2020-05-10	210
D54	dir	Dirichlet distribution	JoramSoch	2020-05-10	244
D55	dist	Probability distribution	JoramSoch	2020-05-17	16
D56	dist-joint	Joint probability distribution	JoramSoch	2020-05-17	17

D57	dist-marg	Marginal probability distribution	JoramSoch	2020-05-17	17
D58	dist-cond	Conditional probability distribution	JoramSoch	2020-05-17	17
D59	llf	Log-likelihood function	JoramSoch	2020-05-17	113
D60	mle	Maximum likelihood estimation	JoramSoch	2020-05-15	113
D61	mll	Maximum log-likelihood	JoramSoch	2020-05-15	114
D62	poiss	Poisson distribution	JoramSoch	2020-05-25	138
D63	snorm	Standard normal distribution	JoramSoch	2020-05-26	152
D64	sgam	Standard gamma distribution	JoramSoch	2020-05-26	185
D65	rvar	Random variable	JoramSoch	2020-05-27	3
D66	rvec	Random vector	JoramSoch	2020-05-27	4
D67	rmat	Random matrix	JoramSoch	2020-05-27	4
D68	cgf	Cumulant-generating function	JoramSoch	2020-05-31	40
D69	pgf	Probability-generating function	JoramSoch	2020-05-31	40
D70	cov	Covariance	JoramSoch	2020-06-02	60
D71	corr	Correlation	JoramSoch	2020-06-02	68
D72	covmat	Covariance matrix	JoramSoch	2020-06-06	63
D73	corrmat	Correlation matrix	JoramSoch	2020-06-06	70
D74	precmat	Precision matrix	JoramSoch	2020-06-06	66
D75	ind	Statistical independence	JoramSoch	2020-06-06	7
D76	logreg	Logistic regression	JoramSoch	2020-06-28	401
D77	beta-data	Beta-distributed data	JoramSoch	2020-06-28	387
D78	bin-data	Binomial observations	JoramSoch	2020-07-07	393
D79	mult-data	Multinomial observations	JoramSoch	2020-07-07	397
D80	lfe	Log family evidence	JoramSoch	2020-07-13	422
D81	emat	Estimation matrix	JoramSoch	2020-07-22	337
D82	pmat	Projection matrix	JoramSoch	2020-07-22	337
D83	rformat	Residual-forming matrix	JoramSoch	2020-07-22	337
D84	lbf	Log Bayes factor	JoramSoch	2020-07-22	425
D85	ent-cross	Cross-entropy	JoramSoch	2020-07-28	83
D86	dent-cross	Differential cross-entropy	JoramSoch	2020-07-28	93

D87	pmp	Posterior model probability	JoramSoch	2020-07-28	431
D88	duni	Discrete uniform distribution	JoramSoch	2020-07-28	132
D89	bma	Bayesian model averaging	JoramSoch	2020-08-03	435
D90	mom	Moment	JoramSoch	2020-08-19	74
D91	fwhm	Full width at half maximum	JoramSoch	2020-08-19	73
D92	bf	Bayes factor	tomfaulkenberry	2020-08-26	427
D93	encm	Encompassing model	tomfaulkenberry	2020-09-02	431
D94	std	Standard deviation	JoramSoch	2020-09-03	73
D95	wald	Wald distribution	tomfaulkenberry	2020-09-04	216
D96	const	Constant	JoramSoch	2020-09-09	4
D97	mom-raw	Raw moment	JoramSoch	2020-10-08	76
D98	mom-cent	Central moment	JoramSoch	2020-10-08	78
D99	mom-stand	Standardized moment	JoramSoch	2020-10-08	79
D100	chi2	Chi-squared distribution	kjpetrykowski	2020-10-13	204
D101	med	Median	JoramSoch	2020-10-15	72
D102	mode	Mode	JoramSoch	2020-10-15	72
D103	prob-exc	Exceedance probability	JoramSoch	2020-10-22	10
D104	dir-data	Dirichlet-distributed data	JoramSoch	2020-10-22	389
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	5
D107	min	Minimum	JoramSoch	2020-11-12	73
D108	max	Maximum	JoramSoch	2020-11-12	74
D109	rexp	Random experiment	JoramSoch	2020-11-19	2
D110	reve	Random event	JoramSoch	2020-11-19	3
D111	cvlme	Cross-validated log model evidence	JoramSoch	2020-11-19	420
D112	ind-cond	Conditional independence	JoramSoch	2020-11-19	8
D113	uplme	Uniform-prior log model evidence	JoramSoch	2020-11-25	420
D114	ebhme	Empirical Bayesian log model evidence	JoramSoch	2020-11-25	421

