

**Data Analysis –  
An Introduction to Parameter Estimation, Modeling,  
& Hypothesis Testing  
Lecture Notes**

**Dieter Wolf-Gladrow**  
Alfred Wegener Institute  
- Helmholtz Centre for Polar and Marine Research,  
Postfach 12 01 61  
D-27515 Bremerhaven, Federal Republic of Germany  
[Dieter.Wolf-Gladrow@awi.de](mailto:Dieter.Wolf-Gladrow@awi.de)

July 14, 2025



# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
1.1	Preface (short version) . . . . .	13
1.2	Example 1: Neutrinos . . . . .	14
1.2.1	A simple model for the observed frequency distribution: Poisson . . . . .	15
1.2.2	Estimate the mean rate of neutrino detection from data . . . . .	17
1.2.3	Comparison between relative frequencies and Poisson probabilities . . . . .	17
1.3	Probability . . . . .	20
1.3.1	Bayes' Theorem in the context of estimation . . . . .	22
1.4	Example 1 (neutrinos) revisited: Bayesian approach . . . . .	23
1.5	Example 2: Temperature at the freezing point? . . . . .	25
1.5.1	Null Hypothesis Significance Test (NHST): <i>t</i> -test . . . . .	28
1.5.2	Monte Carlo simulation: estimate <i>t</i> -distribution . . . . .	29
1.5.3	Rejection regions, <i>p</i> -value, & decision . . . . .	29
1.6	Bayesian <i>t</i> -test . . . . .	34
1.6.1	NHST versus Bayesian approach . . . . .	35
1.7	Example 3: straight line fitting . . . . .	36
1.8	Summary & Outlook . . . . .	39
<b>2</b>	<b>Look at your data: graphical analysis</b>	<b>41</b>
2.1	Scatterplots, histograms, box plots, transparent box plot . . . . .	42
2.1.1	Box-and-whisker plots ('box plots') . . . . .	46
2.2	Bubble plots . . . . .	48
2.3	Maps: image . . . . .	49
2.4	Heatmaps: 'plot a matrix' . . . . .	50
2.5	Exercises . . . . .	52
<b>3</b>	<b>Mean, variance, random sampling</b>	<b>53</b>
3.1	Central tendency: mean, median, ... . . . . .	53
3.1.1	Arithmetic mean or sample mean . . . . .	53
3.1.2	Median . . . . .	54
3.1.3	Geometric mean (*) . . . . .	54
3.1.4	Harmonic mean (*) . . . . .	54

3.1.5 Winsorized mean (*) . . . . .	54
3.2 Data dispersion: variance, standard deviation, MADN . . . . .	57
3.2.1 Sample variance . . . . .	57
3.2.2 Sample standard deviation . . . . .	57
3.2.3 MADN . . . . .	57
3.3 Covariance . . . . .	58
3.3.1 A simple example . . . . .	58
3.3.2 An example with negative covariance . . . . .	60
3.4 Anscombe's quartet . . . . .	61
3.4.1 Covariance matrix . . . . .	62
3.5 Correlation . . . . .	63
3.5.1 Pearson's correlation coefficient $r$ . . . . .	63
3.5.2 Correlation matrix . . . . .	63
3.5.3 Other correlation coefficients: Spearman, Kendall . . . . .	64
3.6 Degrees of freedom . . . . .	66
<b>4 Probabilities: concept, rules, assignment</b> . . . . .	<b>69</b>
4.1 The basic rules of probabilities . . . . .	70
4.1.1 The sum rule of probabilities . . . . .	70
4.1.2 The generalized sum rule of probabilities . . . . .	70
4.1.3 The product rule of probabilities . . . . .	72
4.1.4 Bayes' Theorem . . . . .	74
4.2 The Monty Hall problem . . . . .	75
4.2.1 Initial assignment of probabilities . . . . .	77
4.2.2 The host opens a door: assignment using additional information . . . . .	77
4.2.3 Apply Bayes' Theorem and marginalization . . . . .	79
4.2.4 Final result & discussion . . . . .	80
4.2.5 Take-home message: probability rules & Monty Hall . . . . .	81
4.3 Assigning probabilities . . . . .	82
4.4 The Principle of Indifference . . . . .	84
4.5 Maximum Entropy Principle . . . . .	85
4.5.1 Tossing a coin . . . . .	85
4.5.2 The unbiased die . . . . .	86
4.5.3 A loaded die . . . . .	87
4.5.4 The discrete exponential distribution . . . . .	90
4.6 Maximum Entropy Principle applied to PDFs . . . . .	90
<b>5 Random numbers</b> . . . . .	<b>93</b>
5.1 Generation of (pseudo-)random numbers in R . . . . .	93

<i>Contents</i>	3
<b>6 PDs &amp; PDFs</b>	<b>97</b>
6.1 Statistical populations: discrete versus continuous . . . . .	98
6.2 Expected value, expectation: mean & variance of PDs & PDFs . . . . .	101
6.2.1 Examples for mean values & variances (*) . . . . .	101
6.2.2 Expectation of the product of two stochastic variables . . . . .	102
6.2.3 How are expected values and sample mean & variance related to each other? . . . . .	103
6.2.4 Median of distributions (*) . . . . .	103
6.2.5 Mode . . . . .	103
6.3 Probability distributions: examples . . . . .	106
6.3.1 Discrete uniform distributions . . . . .	106
6.3.2 Binomial distribution . . . . .	107
6.3.3 Poisson distribution . . . . .	109
6.4 Probability Density Functions (PDFs): examples . . . . .	112
6.4.1 The normal distribution . . . . .	112
6.4.2 The uniform PDF . . . . .	120
6.4.3 Student's <i>t</i> -distribution . . . . .	122
6.4.4 The <i>F</i> distribution . . . . .	125
6.5 PDFs: change of variables . . . . .	128
6.6 Multivariate (joint) PDFs . . . . .	130
6.6.1 Constraints for multivariate PDFs . . . . .	130
6.6.2 Expectations, means, and variances . . . . .	131
6.6.3 Marginal PDFs . . . . .	131
<b>7 Central Limit Theorem (CLT)</b>	<b>133</b>
7.1 Normal distribution: sum of random numbers . . . . .	134
7.2 Central Limit Theorem: summary & take-home message . . . . .	137
<b>8 Uncertainty</b>	<b>139</b>
8.1 Propagation of uncertainty . . . . .	141
8.1.1 Law of propagation of uncertainty for $q = f(x)$ . . . . .	141
8.1.2 Law of propagation of uncertainty (general case) . . . . .	141
8.1.3 Law of propagation of uncertainty for $q = f(x, y)$ . . . . .	142
8.1.4 Propagation of uncertainty: Monte Carlo simulations . . . . .	143
<b>9 Likelihood and likelihood function</b>	<b>151</b>
9.1 Likelihood function for simple linear regression . . . . .	152
9.2 Likelihood function for the Poisson distribution . . . . .	154

<b>10 Point estimators</b>	<b>157</b>
10.1 A list of point estimators and their properties (for speed-readers) . . . . .	160
10.2 How good are estimates of the mean $\mu$ ? . . . . .	161
10.2.1 Estimate mean of standard normal distribution; standard error of the mean . . . . .	161
10.2.2 Abuse of the standard error or what's the size of the Emperor of China? . . . . .	162
10.3 Estimate the variance of the standard normal distribution . . . . .	163
10.3.1 Four different estimators of the variance . . . . .	165
10.4 Estimators for covariance & correlation . . . . .	169
10.4.1 Monte Carlo simulations . . . . .	170
10.5 Estimate rate parameter $\lambda$ of Poisson distribution . . . . .	175
10.6 Estimate of success in single trial (binomial distribution) . . . . .	178
10.7 Estimate mean of $F$ distribution (*) . . . . .	179
10.8 Robust estimators for central tendency & dispersion . . . . .	180
10.8.1 Robust estimation of the dispersion: MAD & MADN . . . . .	183
10.9 Estimate uncertainty via bootstrapping . . . . .	186
<b>11 Parameter estimation: the Bayesian approach</b>	<b>189</b>
11.1 Bayes' theorem in the context of parameter estimation . . . . .	190
11.2 Estimating the mean of a normal population with known variance . . . . .	191
11.3 Estimate the variance of a normal population for known mean value . . . . .	194
11.4 Conjugate distributions . . . . .	195
11.5 Marginal posterior distribution for $\mu$ : normal population, $\mu$ & $\sigma^2$ unknown . . . . .	197
11.6 Reporting the results of Bayesian parameter estimation . . . . .	200
11.7 Mean Squared Error (MSE) . . . . .	200
<b>12 Hypothesis testing</b>	<b>203</b>
12.1 Schools of thought: NHST versus Bayesian . . . . .	204
12.1.1 Example . . . . .	204
12.1.2 Sample mean, standard deviation, standard error of the mean . . . . .	205
12.1.3 Null Hypothesis Significance Testing (NHST): $t$ -test . . . . .	207
12.1.4 Bayesian $t$ -test . . . . .	210
12.1.5 Comparing $t$ -test and Bayesian $t$ -test . . . . .	213
12.2 Which test to apply for which question or hypothesis? . . . . .	227
12.3 Equal means? More than two samples: ANOVA . . . . .	229
12.3.1 ANOVA: NHST . . . . .	229
12.3.2 ANOVA: Bayesian . . . . .	233
12.4 Sample from a specified distribution? Normality tests etc. . . . .	234
12.4.1 Sample from normal distribution? The Shapiro-Wilk test . . . . .	235
12.4.2 R codes: Shapiro-Wilk, KS, and Lilliefors test . . . . .	236
12.4.3 Kolmogorov-Smirnov test . . . . .	236
12.4.4 Sample from normal distribution? The Lilliefors test . . . . .	240
12.4.5 Paul observed edelweiss . . . . .	243

<b>13 Beyond hypothesis testing: Bayesian inference</b>	<b>249</b>
13.1 Analysis based on summarized data . . . . .	250
13.1.1 Frequentistic approach: null hypothesis significance test (NHST) . . . . .	250
13.1.2 Why using lognormal PDFs for abundances? . . . . .	250
13.1.3 Frequentistic approach: lognormal confidence intervals . . . . .	252
13.1.4 Frequentistic approach: confidence interval for the difference . . . . .	254
13.1.5 Likelihood inference . . . . .	255
13.1.6 Likelihood inference: profile likelihood of the difference, $d$ . . . . .	256
13.1.7 Bayesian inference: joint posterior for flat prior . . . . .	257
13.1.8 Bayesian inference: marginal posteriors for flat prior . . . . .	258
13.1.9 Bayesian inference: change in abundance for flat prior . . . . .	259
13.1.10 Bayesian inference: informative priors . . . . .	260
13.2 Inference based on original data . . . . .	262
<b>14 Linear models: straight lines</b>	<b>265</b>
14.1 Fitting a straight line to an artificial data set: simple linear regression . . . . .	266
14.2 How to estimate intercept & slope: least squares . . . . .	272
14.2.1 Underlying assumptions for simple straight line fitting . . . . .	273
14.2.2 How to test the underlying assumptions? . . . . .	274
14.3 Confidence bands for simple linear regression . . . . .	278
14.3.1 Confidence interval for a single point . . . . .	278
14.3.2 Scheffé bands (hyperbolic) . . . . .	278
14.3.3 Gafarian bands (straight) . . . . .	280
14.3.4 Naive approximation . . . . .	280
14.4 Straight line through origin . . . . .	283
14.5 Which is the limiting nutrient? Redfield ratio . . . . .	284
14.5.1 World Ocean Atlas data . . . . .	284
14.5.2 Molar nitrate to phosphate ratio: simple linear regression . . . . .	287
14.6 Inverse prediction in the context of paleoproxies . . . . .	289
14.7 Fit exponential function . . . . .	290
14.8 Fit polynomial to data . . . . .	293
<b>15 Errors in <math>y</math> and <math>x</math>: errors in variables (EIV)</b>	<b>295</b>
15.1 Redfield ratios . . . . .	296
15.2 Errors in variables model (EVM) . . . . .	298
15.3 Regression and regression-based methods . . . . .	298
15.3.1 Simple linear regression: $y$ on $x$ . . . . .	299
15.3.2 Simple linear regression: via $x$ on $y$ . . . . .	300
15.3.3 The 'bisection' or 'double regression' method . . . . .	301
15.3.4 The 'geometric' line . . . . .	301
15.3.5 Comparison of bisection and geometric lines . . . . .	304
15.3.6 Summary: regression and regression-based methods . . . . .	305

<b>16 Multiple linear regression (MLR)</b>	<b>311</b>
16.1 Multiple linear regression: a simple example . . . . .	311
16.2 Multiple linear regression: a more difficult example . . . . .	314
16.2.1 Is less more? . . . . .	316
16.2.2 Fitting the noise? . . . . .	317
16.2.3 Akaike information criterion (AIC, AICc) and other information-theoretic approaches . . . . .	318
16.2.4 Dredging another example . . . . .	320
16.3 Wilkinson notation . . . . .	322
<b>17 Collinearity</b>	<b>323</b>
17.1 Effects of collinearity . . . . .	323
17.2 Acetylene data, choice of predictors, and unit length scaling . . . . .	326
17.2.1 The acetylene data . . . . .	326
17.2.2 Extended set of predictors: interaction and quadratic terms . . . . .	328
17.2.3 Unit length scaling . . . . .	329
17.3 The Webster et al. (1974) data . . . . .	332
17.4 Diagnostic of collinearity and multicollinearity . . . . .	334
17.4.1 Inspection of the correlation matrix . . . . .	334
17.4.2 Variance inflation factors . . . . .	334
17.4.3 Eigensystem analysis of the correlation matrix . . . . .	334
17.4.4 Singular value decomposition of the predictor matrix . . . . .	334
17.4.5 Summary . . . . .	335
17.5 Multiple linear regression of the acetylene data . . . . .	336
17.5.1 Multiple linear regression of the acetylene data (not scaled) . . . . .	336
17.5.2 Multiple linear regression of the scaled acetylene data . . . . .	337
17.5.3 Predicted response values . . . . .	337
17.5.4 Extrapolation . . . . .	340
17.5.5 Summary . . . . .	341
17.6 Ridge regression . . . . .	342
17.6.1 Ridge regression of the acetylene data: choice of predictors and algorithm . . . . .	342
17.6.2 How do the slopes change with $k$ ? Ridge trace . . . . .	343
17.6.3 Choice of $k$ , fit of predicted response data, and extrapolation . . . . .	345
17.6.4 Justification for ridge regression . . . . .	348
17.6.5 Summary . . . . .	348
<b>18 Linear modeling including factors</b>	<b>349</b>
18.1 Example 1: Parties and drinks . . . . .	349
18.2 Example 2: Sleep deprivation . . . . .	352
<b>19 Mixed effects models</b>	<b>357</b>
19.1 Sleep deprivation . . . . .	357

<i>Contents</i>	7
<b>20 Least squares for non-linear models</b>	<b>361</b>
20.1 Non-linear regression/least squares . . . . .	363
20.2 Lineweaver-Burk transformation . . . . .	367
<b>21 Stochastic count data: Poisson models</b>	<b>375</b>
21.1 Fatal horse kicks . . . . .	375
21.2 Estimating the R-value of the COVID-19 pandemic . . . . .	379
<b>22 Generalized Linear Modeling (GLM)</b>	<b>387</b>
22.1 Poisson regression . . . . .	388
22.1.1 Poisson regression: an example . . . . .	388
22.1.2 Poisson regression: Zuur et al. (2007) species richness data . . . . .	391
22.2 Logistic regression . . . . .	394
22.2.1 Logistic regression: an example using artificial data . . . . .	396
<b>A Probabilities</b>	<b>401</b>
A.1 Frequencies & probabilities: the law of large numbers . . . . .	401
A.1.1 Rolling a fair (unbiased) die . . . . .	401
A.2 Application of Bayes' Theorem: heroin addiction . . . . .	410
A.3 Calculation of the Lagrange multipliers for the loaded die . . . . .	411
A.3.1 Method 1: derive and solve the 'z-equation' . . . . .	411
A.3.2 Method 2: brute force numerical solution . . . . .	413
A.3.3 Method 3: iterative solution . . . . .	413
A.4 The loaded die once again . . . . .	416
A.5 Lagrange multipliers for the discrete exponential function . . . . .	417
A.6 More MaxEnt distributions . . . . .	417
A.7 Mean kinetic energy for particles with three possible speeds (MaxEnt) . . . . .	418
A.8 Derivation of the normal PDF by MaxEnt . . . . .	420
A.9 Relative entropy, cross-entropy, directed divergence (*) . . . . .	422
A.9.1 Equilibrium distributions for the D2Q9 lattice Boltzmann model (*) . . . . .	422
A.10 Probability for '4' in next throw of a die (*) . . . . .	427
A.10.1 Derive analytical solution . . . . .	427
A.10.2 A numerical example . . . . .	429
<b>B Random numbers</b>	<b>431</b>
B.1 Generation of random numbers from other distributions . . . . .	431
B.1.1 Example: random numbers from the tent distribution . . . . .	431
B.1.2 Random numbers from discrete distributions . . . . .	437

<b>C PDs &amp; PDFs (<sup>ref</sup>)</b>	<b>439</b>
C.1 Most common univariate distributions and some of their relationships . . . . .	439
C.2 Probability distributions (PDs) . . . . .	442
C.2.1 From the binomial to the Poisson distribution . . . . .	442
C.2.2 Zero Inflated Poisson distribution . . . . .	442
C.2.3 Zero truncated Poisson distribution . . . . .	444
C.2.4 Hypergeometric distribution: sampling without replacement . . . . .	445
C.2.5 Geometric probability distribution . . . . .	446
C.2.6 Broken stick distribution . . . . .	447
C.2.7 Negative binomial distribution . . . . .	447
C.3 Probability density functions: give me more . . . . .	451
C.3.1 Half-normal distribution . . . . .	451
C.3.2 The non-standardized Student's <i>t</i> -distribution . . . . .	453
C.3.3 Chi-squared ( $\chi^2$ ) distribution . . . . .	455
C.3.4 Cauchy distribution . . . . .	458
C.3.5 Scaled inverse $\chi^2$ & inverse-gamma distribution . . . . .	460
C.3.6 Kolmogorov-Smirnov CDF & PDF . . . . .	462
C.3.7 Inverse chi-squared ( $\chi^2$ ) distribution . . . . .	464
C.3.8 Lognormal PDF . . . . .	466
C.3.9 $\beta$ distribution . . . . .	470
C.3.10 Exponential PDF . . . . .	472
C.3.11 Gamma PDF: $\text{Gamma}(x; \alpha, \beta)$ . . . . .	473
C.3.12 Upper & lower incomplete $\Gamma$ function (*) . . . . .	474
C.3.13 Family of Student PDFs . . . . .	475
C.4 Joint PDFs: polynomial example (*) . . . . .	476
<b>D Uncertainty</b>	<b>479</b>
D.1 Paradigm shift . . . . .	479
D.2 Taylor & Kuyatt (1994) definitions . . . . .	480
<b>E Monte Carlo simulations</b>	<b>483</b>
E.1 Monte Carlo integration (*) . . . . .	483
E.1.1 Importance sampling . . . . .	487
E.2 Monte Carlo: CDF of normal PDF . . . . .	487
E.3 Estimate t distribution by Monte Carlo simulation (alternative code) . . . . .	489
<b>F Point estimators</b>	<b>491</b>
F.1 Estimate mean & variance (analytical results) . . . . .	491
F.1.1 Unbiased estimators for $\mu$ & $\sigma^2$ . . . . .	491
F.2 Estimators for the standard deviation (*) . . . . .	493
F.3 How to find point estimators? Examples . . . . .	494
F.3.1 Methods of moments . . . . .	494
F.3.2 Maximum likelihood estimators (MLEs) (*) . . . . .	497

<i>Contents</i>	9
<b>G Parameter estimation: the Bayesian approach (Appendix)</b>	<b>499</b>
G.1 Variance of a normal population for known mean value: conjugate prior (*) . . . . .	499
<b>H Hypothesis testing (Appendix)</b>	<b>501</b>
H.1 Fisher's scale of evidence against the null hypothesis . . . . .	501
H.2 Two-sample $t$ test . . . . .	501
H.3 Equal variances? ( <sup>ref</sup> ) . . . . .	503
H.3.1 Equal variances? The two-sample two-tailed variance ratio test . . . . .	503
H.3.2 Equal variances? Count data (*) . . . . .	507
H.3.3 Monte Carlo estimate of the density for the test statistic $F$ (*) . . . . .	508
H.3.4 Monte Carlo simulations: sampling from discrete uniform populations (*) . . . . .	510
H.3.5 Equal variances? Two or more samples: Levene test . . . . .	511
H.3.6 Equal variances? More than two samples: Bartlett test . . . . .	512
H.3.7 Equal variances? More than two samples: Fligner-Killeen test (*) . . . . .	512
H.4 Goodness-of-fit test: Mendelian factors? . . . . .	512
H.5 Correlation significantly different from zero? . . . . .	515
H.5.1 Does the test statistic $t_{\text{cor}}$ follow the $t$ -distribution? (*) . . . . .	518
H.6 Permutation test on correlation coefficients . . . . .	519
H.7 $\chi^2$ test . . . . .	521
H.8 Pedestrian way of KS-test (Edelweiss) (*) . . . . .	523
H.9 Fair/unbiased coin? Bayesian approach . . . . .	526
H.9.1 Slightly different derivation (*) . . . . .	530
H.10 Sample from Poisson or from zero inflated Poisson distribution? . . . . .	531
H.10.1 Zero inflated Poisson (ZIP) distributions . . . . .	534
H.10.2 $H_0, H_1$ , priors, marginal likelihoods, Bayes factor (*) . . . . .	536
H.10.3 Choice of priors (discussion) . . . . .	537
H.10.4 Formulation as an estimation problem . . . . .	538
H.10.5 Likelihoods of Poisson and zero inflated Poisson distributions (*) . . . . .	540
H.10.6 Marginal likelihoods for exponential prior (*) . . . . .	540
H.10.7 Bayes factor for exponential prior (*) . . . . .	541
H.10.8 Marginal likelihoods for Jeffreys' prior (*) . . . . .	541
H.10.9 Bayes factor for Jeffreys' prior (*) . . . . .	542
H.11 Bayesian $t$ -test: details (*) . . . . .	542
H.11.1 Likelihoods for $H_0$ and $H_1$ . . . . .	542
H.11.2 Marginal likelihood for $H_0$ using Jeffreys' prior . . . . .	543
H.11.3 Marginal likelihood for $H_1$ using Jeffreys' & Cauchy prior . . . . .	545
H.11.4 $\sigma^2$ known, no prior (Rouder et al., 2009) . . . . .	546
H.11.5 Normal prior for $\mu, \sigma^2$ known (Rouder et al., 2009) . . . . .	548
H.11.6 Unit-information prior (Rouder et al., 2009) . . . . .	550
H.12 Wilcoxon-Mann-Whitney test . . . . .	552

H.13 Wilcoxon paired-sample test . . . . .	556
H.13.1 Wilcoxon (1945) example . . . . .	556
H.13.2 Wilcoxon paired-sample test: Zar (2010, Example 9.4) . . . . .	557
H.14 Wilcoxon-Mann-Whitney test: calculation of probabilities (*) . . . . .	559
H.14.1 Normal approximation (*) . . . . .	559
H.14.2 Recurrence relations (*) . . . . .	559
H.14.3 Generating function (*) . . . . .	561
H.14.4 Monte Carlo (*) . . . . .	564
H.14.5 Critical values of the Wilcoxon paired-sample test (*) . . . . .	568
H.15 Zar (2010, Example 8.1): the <i>t</i> -test procedure . . . . .	570
H.16 The Lilliefors distribution: estimate CDF by Monte Carlo simulation . . . . .	572
H.17 Hypotheses testing: history . . . . .	575
<b>I Beyond hypothesis testing: Bayesian inference</b>	<b>577</b>
<b>J Straight line fitting</b>	<b>587</b>
J.1 Simple linear regression: Bayesian approach . . . . .	587
J.2 Proof of the Gauss-Markov theorem (*) . . . . .	593
J.3 From Bayes' Theorem to least-squares (*) . . . . .	595
J.4 Analytic solution of the least squares straight line problem (*) . . . . .	599
J.5 Fit polynomial to data . . . . .	600
<b>K Errors in variables (Appendix)</b>	<b>601</b>
K.1 Maximum likelihood estimation (MLE) aka Deming regression . . . . .	601
K.2 Markov Chain Monte Carlo (MCMC) . . . . .	602
K.2.1 Start values and prior parameters . . . . .	605
K.2.2 Results . . . . .	606
K.3 Artificial data sets: regressions & MCMC . . . . .	616
K.3.1 Start values and prior parameters for MCMC . . . . .	616
K.3.2 Results . . . . .	617
K.3.3 Discussion of MCMC results and summary . . . . .	626
K.4 Orthogonal linear regression . . . . .	627
K.5 Splitting methods . . . . .	630
K.5.1 Split up sorted data into two groups (Wald, 1940) . . . . .	630
K.5.2 Split up sorted data into three groups (Bartlett, 1949) . . . . .	630
K.5.3 Split up sorted data into three groups (Gibson & Jowett, 1957) . . . . .	631
K.5.4 Application of the splitting methods to an artificial data set . . . . .	631

<i>Contents</i>	11
<b>L Collinearity</b>	<b>635</b>
L.1 From slopes for scaled data to intercept and slopes for the original data (*)	635
L.2 Acetylene data: Montgomery & Peck (1982) choice of predictors	636
L.3 MLR of acetylene data	638
L.3.1 Extrapolation	639
L.4 MLR and ridge regression for Montgomery & Peck (1982) choice of predictors (*)	640
<b>M Priors (appendix)</b>	<b>645</b>
M.1 History of priors, especially ‘non-informative’	645
M.2 Numerical estimation of reference priors	646
M.2.1 Estimation of the rate of an exponential population	647
M.2.2 Numerical estimation of reference prior for exponential population	650
M.2.3 Estimation of the mode of an asymmetric triangular distribution	652
M.2.4 Numerical estimation of a reference prior for asymmetric triangular PDF	656
M.2.5 Analyzing the posterior and estimating the mode	657
<b>N R: tips &amp; tricks</b>	<b>663</b>
N.1 Communication with the operating system	663
N.2 Assign and print	663
N.3 Generate arrays when length is not known (‘dynamic array size’)	664
N.4 Missing functions/routines: <code>erf()</code> , <code>erf.inv()</code> , ...	664
N.5 Find indices of minimum values	664
N.6 Numerical integration in 1D: <code>integrate()</code>	664
N.7 95% intervals: <code>quantile()</code>	665
N.8 Plots	666
N.8.1 How to change position of axis labels	666
N.8.2 How get rid of axes, axes labels, and frame	667
N.8.3 Mathematical annotation	667
N.9 How to obtain specific output from <code>summary()</code>	668
N.10 How to get rid of NAs? <code>x[!is.na(x)]</code>	668
N.11 Data formats	668
N.11.1 Factors	668
N.11.2 Lists	668
N.11.3 Data frames	669
N.11.4 Application of data frames: ANOVA & Tukey HSD post-hoc test	669
N.12 Read from Excel file	669
N.13 Restrict samples by applying simple conditions	669
N.14 If ... else ...	670
N.15 Obtain source code of library routines	670

---

<b>O Notation &amp; Abbreviations</b>	<b>671</b>
O.1 Glossary . . . . .	673
<b>P Postface: long version of preface</b>	<b>677</b>

# Chapter 1

## Introduction

*In this introductory chapter a few examples will be discussed. Various methods will be applied and the corresponding concepts will be mentioned, however, without all details: try to get a smell! These examples, methods, concepts are presented also as motivation for the following chapters.*

### 1.1 Preface (short version)

*Who wants read a long preface where the author tells about the history how writing started, what topics are not covered and why, which text books he recommends, – and finally a long list of people in the acknowledgement? Those who want to know at least some of these details can find the long 'preface' in the Appendix (Chapter P).*

Instead of reading a long preface we will jump into the topic. Here is what you can expect in the introductory chapter:

1. We will start with an example from astrophysics: the data set is small and easy to visualize, we will formulate a few questions, develop a statistical model that can 'explain' the data, and estimate the model parameter.
2. Probability is the language in the land of uncertainty. The three basic rules of probabilities are presented and Bayes' Theorem is derived from the product rule.
3. Bayes' Theorem is applied in parameter estimation.
4. Hypotheses testing is in Example 2 and mention different schools of thought (Bayesian versus frequentist).
5. At the end of the chapter an overview about the other parts of the script/book will be given (Fig. 1.17).
6. Sections marked by a star in brackets, (\*), can be skipped in first or even second reading and might be for the mathematically inclined only.

**How to cite the script?** The script will be available on the Web. Please feel free to use and share it with friends and colleagues. For citation of the script I suggest the following format:

Wolf-Gladrow, D.A., *Data Analysis – An Introduction to Parameter Estimation, Modeling, & Hypothesis Testing – Lecture Notes*, version July 14, 2025. [available on the web]

## 1.2 Example 1: Neutrinos

In the following example from astrophysics<sup>1</sup> we will analyze a small data set, namely 10 pairs of values (Table 1.1). It allows us to introduce **estimation** and **modeling** based on the assumption of a **probability distribution**.

Neutrinos were detected by the Irvine-Michigan-Brookhaven experiment on 23 February 1987. The results (counts over 10 s intervals) are listed in Table 1.1. In most of the intervals (frequency = 1042) no neutrino was

NOE	0	1	2	3	4	5	6	7	8	9
Frequency	1042	860	307	78	15	3	0	0	0	1

Table 1.1: NOE = number of events (= neutrinos detected within 10 s intervals; Barlow, 1999). NOE = 4, frequency = 15  $\Leftrightarrow$  4 neutrinos were observed in 15 intervals of 10 s.

detected. However, in 860 intervals 1 neutrino was detected. The frequency decreases with increasing number of neutrinos and no detection of 6, 7, or 8 neutrinos is reported. However, there is one interval with 9 neutrinos which might look like an ‘outlier’ – not fitting in the decreasing trend. We will come back to this data point. In addition to inspection of data tables it is useful, especially for larger data sets, to visualize the data. A scatter plot of the data is shown in Fig. 1.1.

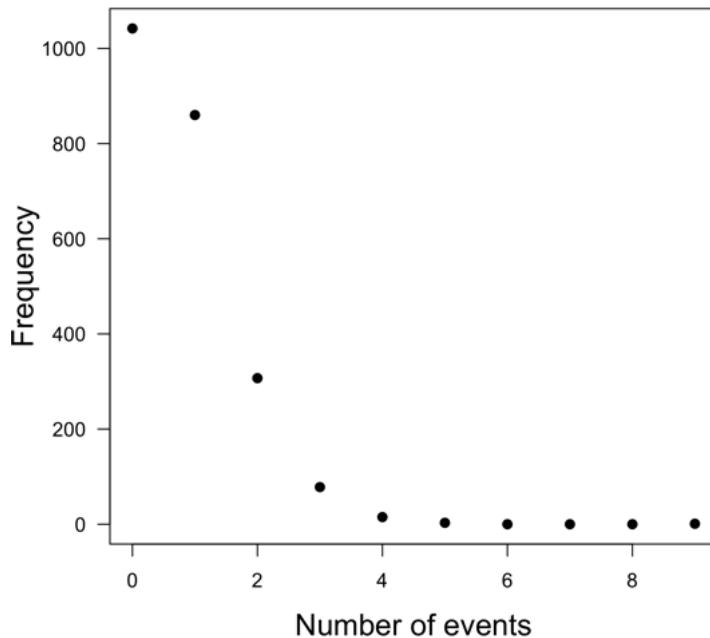


Figure 1.1: Neutrinos detected by the Irvine-Michigan-Brookhaven experiment on 23 February 1987 (compare Table 1.1). [NeutrinosData.R](#)

Observations are usually performed to answer certain questions. However, when data are available one could also ask: What else could I do with the data? Or ‘What else are the data telling us?’ although data are usually a bit shy, don’t talk to much, and thus *we* have to ask.

In the following we will address the following questions:

<sup>1</sup>Even if astrophysics is not your topic, it is highly recommended to follow this example. The necessary background information about neutrinos and supernovae is provided in an infobox (1).

1. Can we describe the observed frequency distribution by a simple model?
2. How to estimate the model parameter(s) from data?
3. Is the observation of 9 neutrinos in a single 10 s interval an outlier?

### 1: Neutrinos & supernova

**Neutrinos** are – still somewhat mysterious – electrically neutral ‘particles’ (excited states of quantum fields) with a mass much smaller than that of the electron. They were postulated by Wolfgang Pauli in 1930 to explain how beta decay could conserve energy, momentum, and spin. In order to distinguish it from another electrically neutral particle, the heavy neutron discovered by James Chadwick in 1932, the Italian physicist Eduardo Amaldi in a conversation with Enrico Fermi coined it neutrino, the Italian equivalent to ‘little neutral particle’. Neutrinos are only affected by the weak interaction, i.e. not by electromagnetic or strong forces, and gravitation can be neglected. They can easily traverse us or even the whole Earth. This makes it difficult to detect neutrinos and it took 26 years before Pauli’s proposition could be confirmed (Cowan et al., 1956). Neutrinos are generated as a side-product of fusion inside stars (including our Sun) and –rarely, however, then in great quantities – during explosions of stars (supernova).

A **supernova** is a luminous explosion of a massive star (according to Subrahmanyan Chandrasekhar more than 1.44 solar masses are required for normal type Ia supernova) during its last evolutionary stages. The peak optical luminosity of a supernova can be comparable to that of an entire galaxy and thus the explosion of far away stars can make them suddenly visible even to the naked eye: they appear as new (Latin: nova) stars. “The widely observed supernova SN 1054 produced the Crab Nebula. Supernovae SN 1572 and SN 1604, the latest to be observed with the naked eye in the Milky Way galaxy, had notable effects on the development of astronomy in Europe because they were used to argue against the Aristotelian idea that the universe beyond the Moon and planets was static and unchanging.” (Wikipedia, accessed 28 March 2022).

Still somewhat mysterious: it is not known whether the neutrino is identical with its antiparticle or not. Mass much smaller than that of the electron: The latest estimates for the electron neutrino are in the range of  $0.1 \text{ eV}/c^2$  or 5 million times smaller than that of the electron.

#### 1.2.1 A simple model for the observed frequency distribution: Poisson

The frequency distribution shown in Fig. 1.1 decreases monotonically with increasing number of neutrinos per interval. The decrease is not linear, i.e. a straight line would not be a ‘good fit’ and, anyway, continuous functions could at best yield envelopes to a discrete relationship.

Selecting an appropriate model requires some experience – and sometimes some creativity/imagination/fantasy. Our neutrino problem is a relative simple case. The detection of neutrinos follows a Poisson process (Infobox 2), i.e. it is a random process that can be characterized by a single parameter, namely the mean rate of detection,  $\lambda$ . By observation one obtains *count data*<sup>2</sup>  $k = 0, 1, 2, 3, \dots$ , i.e. non-negative integers, that appear to vary in a random way. The probabilities for the different outcomes are given by the Poisson distribution

$$p(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (1.1)$$

(Fig. 1.2). For samples from a Poisson process, the *relative frequencies*  $f_k = \frac{\text{frequency}_k}{\sum_j \text{frequency}_j}$  follow the same distribution except for deviations due to random sampling; the relative deviations can be large for small frequencies. The observed frequency distribution of detected neutrinos (Fig. 1.1) shows its maximum at zero neutrinos ( $k = 0$ ) and thus the mean rate  $\lambda$  of the Poisson model should be small ( $< 1$ ).

Before we continue we summarize our model assumptions:

1. Neutrinos stem from all kind of cosmic sources, especially from the interior of stars including our Sun.
2. The number of incoming neutrinos is approximately constant.

<sup>2</sup>Count data are quite common in ecology where one counts the number of individuals of certain animal or plant species.

## 2: Poisson process & distribution

A random process that yield non-negative integers (0, 1, 2, ...) with a constant mean rate (and no other regularities or constraints) is called a Poisson process. One example in physics is radioactive decay. "In counting the  $\alpha$  particles from radioactive substances either by the scintillation or electric method, it is observed that, while the average number of particles from a steady source is nearly constant, when a large number is counted, the number appearing in a given short interval is subject to large fluctuations. These variations are especially noticeable when only a few scintillations appear per minute. For example, during a considerable interval it may happen that no  $\alpha$  particle appears; then follows a group of  $\alpha$  particles in rapid succession; then an occasional  $\alpha$  particle, and so on." Rutherford et al. (1910). The probabilities for the different number  $k$  of  $\alpha$  particles per interval follows the Poisson distribution

$$\text{Poisson}(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (1.2)$$

where  $\lambda$  is the mean rate. What is the reason that the random ('fluctuating') outcome obeys such a relative simple probability distribution? Randomness is due to the fact that the unstable nuclei in Rutherford's radioactive material act independently from each other and the mean rate is approximately constant because only a tiny fraction of the unstable nuclei decay during an experiment (the most abundant polonium isotope  $^{210}\text{Po}$  possess a half-life time of 138 days).

3. Only a tiny fraction (factor  $10^{-15}$ ) of the incoming neutrinos is detected. Although the mean rate of detected neutrinos is constant, the small number of detected neutrinos in 10 s intervals varies in a random way.

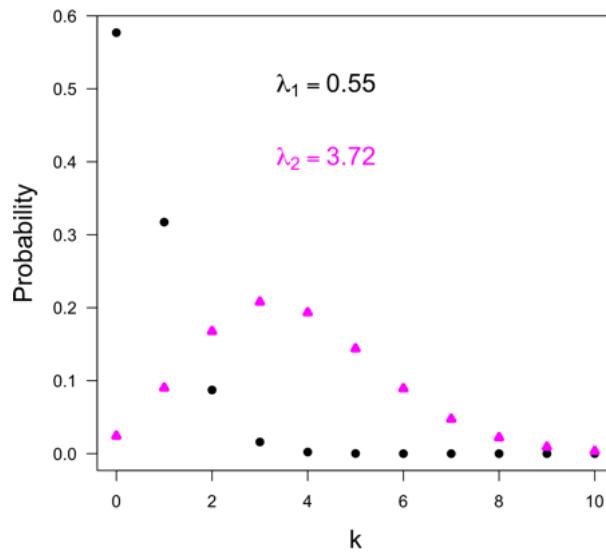


Figure 1.2: Two Poisson probability distributions: for small mean rates (here:  $\lambda_1 = 0.55$ , black dots) the maximum is located at  $k = 0$ ; for larger mean rates (here:  $\lambda_2 = 3.72$ , magenta triangles) the maximum is located at  $k > 0$ , actually close to the mean rate. [PDsPDFsPoisson2Examples.R](#)

### 1.2.2 Estimate the mean rate of neutrino detection from data

We now would like to estimate<sup>3</sup> the mean rate  $\lambda$  of our Poisson model from the observations. The Poisson distribution has the special property that  $\lambda$  is the expectation of  $k$  ('mean  $k'$ )

$$E[k] = \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \quad (1.3)$$

as well as the variance defined by

$$\text{Var} = E[(k - \lambda)^2] = \sum_{k=0}^{\infty} (k - \lambda)^2 p_k = \sum_{k=0}^{\infty} (k - \lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda. \quad (1.4)$$

Intuition would tell us that  $\lambda$  could be estimated by the sample mean or sample variance or by a combination of both. So called *point estimators* – procedures to estimate the value of single quantities of interest – will be discussed in detail in Chapter 10; for estimators of the mean rate see especially Section 10.5. Here we will apply the sample mean – replacing the probabilities  $p_k$  in Eq. 1.3 by the relative frequencies  $f_k / \sum_k f_k$  – because it can be shown that it is a better estimator than the sample variance. This yields the estimate

$$\hat{\lambda} = \frac{\sum_{k=0}^{\infty} k \cdot f_k}{\sum_{k=0}^{\infty} f_k} = \frac{\sum_{k=1}^9 k \cdot f_k}{\sum_{k=0}^9 f_k} = \frac{1792}{2306} = 0.777 \quad (1.5)$$

whereby the little hat on top of  $\lambda$  denotes an estimate.

### 1.2.3 Comparison between relative frequencies and Poisson probabilities

We can now compare the relative frequencies with the Poisson probabilities based on  $\hat{\lambda} = 0.777$  (Fig. 1.3): at least for small  $k$  the relative frequencies are close to the Poisson probabilities. For larger  $k$  the graphical comparison on the linear scale is inapt because both relative frequencies and probabilities are small and their difference cannot be seen. In order to make differences of the small values recognizable we switch to the logarithmic scale (Fig. 1.4). The Poisson probabilities for observing  $k = 6, 7$ , or  $8$  neutrinos within an 10 s interval is below  $1.5 \cdot 10^{-4}$  and no such case has been observed within 2306 intervals. For  $k = 9$  the Poisson probability is  $1.3 \cdot 10^{-7}$  whereas the relative frequency is 0.00043 or more than 3000 times larger than the Poisson probability. This large mismatch could be explained in two ways:

1. An error in the data.
2. A violation of one of the model assumptions.

When (1) can be ruled out (which is the case here) one has to search for a violation of the model assumptions.

Final remark:

“About  $10^{46}$  joules, approximately 10% of the star’s rest mass, is converted into a ten-second burst of neutrinos which is the main output of the event.” (Wikipedia, accessed 28 March 2022; based on Woosley & Janka, 2005)

Further reading: Loredo (1990), Loredo & Lamb (2002)

---

<sup>3</sup>In statistics ‘estimation’ is a technical term. An estimator is a procedure to calculate a value for the interesting quantity from the data.

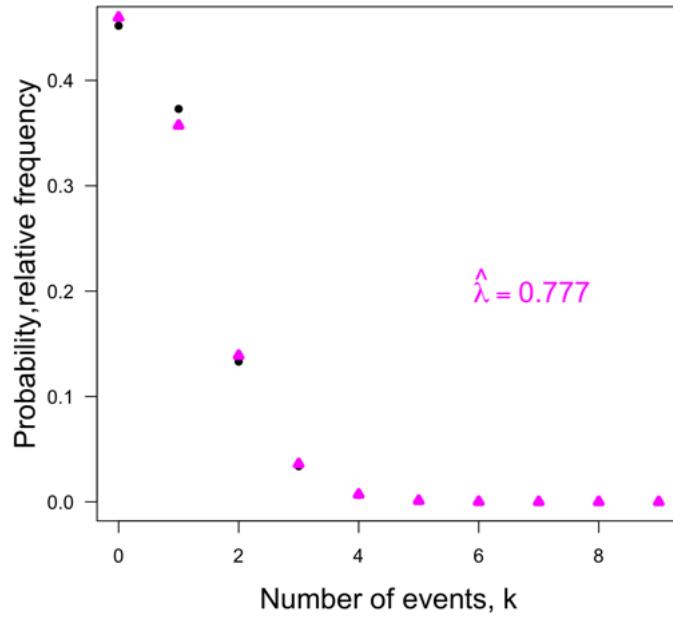


Figure 1.3: Relative frequencies (black dots) and Poisson probabilities (magenta triangles)  
[NeutrinosMeanPoisson.R](#)

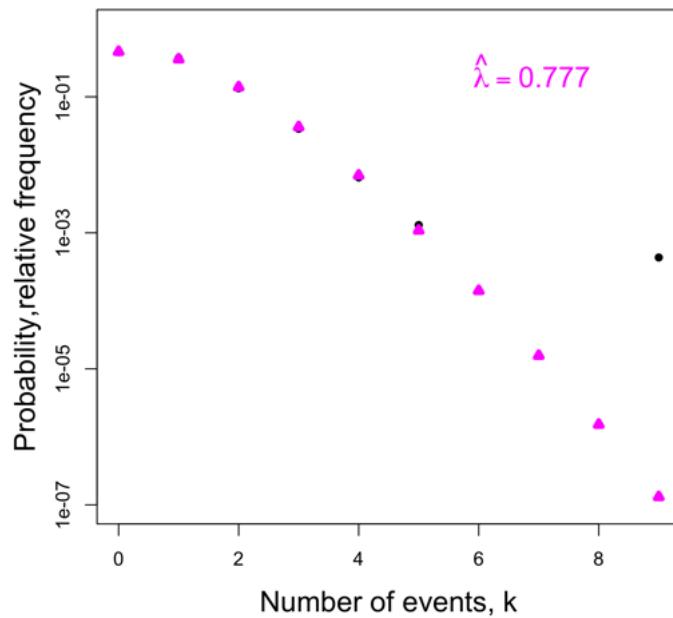


Figure 1.4: Relative frequencies (black dots) and Poisson probabilities (magenta triangles): logarithmic scale.  
[NeutrinosMeanPoissonLog.R](#)

**Music:**

Marteria (1982): Supernova

### 1.3 Probability

Probability is the native language in the land of uncertainty. We encountered probabilities already in Example 1 where we studied a model for the frequency distribution of detected neutrinos. Even if the mean rate of events would be constant – as in the case of neutrinos without the exception of a ‘close-by’ supernova or in other processes studied later-on (Section 21) – the actual observations would fluctuate. In the Bayesian approach, probability can be also used for single events when our information is limited. In this section we will just list some basic properties of probabilities, probability densities, give the basic rules of probabilities (how to ‘add’ and ‘multiply’ probabilities), derive Bayes’ Theorem, and show how this theorem can be used in parameter estimation and hypothesis testing. Different schools of thought with respect to the concept and applicability of probability (‘frequentist versus Bayesian’) and the problem of assignment of probabilities will be discussed in Chapter 4.

Probabilities,  $P_k$ , for certain events obey the following rules:

1.  $0 \leq P_k \leq 1$  non-negative, upper limit
2.  $\sum_k P_k = 1$  sum over all exclusive cases

Probability example 1: fair coin,  $P_1 = P_{\text{head}} = 1/2 = P_{\text{tail}} = P_2$  and  $\sum_k P_k = 1/2 + 1/2 = 1$

Probability example 2: unbiased die with 6 sides,  $P_k = 1/6$ ,  $k = 1, 2, \dots, 6$ ,  $\sum_k P_k = 6 \cdot 1/6 = 1$

Probability example 3: Poisson process,  $P_k(\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$ ,  $\sum_{k=0}^{\infty} P_k(\lambda) = 1$

Conditional probabilities: In the Bayesian approach all probabilities are conditional probabilities, i.e. their value depends on what we know. Thus for the fair coin it would be more precise to write

$$P(\text{head} | \text{fair coin}) = 1/2 \quad (1.6)$$

where our background knowledge (‘fair coin’) is written after the vertical line. All statements behind the vertical line are assumed to be true and the probability value is conditional to this. If the background information is obvious from the context, the notation is often simplified as in the three probability examples given above. In general expressions – like in the following three basic rules of probabilities – we will indicate the background information by a capital  $I$ .

Probabilities obey three rules:

1. Sum rule:

$$P(A|I) + P(\text{not } A|I) = 1 \quad (1.7)$$

the probability for proposition  $A$  (‘todays maximum temperature will be above 30°C’) plus the probability for proposition not  $A$  (‘todays maximum temperature will not be above 30°C’) is 1 (= 100% = certain).

2. Generalized sum rule:

$$P(A \text{ (inclusive)} \text{ or } B|I) = P(A|I) + P(B|I) - P(A \text{ and } B|I) \quad (1.8)$$

the probability that propositions  $A$  (‘todays maximum temperature will be above 25°C’) or (meant here inclusively, i.e. both  $A$  and  $B$  can be true)  $B$  (‘todays maximum temperature will be above 30°C’) is equal to the sum of the probabilities for  $A$  is true and for  $B$  is true minus the probability that  $A$  and  $B$  are true at the same time.

3. Product rule:

$$P(A \text{ and } B|I) = P(B|A \text{ and } I) P(A|I) = P(A|B \text{ and } I) P(B|I) \quad (1.9)$$

the probability for  $A$  and  $B$  to be true is equal to the product of the probability that  $B$  is true under the condition that  $A$  is true and the probability that  $A$  is true; because of the symmetry –  $A$  and  $B$  is true is the same as  $B$  and  $A$  is true –  $A$  and  $B$  can be interchanged.

**Bayes' Theorem:** From the symmetry of the product rule one immediately derives

$$P(B|A \text{ and } I) = \frac{P(A|B \text{ and } I) P(B|I)}{P(A|I)} \quad (1.10)$$

In Example 1 (neutrinos) we investigated [count data](#), that could take values  $0, 1, 2, 3, \dots$ , and described the frequency spectrum by a Poisson distribution. Probabilities are appropriate when the system is *discrete* (in contrast to *continuous*), i.e. when the interesting quantities are from a discrete set.<sup>4</sup> However, many interesting quantities, as for example, temperature are from a continuous set. To deal with such cases, *probability densities*<sup>5</sup> are introduced. Probabilities can be obtained from a probability density function by integration (Fig. 1.5).

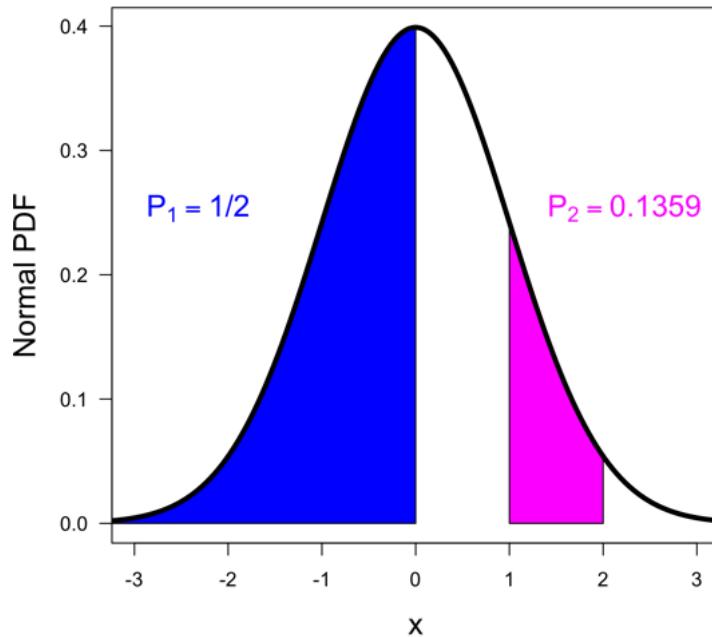


Figure 1.5: The standard normal probability density function (normal PDF, black solid line): obviously ('integration by eye'), half of the probability is located between  $x = -\infty$  and  $x = 0$  (blue area,  $P_1 = 1/2$ ). The probability between  $x = 1$  and  $x = 2$  is calculated by integration over the normal PDF:  $P_2 = \int_1^2 \text{Normal}(x) dx \approx 0.1359$ . [NormalPDFprobabilities.R](#)

The product rule and Bayes' Theorem apply also to probability densities (Zellner, 1971). Probability densities  $f(x)$  are  $\geq 0$ , however, in contrast to probabilities, they are not limited to 1. Instead of the normalization for probability distributions (PDs), probability density functions (PDFs) are normalized by integration

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (1.11)$$

<sup>4</sup>Not necessary integers; half-integers etc. would also do!

<sup>5</sup>There is an analogy to mass density which is a local property of a substance, as for example air. The density of air varies with height above sea level. In order to obtain the mass of an air column one has to integrate the variable mass density over the volume of the column.

### 1.3.1 Bayes' Theorem in the context of estimation

In the context of point estimation, one substitutes  $A = \text{data}$  (given) and  $B = \text{parameter}$  (to be estimated) to obtain

$$P(\text{parameter}|\text{data}) = \frac{P(\text{data}|\text{parameter}) \cdot P(\text{parameter}|I)}{P(\text{data}|I)} \quad (1.12)$$

In this context, the various terms have certain meanings and corresponding names:

- **posterior**  $P(\text{parameter}|\text{data})$ : PD or PDF for the parameter conditional on the given data ('in the light of data');
- **likelihood**  $P(\text{data}|\text{parameter})$ : 'how likely is it to observe the data when the parameter value is given (known)';
- **prior**  $P(\text{parameter}|I)$ : PD or PDF for parameter value before (prior to) looking (or even before sampling) the data;
- $P(\text{data}|I)$  would be a PD or PDF for observing the data before sampling the data and without any model which seems impossible; in the context of estimation, this term is considered as a constant that is used to normalize the posterior.

This leads to

$$\underbrace{P(\text{parameter}|\text{data})}_{\text{posterior}} \propto \underbrace{P(\text{data}|\text{parameter})}_{\text{likelihood}} \cdot \underbrace{P(\text{parameter}|I)}_{\text{prior}} \quad (1.13)$$

The last step – which is characteristic for the Bayesian approach – is a switch of perspective: after deriving the likelihood (a mathematical expression with data and parameter as 'arguments'), the mathematical expression is kept, however, the 'arguments' are interchanged. This leads to the **likelihood function**  $LF(\text{parameter}|\text{data})$  where the parameter value is considered as conditional on the given data. Thus finally one obtains

$$\underbrace{P(\text{parameter}|\text{data})}_{\text{posterior}} \propto \underbrace{LF(\text{parameter}|\text{data})}_{\text{likelihood function}} \cdot \underbrace{P(\text{parameter}|I)}_{\text{prior}} \quad (1.14)$$

The result of Bayesian parameter estimation is the posterior: a whole PD or PDF for the parameter instead of a single value for the estimate. From the posterior one can calculate a measure for the central tendency as, for example, the mean (expectation), mode, or median of the posterior, and a standard deviation or a 95% credible set as measures for the parameter uncertainty. We will revisit Example 1 (neutrinos) to illustrate the Bayesian approach and to estimate the mean rate parameter and its uncertainty.

Later-on we will encounter models with more than one parameter. The estimation formula Eq. (1.14) can be easily generalized by introducing the plural:

$$\underbrace{P(\text{parameters}|\text{data})}_{\text{posterior}} \propto \underbrace{LF(\text{parameters}|\text{data})}_{\text{likelihood function}} \cdot \underbrace{P(\text{parameters}|I)}_{\text{prior}}. \quad (1.15)$$

However, we will see later that not only the computational demand is larger but also properties of the data (for example, collinearity in multivariate data sets) can cause problems.

## 1.4 Example 1 (neutrinos) revisited: Bayesian approach

In order to estimate the uncertainty of the mean rate parameter  $\lambda$  we will apply the Bayesian approach to parameter estimation.

The result of the Bayesian point estimation is the posterior for the parameter (here:  $\lambda$ ). The posterior is proportional to the likelihood function and the prior. Thus we have to assign a prior and to formulate the likelihood function based on the data and our model (here: Poisson).

The prior expresses our prior information – or often ignorance – about the parameter. Because we don't possess any prior information we apply a non-informative prior by using a constant as prior – a so-called 'flat' prior.<sup>6</sup>

The likelihood for a single data point  $x_i$  is given by the Poisson model

$$L(x_i|\lambda) = P(x_i|\lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \quad (1.16)$$

The likelihood for observing two different data points that are independent from each other is equal to the product of the two single point likelihoods:

$$L(x_i, x_j|\lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \frac{\lambda^{x_j}}{x_j!} e^{-\lambda} = \frac{\lambda^{x_i+x_j}}{x_i! x_j!} e^{-2\lambda} \quad (1.17)$$

This can be generalized for  $n$  independent observations

$$L(x_1, x_2, \dots, x_n|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda} = \frac{\lambda^s}{\prod_{i=1}^n x_i!} e^{-n\lambda} \quad (1.18)$$

where  $n = \sum_{i=0}^m f_i$  (here:  $m = 9$ ) is the total number of observed frequencies  $f_i$  with  $s = \sum_{i=0}^m i \cdot f_i = \sum_{i=1}^m i \cdot f_i$ . Thus the likelihood function for  $\lambda$  reads

$$LF(\lambda|x_1, x_2, \dots, x_n) = \frac{\lambda^s}{\prod_{i=1}^n x_i!} e^{-n\lambda} \quad (1.19)$$

This function has to be normalized with respect to  $\lambda$ , i.e. one has to find the constant  $c$  such that

$$c \int_0^\infty LF(\lambda|x_1, x_2, \dots, x_n) d\lambda = c \int_0^\infty \frac{\lambda^s}{\prod_{i=1}^n x_i!} e^{-n\lambda} d\lambda = \frac{c}{\prod_{i=1}^n x_i!} \int_0^\infty \lambda^s e^{-n\lambda} d\lambda = 1 \quad (1.20)$$

Instead of performing the integration one can try to find a PDF that has the desired dependency  $\lambda^s e^{-n\lambda}$ . The gamma PDF (Section C.3.11)

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (1.21)$$

possess this dependency and thus – with  $\alpha = s + 1$ ,  $\beta = n$  – is the desired posterior (Fig. 1.6). The central tendency can be characterized by the mean (expectation of  $\lambda$ ) of the posterior:

$$\hat{\lambda} = \frac{\alpha}{\beta} = \frac{s+1}{n} = \frac{s+1}{n} \pm \frac{\sqrt{s+1}}{n} = \frac{1793}{2306} \pm \frac{\sqrt{1793}}{2306} = 0.778 \pm 0.018 \quad (1.22)$$

Please note that the Bayesian estimate of the central tendency based on flat prior,  $\hat{\lambda}_B = \frac{s+1}{n}$ , is always larger than the estimate using the sample mean,  $\hat{\lambda}_{\text{sample mean}} = \frac{s}{n}$ . However, for large sample sizes ( $s$  and  $n \gg 1$ ) the difference is small or tiny.

<sup>6</sup>A prior  $f(\theta) = \text{constant}$  over an infinite range from 0 to  $\infty$  can not be normalized and thus the constant can not be specified. Fortunately, the specification of this constant is not required because in the end the posterior has to be normalized anyway. Priors that can not be normalized are called *improper*. Improper priors may work (as in the current example), however, in some cases might lead to posteriors that can not be normalized.

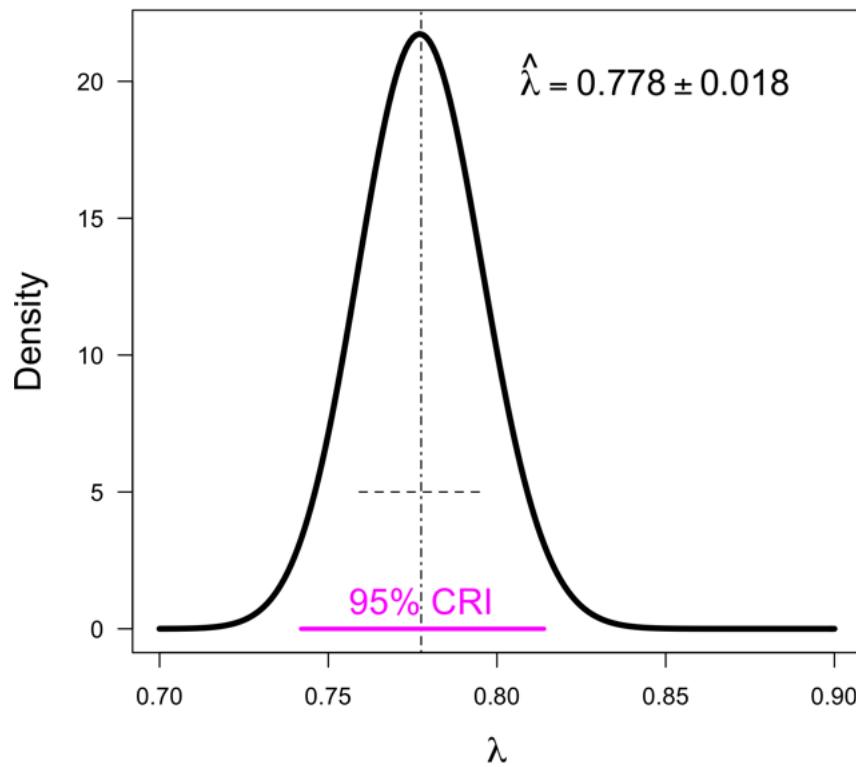


Figure 1.6: The posterior for the neutrino problem is the gamma PDF (black solid line) with  $\alpha = 1793$  and  $\beta = 2306$ . The central tendency can be characterized by the mean  $\alpha/\beta = 0.778$  (dash-dotted vertical line) with a standard deviation of  $\sqrt{\alpha}/\beta = 0.018$ . The 95% credibility interval (CRI) ranges from 0.742 to 0.814 (magenta horizontal line). [NeutrinosBayesianFlat.R](#)

## 1.5 Example 2: Temperature at the freezing point?

After modeling and estimation (Example 1), hypotheses testing will be introduced here using a small set of temperature data. We want to know whether the true temperature is at a certain value, namely the freezing point of water at  $0^{\circ}\text{C}$ . Hypotheses testing is still a minefield because different approaches (still) exist. In the – in the 20th century most commonly applied – Null Hypothesis Significance Test (NHST) a level of significance ( $p$ -value) is calculated for a null hypothesis  $H_0$  and  $H_0$  can be rejected or not based on the  $p$ -value. This approach has been criticized from the beginning and is now even banished from a number of scientific journals (for example, *Basic and Applied Social Psychology*; Trafimow, 2014; Trafimow & Marks, 2015). But it's still in use,  $p$ -value show up again and again, and thus one should know how it works. In the Bayesian approach to hypotheses testing two hypotheses (null and alternative or working hypothesis) are formulated and the probabilities for both are compared to each other.

"Statisticians tend to use graphs, but non-statisticians seem to prefer tests."

Zuur et al. (2009, p. 543)

The following temperatures ( $^{\circ}\text{C}$ ) have been measured

$$x = \{1.5, 0.3, 1.8, -1.4, 0.8, 3.0, -0.3, 0.2, -0.4, 1.9, 0.0, 0.3, -1.0, 1.2, 3.8, 0.5, -0.8, 2.0, 1.1, 1.2, -0.4, 2.7, 0.5, -1.4, 1.1\}$$

We want to know whether or not the temperature is at the freezing point of water at  $0^{\circ}\text{C}$ . Before doing any calculations let's look at the data (Fig. 1.7).