D115	vblme	Variational Bayesian log model evidence	JoramSoch	2020-11-25	422
D116	prior-flat	Flat, hard and soft prior distribution	JoramSoch	2020-12-02	125
D117	prior-uni	Uniform and non-uniform prior distribution	JoramSoch	2020-12-02	125
D118	prior-inf	Informative and non-informative prior distribution	JoramSoch	2020-12-02	125
D119	prior-emp	Empirical and theoretical prior distribution	JoramSoch	2020-12-02	126
D120	prior-conj	Conjugate and non-conjugate prior distribution	JoramSoch	2020-12-02	126
D121	prior-maxent	Maximum entropy prior distribution	JoramSoch	2020-12-02	126
D122	prior-eb	Empirical Bayes prior distribution	JoramSoch	2020-12-02	127
D123	prior-ref	Reference prior distribution	JoramSoch	2020-12-02	127
D124	ug	Univariate Gaussian	JoramSoch	2021-03-03	260
D125	h0	Null hypothesis	JoramSoch	2021-03-12	117
D126	h1	Alternative hypothesis	JoramSoch	2021-03-12	117
D127	hyp	Statistical hypothesis	JoramSoch	2021-03-19	115
D128	hyp-simp	Simple and composite hypothesis	JoramSoch	2021-03-19	115
D129	hyp-point	Point and set hypothesis	JoramSoch	2021-03-19	115
D130	test	Statistical hypothesis test	JoramSoch	2021-03-19	116
D131	tstat	Test statistic	JoramSoch	2021-03-19	118
D132	size	Size of a statistical test	JoramSoch	2021-03-19	118
D133	alpha	Significance level	JoramSoch	2021-03-19	119
D134	cval	Critical value	JoramSoch	2021-03-19	120
D135	pval	p-value	JoramSoch	2021-03-19	120
D136	ugkv	Univariate Gaussian with known variance	JoramSoch	2021-03-23	275
D137	power	Power of a statistical test	JoramSoch	2021-03-31	119
D138	hyp-tail	One-tailed and two-tailed hypothesis	JoramSoch	2021-03-31	116
D139	test-tail	One-tailed and two-tailed test	JoramSoch	2021-03-31	118

D140	dist-samp	Sampling distribution	JoramSoch	2021-03-31	18
D141	cdf-joint	Joint cumulative distribution function	JoramSoch	2020-04-07	35
D142	mean-samp	Sample mean	JoramSoch	2021-04-16	41
D143	var-samp	Sample variance	JoramSoch	2021-04-16	53
D144	cov-samp	Sample covariance	ciaranmci	2021-04-21	61
D145	prec	Precision	JoramSoch	2020-04-21	60
D146	f	F-distribution	JoramSoch	2020-04-21	207
D147	t	t-distribution	JoramSoch	2021-04-21	181
D148	mvt	Multivariate t-distribution	JoramSoch	2020-04-21	231
D149	eb	Empirical Bayes	JoramSoch	2021-04-29	129
D150	vb	Variational Bayes	JoramSoch	2021-04-29	130
D151	mome	Method-of-moments estimation	JoramSoch	2021-04-29	114
D152	nst	Non-standardized t-distribution	JoramSoch	2021-05-20	181
D153	covmat-samp	Sample covariance matrix	JoramSoch	2021-05-20	64
D154	mean-rvec	Expected value of a random vector	JoramSoch	2021-07-08	52
D155	mean-rmat	Expected value of a random matrix	JoramSoch	2021-07-08	53
D156	exc	Mutual exclusivity	JoramSoch	2021-07-23	10
D157	sun	Standard uniform distribution	JoramSoch	2021-07-23	146
D158	prob-ax	Kolmogorov axioms of probability	JoramSoch	2021-07-30	11
D159	cf	Characteristic function	JoramSoch	2021-09-22	36
D160	tglm	Transformed general linear model	JoramSoch	2021-10-21	358
D161	iglm	Inverse general linear model	JoramSoch	2021-10-21	361
D162	cfm	Corresponding forward model	JoramSoch	2021-10-21	364
D163	slr	Simple linear regression	JoramSoch	2021-10-27	296
D164	regline	Regression line	JoramSoch	2021-10-27	311
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	3

D168	corr-samp	Sample correlation coefficient	JoramSoch	2021-12-14	69
D169	corrmat-samp	Sample correlation matrix	JoramSoch	2021-12-14	71
D170	lognorm	Log-normal distribution	majapavlo	2022-02-07	??
D171	aicc	Corrected Akaike information criterion	JoramSoch	2022-02-11	??
D172	dev	Deviance	JoramSoch	2022-03-01	??
D173	mse	Mean squared error	JoramSoch	2022-03-27	??
D174	ci	Confidence interval	JoramSoch	2022-03-27	??

3 Proof by Topic

A

- Accuracy and complexity for the univariate Gaussian, 274
- Accuracy and complexity for the univariate Gaussian with known variance, 286
- Addition law of probability, 14
- Addition of the differential entropy upon multiplication with a constant, 88
- Addition of the differential entropy upon multiplication with invertible matrix, 89
- Additivity of the Kullback-Leibler divergence for independent distributions, 105
- Additivity of the variance for independent random variables, 59
- Akaike information criterion for multiple linear regression, ??

B

- Bayes' rule, 128
- Bayes' theorem, 128
- Bayesian information criterion for multiple linear regression, ??
- Bayesian model averaging in terms of log model evidences, 436
- Best linear unbiased estimator for the inverse general linear model, 362

C

- Characteristic function of a function of a random variable, 37
- Chi-squared distribution is a special case of gamma distribution, 204
- Concavity of the Shannon entropy, 81
- Conditional distributions of the multivariate normal distribution, 226
- Conditional distributions of the normal-gamma distribution, 242
- Conjugate prior distribution for Bayesian linear regression, 344
- Conjugate prior distribution for binomial observations, 393
- Conjugate prior distribution for multinomial observations, 397
- Conjugate prior distribution for multivariate Bayesian linear regression, 366
- Conjugate prior distribution for Poisson-distributed data, 375
- Conjugate prior distribution for the Poisson distribution with exposure values, 381
- Conjugate prior distribution for the univariate Gaussian, 266
- Conjugate prior distribution for the univariate Gaussian with known variance, 280
- Construction of confidence intervals using Wilks' theorem, 111
- Construction of unbiased estimator for variance, 408
- Convexity of the cross-entropy, 83
- Convexity of the Kullback-Leibler divergence, 105
- Corrected Akaike information criterion converges to uncorrected Akaike information criterion when infinite data are available, ??
- Corrected Akaike information criterion for multiple linear regression, ??
- Corrected Akaike information criterion in terms of maximum log-likelihood, ??
- Correlation always falls between -1 and +1, 68
- Correlation coefficient in terms of standard scores, 70
- Covariance of independent random variables, 62
- Cross-validated log Bayes factor for the univariate Gaussian with known variance, 293
- Cross-validated log model evidence for the univariate Gaussian with known variance, 291
- Cumulative distribution function in terms of probability density function of a continuous random variable, 32

- Cumulative distribution function in terms of probability mass function of a discrete random variable, 31
- Cumulative distribution function of a strictly decreasing function of a random variable, 30
- Cumulative distribution function of a strictly increasing function of a random variable, 30
- Cumulative distribution function of a sum of independent random variables, 29
- Cumulative distribution function of the beta distribution, 212
- Cumulative distribution function of the continuous uniform distribution, 147
- Cumulative distribution function of the discrete uniform distribution, 133
- Cumulative distribution function of the exponential distribution, 199
- Cumulative distribution function of the gamma distribution, 188
- Cumulative distribution function of the normal distribution, 162