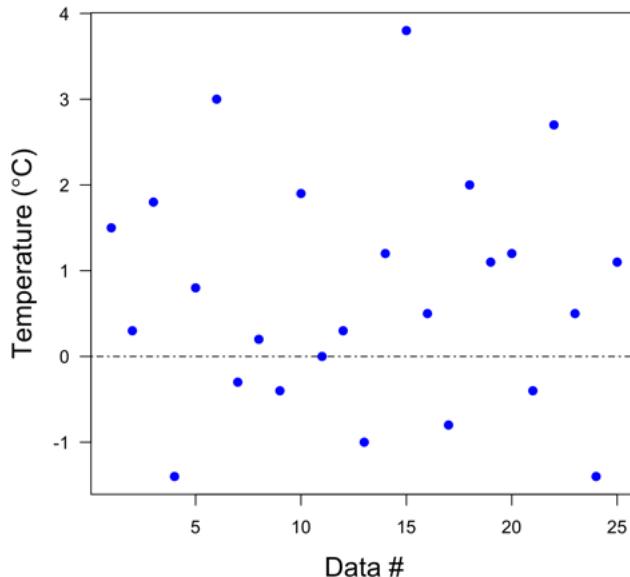


Figure 1.7: Observed temperatures (blue dots) and hypothesized true mean  $\mu_0 = 0^{\circ}\text{C}$  (black dash-dotted line). The hypothesized temperature  $\mu_0 = 0^{\circ}\text{C}$  lies inside the cloud of data, however, more data lie above  $\mu_0 = 0^{\circ}\text{C}$  than below which might speak against  $\mu = \mu_0$  where  $\mu$  is the true mean of the population from which we sampled. [BayesianHyp-t-test-Data.R](#)

As the next step of analysis we calculate the sample mean  $\bar{x} = 0.73^\circ\text{C}$ . The sample mean is an *unbiased estimator of the true mean  $\mu$*  (estimators will be presented in Chapter 10):

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0.73^\circ\text{C} \quad (1.23)$$

where the little hat  $\hat{\cdot}$  atop  $\mu$  stands for 'estimate'.

Can we answer our question based on the estimate of  $\mu$ ? No, because the sample mean is almost always different from the true mean. What is missing is an estimate of the uncertainty of our estimate or of the uncertainty of the difference between estimated and hypothesized mean.

How to measure uncertainties? The observed data show quite a bit of spread (Fig. 1.7) that can be measured by the standard deviation,  $\sigma$ , or by the variance,  $\sigma^2$ . The true variance  $\sigma^2$  can be estimated by the sample variance  $s^2$ :

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.24)$$

However, this is a measure for the variance of the population from which we sample and not the uncertainty in our estimate for the mean. With increasing sample size  $n$ ,  $\hat{\sigma}^2$  will vary only slightly (random sample!) but stay more or less at a certain level (namely close to the true variance  $\sigma^2$ ). We have the expectation (from common sense) that the uncertainty of the mean should decrease with increasing sample size. This is indeed the case, as will be shown in a second.

When estimating the population mean by calculating the sample mean, one adds up random values from a statistical population. These values are usually different from the true mean  $\mu$ , some are smaller, some are larger. By adding up these values with negative or positive deviations from the true mean the deviations partially compensate each other. This compensation works better and better with increasing sample size. One can show<sup>7</sup> that the uncertainty in the estimate of the mean can be estimated by  $\sigma/\sqrt{n}$ , i.e. it decreases with one over the square of the sample size. The quantity  $s/\sqrt{n}$  ( $s$  is the estimate of  $\sigma$ ) is called the **standard error of the mean (SE)**; it can be calculated from the data. For our temperature data one obtains  $\text{SE} \approx 0.27^\circ\text{C}$ . We will add lines for  $\bar{x}$  (blue solid line) and  $\bar{x} \pm \text{SE}$  (red dashed lines) to our data plot (Fig. 1.8). The hypothesized  $\mu_0 = 0^\circ\text{C}$  lies clearly outside the range  $\bar{x} \pm \text{SE}$ , actually the difference  $\bar{x} - \mu_0$  is about 2.7 SE. Physicists would report a difference of 2.7 standard errors (or in their notation 2.7  $\sigma$  where  $\sigma$  stands for SE) and might reject the hypothesis  $\mu = \mu_0$  when the difference is larger than 2 SE. Depending on what is at stake they shift this rejection boundary to 3 or even 5 SE. This finishes already our first approach to answering our initial question. This simple estimation approach works because we know how the uncertainty falls off with increasing sample size. It is remarkable that this estimate  $s/\sqrt{n}$  applies to samples from all kind of distributions as long as their variances are finite (Casella & Berger, 2002, p. 213–214, Theorem 5.2.6).

---

<sup>7</sup>Casella & Berger, 2002, p. 213–214, Theorem 5.2.6, for any (not just normal!) distribution with finite variance, i.e.  $\sigma^2 < \infty$ ; it does not apply, for example, to the Cauchy distribution. Compare Section F.1 for details.

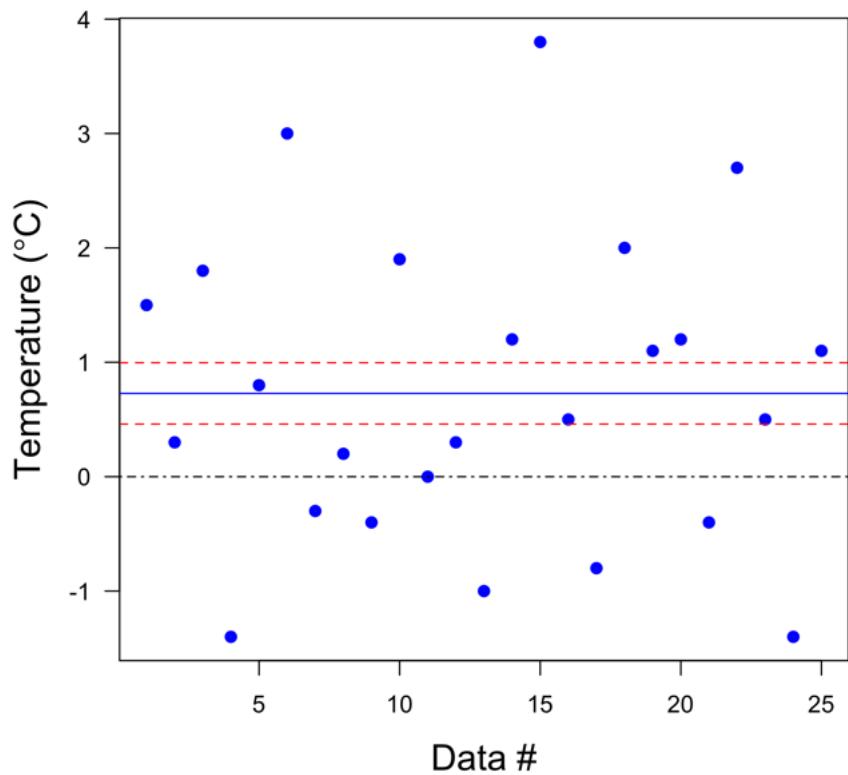


Figure 1.8: Observed temperatures (blue dots) and hypothesized mean  $\mu_0 = 0^\circ\text{C}$  (black dash-dotted line); the sample mean  $\bar{x}$  at  $0.73^\circ\text{C}$  (blue line) and the sample mean  $\pm$  one standard error of the mean (red dashed lines). The hypothesized  $\mu_0 = 0^\circ\text{C}$  lies clearly outside the range  $\bar{x} \pm \text{SE}$ , actually the difference  $|\bar{x} - \mu_0|$  is about  $2.7 \text{ SE}$ .

[Bayesian-t-test-SE.R](#)

### 1.5.1 Null Hypothesis Significance Test (NHST): *t*-test

Applying the *t*-test in R is easy:

```
t.test(x)
```

The output reads

```
One Sample t-test
```

```
data: x
t = 2.7128, df = 24, p-value = 0.01215
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.1741312 1.2818688
sample estimates: mean of x 0.728
```

Explanation of the output:

1. 'One Sample t-test': yes, we use one sample only and apply the *t*-test
2. 'data: x': the data set used is denoted 'x'
3. 't = 2.7128': *t* is the so-called *test statistic*, its value 2.7128 has been calculated from the data
4. 'df = 24': 'df' stands for *degrees of freedom*; its value 24 is given by the number of data (25) minus 1 because one degree of freedom is 'used up' in calculating the test statistic *t* (compare Section 3.6 for details)
5. 'p-value = 0.01215': this is the famous *p*-value or - with its civil name – the *observed level of significance* (don't confuse with the *chosen level of significance*,  $\alpha$ , see below).
6. 'alternative hypothesis: true mean is not equal to 0': although in NHST one is testing the null hypothesis  $H_0$  'true mean is equal to 0', the output is given us the alternative hypothesis which has not been tested; this is a bit strange, to say the least.
7. '95 percent confidence interval: 0.1741312 1.2818688': although we did not ask for it, the *t*-test routine calculates an 95% *confidence interval* (CI) of  $0.17 \leq \mu \leq 1.28$  with the (frequentist) meaning that if we would sample again from the same statistical population in 95% of the cases the estimated interval would include the true mean value  $\mu$ ; please note that this is a probability for the interval and not for  $\mu$ , i.e. it does not imply that  $\mu$  lies with 95% probability within the confidence interval. The latter statement will be addressed by the *credibility set* in the Bayesian approach to hypotheses testing.
8. 'sample estimates: mean of x 0.728': the sample mean of *x* is  $\bar{x} = 0.728$ .

In the following we will explain how to calculate *t*, *p*-value, and the limits of the 95% confidence interval from the data and the assumptions made in the null hypothesis  $H_0$ .

In contrast to our simple question 'Is the temperature at the freezing point?' the null hypothesis in NHST is more specific in that it includes an assumption about the statistical population from which we take a random sample:

**Null hypothesis  $H_0$ :** the data *x* have been randomly sampled from a normal distribution with true mean  $\mu = 0$  and unknown standard deviation  $\sigma$ .

Why does one need this assumption about the population?

Answer: in order to calculate

- (1) the distribution of the test statistic *t* when  $H_0$  is true (Student's *t*-distribution) and
- (2) the *p*-value.

**Test statistic:** The test statistic should allow us to infer something about  $H_0$ . It is often derived by intuition and seen as a plausible choice.<sup>8</sup> In the case of the one-sample  $t$ -test the test statistic is called  $t$  and defined by

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{0.73 - 0}{\frac{1.34}{\sqrt{25}}} = 2.713 = t_{\text{obs}} \quad (1.25)$$

i.e. it is the deviation of the estimated mean  $\bar{x}$  from the hypothesized true mean  $\mu = 0$  weighted by 1 over the estimated standard error of the mean. Large deviations between the estimated and hypothesized true mean would speak against  $H_0$ , however, one has to measure this difference in units of the estimated standard error. The result is identical to the output of the `t.test()` routine.

Our next task is to calculate the  $p$ -value. For this purpose one has to know how the test statistic  $t$  is distributed when  $H_0$  is true. Although an analytic expression for the  $t$ -distribution is available, we will estimate this distribution by a *Monte Carlo simulation*, a method that can be also applied when analytic expressions are not available and that will be used again and again in other chapters.

### 1.5.2 Monte Carlo simulation: estimate $t$ -distribution

In the following Monte Carlo calculation the random sampling from a normal distribution with mean  $\mu = 0$  and a fixed but arbitrary (in the real problem unknown) standard deviation  $\sigma$  is simulated by

1. taking many times (for example  $M = 1000$  times) random samples of size  $n = 25$  (equal to the sample size in the data set of Example 2);
2. calculating the test statistic  $t$  for each sample;
3. plotting the values of the test statistic  $t$  in the form of a histogram (Fig. 1.9);
4. and finally estimating the *density* (short for probability density function, PDF; Fig. 1.10) from the Monte Carlo  $t$  values.

The Monte Carlo simulation gives us an idea where the Student- $t$  or, short,  $t$  distribution comes from. Many other PDFs and CDFs stem from other hypotheses tests and their associated test statistics. Some of these distributions have been derived in analytic form by mathematicians – as, for example, the F and  $\chi^2$  distributions – whereas others have to be estimated by numerical methods, especially Monte Carlo methods – as, for example, the distributions for the Kolmogorov-Smirnov and for the Lilliefors test.

### 1.5.3 Rejection regions, $p$ -value, & decision

From now on we will proceed with the analytic expression for the  $t$ -distribution.

---

<sup>8</sup>Well chosen test statistics have desirable properties such as ‘sufficiency’, i.e. they encompass all essential information contained in the sample (for details compare, for example, Casella & Berger, 2002).

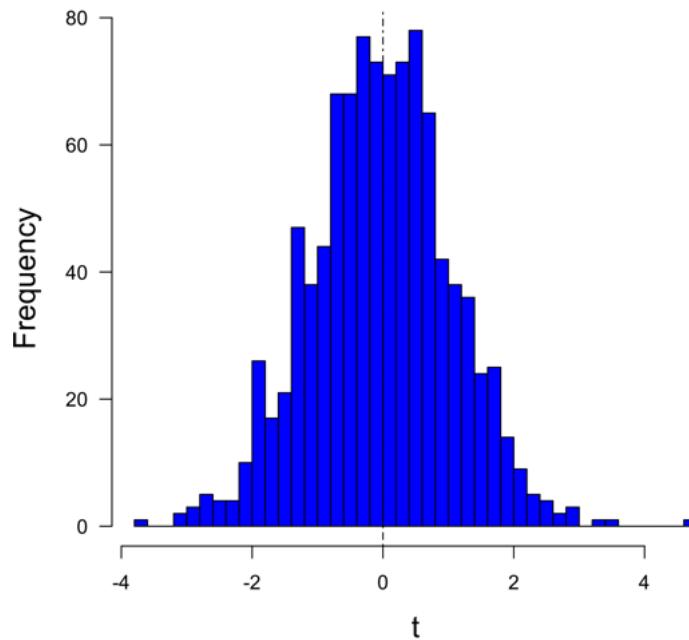


Figure 1.9: Histogram of  $M = 1000$   $t$  values from Monte Carlo simulations. The  $t$  value for the true mean value  $\mu = 0$  is  $t = 0$  (indicated by the black dash-dotted vertical line). As expected, the  $t$  values are concentrated around  $t = 0$  and their numbers fall off to both sides symmetrically (except for some randomness). Values beyond  $t = \pm 3$  are rare.

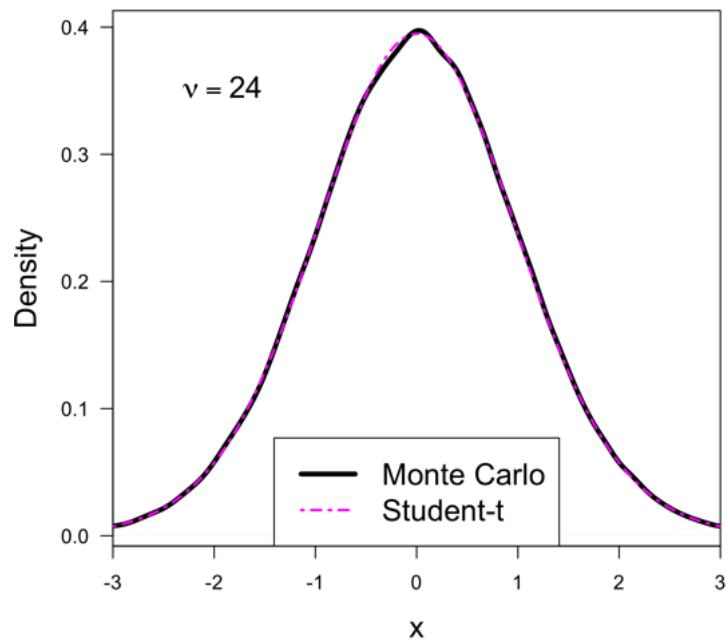


Figure 1.10: Estimate of the density (probability density function, PDF) of the  $t$ -distribution based on  $M = 10^5$   $t$  values from a Monte Carlo simulation (black solid line) and Student's  $t$ -distribution for  $\nu = 24$  degrees of freedom (magenta broken line): the differences between the Monte Carlo estimate and the analytical curve is very small and would further decrease with increasing  $M$ .

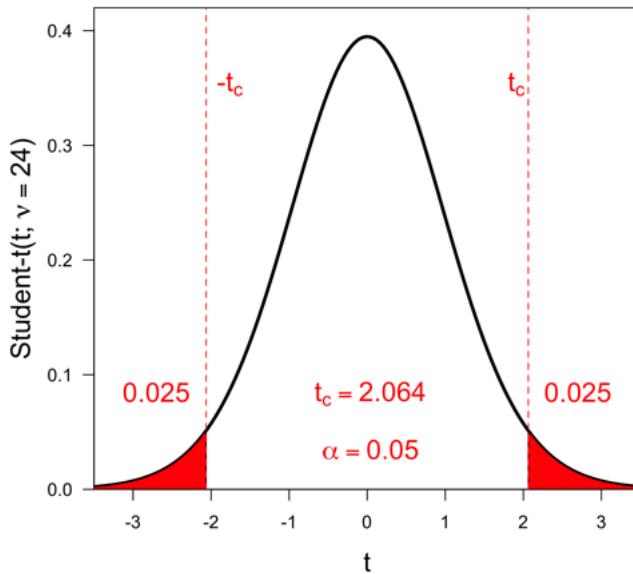


Figure 1.11: The  $t$ -distribution for  $\nu = 24$  (black line) and the rejection region (red area) for the level of significance  $\alpha = 0.05$ . In the two-side  $t$ -test the rejection region actually consists of two parts, one in each tail, with an area of  $\alpha/2 = 0.025$  each.  $t_c = t_{\alpha(2), \nu}$  is the critical  $t$  value for the two-sided  $t$ -test.

[NHST-t-test-RejectionRegion.R](#)

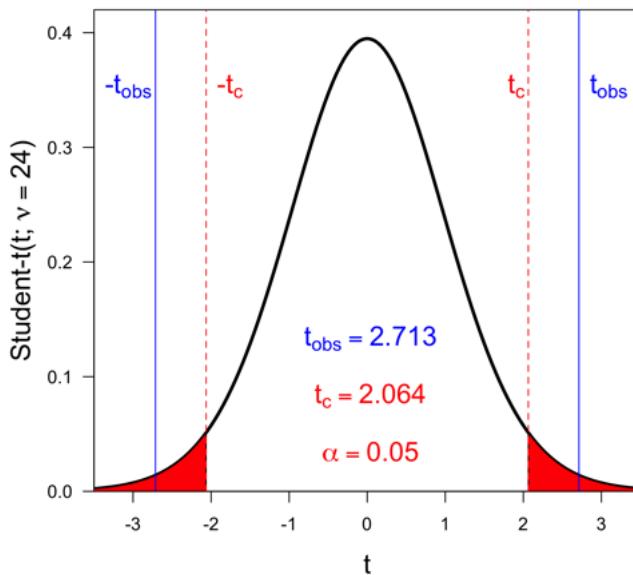


Figure 1.12: The  $t$ -distribution for  $\nu = 24$  (black line) and the observed  $t$ -value (blue vertical lines for  $\pm t_{\text{obs}} = \pm 2.713$ ). The observed  $t$ -value  $t_{\text{obs}} = 2.713$  falls into the right-tail rejection region which speaks against  $H_0$ .

[NHST-t-test-RRtObs.R](#)

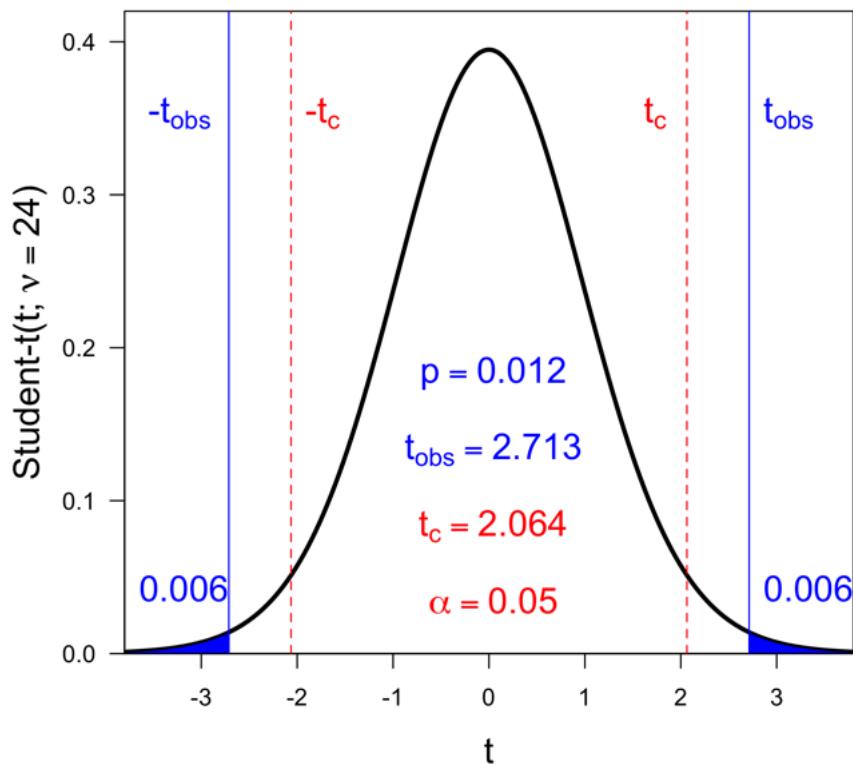


Figure 1.13: The  $p$ -value is the probability to observe  $t_{\text{obs}} = 2.713$  or more extreme values, i.e.  $t \geq t_{\text{obs}} = 2.713$  or  $t \leq -t_{\text{obs}} = -2.713$ . These are the two blue areas which each contribute a probability of about 0.006. Adding up these two probabilities yields the  $p$ -value 0.012. [NHST-t-test-pvalue.R](#)

**Making a decision:** The observed  $t$  value,  $t_{\text{obs}}$ , is located in the rejection region of the  $t$  distribution (red area) and the probability to observe such a  $t$  value or more extreme  $t$  values (blue regions) is  $p = 0.012$  ( $p$ -value, observed level of significance) and thus smaller than the chosen level of significance  $\alpha = 0.05$ . Based on this (and may be additional evidence) the null hypothesis is rejected.

Problems with NHST will be discussed in Section 12.1.5.

The limits of the 95% confidence interval are estimated by  $\bar{x} \pm t_{0.05(2),v} \cdot s / \sqrt{n} = 0.7280 \pm 2.0639 \cdot 1.3418 / 5$  where  $t_{0.05(2),v}$  is the two-side critical value.

## 1.6 Bayesian *t*-test

Applying the Bayesian *t*-test in R is easy:

```
out = ttestBF(x)
```

The output of **ttestBF()** is

```
Bayes factor analysis
-----
[1] Alt., r=0.707 : 4.060789 <U+00B1>0%
Against denominator:
 Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
```

### Explanation of the output:

1. 'Bayes factor analysis': the routine calculates the Bayes factor (for details compare Section [12.1.4](#))
2. '4.060789' is the Bayes factor  $B_{10}$ , i.e. the ratio of the probabilities for the working or alternative hypothesis  $H_1$  and the null hypothesis  $H_0$
3. 'Against denominator: Null, mu = 0' is a bit cryptic hint to the null hypothesis  $H_0$
4. 'Bayes factor type: BFoneSample, JZS': yes, it is a one-sample test; JZS refers to the use of the Jeffreys-Zellner-Siow prior (Rouder et al., 2009)

Instead of the *p*-value, the Bayes factor  $B_{10}$  is the important quantity that has been calculated from the data for given hypotheses  $H_0$  and  $H_1$ . [How to interpret this factor?](#)  $B_{10}$  is much larger than 1 ( $\gg 1$ ) when the probability for  $H_1$  is much larger than that for  $H_0$  and thus the data would speak for  $H_1$  and against  $H_0$ . Conversely,  $B_{10}$  is much smaller than 1 ( $\ll 1$ ) when the probability for  $H_0$  is much larger than that for  $H_1$  and thus the data would speak for  $H_0$  and against  $H_1$ . [However, where to place the limits?](#) Which values are just larger or already much larger than 1? Jeffreys (1961) has proposed the following scales of evidence for  $B_{10}$ : ( $1/\sqrt{10} \approx 0.316$ ,  $\sqrt{10} \approx 3.16$ )

$B_{10} > 10$	strong evidence against $H_0$ (for $H_1$ )
$3.16 < B_{10} < 10$	substantial evidence against $H_0$ (for $H_1$ )
$1 < B_{10} < 3.16$	slight evidence against $H_0$ (for $H_1$ )
$0.316 < B_{10} < 1$	slight evidence against $H_1$ (for $H_0$ )
$0.1 < B_{10} < 0.316$	substantial evidence against $H_1$ (for $H_0$ )
$B_{10} < 0.1$	strong evidence against $H_1$ (for $H_0$ ).

The calculated value  $B_{10} = 4.06$  falls into the range  $3.16 < B_{10} < 10$  providing substantial evidence against  $H_0$  and substantial evidence for  $H_1$ .

The theoretical background for Bayesian testing and the Bayesian *t*-test in particular can be found in Sections [12.1.4](#) and [H.11](#).

### 1.6.1 NHST versus Bayesian approach

In Null Hypothesis Significance Testing (NHST) only one hypothesis, namely the null hypothesis  $H_0$ , is used to calculate the  $p$ -value. Although the alternative hypothesis  $H_a$  is often mentioned (as, for example, in the output of the routine `t.test()`), it is never worked out in terms of assumptions about the statistical population and a corresponding probability ( $p$ -value) for  $H_a$ . As a consequence one can only make (restricted) conclusions about  $H_0$ , namely reject or not reject  $H_0$ . NHST is further plagued by its inconsistency: even at large sample sizes  $\alpha \cdot 100\%$  of the true null hypotheses yield  $p$ -values smaller than  $\alpha$  and thus will be rejected at the chosen  $\alpha$  level.

In the Bayesian approach two hypotheses, namely the null hypothesis  $H_0$  and the alternative or working hypothesis  $H_1$ , are formulated and treating in the same way (formulating likelihoods and assigning priors). Conceptually and mathematically this is more demanding than NHST. In the Bayesian approach the question 'Which hypothesis is more probable,  $H_0$  or  $H_1$ ?' is different from the NHST question 'Is  $H_0$  true or false?' and thus different decisions based on the same data set are possible. Of course, when a blind man can see that  $H_0$  is false, both approaches usually lead to the 'same'<sup>9</sup> decision.

Recommendation: use the Bayesian approach whenever possible<sup>10</sup> and apply NHST with care (compare with what common sense is telling you).

<sup>9</sup>Although differences remain in that the Bayesian approach gives 'evidence' whereas NHST gives 'significance'.

<sup>10</sup>Unfortunately, for many null hypothesis significance tests Bayesian alternatives are not yet available (status: beginning of 2022). The package **BayesFactor** (update 13 December 2021) contains in addition to Bayesian *t*-tests also a Bayesian alternative for ANOVA.

## 1.7 Example 3: straight line fitting

In Example 1 we already developed a statistical model. The Poisson model applies to count data (discrete). The single model parameter, namely the mean rate  $\lambda$ , and its uncertainty has been estimated from the data (Section 1.4). In the current section we will perform some modeling and estimation again: we will fit a straight line to data from a continuous population and estimate two model parameters, namely intercept and slope, and their uncertainties from the data. Under certain conditions, the application of the least-squares method is justified. This method is very powerful and is applied also in more complicated (multi-linear, non-linear) modeling problems.

The following pairs  $(x_j, y_j)$  of data are given (Fig. 1.14):

$$\begin{aligned} x &= \{1.00, 1.14, 1.28, 1.41, 1.55, 1.69, 1.83, 1.97, 2.10, 2.24, 2.38, 2.52, 2.66, 2.79, 2.93, 3.07, \\ &\quad 3.21, 3.34, 3.48, 3.62, 3.76, 3.90, 4.03, 4.17, 4.31, 4.45, 4.59, 4.72, 4.86, 5.00\} \\ y &= \{6.82, 5.73, 6.88, 6.49, 6.37, 6.21, 7.29, 6.74, 5.47, 3.89, 5.01, 5.14, 4.74, 2.32, 3.96, 2.80, \\ &\quad 4.94, 4.56, 6.33, 3.40, 3.69, 3.80, 3.65, 2.62, 3.06, 3.28, 2.71, 0.84, 0.56, 2.65\} \end{aligned}$$

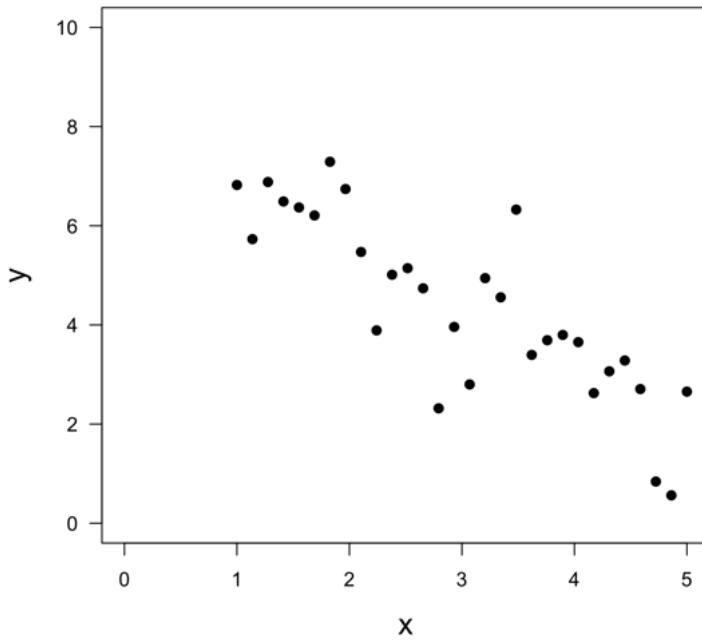


Figure 1.14: Plot of the  $n = 30$  data pairs  $(x_j, y_j)$ . Guessing: a straight line might be a good model for the given data. [StraightLineIntro1.R](#)

Both  $x$  and  $y$  are assumed to be continuous variables. We will assume that  $x$  is a non-stochastic variable (non-random, values known with negligible uncertainty) whereas  $y$  is a stochastic variable, i.e. it contains some noise (additive to its exact values that we do not know). The noise is assumed to be from a normal distribution with zero mean and unknown variance that is independent of  $x$ , i.e.  $\text{Normal}(\mu = 0, \sigma^2) = N(\mu = 0, \sigma^2)$ .<sup>11</sup>

<sup>11</sup>Remarks: (1) Because a single variance  $\sigma^2$  applies for all  $x$  there should be no pattern in the noise that becomes visible when calculating the residuals. (2) Least-squares is justified also under weaker conditions: compare the Gauss-Markov theorem (Section 14.2.1).

A straight line can be fitted to the data by a call of the **R** routine **lm()** for linear modeling<sup>12</sup>:  
**lm(y ~ x)**

Note the order of the argument: the *response* *y* is '*related to*' (represented by a tilde sign,  $\sim$ ) to the *predictor* *x*.

The output of **lm(y ~ x)** reads:  
Call: **lm(formula = y ~ x)**

Coefficients:

(Intercept)	x
8.140	-1.247

### Explaining the output:

1. 'Call: **lm(formula = y ~ x)**' is just an echo of our call
2. 'Coefficients': the model parameters are called coefficients
3. '(Intercept) 8.140': the intercept value has been estimated to equal 8.140; note that the term 'Intercept' is always displayed in brackets to discern it from coefficients of predictor variables
4. 'x -1.247': the coefficient of *x* is estimated to be equal to -1.247; this is the slope of the straight line and, as expected, it is negative because the slope leads us downhill

Intercept and slope are already enough to draw the fitted straight line

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}x = 8.140 - 1.247 x \quad (1.26)$$

through the data cloud (Fig. 1.15). However, we also like to know: What are the uncertainties in the estimates of the intercept and the slope? The **lm()** routine is a bit shy to give as the answer, but the call **summary(lm(y ~ x))** gives us the desired information (and more!). Part of the output reads:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.1399	0.4959	16.41	6.72e-16 ***
x	-1.2472	0.1536	-8.12	7.69e-09 ***

### Explaining the output:

1. In addition to the estimates of intercept and slope (1. numerical column), the uncertainties are given under the heading Std. Error: they read 0.4959 for the intercept and 0.1536 for the slope.
2. The *t* values (3. numerical column) indicated that a *t* test has been applied (although we didn't ask for it!): the null hypothesis  $H_0$  is 'the coefficient is zero'.
3. The *p*-values for both *t* tests (4. numerical column) are both tiny and thus both null hypotheses are rejected on the chosen significance level of  $\alpha = 0.05$ .
4. Consequently, the results are 'highly significant' indicated by the three little stars.
5. Remark: The results from the *t* test can already be anticipated by the fact that the uncertainties are much smaller than the magnitude of the coefficients, making it obvious zero coefficients are out of question (except for tiny probabilities).

One would report the following estimates:

$\hat{\beta}_0 = 8.14 \pm 0.50$  intercept

$\hat{\beta} = -1.25 \pm 0.15$  slope.

<sup>12</sup>The routine **lm()** is very powerful and can do much more than fitting a straight line. It can, for example, also be applied to multivariate data sets.

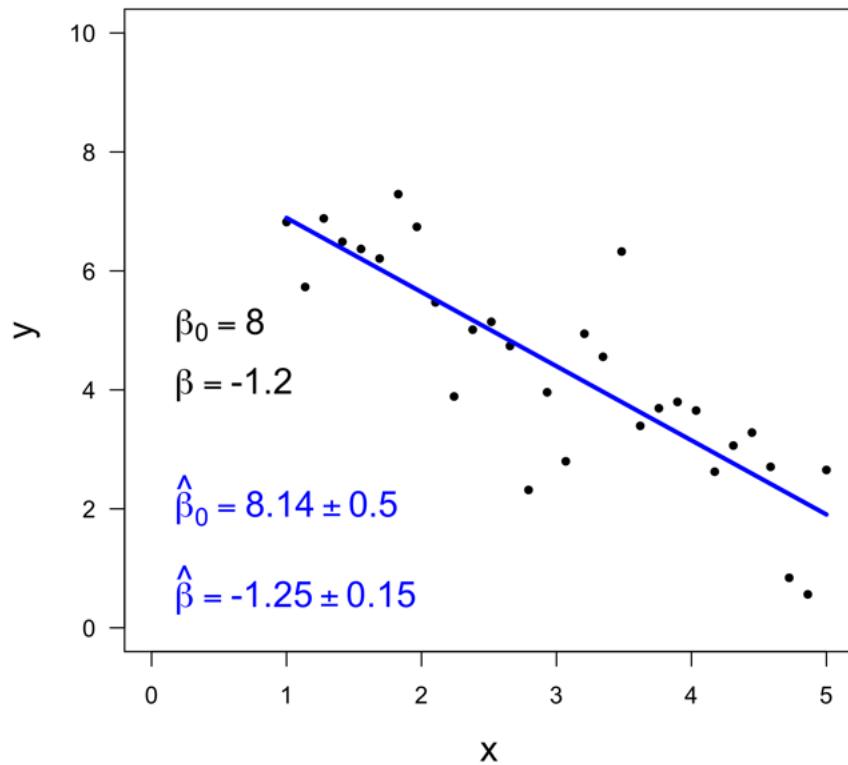


Figure 1.15: The data (black dots) and the fitted straight line (blue solid line). The estimates of the intercept,  $\hat{\beta}_0 = 8.14 \pm 0.50$ , and the slope,  $\hat{\beta} = -1.25 \pm 0.15$ , encompass the exact values (which are not available in real world cases)  $\beta_0 = 8$  and  $\beta = -1.2$  in their uncertainty ranges.

The straight line fitting is based on the [least-squares method](#) that is illustrated in Fig. 1.16. The deviations in  $y$ -direction between data points and a straight line can be positive (data point above the line) or negative (data point below the line). In order that every deviation counts one square the deviations (represented by the yellow squares). The squared differences for all data points are summed up (total area of yellow squares). Of course, the sum of squares depends on the position of the straight line, i.e. its intercept and slope. By varying the intercept and slope one tries to find the minimum of the sum of squares. The intercept and slope that belong to the straight line that minimizes the sum of squares are our searched for least-squares values.

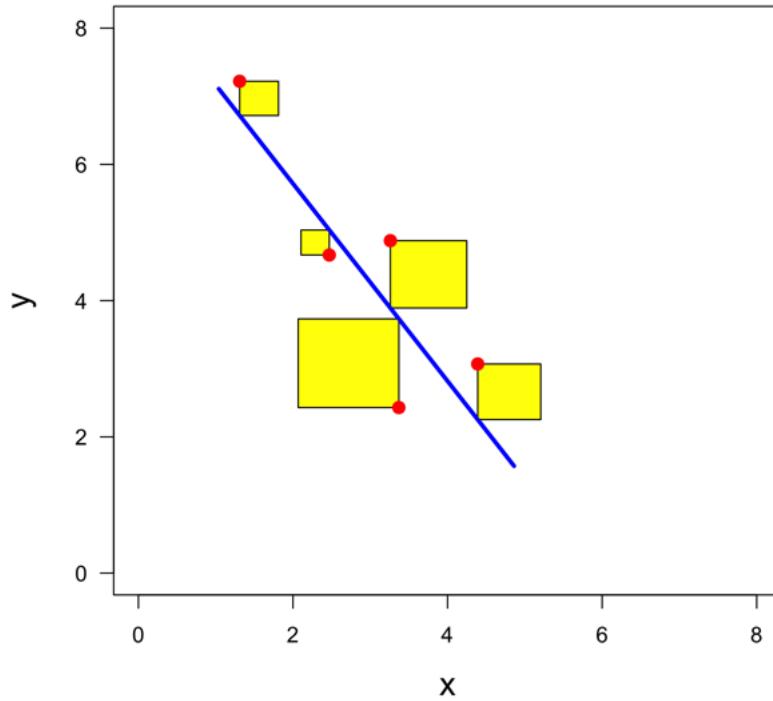


Figure 1.16: The estimated optimal straight line (blue solid line), five out of 30 data points  $(x_k, y_k)$  (red dots), and the corresponding squares of the distances  $d_k$  (yellow squares). The least squares method minimizes the total area of the yellow squares (sum of squares) by varying the intercept and slope of the straight line. [LSconcept.R](#)

## 1.8 Summary & Outlook

In this introductory chapter we have discussed three examples with relative small data sets that are easy to display and grasp. This allowed us to introduce certain concepts – statistical model, estimation (as a technical term), discrete versus continuous, probability distributions, probability density functions,  $p$ -value, Bayes factor – and methods – modeling, parameter estimation, hypotheses testing, Monte Carlo simulation, least-squares. The methods were applied to data sets by calling certain **R** routines and the output from these calls was interpreted. Many details about the theoretical background were left out but will be addressed in further chapters. Fig. 1.17 gives a rough overview about the topics that will be presented in the chapters that will follow.

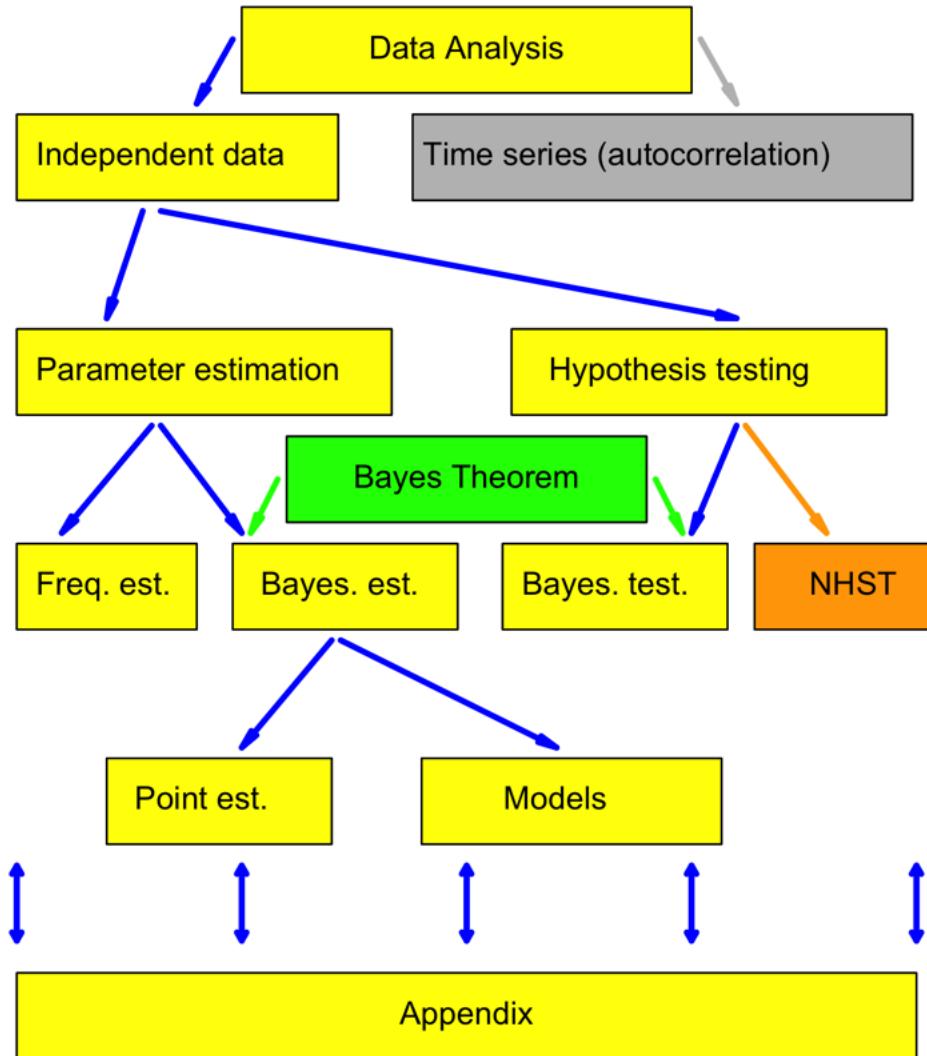


Figure 1.17: Overview: covered topic (yellow), orange (covered, not recommended but still in use), (almost) not covered (grey), green (theoretical background); 'Bayes Theorem' stands for the basic rules of probability and the application of Bayes Theorem in parameter estimation and hypothesis testing; NHST = Null Hypothesis Significance Testing. Note: not all topics could be covered by this overview graph. [Overview22.R](#)

## Chapter 2

# Look at your data: graphical analysis

"A plot is worth a thousand numbers" (a variant of "A picture is worth a thousand words")

"The greatest value of a picture is when it forces us to notice what we never expected to see."  
John Tukey (1977)

*Looking at data is an essential part of data analysis: it can give blueinsights into what is going on. Many guesses mentioned in this script are based on '[graphical analysis](#)'. For nice examples of graphical analysis compare Spiegelhalter (2019), especially the murder case with more than 200 victims. Graphical displays of data can give a much better view on data than a look at data values in lists or tables. Graphical displays of complex or high-dimensional data can be challenging and thus inspection of subsets (slices, sections, projections) may be helpful. In the current chapter various ways to visualize data are shown, however, the following is not an exhaustive list of graph types.*

## 2.1 Scatterplots, histograms, box plots, transparent box plot

Suppose the following data are given:

$$x_1 = \{-0.68644, -0.82379, -0.98416, -2.02230, -0.43507, -0.76655, 1.22178, -0.76655, 1.22178, -0.76655, 1.22178, 0.09767, -0.93391, -1.23458, 0.09188, 0.56736, -0.55276, -0.07969, 0.11767, 2.07541, 1.76443, 0.60249, -1.29916, -0.30322, -0.77935, -0.97190, 0.84580, 0.28698, 1.15160, 0.35533, 0.32936, 1.68584, 0.18260, 1.93600\}$$

First, we display the data in the form of a scatterplot (Fig. 2.1).

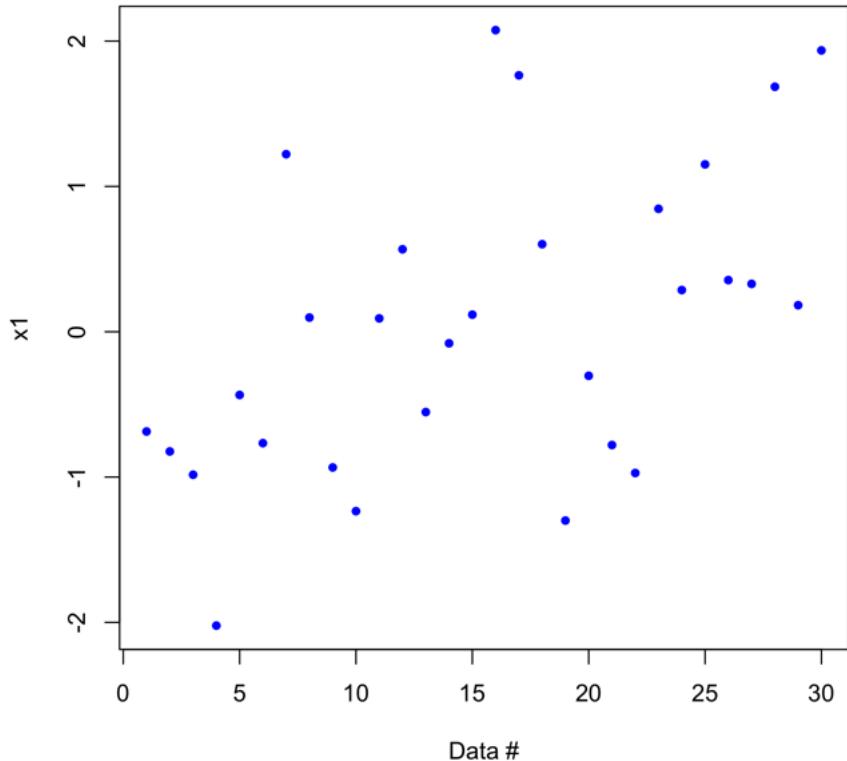


Figure 2.1: Scatterplot of the data  $x_1$ . R-code: [LookScatterQuickAndDirty.R](#)

The quick and dirty plot Fig. 2.1 can be improved (larger axis labels) for presentations or publications by a few more calls:

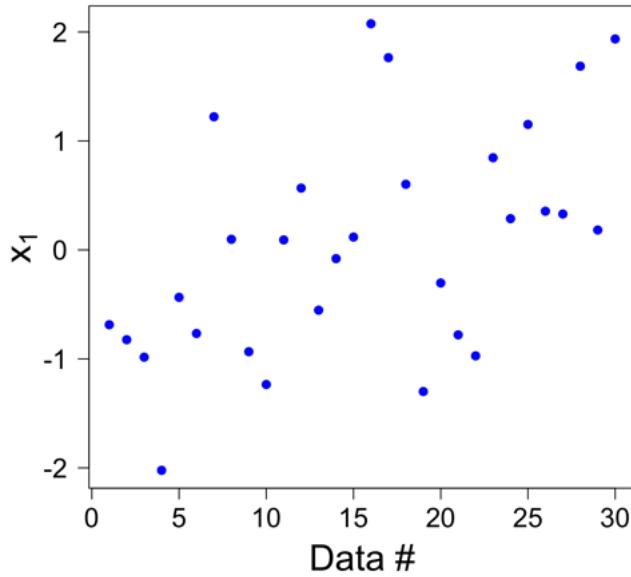


Figure 2.2: Scatterplot of the data  $x_1$  (improved version). [R-code: LookScatter.R](#)

Scatterplots give us some impression of the sample size (number of data), the central tendency and the dispersion (we might guess what's the size of the sample mean or of the standard deviation) and we can read the range (maximum - minimum) of the data. Although we can guess from a scatter plot how the data are distributed over the range, this distribution can be visualized much better by a histogram where the range is broken up into equidistant subranges, the values are counted for each subrange, and finally plotted in form of bins (Figs. 2.3 – 2.4).

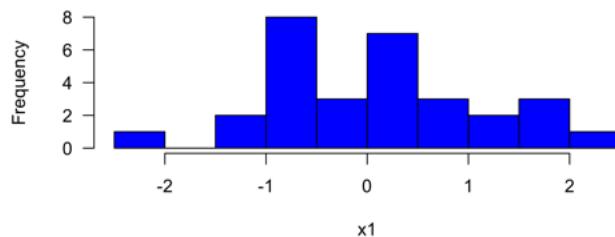
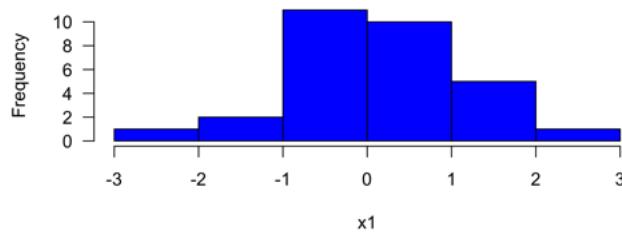


Figure 2.3: Histogram of the data  $x_1$  (quick & dirty version): upper and lower panel with same data, however, different breaks.

R-code: [LookHistogramQuickAndDirty.R](#)

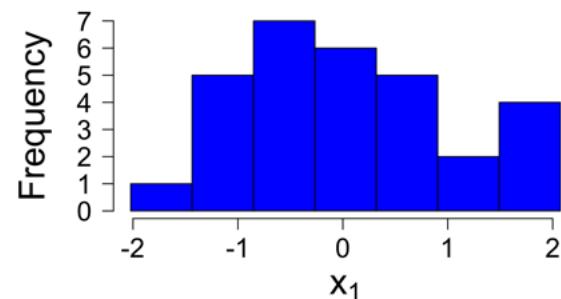
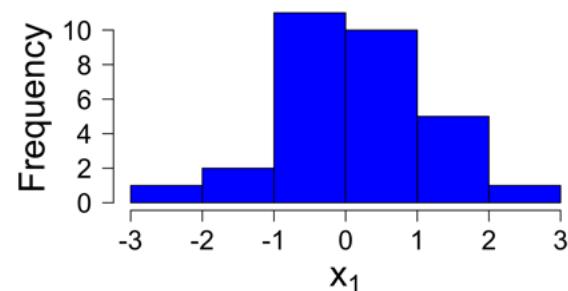
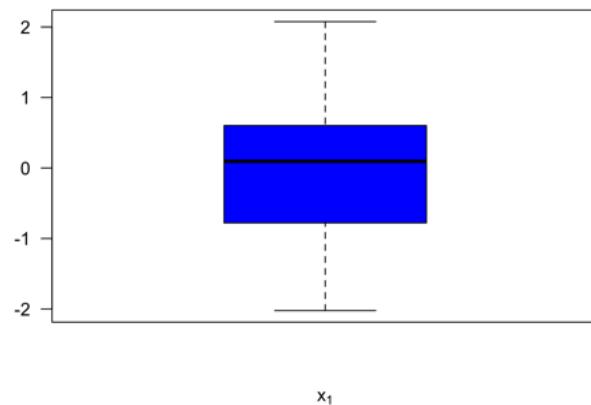
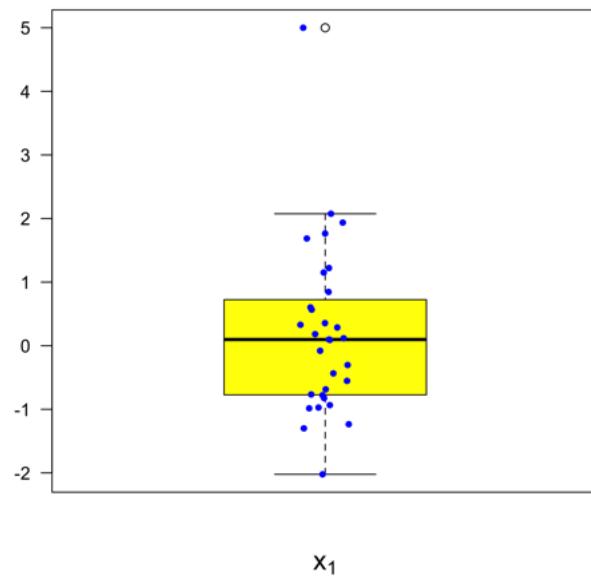


Figure 2.4: Histogram of the data  $x_1$ : upper and lower panel with same data, however, different breaks.

R-code: [LookHistogram.R](#)

Figure 2.5: Box plot of the data  $x_1$ . [LookBoxplot.R](#)Figure 2.6: Transparent box plot of  $x_1$  and an additional data point ('outlier').  
R-code: [LookTransparentBoxplot.R](#)

### 2.1.1 Box-and-whisker plots ('box plots')

Box-and-whisker plots (or 'box plots' for short) were introduced in the 1970ies by John Tukey (Tukey, 1970, 1977). They allow a quick look at the central tendency (median) and spread of your data (helps to identify 'outliers'). The plot makes no assumptions about the underlying statistical distribution. What is shown in box-and-whisker plots (Fig. 2.7)? The numerical values given below are based on the data set used in Fig. 2.6.

1. The horizontal line within the box indicates the median,  $Q_2 = 0.0977$ , (a robust estimate of the central tendency) of the data. The sample mean (arithmetic mean)  $\bar{x} = 0.2077$  is different from the median.
2. The lower hinge of the box,  $Q_1 = -0.773$ , is defined by the 25% quantile of the data, i.e. 25% of the data lie below  $Q_1$ .
3. The upper hinge of the box,  $Q_3 = 0.7241$ , is defined by the 75% quantile of the data, i.e. 75% of the data lie above  $Q_3$ .
4. Note: For even sample size  $n$ , the actual values of  $Q_1$  and  $Q_3$  provided by R routine `boxplot()` can deviate slightly from the 25% and 75% quantiles.
5. The box covers 50% of the data.
6. The interquartile range (IQR) is defined by  $IQR = Q_3 - Q_1 = 1.4971$ .
7. Various conventions for the definition of the whiskers have been proposed and are in use. They have in common that the whiskers must end at an observed data point (this can be quickly checked in transparent box plots). The convention used in the R routine `boxplot()` is as follows. The upper whisker,  $UW = 2.0754$ , is defined by the largest data point that lies in the range between  $Q_3$  und  $Q_3$  plus 1.5 times the interquartile range (IQR); the lower whisker,  $LW = -2.0223$ , is defined by the smallest data point that lies in the range between  $Q_1$  und  $Q_1$  minus 1.5 times the interquartile range (IQR). Thus the differences  $Q_1 - LW = 1.2493$  and  $UW - Q_3 = 1.3513$  are different from each.

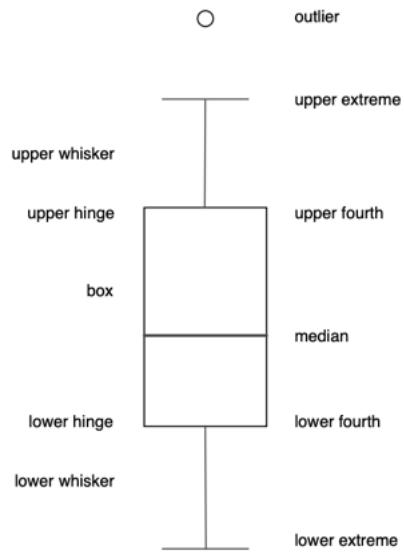


Figure 2.7: Wickham & Stryjewski (2011, Fig. 1): Construction of a box plot. Labels on the left give names for graphic elements, labels on the right give the corresponding summary statistics. Note that the convention for the definition of the whiskers in this figure (by the extreme data points) is different from the convention used by the R routine `boxplot()`.

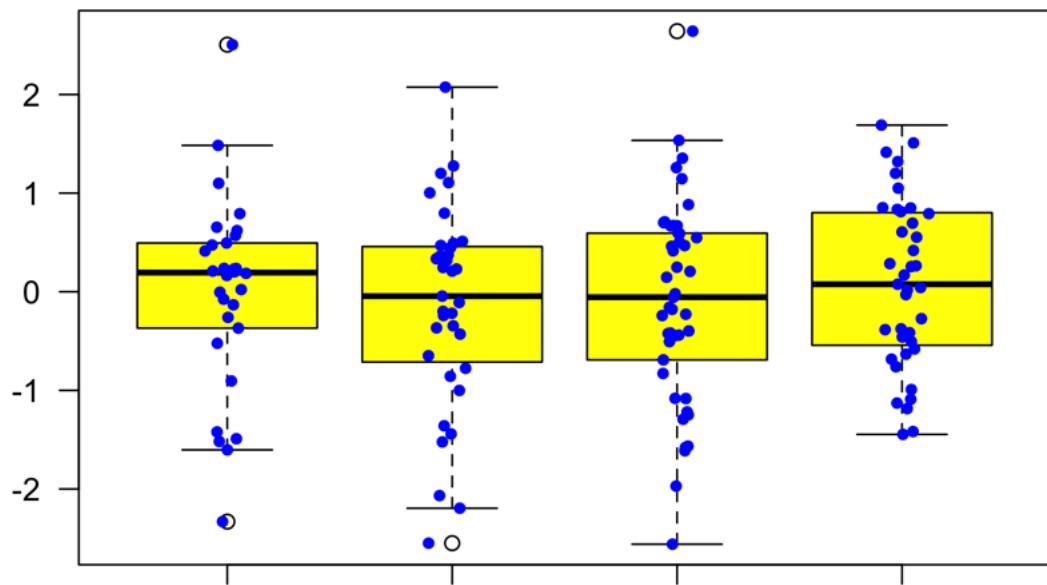


Figure 2.8: Transparent box plot of four samples with different sample sizes, all from the standard normal distribution. [LookMultipleBoxPlot.R](#)

## 2.2 Bubble plots

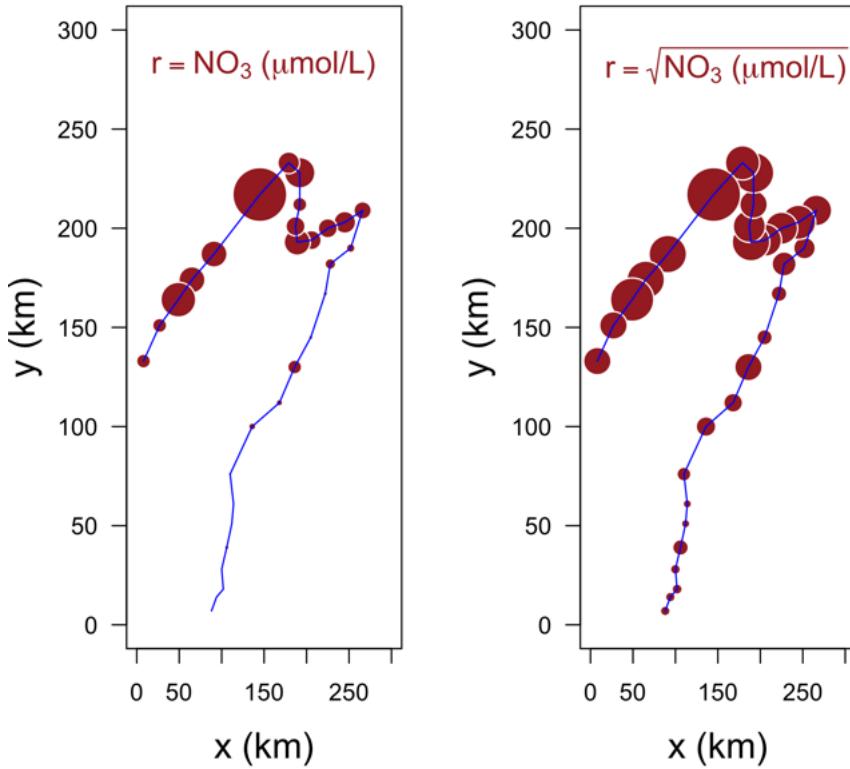


Figure 2.9: The plots show nitrate concentrations ( $\mu\text{mol L}^{-1}$ ) in the form of bubbles (radius  $r$ ) in the Doubs river (Borcard et al., 2011): radius (left panel) or area (right panel) of the bubbles are proportional to the nitrate concentrations. The ‘radius’ scaling (left panel) amplifies the differences in concentrations whereas ‘area scaling’ (right panel) gives a more realistic impression of the concentrations.

R-code: [LookBubblePlots.R](#)

## 2.3 Maps: image

Nowadays many environmental data are publicly available via internet from various data centers. An example from an oceanographic database are the concentrations of phosphate, nitrate, etc. on a  $1^\circ \times 1^\circ$  grid for 33 vertical levels. Below we load such a large data set and generate a color map of the surface concentration of phosphate.

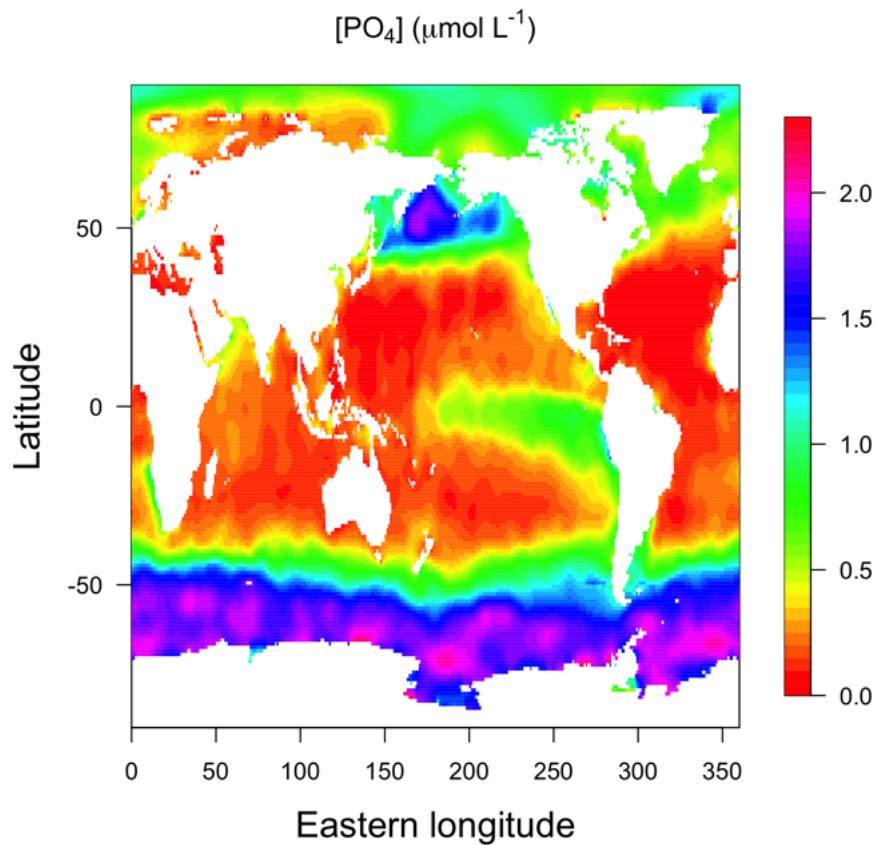


Figure 2.10: Annual mean ocean surface phosphate concentration,  $[\text{PO}_4]$  ( $\mu\text{mol L}^{-1}$ ).

[R-code: LookColorMapPO4.R](#)

## 2.4 Heatmaps: 'plot a matrix'

Reading and analyzing matrices can be tedious when their sizes exceed certain limits. Thus one may ask 'Can we plot matrices?' Yes, we can! In this section we will plot a correlation matrix.