D

- Derivation of Bayesian model averaging, 436
- Derivation of R^2 and adjusted R^2 , 410
- Derivation of the Bayesian information criterion, 415
- Derivation of the log Bayes factor, 426
- Derivation of the log family evidence, 423
- Derivation of the log model evidence, 418
- Derivation of the posterior model probability, 432
- Deviance for multiple linear regression, ??
- Deviance information criterion for multiple linear regression, ??
- Differential entropy can be negative, 86
- Differential entropy of the gamma distribution, 195
- Differential entropy of the multivariate normal distribution, 221
- Differential entropy of the normal distribution, 175
- Differential entropy of the normal-gamma distribution, 236
- Distribution of parameter estimates for simple linear regression, 307
- Distribution of the inverse general linear model, 361
- Distribution of the transformed general linear model, 359
- Distributional transformation using cumulative distribution function, 34

E

- Effects of mean-centering on parameter estimates for simple linear regression, 309
- Encompassing Prior Method for computing Bayes Factors, 430
- Equivalence of matrix-normal distribution and multivariate normal distribution, 251
- Equivalence of parameter estimates from the transformed general linear model, 360
- Exceedance probabilities for the Dirichlet distribution, 246
- Existence of a corresponding forward model, 365
- Expectation of a quadratic form, 48
- Expectation of parameter estimates for simple linear regression, 302
- Expectation of the cross-validated log Bayes factor for the univariate Gaussian with known variance, 294
- Expectation of the log Bayes factor for the univariate Gaussian with known variance, 289
- Expected value of a non-negative random variable, 41
- Expected value of the trace of a matrix, 48
- Expected value of x times $\ln(x)$ for a gamma distribution, 194
- Exponential distribution is a special case of gamma distribution, 198

- Expression of the cumulative distribution function of the normal distribution without the error function, 164
- Extreme points of the probability density function of the normal distribution, 173

F

- First central moment is zero, 78
- First raw moment is mean, 77
- Full width at half maximum for the normal distribution, 172

G

- Gaussian integral, 159
- Gibbs' inequality, 84

I

- Inflection points of the probability density function of the normal distribution, 174
- Invariance of the differential entropy under addition of a constant, 87
- Invariance of the Kullback-Leibler divergence under parameter transformation, 106
- Invariance of the variance under addition of a constant, 56
- Inverse transformation method using cumulative distribution function, 33

J

- Joint likelihood is the product of likelihood function and prior density, 122

K

- Kullback-Leibler divergence for the Dirichlet distribution, 245
- Kullback-Leibler divergence for the gamma distribution, 196
- Kullback-Leibler divergence for the matrix-normal distribution, 252
- Kullback-Leibler divergence for the multivariate normal distribution, 223
- Kullback-Leibler divergence for the normal distribution, 176
- Kullback-Leibler divergence for the normal-gamma distribution, 238
- Kullback-Leibler divergence for the Wishart distribution, 256

L

- Law of the unconscious statistician, 50
- Law of total covariance, 63
- Law of total expectation, 49
- Law of total probability, 15
- Law of total variance, 59
- Linear combination of independent normal random variables, 179
- Linear transformation theorem for the matrix-normal distribution, 253
- Linear transformation theorem for the moment-generating function, 38
- Linear transformation theorem for the multivariate normal distribution, 224
- Linearity of the expected value, 43
- Log Bayes factor for the univariate Gaussian with known variance, 288
- Log Bayes factor in terms of log model evidences, 426
- Log family evidences in terms of log model evidences, 424
- Log model evidence for Bayesian linear regression, 348
- Log model evidence for binomial observations, 395
- Log model evidence for multinomial observations, 399

- Log model evidence for multivariate Bayesian linear regression, 370
- Log model evidence for Poisson-distributed data, 377
- Log model evidence for the Poisson distribution with exposure values, 384
- Log model evidence for the univariate Gaussian, 271
- Log model evidence for the univariate Gaussian with known variance, 285
- Log model evidence in terms of prior and posterior distribution, ??
- Log sum inequality, 85
- Log-odds and probability in logistic regression, 403
- Logarithmic expectation of the gamma distribution, 192

M

- Marginal distributions of the multivariate normal distribution, 225
- Marginal distributions of the normal-gamma distribution, 239
- Marginal likelihood is a definite integral of joint likelihood, 124
- Maximum likelihood estimation can result in biased estimates, ??
- Maximum likelihood estimation for Dirichlet-distributed data, 389
- Maximum likelihood estimation for multiple linear regression, 342
- Maximum likelihood estimation for Poisson-distributed data, 373
- Maximum likelihood estimation for simple linear regression, 321
- Maximum likelihood estimation for simple linear regression, 324
- Maximum likelihood estimation for the general linear model, 356
- Maximum likelihood estimation for the Poisson distribution with exposure values, 379
- Maximum likelihood estimation for the univariate Gaussian, 260
- Maximum likelihood estimation for the univariate Gaussian with known variance, 276
- Maximum likelihood estimator of variance is biased, 406
- Maximum log-likelihood for multiple linear regression, ??
- Mean of the Bernoulli distribution, 135
- Mean of the beta distribution, 213
- Mean of the binomial distribution, 137
- Mean of the categorical distribution, 142
- Mean of the continuous uniform distribution, 149
- Mean of the exponential distribution, 201
- Mean of the gamma distribution, 190
- Mean of the multinomial distribution, 144
- Mean of the normal distribution, 167
- Mean of the normal-gamma distribution, 235
- Mean of the Poisson distribution, 139
- Mean of the Wald distribution, 218
- Median of the continuous uniform distribution, 151
- Median of the exponential distribution, 202
- Median of the log-normal distribution, ??
- Median of the normal distribution, 168
- Method of moments for beta-distributed data, 387
- Mode of the continuous uniform distribution, 151
- Mode of the exponential distribution, 203
- Mode of the log-normal distribution, ??
- Mode of the normal distribution, 169
- Moment in terms of moment-generating function, 75

- Moment-generating function of a function of a random variable, 38
- Moment-generating function of linear combination of independent random variables, 39
- Moment-generating function of the beta distribution, 211
- Moment-generating function of the normal distribution, 161
- Moment-generating function of the Wald distribution, 216
- Moments of the chi-squared distribution, 207
- Monotonicity of probability, 11
- Monotonicity of the expected value, 45

N

- Necessary and sufficient condition for independence of multivariate normal random variables, 230
- Non-invariance of the differential entropy under change of variables, 91
- (Non-)Multiplicativity of the expected value, 46
- Non-negativity of the expected value, 42
- Non-negativity of the Kullback-Leibler divergence, 102
- Non-negativity of the Kullback-Leibler divergence, 102
- Non-negativity of the Shannon entropy, 80
- Non-negativity of the variance, 54
- Non-symmetry of the Kullback-Leibler divergence, 103
- Normal distribution maximizes differential entropy for fixed variance, 178

O

- One-sample t-test for independent observations, 262
- One-sample z-test for independent observations, 277
- Ordinary least squares for multiple linear regression, 332
- Ordinary least squares for multiple linear regression, 333
- Ordinary least squares for simple linear regression, 298
- Ordinary least squares for simple linear regression, 300
- Ordinary least squares for the general linear model, 354

P

- Paired t-test for dependent observations, 265
- Paired z-test for dependent observations, 280
- Parameters of the corresponding forward model, 364
- Partition of a covariance matrix into expected values, 64
- Partition of covariance into expected values, 61
- Partition of sums of squares in ordinary least squares, 335
- Partition of the log model evidence into accuracy and complexity, 419
- Partition of the mean squared error into bias and variance, 110
- Partition of variance into expected values, 54
- Posterior credibility region against the omnibus null hypothesis for Bayesian linear regression, 352
- Posterior density is proportional to joint likelihood, 123
- Posterior distribution for Bayesian linear regression, 346
- Posterior distribution for binomial observations, 394
- Posterior distribution for multinomial observations, 398
- Posterior distribution for multivariate Bayesian linear regression, 368
- Posterior distribution for Poisson-distributed data, 376
- Posterior distribution for the Poisson distribution with exposure values, 383