Music: Can: Paperhouse <https://www.youtube.com/watch?v=m5SyB1oMwsM>

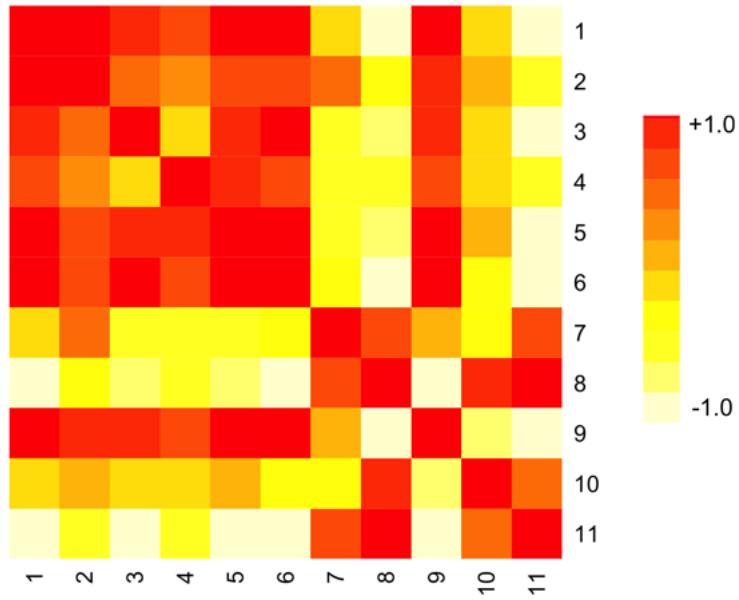


Figure 2.11: Heatmap of a correlation matrix: correlation values can vary between  $-1$  (light yellow) and  $+1$  (red); the diagonal elements of a correlation matrix are all equal to  $+1$  (correlation of each data vector with itself); the correlation matrix and thus the graph are symmetric with respect to the diagonal. The correlation matrix considered here possesses some high correlations outside the diagonal. Remark: the R routine `heatmap()` applied here can do much more, for example, plot associated dendograms.

R-code: [LookHeatmap.R](#)

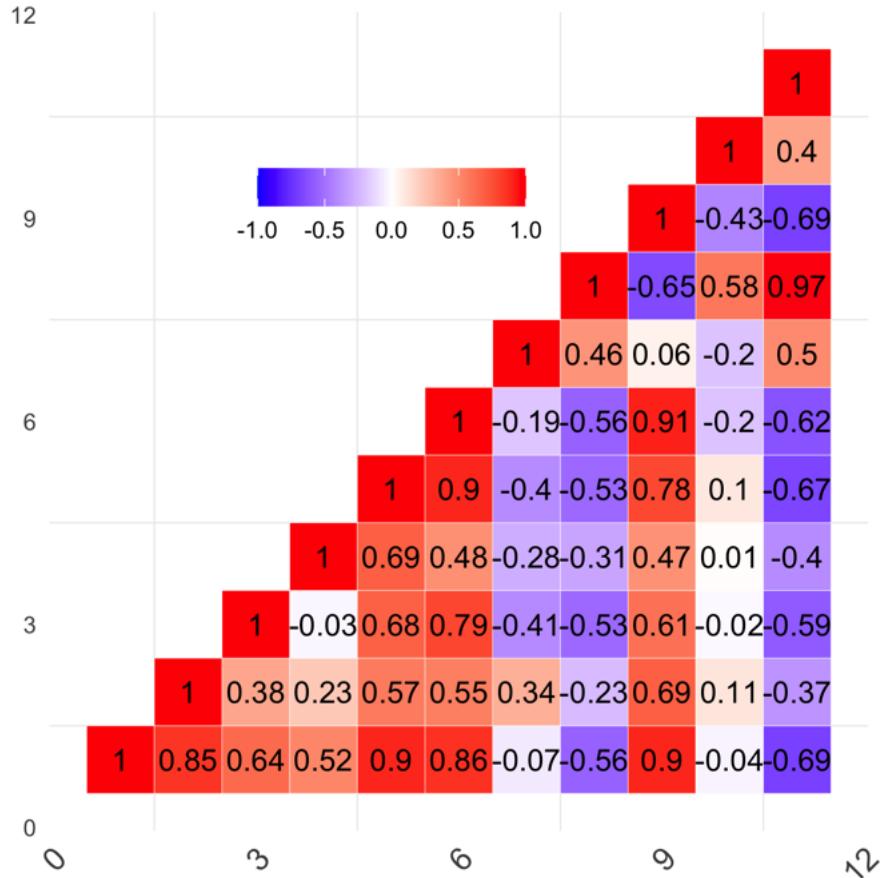


Figure 2.12: Alternative plot (using the package `ggplot2`) of a correlation matrix: only 'half' of the matrix is plotted (because of the symmetry this contains all relevant information); correlation values can vary between  $-1$  (blue) and  $+1$  (red); the diagonal elements of a correlation matrix are all  $+1$  (perfect correlation).  
`install.packages('reshape2')`

R-code: [LookHeatmapSymMatrix.R](#); [plotCorMatrix.R](#)

## 2.5 Exercises

### Exercise 1 Transparent box plot for random sample from F-distribution

Generate a random sample of size  $n = 30$  from the F-distribution with degrees of freedom  $v_1 = 15$  and  $v_2 = 3$  and produce a transparent box plot. Discuss the resulting plot.

### Exercise 2 Mean & variance & correlation

Analyze the data listed in Table 2.1 by calculating the mean and variance for each  $x_k$  and  $y_k$  and the correlation between  $x_k$  and  $y_k$  before plotting the data. Then fit straight lines to the data (simple linear regression; compare Chapter 14) and plot data and regression lines.

$x_1$	$y_1$	$x_2$	$y_2$	$x_3$	$y_3$	$x_4$	$y_4$
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Table 2.1: Sets of paired data

Further reading: Weissgerber et al. (2015): Beyond bar and line graphs ...

# Chapter 3

## Mean, variance, random sampling

*There is randomness in the world and thus if you measure a quantity  $X$  several times the results usually differ (often only slightly) from each other. In the current chapter we will discuss various measures of the so-called central tendency (the most important are the arithmetic mean and the median) and the dispersion (spread) of the data (variance, standard deviation, MADN). Although this part of data analysis is often called 'descriptive statistics' in textbooks, we will see that 'parameter estimation' is the better notation.*

*Additional concepts introduced in this chapter are: statistical population, random sample, estimator. Monte Carlo simulations will be applied to evaluate estimators.*

### 3.1 Central tendency: mean, median, ...

The aim is to find a typical or central value of a sample. Various measures of central tendency have been proposed: arithmetic mean, median, geometric mean, harmonic mean, weighted arithmetic mean, truncated mean, midrange, midhinge, trimean, Winsorized mean. In this section we will discuss most of these concepts and apply them to the following sample with  $n = 9$  data points:

$$x = \{2.480, 0.668, 0.542, 1.770, 2.678, 2.931, 0.295, 4.656, 1.825\}. \quad (3.1)$$

#### 3.1.1 Arithmetic mean or sample mean

The **sample mean** or arithmetic mean (or mean for short) is defined by

$$\text{sample mean} = \text{arithmetic mean} = \text{mean} = \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.2)$$

For our sample we obtain  $\bar{x} = 1.983$ . The mean is the most commonly used measure for the central tendency. It is an unbiased estimator (compare Chapter 10) for the true mean  $\mu$  of the **statistical population** (or 'population' for short) from which we took the sample. The (usually unknown) true population can be described by a probability distributions (PD; discrete) or by a probability density functions (PDF; continuous); compare Chapter 6. True values are denoted by Greek letters. Estimates of the true values are denoted by Latin letters or by adding a hat on top of the Greek letters, i.e.  $\hat{\mu} \equiv \bar{x}$ . The arithmetic mean, while unbiased, is sensitive to outliers, i.e. in contrast to other estimators of the central tendency such as the median or the Winsorized mean it is not a robust estimator.

### 3.1.2 Median

In order to determine the median one has to sort the data of the sample in ascending or descending order (it does not matter which):

$$\mathbf{x}_s = \{0.295, 0.542, 0.668, 1.770, \textcolor{blue}{1.825}, 2.480, 2.678, 2.931, 4.656\}. \quad (3.3)$$

The median is the value in the middle of the ordered (sorted) data set  $\mathbf{x}_s$ . For our sample the median is 1.825. If the number of data  $n$  is even one uses the arithmetic mean of the two middle values as median:

$$\mathbf{y}_s = \{0.295, 0.542, 0.668, 1.770, \textcolor{blue}{1.825}, \textcolor{blue}{2.480}, 2.678, 2.931, 4.656, 4.811\} \quad (3.4)$$

The median for the data set  $\mathbf{y}_s$  is  $(1.825 + 2.480)/2 = 2.1525$ . The median is an unbiased robust estimator of the mean  $\mu$  of normally distributed populations with, however, a larger variance than the sample mean (Exercise 3). Remark: in R the median of  $\mathbf{x}$  can be determined by calling `median(x)`, i.e. the data do not have to be sorted beforehand.

### 3.1.3 Geometric mean (\*)

The geometric mean is defined by

$$\text{geometric mean} := (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n} = \left( \prod_{i=1}^n x_i \right)^{1/n}. \quad (3.5)$$

Obviously, this definition makes sense only when all  $x_i \neq 0$ ; if at least one  $x_i = 0$ , the geometric mean would yield 0 which is not a good measure of the central tendency when most data points are different from 0. If an odd number of the  $x_i$  are negative, the product of all  $x_i$  is negative and one would obtain an imaginary number for the geometric mean.

Usually, the geometric mean is applied only when all data are positive numbers, i.e. all  $x_i > 0$ . For  $n = 2$  the geometric mean is equal to the length of the diagonal of a rectangle with side lengths  $x_1$  and  $x_2$ . For  $n = 3$  it is equal to the length of the diagonal of a cuboid with side lengths  $x_1$ ,  $x_2$ , and  $x_3$ , etcetera. For our sample (3.1) the geometric mean is 1.466. The geometric mean is sometimes used when data are (approximately) log-normal distributed (for example, Daniels et al. 2018).

### 3.1.4 Harmonic mean (\*)

The harmonic mean is defined by

$$\text{harmonic mean} := \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}. \quad (3.6)$$

Obviously, the definition makes sense only when all  $x_i \neq 0$ . It is usually only applied when all  $x_i > 0$ . For our sample the harmonic mean is 1.020. Examples for application of the harmonic mean: (1) calculate the average speed of a vehicle (Exercise 4) or (2) replace  $n$  parallel resistors by a single resistor (electrical engineering; Exercise 5).

### 3.1.5 Winsorized mean (\*)

In order to get rid of (possible) outliers, the lowest sample point is replaced by the second lowest one and the largest sample point is replaced by the second largest one:

$$\mathbf{x}_{w1} = \{\textcolor{blue}{0.295}, 0.295, 0.542, 0.668, 1.770, 1.825, 2.480, 2.678, 2.931, \textcolor{blue}{2.931}\}. \quad (3.7)$$

Subsequently, the arithmetic mean of the modified sample is calculated. For our sample we obtain the Winsorized mean = 1.762. Instead of modifying just one value at each end of the sorted data set, one may modify  $k$  values at each end. When modifying all values except for the middle one(s), one obtains the median.

**Exercise 3 Median of the Poisson distribution (\*)**

No exact expression is known for the median of the Poisson distribution. It lies between

$$\lambda - \log 2 < \text{median} < \lambda + \frac{1}{3} \quad (3.8)$$

(Adell & Jodrá, 2005). Apply Monte Carlo simulations to investigate how good these limits are for  $\lambda$  between 0.5 and 3 in steps of 0.1.

**Exercise 4 Average speed**

On a trip you travel half the distance with speed  $90 \text{ km h}^{-1}$  and the other half with  $10 \text{ km h}^{-1}$ . What is the average speed during your trip? Show that the average speed can be calculated by the harmonic mean.

**Exercise 5 Parallel resistors**

Two parallel resistors with 600 and 400 ohm, respectively, should be replaced by a two resistors with equal resistance while keeping the electric current at the same level. What's the resistance of these resistors?

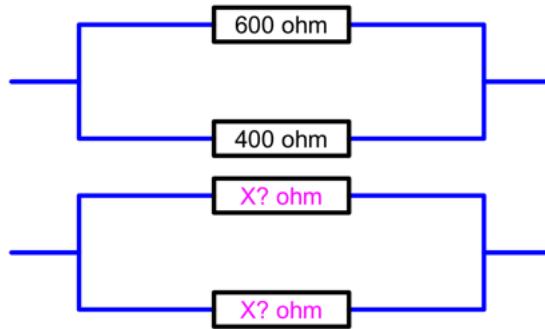


Figure 3.1: Two parallel resistors [MVCCexerciseResistors.R](#)

## 3.2 Data dispersion: variance, standard deviation, MADN

The dispersion of data is the second most important characteristic of a data set. It can be measured by the variance, the standard deviation, or, in a more robust way, by the 'Normalized Median Absolute Deviation about the Median' (MADN).

### 3.2.1 Sample variance

The sample variance  $s^2$  is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{\nu} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.9)$$

where  $\nu = n - 1$  is the 'degrees of freedom'<sup>1</sup>. The sample variance  $s^2$  is an unbiased estimator of the true variance  $\sigma^2$  given that it exists<sup>2</sup>. The units of the variance are different from those of the data, except in cases where the quantities are dimensionless. For the sample (3.1)  $s^2 = 1.937$ .

### 3.2.2 Sample standard deviation

The standard deviation of the sample is given by the square root of the sample variance

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{\nu} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.10)$$

The standard deviation has the same units as the data or the mean. Therefore, it is easy to interpret as a measure of dispersion. For the sample in (3.1)  $s = 1.392$ .

### 3.2.3 MADN

The 'Normalized Median Absolute Deviation about the Median' (MADN) is a robust estimator of the true standard deviation  $\sigma$ . It is defined by

$$\text{MADN} = \text{median} |x - \text{median}(x)| / 0.6745 \quad (3.11)$$

where the factor  $1/0.6745$  makes MADN an unbiased estimator of the standard deviation for normal distributions and large sample sizes (for a more detailed discussion see Section 10.8.1). For the sample in (3.1) MADN = 1.640.

---

<sup>1</sup>Degrees of freedom = number of data minus number of constraints (Section 3.6); here the sample mean is the only constraint.

<sup>2</sup>The true variance  $\sigma^2$  is the variance of the statistical population from which we sampled. True values are denoted by Greek letters. The corresponding estimates are denoted by Latin letters or by a hat atop the Greek letters, as, for example,  $s^2 \equiv \hat{\sigma}^2$ . Some populations do not possess variances; an example is the Cauchy PDF (Section C.3.4).

### 3.3 Covariance

So far we have discussed various statistical concepts (mean, variance, etc.) that characterize single data sets. Now we will introduce concepts for the relationships between two or more data sets. The covariation of data pairs  $x_k, y_k$  ( $k = 1, 2, \dots, n$ ) can be measured by the covariance or, after scaling, by the correlation. For more than two data sets it is useful to summarize the covariances or correlations of the various sample pairs in the form of a covariance or correlation matrix, respectively. Non-parametric correlations based on the ranks of data will be discussed as well.

#### 3.3.1 A simple example

As a simple example let us consider the following data sets

$$x = \{5.1, 4.8, 1.8, 4.5, 6.6, 2.2, 5.7, 5.7, 5.9, 9.7\} \quad (3.12)$$

$$y = \{7.8, 8.2, 3.5, 6.7, 8.9, 3.2, 8.7, 8.5, 7.2, 14.2\} \quad (3.13)$$

The data are plotted in Fig. 3.2 (blue circles) together with the positions of the sample mean values  $\bar{x} \equiv \hat{\mu}_x$  and  $\bar{y} \equiv \hat{\mu}_y$  (indicated by the red lines). The scatterplot 3.2 shows that  $x$  and  $y$  covary, with low values of  $y_k$  usually corresponding to low values of  $x_k$  and vice versa. Most data pairs  $(x_k, y_k)$  that follow this rule lie in the upper right quadrant (defined by  $x_k > \hat{\mu}_x$  and  $y_k > \hat{\mu}_y$ ) or in the lower left quadrant (defined by  $x_k < \hat{\mu}_x$  and  $y_k < \hat{\mu}_y$ ).

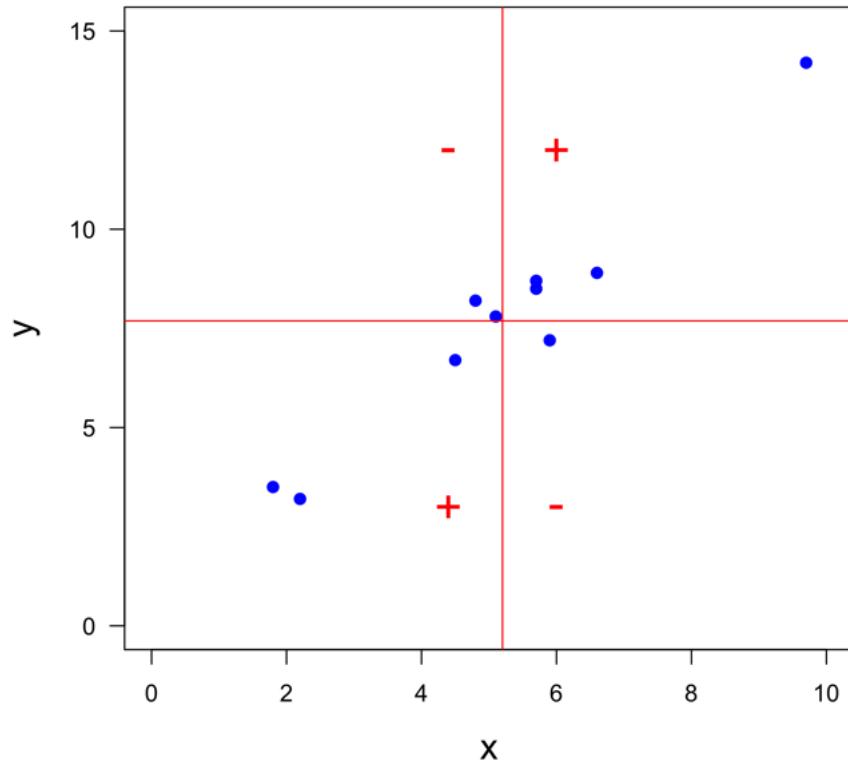


Figure 3.2: Data pairs  $(x_k, y_k)$  (blue circles) and the positions of the sample mean values  $\bar{x}$  and  $\bar{y}$  (indicated by red lines). [MVCexampleCovariance.R](#)

The covariance for two samples  $x$  and  $y$  is 'defined'<sup>3</sup> by

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (3.14)$$

The differences  $(x_k - \bar{x})$  and  $(y_k - \bar{y})$  can be positive (for values larger than the corresponding sample mean) or negative (for values smaller than the corresponding sample mean). The products  $(x_k - \bar{x})(y_k - \bar{y})$  are positive in the upper right and the lower left quadrants and negative in the upper left and the lower right quadrants (Fig. 3.2). They are large when the data pairs are far apart from the mean values (red lines). For the data set given above most data pairs lie in the positive quadrants, many quite distant from the mean values, consequently the covariance is positive. The covariance value is difficult to interpret without some reference to compare it with. In addition, the covariance units are those of the product of  $x$  and  $y$  (for example °C m) which is not easy to grasp.

---

<sup>3</sup>The expression (3.14) is actually an estimator of the true covariance (von Storch & Zwiers, 2003, p. 83, Eq. 5.13).

### 3.3.2 An example with negative covariance

As another example let us consider the following data set <sup>4</sup>

$$x = \{0.28, 0.37, 0.45, 0.84, 0.76, 0.46, 0.58, 1, 0.66, 1.08, 0.95\} \quad (3.15)$$

$$y = \{66.8, 75.3, 70.6, 15.2, 27.6, 27.3, 37, 45.1, 32.2, 57.8, 24.5\} \quad (3.16)$$

where  $x_k$  are concentrations of phosphate ( $\text{PO}_4$ ;  $\mu\text{mol L}^{-1}$ ) and  $y_k$  are concentrations of vitamin B<sub>12</sub> ( $\text{pmol L}^{-1}$ ).

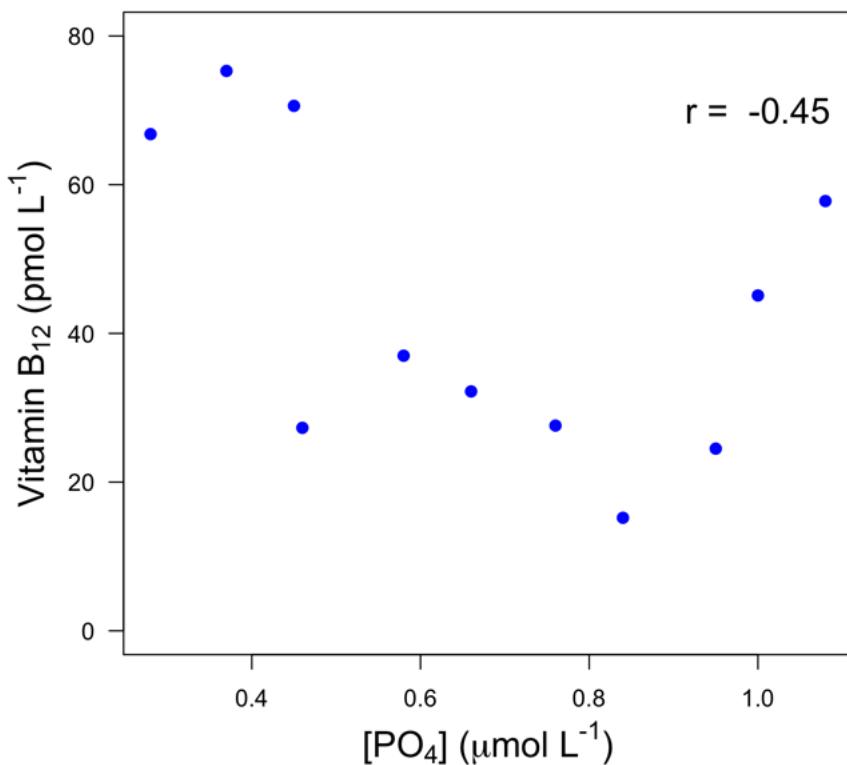


Figure 3.3: Vitamin B<sub>12</sub> versus phosphate: the concentration of vitamin B<sub>12</sub> is high at low phosphate concentrations and lower towards higher phosphate concentrations. The relationship between these two data sets can be quantified by the covariance ( $\text{cov}(x, y) = -2.57$ ) and the correlation ( $\text{cor}(x, y) = -0.45$ ).

[MVCexampleCovSanudo.R](#)

---

<sup>4</sup>Sañudo-Wilhelmy et al. (2006)

## 3.4 Anscombe's quartet

Anscombe (1973) published 4 sets of data pairs that look very different from each other, but possess the same values of covariance (Fig. 3.4). This example shows the limitation of covariance analysis: this 'linear concept' fails to indicate difference in case of strong non-linearities.

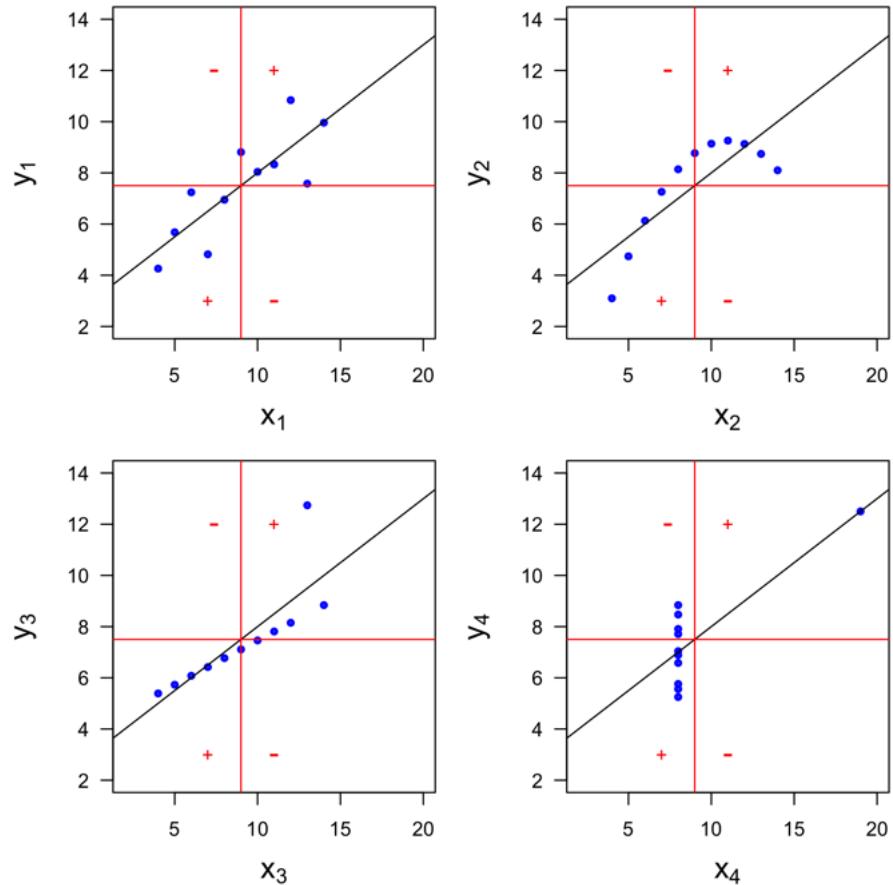


Figure 3.4: Anscombe's 4 sets of data pairs (Anscombe's quartet) that have the same values of covariance (Anscombe, 1973). [MVCCexampleCovAnscombe.R](#)

### 3.4.1 Covariance matrix

For more than two samples of the same size, one can calculate the covariances between each pair of samples including the covariance with itself,  $\text{cov}(x_i, x_i)$ . The results can be displayed in the form of a table called 'covariance matrix'. For an example with 4 samples (see below) the covariance matrix reads

$$\text{cov}(X) = \begin{pmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) & \text{cov}(x_1, x_4) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \text{cov}(x_2, x_3) & \text{cov}(x_2, x_4) \\ \text{cov}(x_3, x_1) & \text{cov}(x_3, x_2) & \text{cov}(x_3, x_3) & \text{cov}(x_3, x_4) \\ \text{cov}(x_4, x_1) & \text{cov}(x_4, x_2) & \text{cov}(x_4, x_3) & \text{cov}(x_4, x_4) \end{pmatrix} \quad (3.17)$$

$$= \begin{pmatrix} 916.7 & 460.1 & -430.7 & -107.2 \\ 460.1 & 248.6 & -211.1 & -85.3 \\ -430.7 & -211.1 & 295.1 & 41.3 \\ -107.2 & -85.3 & 41.3 & 797.0 \end{pmatrix} \quad (3.18)$$

The covariance matrix is **symmetric** because the result does not depend on the order of the samples:

$$\text{cov}(x_i, x_j) = \text{cov}(x_j, x_i). \quad (3.19)$$

The diagonal terms are always positive ( $\text{cov}(x_i, x_i) > 0$ ) except when  $x_i \equiv 0$  and correspond to the variance of each sample (Eq. 3.9). Elements of the covariance matrix with positive sign indicate covariation, with negative sign anti-covariation. For example,  $x_2$  covaries with  $x_1$ , and  $x_3$  and  $x_4$  anti-covary with  $x_1$ . As mentioned previously, without a reference covariances are difficult to interpret. Suppose the 4 samples have units h, m, °C, and kg. The elements of the covariance matrix will have units h m, h °C, h kg, m °C, °C kg etc. . . a mixed bag of values with strange units. The solution to this problem is scaling each sample by its standard deviation in order to obtain dimensionless quantities (see next section on 'correlation').

## 3.5 Correlation

Although the covariance gives important hints about the relationships between different data sets, it is often difficult to interpret covariance values because of the units and missing reference points (scales). The solution is to get rid of the units (generate dimensionless quantities) and to apply scaling. Division of the covariance by the standard deviations accomplishes this in a single step leading to Pearson's correlation coefficient  $r$ . An alternative approach is to first rank each data set (ranks have no units) and then calculate the correlation of ranks leading to Spearman's correlation coefficient  $r_s$ .

### 3.5.1 Pearson's correlation coefficient $r$

Pearson's correlation coefficient  $r$  is defined by

$$r = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}} \quad (3.20)$$

where  $\hat{\sigma}_x$  and  $\hat{\sigma}_y$  are the sample standard deviations. Pearson's correlation coefficient is dimensionless and can vary between -1 (perfect anti-correlation) and +1 (perfect correlation). Conventionally one speaks of correlation for  $r > 0.3$ , anti-correlation for  $r < -0.3$ , and weak or no correlation for  $-0.3 \leq r \leq +0.3$ . If authors mention correlation coefficients without further specification (Pearson, Spearman, ...) it usually refers to Pearson's correlation coefficient.

In linear regression (Chapter 14), a measure of the goodness-of-fit of the straight line fit,  $y(x) = \beta_0 + \beta_1 x$ , to the data pairs  $(x_i, y_i)$  is given by the coefficient of determination  $r^2$  which measures the ratio of total variation in dataset  $\mathbf{y}$  to the amount of variation explained by the fitted line. One can show that the coefficient of determination  $r^2$  is equal to the square of Pearson's correlation coefficient  $r$  for the data pairs  $(x_i, y_i)$  (Casella & Berger, 2002, p. 556-557).

### 3.5.2 Correlation matrix

For more than two samples of same size, one can calculate the correlations between all combinations of two samples including the correlation of a sample with itself,  $\text{cor}(x_i, x_i)$ . The results of can be displayed in the form of a table, the so-called 'correlation matrix' (analogous to the covariance matrix). For the example with 4 samples (given in Section 3.4.1) the correlation matrix is

$$\text{cor}(\mathbf{X}) = \begin{pmatrix} \text{cor}(x_1, x_1) & \text{cor}(x_1, x_2) & \text{cor}(x_1, x_3) & \text{cor}(x_1, x_4) \\ \text{cor}(x_2, x_1) & \text{cor}(x_2, x_2) & \text{cor}(x_2, x_3) & \text{cor}(x_2, x_4) \\ \text{cor}(x_3, x_1) & \text{cor}(x_3, x_2) & \text{cor}(x_3, x_3) & \text{cor}(x_3, x_4) \\ \text{cor}(x_4, x_1) & \text{cor}(x_4, x_2) & \text{cor}(x_4, x_3) & \text{cor}(x_4, x_4) \end{pmatrix} \quad (3.21)$$

$$= \begin{pmatrix} 1.000 & 0.964 & -0.828 & -0.125 \\ 0.964 & 1.000 & -0.779 & -0.192 \\ -0.828 & -0.779 & 1.000 & 0.085 \\ -0.125 & -0.192 & 0.085 & 1.000 \end{pmatrix} \quad (3.22)$$

The correlation matrix is symmetric<sup>5</sup>:

$$\text{cor}(x_i, x_j) = \text{cor}(x_j, x_i). \quad (3.23)$$

---

<sup>5</sup>Same argument as for covariance matrix.

The diagonal terms are all equal to 1<sup>6</sup>

$$\text{cor}(x_i, x_i) = 1 \quad (3.24)$$

Elements of the correlation matrix with positive sign indicate positive correlation, with negative sign anti-correlation. For example,  $x_2$  is correlated to  $x_1$ ,  $x_3$  is anti-correlated to  $x_1$ ,  $x_4$  is uncorrelated to all other samples. The correlation matrix is easy to interpret because all elements are dimensionless and can vary over the same range from -1 to +1.

### 3.5.3 Other correlation coefficients: Spearman, Kendall

**Spearman's rank correlation coefficient  $\rho$  or  $r_s$**  is a nonparametric measure of rank correlation. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables:

$$\rho \equiv r_s := \frac{\text{cov}(\text{rank}_x, \text{rank}_y)}{\sigma_{\text{rank}_x} \sigma_{\text{rank}_y}} \quad (3.25)$$

"While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other." (Wikipedia)

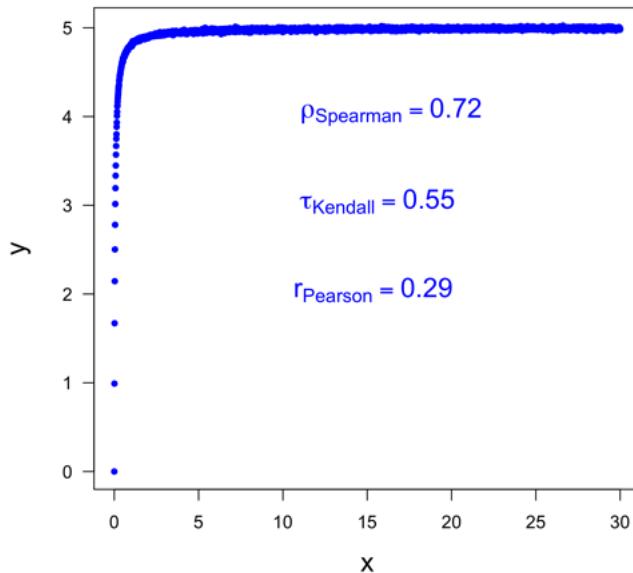


Figure 3.5: Example with small Pearson correlation coefficient ( $r = 0.29$ , 'no linear relationship over the whole  $x$ -range') and high Spearman correlation coefficient ( $\rho = 0.72$ , 'monotonic relationship').

[MVCCorSpearmanPearson.R](#)

---

<sup>6</sup>Except for the pathological case of a sample where all data point have the same value and as a consequence with standard deviation equal to 0 in which case the correlation is not defined.

**Kendall rank correlation coefficient** (also called **Kendall's tau coefficient**): "Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of -1) rank between the two variables." (Wikipedia)

Definition: "Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be a set of observations of the joint random variables  $X$  and  $Y$  respectively, such that all the values of  $(x_i)$  and  $(y_i)$  are unique. Any pair of observations  $(x_i, y_i)$  and  $(x_j, y_j)$ , where  $i \neq j$ , are said to be *concordant* if the ranks for both elements agree: that is, if both  $x_i > x_j$  and  $y_i > y_j$ ; or if both  $x_i < x_j$  and  $y_i < y_j$ . They are said to be *discordant*, if  $x_i > x_j$  and  $y_i < y_j$ ; or if  $x_i < x_j$  and  $y_i > y_j$ . If  $x_i = x_j$  or  $y_i = y_j$ , the pair is neither concordant nor discordant."

The Kendall  $\tau$  coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2} \quad (3.26)$$

" (Wikipedia)

Kendall's  $\tau$  can vary between -1 (perfect anti-correlation) and +1 (perfect correlation).

### Exercise 6 Spearman correlation matrix

Calculate the correlation matrix based on the Spearman correlation coefficient for the data set used in Section 3.5.2.

## 3.6 Degrees of freedom

"A concept of central importance to modern statistical theory which few textbooks have attempted to clarify is that of 'degrees of freedom'. . . ."

Not only do most texts omit all mention of the concept but many actually give incorrect formulas and procedures because of ignoring it."

Walker (1940)

The concept of 'degrees of freedom' is used in statistics in various contexts. Despite this fact, textbooks usually do not give a proper definition. I did not find a simple definition for degrees of freedom that can be applied in all situations. However, the following approach will guide us through the first examples.<sup>7</sup> We will come back to the topic of degrees of freedom in the context of point estimators (Chapter 10).

**Definition** According to Walker (1940), the 'degrees of freedom' are given by the number of independent terms or data minus the number of constraints, or, in short,

$$\text{degrees of freedom} = \# \text{ of data} - \# \text{ of constraints}. \quad (3.27)$$

This definition makes no sense without defining the 'constraints'. Instead of trying to give further definitions let us look at some examples.

**Examples:**

1. Sample mean: no constraint  $\Rightarrow$  degrees of freedom = number of data:  $\nu = n$

$$\hat{\mu} \equiv \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{\nu} \sum_{k=1}^n x_k; \quad \nu = n \quad (3.28)$$

2. Estimate of variance when the true mean  $\mu$  is not known: for the estimate one has to calculate the sample mean which counts as one constraint. Thus the degrees of freedom  $\nu$  is  $n - 1$ :

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{\nu} \sum_{k=1}^n (x_k - \bar{x})^2; \quad \nu = n - 1 \quad (3.29)$$

3. Estimate of variance when the true mean  $\mu$  is known: for the estimate one does not have to calculate the sample mean and thus the degrees of freedom  $\nu$  is equal to  $n$ :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2 = \frac{1}{\nu} \sum_{k=1}^n (x_k - \mu)^2; \quad \nu = n \quad (3.30)$$

4. Two-sample t-test: for the calculation of the test statistic  $t$  one has to calculate two variances without knowing the true means and thus the degrees of freedom  $\nu$  is given by the total number of data  $n = n_1 + n_2$  ( $n_1$  = number of data in sample 1,  $n_2$  = number of data in sample 2) minus 2, i.e.  $\nu = n_1 + n_2 - 2$ .

5. One-sample t-test: for the calculation of the test statistic  $t$  one has to calculate a variance without knowing the true mean and thus the degrees of freedom  $\nu$  is given by number of data  $n$  minus 1, i.e.  $\nu = n - 1$ .

Non-integer values for degrees of freedom occur in the Fisher-Behrens or Welch version of the t-test. They are used when a prerequisite of the two-sample t-test, namely the equality of the true variances, is violated (compare, for instance, Zar, 2010).

**Historical note:** According to Walker (1940), the concept of degrees of freedom was already known to Carl Friedrich Gauß (Carolus Fridericus Gauss, 1777-1855) and applied, for example, 'in his classical work on the *Theory of the Combination of Observations* (Theoria Combinationis Observationum Erroribus Minimis Obnoxiae)'. However, its importance was not recognized in statistics until 1915, when Fisher (1915) and Greenwood

---

<sup>7</sup>If you do not immediately understand the degrees of freedom in more involved situations: Welcome to the club!

& Yule (1915) used degrees of freedom that were different from the number of data. According to David (1995; cited in Zar, 2010, p. 471), the term 'degrees of freedom' was used for the first time in statistics by Fisher (1922b). Fisher used this concept in his book *Statistical Methods for Research Workers* (1925), however the term 'degrees of freedom' does not appear in the index. Although the term was used in statistics textbooks after 1925, the concept was usually not properly explained and not be found in the index. In 1940, Helen Walker wrote a nice article about the history and meaning of the concept. Unfortunately, the situation has not much improved: proper explanations of 'degrees of freedom' are usually missing in textbooks. In the excellent text about the history of statistical concepts and methods by Stigler (1999) one finds a chapter on 'Karl Pearson and Degrees of Freedom (Chapter 19), however, no detailed history of the concept (although p. 348–352 contains interesting information complementary to Walker, 1940).

**Further reading** (degrees of freedom): Good (1973), Yu (2011)



## Chapter 4

# Probabilities: concept, rules, assignment

Probability is an essential concept for data analysis. The first ('classical') definition of probability  $p$  was given by Jacob Bernoulli (1713) as

$$p = \frac{M}{N} = \frac{\text{number of cases favorable to proposition } A}{\text{total number of equally possible cases}} \quad (4.1)$$

An example is  $A = \text{'obtain a sum of 11'}$  when rolling two dice. The number of favorable cases are two, namely (5,6) and (6,5), and the total umber of equally possible cases are  $6 \times 6 = 36$ , namely (1,1), (1,2), ..., (1,6), (2,1), (2,2), ..., (6,6). Thus  $p$  is  $2/36 = 1/18$ . Bernoulli clearly recognized the **limited applicability** of his definition<sup>1</sup> and considered estimating probabilities by observing frequencies in many trials.

Currently there are **various schools of thought** that propose and use different probability concepts: (1) According to Richard von Mises (1919, 1928) probabilities are defined as limits of relative frequencies; for flipping a coin, for example, this would read  $p_{\text{head}} = \lim_{n \rightarrow \infty} \text{number of heads } h / \text{number of flips } n$ . Members of this school are often called frequentists. (2) In 1933, A. Kolmogorov proposed probabilities that are considered as a measure<sup>2</sup> of subsets and that follow certain axioms ('axiomatization of probability theory'). (3) The Bayesian school considers probabilities as plausibilities based on the available, however, usually limited information. Jaynes (2003) derives, following Cox (1946), the basic rules of probabilities (product rule, sum rule, generalized sum rule) from a few desiderata whereby consistency of reasoning plays an essential role.

All three schools of thought agree that probabilities  $p$  are real numbers that can vary between 0 and 1, whereby  $p = 0$  means false/impossible and  $p = 1$  true/certain.

Jaynes considers methods for assigning probabilities such as the Principle of Indifference or the Maximum Entropy Principle as part of probability theory; this is missing from Kolmogorov's approach and goes beyond the frequentist's approach. The name 'Bayesian' for this school of thought is related to Bayes' Theorem, that can be easily derived from the product rule of probabilities and that plays an important role in applications of probability theory to data analysis.

Despite the conceptual differences between the approaches by Kolmogorov and Jaynes, the basic equations (Kolmogorov axioms, basic rules of the Bayesians) can be related to each other (compare discussion in Jaynes, 2003, Appendix A1). The Bayesian approach can be even applied in situations where repetition of experiments is not possible (compare observation of neutrinos from the supernova SN 1987A, Section 1.2) and where frequencies from many trials are not available.

The Bayesian approach is in a sense broader than the other approaches in that it can handle not only problems where randomness plays a role ('stochastic processes'), but also where limited information is due to other reasons. According to Bretthorst (1996) '**probability ... [is] a state of knowledge, not a state of nature**'.

---

<sup>1</sup>'But here, finally, we seem to have met our problem, since this may be done only in a very few cases and almost nowhere other than in games of chance the inventors of which, in order to provide equal chances for the players, took pains to set up so that the numbers of cases would be known and — so that all these cases could happen with equal ease.' Bernoulli cited after Jaynes (1978).

<sup>2</sup>Measure theory is a branch of mathematics that assigns a number to sets and thereby generalizes the intuitive notions of length, area, and volume. It was developed in the late 19th and early 20th century. Measure theory can even deal with somewhat pathological looking sets such as, for example, the Cantor set.

## 4.1 The basic rules of probabilities

There are three basic rules of probabilities: (1) the sum rule, (2) the generalized sum rule, and (3) the product rule. From the product rule one can immediately derive Bayes' Theorem that will later be used in the context of parameter estimation and hypothesis testing.

**Remark:** The discussion of the basic rules of probability is based on the great book by Jaynes (2003).

**Notation:**  $P(A|I)$  is a **conditional probability** for the **proposition A** given the **background information I**. The two 'arguments' of  $P(|)$  are quite different from each other:  $A$  is a proposition for which we seek its probability, whereas  $I$  is true or assumed to be true. The order of the two arguments of  $P(|)$  is essential. The two arguments are divided by a vertical bar | which is read always as 'given', i.e. probability for  $A$  given  $I$  is true.

### 4.1.1 The sum rule of probabilities

The sum rule is arguably the most simple of the three basic rules of probabilities. It states that the sum of the probability for proposition  $A$  and that for not  $A$  add up to 1.

$$P(A|I) + P(\text{not } A|I) = 1 \quad (4.2)$$

or

$$P(A|I) + P(\bar{A}|I) = 1 \quad (4.3)$$

where the bar over  $A$  indicates the **negation** of  $A$ .

### 4.1.2 The generalized sum rule of probabilities

The **generalized sum rule** reads (Jaynes, 2003)

$$P(A \text{ (inclusive) or } B|I) = P(A|I) + P(B|I) - P(A \text{ and } B|I) \quad (4.4)$$

$$P(A \cup B|I) = P(A|I) + P(B|I) - P(A \cap B|I) \quad (4.5)$$

A few remarks are in order:

1. The meaning of 'inclusive or' (represented by the union symbol  $\cup$ ) is as follows:  $A \cup B$  is true if any of the three alternatives is given
  - (a)  $A$  is true and  $B$  is false,
  - (b)  $A$  is false and  $B$  is true
  - (c) Both  $A$  and  $B$  are true

whereas it is false when both  $A$  and  $B$  are false.<sup>3</sup>

2. The generalized sum rule is not just a sum of nonnegative terms (probabilities) but includes also the subtraction of a probability. Why is that necessary?

<sup>3</sup>Note that 'exclusive or' means:  $A$  exclusive or  $B$  is true if any of the two alternatives (a) ' $A$  is true and  $B$  is false' or (b) ' $A$  is false and  $B$  is true' is given and it is false when  $A$  and  $B$  are both true or both false.

3. A simple answer is provided by recalling the constraint that probabilities have to be smaller or equal to 1. If adding up two probabilities, one might end up with a value larger than 1. Thus a sum of two positive terms can not work in general.
4. The generalized sum rule is easy to remember by an analogue from set theory<sup>4</sup> (Fig. 4.1).

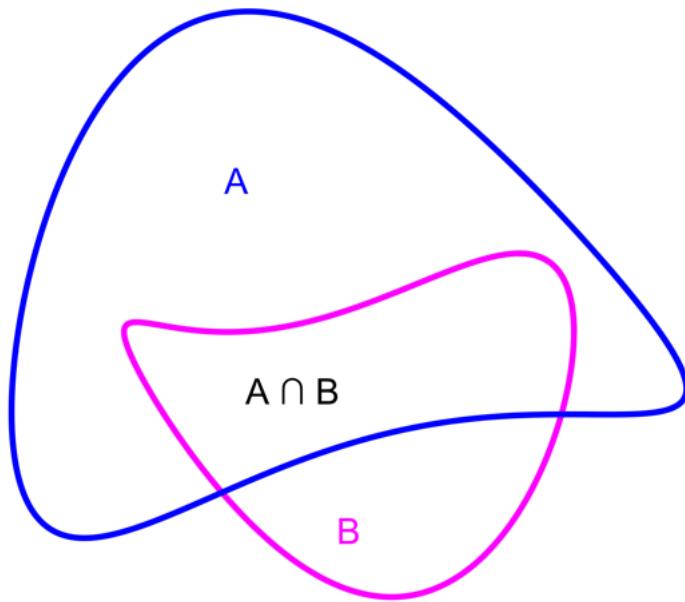


Figure 4.1: Venn diagram as crib (mnemonic) to remember the generalized sum rule of probabilities. One can identify probabilities with areas of sets. Set A is the domain inside the blue curve, set B the inside of the magenta curve. The area of the union between A and B is given by the sum of the area inside the blue curve plus the area inside the red curve minus the overlap between the sets A and B which would be otherwise counted twice, i.e.  $\text{area}_{A \cup B} = \text{area}_A + \text{area}_B - \text{area}_{A \cap B}$  where  $\cap$  is the sign for (inter)section of two sets (the overlap). When replacing 'area' by 'probability', one obtains the generalized sum rule of probabilities. [ProbVenn.R](#)

---

<sup>4</sup>For a critical view of the Venn diagram (Fig. 4.1) compare Jaynes, 2003, p. 47-49.

### 4.1.3 The product rule of probabilities

The **product rule of probabilities** reads (Jaynes, 2003)

$$P(A \text{ and } B|I) = P(B|A \text{ and } I) P(A|I) = P(A|B \text{ and } I) P(B|I) \quad (4.6)$$

or in the notation of set theory ( $\cap$  = section)

$$P(A \cap B|I) = P(B|A \cap I) P(A|I) = P(A|B \cap I) P(B|I) \quad (4.7)$$

or in words

The probability for 'both  $A$  and  $B$  are true' given ' $I$  is true' is the product of the probability for ' $A$  is true' given ' $I$  is true' times the probability for ' $B$  is true' given ' $A$  and  $I$  are true'.

The product rule of probabilities is **symmetric in  $A$  and  $B$**  because 'both  $A$  and  $B$  are true' is the same as 'both  $B$  and  $A$  are true'. This symmetry has been used already in Eqs. 4.6 and 4.7 (second equal signs) and it will be essential for the derivation of Bayes' Theorem.

Various textbooks of statistics give a different (more simple) equation, namely,

$$P(A \cap B|I) = P(B|I) P(A|I) = P(A|I) P(B|I) \quad (4.8)$$

or, leaving out the second argument (which is the same in all probabilities in Eq. 4.8, namely the background information  $I$ ),

$$P(A \cap B) = P(B) \cdot P(A) = P(A) \cdot P(B) \quad (4.9)$$

and also call it the 'product rule' given the restriction of independence between  $A$  and  $B$  or forget to mention this restriction.<sup>5</sup> We will call (Eq. 4.9) the **simplified product rule**. This brings us to the question '**Why is the product rule in general more complicated than Eq. 4.9?**'. The answer is '**Eq. 4.9 is not consistent with common sense**'.

Proposition  $A$  might be very plausible given  $I$ , and  $B$  might be very plausible given  $I$ ; but  $A \cap B$  could still be either very plausible or very implausible. Let us consider the following example: It is quite plausible that the next person you meet has a **left eye that is brown**<sup>6</sup> (proposition  $A$  with  $P(A|I) \approx 9/10$ ). And it is also quite plausible that the next person you meet has a **right eye that is blue** (proposition  $B$  with  $P(B|I) \approx 1/18$ ). But the possibility that the next person you meet has a left brown eye and a right blue eye (proposition  $A \cap B$ ) is not  $1/20$ ; it is almost zero. Almost . . .

<sup>5</sup>They don't tell the reader what to do when  $A$  and  $B$  are not independent of each other.

<sup>6</sup>Here 'brown' refers to a dark eye color. If one considers the worldwide human population, by far the most people possess dark eyes. Blue eye color is very high in Finland and Estonia (about 90%) and rare, for example, in Algeria, Morocco, and Tunisia (below 3%); compare: references in Wikipedia 'Eye color'.



Figure 4.2: David Bowie's eyes & the product rule of probabilities

The product rule of probabilities is applied, for example, in Section 4.2 and in Exercise 63.

In application we will indeed often use the simplified product rule (Eq. 4.8 or Eq. 4.9) because we either **know** or just **assume** that propositions  $A, B, \dots$  are independent of each other. Independence can very much simplify data analysis. However, one has to make sure (for example, in the design of experiments or sampling procedures) that the data are independent of each other.<sup>7</sup>

<sup>7</sup>Independence is usually not given in **time series** (data sampled from a **single system** at successive time points or locations with small spatial distances) and thus analysis of such data require special statistical methods (not covered here; compare, for example, Jenkins & Watts, 1968, Bendat & Piersol, 1972, Dettinger, Michael & Ghil, 1998, Cowpertwait & Metcalfe, 2009, Mudelsee, 2014).

#### 4.1.4 Bayes' Theorem

From the symmetry of the sum rule (Eq. 4.10)

$$P(A \cap B|I) = P(B|A \cap I) P(A|I) = P(A|B \cap I) P(B|I) \quad (4.10)$$

it follows that

$$P(B|A \cap I) = \frac{P(A|B \cap I) P(B|I)}{P(A|I)} \quad \text{Bayes' Theorem} \quad (4.11)$$

This easy to derive and innocent looking formula is called **Bayes' Theorem**. It plays an essential role in estimation of parameters, hypothesis testing, and model selection. In parameter estimation, proposition  $A$  consists of the data (observations) and  $B$  is a model including its model parameters (for example, intercept and slope in simple linear regression). Bayes' Theorem is then used in the form

$$P(\text{parameter}|\text{data}) \propto P(\text{data}|\text{parameter}) \cdot P(\text{parameter}|I) \quad (4.12)$$

where

- **posterior  $P(\text{parameter}|\text{data})$**  = probability density function (PDF) or probability distribution (PD) for the model parameters in the light of data;
- **likelihood  $P(\text{data}|\text{parameter})$**  = PDF or PD for the data given a particular model (here characterized by its parameters); 'how likely is it to observe the data as a sample from a given population?';
- **prior  $P(\text{parameter}|I)$**  = PDF or PD or non-normalized function of parameter(s) describing our knowledge about, respectively ignorance of, the model parameters based on our background knowledge, however, without ('prior to') looking at the data.

In the context of parameter estimation the term  $P(A|I) = P(\text{data}|I)$  is difficult to assign (without a given model one can not assign a PDF or PD for the data) and used only to normalize the posterior (for more details, compare application to simple linear regression in Chapter 14).

**Further reading (Bayes' Theorem):** Efron (2013)

## 4.2 The Monty Hall problem

The Monty Hall problem is used to illustrate the application of the basic rules of probabilities and Bayes' Theorem. In addition, the assignment of probabilities using the Principle of Indifference and the method of marginalization are introduced.

**Monty Hall problem:** "Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the other doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No.2?' Is it to your advantage to take the switch?"<sup>8</sup>

**Music:** The Doors: Strange Days



Figure 4.3: The doors of the Monty Hall problem.

Source: <https://de.cleapng.com/png-pkarfh/download-png.html>

<sup>8</sup>Hint: The host does nothing that would end the game, i.e. he is not opening the candidate's door or opening the door with the car behind.



Figure 4.4: Monty Hall (1975) Source: [www.letsmakeadeal.com/images/mh-1975.jpg](http://www.letsmakeadeal.com/images/mh-1975.jpg)

### 4.2.1 Initial assignment of probabilities

When asked by the host which door to choose, I have no clue where to find the car. So my choice of door No. 1 is as good as any other. I know that the chances to win the car are  $1/3$  or 33%. How do I know? I'm applying the **Principle of Indifference** which states that, when there are  $n$  different cases that are mutually exclusive and exhaustive and when there is no reason to give more or less probability to any one of the cases, all probabilities should be the same and equal to  $1/n$  (compare Section 4.4 for more details). In the current context, mutually exclusive means, when the car is located behind a particular door, it cannot be behind one of the other two; exhaustive means that the car is located for sure behind one of the three doors, there is a single car, and there do not exist more than the three doors in the current context.

Accordingly, I pick one of the doors, say door No. 1. The probability for choosing door No. 1 is denoted by  $P(M_1|I)$  and assigned a value of  $1/3$

$$P(M_1|I) = 1/3. \quad (4.13)$$

It is a conditional probability for the proposition ' $M_1$  = my choice is door No. 1' given the background information  $I$  (= rules of the game, including the information of 1 car behind one of 3 doors). Please note that the statements behind the vertical bar are true or supposed to be true. The probabilities of me choosing the other doors are  $1/3$  as well:

$$P(M_2|I) = 1/3 = P(M_3|I). \quad (4.14)$$

My choice (proposition  $M_1$ ) has no influence on the probability that the car is behind any particular door (proposition  $C_v$ ,  $v = 1, 2$ , or 3) and thus all probabilities are equal (please check that the three probabilities are mutually exclusive and exhaustive):

$$P(C_1|M_1 \cap I) = P(C_2|M_1 \cap I) = P(C_3|M_1 \cap I) = 1/3. \quad (4.15)$$

Please note, that  $M_1$  is now behind the vertical bar and thus considered true or a given fact.  $P(C_1|M_1 \cap I)$  is the probability for the proposition 'car is behind door No. 1' given the background information  $I$  and my choice of door No. 1 ( $M_1$ ).

### 4.2.2 The host opens a door: assignment using additional information

After my choice of door No. 1 the host tries to help me by opening door No. 3 exposing a goat (Fig. 4.5). Thus only two closed doors are left. My chances to win the car have been increased. Two doors could suggest a 50-50 chance. Let's see ... We are left with two propositions ( $C_1$  = car is behind door No. 1,  $C_2$  = car is behind door No. 2) and we have additional information ( $S_3$  = host opened door No. 3 with a goat behind) compared to the initial situation. How will this change the assignment of probabilities?

We are left with two probabilities:

$P(C_1|M_1 \cap S_3 \cap I)$  = probability for proposition 'car is behind No. 1' given my choice was door No. 1 ( $M_1$ ), the host opened door No. 3 ( $S_3$ ), and the background information  $I$  and

$P(C_2|M_1 \cap S_3 \cap I)$  = probability for proposition 'car is behind No. 2' given my choice was door No. 1 ( $M_1$ ), the host opened door No. 3 ( $S_3$ ), and the background information  $I$ .

The two cases 'car is behind door No. 1' and 'car is behind door No. 2' are mutually exclusive and exhaustive and thus the associated probabilities have to add up to 1:

$$P(C_1|M_1 \cap S_3 \cap I) + P(C_2|M_1 \cap S_3 \cap I) = 1. \quad (4.16)$$

The assignment of these probabilities is not that obvious because my choice of door No. 1 could have had an influence on the host: if the car is indeed behind door No. 1 he would have had to choose between opening of door No. 2 or No. 3, whereas if the car is behind door No. 2 he would have had no choice as to open door No. 3 because door No. 1 is 'blocked' by my choice and No. 2 is 'blocked' by the car.



Figure 4.5: The goat behind door No. 3.

### 4.2.3 Apply Bayes' Theorem and marginalization

For the assignment of the probability  $P(C_2 | M_1 \cap S_3 \cap I)$  we will apply Bayes' Theorem

$$P(B | A \cap I) = \frac{P(A | B \cap I) P(B | I)}{P(A | I)} \quad (4.17)$$

with the following substitutions

$$B \Rightarrow C_2 \quad (4.18)$$

$$A \Rightarrow S_3 \quad (4.19)$$

$$I \Rightarrow M_1 \cap I \quad (4.20)$$

(my choice of door No. 1,  $M_1$ , is combined with the background information,  $I$ ) yielding

$$P(C_2 | S_3 \cap M_1 \cap I) = \frac{P(S_3 | C_2 \cap M_1 \cap I) P(C_2 | M_1 \cap I)}{P(S_3 | M_1 \cap I)}. \quad (4.21)$$

We hope that we will be able to assign values to the probability terms on the right-hand-side of Eq. 4.21. We assigned already  $P(C_2 | M_1 \cap I) = 1/3$ .

The most difficult task is assigning a value to  $P(S_3 | M_1 \cap I)$  = probability for the proposition 'host opens door No. 3' given my choice of door No. 1,  $M_1$ , and the background information,  $I$ . We will apply a method called **marginalization** based on the fact that the probability for a proposition (here:  $S_3$ ) does not change by combining ('section') it with another proposition that is true for sure (here: the car is behind door No. 1, 2, or 3 =  $C_1 \cup C_2 \cup C_3$ ; note that  $C_1$ ,  $C_2$ , and  $C_3$  are mutually exclusive and exhaustive):

$$P(S_3 | M_1 \cap I) = P(S_3 \cap (C_1 \cup C_2 \cup C_3) | M_1 \cap I) \quad (4.22)$$

The right-hand-side can be split by applying the generalized sum (a) and the product rule (b) of probabilities:

$$P(S_3 | M_1 \cap I) = P(S_3 \cap (C_1 \cup C_2 \cup C_3) | M_1 \cap I) \quad (4.23)$$

$$\underbrace{=}_{(a)} P(S_3 \cap C_1 | M_1 \cap I) + P(S_3 \cap C_2 | M_1 \cap I) + P(S_3 \cap C_3 | M_1 \cap I) \quad (4.24)$$

$$\underbrace{=}_{(b)} P(S_3 | C_1 \cap M_1 \cap I) P(C_1 | M_1 \cap I) + P(S_3 | C_2 \cap M_1 \cap I) P(C_2 | M_1 \cap I) \\ + P(S_3 | C_3 \cap M_1 \cap I) P(C_3 | M_1 \cap I). \quad (4.25)$$

Of the six terms on the right-hand-side we assigned already values to the following three (Eq. 4.15):

$$P(C_1 | M_1 \cap I) = P(C_2 | M_1 \cap I) = P(C_3 | M_1 \cap I) = 1/3. \quad (4.26)$$

Up to now we could always apply the Principle of Indifference and thus all assignments are equal to 1/3.

Now we have to assign values to  $P(S_3 | C_k \cap M_1 \cap I)$  with  $k = 1, 2, 3$ :

$P(S_3 | C_1 \cap M_1 \cap I)$  is the probability that the host opens door No. 3 ( $S_3$ ) given the fact that the car is behind door No. 1 ( $C_1$ ), that my first choice was door No. 1 ( $M_1$ ), and given the rules of the game ( $I$ ): this probability is 1/2 because the host had to choose between door No. 2 and No. 3; please note that in the current context doors No. 2 and No. 3 are mutually exclusive and exhaustive and the Principle of Indifference can be applied.  $\Rightarrow P(S_3 | C_1 \cap M_1 \cap I) = 1/2$ .

$P(S_3 | C_2 \cap M_1 \cap I)$  is the probability that the host opens door No. 3 ( $S_3$ ) given the fact that the car is behind door No. 2 ( $C_2$ ), that my first choice was door No. 1 ( $M_1$ ), and given the rules of the game ( $I$ ): this probability is 1 because the host has to open door No. 3 (door No. 1 is 'my door' and the car is behind door No. 2).  $\Rightarrow P(S_3 | C_2 \cap M_1 \cap I) = 1$ .

$P(S_3 | C_3 \cap M_1 \cap I)$  is the probability that the host opens door No. 3 ( $S_3$ ) given the fact that the car is behind door No. 3 ( $C_3$ ), that my first choice was door No. 1 ( $M_1$ ), and given the rules of the game ( $I$ ): this probability

is 0 because the host never opens the door with the car behind.  $\Rightarrow P(S_3|C_3 \cap M_1 \cap I) = 0$ . Combining the six probabilities, one obtains

$$\begin{aligned} P(S_3|M_1 \cap I) &= P(S_3|C_1 \cap M_1 \cap I) P(C_1|M_1 \cap I) + P(S_3|C_2 \cap M_1 \cap I) P(C_2|M_1 \cap I) \\ &\quad + P(S_3|C_3 \cap M_1 \cap I) P(C_3|M_1 \cap I) \\ &= (1/2 + 1 + 0) 1/3 = 1/2. \end{aligned} \quad (4.27)$$

#### 4.2.4 Final result & discussion

Inserting all terms into Bayes' Theorem yields

$$P(C_2|M_1 \cap S_3 \cap I) = \frac{P(S_3|C_2 \cap M_1 \cap I) P(C_2|M_1 \cap I)}{P(S_3|M_1 \cap I)} \quad (4.28)$$

$$= \frac{1 \cdot 1/3}{1/2} = \frac{2/3}{1/2} \quad (4.29)$$

and thus

$$P(C_1|M_1 \cap S_3 \cap I) = 1 - P(C_2|M_1 \cap S_3 \cap I) = 1/3. \quad (4.30)$$

#### If I change from door No. 1 to No. 2, my chance to win the car doubles!

How can we grasp this unexpected result?