- Posterior distribution for the univariate Gaussian, 268
- Posterior distribution for the univariate Gaussian with known variance, 282
- Posterior model probabilities in terms of Bayes factors, 432
- Posterior model probabilities in terms of log model evidences, 434
- Posterior model probability in terms of log Bayes factor, 433
- Posterior probability of the alternative hypothesis for Bayesian linear regression, 350
- Probability and log-odds in logistic regression, 402
- Probability density function is first derivative of cumulative distribution function, 28
- Probability density function of a linear function of a continuous random vector, 27
- Probability density function of a strictly decreasing function of a continuous random variable, 23
- Probability density function of a strictly increasing function of a continuous random variable, 22
- Probability density function of a sum of independent discrete random variables, 21
- Probability density function of an invertible function of a continuous random vector, 25
- Probability density function of the beta distribution, 210
- Probability density function of the chi-squared distribution, 205
- Probability density function of the continuous uniform distribution, 146
- Probability density function of the Dirichlet distribution, 244
- Probability density function of the exponential distribution, 199
- Probability density function of the F-distribution, 208
- Probability density function of the gamma distribution, 188
- Probability density function of the log-normal distribution, ??
- Probability density function of the matrix-normal distribution, 250
- Probability density function of the multivariate normal distribution, 221
- Probability density function of the normal distribution, 161
- Probability density function of the normal-gamma distribution, 234
- Probability density function of the t-distribution, 183
- Probability density function of the Wald distribution, 216
- Probability integral transform using cumulative distribution function, 33
- Probability mass function of a strictly decreasing function of a discrete random variable, 20
- Probability mass function of a strictly increasing function of a discrete random variable, 19
- Probability mass function of a sum of independent discrete random variables, 18
- Probability mass function of an invertible function of a random vector, 20
- Probability mass function of the Bernoulli distribution, 135
- Probability mass function of the binomial distribution, 136
- Probability mass function of the categorical distribution, 142
- Probability mass function of the discrete uniform distribution, 132
- Probability mass function of the multinomial distribution, 143
- Probability mass function of the Poisson distribution, 138
- Probability of exhaustive events, 16
- Probability of exhaustive events, ??
- Probability of the complement, 13
- Probability of the empty set, 12
- Probability under mutual exclusivity, 10
- Probability under statistical independence, 9
- Projection matrix and residual-forming matrix are idempotent, 339
- Projection of a data point to the regression line, 312

- Quantile function is inverse of strictly monotonically increasing cumulative distribution function, 35
- Quantile function of the continuous uniform distribution, 148
- Quantile function of the discrete uniform distribution, 134
- Quantile function of the exponential distribution, 200
- Quantile function of the gamma distribution, 189
- Quantile function of the normal distribution, 166

R

- Range of probability, 13
- Range of the variance of the Bernoulli distribution, ??
- Range of the variance of the binomial distribution, ??
- Relation of continuous Kullback-Leibler divergence to differential entropy, 108
- Relation of continuous mutual information to joint and conditional differential entropy, 100
- Relation of continuous mutual information to marginal and conditional differential entropy, 98
- Relation of continuous mutual information to marginal and joint differential entropy, 99
- Relation of Kullback-Leibler divergence to entropy, 107
- Relation of mutual information to joint and conditional entropy, 96
- Relation of mutual information to marginal and conditional entropy, 94
- Relation of mutual information to marginal and joint entropy, 95
- Relationship between coefficient of determination and correlation coefficient in simple linear regression, 330
- Relationship between correlation coefficient and slope estimate in simple linear regression, 329
- Relationship between covariance and correlation, 62
- Relationship between covariance matrix and correlation matrix, 65
- Relationship between gamma distribution and standard gamma distribution, 186
- Relationship between gamma distribution and standard gamma distribution, 187
- Relationship between multivariate t-distribution and F-distribution, 232
- Relationship between non-standardized t-distribution and t-distribution, 182
- Relationship between normal distribution and chi-squared distribution, 155
- Relationship between normal distribution and standard normal distribution, 152
- Relationship between normal distribution and standard normal distribution, 153
- Relationship between normal distribution and standard normal distribution, 154
- Relationship between normal distribution and t-distribution, 157
- Relationship between precision matrix and correlation matrix, 67
- Relationship between R^2 and maximum log-likelihood, 411
- Relationship between residual variance and sample variance in simple linear regression, 327
- Relationship between second raw moment, variance and mean, 77
- Relationship between signal-to-noise ratio and R^2 , 413