"Suppose that the host always opens a door with a goat. If the player's first door is incorrect (contains a goat), then the host has no choice and must open the other door with a goat. Then, if the player switches, she wins. On the other hand, if the player's first door is correct and she switches, then of course she loses. Thus, we see that if the player always switches, then she wins if and only if her first choice is incorrect, an event that obviously has probability 2/3. If the player never switches, then she wins if and only if her first choice is correct, an event with probability 1/3." [http://www.ds.unifi.it/VL/VL\\_EN/games/games6.html](http://www.ds.unifi.it/VL/VL_EN/games/games6.html)

1. Repeat the game 100 times!
2. A game with 100 doors: You choose door No. 1 and the host opens 98 of the other doors except for door 53. Would you change?

Suppose a guest comes in after the host has already opened door No. 3. The late guest does not know anything about what happened before. Which door (No. 1 or No. 2) should he pick?

The host, I, and the late guest have different information and thus different probabilities are assigned:

1. The host knows where the car is located and thus he assigns probabilities 1 (certain) or 0 (impossible).
2. Before the host opens a door I have no clue and thus I assign equal probabilities (1/3) to all doors.
3. After the host has opened a door, I change probabilities because there are less doors to choose between (this suggests the fifty-fifty answer which is wrong) and because of the action of the host that is constrained by my initial choice.
4. The late guest has no clue (same as I initially) in front of two (instead of three) doors: he has a fifty-fifty chance.

### 4.2.5 Take-home message: probability rules & Monty Hall

Generalized sum rule (memorize using the Venn diagram, Fig. 4.1):

$$P(A \cup B|I) = P(A|I) + P(B|I) - P(A \cap B|I)$$

Product rule:

$$P(A \cap B|I) = P(B|A \cap I) P(A|I) = P(A|B \cap I) P(B|I)$$

Bayes' Theorem (follows from symmetry of product rule):

$$P(B|A \cap I) = \frac{P(A|B \cap I) P(B|I)}{P(A|I)}$$

Principle of Indifference:  $P = 1/n$  (equal probabilities)

Marginalization

**Further reading:** [https://en.wikipedia.org/wiki/Monty\\_Hall\\_problem](https://en.wikipedia.org/wiki/Monty_Hall_problem)  
Cox (1946), Jaynes (2003), Pearl & Mackenzie (2019)

### 4.3 Assigning probabilities

Assigning probabilities or probability densities to propositions is an [open-ended problem](#) (Jaynes, 2003). However, various methods exist for assigning probabilities or probability densities in particular circumstances. The [Principle of Indifference](#) can often be applied to discrete propositions, and, with some care also to assign probability densities for continuous systems. The [Maximum Entropy Principle \(MaxEnt\)](#) is an extremely powerful method to assign probabilities and probability densities taking into account various constraints such as normalization, mean value, variance, or, other expectations. It can be shown that the Principle of Indifference follows from MaxEnt. Even more general than MaxEnt is the principle of Maximum Relative Entropy (MaxRelEnt) (Fig. 4.6) which can take into account priors. MaxRelEnt is discussed in the Appendix (Section A.9).

MaxEnt was proposed by Jaynes (1957a,b). It was first based on the Shannon entropy (Shannon, 1948). Jaynes (2003) showed that various approaches alternative to the Shannon entropy do not work (for simple examples this would lead to assignments of negative values which is impossible for probabilities) whereas MaxEnt always led to plausible results. MaxEnt was applied also to probability density functions. MaxRelEnt was proposed by Kullback & Leibler (1951; Kullback, 1959). Finally, an axiomatic derivation of both MaxEnt and MaxRelEnt for both probability distributions and probability density functions was given by Shore & Johnson (1980).

What else can be found in our toolbox for assigning probabilities? The basic rules of probabilities discussed in the previous section! In combination with the Principle of Indifference one can derive, for example, the binomial and the hypergeometric probability distribution; for details compare, for example, Jaynes (2003, Chapter 3).

**Movie:** Whatever works (Woody Allan, 2009)

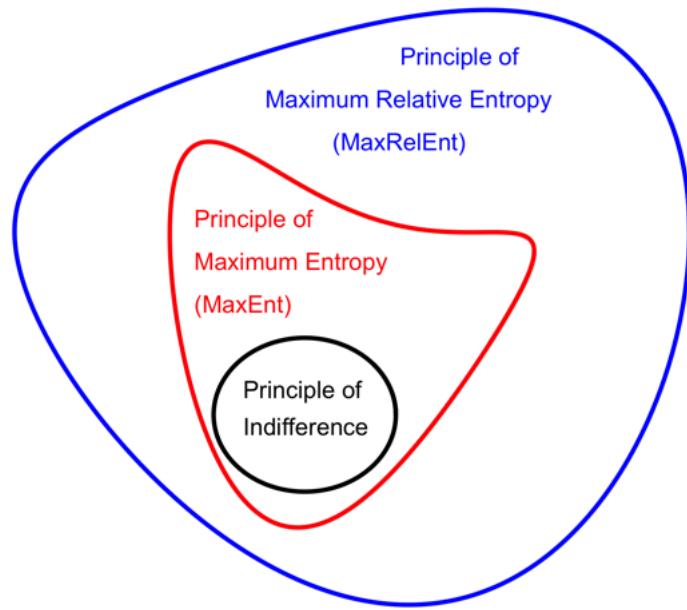


Figure 4.6: Range of application of various principles to assign probability distributions. Most general is the Principle of Maximum Relative Entropy (MaxRelEnt). From this principle one can derive the Principle of Maximum Entropy (MaxEnt) by assuming a flat prior. The Principle of Indifference has a much smaller range of application in that it can provide uniform distributions only. Its application in case of continuous probability distributions requires great care (compare, for example, the discussion of Bertrand's problem by Jaynes, 2003). [ProbMaxEntPrinciples.R](#)

## 4.4 The Principle of Indifference

The Principle of Indifference<sup>9</sup> states that, if there is no reason to assign more or less probability to any proposition (outcome) under consideration, then all probabilities should be equal, and, if there are in total  $n$  mutually exclusive and exhaustive propositions (outcomes), the probability for each single proposition is  $1/n$  (discrete uniform distribution; sum over all probabilities has to be 1). Although the name of this principle is not widely known, we apply it every day because it is part of our '[common sense](#)'.

**Remark:** The principle is based on [symmetry](#) (no variation under interchange of proposition).

**Application of the Principle of Indifference (examples):**

1. Flipping a coin: head and tail are mutually exclusive (head and tail can not occur at the same time) and exclusive (there are no other possible outcomes here; if the coin ends up lying on its edge, it is not considered as a proper flip); if there is no reason for assuming higher or lower probability for head ( $p_{\text{head}}$ ) compared to tail ( $p_{\text{tail}}$ ), both probabilities should be equal; one speaks of an 'unbiased coin';  $\Rightarrow p_{\text{head}} = p = 1/2 = p_{\text{tail}}$ .<sup>10</sup>
2. Unbiased die: the 6 possible outcomes '1', '2', ..., '6' are mutually exclusive and exhaustive; if there is no reason to assign more or less probability to one of the 6 possible outcomes, all probabilities should be equal  $\Rightarrow p = 1/6$ .
3. Monty Hall problem (before the host opens one of the three doors): A host stands in front of three closed doors. We are told that there is a car behind one of the three doors and goats behind the two other doors; the proposals 'car is behind door  $k$ ', where  $k$  can be 1, 2, or 3, are mutually exclusive and exhaustive; we have no clue (smell, noise, ...) where to find the car or the goats and thus we assign equal probabilities  $p_k$  to the proposals 'car is behind door  $k$ '  $\Rightarrow p_k = 1/3$ .

---

<sup>9</sup>John Maynard Keynes (1921) renamed the 'Principle of Insufficient Reason' to 'Principle of Indifference'. In German the old name 'Prinzip vom unzureichenden Grunde' is still in use.

<sup>10</sup>In reality, tossing coins is more tricky: 'Thus, anyone familiar with the law of conservation of angular momentum can, after some practice, cheat at the usual coin-tossing and call his shots with 100% accuracy. You can obtain any frequency of heads you want – *and the bias of the coin has no influence at all on the results!*' (Jaynes, 2003)

## 4.5 Maximum Entropy Principle

The Shannon entropy of a (discrete) probability distribution  $p_j, j = 1, 2, \dots, L$  (with  $L$  either finite or  $\infty$  as, for example, in the Poisson distribution) is defined by

$$S = - \sum_j p_j \ln p_j \quad (4.31)$$

The Principle of Maximum Entropy states that probability distributions (PDs) can be found in case of limited information by maximizing the Shannon entropy  $S$  (Eq. 4.31) under constraints. An obvious constraint for a PD is its normalization, i.e. the sum over all probabilities has to be 1. Further constraints can be the mean or the variance of the PD (in general: expectations, Section 6.2). Coupling of all constraints by Lagrange multipliers to the Shannon entropy  $S$  yields the so-called Lagrange equation  $\mathcal{L}$  (compare examples below).

MaxEnt can also be applied for the assignment of PDFs. For this aim, one has to replace the (discrete) Shannon entropy by

$$S_c = - \int p(x) \ln p(x) dx \quad (4.32)$$

On first sight the MaxEnt method for assigning probabilities might look to be obscure or, at least, seems to use a bit of magic. However, the examples discussed below will hopefully convince you that it is quite useful because it yields valuable results.

### 4.5.1 Tossing a coin

Consider a coin with probabilities  $p_{\text{head}} \equiv p_1$  and  $p_{\text{tail}} \equiv p_2$  for head and tail, respectively. The Shannon entropy for a probability distribution with two cases reads

$$S = - \sum_{j=1}^2 p_j \ln p_j = -p_1 \ln p_1 - p_2 \ln p_2 \quad (4.33)$$

Note that we use the natural logarithm here because this will make differentiation with respect to the  $p_j$  easier. Which probability distribution would result from maximizing  $S$ ? A necessary condition for a maximum is the vanishing of the first derivatives of the entropy  $S(p_1, p_2)$  with respects to its arguments, namely  $p_1$  and  $p_2$ :

$$\left( \frac{\partial S(p_1, p_2)}{\partial p_1} \right)_{p_2 = \text{constant}} = -\ln p_1 - \frac{p_1}{p_1} = -1 - \ln p_1 \stackrel{!}{=} 0 \quad (4.34)$$

$$\left( \frac{\partial S(p_1, p_2)}{\partial p_2} \right)_{p_1 = \text{constant}} = -\ln p_2 - \frac{p_2}{p_2} = -1 - \ln p_2 \stackrel{!}{=} 0 \quad (4.35)$$

Thus the solution reads

$$p_1 = e^{-1} = p_2 \approx 0.3679 \quad (4.36)$$

The good news is: the probabilities are equal to each other (as expected by the Principle of Indifference) and in the allowed range between zero and one. However, the probabilities do not add up to one as required for a probability distribution. What is missing? Obviously, it's the normalization constraint

$$\sum_j p_j = 1 = p_1 + p_2 \quad (4.37)$$

How can the normalization constraint be taken in account? One couples the constraint by a Lagrange multiplier  $\lambda$  to the entropy and obtains the Lagrange function  $\mathcal{L}(p_1, p_2; \lambda)$

$$\mathcal{L}(p_1, p_2; \lambda) = S + \lambda \sum_{j=1}^2 p_j = -p_1 \ln p_1 - p_2 \ln p_2 + \lambda (p_1 + p_2) \quad (4.38)$$

Note that one can leave out additive constants like, for example, the 1 of the normalization constraint, because such constants have no influence on the derivatives of  $\mathcal{L}$  with respect to  $p_j$ . The Lagrange function  $\mathcal{L}$  depends on two variables, namely  $p_1$  and  $p_2$  and one Lagrange parameter  $\lambda$ . A necessary condition for a maximum is the vanishing of the first derivatives of the Lagrange function  $\mathcal{L}(p_1, p_2; \lambda)$  with respects to the probabilities  $p_j$ ,  $j = 1, 2$ :

$$\left( \frac{\partial \mathcal{L}(p_1, p_2; \lambda)}{\partial p_1} \right)_{p_2 \text{ and } \lambda \text{ constant}} = -1 - \ln p_1 + \lambda \stackrel{!}{=} 0 \quad (4.39)$$

$$\left( \frac{\partial \mathcal{L}(p_1, p_2; \lambda)}{\partial p_2} \right)_{p_1 \text{ and } \lambda \text{ constant}} = -1 - \ln p_2 + \lambda \stackrel{!}{=} 0 \quad (4.40)$$

The two equation can be easily solved for  $p_j$

$$p_j = p_j(\lambda) = e^{-1+\lambda} \quad j = 1, 2 \quad (4.41)$$

The right hand side of this equation is independent of  $j$  ( $\lambda$  is a constant not depending on  $j$ ) and thus all probabilities  $p_j$  are equal:  $p_1 = p_2$ . Because of the normalization constraint ( $p_1 + p_2 = 1$ ) they are equal to  $1/2$ . More formally<sup>11</sup> one can calculate the value of the Lagrange multiplier from the constraint:

$$1 = \sum_{j=1}^2 p_j = \sum_{j=1}^2 e^{-1+\lambda} = 2 e^{-1+\lambda} \quad (4.42)$$

$$\Rightarrow e^{-1+\lambda} = 1/2 \Rightarrow -1 + \lambda = \ln(1/2) = -\ln 2 \quad (4.43)$$

and finally

$$\lambda = 1 - \ln 2 \approx 0.3069 \quad (4.44)$$

and

$$p_j = e^{-1+\lambda} = e^{-\ln 2} = 2^{-1} = 1/2 \quad (4.45)$$

**Take-home message from the coin example** The uniform distribution of probabilities,  $p_j$ , can be obtained by maximizing the Lagrange function

$$\mathcal{L}(p_1, p_2; \lambda) = - \sum_{j=1}^2 p_j \ln p_j + \lambda \sum_{j=1}^2 p_j \quad (4.46)$$

which is the sum of the Shannon entropy and the product of a Lagrange multiplier,  $\lambda$ , times the normalization condition for probability distributions.

In the case of the coin the result from applying MaxEnt is consistent with common sense and identical to the result from the Principle of Indifference. However, the Principle of Maximum Entropy has a much wider range of application than the Principle of Indifference.

### 4.5.2 The unbiased die

Consider a common die with 6 faces.  $p_j$  is the probability to obtain face  $j$  when tossing the die. If we have no additional information about the die, we can apply MaxEnt:

$$\mathcal{L} = \underbrace{- \sum_{j=1}^6 p_j \ln p_j}_{\text{Shannon entropy}} + \lambda \underbrace{\sum_{j=1}^6 p_j}_{\text{normalization constraint}} \quad (4.47)$$

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - 1 + \lambda \stackrel{!}{=} 0 \quad (4.48)$$

---

<sup>11</sup>This is superfluous here but required for more complicated problems.

$$p_j = p_j(\lambda) = e^{-1+\lambda} \quad (4.49)$$

from which we can infer immediately that all  $p_j$  are all equal to each other and thus  $p_j = 1/6$  (from normalization).

More formally we can calculate the value of the Lagrange multiplier from the constraint:

$$1 = \sum_{j=1}^6 p_j = \sum_{j=1}^6 e^{-1+\lambda} = 6e^{-1+\lambda} \quad (4.50)$$

$$\Rightarrow e^{-1+\lambda} = 1/6 \Rightarrow -1 + \lambda = \ln(1/6) = -\ln 6 \quad (4.51)$$

and finally

$$\lambda = 1 - \ln 6 = -0.7918 \quad (4.52)$$

and

$$p_j = e^{-1+\lambda} = e^{-\ln 6} = 6^{-1} = 1/6 \quad (4.53)$$

This is another trivial example for the application of MaxEnt because it yields the identical result as the Principle of Indifference.

### 4.5.3 A loaded die

So far we have solved problems that could be solved much faster by application of the Principle of Indifference. The Maximum Entropy Principle yielded identical results. Now we will consider problems where not all probabilities are equal to each other and thus the Principle of Indifference is not applicable anymore.

A simple example is a loaded die with a mean  $\mu = 3.8$  instead of  $\mu = 3.5$  for the unbiased die. In order to obtain a mean above 3.5 one or more of the probabilities  $p_4$ ,  $p_5$ , and  $p_6$  have to be larger than 1/6. The constraint 'mean  $\mu = 3.8$ ' is not enough to uniquely assign all six probabilities. The following three probability distributions, for example, all fulfill the constraint  $\mu = 3.8$ :

$$p_a = \{0.1467, 0.1467, 0.1467, 0.1467, 0.1467, 0.2667\} \quad (4.54)$$

$$p_b = \{0.1417, 0.1417, 0.1417, 0.1417, 0.2167, 0.2167\} \quad (4.55)$$

$$p_c = \{0.1333, 0.1333, 0.1333, 0.2, 0.2, 0.2\} \quad (4.56)$$

Further solutions are calculated in Exercise 55.

**What do we want?** From the three distributions given above we can conclude that the problem of assigning probability distributions that respect certain constraints has **no unique solution**. In the loaded die example we used simple **arbitrary** procedures to produce solutions resulting in relative large deviations of single probabilities from the distribution for the unbiased case.

**What we need is a general procedure that**

1. can handle almost any constraint,
2. distributes the burden of change over the whole distribution and thereby
3. limits the amount of change for each individual  $p_j$ ,
4. ensures non-negative values for all  $p_j$ ,
5. does not include arbitrary choices not requested by constraints.

Let's see how the Maximum Entropy Principle performs.

**Solution of the loaded die problem by application of the Maximum Entropy Principle**

The Shannon entropy plus constraints for the loaded die problem leads to the Lagrange function

$$\mathcal{L}(p_1, \dots, p_6; \lambda_1, \lambda_2) = \underbrace{-\sum_{j=1}^6 p_j \ln p_j}_{\text{Shannon entropy}} + \lambda_1 \underbrace{\sum_{j=1}^6 p_j}_{\text{normalization}} + \lambda_2 \underbrace{\sum_{j=1}^6 j \cdot p_j}_{\text{mean value}} \quad (4.57)$$

where  $\lambda_1$  and  $\lambda_2$  are the Lagrange multipliers.

The first derivatives of  $\mathcal{L}$  with respect to the probabilities  $p_j$  are set to zero (necessary condition for maximum)

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - 1 + \lambda_1 + j \lambda_2 = 0. \quad (4.58)$$

These equations are readily solved:

$$p_j = p_j(\lambda_1, \lambda_2) = e^{-1+\lambda_1+j\lambda_2} \quad (4.59)$$

**Note that the solution for the  $p_j$  in the form of exponential functions ensures non-negative values for all probabilities.** For  $\lambda_2 \neq 0$  the probabilities  $p_j$  vary with  $j$ .

**Calculation of the Lagrange multipliers** The Lagrange multipliers  $\lambda_1$  and  $\lambda_2$  can be calculated<sup>12</sup> from the constraints

$$\sum_{j=1}^6 p_j = \sum_{j=1}^6 e^{-1+\lambda_1+j\lambda_2} = 1 \quad (4.60)$$

$$\sum_{j=1}^6 j \cdot p_j = \sum_{j=1}^6 j \cdot e^{-1+\lambda_1+j\lambda_2} = \mu = 3.8 \quad (4.61)$$

These two equations form a coupled system ( $\lambda_1$  and  $\lambda_2$  occur in both equations) of nonlinear equations ( $\lambda_1$  and  $\lambda_2$  occur as arguments of exponential functions). Systems of non-linear equations are generally very difficult to solve<sup>13</sup>. The details of calculating the Lagrange multipliers are discussed in the Appendix (Section A.3). The results are displayed in Fig. 4.7. The deviations from 1/6 are distributed over all probabilities, they are relative small and do not vary in a linear way.

**Conclusions from the loaded die example:**

**MaxEnt works very well in the case of the loaded die (where the Principle of Indifference is not applicable). It is able**

1. to handle both constraints (normalization, mean) simultaneously,
2. to distribute the burden of change over the whole distribution and thereby
3. to limit the amount of change for each individual  $p_j$ ,
4. to ensure non-negative values for all  $p_j$  (the exponential function yields non-negative values for all real arguments),

**and does not include arbitrary choices not requested by constraints.**

<sup>12</sup>Without proof: the solution is unique.

<sup>13</sup>"There are *no* good, general methods for solving systems of more than one nonlinear equation. Furthermore, it is not hard to see why (very likely) there *never will be* any good, general methods ..." (Press et al., 1986, p. 269)

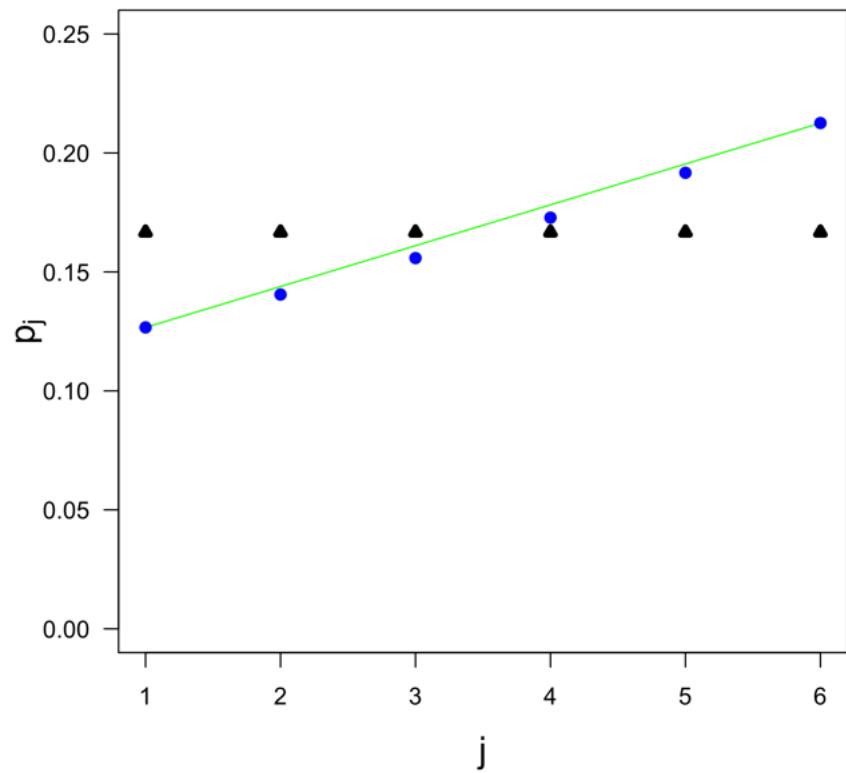


Figure 4.7: The probabilities for the load die derived from MaxEnt read  $p = \{0.1267, 0.1405, 0.1558, 0.1728, 0.1917, 0.2126\}$  (blue dots). The deviations from the probabilities for the unbiased die (black triangles) are small and do not vary in a linear way (the probabilities for the loaded die do not all lie on the straight line connecting  $p_1$  and  $p_6$ , green solid line). [ProbLagrangeMultipliersLoadedDie.R](#)

#### 4.5.4 The discrete exponential distribution

We seek for a PD that can take non-negative values  $j$ , i.e.  $j = 0, 1, 2, \dots$  and has mean value  $\mu$ . The Lagrange function reads

$$\mathcal{L} = \sum_{j=0}^{\infty} (-p_j \ln p_j + \lambda_0 p_j + \lambda_1 j p_j). \quad (4.62)$$

A necessary condition for a maximum is the vanishing of the partial derivatives of  $\mathcal{L}$  with respect to the probabilities  $p_j$ :

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - \underbrace{p_j \frac{1}{p_j}}_{=-1} + \lambda_0 + \lambda_1 j = 0 \quad \text{for } j = 0, 1, 2, \dots \quad (4.63)$$

and thus

$$p_j = e^{\lambda_0 - 1 + \lambda_1 j} \quad (4.64)$$

The Lagrange multipliers are calculated from the constraints

$$\sum_{j=0}^{\infty} p_j = \sum_{j=0}^{\infty} e^{\lambda_0 - 1 + \lambda_1 j} = 1 \quad (4.65)$$

$$\sum_{j=0}^{\infty} j p_j = \sum_{j=0}^{\infty} j e^{\lambda_0 - 1 + \lambda_1 j} = \mu \quad (4.66)$$

$$(4.67)$$

(compare Section A.5 for details). Finally one obtains

$$p_j = \frac{1}{1+\mu} \left( \frac{\mu}{1+\mu} \right)^j \quad j = 0, 1, 2, \dots \quad (4.68)$$

The probabilities  $p_j, j = 0, 1, 2, \dots, 20$  for  $\mu = 2.5$  are shown in Fig. 4.8.

## 4.6 Maximum Entropy Principle applied to PDFs

So far we have applied the Maximum Entropy Principle (MaxEnt) to discrete problems, i.e. to derive probability distributions. However, the range of MaxEnt is much wider as shown by the following example. Here we ask: What is the form of a continuous distribution  $p(x)$  characterized by its mean,  $\mu$ , and its variance,  $\sigma^2$ , alone?

Usually continuous distributions cannot be characterized by just two of their moments (mean, variance). It is easy to find or construct distributions with the same mean and variance that differ in other moments. The construction of a distribution from a limited number of moments is a classical example of insufficient knowledge and lends itself to the application of the Maximum Entropy Principle.

The main result of the section can be formulated as a 'theorem':

"The normal distribution  $\mathcal{N}(\mu, \sigma^2)$  has maximum entropy among all real-valued distributions with specified mean  $\mu$  and standard deviation  $\sigma$ . Therefore, if all you know about a distribution is its mean and standard deviation, it is often reasonable to assume that the distribution is normal." (Wikipedia, Maximum entropy probability distribution)

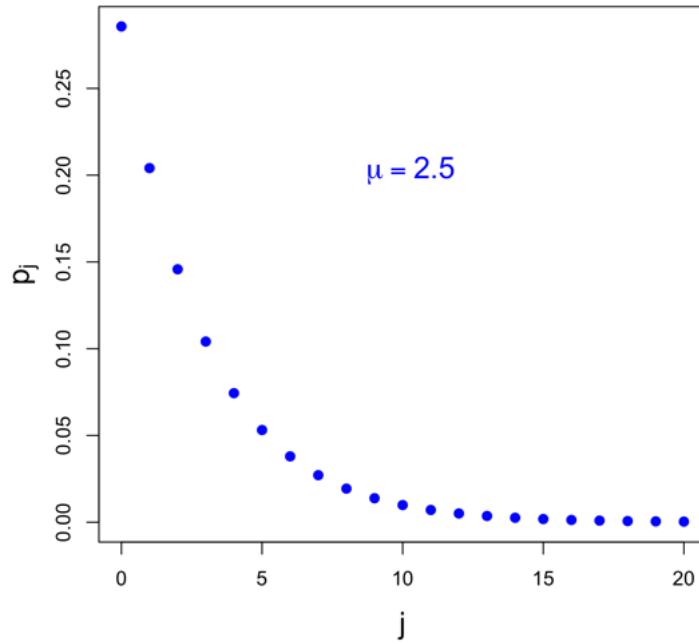


Figure 4.8: Discrete exponential PD: the probabilities  $p_j, j = 0, 1, 2, \dots, 20$  for  $\mu = 2.5$ . [MaxEntExpPD.R](#)

### Exercise 7 Joint birthday?

A randomly composed group of  $N = 26$  people meet in a 'Marine Biology' course at the University of Bremen. What's the probability of at least two persons have birthdays on the same date (date = same day, omitting year of birth)?

Remarks:

- (1) Exclude twins  $\Rightarrow$  birthdays are independent of each other.
- (2) For the sake of simplicity February 29 does not exist, i.e. the year has 365 days.
- (3) Assume that birthdays are uniformly distributed.

Hint: first calculate the probability that the dates of birthdays are different.

### Exercise 8 Joint birthdays (more than 2)

A randomly composed group of  $N = 26$  people meet in a 'Marine Biology' course at the University of Bremen. What's the probability of two persons have birthdays on the same date (date = same day, omitting year of birth) and another two persons have birthdays on another same date or that four persons have birthdays on the same date?

Remarks:

- (1) Exclude twins  $\Rightarrow$  birthdays are independent of each other.
- (2) For the sake of simplicity February 29 does not exist, i.e. the year has 365 days.
- (3) Assume that birthdays are uniformly distributed.

Hint: estimate probability by Monte Carlo simulation.

### Exercise 9 Double six in 24 trials?

Suppose you role a pair of fair (unbiased) dice 24 times. What is the probability to obtain a double six at least one time? Which basic rules of probability can be applied to solve this problem?



# Chapter 5

## Random numbers

More correctly, this chapter should read ‘generation of pseudo-random numbers’. The random numbers are generated by computer algorithms.

### 5.1 Generation of (pseudo-)random numbers in R

R provides various routines to generate random numbers from standard PDs (discrete uniform, Poisson, binomial, ...) and PDFs (uniform, normal,  $t$ ,  $F$ ,  $\chi^2$ , ...) by a simple call, i.e. one line of code. The generation of random numbers from other distributions is discussed in the Appendix (Section B.1).

**set.seed(seed, kind = NULL, normal.kind = NULL)**

**set.seed(1953)** sets a seed for the generation of (pseudo-)random numbers

Random numbers generated by computers are not truly random because they are generated by deterministic algorithms and thus they are more precisely called pseudo-random numbers. The various algorithms for random number populations require one or more initial values that are either created from the current time and the process ID (when no seed is set) or by calling `set.seed(k)` where  $k$  is an integer. In the first case, the initial values of the random number generators change every time you call a routine for generating random numbers as, for example, `rnorm(10)` and you will get different random numbers again and again with each call of `rnorm(10)`. On the other hand, by calling **set.seed(k)** with a particular integer  $k$  one will always get the identical sequence of pseudo-random numbers. In applications as, for example, Monte Carlo simulations, it is not necessary to call `set.seed`. However, we will call `set.seed` in the text most of the time in order to obtain reproducible (identical) results<sup>1</sup>.

**sample(x,size,replace=TRUE)**

`xrange = seq(1,6); sample(x=xrange,1000,replace=TRUE)` generates 1000 random numbers from the (discrete) uniform PD for event space  $\{1, 2, 3, 4, 5, 6\}$  of rolling a fair die (equal probabilities).

**rpois(n, lambda)**

`rpois(5, 2.3)` generates 5 random numbers from the **Poisson distribution** with mean rate of events  $\lambda = 2.3$ .

**rbinom(n,size,prob)**

`rbinom(5,3,0.6)` generates 5 random numbers from the **binomial distribution** with number of trials  $k = 3$  and probability of success in a single trial  $p = 0.6$ .

**runif(n, min = 0, max = 1)**

`runif(5)` generates 5 random numbers from the **uniform PDF** between 0 and 1

`runif(7,min=1.8, max=1.9)` generates 7 random numbers from the uniform PDF between 1.8 and 1.9

**rnorm(n, mean = 0, sd = 1)**

`rnorm(5)` generates 5 random numbers from the **standard normal distribution**

---

<sup>1</sup>This is extremely useful when teaching and all students work on their own computers.

`rnorm(6,1,3)` generates 6 random numbers from the [normal distribution](#) with mean  $\mu = 1$  and standard deviation  $\sigma = 3$

`rt(n, df, ncp)`

`rt(5,3)` generates 5 random numbers from the [t distribution](#) with  $\nu = 3$  degrees of freedom

`rf(n, df1, df2, ncp)`

`rf(5,4,3)` generates 5 random numbers from the [F distribution](#) with degrees of freedom  $\nu_1 = 5$  and  $\nu_2 = 3$

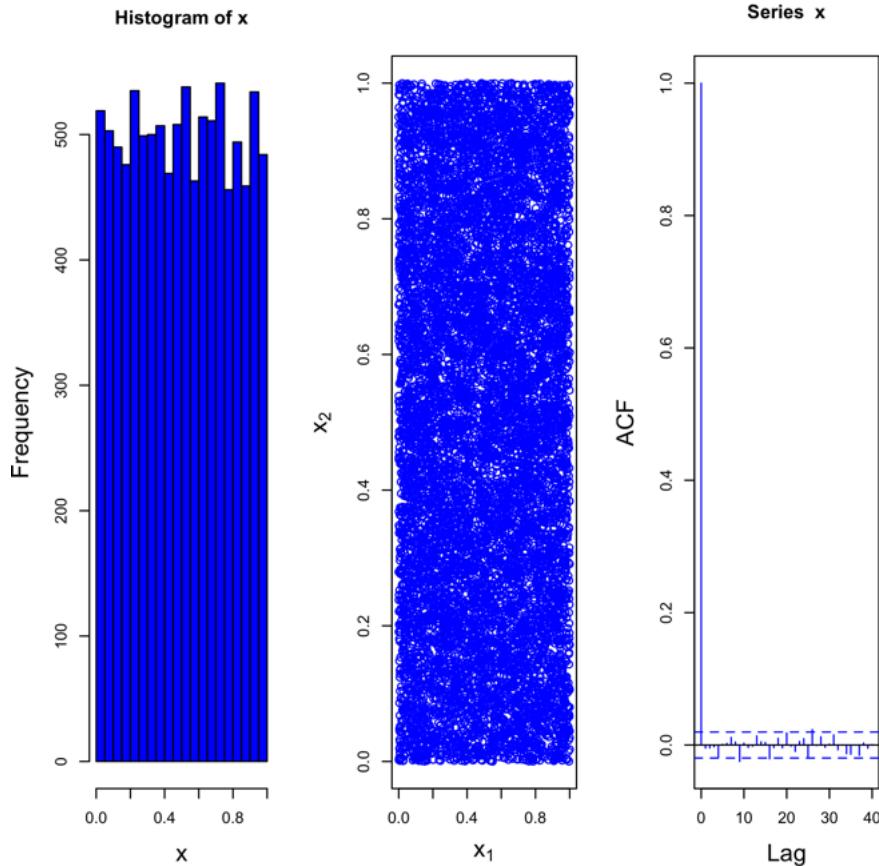


Figure 5.1: The  $10^4$  pseudo-random numbers generated with **R** routine `runif()` do indeed look random: (1) flat histogram, (2) data pairs fill the plane, (3) vanishing auto-correlation (from Robert & Casella (2009, p. 43, Fig. 2.1 modified: random seed used & blue color added). [RandomUnifLooksRandom.R](#)

**Exercise 10 Random numbers: rolling a die**

*Generate 10000 random numbers for the rolling of a die and show the results in the form of a histogram.*

**Exercise 11 Random numbers: multiplication by  $c > 0$  or addition of  $q$** 

- (1) How does multiplication of random numbers from a chosen population by a constant factor  $c > 0$  change their sample mean, variance, and standard deviation?
- (2) How does addition of random numbers from a chosen population by a constant  $q$  change their sample mean, variance, and standard deviation?

# Chapter 6

## PDs & PDFs

*In the current chapter various (discrete) probability distributions (PDs) and (continuous) probability density functions (PDFs) will be presented. The goal is to give an overview on the most important and commonly used PDs (uniform, Poisson, binomial, ...) and PDFs (normal, uniform, t,  $\chi^2$ , ...) with plots, mathematical formulas and characteristic properties (mean, median, variance).*

**Further reading:** Leemis, Relationships among common univariate distributions (1986)

## 6.1 Statistical populations: discrete versus continuous

When rolling a die, the outcome can be 1, 2, 3, 4, 5, or 6. The set of numbers  $\{1, 2, 3, 4, 5, 6\}$  is called the **sample space**<sup>1</sup>. For the die, the sample space is discrete (in contrast to continuous) and one can assign to each element  $x_k$  of the sample space a probability  $0 \leq p(x_k) \equiv p_k \leq 1$ . If the elements of the sample space are mutually exclusive and exhaustive (which is the case here for the sample space of the die and also for other systems considered below) the sum of the probabilities  $p_k$  has to add up to 1 (= it is certain that one of the outcomes will be realized). The set of probabilities corresponding to the sample space is called a **probability distribution**; for the die it reads  $\{p_1, p_2, p_3, p_4, p_5, p_6\}$ . Probability distributions are discrete sets that fulfill two properties<sup>2</sup>

$$0 \leq p_k \leq 1 \quad \text{for all } k \text{ (a general property of probabilities)} \quad (6.1)$$

$$\sum_k p_k = 1 \quad \text{normalization condition} \quad (6.2)$$

A **discrete statistical population** (population for short) is described completely by the sample space and the corresponding probability distribution.

Temperature or salinity are continuous quantities and thus their sample space is continuous<sup>3</sup>. Single outcomes on a continuous scale can not be described by probabilities ('a single point would have probability zero'). Instead of probabilities one uses **probability density functions (PDFs)**  $f(x)$  with the following properties

$$f(x) \geq 0 \quad \text{probability density functions are non-negative} \quad (6.3)$$

$$\int f(x) dx = 1 \quad \text{normalization condition.} \quad (6.4)$$

An example is the standard normal distribution with probability density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (6.5)$$

which is obviously  $\geq 0$  (the constraint in Eq. 6.3); constraint Eq. 6.4 is 'responsible' for the factor  $1/\sqrt{2\pi}$  ('normalization constant') in front of the exponential function.

Probabilities can be calculated from probability densities by integration:

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x) dx \quad (6.6)$$

is the probability to find  $x$  in the interval  $[x_1, x_2]$ ; this is the area under the density curve (between the curve and the  $x$ -axis). Because of the symmetry of the standard normal distribution with respect to  $x = 0$ , half of the probability is located in the interval from minus infinity to zero ( $[-\infty, 0]$ , Fig. 6.1) or in terms of integration

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{x^2}{2}} dx = \frac{1}{2} \quad (6.7)$$

**Cumulative probability distribution functions (CDFs):** We have already seen that probabilities can be calculated from probability density functions (PDFs) by integration. If one starts integration at the lower limit of

<sup>1</sup>A sample consists of elements from the sample space and thus, for the die, can not be 7 or 3.1.

<sup>2</sup>Vice versa, any discrete set obeying the two constraints given above could (depending on the context) be interpreted as a probability distribution.

<sup>3</sup>Temperature and salinity are properties of macroscopic systems and thus we can neglect here the underlying discrete nature (molecules!). We will also neglect the discretization caused by measuring devices and the finite number of digits in reading and storing data.

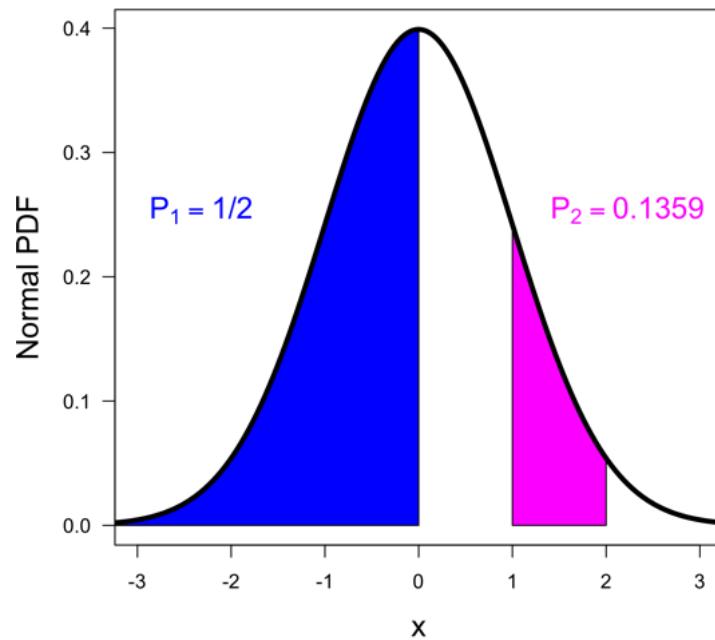


Figure 6.1: The standard normal probability density function (normal PDF, black solid line): obviously ('integration by eye'), half of the probability is located between  $x = -\infty$  and  $x = 0$  (blue area,  $P_1 = 1/2$ ). The probability between  $x = 1$  and  $x = 2$  is calculated by integration over the normal PDF:  $P_2 = \int_1^2 \text{Normal}(x) dx \approx 0.1359$ . [NormalPDFprobabilities.R](#)

the sample space ( $x_L = -\infty$  for the normal distribution), one obtains the cumulative probability distribution function (CDF)

$$\text{CDF}_f(x) = \int_{x_L}^x f(y; \dots) dy \quad (6.8)$$

which is a function of the upper integration boundary  $x$ . The CDF at the lower limit of the sample space is zero ( $\text{CDF}_f(x_L) = 0$ , no probability below – 'left of' –  $x_L$ ) and equal to one at the upper limit of the sample space ( $\text{CDF}_f(x_U) = 1$ , no probability above – 'right of' –  $x_U$ ) because of the normalization condition (for the CDF of the standard normal PDF compare Fig. 6.8).

**Quantile function:** The inverse of the cumulative probability density function is called the quantile function. It is denoted by  $\text{CDF}_f^{-1}(p) \equiv Q_f(p)$ . The quantile function for the probability  $p$  gives the position  $x$  at which the cumulative probability density function is equal to  $p$  or, in other words, the probability below – 'left of' –  $x$  is equal to  $p$  (compare Subsection 6.4.1 for an example).

## 6.2 Expected value, expectation: mean & variance of PDs & PDFs

The expected value or expectation  $E[X]$  of a discrete stochastic variable  $X$  is defined as the weighted sum  $\sum_i X_i p(X_i; \dots)$  where the probabilities of the discrete distribution  $p(X_i; \dots)$  are the weights and the sum runs over all possible values  $X_i$ . Analogously, the expectation  $E[X]$  of a continuous stochastic variable  $X$  is defined as the integral  $\int X q(X; \dots) dX$  where the probability density  $q(X; \dots)$  times  $dX$  are the weights and the integral runs over all possible  $X$  values. The expectation of  $X$  is also called the 'mean'<sup>4</sup>. Expectations  $E[\cdot]$  can also be calculated for functions of  $X$ ,  $f(X)$ :

$$E[f(X)] = \sum_i f(X_i) p(X_i; \dots) \quad (6.9)$$

$$E[f(X)] = \int f(X) q(X; \dots) dX. \quad (6.10)$$

The expectations of  $f(X_i) = (X_i - \mu)^2$  and  $f(X) = (X - \mu)^2$ , respectively, are the variances of  $X_i$  and  $X$ , respectively.

Casella & Berger (2002) further require that  $E[|f(X)|]$  is  $< \infty$  ('exists' in mathematical parlance). As a consequence of this additional condition, the Cauchy PDF possesses no mean value (the variance of the Cauchy PDF does not exist already without this additional condition). Please note that various other textbook authors (for example, Barlow, 1989, von Storch & Zwiers, 2001) do not mention this additional condition.

**Music:** The Rolling Stones: No expectation (from the album Beggars Banquet, 1968)

### 6.2.1 Examples for mean values & variances (\*)

(1) For the standard normal distribution one obtains

$$E[x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{x^2}{2}} dx = 0 \quad (6.11)$$

because the integrand  $r(x)$  is an odd function, i.e.  $r(-x) = -r(x)$ . I.e. the mean  $\mu = E[x] = 0$  (no surprise or 'as expected').

(2) The mean of  $x$  over the  $F$  distribution  $\mathcal{F}(x; \nu_1 = 15; \nu_2 = 5)$  is defined by

$$E[x] = \int_0^{+\infty} x \mathcal{F}(x; \nu_1 = 15; \nu_2 = 5) dx. \quad (6.12)$$

Numerical evaluation of the integral yields  $\mu = E[x] \approx 5/3$  which is consistent with the analytical expression  $\mu = \mu(\nu_2) = \frac{\nu_2}{\nu_2 - 2}$ .

(3) The mean of the Poisson distribution is defined by

$$E[k] = \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad (6.13)$$

The Taylor expansion of  $e^\lambda$  about  $\lambda = 0$  reads

$$e^\lambda = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad (6.14)$$

and thus

$$E[x] = \lambda e^{-\lambda} e^\lambda = \lambda \quad (6.15)$$

---

<sup>4</sup>The expectation of  $X$  is the true mean of  $X$ , whereas the sample mean  $\bar{X}$  provides only an estimate of the true mean.

**Examples for variances**  $\sigma^2 = E[(X - \mu)^2]$ :

The **variance**  $\sigma^2$  of a continuous variable is defined by the expectation

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 q(x; \dots) dx. \quad (6.16)$$

(1) For the standard normal distribution one obtains

$$E[(x - \mu)^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x - \mu)^2 e^{-\frac{x^2}{2}} dx \quad (6.17)$$

$$= \frac{1}{\sqrt{2\pi}} \left[ \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx - 2\mu \underbrace{\int_{-\infty}^{+\infty} x e^{-\frac{x^2}{2}} dx}_{=0} + \underbrace{\mu^2}_{=0} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right] = 1 \quad (6.18)$$

(2) The variance of the  $F$  distribution  $\mathcal{F}(x; \nu_1 = 15; \nu_2 = 5)$  is defined by

$$E[(x - 5/3)^2] = \int_0^{+\infty} (x - 5/3)^2 \mathcal{F}(x; \nu_1 = 15; \nu_2 = 5) dx. \quad (6.19)$$

Numerical evaluation of the integral yields  $6.667 \approx 20/3$  which is identical to the analytical expression

$$\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \text{for } \nu_2 > 4. \quad (6.20)$$

(3) The variance of the standard uniform distribution is defined by

$$E[(x - 1/2)^2] = \int_0^1 (x - 1/2)^2 dx = \int_0^1 (x^2 - x + 1/4) dx \quad (6.21)$$

$$= \left. \frac{x^3}{3} - \frac{x^2}{2} + \frac{x}{4} \right|_0^1 = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{4 - 6 + 3}{12} = \frac{1}{12} \quad (6.22)$$

### Exercise 12 Variance of the Poisson distribution (\*)

Show that

$$E[(k - \lambda)^2] = \sum_{k=0}^{\infty} (k - \lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \quad (6.23)$$

### Exercise 13 Expectation of $x$ with respect to the exponential PDF (\*)

Calculate the expectation  $E[x]$  with respect to the exponential PDF

$$f(x; \lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x < \infty, \quad \lambda > 0. \quad (6.24)$$

### Exercise 14 Expectation $E[\sum_i g(X_i)]$ (\*)

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population and let  $g(x)$  be a function such that  $E[g(X_1)]$  exists. Show that

$$E \left[ \left( \sum_{i=1}^n g(X_i) \right) \right] = n E[(g(X_1))]. \quad (6.25)$$

## 6.2.2 Expectation of the product of two stochastic variables

The expectation of the product of two stochastic variables  $X$  and  $Y$  is given by

$$E(XY) = E(X) E(Y) + \text{Cov}(XY) \quad (6.26)$$

and thus for  $X = Y$

$$E(X^2) = (E(X))^2 + \text{Var}(X) = \mu^2 + \sigma^2$$

In words: The expectation of the random variable  $X^2$  is given by the sum of the squared mean,  $\mu^2$ , and the variance,  $\sigma^2$ . For the sample mean,  $\bar{X}$ , the mean is  $\mu$ , however, the variance is smaller than  $\sigma^2$ , namely,  $\frac{\sigma^2}{n}$ . Thus the expectation of the sample mean squared is  $\mu^2 + \frac{\sigma^2}{n}$ .

### 6.2.3 How are expected values and sample mean & variance related to each other?

On first sight, the formulas for the sample mean and the expectation of  $x$  look quite different from each other:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \quad (6.27)$$

$$E[x] = \sum_i x_i p(x_i; \dots) \quad (6.28)$$

Whereas the sample mean is an arithmetic mean, the expectation is a weighted mean with weights given by the probability distribution (for example, the Poisson distribution). Another interpretation of the sample mean is a weighted sum where all weights are equal to  $1/n$ . So, we have here two weighted means, however, with different weights. However, the summands are different as well: whereas the  $x_k$  in the sample mean are random numbers from the statistical population  $\mathcal{P}$ , the  $x_i$  in the expectation formula run through all conceivable outcomes from the same population. How can these two quantities give similar results? Or in other words: How can the sample mean  $\bar{x}$  be an (unbiased) estimator of the mean  $\mu = E[x]$ ? The reason is simple: values  $x_k$  with a high probability  $p(x_k)$  occur more often than values  $x_m$  with a low probability  $p(x_m)$  and thus they contribute more (more often, with a higher probability) to the sample mean. In other words, the varying weights  $p(x_i)$  in the expectation correspond to varying frequencies (= number of occurrences) of certain values  $x_i$  in the sample mean. This kind of reasoning can be extended to PDFs. The same applies to the variance. For more details compare Chapter 10 on point estimators.

### 6.2.4 Median of distributions (\*)

"A median of a distribution is a value  $m$  such that  $P(X \leq m) \geq 1/2$  and  $P(X \geq m) \geq 1/2$ ." (Casella & Berger, 2002, p. 78) In the definition  $P(X)$  is the cumulative distribution function.

**Examples:**

1. Normal pdfs: median = mean, i.e.  $m = \mu$ .
2. The median of the Cauchy PDF (Eq. C.33) is equal to the parameter  $\theta$ .

### 6.2.5 Mode

The mode of the probability distribution  $p(x)$  is defined as the  $x$ -value with the highest probability (for example, upper left panel of Fig. 6.2). The mode of the probability density function  $f(x)$  is defined as the  $x$ -value with the highest probability density (lower left panel of Fig. 6.2). Some distributions posses several (local) maxima of probability or probability density (right panels of Fig. 6.2) and thus one speaks of bimodal (in case of two modes) or multimodal (more than two modes) distributions; distributions with a single mode are called unimodal. For symmetric unimodal distributions the mean, median, and mode are identical. However, this is not the case for asymmetric distributions as, for example, for the F-distribution with  $v_1 = 15$  and  $v_2 = 3$  degrees of freedom (Fig. 6.2 lower left panel) where the mode =  $\frac{v_1 - 2}{v_1} \frac{v_2}{v_2 + 2} = 0.52$  is small than the median  $\approx 1.21$  and the mean =  $\frac{v_2}{v_2 - 2} = 3$ . For samples from discrete population the mode can be defined analogously

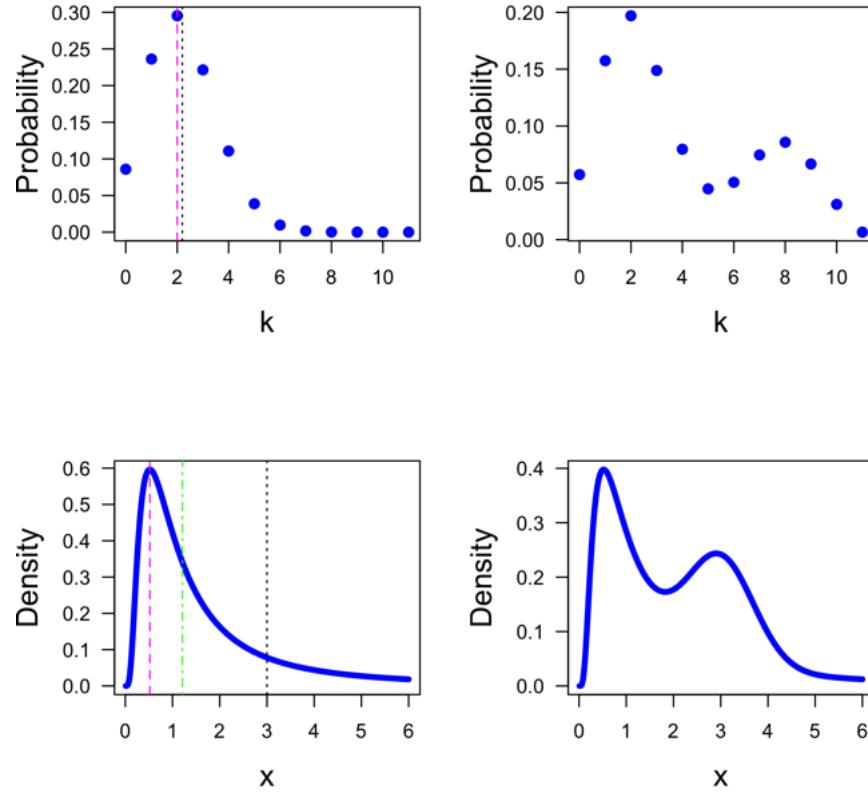


Figure 6.2: Mode of a distribution. Upper left panel: the mode (and median) of the binomial distribution with probability of  $p = 0.2$  for success in a single trial and  $n = 11$  trials (blue dots) is 2 (indicated by magenta vertical broken line) whereas its mean is 2.2 (indicated by black vertical dotted line). Upper right panel: a bimodal discrete distribution. Lower left panel: the mode of the F-distribution with  $\nu_1 = 15$  and  $\nu_2 = 3$  degrees of freedom (blue solid line) is 0.52 (indicated by magenta vertical broken line), its median is  $\approx 1.21$  (indicated by green vertical dash-dotted line) and its mean is 3 (indicated by black vertical dotted line). Lower right panel: a bimodal probability density function.

R code: [ModeOfDistribution.R](#)

as the  $x$  value observed with highest frequency. For samples from continuous populations such a definition can not be applied because usually all observations are (slightly) different from each other and thus occur only once. Therefore the observed values first have to be binned in an appropriate way (histogram, boundaries of bins have to be chosen) leading to discrete values (number of observations in bins) that can be treated like data from a discrete population. Alternatively, one can estimate the density of  $f(x)$  and use the maximum of this estimate as an estimator of the mode.

## 6.3 Probability distributions: examples

Here we will present a few examples of probability distributions (PDs) – uniform, binomial, Poisson – with formulas, graphs, expressions for mean and variance. Various other PDs can be found in the appendix.

### 6.3.1 Discrete uniform distributions

$$\text{Uniform}(k; N) = \frac{1}{N} \quad \text{for } k = 1, 2, \dots, N \quad (6.29)$$

where  $k$  is the actual value out of  $N$  possible values. **Some properties of discrete uniform distributions:**

**Mean:**  $\mu = \frac{N+1}{2}$

**Variance:**  $\sigma^2 = \frac{(N-1)(N+1)}{12} = \frac{N^2-1}{12}$

Discrete uniform distributions between  $A$  and  $B$

$$\text{Uniform}_{AB}(k) = \frac{1}{N} \quad \text{for } k = A, A+1, A+2, \dots, B \quad (6.30)$$

with  $N = B - A + 1$

**Mean:**  $\mu = \frac{A+B}{2}$

**Variance:**  $\sigma^2 = \frac{(B-A+1)^2-1}{12} = \frac{N^2-1}{12}$

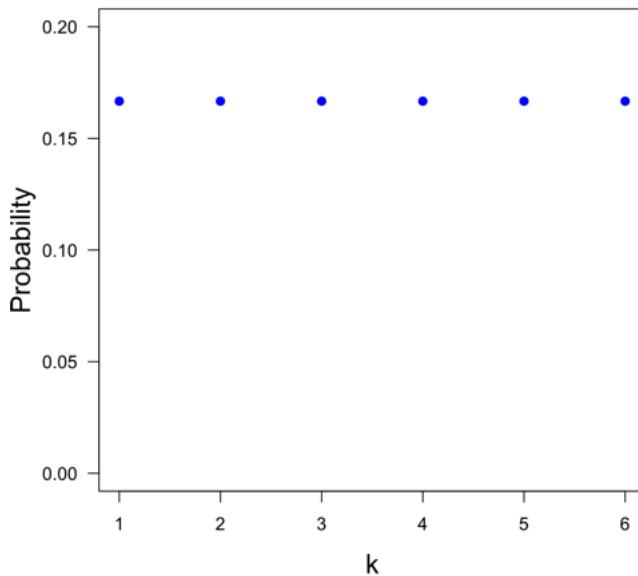


Figure 6.3: Discrete uniform distribution for the sample space  $\{1, 2, 3, 4, 5, 6\}$  (fair die).  
[PDsPDFsUniformDiscrete6.R](#)

### 6.3.2 Binomial distribution

A random variable  $m$  has a *Bernoulli probability distribution* if

$$m = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1 - p) \end{cases} \quad (6.31)$$

where  $0 \leq p \leq 1$  is the probability of success in a single trial<sup>5</sup>. Instead of  $m = 1$  and  $m = 0$  one could use as well 'head' and 'tail' or 'black' and 'white' or  $\dots$ . If repeating the flipping of a coin  $n$  times, one can ask 'what is the probability for  $k$  times head?'<sup>6</sup>. The answer is given by the *binomial probability distribution* (Fig. 6.4)

$$\text{Binom}(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k} \quad (6.32)$$

where  $k$  runs in steps of 1 from zero to  $n$ . The binomial distribution has two parameters, namely  $n$  the number of flips (trials) and the probability  $p$  for success in a single trial.<sup>7</sup> A nice application of the binomial distribution can be found in Husmann & Klaas (2022) who estimate the growth rate of diatoms by staining and counting.

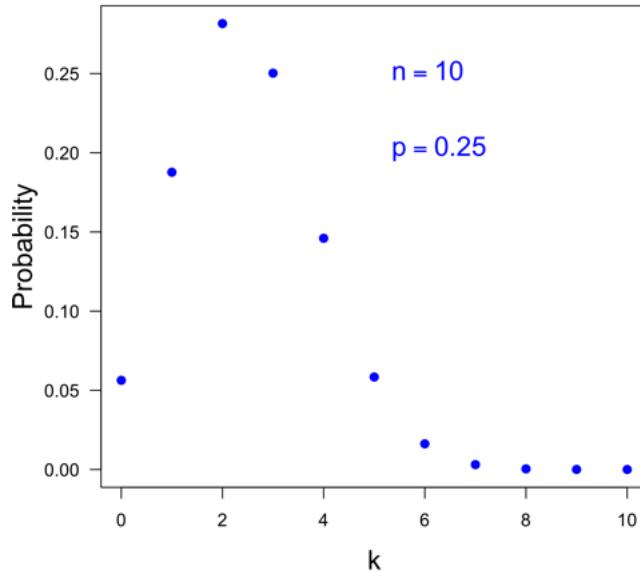


Figure 6.4: Binomial probability distribution  $\text{Binom}(k; n = 10, p = 0.25)$  [PDsPDFsBinomPD10n.R](#)

#### Some properties of the binomial distribution:

**Mean:**  $\mu = n \cdot p$

**Variance:**  $\sigma^2 = n \cdot p (1 - p)$

The integral of the binomial PD over  $dp$  has a simple solution (that is independent of  $k$ ):

$$\int_0^1 \text{Binom}(k; n, p) dp = \int_0^1 \binom{n}{k} p^k (1 - p)^{n-k} dp = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k} dp = \frac{1}{n+1} \quad (6.33)$$

<sup>5</sup>The outcome 'm=1' is called 'success'

<sup>6</sup>Sampling with replacement

<sup>7</sup>Parameters of the distributions are shown in the argument list after the semi-colon:  $\text{Binom}(k; n, p)$ .

Useful identity (including normalization):

$$1 = (p + (1 - p))^n = \sum_{k=0}^n \text{Binom}(k; n, p) \quad (6.34)$$

### 3: Factorial & gamma function

The **factorial** of  $n$  ( $n = 0, 1, 2, \dots$  is a non-negative integer) is defined by

$$n! = 1 \cdot 2 \cdot \dots \cdot n \quad (6.35)$$

and thus  $1! = 1, 2! = 2, 3! = 6, 4! = 24, 5! = 120$ . In some formulas (for example, the binomial distribution Eq. 6.32) the factorial of zero can occur. Its value cannot be derived from Eq. 6.35 and thus requires a special definition: the factorial of 0 is equal to 1, i.e.  $0! = 1$ .

The **gamma function**  $\Gamma(z)$  can be seen as a generalization of the factorial to a larger set of arguments (actually to complex values). The gamma function can be defined by the following integral

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad (6.36)$$

where  $z$  is a complex number and  $x$  is real. Special value:  $\Gamma(1/2) = \sqrt{\pi} \approx 1.772454$

The connection between the gamma function and the factorial is given by

$$\Gamma(n+1) = n! \quad (6.37)$$

where  $n$  is a nonnegative integer.

### Exercise 15 Success or failure?

Suppose you are rolling a fair die two times and count a '6' as success (S) and all other outcomes as failure (F). What are the probabilities for the cases SS, SF, FS, and FF? [Hint: you can apply the simplified product rule (Eq. 4.9). Why?] Show that the probabilities for SS, (SF or FS), and FF obey the binomial distribution  $\text{Binom}(k; n, p)$  for  $k = 0, 1, 2$ ,  $n = 2$ ,  $p = 1/6$ .

### 6.3.3 Poisson distribution

The Poisson distribution<sup>8</sup>

$$\text{Poisson}(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k = 0, 1, 2, \dots \quad (6.38)$$

is a discrete probability distribution that expresses the probability for a given number of events  $k$  occurring in a fixed interval of time and/or space if these events occur **randomly with the only constraint of an average rate  $\lambda$** . The Poisson distribution possesses the interesting property that its variance,  $\sigma^2$ , has the same value as its mean,  $\mu$ :  $\sigma^2 = \lambda = \mu$ . This is only possible because the mean rate  $\lambda$  is considered as a dimensionless quantity<sup>9</sup>. The fact that the variance and the mean of the Poisson distribution are equal makes this PD a bit more 'rigid' (less flexible) compared to the binomial distributions which possess two adjustable parameters. This can lead to the problem of **over- or underdispersion** where a data set shows much larger or lower variability (variance, statistical dispersion) than expected from a Poisson distribution that fits the sample mean.

The Poisson distribution can be derived from the binomial distribution in the limit of a small probability of success,  $p$ , and a large number of trials,  $n$ , where  $p \cdot n = \lambda$  is kept constant (compare Section C.2.1 for mathematical details).

**Some properties of the Poisson distribution:** **Mean:**  $\mu = \lambda$ ; **Variance:**  $\sigma^2 = \lambda$

For applications of the Poisson distribution compare Sections 1.2, 1.4, and 21.

---

<sup>8</sup>Named after the French mathematician Siméon Denis Poisson (1781-1840).

<sup>9</sup>I.e.  $\lambda = 2.5$  means 2.5 events on average for the interval considered and *not* something like 2.5 events per hour (even if the length of the interval would be one hour).