S

- Sampling from the matrix-normal distribution, 255
- Savage-Dickey Density Ratio for computing Bayes Factors, 428
- Scaling of the variance upon multiplication with a constant, 57
- Second central moment is variance, 78
- Simple linear regression is a special case of multiple linear regression, 297
- Sums of squares for simple linear regression, 313

T

- The p-value follows a uniform distribution under the null hypothesis, ??
- The regression line goes through the center of mass point, 311
- The residuals and the covariate are uncorrelated in simple linear regression, 326
- The sum of residuals is zero in simple linear regression, 325
- Transformation matrices for ordinary least squares, 337
- Transformation matrices for simple linear regression, 315
- Transitivity of Bayes Factors, 428
- Transposition of a matrix-normal random variable, 254
- Two-sample t-test for independent observations, 264
- Two-sample z-test for independent observations, 278

V

- Variance of constant is zero, 55
- Variance of parameter estimates for simple linear regression, 304
- Variance of the Bernoulli distribution, ??
- Variance of the beta distribution, 214
- Variance of the binomial distribution, ??
- Variance of the gamma distribution, 191
- Variance of the linear combination of two random variables, 58
- Variance of the normal distribution, 170
- Variance of the Poisson distribution, 140
- Variance of the sum of two random variables, 57
- Variance of the Wald distribution, 219

W

- Weighted least squares for multiple linear regression, 340
- Weighted least squares for multiple linear regression, 341
- Weighted least squares for simple linear regression, 318
- Weighted least squares for simple linear regression, 320
- Weighted least squares for the general linear model, 355

4 Definition by Topic

A

- Akaike information criterion, 415
- Alternative hypothesis, 117

B

- Bayes factor, 427
- Bayesian information criterion, 415
- Bayesian model averaging, 435
- Bernoulli distribution, 135
- Beta distribution, 210
- Beta-distributed data, 387
- Binomial distribution, 136
- Binomial observations, 393

C

- Categorical distribution, 142
- Central moment, 78
- Characteristic function, 36
- Chi-squared distribution, 204
- Coefficient of determination, 409
- Conditional differential entropy, 92
- Conditional entropy, 82
- Conditional independence, 8
- Conditional probability distribution, 17
- Confidence interval, ??
- Conjugate and non-conjugate prior distribution, 126
- Constant, 4
- Continuous uniform distribution, 146
- Corrected Akaike information criterion, ??
- Correlation, 68
- Correlation matrix, 70
- Corresponding forward model, 364
- Covariance, 60
- Covariance matrix, 63
- Critical value, 120
- Cross-entropy, 83
- Cross-validated log model evidence, 420
- Cumulant-generating function, 40
- Cumulative distribution function, 28

D

- Deviance, ??
- Deviance information criterion, 417
- Differential cross-entropy, 93
- Differential entropy, 86
- Dirichlet distribution, 244

- Dirichlet-distributed data, 389
- Discrete and continuous random variable, 5
- Discrete uniform distribution, 132

E

- Empirical and theoretical prior distribution, 126
- Empirical Bayes, 129
- Empirical Bayes prior distribution, 127
- Empirical Bayesian log model evidence, 421
- Encompassing model, 431
- Estimation matrix, 337
- Event space, 2
- Exceedance probability, 10
- Expected value, 41
- Expected value of a random matrix, 53
- Expected value of a random vector, 52
- Explained sum of squares, 334
- Exponential distribution, 198