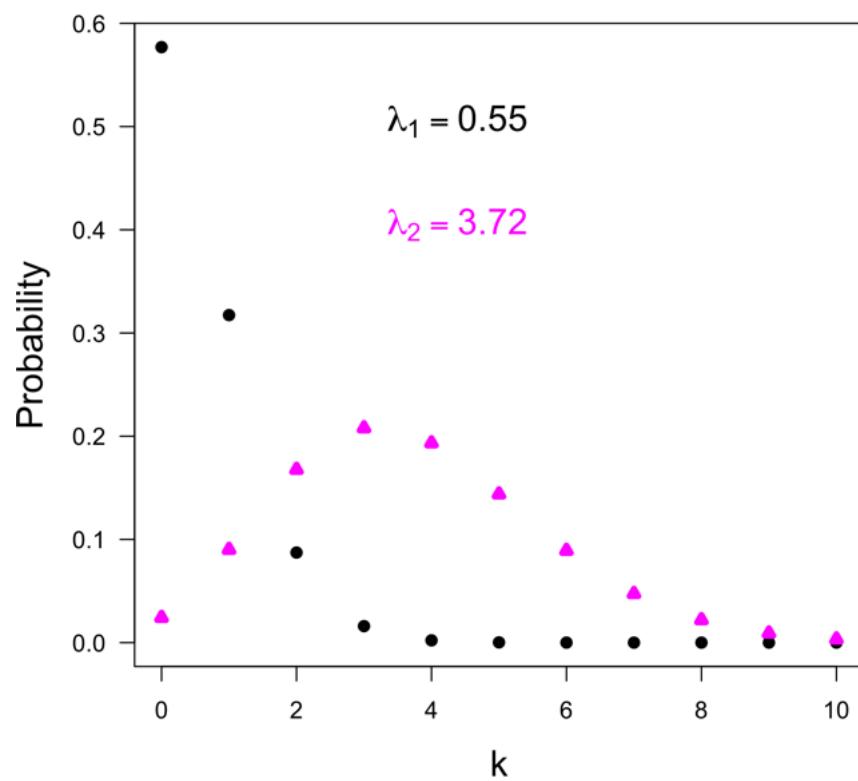


Figure 6.5: Poisson distributions for two different mean rates. [PDsPDFsPoisson2Examples.R](#)

**Exercise 16 Comparison of binomial & Poisson distribution**

Calculate the mean value  $\mu$  for the binomial distribution  $\text{Binom}(k; n = 10, p = 0.27)$  and use this value as the mean rate  $\lambda$  for a Poisson distribution. Plot both distributions for  $k = 0, 1, \dots, 10$  and compare the results.

**Exercise 17 Corona vaccine**

Currently (July, 2020), more than 20 Covid-19 vaccine candidates are under investigation in clinical studies. If one assumes that there is a 10% chance that a candidate is effective, what is the chance to find 1, 2, or more effective vaccines out of 20 candidates? What is the chance that all candidates fail?

## 6.4 Probability Density Functions (PDFs): examples

In this section we present a few often applied PDFs: normal, uniform, Student-t, and F. More PDFs can be found in the appendix.

### 6.4.1 The normal distribution

The normal<sup>10</sup> distribution is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \text{Normal}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6.39)$$

where  $\mu$  is the mean value,  $\sigma$  the standard deviation, and  $\sigma^2$  the variance.<sup>11</sup> For  $\mu = 0$  and  $\sigma = 1$ , Eq. 6.39 simplifies to (Fig. 6.6)

$$\mathcal{N}(x; \mu = 0, \sigma^2 = 1) \equiv \mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (6.40)$$

which is called **standard normal distribution**.

The maximum of normal distributions is located as  $x = \mu$  and amounts to  $\frac{1}{\sqrt{2\pi\sigma^2}}$ .

---

<sup>10</sup>The normal distribution is also called Gaussian distribution.

<sup>11</sup>Note that some authors use the standard deviation  $\sigma$  as argument, i.e.  $\mathcal{N}(x; \mu, \sigma)$ , whereas others use the variance  $\sigma^2$ , i.e.  $\mathcal{N}(x; \mu, \sigma^2)$ . In order to avoid confusion by statements like  $\mathcal{N}(x; 1.3, 4)$  where it's not clear whether  $\sigma = 4$  or  $\sigma^2 = 4$  we will use instead the notations  $\mathcal{N}(x; 1.3, 2^2)$  or  $\mathcal{N}(x; \mu = 1.3, \sigma^2 = 4)$ . For  $\sigma = 1$  no confusion is expected when writing  $\mathcal{N}(x; 1.3, 1)$ .

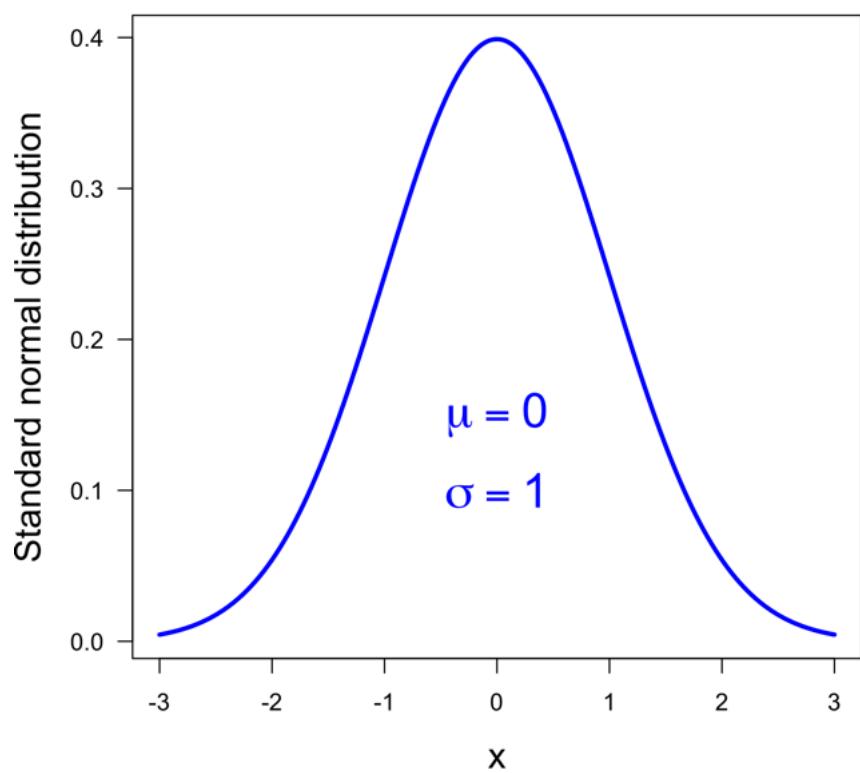


Figure 6.6: Standard normal distribution (Eq. 6.40) [PDsPDFsNormalStandard.R](#)

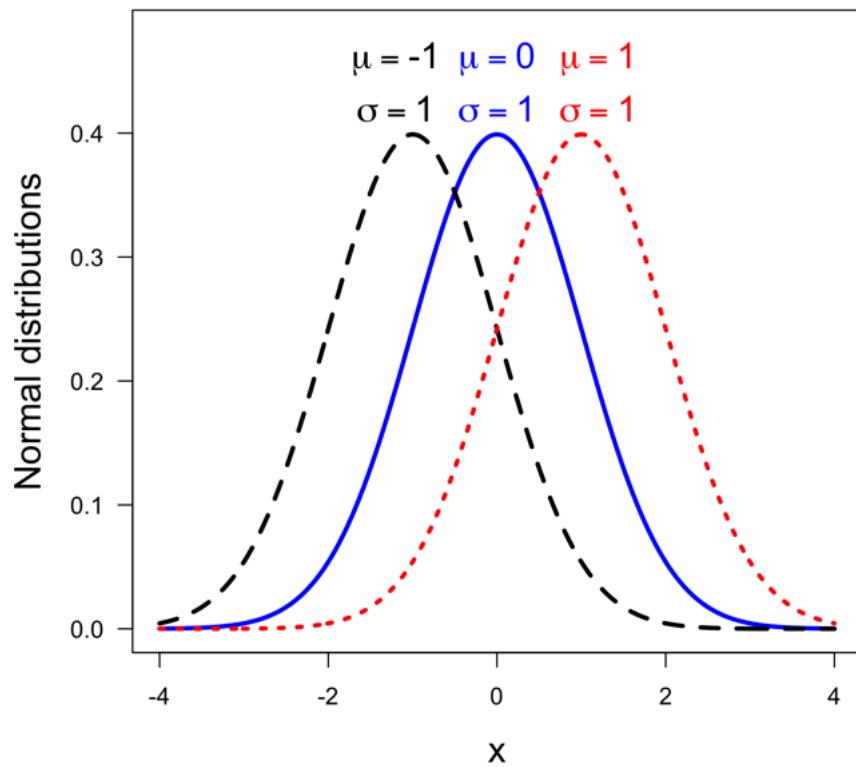


Figure 6.7: Normal distributions with different standard mean values (Eq. 6.39;  $\sigma = 1$ ): The blue solid curve shows the standard normal distribution. For  $\mu < 0$  (here:  $\mu = -1$ , black dashed line) the distribution is shifted to the left. For  $\mu > 0$  (here:  $\mu = +1$ , red dash-dotted line) the distribution is shifted to the right. [PDsPDFsNormalDifferentMeans.R](#)

The cumulative distribution function (CDF) of the normal distribution (Fig. 6.8) is given by

$$\mathcal{N}_{\text{CDF}}(x; \mu, \sigma^2) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2\sigma^2}}\right) \quad (6.41)$$

where  $\operatorname{erf}(x)$  is the error function. The CDF of the standard normal PDF

$$\mathcal{N}_{\text{CDF}}(x) = \Phi(x) == \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \quad (6.42)$$

is often denoted as  $\Phi(x)$ .

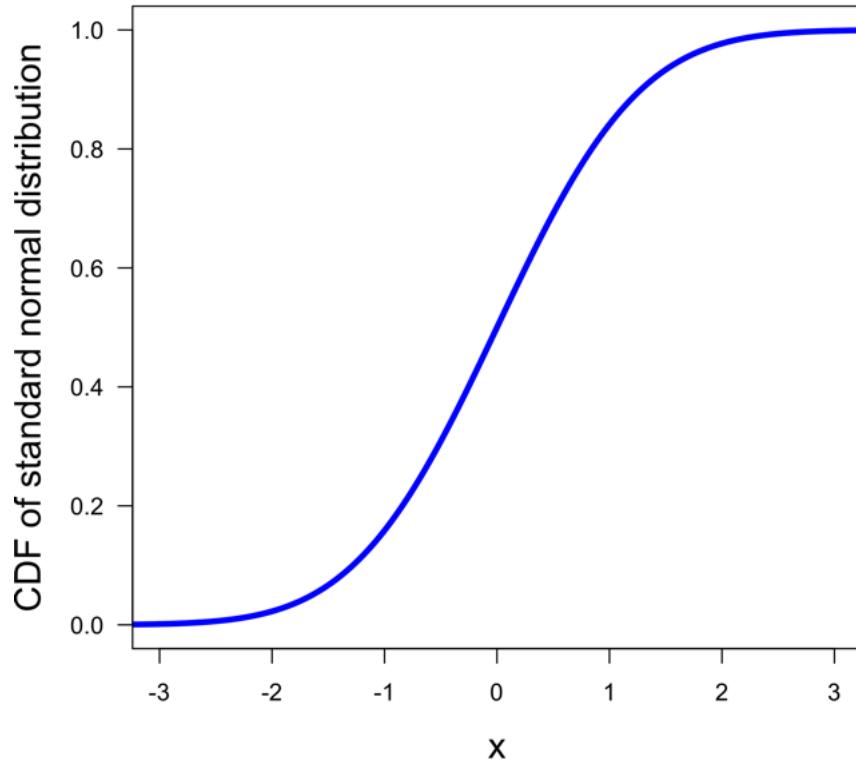


Figure 6.8: CDF of the standard normal distribution (Eq. 6.41 for  $\mu = 0, \sigma^2 = 1$ ). PDsPDFsNormalCDF.R

The quantile function of the normal distribution is given by

$$\text{CDF}_{\mathcal{N}}^{-1}(p) \equiv Q_{\mathcal{N}}(p) = \mu + \sqrt{2\sigma^2} \operatorname{erf}^{-1}(2p - 1), \quad p \in (0, 1). \quad (6.43)$$

For the standard normal distribution Eq. (6.43) simplifies to

$$\text{CDF}_{\mathcal{N},st}^{-1}(p) \equiv Q_{\mathcal{N},st}(p) \equiv \Phi^{-1}(p) = \sqrt{2} \operatorname{erf}^{-1}(2p - 1), \quad p \in (0, 1) \quad (6.44)$$

which is called the [probit function](#)<sup>12</sup> (Fig. 6.9).

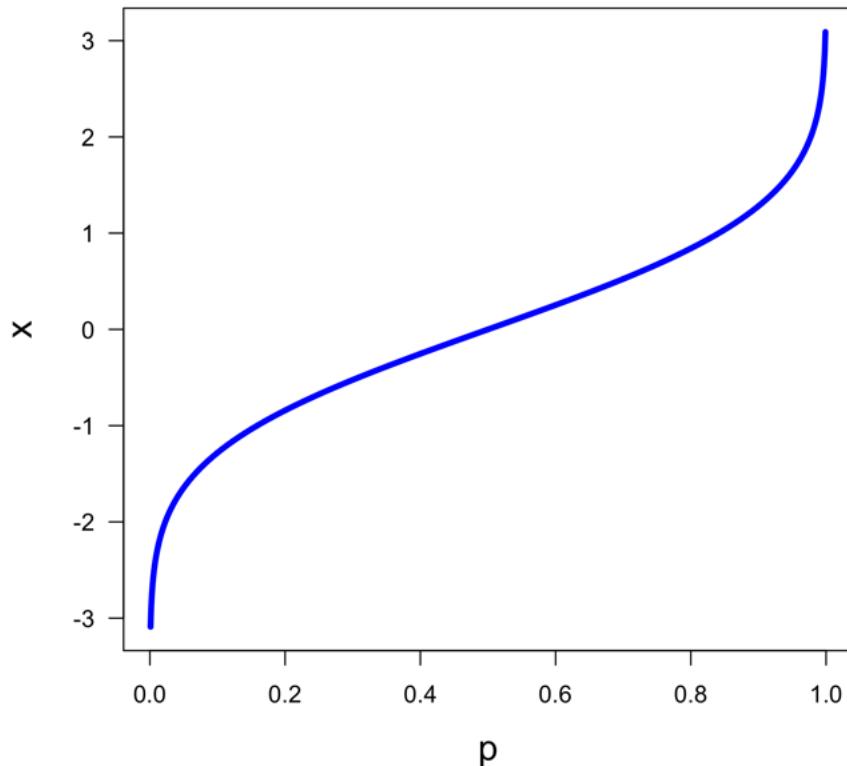


Figure 6.9: The quantile function of the standard normal distribution also called probit function (Eq. 6.44):  $x \rightarrow -\infty$  for  $p \rightarrow 0$ ,  $x = 0$  for  $p = 1/2$ , and  $x \rightarrow \infty$  for  $p \rightarrow 1$  (the plot is restricted to the interval  $0.001 \leq p \leq 0.999$ ).  
[PDsPDFsNormalQuantileFct.R](#)

---

<sup>12</sup>'Probit' is an abbreviation of 'probability unit' (Bliss, 1934).

**4: Normal distribution: density, CDF, quantile, random in R**

R provides 4 routines related to the normal PDF:

**dnorm(x,mean=0, sd=1)** is the standard normal density function that can be called just by **dnorm(x)** because mean = 0 and standard deviation (sd) = 1 are the default values. The call **dnorm(x,mean=1.3, sd=2)** would be the normal density function with mean  $\mu = 1.3$  and standard deviation  $\sigma = 2$ .

**pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)** is the cumulative distribution function for the standard normal distribution (Fig. 6.8). It can be called just by **pnorm(q)** whereby  $q$  can vary between  $-\infty$  and  $\infty$ : **pnorm(q) = 0** for  $q = -\infty$ , **pnorm(q) = 1/2** for  $q = 0$ , and **pnorm(q) = 1** for  $q = \infty$ .

**qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)** is the quantile function for the standard normal distribution (Fig. 6.9). It can be called just by **qnorm(p)** whereby  $p$  can vary between 0 and 1: **qnorm(0) =  $-\infty$** , **qnorm(1/2) = 0**, **qnorm(1) =  $\infty$** .

**rnorm(n, mean = 0, sd = 1)** is a routine for generating  $n$  random numbers from the standard normal distribution. It can be called just by **rnorm(n)**.

**Remark:** R provides PDs and PDFs ('densities', names always start with 'd'), CDFs (names always start with 'p'), quantile functions (names always start with 'q'), and random number generators (names always start with 'r') for various other PDFs or PDs. Example:  $t$ -distribution with density **dt()**, CDF **pt()**, quantile function **qt()**, and random number generator **rt()**.

**Exercise 18 Properties of the normal PDF**

Generate a graph that displays the standard normal PDF with some of its properties:  $\mu$ ,  $\sigma$ , maximum value, areas for ranges  $\pm n \sigma$  with  $n = 1, 2, 3, 66\%, 95\%$  and  $99\%$  level.

Normal distributions possess various **amazing properties**:

- The difference  $D = X - Y$  between two random numbers  $X$  and  $Y$  from the standard normal distribution  $\mathcal{N}(X; \mu = 0, \sigma^2 = 1)$  follows the normal distribution

$$\mathcal{N}(D; \mu = 0, \sigma^2 = 2) \quad (6.45)$$

(i.e., normal distribution with a larger variance).

- The difference  $D = X - Y$  between two random numbers  $X$  and  $Y$  from normal distributions  $\mathcal{N}(X; \mu_X, \sigma_X^2)$  and  $\mathcal{N}(Y; \mu_Y, \sigma_Y^2)$  follows the normal distribution

$$\mathcal{N}(D; \mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2). \quad (6.46)$$

- The linear combination  $Z = c_1 X + c_2 Y$  ( $c_1, c_2$  real numbers) of two random numbers  $X$  and  $Y$  from normal distributions  $\mathcal{N}(X; \mu_X, \sigma_X^2)$  and  $\mathcal{N}(Y; \mu_Y, \sigma_Y^2)$  follows the normal distribution

$$\mathcal{N}(Z; c_1 \mu_X + c_2 \mu_Y, c_1^2 \sigma_X^2 + c_2^2 \sigma_Y^2). \quad (6.47)$$

Eqs. 6.45 and 6.46 are special cases of Eq. 6.47.

- The product of two normal distributions with means  $\mu_1$  and  $\mu_2$ , respectively, and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, is proportional (not normalized!) to a normal distribution with mean

$$\mu_3 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\mu_1 h_1 + \mu_2 h_2}{h_1 + h_2} \quad (6.48)$$

and variance

$$\sigma_3^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} \quad (6.49)$$

(the inverse of a variance,  $h_k = 1/\sigma_k^2$ , is called 'precision').

### Exercise 19 Product of two normal distributions (\*)

(1) Show that the product of two normal distributions is proportional to a normal distribution, i.e.

$$\text{normal}(x; \mu_1, \sigma_1) \text{ normal}(x; \mu_1, \sigma_1) \propto \text{normal}(x; \mu_3, \sigma_3) \quad (6.50)$$

with

$$\mu_3 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \quad \text{and} \quad \sigma_3^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} \quad (6.51)$$

(2) Discuss the special cases:

- $\sigma_1 = \sigma = \sigma_2$
- $\mu_1 = \mu = \mu_2$
- product of two identical normal distributions.

Hint: for the derivation of Eq. 6.51 introduce the very useful abbreviation  $h_k = 1/\sigma_k^2$  (called 'precision' parameter, not the best choice of notation).

**Why do we encounter normal distributions so often?** There could be several reasons:

1. Sums of random numbers (noise) leads to values that are normally distributed. This is an essential result of the Central Limit Theorem (CLT, Chapter 7).

2. '... any function with a single rounded maximum raised to a higher and higher power, goes into a Gaussian function' (Jaynes, 2003, p. 189; compare Exercise 22). Such functions with a sharp peak often result from likelihood functions (products of single likelihood functions when data are independent from each other) that contribute essentially to posterior PDFs.
3. Some samples looks as if they could stem from a normal distribution, however, they may actually be from other symmetric distributions and the sample size is too small to apply appropriate tests allowing rejection of the null hypothesis  $H_0$ : 'sample stems from normal distribution'.

**Music:** We are normal (Bonzo Dog Doo-Dah Band):  
[https://www.youtube.com/watch?v=Gn3hlyCv\\_f8](https://www.youtube.com/watch?v=Gn3hlyCv_f8))

#### Exercise 20 Difference & linear combination of random numbers from normal distributions (\*)

*Study the distributions of*

(a) the difference  $D = X - Y$  between two random variables  $X$  and  $Y$  from normal distributions  $\mathcal{N}(X; \mu_X, \sigma_X^2)$  and  $\mathcal{N}(Y; \mu_Y, \sigma_Y^2)$  with  $\mu_X = 2, \sigma_X^2 = 4, \mu_Y = 1, \sigma_Y^2 = 9$ .

(b) the linear combination  $Z = c_1 X + c_2 Y$  ( $c_1 = 3.2, c_2 = 2.7$ ) of two random numbers  $X$  and  $Y$  from normal distributions  $\mathcal{N}(X; \mu_X, \sigma_X^2)$  and  $\mathcal{N}(Y; \mu_Y, \sigma_Y^2)$  with  $\mu_1 = 2, \mu_2 = 1.5, \sigma_1^2 = 4, \sigma_2^2 = 1$

*Do this using Monte Carlo simulations. How many Monte Carlo runs are typically necessary to obtain an estimate of the mean and variance that deviates by less than 1% from the theoretical values?*

#### Exercise 21 Difference of random numbers from two standard normal distributions (\*)

*Derive Eq. 6.45. Hint: generalize the product rule of probabilities to probability densities, then integrate over a nuisance parameter.*

#### Exercise 22 High powers of functions with single rounded maximum

*Jaynes (2003, p. 189) remarked that 'any function with a single rounded maximum raised to a higher and higher power, goes into a Gaussian function'.  $y(x) = \sin x$  in the range  $0 \leq x \leq \pi$  is such a function.*

- (1) Plot  $y(x)$ ,  $[y(x)]^5$ , and  $[y(x)]^{10}$ .
- (2) Normalize  $[y(x)]^{10}$ , construct a normal approximation, and plot both PDFs.

### 6.4.2 The uniform PDF

The standard uniform PDF (Fig. 6.10) is given by

$$\mathcal{U}(x) = \text{Uniform}(x) = 1 \quad \text{for} \quad 0 \leq x \leq 1 \quad (0 \text{ otherwise}) \quad (6.52)$$

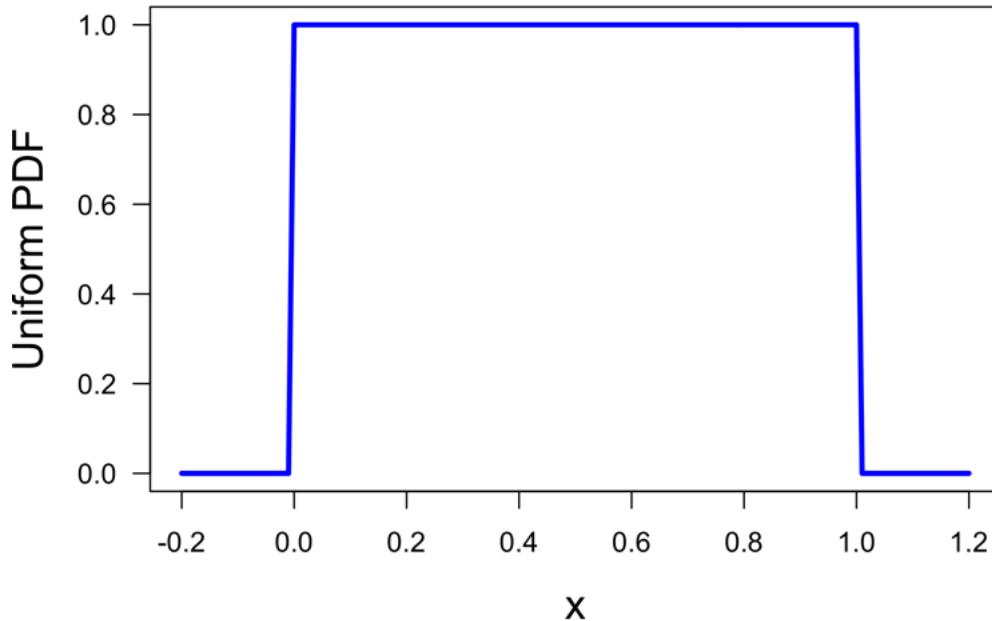


Figure 6.10: The standard uniform PDF:  $\mu = 1/2$ ,  $\sigma^2 = 1/12 \approx 0.0833$  [PDsPDFsUniformPDF.R](#)

**Some properties of the standard uniform distribution:**

**Mean:**  $\mu = 1/2$

**Variance:**  $\sigma^2 = 1/12 \approx 0.0833$

The general uniform PDF is given by

$$\mathcal{U}(x; \alpha, \beta) = \frac{1}{\beta - \alpha} \quad \text{for} \quad \alpha \leq x \leq \beta \quad (0 \text{ otherwise}) \quad (6.53)$$

**Some properties of the general uniform PDF:**

**Mean:**  $\mu = \frac{\alpha + \beta}{2}$

**Variance:**  $\sigma^2 = \frac{(\beta - \alpha)^2}{12}$

**The standardized uniform distribution.** For some applications<sup>13</sup> it is of interest to consider a uniform distribution with the same mean  $\mu$  and variance  $\sigma^2$  as the standard normal distribution, i.e.  $\mu = 0$  and  $\sigma^2 = 1$ ; this distribution will be called the 'standardized uniform distribution' or 'standardized uniform PDF'. The mean value of a uniform PDF is zero when it is symmetric about  $x = 0$ , i.e.  $\beta > 0$  and  $\alpha = -\beta$ , which implies  $\sigma^2 = 4\beta^2/12 = \beta^2/3$  and thus for  $\sigma^2 = 1$  one obtains  $\beta = \sqrt{3} \approx 1.732$ .

<sup>13</sup>For example, application of the Kolmogorov-Smirnov test Section 12.4.3.

**Exercise 23 CDFs of standardized uniform and standard normal PDFs**

*Plot the CDFs of the standardized uniform and the standard normal PDF and find the maximal absolute difference between these two CDFs.*

### 6.4.3 Student's $t$ -distribution

Student's  $t$ -distribution<sup>14</sup> or simply the  $t$  distribution shows up in significance testing (for example,  $t$ -test, Section 12.1.3) as well as in the Bayesian approach (for example, as the marginal posterior for the mean of a normal distribution when assigning a non-informative prior, Gelman et al. 2020, p. 64–66). It is given by

$$\mathcal{T}(t; \nu) = \text{Student-t}(t; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (6.54)$$

( $-\infty < t < +\infty$  and degrees of freedom  $\nu > 0$ ). It is a PDF that describes the distribution of the test statistic  $t$  ( $t$ -test).

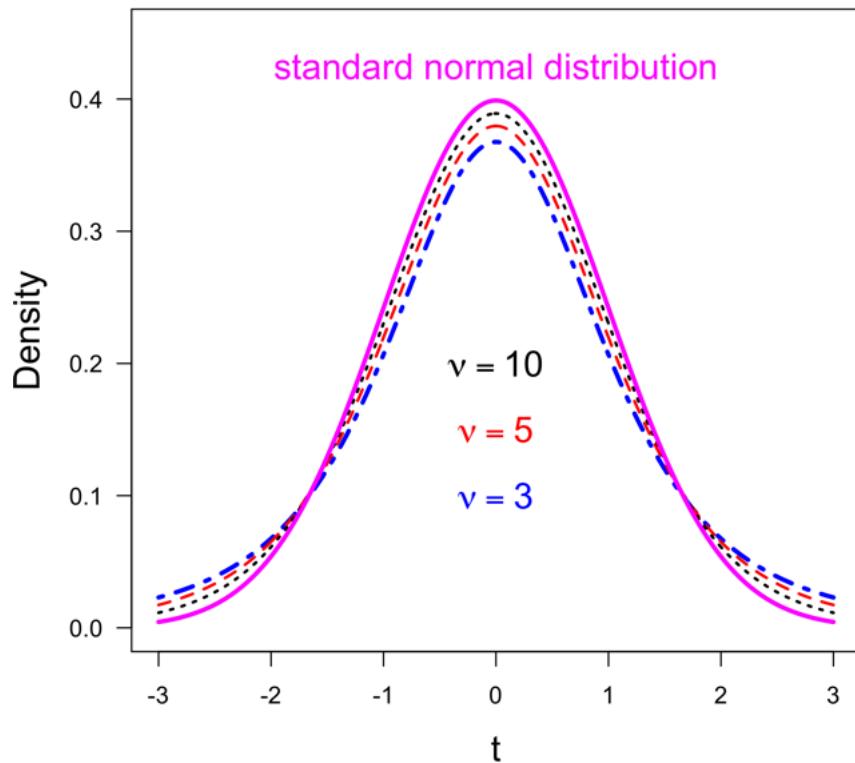


Figure 6.11: The  $t$ -distributions for various degrees of freedom,  $\nu$ , and the standard normal distribution (magenta solid line). The  $t$ -distributions have fatter tails than the standard normal distribution, especially when the degrees of freedom are small. With increasing number of degrees of freedom the  $t$ -distribution approaches the normal distribution. [PDsPDFsStudent-t-Normal.R](#)

**Some properties of  $t$ -distributions:**

**Mean:**  $\mu = 0$  for  $\nu > 1$ .

**Variance:**  $\sigma^2 = \frac{\nu}{\nu - 2}$  for  $\nu > 2$ .

<sup>14</sup>The distribution was derived for the first time by Helmert (1875, 1876a,b) and Lüroth (1976). It was named after William Sealy Gosset who worked for the Guinness brewery and published under the pseudonym 'Student'.

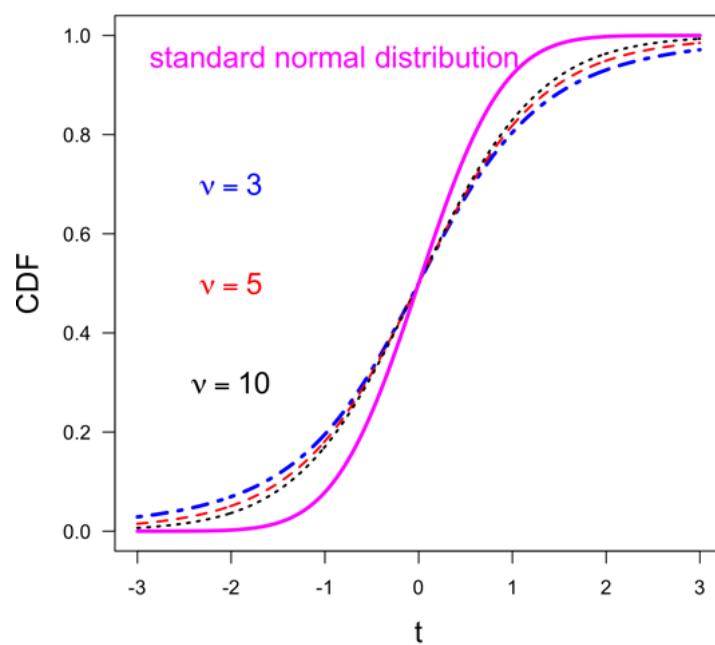


Figure 6.12: CDFs of  $t$ -distributions for various degrees of freedom,  $\nu$ , and the standard normal distribution (magenta solid line). [PDsPDFsStudent-t-NormalCDFs.R](#)

**Relationship between  $t$  and (certain)  $F$ -distributions:**

$$F_{\alpha(1),1,\nu} = t_{\alpha(2),\nu}^2 \text{ (Zar, 2010, p. 347).}$$

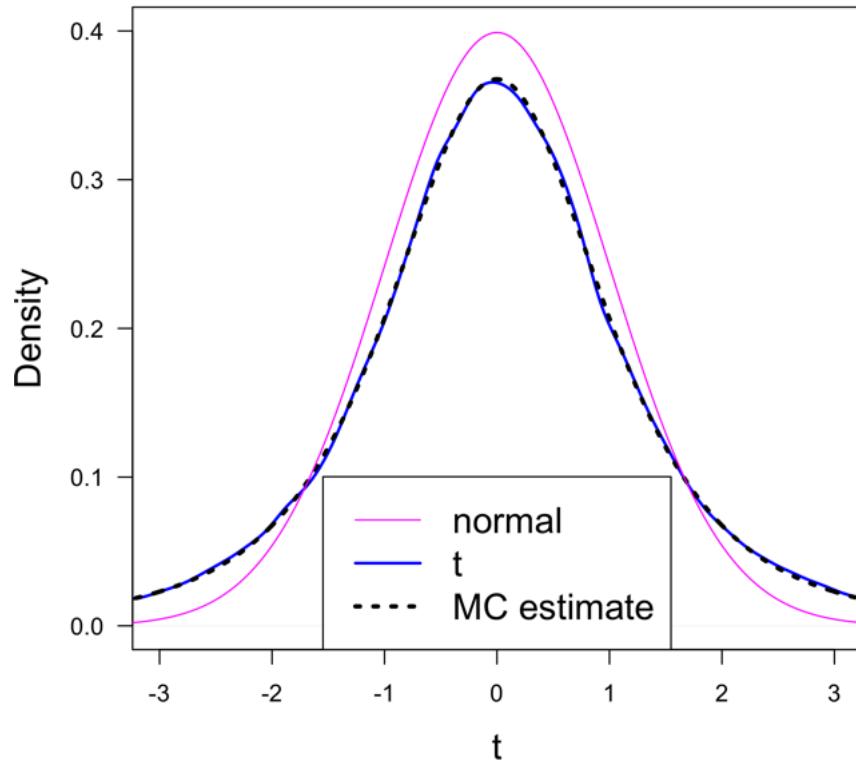


Figure 6.13: The  $t$  distribution  $f(t; \nu = 3)$  estimated from  $10^5$  Monte Carlo runs (blue solid line) compared to the exact  $t$  distribution (black dashed line) and the standard normal distribution (magenta solid line).  
[PDsPDFsStudent-tMonteCarlo.R](#)

#### 6.4.4 The $F$ distribution

The  $F$  distribution<sup>15</sup>  $\mathcal{F}(x; \nu_1, \nu_2)$  is the PDF of the test statistic  $x$  in various hypotheses tests (ANOVA, variance ratio test). It reads

$$\mathcal{F}(x; \nu_1, \nu_2) = \frac{\sqrt{\frac{(\nu_1 x)^{\nu_1} \nu_2^{\nu_2}}{(\nu_1 x + \nu_2)^{\nu_1 + \nu_2}}}}{x B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \quad (6.55)$$

where

$$B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) = \int_0^1 t^{\nu_1-1} (1-t)^{\nu_2-1} dt \quad (6.56)$$

is the beta function (also called the Euler integral). The argument  $x$  can vary between 0 and  $+\infty$ ;  $\nu_1$  and  $\nu_2$  are two degrees of freedom (for example, in ANOVA related to the number of groups and the number of data, respectively).

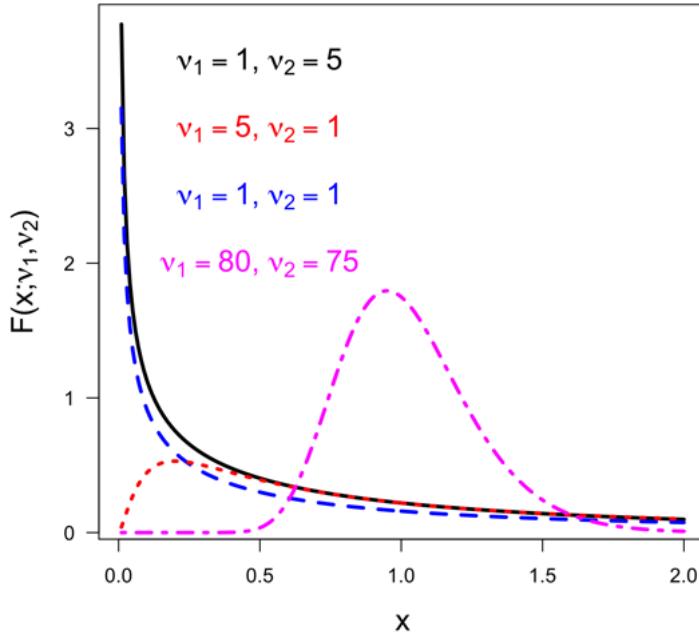


Figure 6.14: The  $F$  distribution for various combinations of degrees of freedom  $\nu_1$  and  $\nu_2$ . For large values of  $\nu_1$  and  $\nu_2$  the  $F$  distribution looks similar to a normal distribution (compare Exercise 24). [PDsPDFs-F-PDFs.R](#)

$\mathcal{F}$  is not symmetric in  $\nu_1$  and  $\nu_2$ , i.e. for  $\nu_1 \neq \nu_2$  in general  $\mathcal{F}(x; \nu_1, \nu_2) \neq \mathcal{F}(x; \nu_2, \nu_1)$  (Fig. 6.14).

The cumulative probability distribution function (CDF) is given by the (regularized) incomplete beta function

$$\mathcal{F}_{\text{CDF}}(x; \nu_1, \nu_2) = I_{\frac{\nu_1 x}{\nu_1 x + \nu_2}}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) \quad (6.57)$$

---

<sup>15</sup>F distribution: also known as Fisher-Snedecor distribution (after Ronald Fisher and George W. Snedecor).

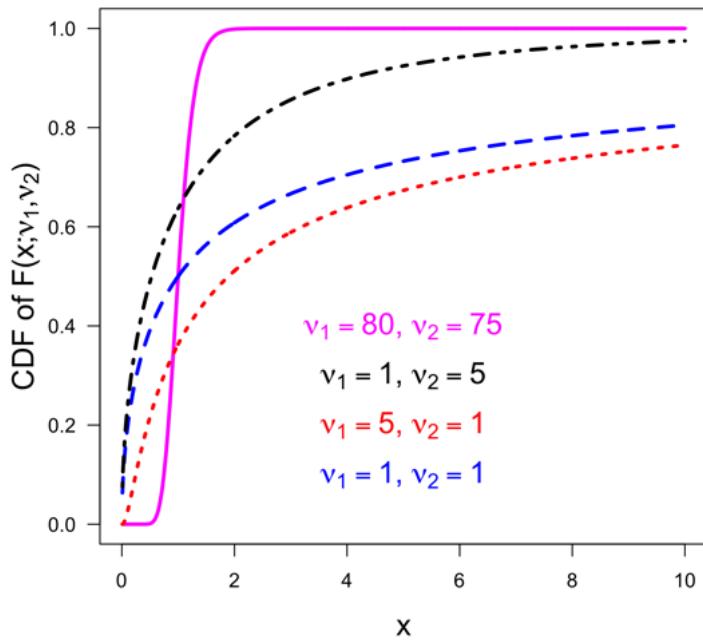


Figure 6.15: CDFs of the  $F$  distributions shown in Fig. 6.14. For small degrees of freedom the CDFs approach 1 quite slowly because a lot of probability is located in the right tail of the  $F$  distributions. [PDsPDFs-F-CDFs.R](#)

**Some properties of  $F$  distributions:**

**Mean:**  $\mu = \frac{\nu_2}{\nu_2 - 2}$  for  $\nu_2 > 2$ .

**Variance:**  $\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$  for  $\nu_2 > 4$ .

**Relationship between  $t$  and (certain)  $F$  distributions:**

$$F_{\alpha(1),1,\nu} = t_{\alpha(2),\nu}^2 \quad (\text{Zar, 2010, p. 347})$$

**Exercise 24 Compare  $F$  and normal distribution**

The  $F$  distribution for large values of the degrees of freedom  $\nu_1$  and  $\nu_2$  looks similar to a normal distribution. Calculate the mean ( $\mu$ ) and the standard deviation ( $\sigma$ ) for the  $F$  distribution with  $\nu_1 = 100 = \nu_2$ , plot  $\mathcal{F}(x; \nu_1, \nu_2)$  and  $\mathcal{N}(x; \mu, \sigma^2)$  and discuss the similarities and differences between the two curves.

## 6.5 PDFs: change of variables

What happens to a probability density  $p(x; \dots)$  if we change the variable  $x$  to  $f(x)$ ? In other words: How does  $p(f(x); \dots)$  look like? The answer

$$p(f(x); \dots) = p(x; \dots) / |df/dx| \quad (6.58)$$

follows from the requirement that probability contained in a differential area must be invariant under change of variables:

$$|p(f(x); \dots) df| = |p(x; \dots) dx|. \quad (6.59)$$

As an example we consider the statement 'the variable  $\frac{\mu - \bar{y}}{\sqrt{s^2/n}}$  follows the  $t$  distribution with  $v = n - 1$  degrees of freedom'.<sup>16</sup> But how is  $\mu$  distributed? We know that  $t(\mu) = \frac{\mu - \bar{y}}{\sqrt{s^2/n}}$  is  $t$ -distributed. The derivative reads  $|dt/d\mu| = |1/\sqrt{s^2/n}|$  because the statistics  $\bar{y}$  and  $s^2$  are constants;  $1/\sqrt{s^2/n} \approx 1.58$  is positive and thus we can leave out the absolute sign. Thus

$$p(\mu | n, \bar{y}, s^2) = p(t | v) |dt/d\mu| = p\left(t \frac{\mu - \bar{y}}{\sqrt{s^2/n}} | v\right) / \sqrt{s^2/n}. \quad (6.60)$$

For  $n = 5$ ,  $\bar{y} = 1.7$ ,  $s^2 = 2$  one expects a unimodal distribution with mode at  $\bar{y} = 1.7$ , maximal value = maximum of  $t$ -distribution ( $\approx 0.375$ )  $\times$  the scaling factor  $1/\sqrt{s^2/n} \approx 0.375 \cdot 1.58 \approx 0.6$  (Fig. 6.16): the  $t$ -distribution (black dashed line) is shifted to the right and 'compressed from both sides', i.e. the spread (variance) is reduced while the maximal value is higher.

When more than one parameter is transformed, i.e.  $y_1 = f_1(x_1, x_2, \dots, x_m)$ , ...,  $y_m = f_m(x_1, x_2, \dots, x_m)$ , the magnitude of the derivative (Eq. 6.59) has to be replaced by the determinant of the Jacobian of the inverse transformations  $f_j^{-1}$ . The Jacobian of the inverse transformations  $f_j^{-1}$  is the matrix of all partial derivatives  $\frac{\partial f_j^{-1}}{\partial y_i}$ ,  $i, j = 1, \dots, m$ .

---

<sup>16</sup>This is the marginal posterior distribution when analyzing (using the Bayesian approach with the non-informative prior  $p(\mu, \sigma^2) \propto 1/\sigma^2$ ) a sample  $y$  of size  $n$  from a normal population with unknown mean  $\mu$  and unknown variance  $\sigma^2$ ;  $\bar{y}$  is the sample mean and  $s^2$  is the sample variance. Compare, for example, Zellner (1971) or Gelman et al., (2020) for a detailed derivation.

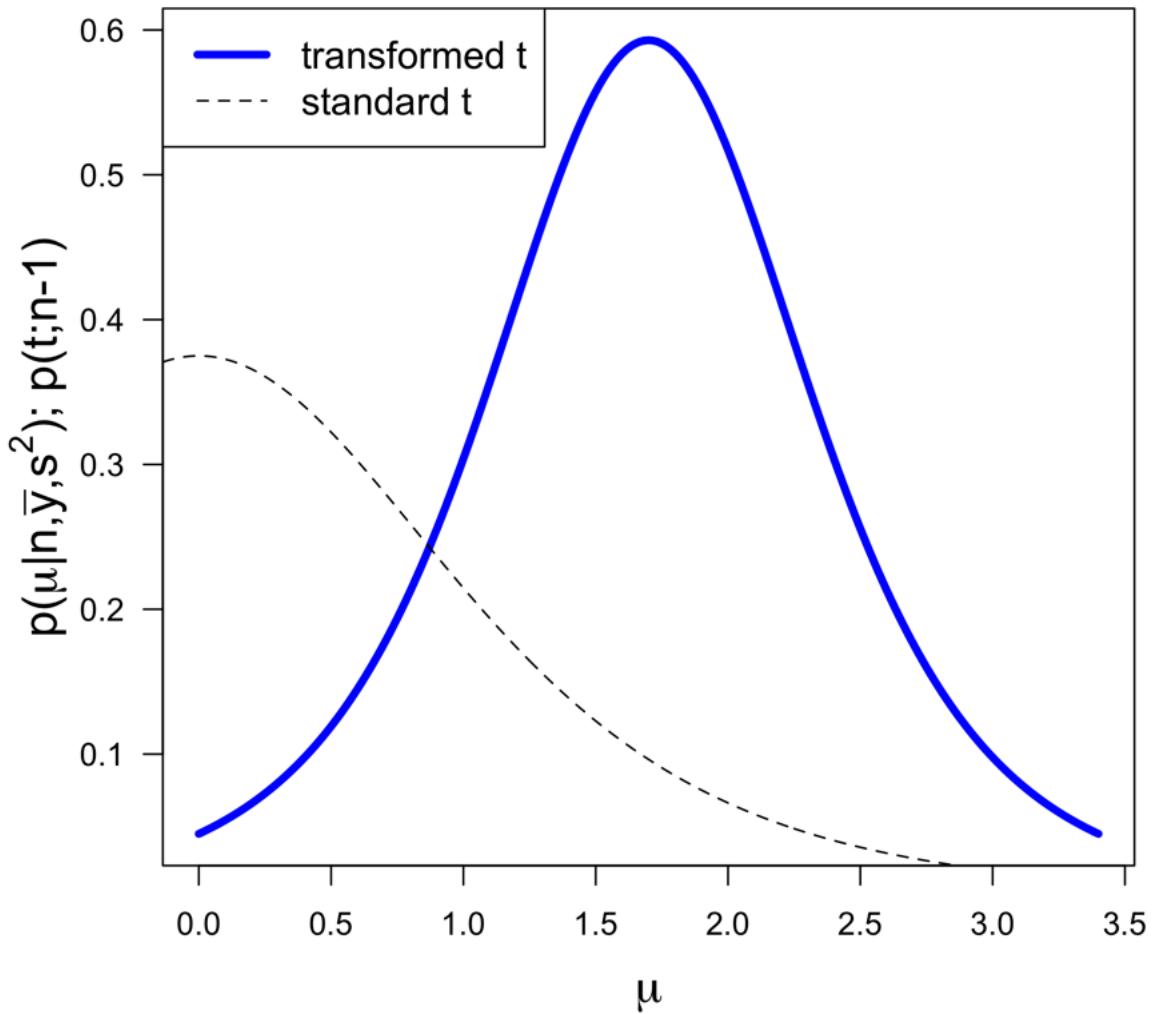


Figure 6.16: The posterior PDF  $p(\mu|n, \bar{y}, s^2)$  for  $n = 5, \bar{y} = 1.7, s^2 = 2$  (blue solid line). One expects a unimodal distribution with mode at  $\bar{y} = 1.7$ , maximal value = maximum of  $t$ -distribution ( $\approx 0.375$ )  $\times$  the scaling factor  $1/\sqrt{s^2/n} = 1.58 \approx 0.375 \cdot 1.58 \approx 0.6$ : the  $t$ -distribution (black dashed line) is shifted to the right and ‘compressed from both sides’, i.e. the spread (variance) is reduced while the maximal value is higher.

[PDsPDFsTransformPDFs.R](#)

## 6.6 Multivariate (joint) PDFs

Up to now, we considered PDFs that were dependent of one variable only (often denoted by  $x$  and varying between  $-\infty$  and  $+\infty$ ) plus parameters (for example,  $\mu$  and  $\sigma$  for the normal PDF). These functions are **univariate PDFs**, although this term is rarely used. In this section we will introduce **multivariate PDFs**, i.e. the function will depend on two ( $x, y$ ) or more variables.

**Example 1:** The bivariate normal PDF with means  $\mu_x$  and  $\mu_y$ , variances  $\sigma_x^2$  and  $\sigma_y^2$ , and correlation  $\rho$  is given by

$$f(x, y; \mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{(1-\rho^2)} \left[ \frac{(x-\mu_x)^2}{2\sigma_x^2} - \rho \frac{x-\mu_x}{\sigma_x} \frac{y-\mu_y}{\sigma_y} + \frac{(y-\mu_y)^2}{2\sigma_y^2} \right]} \quad (6.61)$$

This multivariate distribution possesses various nice properties:

1. The correlation between  $x$  and  $y$  is  $\rho$ .
2. The marginal distribution of  $x$ , i.e.  $f(x; \mu_x, \sigma_x^2) = \int_{-\infty}^{+\infty} f(x, y; \dots) dy$ , is the normal distribution  $\mathcal{N}(x; \mu_x, \sigma_x^2)$ .
3. The marginal distribution of  $y$  is the normal distribution  $\mathcal{N}(y; \mu_y, \sigma_y^2)$ .

(Casella & Berger, 2002, p. 175).

**Example 2:** The bivariate normal distribution for uncorrelated variables  $x$  and  $y$  is given by

$$f(x, y; \mu_x, \sigma_x, \mu_y, \sigma_y) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (6.62)$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (6.63)$$

(this is a special case of Eq. 6.61:  $\rho = 0$ ).

### 6.6.1 Constraints for multivariate PDFs

The constraints for multivariate PDFs are obvious generalizations of the constraints for univariate PDFs, namely the non-negativity of the distribution function and its normalization to one. For a multivariate PDF of two independent variables,  $f(x, y)$ , the constraints read

$$f(x, y) \geq 0 \quad (6.64)$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1 \quad (6.65)$$

## 6.6.2 Expectations, means, and variances

The expectation of the function  $g(x, y)$  with respect to the multivariate PDF  $f(x, y)$  is defined by

$$E_f[g(x, y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy \quad (6.66)$$

if the integral exists<sup>17</sup>. The expectations of  $x$  and  $y$  are the mean values:

$$\mu_x = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f(x, y) dx dy \quad (6.67)$$

$$\mu_y = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x, y) dx dy \quad (6.68)$$

The expectations of  $(x - \mu_x)^2$ ,  $(y - \mu_y)^2$ , and  $(x - \mu_x)(y - \mu_y)$  are the variances and the covariance:

$$\sigma_x^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x, y) dx dy \quad (6.69)$$

$$\sigma_y^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (y - \mu_y)^2 f(x, y) dx dy \quad (6.70)$$

$$\sigma_{xy} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy \quad (6.71)$$

## 6.6.3 Marginal PDFs

*Marginal PDFs* can be derived from multivariate PDFs by integration over one or more variables, however, less than the total number of independent variables.

**Example** The marginal PDFs of the bivariate normal distribution (correlation  $\rho = 0$ ) are defined by

$$f(x; \mu_x, \sigma_x) = \int_{-\infty}^{+\infty} f(x, y; \mu_x, \sigma_x, \mu_y, \sigma_y) dy \quad (6.72)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}} dy \quad (6.73)$$

$$= \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \quad (6.74)$$

and

$$f(y; \mu_y, \sigma_y) = \int_{-\infty}^{+\infty} f(x, y; \mu_x, \sigma_x, \mu_y, \sigma_y) dx \quad (6.75)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}} dx \quad (6.76)$$

$$= \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (6.77)$$

---

<sup>17</sup>Please note that other authors use different notations. Casella & Berger, 2002, p. 144, for example, write  $E_g(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy$ .

### 5: Exponential magic

The normal distribution is a member of the family of exponential distributions. A nice property of normal distributions is the fact that the product of two normal distributions leads – after normalization – again to a normal distribution

$$\text{normal}(x; \mu_1, \sigma_1) \text{normal}(x; \mu_1, \sigma_1) \propto \text{normal}(x; \mu_3, \sigma_3) \quad (6.78)$$

and the parameters  $\mu_3$  and  $\sigma_3$  can be calculated from the parameters  $\mu_1, \sigma_1, \mu_2$ , and  $\sigma_2$  as follows

$$\mu_3 = \frac{\mu_1 + \mu_2}{\frac{\sigma_1^2}{1} + \frac{1}{\sigma_2^2}} \quad \text{and} \quad \frac{1}{\sigma_3^2} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \quad (6.79)$$

(compare Exercise 19). Products of distributions are used to formulate likelihoods, likelihood functions, and posteriors (= product of likelihood function and prior).

Normal functions are, however, not the only members of the exponential family. The expression for the beta distribution, for example,

$$\text{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}, \quad 0 \leq p \leq 1, \quad \alpha > 0, \quad \beta > 0. \quad (6.80)$$

does not contain an exponential function. However, the identity

$$x^\alpha = e^{\alpha \log x} \quad (6.81)$$

allows us to rewrite potentials as exponential functions. From Eq. (6.81) one further derives

$$x^\alpha x^\beta = e^{\alpha \log x} e^{\beta \log x} = e^{(\alpha+\beta) \log x} = x^{\alpha+\beta} \quad (6.82)$$

For beta distributions it is easy to show that

$$\text{Beta}(p; \alpha_1, \beta_1) \text{Beta}(p; \alpha_2, \beta_2) \propto \text{Beta}(p; \alpha_3, \beta_3) \quad (6.83)$$

with  $\alpha_3 = \alpha_1 + \alpha_2 - 1$  and  $\beta_3 = \beta_1 + \beta_2 - 1$ .

Other members of the exponential family: exponential, gamma, chi-squared, Wishart. Not a member: *t*-distribution.

# Chapter 7

## Central Limit Theorem (CLT)

Suppose you take  $N = 10000$  random numbers from the standard uniform distribution, i.e. real numbers from the interval between 0 and 1, and add them up. Common sense tells you that the sum is about  $N \cdot \mu = 5000$  because the mean value of the random numbers is 0.5. 'About' refers to the variations due to randomness: the sum is usually either a bit smaller or larger than 5000. If you repeat this exercise  $M = 1000$  times, you will get sums varying around a mean of 5000. How does the distribution of sums look like and how large is its dispersion (standard deviation)?

The answer to this question is given by the Central Limit Theorem (CLT) stating that the distribution of sums of random numbers approaches a normal distribution with certain values of its mean and variance. In our case, the mean of the sums  $\mu_{\Sigma}$  approaches

$$\mu_{\Sigma} = N \cdot \mu_{\text{std. uniform}} = 10^4 \cdot 0.5 = 5000 \quad (7.1)$$

(this is what common sense tells us), the variance of the sums approaches

$$\sigma_{\Sigma}^2 = N \cdot \sigma_{\text{std. uniform}}^2 = 10^4 \cdot 1/12 \approx 10^5 \cdot 0.0833 \approx 833 \quad (7.2)$$

(this is a bit beyond common sense), and thus  $\sigma_{\Sigma} \approx 28.9$ .

Although the theorem has been postulated already in 1733 by Abraham de Moivre, it took almost 200 years before Pólya (1920) could give a mathematical proof. The CLT shows that adding up contributions from a large number of random processes leads to normal distributions. This could 'explain' – at least in part – the frequent occurrence of (close to) normal distributions in nature as well as in technical devices. However, be aware that distributions encountered in real life can deviate from normal distributions and can, for example, possess 'fatter tails'.

An important take home message from the CLT is: **Don't add up standard deviations!** The standard deviation of the sums distribution has to be estimated by the 'detour via variances', i.e. first add up the variance, then take the square route. A similar formula (although including weights) applies for the propagation of uncertainties (Section 8.1).

**Central Limit Theorem** (Moivre, Laplace, Rayleigh, Edgeworth, Pólya)

The sum  $X$  of  $N$  independent variables  $x_k$ ,  $k = 1, 2, \dots, N$ , each taken from a distribution of mean  $\mu_k$  and variance  $\sigma_k^2$

(a) has an expectation value (mean)

$$\mu = \langle X \rangle = \sum_{k=1}^N \mu_k \quad (7.3)$$

(b) has variance

$$\sigma^2 = \sum_{k=1}^N \sigma_k^2 \quad (7.4)$$

(c) The PDF of  $X$  becomes normal (Gaussian) as  $N \rightarrow \infty$ .

Proof: (a) and (b) is simple; (c) requires Fourier transformation and cumulants (see, for example, Barlow, 1999).

## 7.1 Normal distribution: sum of random numbers

According to the Central Limit Theorem (Chapter 7) the distribution of the sum of  $N$  random numbers approaches a normal distribution for large  $N$  ( $N \rightarrow \infty$ ). In our Monte Carlo simulation we generate  $M = 10^3$  sums over  $N = 10000$  random numbers from the standard uniform PDF ( $0 \leq x \leq 1$ ). The histogram of the resulting sums (Fig. 7.1) looks similar to a random sample from a normal PDF.

The PDF of the sums of random numbers can be estimated from the generated sums with the R routine `density()`. The result (blue solid line in Fig. 7.2) is close to the theoretical distribution (red broken line) derived from the Central Limit Theorem. The agreement between the Monte Carlo derived density and the theoretical PDF can be improved by increasing the number of Monte Carlo runs: Fig. 7.3 shows the result for  $M = 10^5$ .

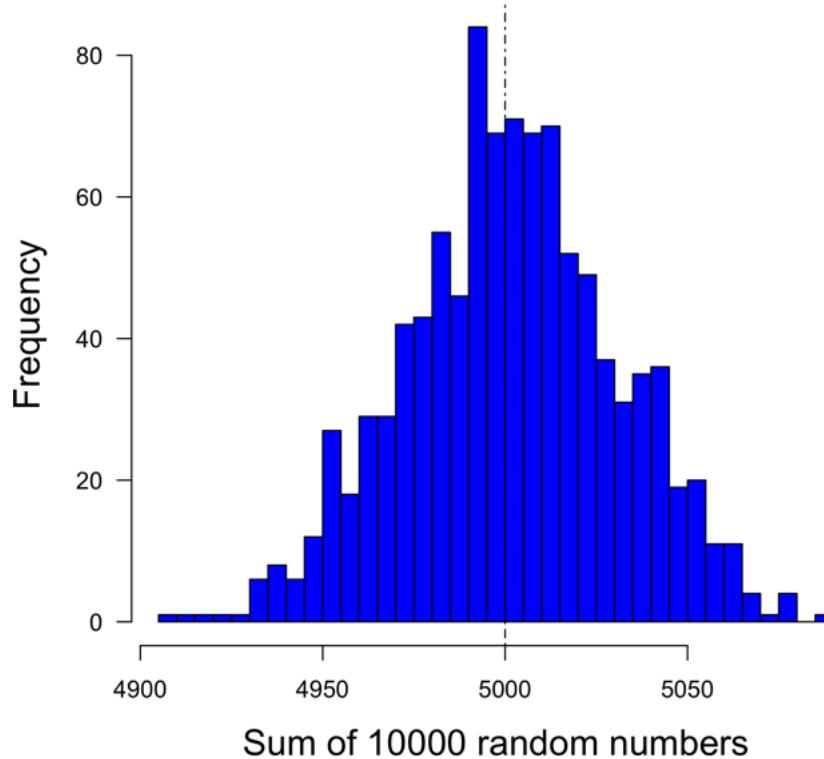


Figure 7.1: Histogram of  $M = 10^3$  sums over  $N = 10000$  random numbers each from the standard uniform distribution. [CLThistMC.R](#)

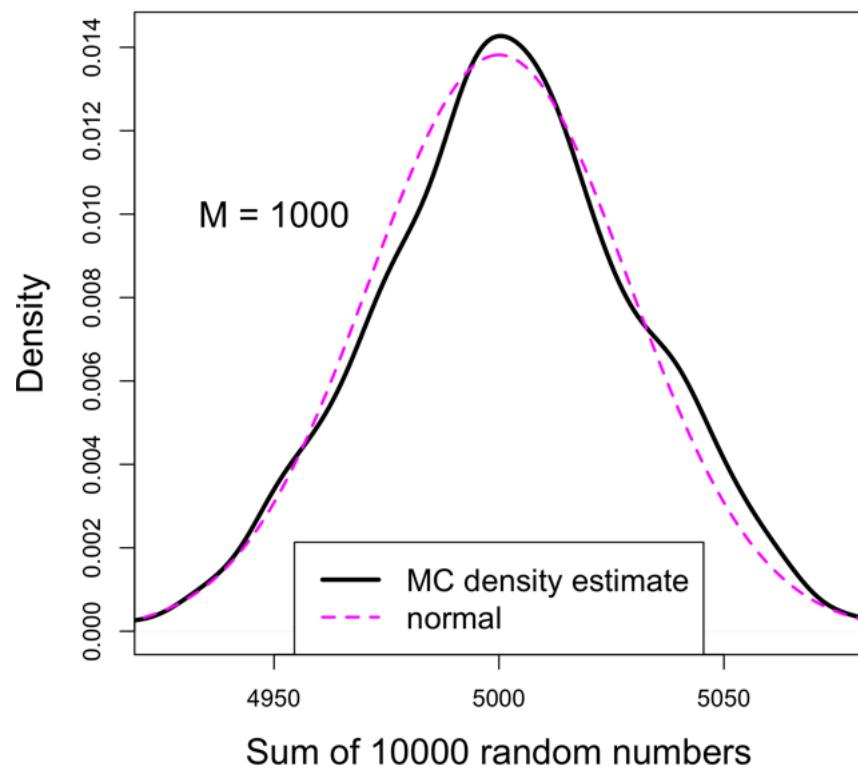


Figure 7.2: Estimate of the density, based on  $M = 10^3$  Monte Carlo runs, for sums over  $N = 10000$  random numbers each from the standard uniform distribution (black line) and the normal distribution with mean  $\mu = 5000$  and standard deviation  $\sigma = \sqrt{N/12} \approx 28.9$  derived from the Central Limit Theorem (magenta broken line). [CLTdensity1e3MC.R](#)

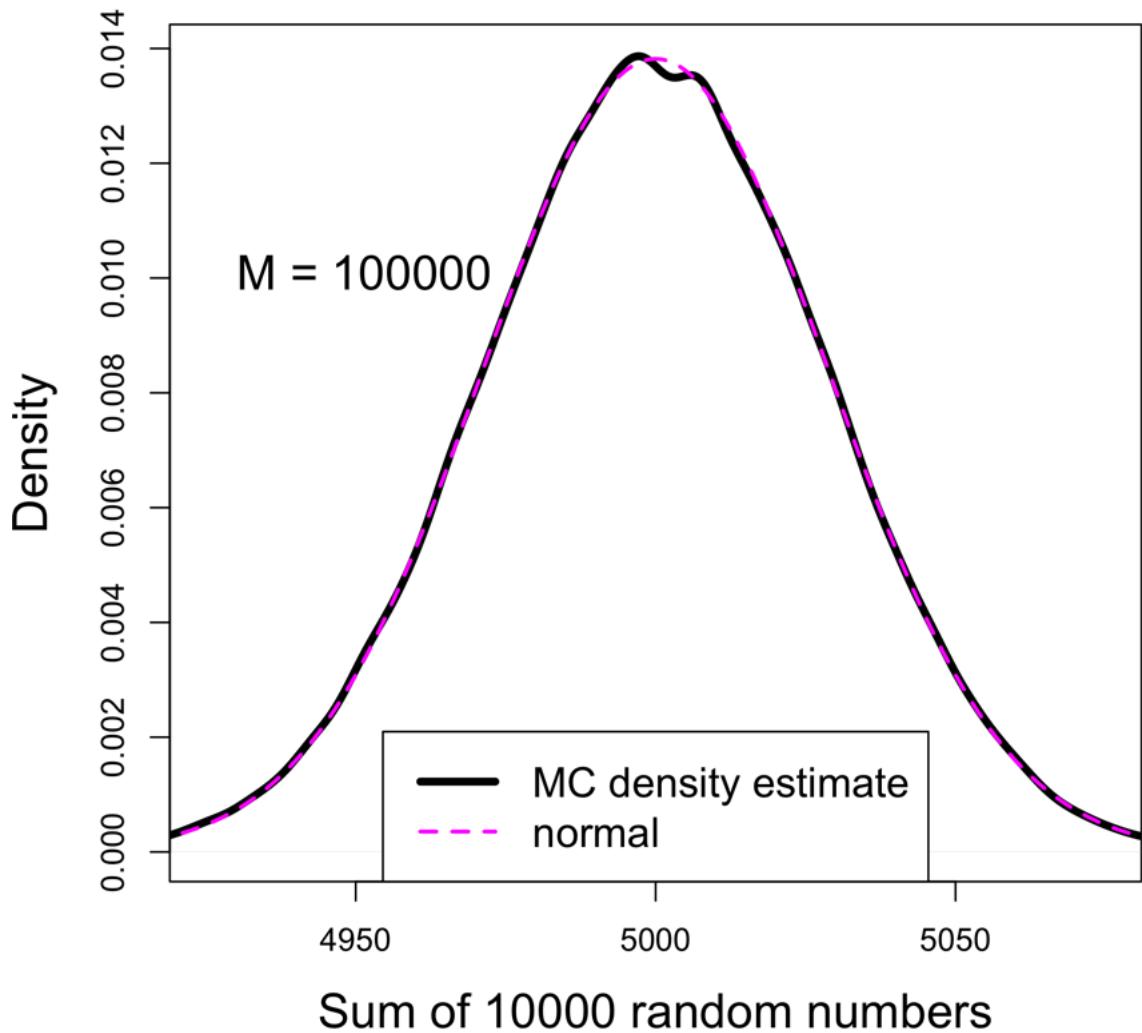


Figure 7.3: Estimate of the density, based on  $M = 10^5$  Monte Carlo runs, for sums over  $N = 10000$  random numbers each from the standard uniform distribution (blue line) and the normal distribution with mean  $\mu = 5000$  and standard deviation  $\sigma = \sqrt{N/12} \approx 28.9$  derived from the Central Limit Theorem (red curve).  
[CLTdensity1e5MC.R](#)

## 7.2 Central Limit Theorem: summary & take-home message

Central Limit Theorem:

- (a) mean of the sums = sum of the means,
- (b) variance of the sums = sum of the variances<sup>1</sup>,
- (c) the probability density function of the sums becomes normal (for large number of independent variables).

**Warning:** The CLT works better in the centre of the distribution than far away from it (Fig. 7.4): relative deviations between the estimated density and the normal distribution are small near the maximum, however, can be large in the tails. The sum of random numbers from the standard uniform PDF (between 0 and 1) is always non-negative whereas a normal PDF with mean  $\mu > 0$  would yield a probability even for negative values.

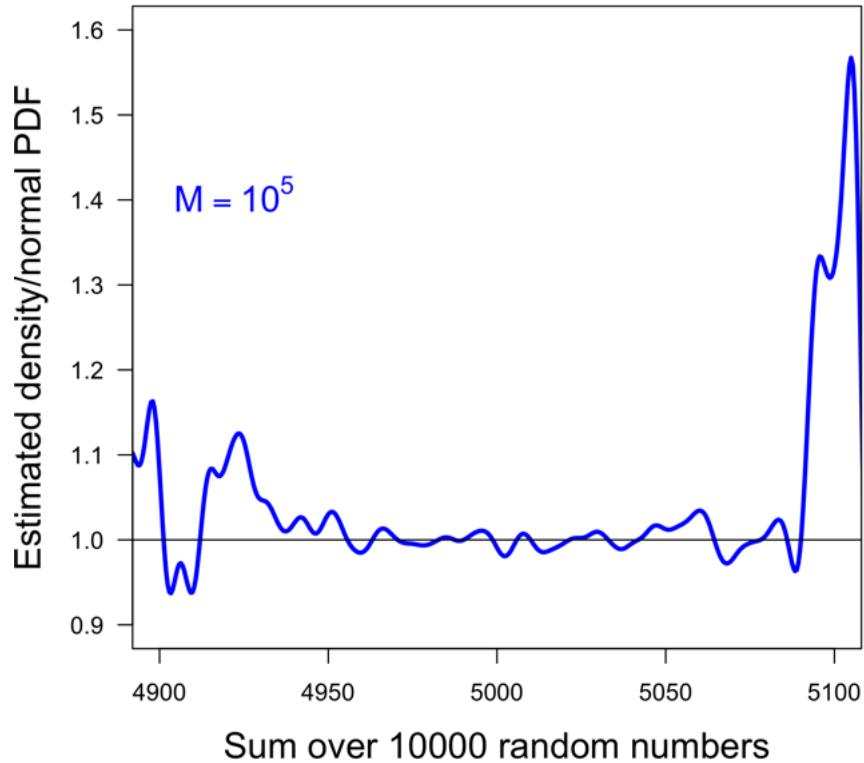


Figure 7.4: Ratio between the estimated and the normal PDF based on the Central Limit Theorem (CLT): relative deviations can be large in the tails (beyond  $\pm 3\sigma$ ). [CLTdeviationsTails.R](#)

---

<sup>1</sup>Please note that the standard deviations do not sum up! The mean and the variance are the most 'natural' parameters to characterize PDFs because they can simply be added up to characterize distribution of the sums.

**Notation:** The term 'central limit theorem' is derived from the German 'zentraler Grenzwertsatz' coined by Pólya (1920). The theorem is central to probability theory. There is no 'central limit'. In the limit  $N \rightarrow \infty$  one obtains the central (called normal or Gaussian) distribution.

**Further reading (diffusion, Central Limit Theorem):**

Narasimhan, T.N., The dichotomous history of diffusion, Physics Today, July, p.48-53, 2009.

# Chapter 8

## Uncertainty

*In the good old days we used to talk about **accuracy** (considered as a quantitative concept, namely the deviation of the measured value from the **true value**) and **precision** (another quantitative concept, namely a measure of the dispersion of repeated measurements). It was easy to estimate the precision, but it was usually very difficult to estimate the accuracy. Two different types of errors were related to these concepts: (1) **systematic errors** leading to systematic differences between true and measured values and (2) **statistical** or **random errors** generating a dispersion of measured values. Systematic errors are notoriously difficult to detect. The statistical errors were often assumed to be normally distributed for continuous quantities or following binomial or Poisson distributions for discrete quantities. Yes, so were the good old days . . .*

*Taylor & Kuyatt (1994) introduced many new concepts and corresponding terminology (including 'accuracy' as a qualitative concept, 'repeatability', 'reproducibility', 'error', 'systematic error', 'random error', etc.). However, the estimate of uncertainty remains a challenge and methods to minimize systematic errors are often very specific for disciplines and research topics (reference materials for total alkalinity is an example, compare Dickson et al., 2003).*

*A measurement of a quantity  $x$  is not complete before the value **and its uncertainty** have been determined, i.e. before one can write  $\hat{x} = \hat{\mu} \pm \hat{\sigma}_{\hat{\mu}}$ , i.e. the estimate of  $x$ ,  $\hat{x}$ , consists of an estimate of the mean,  $\hat{\mu}$ , and an estimate of its uncertainty  $\hat{\sigma}_{\hat{\mu}}$ .<sup>1</sup> The mean is often (using non-robust estimators) estimated by the sample mean  $\bar{x}$  and the uncertainty by the standard error of the mean  $\hat{\sigma}_{\hat{\mu}} = \hat{\sigma} / \sqrt{n} = SE$  where  $\hat{\sigma}$  is the standard deviation of the sample.<sup>2</sup> Knowledge of the uncertainty  $\hat{\sigma}_{\hat{\mu}}$  is essential, for example, when one wants to infer if the true mean  $\mu$  is different from zero or different from another true mean  $\mu_2$  of another population.*

**Further reading:** My chapter on uncertainty scratches only at the surface of the topic with emphasis on a technical aspect (propagation of variances). An excellent introduction to many aspects of uncertainty is given by Spiegelhalter (2024; highly recommended!).

---

<sup>1</sup>Please note that some communities report multiples of  $\hat{\sigma}_{\hat{\mu}}$  as uncertainty (for example, 'two-sigma' uncertainties).

<sup>2</sup>In experimental physics it is sometimes possible to derive also estimates of systematic errors and thus one finds results like "Clear evidence for the production of a neutral boson with a measured mass of  $126.0 \pm 0.4$  (stat)  $\pm 0.4$  (sys) GeV is presented. This observation, which has a significance of 5.9 standard deviations, . . ." (ATLAS Collaboration, 2012) The '5.9 standard deviations' are actually '5.9 standard errors'. Notice that the statistical error (standard error of the mean based on sample standard deviation) has been reduced to the same level as the estimate of the systematic error.

Another example is the mass of the W boson:  $M_W = 80,4335 \pm 6.4_{\text{stat}} \pm 6.9_{\text{syst}} = 80,4335 \pm 9.4 \text{ MeV}/c^2$ . "This measurement is in significant tension with the standard model expectation." (CDF collaboration & Aaltonen et al. 2022) because it deviates by 0.09% (equivalent to  $7\sigma$ ) from the value based on the standard model (Castelvecchi & Gibney, 2022).

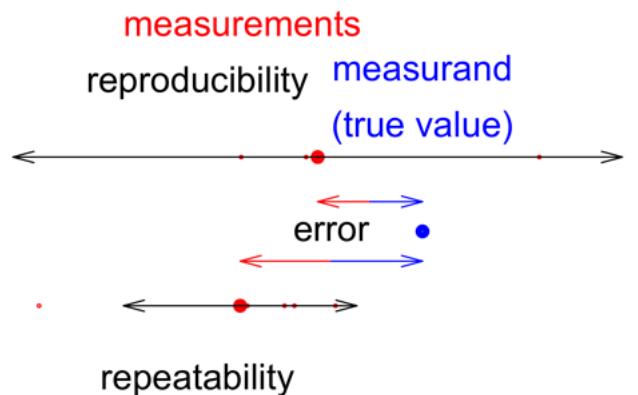


Figure 8.1: True value, uncertainty, & errors: There is only a single true value (big blue dot); unfortunately, we don't know its value. Measurements (small red dots) deviate from the true value because of systematic and random errors. Repeatability is a measure of uncertainty for repeated measurements made by the same person using the same instruments, same set-up etcetera. Reproducibility is a measure of uncertainty for measurements performed by different persons and/or different methods. Please note that a larger value of reproducibility (= larger dispersion) than for repeatability does not necessarily mean that the mean value of reproduced measurements (big red dot) is further away from the true value than the mean value of repeated measurements. [UncertaintyDefs.R](#)

## 8.1 Propagation of uncertainty

Certain quantities can not be measured for various reasons (principally not possible or way too expensive). An example from oceanography is the density of seawater which can be measured in the lab on land, but which is never measured during field studies on board research vessels. Physical oceanographers measure temperature, conductivity, and pressure. Salinity can be derived from conductivity. Finally, density is calculated from temperature, salinity, and pressure. If the unmeasured quantity  $q$  is related to the measured quantities  $x_1, x_2, \dots$  by  $q = f(x_1, x_2, \dots)$ , what's the uncertainty of  $q$  given the uncertainties in  $x_1, x_2, \dots$ ? The 'law of propagation of uncertainty' is used to derive an approximate answer. Monte Carlo simulations can provide more detailed insight especially when  $f(x)$  is highly non-linear.

### 8.1.1 Law of propagation of uncertainty for $q = f(x)$

The law of propagation of uncertainty<sup>3</sup> in the one-dimensional case ( $q = f(x)$ ) reads: the square of the uncertainty of the derived quantity,  $\hat{\sigma}_q^2$ , is proportional to the square of the uncertainty of the original quantity,  $\hat{\sigma}_x^2$ , and the proportionality factor is given by the square of the first derivative of the function relating  $x$  and  $q$  at the estimated value of  $x$ ,  $\hat{x}$ , or as formula

$$\hat{\sigma}_q^2 = \left( \frac{df}{dx} \right)^2 \Big|_{x=\hat{x}} \hat{\sigma}_x^2. \quad (8.1)$$

Suppose one has estimated  $\hat{x} = 5 \pm 0.3$ , the uncertainty of  $q(x)$  can be calculated as follows:

1.  $q = f(x) = x + \text{const} \Rightarrow df/dx = 1 \Rightarrow \hat{\sigma}_q^2 = 1^2 \cdot 0.3^2 = 0.3^2 \Rightarrow \hat{\sigma}_q = 0.3$ , i.e. a simple shift of  $x$  by a constant does not change the uncertainty (consistent with common sense).
2.  $q = f(x) = x^2 \Rightarrow df/dx = 2x \Rightarrow (df/dx = 2x)_{x=5} = 10 \Rightarrow \hat{\sigma}_q^2 = 10^2 \cdot 0.3^2 = 9 \Rightarrow \hat{\sigma}_q = 3$ , i.e. a non-linear function (quadratic) can change the uncertainty quite a bit.
3.  $q = f(x) = e^x \Rightarrow df/dx = e^x \Rightarrow (df/dx = e^x)_{x=5} = e^5 \approx 148.41 \Rightarrow \hat{\sigma}_q^2 = e^{10} \cdot 0.3^2 \approx 1982.4 \Rightarrow \hat{\sigma}_q \approx 44.5$ , i.e. a highly non-linear function (exponential) can change the uncertainty even more.

### 8.1.2 Law of propagation of uncertainty (general case)

In the general case,  $q$  is a function of  $n$  variables  $x_1, x_2, \dots, x_n$ . The squared uncertainty of  $q$  (estimated variance) is given by (Taylor & Kuyatt, 1994)<sup>4</sup>

$$\hat{\sigma}_q^2 = \sum_{i=1}^n \left( \frac{\partial f}{\partial x_i} \right)^2 \hat{\sigma}_{x_i}^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \hat{\sigma}_{x_i, x_j}^2 \quad (8.2)$$

where  $\hat{\sigma}_{x_i, x_j}^2$  is the estimated covariance of  $x_i$  and  $x_j$ . Eq. (8.2) is based on the first-order Taylor series approximation of  $f(x_1, x_2, \dots, x_n)$ . If all  $x_j, x_k, j \neq k$  are independent of each other, the covariances can be neglected and Eq. (8.2) simplifies to

$$\hat{\sigma}_q^2 = \sum_{i=1}^n \left( \frac{\partial f}{\partial x_i} \right)^2 \hat{\sigma}_{x_i}^2 \quad (8.3)$$

or in words: the variance of  $q$  is the weighted sum over the variances of the variables  $x_i$ , whereby the weights are given by the squared partial derivatives of  $f$  with respect to  $x_i$ . The usefulness of variances shows up again: we are allowed to add up variances (as in the Central Limit Theorem), however, we have to take weights into account here. Although in the end we will express the uncertainty in  $q$  as a standard deviation, we are not allowed to add up the standard deviations of the  $x_i$ .

<sup>3</sup>Also called 'law of combination of errors'; German speakers can express it in a single word: 'Fehlerfortpflanzungsgesetz'.

<sup>4</sup>However, a slightly different notation is used here. Further: The partial derivatives have to be evaluated at the estimated values of  $x_1, x_2, \dots, x_n$ ; we leave this out of the notation in Eqs. (8.2), (8.3) etc. in order to avoid clumsy expressions.

### 8.1.3 Law of propagation of uncertainty for $q = f(x, y)$

Let us apply Eq. (8.3) to two-dimensional cases  $q = f(x_1, x_2)$ . In order to get rid of the indices we will use the notation  $x \equiv x_1$  and  $y \equiv x_2$ . The simplified law of propagation of uncertainty reads

$$\hat{\sigma}_q^2 = \left( \frac{\partial f}{\partial x} \right)^2 \hat{\sigma}_x^2 + \left( \frac{\partial f}{\partial y} \right)^2 \hat{\sigma}_y^2 \quad (8.4)$$

In the following examples, we will use  $\hat{x} = 5 \pm 0.3$  and  $\hat{y} = 2 \pm 0.5$ .

1.  $q = f(x, y) = 2x + 5y \Rightarrow \partial f / \partial x = 2, \partial f / \partial y = 5 \Rightarrow \hat{\sigma}_q^2 = 2^2 \cdot 0.3^2 + 5^2 \cdot 0.5^2 = 6.61 \Rightarrow \hat{\sigma}_q \approx 2.6$  and thus  $\hat{q} = 20 + 2.6$ .
2.  $q = f(x) = x \cdot y \Rightarrow \partial f / \partial x = y, \partial f / \partial y = x \Rightarrow \hat{\sigma}_q^2 = 5^2 \cdot 0.3^2 + 2^2 \cdot 0.5^2 = 6.61 \Rightarrow \hat{\sigma}_q \approx 2.6$  and thus  $\hat{q} = 10 + 2.6$ .
3.  $q = f(x) = x/y \Rightarrow \partial f / \partial x = y^{-1}, \partial f / \partial y = -x y^{-2} \Rightarrow \hat{\sigma}_q^2 = 0.5^2 \cdot 0.3^2 + 25/16 \cdot 0.5^2 \approx 0.41 \Rightarrow \hat{\sigma}_q \approx 0.64$  and thus  $\hat{q} = 2.5 + 0.64$ .

A practical example of propagation of uncertainty can be found in the 5. IPCC report (2013). CO<sub>2</sub> emissions from fossil fuel combustion between 1750 to 2011 were estimated by  $375 \pm 30$  Pg C. Over the same time span, deforestation and other land use change led to CO<sub>2</sub> releases of  $180 \pm 80$  Pg C. The sum of these two sources is  $375 + 180 = 555$  Pg C. What's the uncertainty of the sum? One has to simply add the two variance,  $30^2 + 80^2 = 7300$ , and then take the square root to obtain  $\sqrt{7300} \approx 85.4$  Pg C.

### 8.1.4 Propagation of uncertainty: Monte Carlo simulations

The law of propagation of uncertainty Eq. (8.2) is based on the first-order Taylor series approximation. This law often works very well, however, it can fail for highly non-linear functions  $f()$  and/or large uncertainties in the  $x_i$ . In the latter cases Monte Carlo simulations can yield better results.

The basic idea of Monte Carlo simulations for the propagation of uncertainty is quite simple. Suppose  $\hat{x} = 2.1 \pm 0.5$  and assume that the variations around the mean follow a normal distribution. One generates  $M$  ( $M$  = number of Monte Carlo runs) random numbers  $rx$  from a normal distribution with (true) mean  $\mu = 2.1$  and (true) standard deviation  $\sigma = 0.5$ . Next one applies  $f()$  to these random numbers yielding  $M$  random values  $rf$ . Finally one estimates the central tendency and variance from  $rf$ . Below we will discuss three non-linear functions, namely  $f(x) = x^2$ ,  $f(x) = 1/x$ , and  $f(x) = \tanh(3x - 6)$  for  $\hat{x} = 2.1 \pm 0.5$ .

In the **first example** we consider  $q = f(x) = x^2$ . The results of the Monte Carlo simulation with  $M = 10^4$  runs are plotted in the form of a histogram in Fig. 8.2. The law of propagation of uncertainty yields  $\hat{q}_{\text{law-of-prop.}} = 4.41 \pm 2.10$ . One obtains  $\hat{q}_{\text{non-robust}} = 4.633 \pm 2.135$  when using the conventional (non-robust) estimators arithmetic mean and standard deviation of the sample. Robust estimators (median for the central tendency and normalized median absolute deviation (MADN) for the dispersion) yield  $\hat{q}_{\text{robust}} = 4.4 \pm 2.1$  which is closer to the result from the law of propagation of uncertainty. The estimated density from  $rf$  (Fig. 8.3, blue solid line) is approximated reasonably well by a normal distribution with  $\mu = 4.4$  and  $\sigma = 2.1$ , however, deviations in the tails and in the location of the maximum are clearly recognizable.

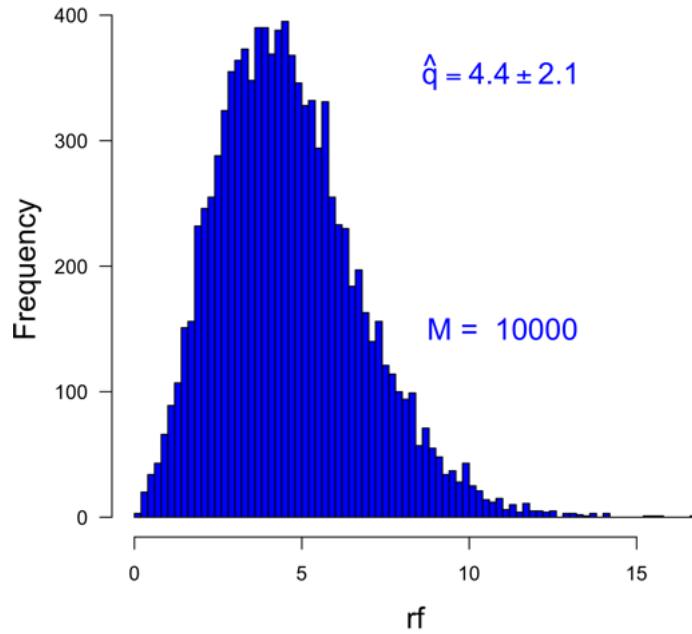


Figure 8.2: Histogram of  $rf = f(rx) = rx^2$  where  $rx$  are  $M = 10^4$  random numbers from a normal distribution with mean  $\mu = 2.1$  and standard deviation  $\sigma = 0.5$ . Robust estimates for the distribution of  $q = f(x) = x^2$  (median for the central tendency and normalized median absolute deviation (MADN) for the dispersion) yield  $\hat{q}_{\text{robust}} = 4.4 \pm 2.1$ . [UncertaintyMCxsquared.R](#)

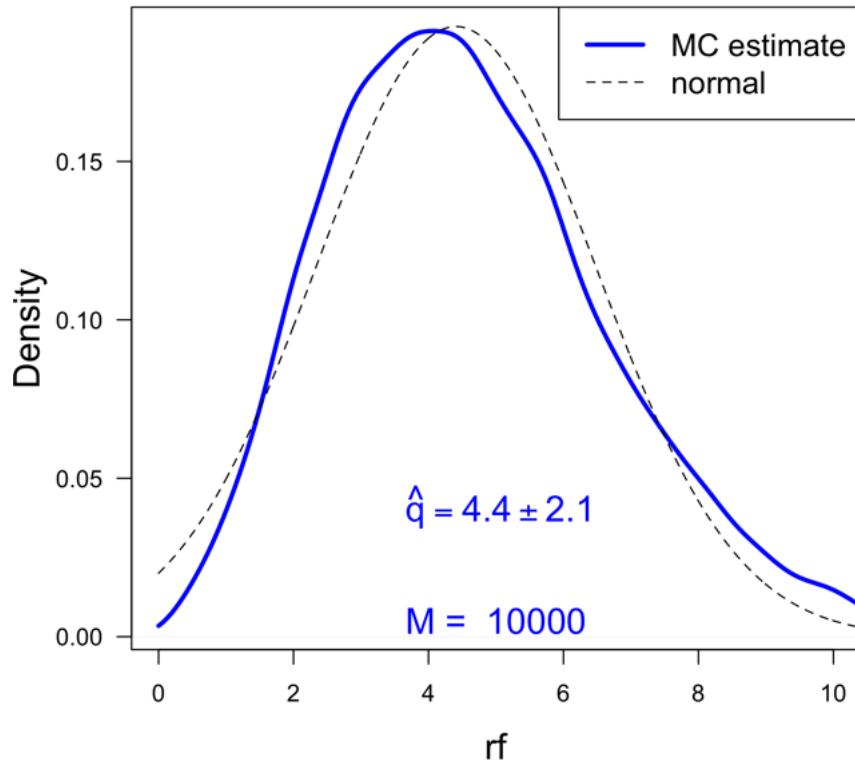


Figure 8.3: Estimate of the density (pdf) of  $q = f(x) = x^2$  based on the results of the Monte Carlo simulation with  $M = 10^4$  runs (blue solid line). A normal distribution with  $\mu = 4.4$  and  $\sigma = 2.1$  (red dashed line) approximates the estimated density reasonably well, however, deviations in the tails and in the location of the maximum are clearly recognizable. [UncertaintyMCxsquared.R](#)

In the **second example** we consider  $q = f(x) = 1/x$  for  $\hat{x} = 2.1 \pm 0.5$ . From the law of propagation of uncertainty one obtains  $\hat{q} = 0.476 \pm 0.113$ . The Monte Carlo simulation with  $M = 10^4$  runs yields  $\hat{q}_{\text{non-robust}} = 0.514 \pm 0.198$  for the sample mean  $\pm$  one standard deviation and  $\hat{q}_{\text{robust}} = 0.476 \pm 0.111$  for the median  $\pm$  MADN. The histogram of the  $rf$  values (Fig. 8.4) indicates deviations from normality which is further supported by the estimate of the pdf ('density'; blue solid line in Fig. 8.5).

Some values of  $rf$  are quite large (far away from the mean) because of division of 1 by small values of  $rx$ . The resulting 'fat' right tail of the distribution makes estimation of mean and standard deviation difficult. Increasing the number  $M$  of Monte Carlo runs is probably also not an option because of the increasing chance for creating even larger values  $rf$  (or even large negative values). Thus we applied **robust estimation**, namely used the **sample median** instead of the sample mean and the **normalized median absolute deviation (MADN)** instead of the standard deviation. Monte Carlo simulation combined with robust estimators (median, MADN) gives an uncertainty result close to the one based on the law of propagation of uncertainty. In addition, the Monte Carlo simulations yield strong deviations from a normal distribution (the law of propagation of uncertainty is silent about resulting distributions) and thus the probability for large deviations is quite high (in our example especially for large positive deviations).

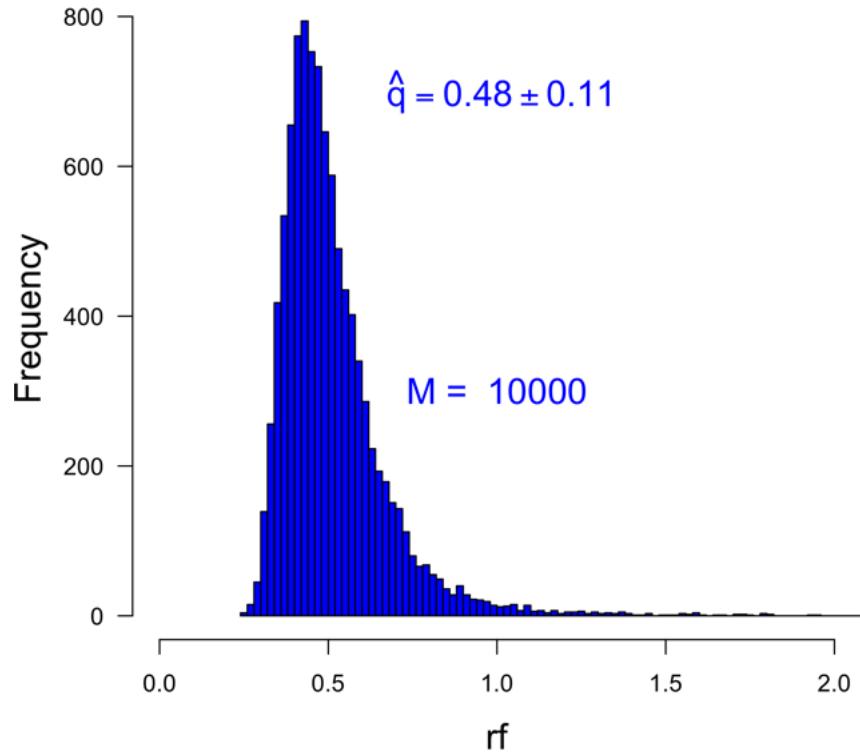


Figure 8.4: Histogram of  $rf = f(rx) = 1/rx$  where  $rx$  are  $M = 10^4$  random numbers from a normal distribution with mean  $\mu = 2.1$  and standard deviation  $\sigma = 0.5$ . Robust estimates for the distribution of  $q = f(x) = 1/x$  (median for the central tendency and normalized median absolute deviation (MADN) for the dispersion) yield  $\hat{q} = 0.48 \pm 0.11$ . [Uncertainty1overx.R](#)

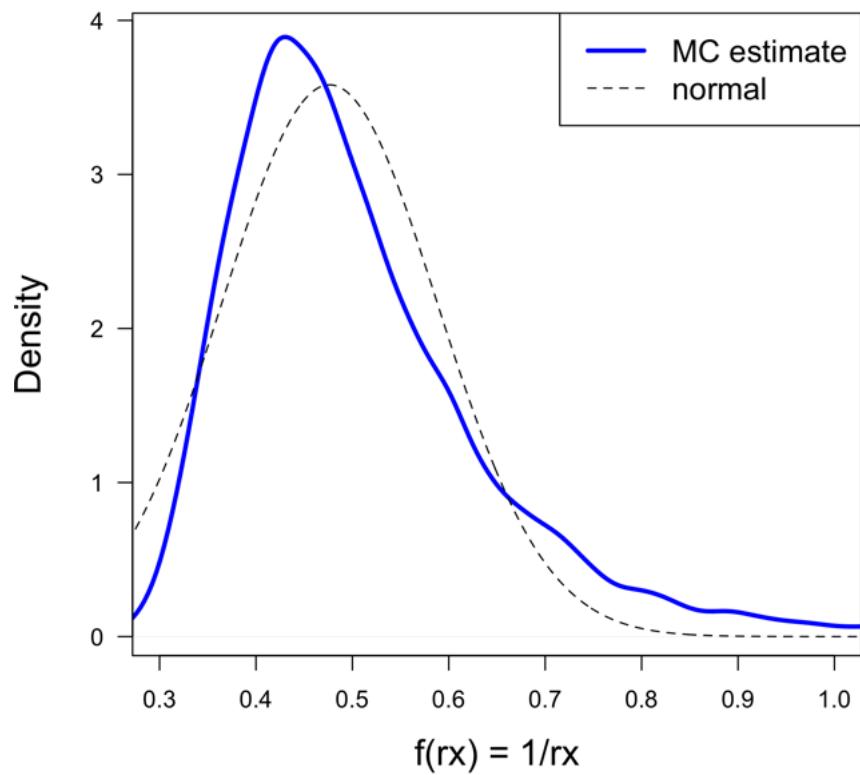


Figure 8.5: Estimate of the density (pdf) of  $q = f(x) = 1/x$  based on the results of the Monte Carlo simulation with  $M = 10^4$  runs (blue solid line). It deviates from a normal distribution with  $\mu = 0.476$  and  $\sigma = 0.111$  (black dashed line). [Uncertainty1overx.R](#)

The [third example](#) with  $f(x) = \tanh(3x - 6)$  and  $\hat{x} = 2.1 \pm 0.5$  is even 'more non-linear' and will give different estimates for both central tendency and uncertainty between results from the law of propagation of uncertainty ( $\hat{q} = 0.291 \pm 1.373$ ) and Monte Carlo simulations ( $\hat{q}_{\text{non-robust}} = 0.130 \pm 0.730$ ) even when applying robust estimators ( $\hat{q}_{\text{robust}} = 0.289 \pm 0.960$ ). The histogram of  $M = 10^4$   $rf$  values (Fig. 8.6) is obviously 'anti-normal' with maxima at  $\pm 1$ ; a maximum near the expected mean value of 0.291 is not recognizable and the uncertainty of  $\pm 1.373$  is larger than the possible range of  $\tanh(-1 \leq f(x) = \tanh(3x - 6) \leq +1)$ . The robust estimation from Monte Carlo simulation yields a central tendency of 0.289 that is close to the one from the law of propagation of uncertainty, however, the estimated uncertainty of  $\pm 0.960$  would also imply  $f(x)$  values outside the allowed range. The estimate of the density (Fig. 8.7) is far away from normal.

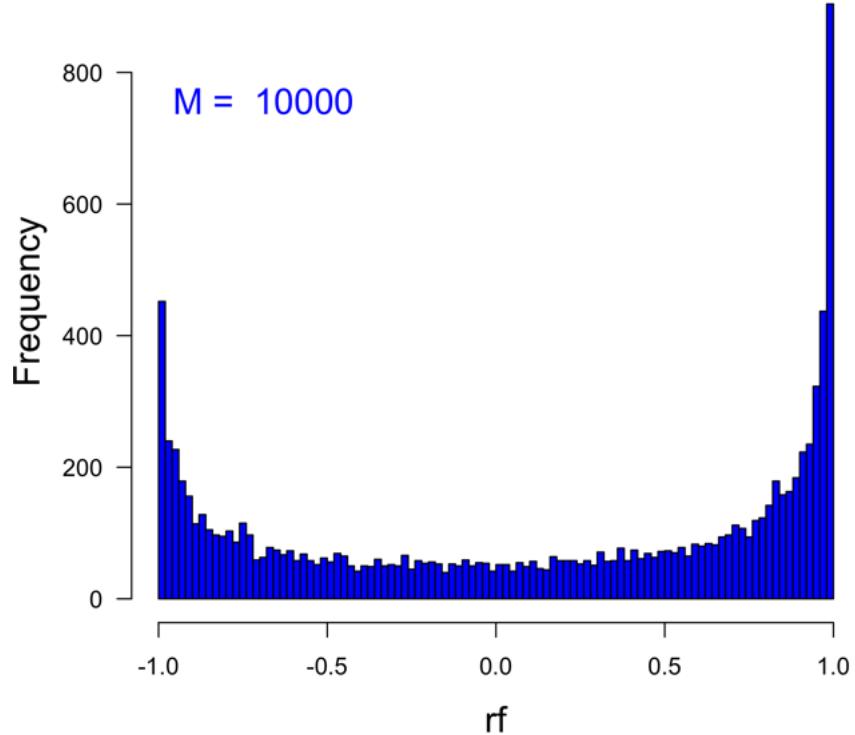


Figure 8.6: Histogram of  $rf = f(rx) = \tanh(3 \cdot rx - 6)$  where  $rx$  are  $M = 10^4$  random numbers from a normal distribution with mean  $\mu = 2.1$  and standard deviation  $\sigma = 0.5$ . Robust estimates for the distribution of  $q = f(x) = \tanh(3x - 6)$  (median for the central tendency and normalized median absolute deviation (MADN) for the dispersion) yield  $\hat{q} = 0.289 \pm 0.960$ . [UncertaintyMCtanh.R](#)

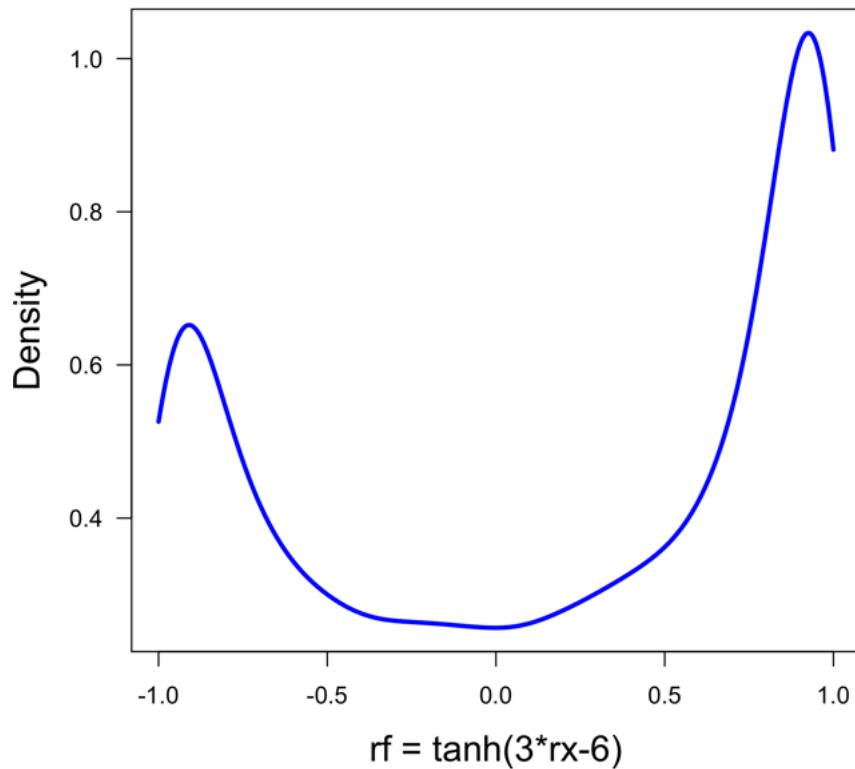


Figure 8.7: Estimate of the density (pdf) of random numbers from the normal PDF with  $\mu = 2.1$  and  $\sigma = 0.5$  propagated via  $q = f(x) = \tanh(3x - 6)$  based on the results of Monte Carlo simulations with  $M = 10^4$  runs. [UncertaintyMCtanh.R](#)

**Summary** Propagation of uncertainty can be calculated via a weighted sum of variances where the weights are given by squares of partial derivatives. An alternative is Monte Carlo simulation which has advantages for highly non-linear functions or complicated relations between measured and derived quantities<sup>5</sup>; it is easy to code and can give additional insights ('non-normal distributions of propagated uncertainties').

---

<sup>5</sup>An example from marine chemistry is the calculation of the carbonate ion concentration from measured values of dissolved inorganic carbon and total alkalinity: this requires finding the roots of a polynomial of fifth order (Zeebe & Wolf-Gladrow, 2001, p. 277).

**Exercise 25**  $f(x) = \tanh(3 \cdot x - 6)$

Plot  $f(x) = \tanh(3 \cdot x - 6)$  over the range  $0 \leq x \leq 4$  and try to explain why  $\hat{x} = 2.1 \pm 0.5$  leads to a very large uncertainty in  $f(x)$ .

**Exercise 26 Propagation of uncertainty for  $q = f(x, y)$**

Calculate the propagation of uncertainty using (a) the law of propagation of uncertainty and (b) Monte Carlo simulations combined with robust estimates of the central tendency and the dispersion for the following functions

(1)  $f(x, y) = x \cdot y$

(2)  $f(x, y) = x/y$

(3)  $f(x, y) = \sin x e^{-y}$

for  $\hat{x} = 2.1 \pm 0.5$  and  $\hat{y} = 1.5 \pm 0.3$ .



# Chapter 9

## Likelihood and likelihood function

The likelihood is a probability distribution (PD, discrete) or probability density function (PDF, continuous) providing the probability for observing data  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  given certain statistical populations<sup>1</sup> that are characterized by a set of parameters  $\boldsymbol{\theta}$ . From the likelihood  $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta})$  one obtains the likelihood function  $L(\boldsymbol{\theta}; \mathbf{x})$  by a switch of perspective: one now considers the observations  $\mathbf{x}$  as given (known) and asks for the probability of the population parameters. Even if the likelihood is discrete, its corresponding likelihood function is continuous when  $\boldsymbol{\theta}$  is varying continuously (which is usually the case; compare, for example, Section 9.2). In contrast to the likelihood, the likelihood function is not normalized to 1 when integrating over  $\boldsymbol{\theta}$ , i.e. it is not a PDF or 'density'. Please note that although powerful methods such as Maximum Likelihood Estimation (MLE), Likelihood Ratio Tests (LRTs) and full Bayesian analysis are based on likelihood functions, the switch of perspective is not accepted by all statisticians.

As a first example of a likelihood let us consider a normally distributed population characterized by the parameters mean  $\mu$  and standard deviation  $\sigma$ , i.e.  $\boldsymbol{\theta} = \{\mu, \sigma\}$ . The likelihood to observe the data point  $x_1$  is given by

$$\mathcal{L}(x_1; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_1 - \mu)^2}{2\sigma^2}} \quad (9.1)$$

where the quantities behind the semicolon in the argument list of  $\mathcal{L}()$  are considered as known parameters<sup>2</sup>.

The likelihood function results from a switch of perspective. One uses the same expression as given on the right-hand-side of Eq. 9.1, however, considers the observation  $x_1$  as given and the parameters  $\boldsymbol{\theta} = \{\mu, \sigma\}$  as variables (Casella & Berger, 2002, p. 290):

$$L(\mu, \sigma | x_1) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_1 - \mu)^2}{2\sigma^2}} \quad (9.2)$$

Please note the change in notation ( $L$  instead of  $\mathcal{L}$ ) and the interchange of arguments: the likelihood function  $L$  is a function of the variables  $\boldsymbol{\theta} = \{\mu, \sigma\}$  and the observation  $x_1$  is considered as a (known) parameter ('switch of perspective'). Please note that the likelihood function is usually not normalized, i.e. the integral over the allowed range of values  $\boldsymbol{\theta} = \{\mu, \sigma\}$  is not equal to one and thus  $L(\mu, \sigma | x_1)$  is not a PDF or 'density': although the right-hand-side of Eq. 9.2 looks like a normal PDF, the interpretation is different!

This switch of perspective forms the basis for a bunch of extremely successful methods. Given a set of data one can find an optimal model by maximizing the likelihood function by varying the model parameters. This approach is called Maximum Likelihood Estimation (MLE). If prior information about the model parameters is available one would maximize the posterior which is essentially (except for normalization) the product of the prior and the likelihood function; this is the full Bayesian approach. Likelihood functions can be used also for hypothesis testing leading to Likelihood Ratio Tests (LRTs).

<sup>1</sup>The statistical populations are usually related to mathematical models as, for example, a straight line, but also include randomness/noise; compare Section 9.1

<sup>2</sup>One might use the notation  $\mathcal{L}(x_1 | \mu, \sigma)$  resembling to the one for conditional probabilities where the quantities right of the vertical bar are considered as true or known for sure.

**Exercise 27 Likelihood function not normalized (\*)**

(1) Show by numerical integration over a large range of  $\mu$  and  $\sigma$  values that the likelihood function Eq. 9.2 for  $x_1 = 0.8$  is not normalized.

(2) How can it be normalized when the allowed standard deviation  $\sigma$  is restricted to an upper limit  $\sigma_{\max} < \infty$ ?

**Exercise 28 Likelihood of the Poisson distribution**

Show that the likelihood of the Poisson distribution is given by

$$\mathcal{L}_{\text{Poisson}}(x|\lambda) = \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \quad (9.3)$$

where  $x = \{x_1, x_2, \dots, x_n\}$  and  $s = \sum_{i=1}^n x_i$ .

**Exercise 29 Likelihood of the zero inflated Poisson distribution**

Show that the likelihood of the zero inflated Poisson (ZIP) distribution is given by

$$\mathcal{L}_{\text{ZIP}}(x|\lambda, p) = \frac{[p + (1-p)e^{-\lambda}]^k (1-p)^{n-k} e^{-(n-k)\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \quad (9.4)$$

where  $x = \{x_1, x_2, \dots, x_n\}$ ,  $k = \sum_{i=1}^n I(x_i = 0)$  (number of zeros in sample),  $I$  is the indicator or characteristic function with  $I(x_i = 0) = 1$  if  $x_i = 0$ ,  $I(x_i = 0) = 0$  if  $x_i \neq 0$ ,  $s = \sum_{i=1}^n x_i$ , and  $0 \leq p \leq 1$  is the zero-inflation parameter.

Hint: Assume that the data have been sorted into two groups: (a) all  $x_i = 0$  and (b) all  $x_i > 0$ .

## 9.1 Likelihood function for simple linear regression

Simple linear regression (Chapter 14) is based on four prerequisites (additive normal noise, homogeneous noise level, independence, 'fixed X') that we will use to construct the likelihood function. Maximizing the likelihood function leads to least squares and finally yields to optimal values of intercept and slope.

In simple linear regression ('regression to the mean') one tries to fit a straight line to a given set of data. Even for an ideal linear relationship between predictor and response, not all data will lie on the line because of noise. Thus the observed response,  $Y$ , is related to the predictor,  $X$ , by

$$Y = \beta_0 + \beta X + \text{normal noise} \quad (9.5)$$

where  $\beta_0$  is the true intercept,  $\beta$  is the true slope, and the noise follows a normal PDF with true mean  $\mu_{\text{noise}} = 0$  and true variance  $\sigma^2$ , i.e. one assumes that the noise level does not vary with  $X$ . The data are assumed to be independent from each other, i.e. observation of a response value  $y_i$  has no influence on the observation of  $y_j$  ( $j \neq i$ ). And finally one considers the predictor  $X$  as a non-stochastic variable<sup>3</sup>, i.e. one does not need to define a statistical population for  $X$ ; in statistical slang this is called 'fixed-X'.

The normal distribution

$$\mathcal{N}(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu-y)^2}{2\sigma^2}}$$

can be rewritten for the difference  $r$  ('residual') between the observation  $y_1$  and the corresponding (at the same

<sup>3</sup>This is at least approximately true when the uncertainties in  $X$  can be neglected. In ecology, time or location can often be determined precisely, whereas number of individuals or species of animal in a certain area can vary quite a bit.

predictor value  $x_1$ ) point  $\mu = \beta_0 + \beta x_1$  on the line:

$$\mathcal{L} \left( y_1; \mu = \underbrace{\beta_0 + \beta x_1}_{\text{on the line}}, \sigma \right) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\left( \underbrace{\beta_0 + \beta x_1}_{\text{prediction}} - \underbrace{y_1}_{\text{obs.}} \right)^2}{2\sigma^2}} \quad (9.6)$$

This is the likelihood,  $\mathcal{L}()$ , for observing  $y_1$  given a linear relationship plus additive normal noise (Eq. 9.5). The likelihood for observing  $y_2$  looks the same except that  $x_1 \rightarrow x_2$  and  $y_1 \rightarrow y_2$ :

$$\mathcal{L} \left( y_2; \mu = \underbrace{\beta_0 + \beta x_2}_{\text{on the line}}, \sigma \right) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\left( \underbrace{\beta_0 + \beta x_2}_{\text{prediction}} - \underbrace{y_2}_{\text{obs.}} \right)^2}{2\sigma^2}} \quad (9.7)$$

Please note that the homogeneous noise level prerequisite is taking into account by using the same standard deviation,  $\sigma$ , as for  $\mathcal{L}(y_1; \dots)$ .

The likelihood to observe  $y_1$  and  $y_2$  can be calculated by applying the product rule of productivity (Section 4.1.3) which takes an especially simple form (Eq. 4.9), namely the product of  $\mathcal{L}(y_1; \dots)$  and  $\mathcal{L}(y_2; \dots)$  (Eqs. 9.6 and 9.7), when the data are independent of each other (independence prerequisite):

$$\mathcal{L}(y_1, y_2; \beta_0, \beta, \sigma) = \mathcal{L} \left( y_1; \mu = \underbrace{\beta_0 + \beta x_1}_{\text{on the line}}, \sigma \right) \mathcal{L} \left( y_2; \mu = \underbrace{\beta_0 + \beta x_2}_{\text{on the line}}, \sigma \right) \quad (9.8)$$

$$= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^2 e^{-\frac{\sum_{m=1}^2 (\beta_0 + \beta x_m - y_m)^2}{2\sigma^2}} \quad (9.9)$$

using

$$c e^a c e^b = c^2 e^{a+b} \quad (9.10)$$

('multiplication of exponential functions = addition of exponents'). The extension to  $n$  data is obvious, yielding

$$\mathcal{L}(y_1, \dots, y_n; \beta_0, \beta, \sigma) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{\sum_{m=1}^n (\beta_0 + \beta x_m - y_m)^2}{2\sigma^2}} \quad (9.11)$$

Now one switches perspective by considering the data  $y_1, \dots, y_n$  as given and the parameters  $\beta_0, \beta, \sigma$  as variables:

$$L(\beta_0, \beta, \sigma; y_1, \dots, y_n) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{\sum_{m=1}^n (\beta_0 + \beta x_m - y_m)^2}{2\sigma^2}} \quad (9.12)$$

is the likelihood function.  $L()$  is maximal when the exponential term

$$\sum_{m=1}^n (\beta_0 + \beta x_m - y_m)^2 \quad (9.13)$$

is minimal: this is least squares!

## 9.2 Likelihood function for the Poisson distribution

The next example for a likelihood function is based on a discrete distribution, namely the Poisson distribution. The single model parameter  $\lambda$  (mean rate) can vary continuously between 0 and  $\infty$  and thus the likelihood function is continuous.

For a single observation  $k_1$  the likelihood function reads

$$L(\lambda | k_1) = \frac{\lambda^{k_1}}{k_1!} e^{-\lambda} \quad (9.14)$$

The likelihood function for any other observation  $k_m$  looks the same except that  $k_1$  is replaced by  $k_m$ . If we assume independence of the data, we can apply the simplified product rule of probabilities and thus obtain for the likelihood function given two observations  $k_1, k_2$ :

$$L(\lambda | k_1, k_2) = \frac{\lambda^{k_1}}{k_1!} e^{-\lambda} \frac{\lambda^{k_2}}{k_2!} e^{-\lambda} = \frac{\lambda^{k_1+k_2}}{k_1! k_2!} e^{-2\lambda} \quad (9.15)$$

This likelihood function is not normalized to 1 (compare Fig. 9.1; numerical integration of  $L(\lambda | k_1 = 3, k_2 = 2)$  over  $d\lambda$  from zero to infinity yields a value of 0.1562).

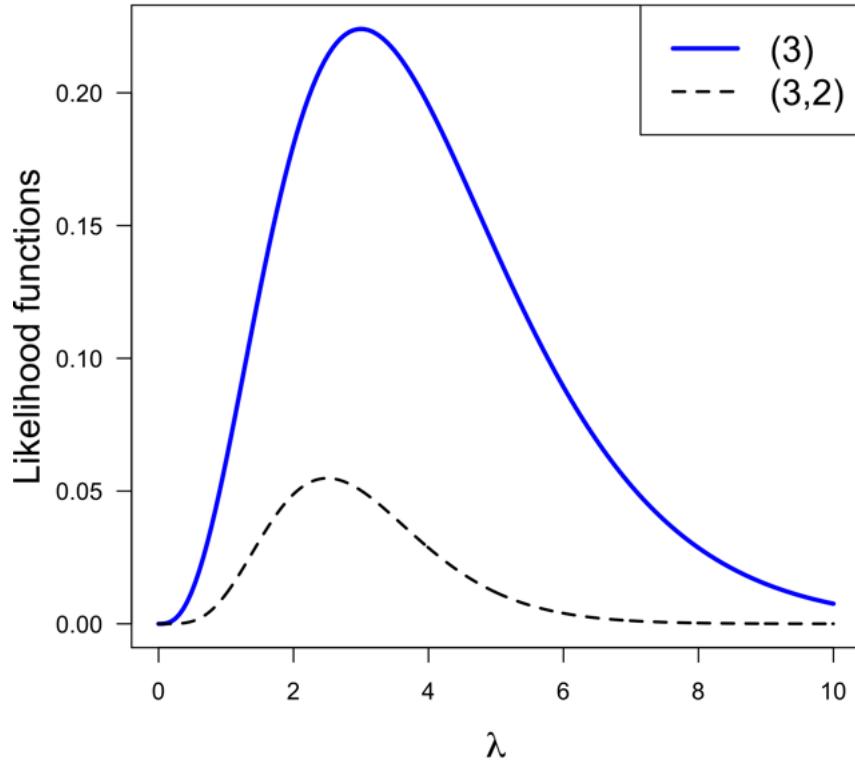


Figure 9.1: The likelihood functions  $L(\lambda | k_1)$  and  $L(\lambda | k_1, k_2)$  for  $k_1 = 3, k_2 = 2$ : even if both  $L(\lambda | k_j), j = 1, 2$  are normalized to one, their product  $L(\lambda | k_1, k_2)$  is not. With more and more data, the maximum value of the (non-normalized!) likelihood functions decreases further and further. [LikelihoodFctPoisson.R](#)

For a sample with  $n$  independent data  $\mathbf{k} = \{k_1, k_2, \dots, k_n\}$  one finally obtains

$$L(\lambda | \mathbf{k}) = \prod_{j=1}^n \frac{\lambda^{k_j}}{k_j!} e^{-\lambda} = e^{-n\lambda} \prod_{j=1}^n \frac{\lambda^{k_j}}{k_j!} \quad (9.16)$$

The likelihood function (Eq. 9.16) possesses a single maximum.

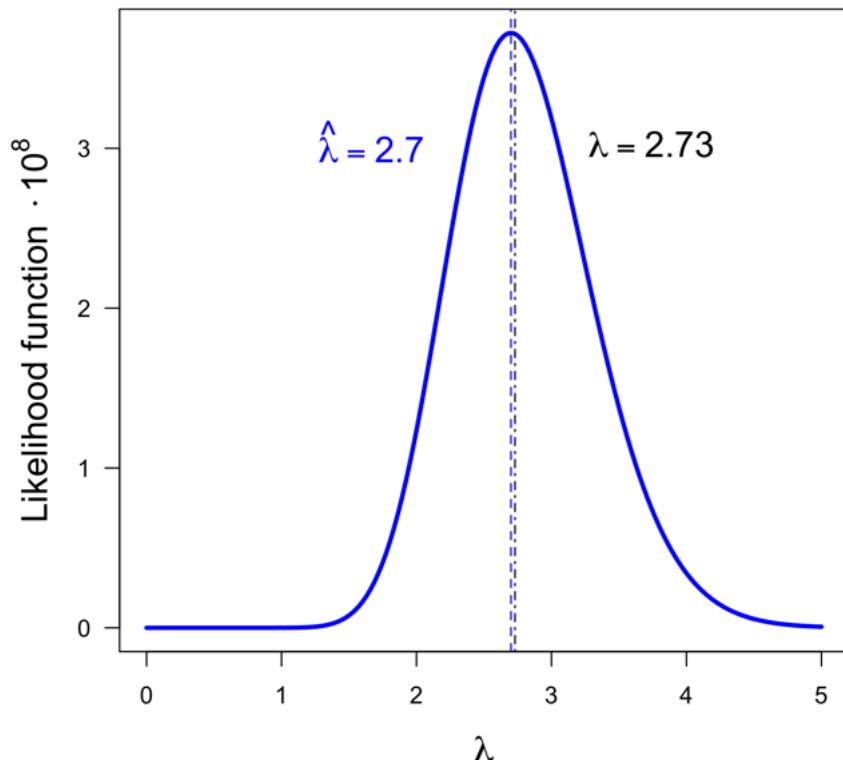


Figure 9.2: Likelihood function (Eq. 9.16, blue solid line) for the Poisson distribution given a sample with  $n = 10$  independent data  $\mathbf{k} = \{3, 2, 1, 2, 3, 1, 3, 3, 3, 6\}$ . The maximum of the likelihood function is located at  $\hat{\lambda} = 2.70$  (dashed blue vertical line); this value is called the maximum likelihood estimate. It is already close to the true  $\lambda = 2.73$  (indicated by the dash-dotted black vertical line) used to generate the data.

[LikelihoodFctPoisson10n.R](#)

**Exercise 30 Normalization of one data point Poisson likelihood***Proof that*

$$L(\lambda | k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

*is normalized to 1.*

# Chapter 10

## Point estimators

"In everyday life, 'estimation' means a rough and imprecise procedure leading to a rough and imprecise result. You 'estimate' when you cannot measure exactly. In statistics, on the other hand, '**estimation**' is a technical term. It means a precise and accurate procedure, leading to a result which may be imprecise but where at least the extent of the imprecision is known."

Roger Barlow (1989, p. 68)

"The following definition of a point estimator may seem unnecessarily vague. . . ."

A point estimator is any function  $W(X_1, \dots, X_n)$  of a sample; . . ."

Casella & Berger (2002)

Statistical populations are usually very large, often containing infinitely<sup>1</sup> many already possible outcomes and a single outcome can be encountered repeatedly, and thus there is usually no chance to measure the whole population. However, despite fundamental or practical (time & money) limitations we would like to know, at least approximately, what are the properties – for example, mean, variance, probability distribution or probability density function – of the population. In this chapter we will discuss point estimators for the mean and the variance of a population and explain why to divide usually by  $n - 1$ , however, sometimes by  $n$ , when estimating the variance. The challenge is to find ways how to squeeze out as much as possible information from a given sample or to estimate the necessary sample size to derive as much information as needed (sampling design).

**Music:** Although 'The Hat' from the Quicksilver Messenger Service LP 'Live at the Winterland Ball Room, December 1973' might look appropriate from the title, I prefer 'Gold and Silver' from the LP 'Quicksilver Messenger Service' (1968).

The basic problem is illustrated in Fig. 10.1: estimators usually yield estimates that are spread around the true value and, often, even the (arithmetic) mean of many estimates is not close to the true value (such estimators are called **biased**). The goal is to find estimators with small or vanishing bias and small dispersion. The dispersion of an estimator  $W$  can be measured by its mean squared error (MSE), see below.

**Definition of bias of estimator and unbiased estimator:** "The **bias** of a point estimator  $W$  of a parameter  $\theta$  is the difference between the expected value of  $W$  and  $\theta$ ; that is  $\text{Bias}_\theta W = E_\theta W - \theta$ . An estimator whose bias is identically (in  $\theta$ ) equal to 0 is called **unbiased** and satisfies  $E_\theta W = \theta$  for all  $\theta$ ." (Casella & Berger, p. 330)

We have encountered the Bayesian approach to point estimation already in the introductory chapter (Section 1.4); more examples will be discussed in Chapter 11. However, here we will investigate various of the commonly used estimators for the central tendency ('mean') and the dispersion ('spread') with the help of Monte Carlo simulations.<sup>2</sup>

---

<sup>1</sup>Or even more than infinitely many members: mathematicians talk about 'uncountable infinite'.

<sup>2</sup>Monte Carlo simulation is our all-purpose method.

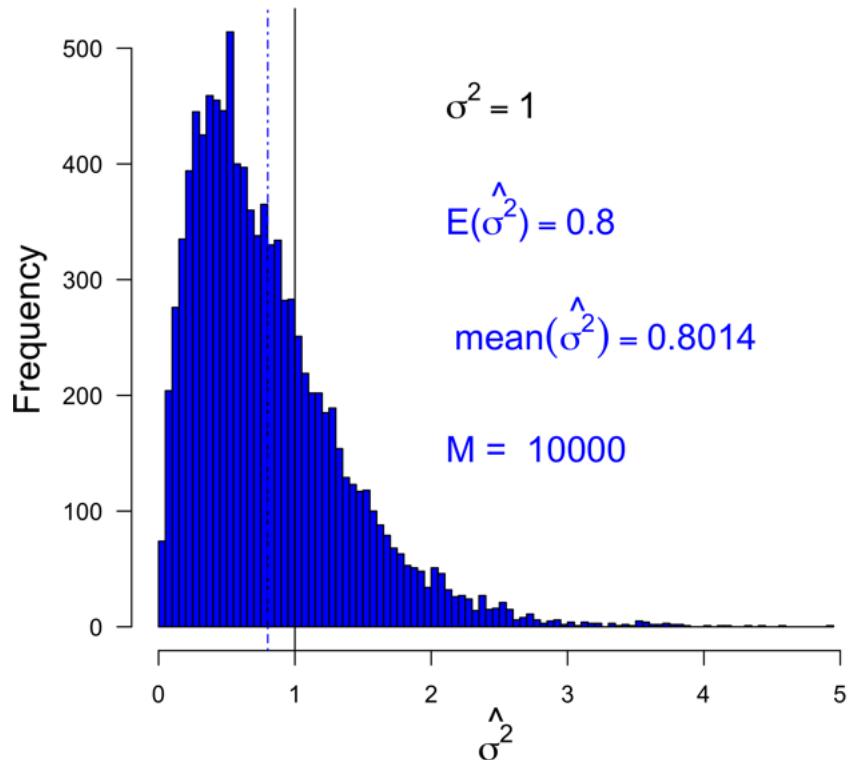


Figure 10.1: Estimating the variance,  $\sigma^2$ , of the standard normal PDF using the maximum likelihood estimator (MLE)  $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n (x_i - \bar{x})^2$ . The histogram shows results of a Monte Carlo simulation with  $M = 10^4$  runs for sample size  $n = 5$ . The MLE estimator is **biased**, i.e. its expectation  $E[\hat{\sigma}^2] = 0.8$  (see text; here estimated by the mean over  $M$  runs = 0.8014, blue dash-dotted vertical line) is different from the true variance  $\sigma^2 = 1$  (black vertical line). Note the large dispersion of the estimates. **The goal is to find estimators with small or vanishing bias and small dispersion.** Division by  $n - 1$  instead of by  $n$  yields an unbiased estimator. [PointEstMLEnormalBiasVar.R](#)

The task of finding 'good'<sup>3</sup> estimators can be split up into two questions (Casella & Berger, 2002): How to derive estimators? How to evaluate estimators?

Estimators can be found by (1) intuition, (2) the method of moments, (3) the maximum likelihood approach, (4) the Bayesian approach, or by (5) the expectation-maximization (EM) algorithm. If an estimator has been found or constructed, it can be evaluated with respect to (1) bias (to be explained below), (2) mean squared error (MSE), (3) sufficiency, (4) performance (efficiency), and (5) optimality.

Here we essentially use (1) intuition (or results from rigorous mathematical investigations) to derive estimators for the mean and variance of populations, and we will investigate their properties by Monte Carlo simulations. Examples for finding estimators by (2) the method of moments or (3) the maximum likelihood approach are given in the appendix (Section F.3).

In addition to randomness, outliers that may be caused by errors in measurements, recording etc., can be a problem for obtaining reliable estimates. For estimating the mean rate  $\lambda$  there are actually infinitely many unbiased estimators available. We will use the best (in the sense of least dispersion) unbiased estimator. In one of the examples, the apparent 'outlier', a value that does not fit at all to a Poisson distribution, turns out not to be an error, actually it is the most important observation. Thus we should be careful to discard outlier without a closer look. Finally, we will discuss the median and the normalized median absolute difference from the median (MADN) as robust estimators for the mean (central tendency) and the standard deviation of statistical populations.

**Further reading (point estimation):** Lehmann & Cassela (2006)

---

<sup>3</sup>'Good' in the sense of small/vanishing bias, small dispersion.

## 10.1 A list of point estimators and their properties (for speed-readers)

Given the data set  $x = \{x_1, \dots, x_n\}$ .