F

- F-distribution, 207
- Flat, hard and soft prior distribution, 125
- Full probability model, 122
- Full width at half maximum, 73

G

- Gamma distribution, 185
- General linear model, 354
- Generative model, 121

I

- Informative and non-informative prior distribution, 125
- Inverse general linear model, 361

J

- Joint cumulative distribution function, 35
- Joint differential entropy, 93
- Joint entropy, 82
- Joint likelihood, 122
- Joint probability, 6
- Joint probability distribution, 17

K

- Kolmogorov axioms of probability, 11
- Kullback-Leibler divergence, 101

L

- Law of conditional probability, 6
- Law of marginal probability, 6

- Likelihood function, 121
- Likelihood function, 121
- Log Bayes factor, 425
- Log family evidence, 422
- Log model evidence, 418
- Log-likelihood function, 113
- Log-normal distribution, ??
- Logistic regression, 401

M

- Marginal likelihood, 124
- Marginal probability distribution, 17
- Matrix-normal distribution, 250
- Maximum, 74
- Maximum entropy prior distribution, 126
- Maximum likelihood estimation, 113
- Maximum log-likelihood, 114
- Mean squared error, ??
- Median, 72
- Method-of-moments estimation, 114
- Minimum, 73
- Mode, 72
- Moment, 74
- Moment-generating function, 37
- Multinomial distribution, 143
- Multinomial observations, 397
- Multiple linear regression, 331
- Multivariate normal distribution, 221
- Multivariate t-distribution, 231
- Mutual exclusivity, 10
- Mutual information, 97
- Mutual information, 97

N

- Non-standardized t-distribution, 181
- Normal distribution, 151
- Normal-gamma distribution, 233
- Null hypothesis, 117

O

- One-tailed and two-tailed hypothesis, 116
- One-tailed and two-tailed test, 118

P

- p-value, 120
- Point and set hypothesis, 115
- Poisson distribution, 138
- Poisson distribution with exposure values, 379

- Poisson-distributed data, 373
- Posterior distribution, 123
- Posterior model probability, 431
- Power of a statistical test, 119
- Precision, 60
- Precision matrix, 66
- Prior distribution, 121
- Probability, 5
- Probability density function, 21
- Probability distribution, 16
- Probability mass function, 18
- Probability space, 3
- Probability-generating function, 40
- Projection matrix, 337

Q

- Quantile function, 35

R

- Random event, 3
- Random experiment, 2
- Random matrix, 4
- Random variable, 3
- Random vector, 4
- Raw moment, 76
- Reference prior distribution, 127
- Regression line, 311
- Residual sum of squares, 334
- Residual variance, 406
- Residual-forming matrix, 337

S

- Sample correlation coefficient, 69
- Sample correlation matrix, 71
- Sample covariance, 61
- Sample covariance matrix, 64
- Sample mean, 41
- Sample space, 2
- Sample variance, 53
- Sampling distribution, 18
- Shannon entropy, 80
- Signal-to-noise ratio, 413
- Significance level, 119
- Simple and composite hypothesis, 115
- Simple linear regression, 296
- Size of a statistical test, 118
- Standard deviation, 73
- Standard gamma distribution, 185

- Standard normal distribution, 152
- Standard uniform distribution, 146
- Standardized moment, 79
- Statistical hypothesis, 115
- Statistical hypothesis test, 116
- Statistical independence, 7

T

- t-distribution, 181
- Test statistic, 118
- Total sum of squares, 334
- Transformed general linear model, 358

U

- Uniform and non-uniform prior distribution, 125
- Uniform-prior log model evidence, 420
- Univariate and multivariate random variable, 5
- Univariate Gaussian, 260
- Univariate Gaussian with known variance, 275

V

- Variance, 53
- Variational Bayes, 130
- Variational Bayesian log model evidence, 422

W

- Wald distribution, 216
- Wishart distribution, 256