- The sample mean

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (10.1)$$

is an unbiased estimator of the true mean  $\mu$ ; the variance of  $\bar{x}$  falls off like  $1/n$ .<sup>4</sup> It is the commonly used non-robust estimator for the mean.<sup>5</sup>

- The sample mean,  $\bar{x}$ , is a **best unbiased estimator** for the mean rate  $\lambda$  of the Poisson distribution<sup>6</sup>.
- The estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (10.2)$$

is the best unbiased estimator for the variance,  $\sigma^2$ , of normal distributions (Casella & Berger, 2003, p. 341). Unfortunately, it is applicable only when the true mean,  $\mu$ , is known (textbook examples!).

- The sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (10.3)$$

is an unbiased estimator of the true variance,  $\sigma^2$ .<sup>7</sup> It is the commonly used non-robust estimator for variances of statistical populations<sup>8</sup>.

- The median (Subsection 3.1.2) is a robust estimator of the mean  $\mu$ .
- The Normalized Median Absolute deviation about the Median (MADN)

$$\text{MADN} = \frac{\text{median} |x - \text{median}(x)|}{0.6745} \quad (10.4)$$

is a robust estimator for the standard deviation  $\sigma$ .<sup>9</sup>

Estimation of other<sup>10</sup> population parameters requires special methods such as maximum likelihood estimation (MLE, Section F.3.2) or Bayes estimates.

---

<sup>4</sup>This is a remarkable result as it is independent of the statistical population, i.e. it applies not only for normal PDFs. The expectation for the variance of  $\mu$  is  $\sigma^2/n$ . Proof: Casella & Berger, 2003, Theorem 5.2.6, p. 213-214.

<sup>5</sup>For robust estimator of the mean compare the median, see below

<sup>6</sup>Casella & Berger (2003, p. 339). There are actually infinitely many unbiased estimator for the mean rate, however, some possess a large dispersion (Exercise)

<sup>7</sup>This is a remarkable result as it is independent of the statistical population, i.e. it applies not only for normal PDFs. Proof: Casella & Berger, 2003, Theorem 5.2.6, p. 213-214. However, even for normal PDFs we do not know whether a better – in the sense of smaller dispersion of the estimates – estimator exists (Casella & Berger, 2003, p. 342). The expectation for the variance of  $S^2$  of normal PDFs is  $2\sigma^4/(n-1)$  (Casella & Berger, 2003, p. 331).

<sup>8</sup>For robust estimator of the variance one may use the square of MADN, see below

<sup>9</sup>Strictly speaking, this is only true for normal PDFs and large sample sizes,  $n$ . For small sample sizes and/or different statistical distributions the ‘magic factor’ 0.6745 has to be modified (compare Section 10.8.1 for detailed discussion).

<sup>10</sup>... other than central tendency or dispersion

## 10.2 How good are estimates of the mean $\mu$ ?

### 10.2.1 Estimate mean of standard normal distribution; standard error of the mean

Estimation of the mean of a normal distribution with true mean  $\mu = 0$  and true standard deviation  $\sigma = 1$  from samples of size  $n = 5$  yields an average estimate of the mean  $\hat{\mu}_a = -0.027 \pm 0.431$  based on 1000 Monte Carlo runs. The little hat,  $\hat{\cdot}$ , on top of  $\mu$  indicates that the mean value has been estimated. This result is consistent with the expectation<sup>11</sup>  $\hat{\mu}_e = \mu \pm \sigma/\sqrt{n} = 0 \pm 0.447$  where  $\sigma/\sqrt{n}$  is called the **standard error of the mean (SE)**. The standard error is the uncertainty of (the estimate of) the mean.

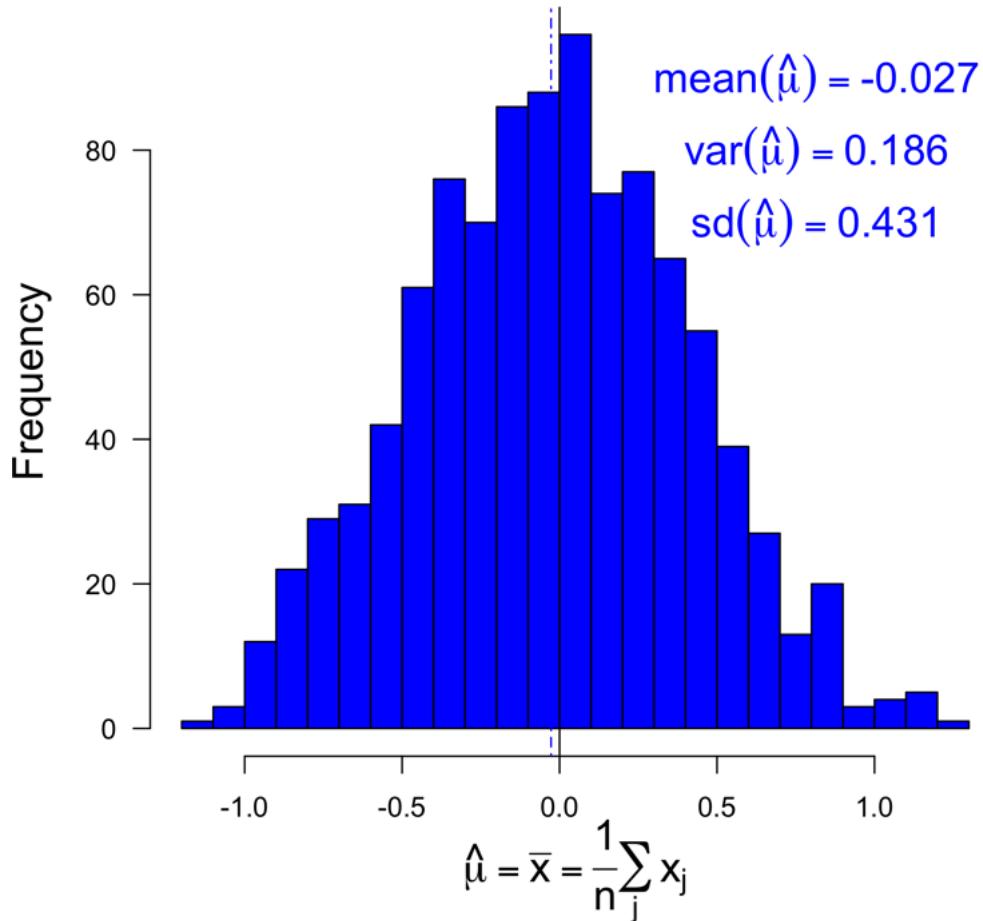


Figure 10.2: Estimate the mean  $\mu$  of the standard normal distribution using the sample mean: histogram of sample means  $\hat{\mu} = \bar{x}$  from  $M = 1000$  Monte Carlo runs and sample size  $n = 5$ . The average estimate is  $\text{mean}(\hat{\mu}) = -0.027$  (blue vertical dash-dotted line) and thus close to the true value  $\mu = 0$  (black vertical line). The variance of sample mean values is 0.186 (standard deviation 0.431) and thus close to the expected value of  $\sigma^2/n = 1/5 = 0.2$  (compare Theorem 5.2.6 in Casella & Berger, 2002). [PointEstMeanNormal.R](#)

<sup>11</sup>Expectation is based on Theorem 5.2.6 in Casella & Berger (2002); compare Section F.1.

## 6: de Moivre's equation

Wainer (2007) calls the relation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (10.5)$$

de Moivre's equation (de Moivre, 1730). It states that the standard deviation of the mean (usually called the standard error of the mean or just the standard error),  $\sigma_{\bar{x}}$ , is given by the standard deviation of the population from which we sample,  $\sigma$ , divided by the square root of the sample size,  $\sqrt{n}$ . Wainer (2007) consider's de Moivre's equation as one of the most 'dangerous' equations in the sense that, if you don't know it, you can easily come to wrong conclusions.

Two famous examples are the incidents of kidney cancer and the variation of scores in math exams with school size. When looking at incidents of kidney cancer in the 3141 counties in the US, one recognizes that 'the counties in which the incident' (measured in cases per 100 000 inhabitants) 'is lowest are mostly rural, sparsely populated, and in traditionally Republican states of the Midwest, the South, and the West' (Kahneman, 2011). An explanation immediately comes to mind: 'this outcome is directly due to the clean living of the rural life-style – no air pollution, no water pollution, access to fresh food without additives and so on' (Wainer, 2007). However, one can also recognize that the counties in which the incident is highest are mostly rural, sparsely populated, and in traditionally Republican states of the Midwest, the South, and the West, often lying next to the counties with lowest incidence. And again an explanation immediately comes to mind: 'this outcome might be directly due to the poverty of the rural lifestyle – no access to good medical care, a high-fat diet, and too much alcohol and tobacco' (Wainer, 2007). These two explanations contradict each other and thus one should look for a better alternative. Here, de Moivre's equation comes in: counties with smaller number of inhabitants posses larger standard deviations and thus are overrepresented in the set of counties with low as well as with high incidence!

The second example concerns the variation of scores in math exams with school size. It has been recognized that small schools are over-represented when it comes to candidates with high scores. Based on this fact, it has been concluded that smaller school size would lead to higher scores. In the 1990s various foundations began supporting small schools or even to split up large schools into several small ones (more than 1 billion US\$ has been spent within a single decade). A detailed analysis (Wainer & Zwerling, 2006; Wainer, 2007) has shown that the success of some of the small schools can be explained by the work of the dangerous de Moivre equation. Linear regression analysis showed that size has no or a slight positive effect on scores.

### 10.2.2 Abuse of the standard error or what's the size of the Emperor of China?

The expression

$$\frac{\sigma}{\sqrt{n}} \quad (10.6)$$

may suggest that one can reduce the uncertainty of the mean to an arbitrary small value if one goes to larger and larger sample size  $n$ . However, this is not consistent with common sense.

A convincing example is given by Jaynes<sup>12</sup>: 'The classical example showing the error of uncritically reasoning here is the old fable about the height of the Emperor of China. Supposing that each person living in China surely knows the height of the Emperor to an accuracy of at least  $\pm 1$  m [ $1.5 \pm 1$  m is a quite good estimate!]; if there are  $N = 1\,000\,000\,000$  inhabitants, then it seems that we could determine his height to an accuracy at least as good as

$$\frac{1}{\sqrt{1\,000\,000\,000}} \text{ m} \approx 3 \cdot 10^{-5} \text{ m} = 0.03 \text{ mm} \quad (10.7)$$

merely by asking each person's opinion and averaging the result.' (Jaynes, 2003, p.258)

This is an absurd result!

What's the solution? The short answer is: There is always a systematic error and the uncertainty is given by the systematic error  $\pm$  the standard error:

$$\text{uncertainty} = \text{systematic error} \pm \text{standard error}. \quad (10.8)$$

For very large sample sizes the standard error approaches zero and thus the uncertainty is given by the systematic error alone. How to obtain or estimate a value for the systematic error? Unfortunately, no general answer is possible because it depends on the respective systems, methods, and measuring devices under consideration. Intercomparisons using different methods and devices might give a clue. Text books on statistics can not give answers here.

<sup>12</sup>Imagine the time before invention of television, internet, smartphones etc. and China with todays population.

## 10.3 Estimate the variance of the standard normal distribution

Now we apply Monte Carlo simulations for estimating the variance of the standard normal distribution. On purpose we choose the small sample size  $n = 5$  in order to see large differences between different estimators that include factor  $1/(n - 1)$  versus  $1/n$ . In the first Monte Carlo simulation we use the R routine `var()` as estimator. The average estimate is  $\hat{\sigma}_1^2 = 1.041$  (Fig. 10.3) and thus close to the true value  $\sigma^2 = 1$ .

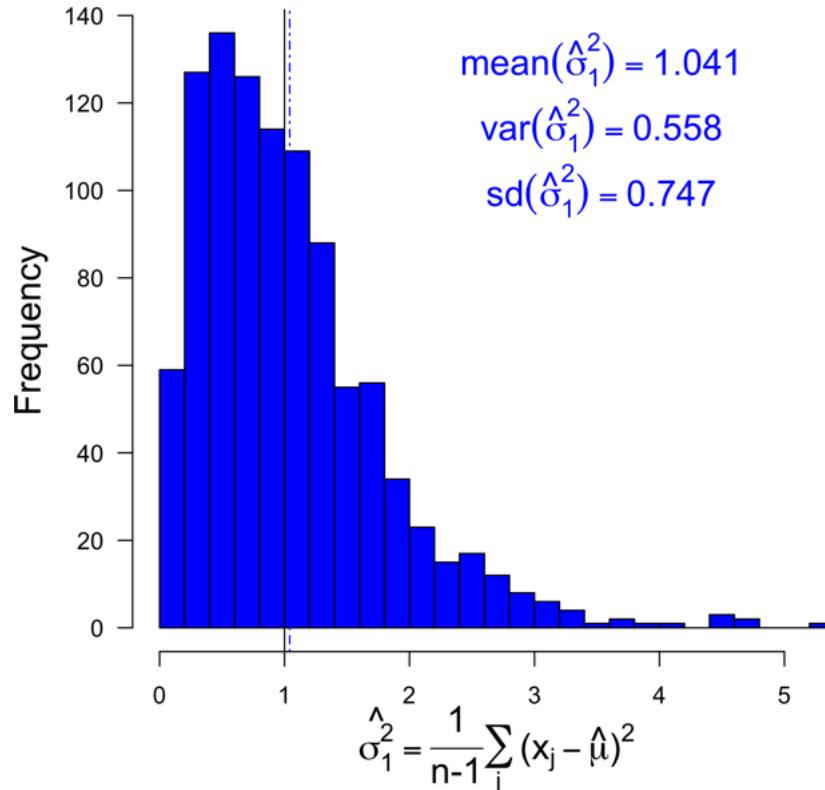


Figure 10.3: Estimate the variance of the standard normal distribution using the sample mean: histogram of sample variances from  $M = 1000$  Monte Carlo runs and sample size  $n = 5$ . The average estimate is  $\hat{\sigma}_1^2 = 1.041$  (blue vertical dash-dotted line) and thus close to the true value  $\sigma^2 = 1$  (black vertical line). The standard deviation of the estimates is 0.747 and thus the uncertainty of  $\hat{\sigma}_1^2$  is quite large (note that Theorem 5.2.6 in Casella & Berger, 2002, is silent on this value and thus we don't know what to expect; however, one may guess that this value varies with  $\sigma^2$ , compare Exercise 31). [PointEstVarNormal.R](#)

**Exercise 31 How does the variance of the estimate of  $\sigma^2$  scale for random samples from normal distributions?**

*Hint: Vary the Monte Carlo simulations for estimating the variance of the standard normal distribution by sampling from normal distributions with different standard deviations  $\sigma$  and try to find a simple scaling law.*

### 10.3.1 Four different estimators of the variance

In this section we will apply four estimators of the variance that differ by (a) dividing by  $(n - 1)$  or by  $n$  and (b) using the sample mean  $\bar{x}$  or the true mean  $\mu$  (Figs. 10.4 to Fig. 10.7).

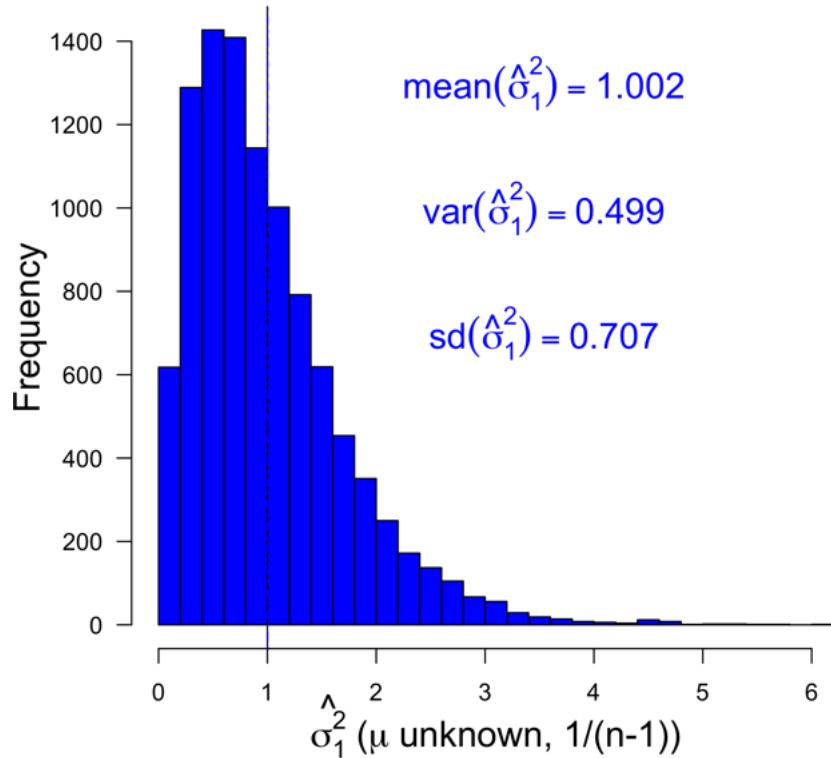


Figure 10.4: Estimator  $\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$ : histogram of 10<sup>4</sup> estimates of the variance of the standard normal distribution for a sample size  $n = 5$ . The average estimate  $1.002 \pm 0.707$  is very close to  $\sigma^2 = 1$  (the estimator is unbiased), however, the standard deviation is quite large (because of the small sample size). [PointEstVarNormalA.R](#)

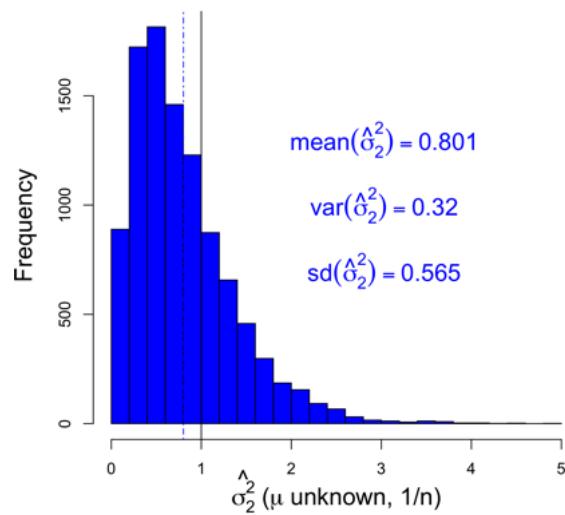


Figure 10.5: Estimator  $\hat{\sigma}_2^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$ : histogram of  $10^4$  estimates of the variance of the standard normal distribution for a sample size  $n = 5$ . The average estimate  $0.801 \pm 0.565$  is significantly lower than  $\sigma^2 = 1$  (the estimator is biased by  $1 : n = 1 : 5$  or 20%); this bias decreases with increasing sample size.

[PointEstVarNormalB.R](#)

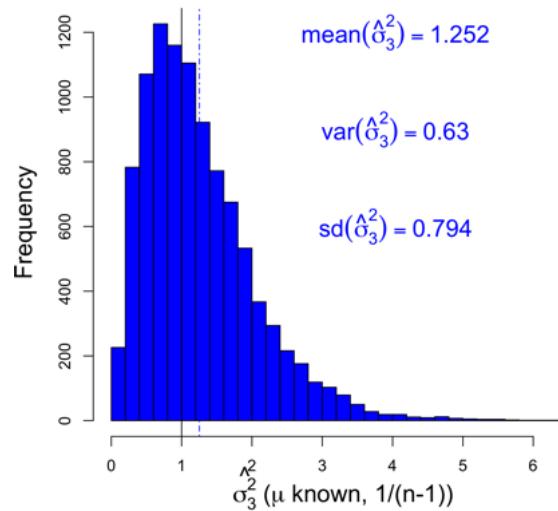


Figure 10.6: Estimator  $\hat{\sigma}_3^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \mu)^2$ : histogram of  $10^4$  estimates of the variance of the standard normal distribution for a sample size  $n = 5$ . The average estimate  $1.252 \pm 0.794$  is significantly higher than  $\sigma^2 = 1$  (the estimator is biased by  $1 : (n-1) = 1 : 4$  or 25%); this bias decreases with increasing sample size.

[PointEstVarNormalC.R](#)

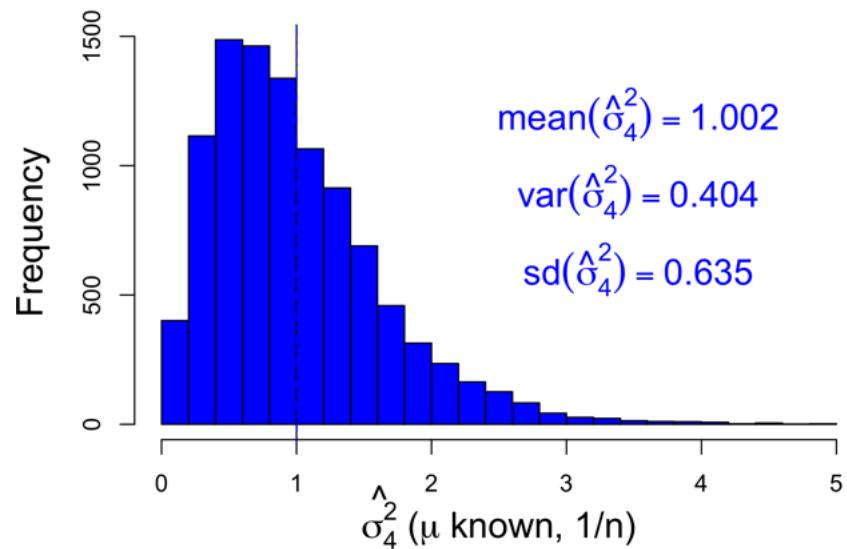


Figure 10.7: Estimator  $\hat{\sigma}_4^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2$ : histogram of  $10^4$  estimates of the variance of the standard normal distribution for a sample size  $n = 5$ . The average estimate  $1.002 \pm 0.635$  is close to  $\sigma^2 = 1$  (the estimator is unbiased). [PointEstVarNormalD.R](#)

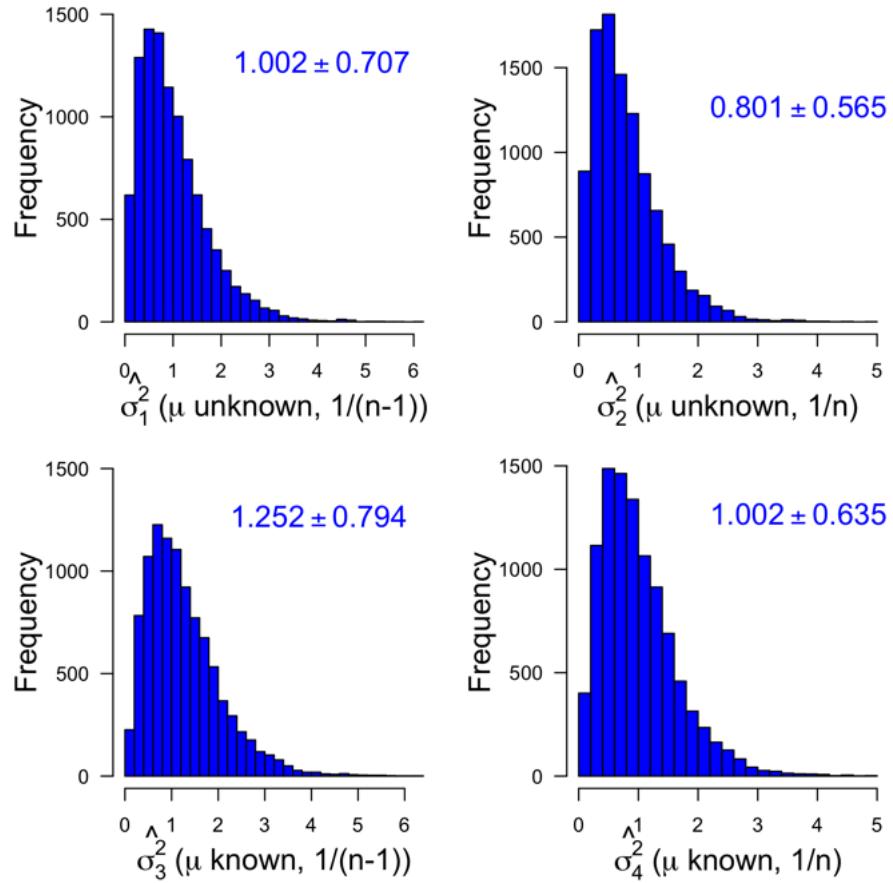


Figure 10.8: Four different estimators of the variance: histograms of  $10^4$  estimates of the variance of the standard normal distribution for a sample size  $n = 5$ . [PointEstVarNormal4.R](#)

The Monte Carlo simulations indicate that the estimator with sample mean  $\bar{x}$  and division by  $(n - 1)$  and the estimator with true population mean  $\mu$  and division by  $n$  are unbiased whereas the two other estimators are biased. Therefore it is recommended to use

$$\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (10.9)$$

and this estimator is implemented in the R routine `var()`<sup>13</sup>. Although

$$\hat{\sigma}_4^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2 \quad (10.10)$$

is also an unbiased estimator, it is rarely applied because the true mean  $\mu$  is usually not known (however, see Section O.1).

## 10.4 Estimators for covariance & correlation

The covariance is usually estimated by

$$\begin{aligned} \widehat{\text{cov}}(x, y) &= \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \\ &= \frac{1}{\nu} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \end{aligned} \quad (10.11)$$

where  $(x_k, y_k)$ ,  $k = 1, 2, \dots, n$  are paired data and  $\bar{x}$  and  $\bar{y}$  are the sample means of  $x$  and  $y$ , respectively. The interesting point here is the value  $\nu = n - 1$  for the degrees of freedom. Why isn't it  $2n - 2$  (number of data,  $2n$ , minus number of constraints, namely 2 mean values)? Or  $n - 2$  = number of products minus number of constraints? Text books usually give no justification.

The estimator (10.11) is at least convenient because it leads to a simple form of the estimator for the (Pearson) correlation  $r$

$$\begin{aligned} \hat{r} &= \frac{\widehat{\text{cov}}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y} \\ &= \frac{1}{n-1} \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}} \\ &= \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}} \end{aligned} \quad (10.12)$$

where all prefactors cancel out. It is the estimator used in the R routine `cov()`.<sup>14</sup>

In the following we will investigate the covariance estimator (10.11) by applying Monte Carlo simulations. The results indicate that the estimator is strongly biased at small sample sizes.

<sup>13</sup>This applies also to other software packages, however, sometimes one has to explicitly choose between division by  $n$  or  $n - 1$ .

<sup>14</sup>The same estimator is most probably used in other statistical software packages.

### 10.4.1 Monte Carlo simulations

The routine **sampleUcorFct0** (R code: [sampleUcorFct.R](#)) generates pairs of random data with correlation  $r$  close to a specified value. These data will be used to estimate the correlation and covariance with the estimators (10.11) and (10.12), respectively. The results as a function of the sample size  $n$  are shown in Figs. 10.9 & 10.10, respectively.

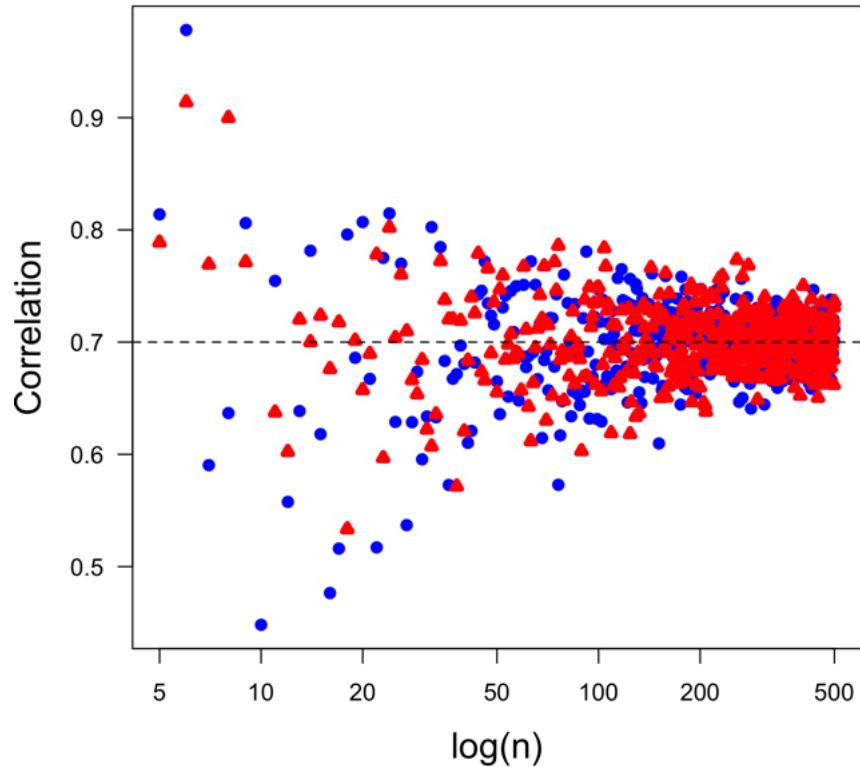


Figure 10.9: Correlations estimated for artificially generated data sets as a function of sample size  $n$  (blue and red symbols for replicate runs). The estimated correlations  $\hat{r}$  are spread around the desired (specified) correlation  $r = 0.7$ . As expected, the spread is larger at small sample sizes. [PointEstCorrelation.R](#)

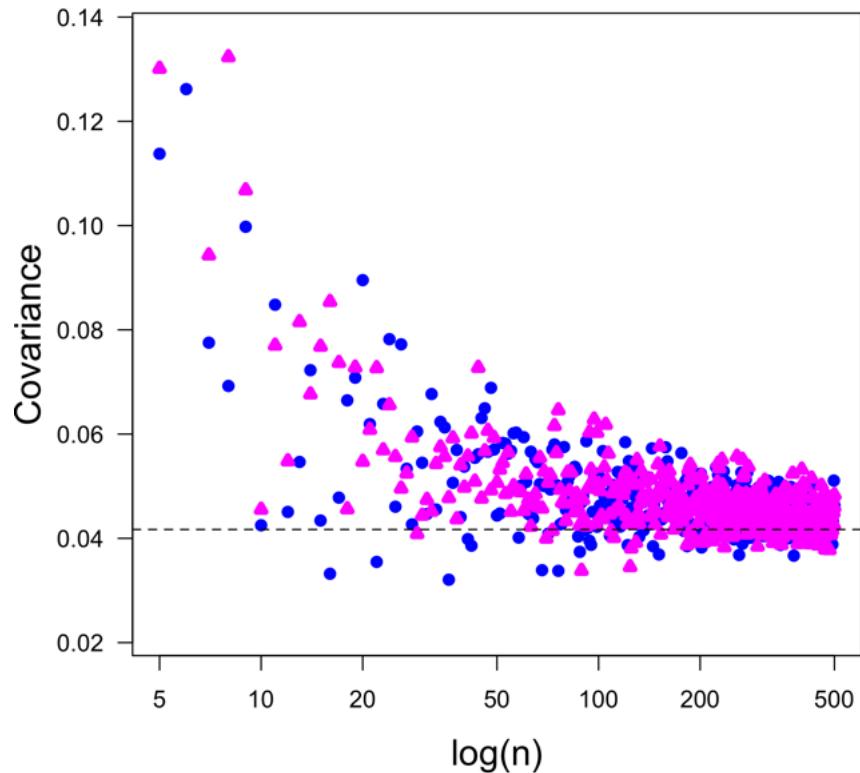


Figure 10.10: Covariances estimated for artificially generated data sets as a function of sample size  $n$  (blue and magenta symbols for replicate runs). The estimated covariances vary with  $n$  and, for small sample sizes, deviate largely towards high values compare to the value of 0.042 estimated for a very large sample size of  $n = 10^5$  (this value can be considered as close to the true value). At large sample sizes the degree of freedoms  $v = n - 1$  or  $v = n$  or  $v = n - 2$  would not make a large difference, whereas at small  $n$  large sample size-dependent corrections would be necessary to obtain estimates that are close to the true value. [PointEstCovariance.R](#)

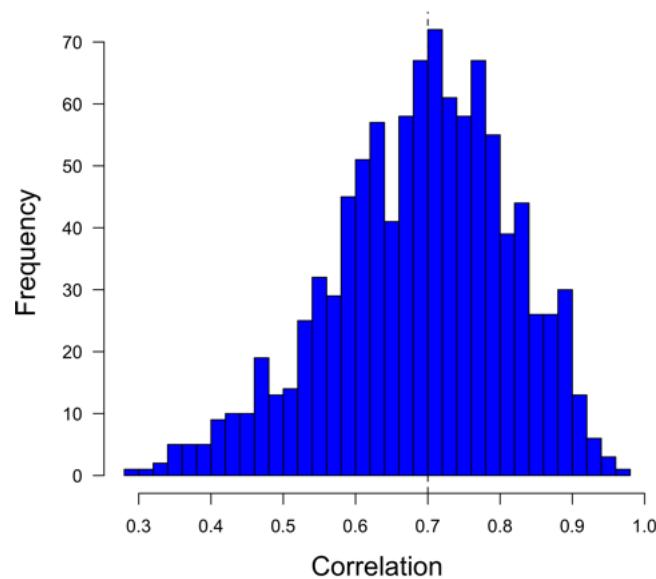


Figure 10.11: Correlations estimated for artificially generated data sets at a small sample size of  $n = 10$ . The distribution of estimates is spread around the desired (specified) correlation  $r = 0.7$ ; it is not symmetric with a long tail towards small correlation values. Exercise: How do the distributions look for  $r = 0$  and  $r = -0.7$ ? First guess, then run Monte Carlo simulations. [PointEstCorrelationSmallN.R](#)

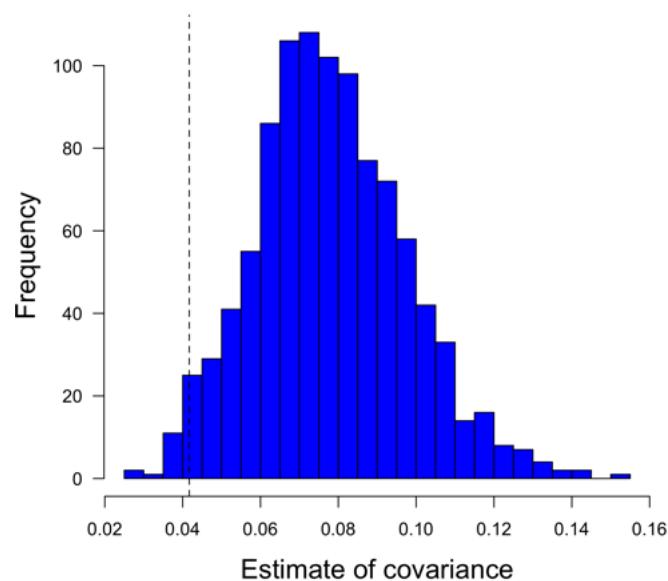


Figure 10.12: Covariances estimated for artificially generated data sets at a small sample size of  $n = 10$ . The distribution of estimates peaks at a value that is much larger (about a factor of 2) than the value of 0.042 (black broken vertical line) estimated for a very large sample size of  $n = 10^5$  (this value can be considered as close to the true value). [PointEstCovarianceSmallN.R](#)

**Summary** Based on our Monte Carlo simulation we can conclude<sup>15</sup> that the estimator for correlations (Eq. 10.12) works fine (Fig. 10.9) and seems to be unbiased (Fig. 10.11). On average the estimator for covariances (Eq. 10.11) shows large deviations (factor 2) from the true value (of which we only estimated an approximate value) at small sample sizes (Fig. 10.9). In order to improve this estimator one would have to derive a sample size-dependent correction factor; similar problems are known from other estimators, compare, for example, the estimator for the standard deviation (Section F.2) or for MADN (Exercise 35). The correction factor most probably also depends on the type of populations from which we sample: a wide field ...

### Exercise 32 Densities of correlated data

The routine `sampleUcorFct()` allows generation of correlated paired data sets  $(x_k, y_k)$ ,  $k = 1, 2, \dots, n$ . Generate a paired data set for large sample size ( $n = 2000$ ), desired (specified) correlation  $r$ , and plot histograms of the two samples  $x$  and  $y$ . Guess the densities of the underlying statistical populations.

---

<sup>15</sup>... however, without mathematical rigor given by a proof ...

## 10.5 Estimate rate parameter $\lambda$ of Poisson distribution

Both the sample mean  $\bar{X}$  and the sample variance  $S^2$  (division by  $n - 1$ ) are unbiased estimators of the rate parameter  $\lambda$  of the Poisson distribution. However, they differ in the variances of their estimates and one can show that

$$\text{Var}_\lambda \bar{X} \leq \text{Var}_\lambda S^2 \quad (10.13)$$

(Casella & Berger, 2002, p. 335). If these two estimators are the only ones available, one would use the sample mean because it possesses the smaller mean squared error (Figs. 10.13 – 10.14 show results of Monte Carlo simulations). However, the class of unbiased estimators for  $\lambda$  is much larger because one might use, for example, any linear combination of sample mean and variance

$$W_a(\bar{X}, S^2) = a \bar{X} + (1 - a) S^2 \quad (10.14)$$

where  $a$  is a constant.

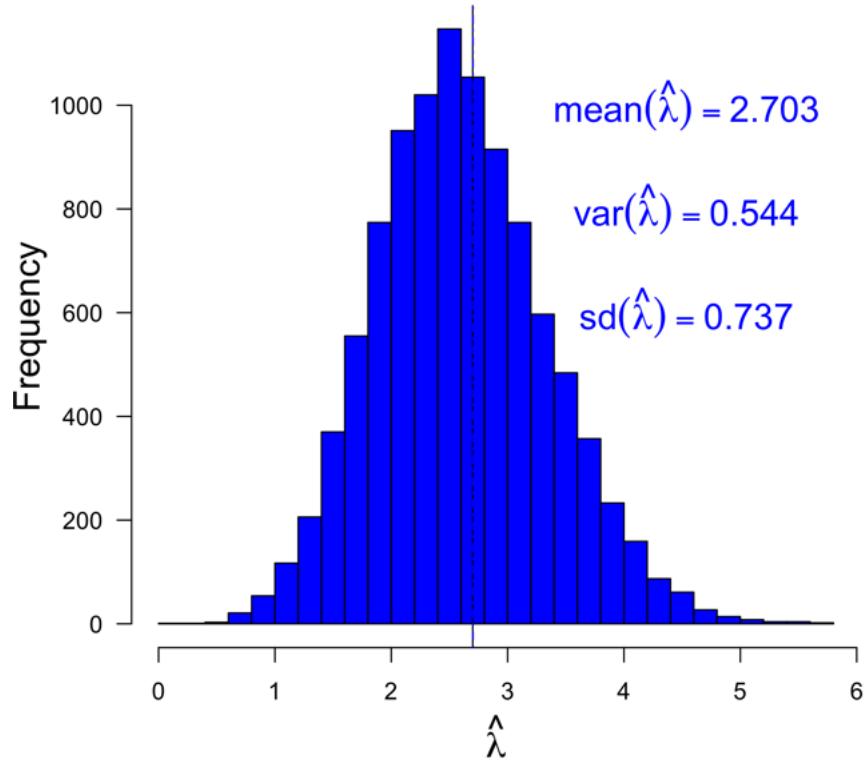


Figure 10.13: Estimates of the rate parameter  $\lambda$  of the Poisson distribution (mean rate  $\lambda = 2.7$ ) using the **sample mean as estimator**: histogram of sample means from  $M = 10^4$  Monte Carlo runs and sample size  $n = 5$ . The average estimate is  $\hat{\lambda}_a = 2.703$  and thus very close to the true value  $\lambda = 2.7$ . The variance of sample mean values is 0.544 and thus close to the expected value of  $\sigma^2/n = \lambda/5 = 0.54$  (compare Theorem 5.2.6 in Casella & Berger, 2002). The distribution of the estimates looks symmetric. [PointEstPoissonSampleMean.R](#)

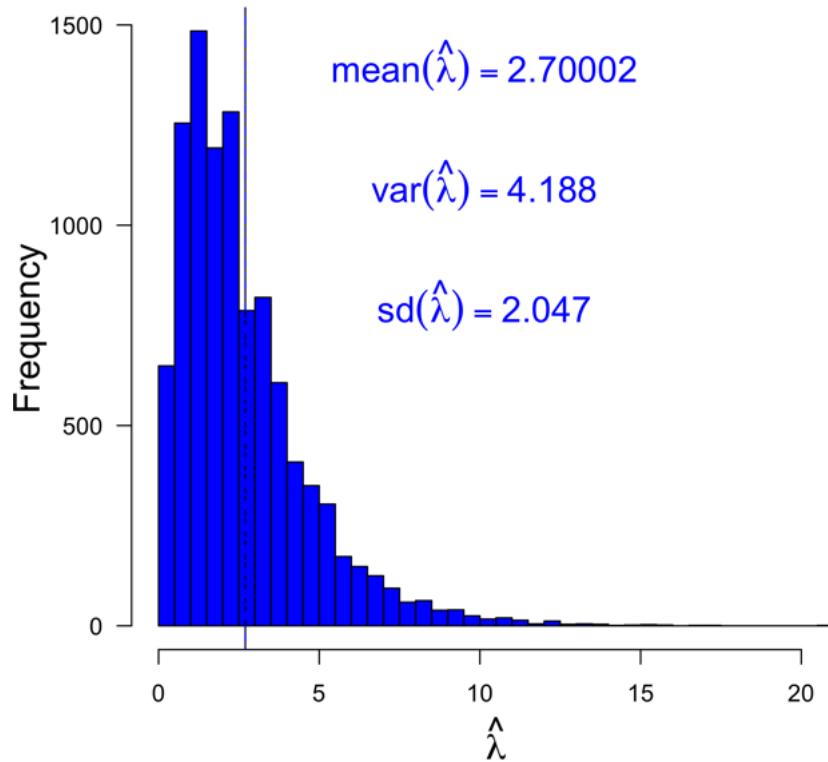


Figure 10.14: Estimates of the rate parameter  $\lambda$  of the Poisson distribution (mean rate  $\lambda = 2.7$ ) using the **sample variance as estimator**: histogram of sample variances from  $M = 10^4$  Monte Carlo runs and sample size  $n = 5$ . The average estimate is  $\hat{\lambda}_a = 2.7002$  and thus very close to the true value  $\lambda = 2.7$ . The variance of sample mean values is 4.19 and thus much larger than the one for estimating  $\lambda$  via the sample mean. The distribution of the estimates is skewed. The large variance of the estimates is a clear disadvantage of this estimator and thus it is recommended to not use it although it is an unbiased estimator. [PointEstPoissonSampleVar.R](#)

**Exercise 33 Variance of estimator  $W_a(\bar{X}, S^2)$** 

Use Monte Carlo simulations to estimate the variances of the unbiased estimators  $W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2$  for  $a$  in the range between 0 and 2 in steps of 0.01. For which  $a$  do you obtain the minimal variance? Is the result in accordance with your expectation?

**Exercise 34 Standard error of the mean: neutrinos**

In the Introduction (Chapter 1) we fitted a Poisson distribution to the neutrino data. The estimation of the mean rate ( $\lambda$ ) and its uncertainty was done using the Bayesian approach: assigning a prior, calculating the likelihood based on the Poisson distribution, switching from likelihood to likelihood function, and, finally, normalization to obtain the posterior which turned out to be a gamma distribution. The mean and standard deviation of this gamma distribution were calculated in Eq. (1.22) and used as estimates for central tendency and dispersion (uncertainty) of the mean rate ( $\lambda$ ) of the Poisson distribution.

Now we will follow a different (more simple) approach to estimate the uncertainty of  $\lambda$  from the data.

- (1) Estimate the standard error of the mean from the neutrino data (Table 1.1).
- (2) Compare the result with the uncertainty estimated in Eq. (1.22): Depending on the result: Why are the estimates so different or so close?

## 10.6 Estimate of success in single trial (binomial distribution)

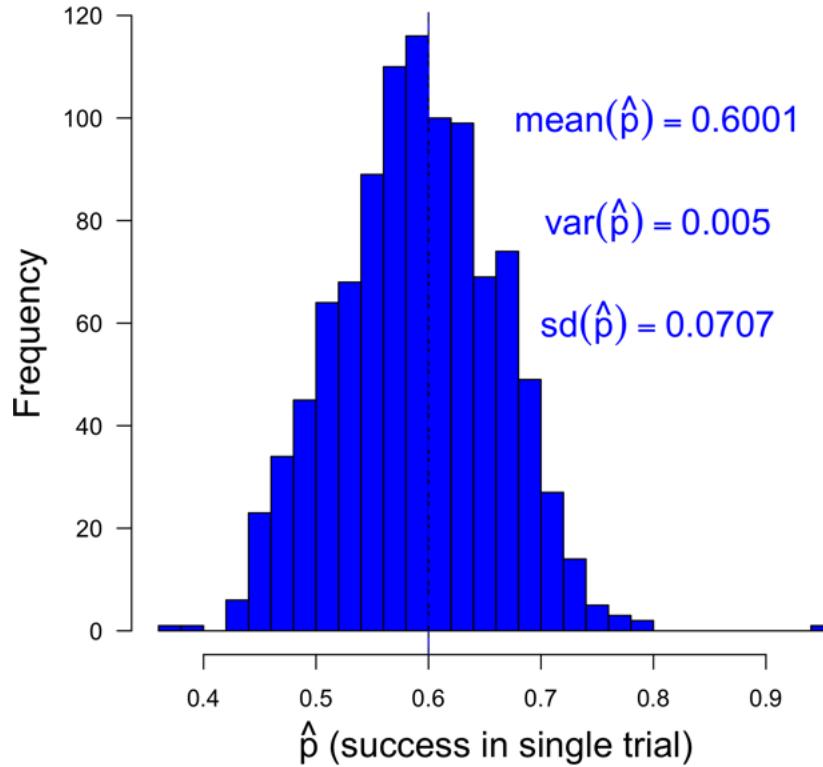


Figure 10.15: Estimate probability for success in a single trial,  $p$ , for random samples of size  $j = 5$  from the binomial distribution  $\text{Binomial}(k; n = 10, p = 0.6)$  (where  $k$  is the actual number of successes in a total number of trials  $n = 10$ ) using  $M = 1000$  Monte Carlo runs. We estimate the true mean  $\mu = n \cdot p = 6$  via the sample mean and obtain on average a mean of  $\hat{\mu} = 6.001$  which is close to the expected value. Estimates of  $p$  can be obtained by simply dividing  $\hat{\mu}$  by the total number of trials, i.e.  $\hat{p} = \hat{\mu}/n = 0.6001$  which is close to the expected value  $p = 0.6$ . The estimated variance of  $p$ ,  $\hat{\sigma}_p^2$ , is 0.0050. This is close to the true variance  $\sigma_p^2 = \frac{\sigma_b^2}{n^2 j} = \frac{n p (1 - p)}{n^2 j} = 0.0048$  where  $\sigma_b^2 = n p (1 - p)$  is the variance of the binomial distribution (Section A.1.1). [PointEstBinomSuccessProb.R](#)

## 10.7 Estimate mean of $F$ distribution (\*)

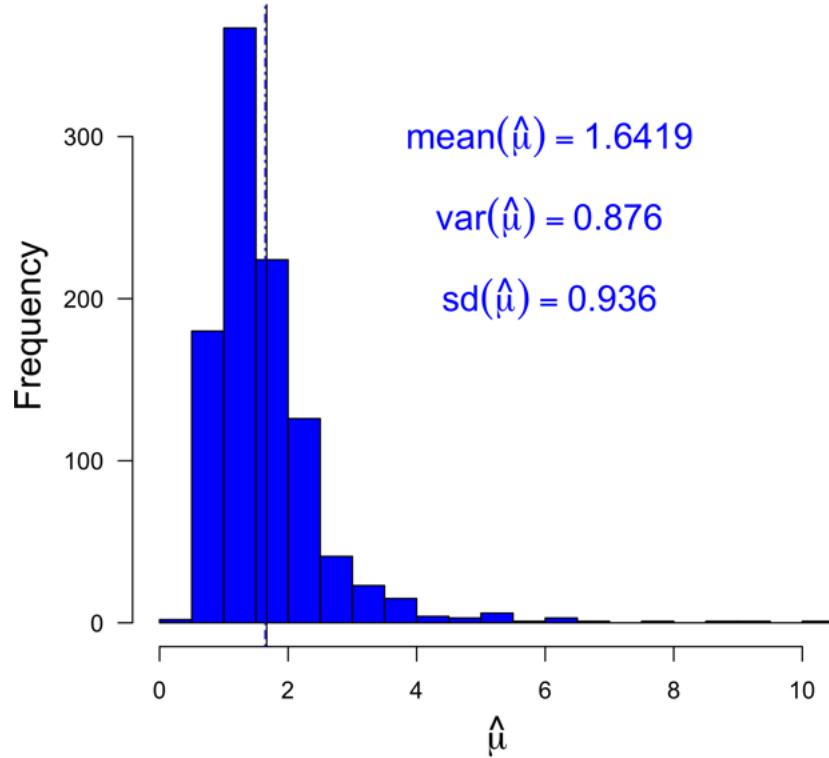


Figure 10.16: Estimation of the mean of the  $F$  distribution  $\mathcal{F}(x; \nu_1 = 15, \nu_2 = 5)$  from random samples of size  $n = 5$  using  $M = 1000$  Monte Carlo runs. We obtain  $\text{mean } \hat{\mu} = 1.6419$  which is very close to the true mean  $\mu = 5/3 \approx 1.6667$  and variance  $\hat{\sigma}_{\hat{\mu}}^2 = 0.876$  which is smaller (Do you have an idea why?) than the true variance  $\sigma_{\mu}^2 = 1.333$ . [PointEstMeanF.R](#)

## 10.8 Robust estimators for central tendency & dispersion

Although sample mean and sample variance are unbiased and efficient estimators of the central tendency ('mean') and dispersion of populations, there is still room for improvement. One reason for it is the stability of the mentioned estimators against outliers: a single outlier can have a major impact on the estimated values for central tendency or dispersion. An alternative are so-called robust estimators as, for example, the median, which can cope much better with outliers.

Given: an **ordered** set of data  $\{x_1, x_2, \dots, x_k, \dots, x_n\}$  with  $x_1 \leq x_2 \leq \dots \leq x_k \leq \dots \leq x_n$ . The median is defined as the value  $z$  that has an equal number of items on either side of it:  $z = x_{(n+1)/2}$  if  $n$  is odd and  $z = (x_{n/2} + x_{1+n/2}) / 2$  if  $n$  is even.

Examples:

$n$  odd:  $\{-1, 1, 7, 19, 22\} \Rightarrow z = 7$  is the median.

$n$  even:  $\{-1, 1, 7, 19, 22, 23\} \Rightarrow z = (x_3 + x_4)/2 = (7 + 19)/2 = 13$  is the median.

### 7: Median of PDFs

We have encountered already the median of a sample. The median of a PDF  $f(x; \dots)$  is defined as the location  $x_{\text{median}}$  where the probabilities left and right of  $x_{\text{median}}$  are equal to each other and thus both equal to  $1/2$ , i.e.

$$\int_{-\infty}^{x_{\text{median}}} f(x; \dots) dx = \int_{x_{\text{median}}}^{\infty} f(x; \dots) dx \quad (10.15)$$

which, in general, is different from the mean,  $\mu$ , defined as the expectation of  $x$ , i.e.

$$E[x] = \int_{-\infty}^{\infty} x f(x; \dots) dx = \mu. \quad (10.16)$$

For symmetric PDF the median is equal to the mean (Fig. 10.17).

Further reading (robust methods): Huber (1964, 1972), Huber & Ronchetti (2011), Maronna et al. (2006)

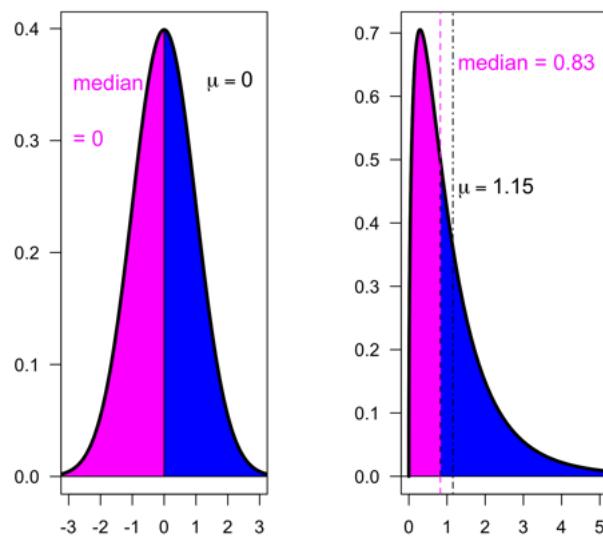


Figure 10.17: For symmetrical distributions – as for example the standard normal PDF (left panel) – the median is equal to the mean,  $\mu$ , whereas for asymmetric distributions – as for example an F-PDF (right panel) – median and mean are different from each other. Both median and mean can be used to characterize the central tendency of a distribution. [PointEstMedianVersusMean.R](#)

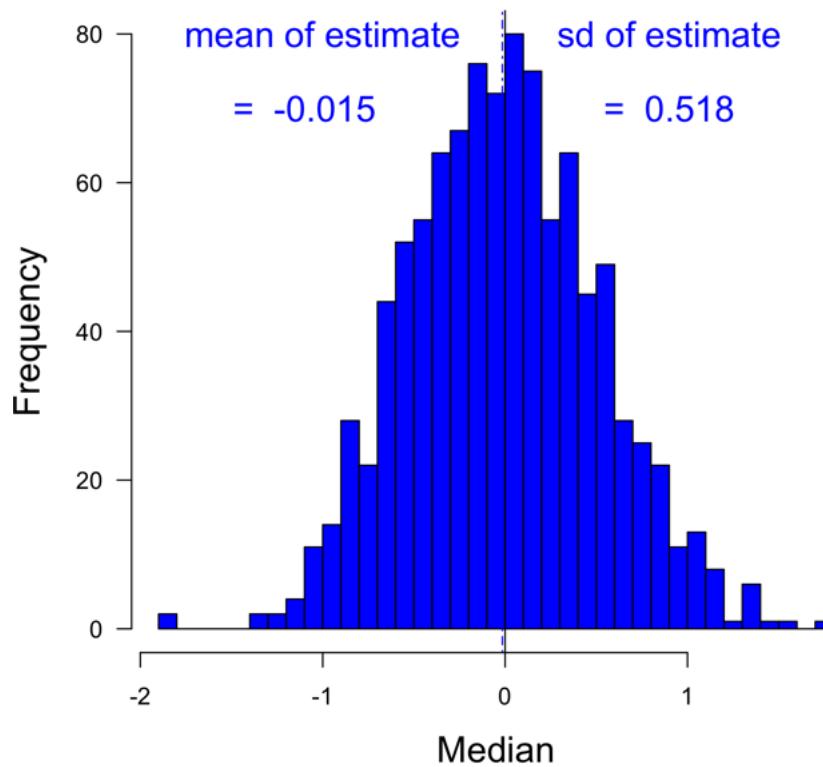


Figure 10.18: Robust estimation of the mean  $\mu$  of the standard normal distribution by the median. For a sample size  $n = 5$  one obtains on average ( $10^3$  Monte Carlo runs)  $\hat{\mu} = -0.015 \pm 0.518$ . For the sample mean as estimator we had obtained  $\hat{\mu} = -0.027 \pm 0.431$ . The slightly larger uncertainty for the estimate based on the median is the prize for the much higher robustness of the estimator. [PointEstRubustMedian.R](#)

### 10.8.1 Robust estimation of the dispersion: MAD & MADN

We will now try to estimate the dispersion of the data using an estimator that applies the median twice, namely, MAD which stands for 'Median Absolute deviation about the Median':

$$\text{MAD} = \text{median} |x - \text{median}(x)| \quad (10.17)$$

(Maronna et al., 2006, p. 5) where  $x = \{x_1, \dots, x_n\}$  is a random sample of size  $n$  from a statistical population. The **median** is used as a robust estimator of the mean the population and its value is subtracted from the sample. The dispersion is then estimated as the **median** of the absolute deviations between the data and the estimated mean of the population.

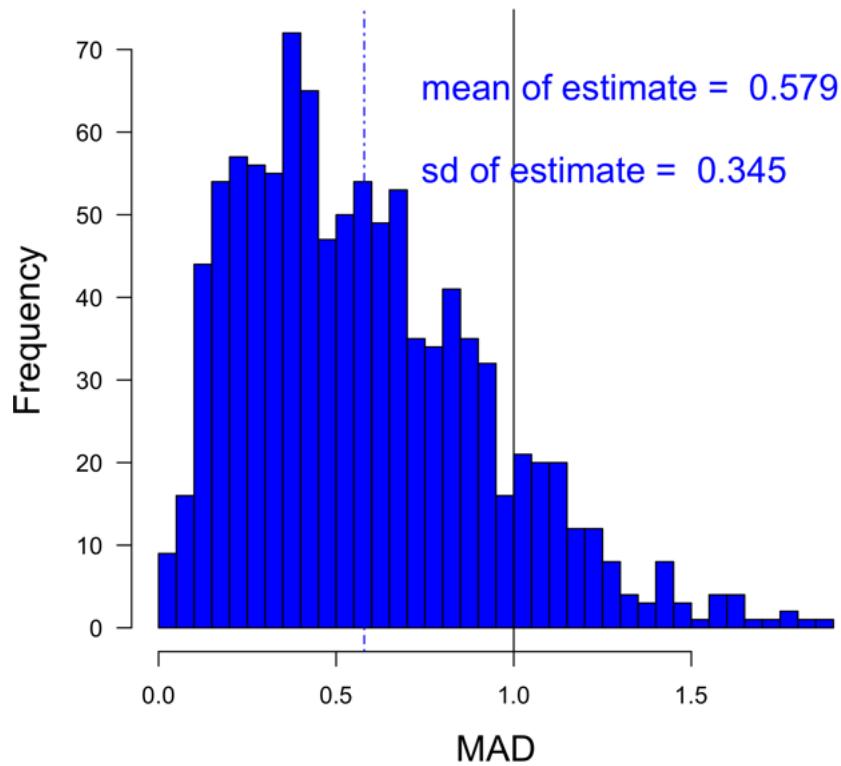


Figure 10.19: Estimate of the dispersion of a population using MAD (Median Absolute deviation about the Median): sample size  $n = 5$ ,  $M = 10^4$  random samples from the standard normal distribution. The average estimate  $0.579 \pm 0.345$  is significantly smaller (strongly biased) than the standard deviation  $\sigma = 1$  of the standard normal distribution. This bias can be corrected by scaling (see MADN, below; Fig. 10.20). [MADnormal1.R](#)

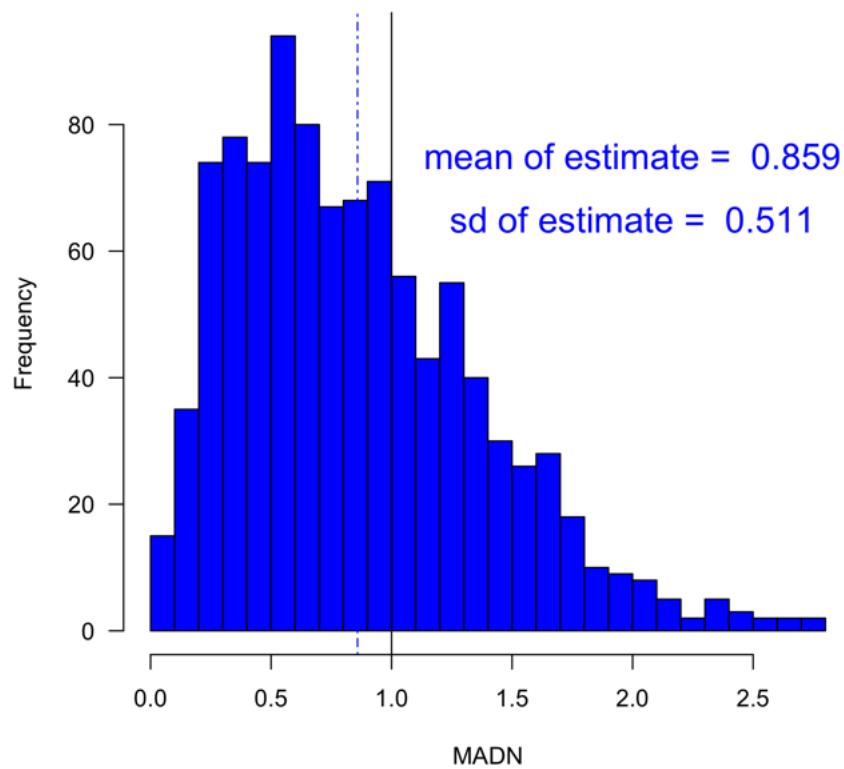


Figure 10.20: Estimate of the dispersion of a population using MADN (Median Absolute deviation about the Median, Normalized): sample size  $n = 5, 10^4$  random samples from the standard normal distribution. The average estimate of  $0.859 \pm 0.511$  is closer to 1 than the MAD estimate, however, still not fully satisfying. Still missing: correction for small sample size. [PointEstMADNnormal.R](#)

Although MAD provides a measure of the dispersion of the standard normal distribution, the average value of the estimate is significantly smaller than the standard deviation of the standard normal PDF ( $0.579 \pm 0.345$  vs.  $\sigma = 1$ ). This difference of measures ('bias') can be reduced by 'scaling' the MAD estimate, i.e. by multiplication of the MAD estimate by a factor. The normalized MAD, denoted by MADN, is defined by  $MAD/0.6745$

$$\text{MADN} := \frac{\text{MAD}}{0.6745} \quad (10.18)$$

(Maronna et al., 2006, p. 5; Wikipedia; the scaling factor 0.6745 is given by  $\sqrt{2}\operatorname{erf}^{-1}(1/2)$ ). However, this correction is obviously too small for small sample sizes: the MADN estimate of  $0.859 \pm 0.511$  (Fig. 10.20) is closer to 1 than the MAD estimate, however, still not fully satisfying. For small sample sizes the scaling factors can be estimated by Monte Carlo simulations (Exercise 35); the size of the scaling factors depends on the sample size  $n$  and on the statistical population from which we sample. The routine `myMADN()` is based on sampling from normal PDFs.

#### **Exercise 35 Scaling of MAD at small sample sizes**

*The scaling of MAD by the factor 1/0.6745 works for large sample sizes, however, is too small for small sample sizes. Use Monte Carlo simulations with sampling from the standard normal distribution to*

- (a) show that the scaling 1/0.6745 works for large sample sizes*
- (b) estimate (Monte Carlo) a size-dependent scaling factor for small sample sizes.*

## 10.9 Estimate uncertainty via bootstrapping

*How can we estimate the uncertainty of a statistic (as for a example the mean) when we have only one sample available (and obtaining more samples is too expensive or impossible)? Bradley Efron and others have developed resampling methods including 'jackknife' and 'bootstrapping' to address this and other problems (including hypothesis testing). In this section we will only give two very simple examples of bootstrapping in order to stimulate your interest in resampling methods.*

**Further reading:** Efron (1979), Efron & Tibshirani (1994); for resampling in the context of errors in variables compare Mudelsee (2023)

**Example 1:** We have taken a random sample  $r$  of size  $n = 30$  from a standard normal distribution with sample mean  $\bar{r} = -0.0013$ . What's the uncertainty (' $1 \sigma$  level') of this estimate of the true mean  $\mu = 0$ . The basic idea of bootstrapping is that the sample contains some information about the distribution from which we sampled and that instead of generating more samples from the original distribution we randomly (re)sample from our single sample in the following way:

1. Draw  $n$  (same as sample size!) random integers  $i$  from the uniform distribution between 1 and  $n$  where one samples with replacement, i.e. some integers can be drawn more than once whereas other are not drawn at all. This is the only random process involved in the bootstrapping procedure.
2. Sample  $n$  values from our single sample  $r$  whereby the integers  $i$  are the indices of our sample vector  $r$ . Because some integers can occur more than once the corresponding  $r$  values are selected several times whereas other  $r$  values are left out. This creates a resample  $r[i]$ .
3. Calculate the mean of the resample  $r[i]$ .
4. Repeating steps 1 to 3 many times ( $B = 1000$ ) yields many different resamples and corresponding mean values. The histogram of the mean values (Fig. 10.21) hints to a normal distribution. The standard deviation of the resamples yields 0.1798. This is an estimate of the ' $1 \sigma$ ' uncertainty of the sample mean:  $\bar{r} = -0.0013 \pm 0.1798$  (although the sample mean of -0.0013 is close to the true mean of zero, its uncertainty is quite large!).

Of course, for the standard normal distribution we know that the sample mean values follow the normal distribution with standard deviation  $1/\sqrt{n} \approx 0.1826$ ; the bootstrap uncertainty estimate of 0.1798 is close to the analytical value.

An alternative way to express uncertainty is by estimating a confidence interval. The ' $1 \sigma$  level' approximately corresponds to the 67% interval, i.e. 67% of the probability of the sample mean distribution lie between -0.1778 and +0.1778 (analytic result). Bootstrapping yields the estimates -0.1855 and 0.1676 as interval limits; these values are slightly asymmetric, however, close to the analytical values. The 95% interval (close to the ' $2 \sigma$  level') yields  $\pm 0.3578$  (analytic) and -0.3584 and +0.3255 (bootstrap estimates). Confidence intervals may be the preferential way to express uncertainty when distributions are asymmetric (see next example).

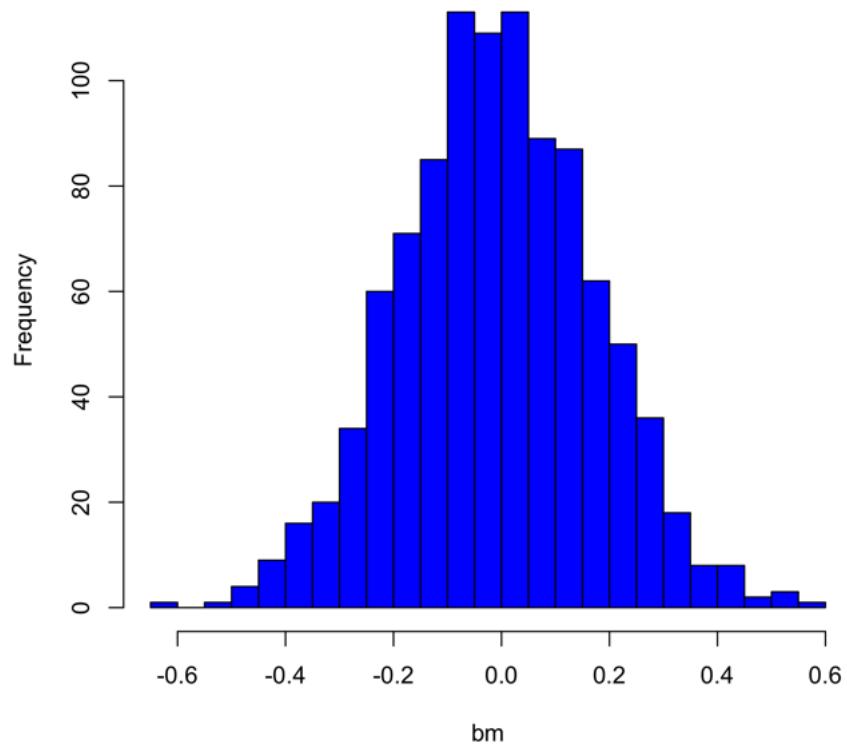


Figure 10.21: Histogram of  $B = 10^3$  bootstrap sample mean values with sample size  $n = 30$  for a sample from the standard normal distribution. [BootstrapMeanNormal.R](#)

**Example 2:** We have taken a random sample  $r$  of size  $n = 30$  from an F distribution with degrees of freedom  $\nu_1 = 15$  and  $\nu_2 = 3$ . Applying the same bootstrapping procedure as in example 1 except for a larger number of resamples  $B = 10^4$  yields an estimate of the 95% interval [1.0861, 2.094] which includes the mean values 1.5442 (bootstrap estimate) and  $\nu_2 / (\nu_2 - 1) = 1.5$  (analytic). I am not aware of an analytic expression for the distribution of sample means from F distributions and thus no comparison with an analytic expression of the 95% interval is possible.

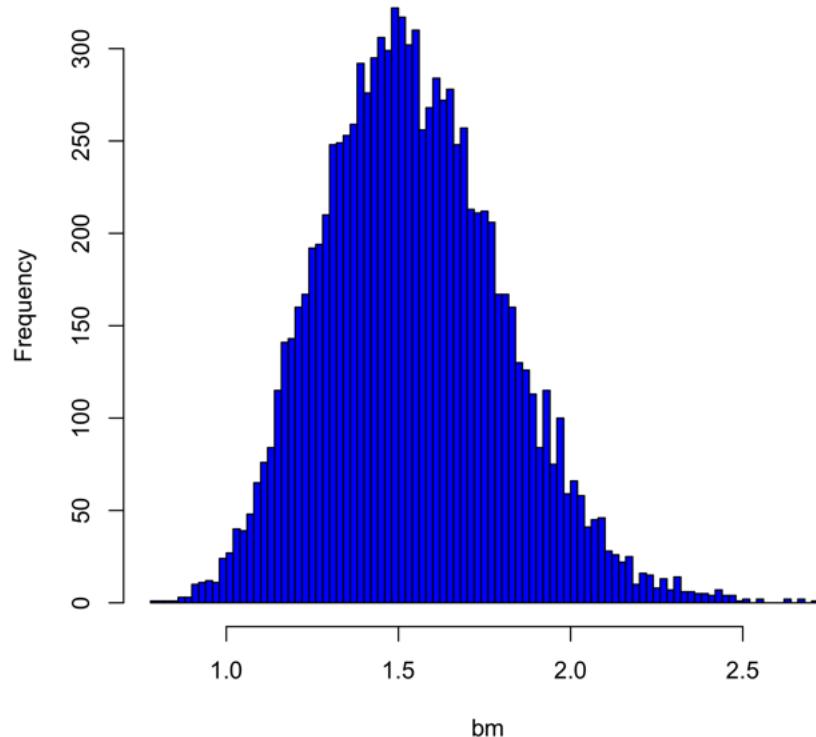


Figure 10.22: Histogram of  $B = 10^4$  bootstrap sample mean values with sample size  $n = 30$  for a sample from the F distribution with  $\nu_1 = 15$  and  $\nu_2 = 3$  degrees of freedom. [BootstrapMeanF.R](#)

Note that for estimating confidence intervals of sample means (and other statistics) no knowledge is necessary of the distribution from which one samples. Of course this makes bootstrapping so strong and attractive.

# Chapter 11

## Parameter estimation: the Bayesian approach

*The Bayesian approach to parameter estimation is based on the consequent use of probability/probability density and especially Bayes' theorem. The goal is to calculate the posterior distribution for the model parameter(s) and to use this distribution for making inference about the parameters as, for example, by calculating the posterior mean, variance, or certain intervals. The posterior is essentially (except for normalization) given by the product of a prior distribution and the likelihood function (Section 9). In Section () we will study some examples like estimating the mean and variance of a normal population using non-informative priors (Box 8). Posterior distributions for other priors and/or other models will be compiled in Section().*

### Literature:

- Zellner (1971) [excellent early approach to Bayesian analysis with often detailed mathematical details]
- Silvia & Skilling (2006) [nice little booklet with good examples]
- Robert (2007) [written for professional statisticians/mathematicians]
- Gelman et al. (2020) [comprehensive, many examples, including complex (hierarchical) models; however, might be more difficult to digest for beginners; mathematical details often missing]

## 11.1 Bayes' theorem in the context of parameter estimation

Bayes' theorem (Eq. 4.11)

$$P(B|A \cap I) = \frac{P(A|B \cap I) P(B|I)}{P(A|I)} \quad (11.1)$$

has been derived from the product rule of probabilities (Eq. 4.10) exploiting the symmetry under interchange of propositions  $A$  and  $B$ . For PDFs the products rules reads (Zellner, 1971)

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}) &= p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y}) \end{aligned} \quad (11.2)$$

and thus

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \quad \text{Bayes' Theorem} \quad (11.3)$$

where  $\mathbf{y}$  are the data ('data vector') and  $\boldsymbol{\theta}$  is the list of model parameters ('parameter vector'). The distribution on the left-hand-side of Bayes' theorem,  $p(\boldsymbol{\theta}|\mathbf{y})$ , is the **posterior distribution** or **posterior** for short. It gives the probability distribution for the model parameters  $\boldsymbol{\theta}$  (left of the vertical bar) in the light of (depending on) the data  $\mathbf{y}$  (right of the vertical bar). The distribution  $p(\mathbf{y}|\boldsymbol{\theta})$  is the **likelihood**: How likely is it to observe the data  $\mathbf{y}$  (left of the vertical bar) given the chosen model with parameters  $\boldsymbol{\theta}$  (right of the vertical bar)? The distribution  $p(\boldsymbol{\theta})$  is the **prior** comprising the knowledge about the model parameters  $\boldsymbol{\theta}$  before (prior to) observing the data.<sup>1</sup> The distribution  $p(\mathbf{y})$  in the denominator could be interpreted as the probability density distribution for observing the data  $\mathbf{y}$  without any assumptions about the population from which to sample. In the context of parameter estimation this term is considered as a **normalization constant**, i.e. one replaces the distribution by a constant value; the constant is chosen in such a way that the posterior PDF is normalized (see below).

In the Bayesian approach one performs a **switch of viewpoint**: instead of interpreting the likelihood  $p(\mathbf{y}|\boldsymbol{\theta})$  as a PDF for observing the data given a model (including parameter values), one uses the same mathematical expression, however, interprets it as the likelihood for the model parameters  $\boldsymbol{\theta}$  given the data  $\mathbf{y}$ .<sup>2</sup> The resulting distribution  $LF(\boldsymbol{\theta}|\mathbf{y})$  is called the **likelihood function**; it is *not* normalized with respect to  $\boldsymbol{\theta}$  and therefore is not a PDF. However, this causes no problems because one has to normalize the posterior anyway. Thus one finally obtains that the **posterior is proportional to the likelihood function times the prior** or

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto LF(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta}) \quad (11.4)$$

for short. It can be normalized as follows

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{LF(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta})}{\int LF(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad (11.5)$$

where the integration is over all possible values of  $\boldsymbol{\theta}$ .

---

<sup>1</sup>The prior is independent of the data; consequently,  $\mathbf{y}$  does not show up in the argument list of the prior.

<sup>2</sup>In the older literature, this switch of perspective is called 'inverse probability'.

## 11.2 Estimating the mean of a normal population with known variance

One of the most useful estimation problems concerns the normal population: What can one say about the mean,  $\mu$ , and the variance,  $\sigma^2$ , of a normal population based on the data  $\mathbf{y} = \{y_1, \dots, y_n\}$  and non-informative priors? We will first address the problem where the variance is assumed to be known and thus only a single parameter, namely the mean, has to be estimated. This and the two-parameter problem will be used to illustrate the Bayesian approach to parameter estimation. Important concepts are location & scale parameters, non-informative priors, likelihood, likelihood function, sufficient statistics, posterior, mean, mode, variance, credible intervals.

We want to infer something about the mean of a normal population from a sample  $\mathbf{y} = \{y_1, \dots, y_n\}$  of  $n$  independent data  $y_k$ . That the data stem from a normal (and not from a  $t$  or  $F$  distribution) is already our first assumption. In this 'textbook example' we know the true value  $\sigma^2 = \sigma_0^2$  of the variance (this is always never the case in 'real world problems'). If we have no prior information about the mean  $\mu$  it can take any value between  $-\infty$  and  $\infty$ . This is characteristic qualifies  $\mu$  as a **location parameter**: it can be located anywhere, we have no clue before taking observations. In the Bayesian approach to parameter estimation one has to assign a prior distribution for the unknown parameter. In order to express our ignorance about the location of  $\mu$  we assign the same probability density value to all possible values of  $\mu$ , i.e. the prior has the same positive value for all  $\mu$  between  $-\infty$  and  $\infty$ . It is called a 'flat' prior because a graph of the prior would look flat. The flat prior can not be normalized to 1 because the integral

$$\int_{-\infty}^{\infty} \text{flat prior}(x) dx = \text{const.} \int_{-\infty}^{\infty} dx \quad (11.6)$$

yields  $\infty$ . Thus this prior is not a PDF; it is called '**improper**'. Despite the normalization problem, the flat prior is used here (and in several other estimation problems) and can yield reasonable results, especially proper posterior distributions. The flat prior represents perfectly our ignorance of  $\mu$  and is thus called a **non-informative prior**.

The other distribution that we have to assign is the likelihood: How likely is it to observe the data  $\mathbf{y} = \{y_1, \dots, y_n\}$  given the assumption that they stem from a normal population with mean  $\mu$  and variance  $\sigma_0^2$ ? The likelihood for observing any  $y_k$  from data set  $\mathbf{y} = \{y_1, \dots, y_n\}$  can be immediately written down because it is the normal PDF

$$L(y_k | \mu, \sigma_0^2) = \mathcal{N}(y_k; \mu, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_k - \mu)^2}{2\sigma_0^2}\right). \quad (11.7)$$

The likelihood for observing two of  $n$  data points, say,  $y_k$  and  $y_m$ ,  $m \neq k$  is in general given by a joint distribution  $p(y_k, y_m | \mu, \sigma_0^2)$  which can be split up using the product rule:

$$p(y_k, y_m | \mu, \sigma_0^2) = p(y_k | y_m, \mu, \sigma_0^2)p(y_m | \mu, \sigma_0^2) = p(y_m | y_k, \mu, \sigma_0^2)p(y_k | \mu, \sigma_0^2). \quad (11.8)$$

We have assumed that the data are independent of each other and thus  $p(y_k | y_m, \mu, \sigma_0^2) = p(y_k | \mu, \sigma_0^2)$  and  $p(y_m | y_k, \mu, \sigma_0^2) = p(y_m | \mu, \sigma_0^2)$  leading to the simplified product rule

$$p(y_k, y_m | \mu, \sigma_0^2) = p(y_k | \mu, \sigma_0^2)p(y_m | \mu, \sigma_0^2). \quad (11.9)$$

Therefore we can write the joint likelihood for  $\mathbf{y}$  as a simple product of normal PDFs for single observations (Zellner, 1971, p. 14–15):

$$L(\mathbf{y} | \mu, \sigma_0^2) = \prod_{k=1}^n \left[ \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_k - \mu)^2}{2\sigma_0^2}\right) \right] = \left( \frac{1}{\sqrt{2\pi\sigma_0^2}} \right)^n \exp\left(-\sum_{k=1}^n \frac{(y_k - \mu)^2}{2\sigma_0^2}\right) \quad (11.10)$$

$$= (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{\nu s^2 + n(\bar{y} - \mu)^2}{2\sigma_0^2}\right) \quad (11.11)$$

where  $\nu = n - 1$  are the degrees of freedom,

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \quad (11.12)$$

is the sample mean and

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2 \quad (11.13)$$

is sample variance. Note that the likelihood (Eq. 11.11) depends on the sample variance  $s^2$ . Should  $s^2$  have an impact on the estimation of the mean despite the fact that we know the true variance  $\sigma_0^2$ ? Should we be worried? The answer is given further below in this section.

The likelihood (Eq. 11.11) depends on the data only via the sample mean and the sample variance, which are called '**sufficient statistics**' because these two quantities are sufficient to calculate the likelihood. The details of the sample ( $n$  values) play a role only as far as they contribute to the two quantities  $\hat{\mu}$  and  $s^2$ . Or in other words, for  $n \gg 2$  most of details in  $y$  don't matter for the likelihood.

We now switch the viewpoint by interchanging the unknown parameter  $\mu$  with the data yielding the **likelihood function**

$$LF(\mu|y, \sigma_0^2) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{\nu s^2 + n(\bar{y} - \mu)^2}{2\sigma_0^2}\right) \quad (11.14)$$

which describes how likely it is that the mean is  $\mu$  given the data and the variance  $\sigma_0^2$ . Multiplication by the flat prior does not change anything with Eq. (11.14). Normalization of Eq. (11.14) is easy because the likelihood function is a constant times  $\exp\left(-\frac{(\mu - \hat{\mu})^2}{2\sigma_0^2/n}\right)$  which implies that the **posterior** is normal with mean  $\hat{\mu}$  and variance  $\sigma_0^2/n$  (Fig. 11.1):

$$p(\mu|y, \sigma_0^2) = (2\pi(\sigma_0^2/n))^{-1/2} \exp\left(-\frac{(\bar{y} - \mu)^2}{2(\sigma_0^2/n)}\right) = \frac{1}{\sqrt{2\pi(\sigma_0^2/n)}} \exp\left(-\frac{(\bar{y} - \mu)^2}{2(\sigma_0^2/n)}\right) \quad (11.15)$$

where  $\sigma_0^2/n$  is the square of the **standard error of the mean**  $\sigma_0/\sqrt{n}$ . Note that the second statistics,  $s^2$ , 'disappears' by normalization (it's a kind of magic'): comparison of the posterior (Eq. 11.15) with the likelihood function (Eq. 11.14) shows that the normalization factor

$$q = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{\nu s^2 + n(\bar{y} - \mu)^2}{2\sigma_0^2}\right)}{(2\pi(\sigma_0^2/n))^{-1/2} \exp\left(-\frac{(\bar{y} - \mu)^2}{2(\sigma_0^2/n)}\right)} \quad (11.16)$$

is given by

$$q = n^{-1/2} (2\pi\sigma_0^2)^{-(n-1)/2} \exp\left(-\frac{\nu s^2}{2\sigma_0^2}\right). \quad (11.17)$$

### Music

Pick up your air guitar and jam with astrophysicist and guitar player Brian May: [https://www.youtube.com/watch?v=0p\\_1QSUsbsM](https://www.youtube.com/watch?v=0p_1QSUsbsM)

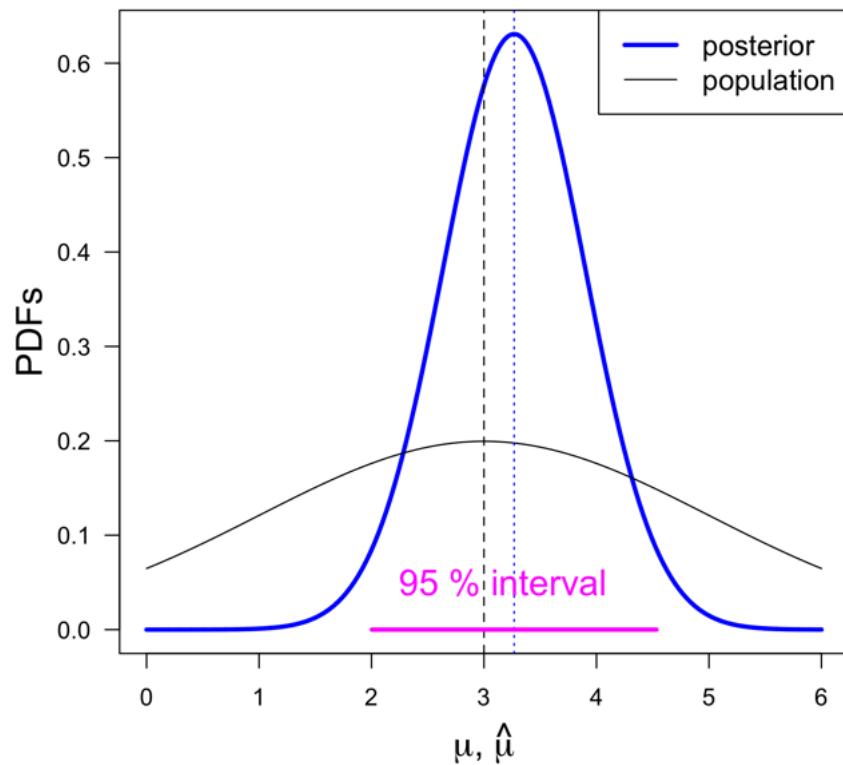


Figure 11.1: Posterior distribution  $p(\mu|y, \sigma_0^2)$  for an artificial data set of  $n = 10$  data with sufficient statistics  $\hat{\mu} = 3.27$  and  $s^2 = 2.84$  (blue solid line) from a normal population with mean  $\mu = 3$  and variance  $\sigma_0^2 = 4$  (thin black line). The PDF of the normal population from which we took a sample is much broader than the posterior. The posterior is normal PDF with mean equal to the estimated mean  $\hat{\mu} = 3.27$  and variance  $\sigma_0^2/n = 0.4$  ( $\sigma_0/\sqrt{n} = 0.6325$  is the standard error of the mean). [BayesianEstNormalVarKnown.R](#)

### Estimation by analyzing the posterior

The posterior (Eq. 11.15) is a probability density function for the unknown parameter  $\mu$  in the light of data. This can be seen as the most complete Bayesian estimation result. However, often one likes to condense this information into one or a few characteristic values. This could be, for example, the mode, the mean, or the median for the central tendency and the variance or standard deviation for the spread (dispersion). For normal PDFs mode, mean, and median give identical values:  $\hat{\mu} = 3.27$  is the sample mean. The standard deviation of the posterior (Eq. 11.15) is the standard error of the mean  $\sigma/\sqrt{n} = 0.63$ . I.e. one would report the estimate  $\hat{\mu} = 3.27 \pm 0.63$  where the uncertainty is usually reported 'on the  $1\sigma$  level' ( $\pm$  one standard deviation). Here the  $\sigma$  refers to the standard deviation of the posterior and not to that of the population from which we sample. Note that some scientific communities use other conventions as, for example,  $\pm 2\sigma$  because this (approximately) indicates the 95% interval.

## 11.3 Estimate the variance of a normal population for known mean value

Another single-parameter problem is the estimation of the variance  $\sigma^2$  of a normal population when the mean value is known, i.e.  $\mu = \mu_0$ . By assigning the Jeffreys prior  $1/\sigma^2$  to the scale parameter  $\sigma^2$ . This problem is a nice exercise for beginners in Bayesian estimation to test their skills (Exercise 36), however, it is less important for direct applications, 'but a building block for more complicated, useful models' (Gelman et al., 2020). Here we just give the result. The posterior is given by either in form (parameterization) of the scaled inverse- $\chi^2$  PDF

$$\text{Inv-}\chi^2(x; \delta, \tau^2) = \frac{(\tau^2 \delta/2)^{\delta/2}}{\Gamma(\delta/2)} x^{-(\delta/2+1)} \exp(-\delta \tau^2/(2x)) \quad (11.18)$$

with  $\delta = n$ ,  $\tau^2 = \frac{\nu s^2}{n}$  or in form (parameterization) of the inverse-gamma PDF

$$p(\sigma^2 | \mathbf{y}, \mu_0) = \text{Inv-gamma}(\sigma^2 | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} \exp\left(\frac{-\beta}{\sigma^2}\right) \quad (11.19)$$

with  $\alpha = n/2$  and  $\beta = \frac{\nu s^2}{2}$ .

### 8: Non-informative priors for location & scale parameters

The normal distribution possesses two parameters, namely the mean,  $\mu$ , and the variance,  $\sigma^2$ . The mean  $\mu$  is a so-called **location parameter** (Where to locate the mean?) that can have values between  $-\infty$  and  $\infty$ . The variance  $\sigma^2$  is a so-called **scale parameter** (measuring the spread of the distribution); it can have values between 0 and  $\infty$ . Note that  $\eta = \log \sigma^2$  can have values between  $-\infty$  and  $\infty$  (same as for the location parameter  $\mu$ ).

"[T]here is a need for explicit rules for selecting prior distributions to represent 'knowing little' or ignorance. Perhaps, surprisingly, filling this need has been a difficult and controversial aspect of Bayesian approach to inference." (Zellner, 1971, p. 42) An answer to this problem has been given by Jeffreys (1961, p. 117): "If the parameter may have any value in a finite range, or from  $-\infty$  to  $+\infty$ , its prior probability should be taken as uniformly distributed. If it arises in such a way that it may conceivably have any value from 0 to  $\infty$ , the prior probability of its logarithm should be taken as uniformly distributed."

Assigning a uniform PDF over a finite range seems to be unproblematic.<sup>3</sup> However, a uniform distribution over the range from  $-\infty$  to  $+\infty$  can not be normalized. Non-normalizable priors are called **improper**. Priors following a uniform PDF are called 'flat' priors.

Jeffreys has shown that from his rule it follows that the non-informative prior for the variance  $\sigma^2$  is proportional to  $1/\sigma^2$ . The prior is improper as well because of the singularity at  $\sigma^2 = 0$  and the infinite range of possible values. It can further be shown that non-informative priors for other parameterizations of the spread as, for example, the standard deviation  $\sigma$  or the precision parameter  $h = 1/\sigma^2$ , read  $1/\sigma^n$  with  $n = 1$  for the standard deviation and  $n = -2$  for the precision parameter.

These results can be expressed as

$$p(\mu)d\mu \propto d\mu \quad (11.20)$$

$$p(\sigma^n)d\sigma^n \propto \frac{d\sigma^n}{\sigma^n} \quad (11.21)$$

Despite the normalization problems for infinite ranges, improper priors are used and can often lead to reasonable results. When applying improper prior one has to find out under which conditions the posterior is proper, i.e. normalizable. Often a small minimum sample size is enough for securing that posteriors are proper.

Further reading: Zellner (1971, p. 41–53) gives an excellent discussion of Jeffreys priors including an information theoretic justification of these priors. Kass & Wasserman (1996), Berger et al. (2009). Further discussion about non-informative and reference priors can be found in Chapter M.

## 11.4 Conjugate distributions

*Conjugate distributions are related to likelihood functions. They are computational convenient and are interpretable as 'additional data' when used as prior.*

A conjugate distribution has been used already in the neutrino example in Section 1.4. The likelihood function for  $\lambda$

$$LF(\lambda|x_1, x_2, \dots, x_n) = \frac{\lambda^s}{\prod_{i=1}^n x_i!} e^{-n\lambda} \quad (11.22)$$

is not normalized with respect to  $\lambda$ . Instead of calculating the integral

$$\int_0^\infty LF(\lambda|x_1, x_2, \dots, x_n) d\lambda = \frac{1}{\prod_{i=1}^n x_i!} \int_0^\infty \lambda^s e^{-n\lambda} d\lambda \quad (11.23)$$

we searched for a PDF has the same form with respect to the dependence on  $\lambda$  as the likelihood function, i.e.  $\lambda^s e^{-n\lambda}$ .

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (11.24)$$

possess this dependency and thus – with  $\alpha = s + 1$ ,  $\beta = n$  – is the desired posterior (Fig. 1.6). This gamma distribution is the posterior for a flat prior. The estimate of  $\lambda$  based on this distribution is given by  $\hat{\lambda}_B = \frac{s+1}{n} = 0.778$ .

Instead of the flat prior which expresses total prior ignorance about  $\lambda$ , one could choose more informative, however, largely diffuse ('broad range') priors. Gamma distributions are convenient for this purpose because the product of two gamma distributions is proportional to another gamma distribution. From

$$\text{Gamma}(\lambda; \alpha, \beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \quad (11.25)$$

and ( $p$  indicates prior)

$$\text{Gamma}(\lambda; \alpha_p, \beta_p) \propto \lambda^{\alpha_p - 1} e^{-\beta_p \lambda} \quad (11.26)$$

it follows that

$$\text{Gamma}(\lambda; \alpha, \beta) \text{Gamma}(\lambda; \alpha_p, \beta_p) \propto \lambda^{\alpha + \alpha_p - 2} e^{-(\beta + \beta_p) \lambda} \propto \text{Gamma}(\lambda; \alpha + \alpha_p - 1, \beta + \beta_p), \quad (11.27)$$

i.e. the posterior is again a gamma distribution, however, with different parameter values. Using again the mean of the posterior as an estimator of  $\lambda$ , one obtains

$$\hat{\lambda} = \frac{\alpha + \alpha_p - 1}{\beta + \beta_p} = \frac{s + \alpha_p}{n + \beta_p}. \quad (11.28)$$

For  $\alpha_p = 1$  and  $\beta_p \rightarrow 0$  the above result is identical to flat prior estimate. For  $\alpha_p > 1$  and  $\beta_p > 0$  the influence of the prior can be interpreted as an increase of an increase of  $s$  by  $\alpha_p - 1$  and  $n$  by  $\beta_p$ . If  $(\alpha_p - 1) \ll (s + 1)$  and  $\beta_p \ll n$ , the impact of the prior is relatively small.

Now we choose an informative prior, namely,  $\text{Gamma}(\lambda; 1, 1)$  (Fig. 11.2), which gives less and less weight with increasing  $\lambda$ . For the neutrino data ( $n = 2306, s = 1792$ )

$$\hat{\lambda} = \frac{s + 1}{n + 1} \pm \frac{\sqrt{s + 1}}{n + 1} = \frac{1793}{2307} \pm \frac{\sqrt{1793}}{2307} = 0.778 \pm 0.018 \quad (11.29)$$

i.e. up to three digits identical to the flat prior estimate which is not surprising given that  $s$  and  $n$  are large compared to the chosen  $\alpha_p = 1$  and  $\beta_p = 1$ .

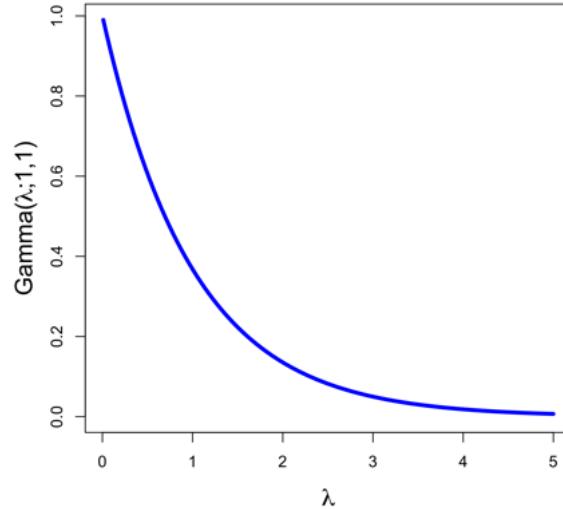


Figure 11.2: Informative prior  $\text{Gamma}(\lambda; \alpha_p = 1, \beta_p = 1)$ . R code: [gamma11.R](#)

"The concept, as well as the term 'conjugate prior', were introduced by Howard Raiffa and Robert Schlaifer in their work on Bayesian decision theory." (Raiffa & Schlaifer, 1961) "A similar concept had been discovered independently by George Alfred Barnard." (Wikipedia) Despite largely increased computational power which allows numerical calculation of posteriors for arbitrary priors, conjugate priors are still attractive because they allow the derivation of analytic results that allow insight into the impact of prior versus data. A list of various conjugate distributions can be found for example under 'Conjugate prior' (Wikipedia, accessed 26 June 2022).

## 11.5 Marginal posterior distribution for $\mu$ : normal population, $\mu$ & $\sigma^2$ unknown

This is our first multi-parameter problem – it is arguably the easiest and most important one. We will apply the non-informative Jeffreys prior  $p(\mu, \sigma^2) \sim 1/\sigma^2$ , calculate the posterior distribution  $p(\mu, \sigma^2|\mathbf{y})$ , split it up into the product  $p(\mu, \sigma^2|\mathbf{y}) = p(\mu|\sigma^2\mathbf{y}) p(\sigma^2|\mathbf{y})$ , and calculate the marginal distribution for  $\mu$ ,  $p(\mu|\mathbf{y})$ , by integrating out the nuisance parameter  $\sigma^2$ . The methods introduced here are of importance for all multi-parameter models.

The posterior is proportional to the product of likelihood function and prior. The likelihood function for our normal problem has been essentially been derived in Section 11.2: we just have to replace in (Eq. 11.14) the in that section assumed to be known variance  $\sigma_0^2$  by the now unknown  $\sigma^2$ :

$$LF(\mu, \sigma^2|\mathbf{y}) \propto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\nu s^2 + n(\bar{y} - \mu)^2}{2\sigma^2}\right) \quad (11.30)$$

where the sample mean  $\bar{y}$  and the sample variance  $s^2$  are the sufficient statistics,  $\nu = n - 1$  are the degrees of freedom.

As mentioned already we will apply the non-informative Jeffreys prior (compare Box 8)

$$p(\mu, \sigma^2) \sim 1/\sigma^2 \quad (11.31)$$

Consequently the **posterior** reads (except for normalization)

$$p(\mu, \sigma^2|\mathbf{y}) \propto \sigma^{-n-2} \exp\left(-\frac{\nu s^2 + n(\bar{y} - \mu)^2}{2\sigma^2}\right) = (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2}\right) \quad (11.32)$$

This is already our first results (identical to Gelman et al., 2020, Eq. 3.2).

**Split up the posterior into the product**  $p(\mu|\sigma^2\mathbf{y}) p(\sigma^2|\mathbf{y})$  Direct sampling from multi-parameter distribution is usually difficult, if not impossible. A way around is to split up the multi-parameter PDF by applying the product rule (Eq. 11.2) into a product of single-parameter distributions. For our two-parameter posterior  $p(\mu, \sigma^2|\mathbf{y})$  there would be two possibilities:

$$p(\mu, \sigma^2|\mathbf{y}) = p(\mu|\sigma^2\mathbf{y}) p(\sigma^2|\mathbf{y}) \quad (11.33)$$

or

$$p(\mu, \sigma^2|\mathbf{y}) = p(\sigma^2|\mu, \mathbf{y}) p(\mu|\mathbf{y}). \quad (11.34)$$

where the first terms on the right-hand-sides are **conditional distributions** (distributions for a single parameter conditional on the knowledge of the other parameter) and the second terms are **marginal distributions** (distributions where one parameter has ‘disappeared’ by integrating out this nuisance parameter). Sampling from the two-parameter posterior is now reduced to the succession of sampling from two single-parameter distributions.

We will go for the first product simply for practical reasons: The conditional distribution  $p(\mu|\sigma^2\mathbf{y})$  has been derived already in Section 11.2. Replacing again  $\sigma_0^2$  by  $\sigma^2$  in (Eq. 11.15) leads to the **normal distribution**

$$p(\mu|\sigma^2\mathbf{y}) \sim \mathcal{N}(\bar{y}, \sigma^2/n). \quad (11.35)$$

The marginal posterior distribution  $p(\sigma^2|\mathbf{y})$  is calculated by integrating out the nuisance parameter  $\mu$  from the posterior

$$p(\sigma^2|\mathbf{y}) \propto \int_{-\infty}^{\infty} (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2}\right) d\mu \quad (11.36)$$

$$\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \quad (11.37)$$

(Gelman et al., 2020, Eq. 3.4) which is proportional to a scaled inverse- $\chi^2$  PDF

$$\sigma^2 | \mathbf{y} \sim \text{Inv-}\chi^2(n - 1, s^2) \quad (11.38)$$

with parameters  $n - 1$  and  $s^2$ .

Thus we have split up the two-parameter posterior into product of a scaled inverse- $\chi^2$  marginal distribution for  $\sigma^2$  and a conditional normal distribution for  $\mu$ . The sampling from the posterior proceeds as follows:

1. Draw a sample  $\mathbf{y}$  from a normal population and calculate the sufficient statistics  $\bar{y}$  and  $s^2$ .
2. Draw a random sample from the scaled inverse- $\chi^2$  distribution for  $\sigma^2$ .
3. Draw a random sample from the conditional normal PDF (Eq. 11.35).
4. Repeating the drawing under 2. and 3.  $M$  times ( $M \gg 1$ ) yields a frequency distribution  $f(\mu, \sigma^2)$  that can, after division by  $M$ , be used as a random sample from the posterior.

The marginal posterior distribution for  $\mu$ ,  $p(\mu | \mathbf{y})$  can be written in the forms

$$p(\mu | \mathbf{y}) = \int \underbrace{p(\mu, \sigma^2 | \mathbf{y})}_{\text{joint posterior}} \quad (11.39)$$

$$= \int \underbrace{p(\mu | \sigma^2, \mathbf{y})}_{\text{normal}} \quad \underbrace{p(\sigma^2 | \mathbf{y})}_{\text{scaled inverse-}\chi^2} \quad d\sigma^2 \quad (11.40)$$

According to Eq. (11.40), the marginal posterior distribution of  $\mu$  can be regarded as a mixture of normal distribution where the scaled inverse- $\chi^2$  distribution provided the weights.

**The non-standardized  $t$  distribution is the analytic form of the marginal posterior distribution of  $\mu$ .** According to Eq. (11.39), the marginal posterior distribution of  $\mu$  can be calculated directly by integrating the joint posterior, (Eq. 11.32), over  $d\sigma^2$ :

$$p(\mu | \mathbf{y}) \propto \int_0^\infty (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2}\right) d\sigma^2 \quad (11.41)$$

One obtains (Gelman et al., 2020, p. 66)

$$p(\mu | \mathbf{y}) \propto \left[1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}\right]^{-n/2} = \left[1 + \frac{1}{\nu} \frac{(\mu - \bar{y})^2}{s^2/n}\right]^{-(\nu+1)/2} \quad (11.42)$$

which after normalization yields the non-standardized  $t$  distribution  $t_\nu(\mu; \bar{y}, s^2/n)$  (Section C.3.2). Instead of using the non-standardized  $t$  for  $\mu$ , one can define the statistic

$$t = t(\mu) = \frac{(\mu - \bar{y})}{s/\sqrt{n}} \quad (11.43)$$

and show that  $t$  follows the standard Student's  $t$  distribution for  $\nu = n - 1$  degrees of freedom (Gelman et al., 2020, p. 66), i.e.

$$p(t | \mathbf{y}) = t_\nu(t) \quad (11.44)$$

where  $\nu = n - 1$ .

The marginal posterior of  $\mu$  for a random sample of small size ( $n = 5$ ) is shown in Fig. 11.3. The 95% interval can be calculated either from the non-standardized  $t$  distribution ( $\bar{y} \pm |t_{\nu-1}(0.025, \bar{y}, \sqrt{s^2/n})|$ ) or from the standard  $t$  distribution ( $\bar{y} \pm |t_{\nu-1}(0.025)|\sqrt{s^2/n}$ ).

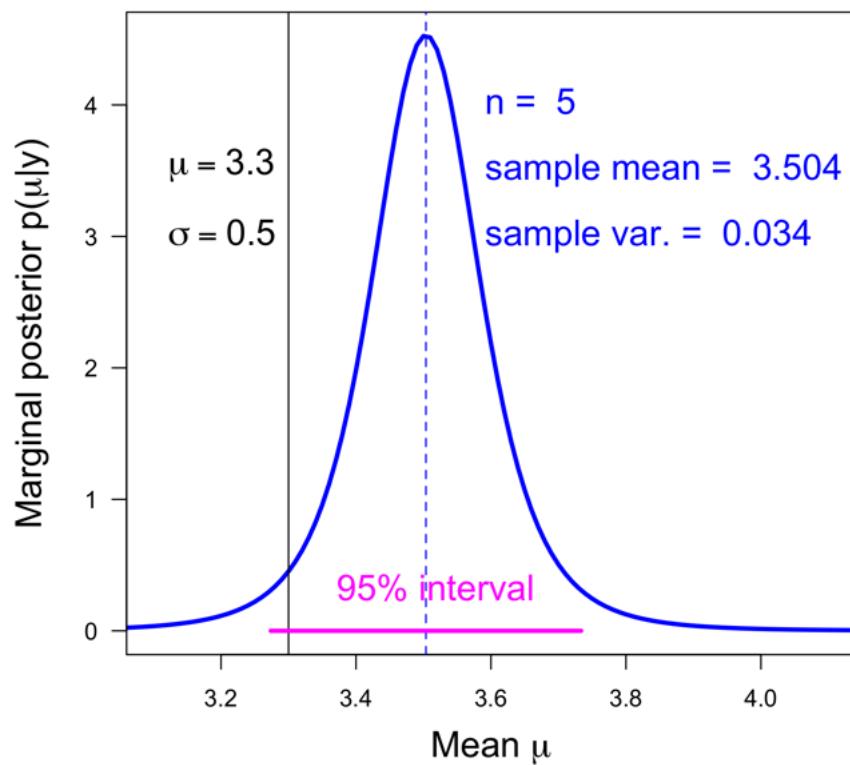


Figure 11.3: Marginal posterior of  $\mu$  (blue solid line) for a random sample  $y$  of small size ( $n = 5$ ) from a normal population with mean  $\mu = 3.3$  and standard deviation  $\sigma = 0.5$ . The sufficient statistics are the sample mean  $\bar{y} = 3.50$  and sample variance  $s^2 = 0.034$  ( $s = 0.19$ ). The 95% interval ranges from 2.99 to 4.02.

[BayesianEstNormalMeanVarUnknown.R](#)

## 11.6 Reporting the results of Bayesian parameter estimation

For the sake of transparency and the possibility of modified (alternative) analyses it is important to provide:

1. **A detailed discussion of the stochastic model assumed to generate the observations.** Why to use a normal or binomial distribution for the population from which we sample? An interesting example for the choice of alternative distributions is discussed in Exercise 40. The likelihood function for a sample should be reported.
2. **Priors** Which priors have been chosen and why? Non-informative or reference priors or informative priors based on what prior information or data.
3. **Sample information** A detailed discussion on how the data were obtained ('methods and data' section in reports); small data sets should be listed in the report (may be in an appendix). Large data sets should be made available (open access) in data repositories such as, for example, PANGEA (<https://www.pangaea.de>) or other members of the World Data System. An excellent example of freely available data is the World Ocean Atlas (<https://www.ncei.noaa.gov/products/world-ocean-atlas>).
4. **Posterior PDFs for parameters of interest** Report the complete posterior PDF and summary characteristics like measures of central density (posterior mean, median, and/or mode) and dispersion (posterior variance, variance matrix, standard deviation, MADN). Intervals (credibility sets) might also be helpful for readers.
5. **Computer code** Provide the computer code used for the Bayesian analysis either in an appendix (for short codes) or in special repositories or developer platforms (for example: <https://github.com>).

**Further reading** Zellner (1971, Section 2.14), Hildreth (1963)

## 11.7 Mean Squared Error (MSE)

In Section 10.2 we have asked for the first time 'How good are estimates of the mean  $\mu$ ?' and we have seen that the sample mean provides an unbiased estimate of the mean. In Section 10.3 we have investigated four different estimators of the variance  $\sigma^2$  of which two are biased and one can rarely be applied because the knowledge of the true mean  $\mu$  is required. The remaining estimator includes the famous  $1/(n - 1)$  factor. However, the amount of bias is not the only measure that can be applied to judge the goodness of estimators. In what follows we will choose the mean squared error (MSE) as our measure of goodness. The goal is to calculate the MSEs for various point estimators and, if accepting MSE as the measure of goodness, to find estimators with small or, if possible, minimum MSE from a defined set of estimators. MSE can be split up in two additive terms, namely the variance of the estimator and its bias squared. In order to minimize MSE a vanishing bias might seem helpful, however, we will see that a (slightly) biased estimator can actually lead to smaller MSE.

References: This section follows largely Casella & Berger (2002, p. 330ff).

The **mean squared error (MSE)** is defined as the expectation of the squared difference between the estimator,  $W$ , and the true value,  $\theta$ ,

$$E(W - \theta)^2 \quad (11.45)$$

The MSE can be split up into two terms, namely the variance of the estimator,  $\text{Var}W$ , and the square of the bias of the estimator,  $(\text{Bias}W)^2$ :

$$\text{MSE} = E(W - \theta)^2 = \text{Var}W + (E\theta - \theta)^2 = \text{Var}W + (\text{Bias}W)^2. \quad (11.46)$$

In words, the definition of the bias of a point estimator  $W$  of a parameter  $\theta$  reads: it is the difference between the expected value of  $W$  and  $\theta$ .

In Section 10.3 we have investigated the estimators for the variance of normal densities

$$\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (11.47)$$

$$\hat{\sigma}_2^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (11.48)$$

which differ by the factor by which one divides the sum of squares:  $n - 1$  or  $n$ . In contrast to the first estimator,  $\hat{\sigma}_1^2$ , which is unbiased, the second estimator,  $\hat{\sigma}_2^2$  is biased. But what about the MSE? Under same conditions (random samples of size  $n$  from a normal distribution with mean  $\mu$  and (finite) variance  $\sigma^2$ , one obtains

$$\text{MSE}_1 = E(\hat{\sigma}_1^2 - \sigma^2)^2 = \frac{2}{n-1} \sigma^4 = \frac{2n^2}{(n-1)n^2} \sigma^4 \quad (11.49)$$

$$\text{MSE}_2 = E(\hat{\sigma}_2^2 - \sigma^2)^2 = \frac{2n-1}{n^2} \sigma^4 = \frac{2n^2 - 3n + 1}{(n-1)n^2} \sigma^4 \quad (11.50)$$

(Casella & Berger, 2002, p. 331)<sup>4</sup>. Obviously,  $\text{MSE}_2$  is smaller than  $\text{MSE}_1$  for all  $n \geq 2$ , i.e. the biased estimator has a smaller MSE. This simple example should only show that unbiased estimators are not always the best estimators when taking MSE as a measure of goodness. The maximum likelihood estimator of the variance,  $\text{MSE}_2$ , is biased but has smaller MSE than the unbiased estimator  $\text{MSE}_1$ . Note that Casella & Berger (2002, p. 332) discuss which estimator to choose when  $\theta$  is a location or scale parameter. For  $\sigma^2$  (a scale parameter) they recommend to stick with the unbiased estimator  $\hat{\sigma}_1^2$ . The biased estimator  $\hat{\sigma}_2^2$  is applied for example in the R routine `AIC()` for calculating the Akaike information criterion (Section O.1).

Expectations of estimators are often not easy to calculate (see, for example, footnote above). However, Monte Carlo simulations can help to find numerical values: `MSEnormalvariance.R`.

---

<sup>4</sup>The calculation of  $E(\hat{\sigma}_1^2 - \sigma^2)^2$  is based on the fact that this expectation is equal to the variance of  $\hat{\sigma}_1^2$  and that for samples from normal distributions  $(n-1)\hat{\sigma}_1^2/\sigma^2$  follows a  $\chi^2$  distribution with  $n-1$  degrees of freedom (von Storch & Zwiers, 2001, p. 77).

**Exercise 36 Variance of normal PDF**

- (1) Derive the posterior for the variance  $\sigma^2$  of a normal population with known mean  $\mu = \mu_0$  based on the Jeffreys prior  $1/\sigma^2$ .
- (2) Estimate the variance  $\sigma^2$  of a normal population with mean  $\mu = \mu_0 = 3$  from a sample of size  $n$  with sample mean  $\hat{\mu} = 3.268$  and sample variance  $s^2 = 2.837$ .

**Exercise 37 Sample mean is MLE for Poisson rate constant (\*)**

The likelihood function for the rate constant  $\lambda$  of the Poisson distribution is proportional to  $\lambda^{t(\mathbf{y})} e^{-n\lambda}$  where  $t(\mathbf{y}) = \sum_{i=1}^n y_i$  is the sufficient statistic. Show that the Maximum Likelihood Estimate (MLE, mode of likelihood function) is equal to the sample mean.

Hint: use the identity  $\lambda = e^{\ln \lambda}$  before calculation the derivative of the likelihood function.

# Chapter 12

## Hypothesis testing

*Hypothesis testing is one of the most applied methods in data analysis.<sup>1</sup> Several schools of thought exist, namely (at least) (1) Null Hypothesis Significance Testing (NHST, or significance testing for short, 'Fisherian' approach, *p* values), (2) Neyman-Pearson approach (two competing hypotheses, 'power' of test), (3) likelihood principle (likelihood ratio tests), (4) Bayesian approach (prior, likelihood, posterior, Bayes' factor). Although the first two approaches have been criticized since a long time, they are still applied today.<sup>2</sup> Bayesian tests have many attractive features, however, easily accessible and useable software became available only recently; the choice of priors is an open-ended problem which can be seen as boon or bane.*

**Further reading (comparing different approaches to hypothesis testing):** Berger (2003), Gerrodette (2011)

**Further reading (Bayes factor functions, BFFs):** Johnson et al. (2023).

**Software for Bayesian tests:** R packages **BayesFactor** & **BFF**

---

<sup>1</sup>Often it is 'over-applied' and other approaches would have yielded more insight!

<sup>2</sup>The discussion is not finished yet: whereas some journals (for example, Basic and Applied Social Psychology; Trafimow, 2014; Trafimow & Marks, 2015) now refuse to accept articles using NHST and some statisticians suggest to abandon statistical significance (for example, McShane et al., 2019) others still suggests improvements of the NHST approach (Benjamin et al., 2018). The generation of 'big data' is a challenge for data analysis and especially for hypothesis testing (Efron, 2010).

## 12.1 Schools of thought: NHST versus Bayesian

Given one sample  $x$  of size  $n$  we can ask whether the sample stems from a population with mean  $\mu$  that is equal to a hypothesized mean  $\mu_0$ . In this section we will discuss how to answer this question by (1) estimation of mean and standard error ('physicists' approach'), (2) applying the one-sample two-sided t-test, (3) applying the one-sample two-sided Bayesian t-test. We will see, that for the example discussed all three approaches come to the same conclusion. However, this is not always the case when applying tests from different schools of thought and replacing hypothesis testing by estimating is not always possible (Section H.4).

### 12.1.1 Example

The following temperatures ( $^{\circ}\text{C}$ ) have been measured

$$x = \{1.5, 0.3, 1.8, -1.4, 0.8, 3.0, -0.3, 0.2, -0.4, 1.9, 0.0, 0.3, -1.0, 1.2, 3.8, 0.5, -0.8, 2.0, 1.1, 1.2, -0.4, 2.7, 0.5, -1.4, 1.1\}$$

We would like to know whether the sample stems from a population with true mean  $\mu_0 = 0^{\circ}\text{C}$ . Before doing any calculations let's look at the data (Fig. 12.1).

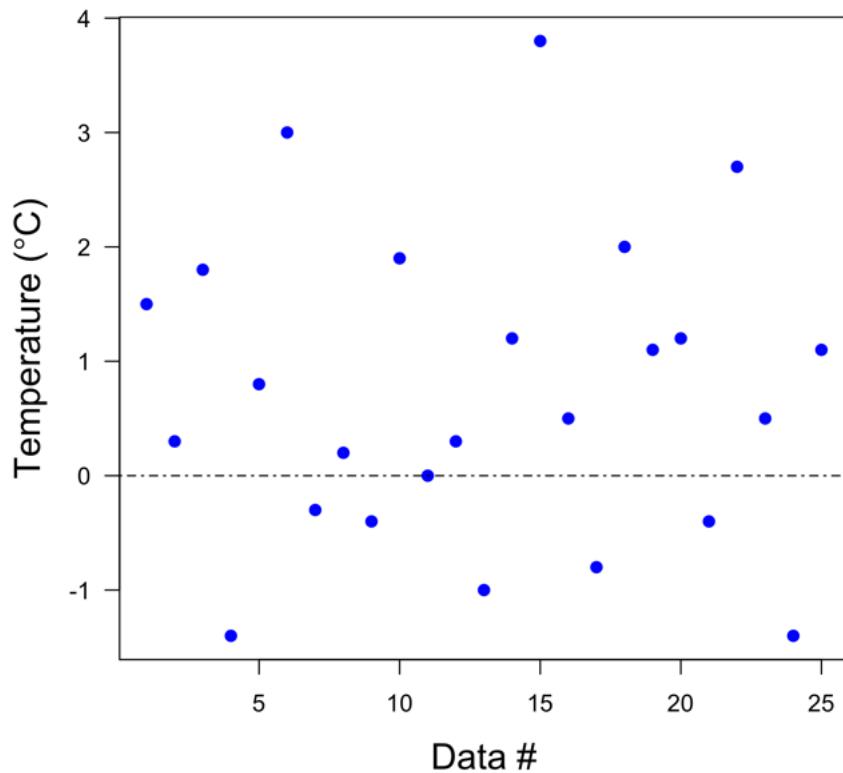


Figure 12.1: Observed temperatures (blue dots) and hypothesized true mean  $\mu_0 = 0^{\circ}\text{C}$  (black dash-dotted line). The hypothesized temperature  $\mu_0 = 0^{\circ}\text{C}$  lies inside the cloud of data, however, more data lie above  $\mu_0 = 0^{\circ}\text{C}$  than below which might speak against  $\mu = \mu_0$  where  $\mu$  is the true mean of the population from which we sampled. [BayesianHyp-t-test-Data.R](#)

### 12.1.2 Sample mean, standard deviation, standard error of the mean

As the next step of analysis we calculate the sample mean  $\bar{x} = 0.73^\circ\text{C}$ . The sample mean is an *unbiased estimator of the true mean  $\mu$*  (estimators will be presented in Chapter 10):

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0.73^\circ\text{C} \quad (12.1)$$

where the little hat  $\hat{\cdot}$  atop  $\mu$  stands for ‘estimate’.

Can we answer our question based on the estimate of  $\mu$ ? No, because the sample mean is almost always different from the true mean. What is missing is an estimate of the uncertainty of our estimate or of the uncertainty of the difference between estimated and hypothesized mean.

How to measure uncertainties? The observed data show quite a bit of spread (Fig. 12.1) that can be measured by the standard deviation,  $\sigma$ , or by the variance,  $\sigma^2$ . The true variance  $\sigma^2$  can be estimated by the sample variance  $s^2$ :

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (12.2)$$

However, this is a measure for the variance of the population from which we sample and not the uncertainty in our estimate for the mean. With increasing sample size  $n$ ,  $\hat{\sigma}^2$  will vary only slightly (random sample!) but stay more or less at a certain level (namely close to the true variance  $\sigma^2$ ). We have the expectation (from common sense) that the uncertainty of the mean should decrease with increasing sample size. This is indeed the case, as will be shown in a second.

When estimating the population mean by calculating the sample mean, one adds up random values from a statistical population. These values are usually different from the true mean  $\mu$ , some are smaller, some are larger. By adding up these values with negative or positive deviations from the true mean the deviations partially compensate each other. This compensation works better and better with increasing sample size. One can show<sup>3</sup> that the uncertainty in the estimate of the mean can be estimated by  $\sigma/\sqrt{n}$ , i.e. it decreases with one over the square of the sample size. The quantity  $s/\sqrt{n}$  ( $s$  is the estimate of  $\sigma$ ) is called the **standard error of the mean (SE)**; it can be calculated from the data. For our temperature data one obtains  $\text{SE} \approx 0.27^\circ\text{C}$ . We will add lines for  $\bar{x}$  (blue solid line) and  $\bar{x} \pm \text{SE}$  (red dashed lines) to our data plot (Fig. 12.2). The hypothesized  $\mu_0 = 0^\circ\text{C}$  lies clearly outside the range  $\bar{x} \pm \text{SE}$ , actually the difference  $\bar{x} - \mu_0$  is about 2.7 SE. Physicists would report a difference of 2.7 standard errors (or in their notation 2.7  $\sigma$  where  $\sigma$  stands for SE) and might reject the hypothesis  $\mu = \mu_0$  when the difference is larger than 2 SE. Depending on what is at stake they shift this rejection boundary to 3 or even 5 SE. This finishes already our first approach to answering our initial question. This simple estimation approach works because we know how the uncertainty falls off with increasing sample size. It is remarkable that this estimate  $s/\sqrt{n}$  applies to samples from all kind of distributions as long as their variances are finite (Casella & Berger, 2002, p. 213–214, Theorem 5.2.6).

---

<sup>3</sup>Casella & Berger, 2002, p. 213–214, Theorem 5.2.6, for any (not just normal!) distribution with finite variance, i.e.  $\sigma^2 < \infty$ ; it does not apply, for example, to the Cauchy distribution. Compare Section F.1 for details.

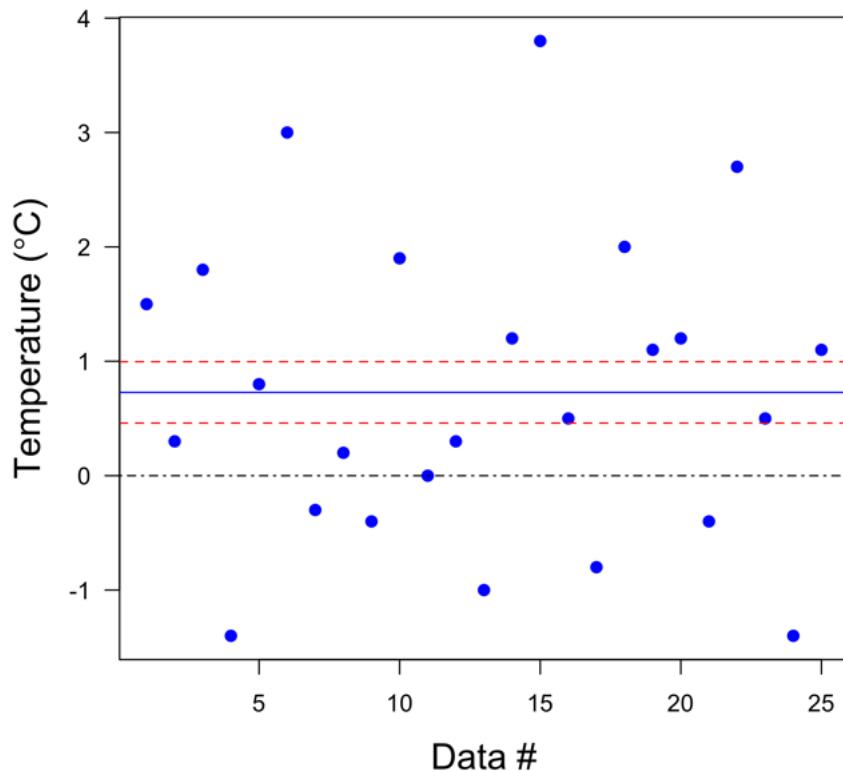


Figure 12.2: Observed temperatures (blue dots) and hypothesized mean  $\mu_0 = 0^\circ\text{C}$  (black dash-dotted line); the sample mean  $\bar{x}$  at  $0.73^\circ\text{C}$  (blue line) and the sample mean  $\pm$  one standard error of the mean (red dashed lines). The hypothesized  $\mu_0 = 0^\circ\text{C}$  lies clearly outside the range  $\bar{x} \pm \text{SE}$ , actually the difference  $|\bar{x} - \mu_0|$  is about  $2.7 \text{ SE}$ .

[Bayesian-t-test-SE.R](#)

### 12.1.3 Null Hypothesis Significance Testing (NHST): $t$ -test

In the Null Hypothesis Significance Testing (NHST) approach<sup>4</sup> one formulates a hypothesis<sup>5</sup> that involves prerequisites about the statistical population from which the sample is taken<sup>6</sup> and suggests a test statistic (a quantity that can be calculated from the data and that may shed light on the question one asks). The distribution of the test statistic is calculated under the assumption that the hypothesis is true. The hypothesis is rejected when the probability for the observed test statistic or for more extreme values, the so-called  $p$ -value, is smaller than a chosen value, the level of significance,  $\alpha$ ;<sup>7</sup> otherwise the null hypothesis is not rejected.

The null hypothesis differs from our initial question (Equal means?) mainly by prerequisites about the statistical population from which the sample is taken. In case of the one-sample  $t$ -test one assumes that the sample stems from a normally distributed statistical population with mean  $\mu_0$  and unknown variance  $\sigma_0^2$ . Such prerequisites are necessary to calculate the distribution of the test statistic (see below). Thus instead of the simple 'Equal means?' the null hypothesis  $H_0$  reads:

- (1) prerequisite:  $x = \{x_1, x_2, \dots, x_n\}$  is a random sample from a normally distributed statistical population  $\mathcal{N}(\mu_0, \sigma_0^2)$  where the mean,  $\mu_0$ , is given, but the variance,  $\sigma_0^2$ , is not known;
- (2) proper hypothesis:  $\mu = \mu_0$  where  $\mu_0$  is a hypothesized mean value and  $\mu$  is the (unknown) true mean of the population from which we sample.

The test further proposes a test statistic that can be calculated from the data and the hypothesized mean value  $\mu_0$ . The test statistic should allow us to infer something about  $H_0$ . It is often derived by intuition and seen as a plausible choice.<sup>8</sup> However, it is not necessarily derived from the basic laws of probability. In the case of the one-sample  $t$ -test the test statistic is called  $t$  and defined by

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}} = \frac{0.73 - 0}{\frac{1.34}{\sqrt{25}}} = 2.713 = t_{\text{obs}} \quad (12.3)$$

This looks familiar because  $t$  is equal to the distance between the sample mean and the hypothesized mean measured in terms of the standard error of the mean ('physicist approach'). However, the NHST approach does not stop here.

How is  $t$  distributed when  $H_0$  is true? Student (1908) solved this problem and derived the  $t$ -distribution (also called Student's  $t$ -distribution)  $t \sim t(t; v)$  where  $v$  is the degrees of freedom<sup>9</sup>. For the one-sample  $t$ -test  $v$  is given by the sample size,  $n$ , minus 1, i.e.  $v = n - 1$  (the single constraint is the calculation of the sample mean).

<sup>4</sup>Largely developed by Ronald Fisher (1890-1962) in the 1920ies. His book from 1925 'Statistical Methods for Research Workers' was very influential; its 14. edition was published in 1973.

<sup>5</sup>Nowadays called null hypothesis  $H_0$ .

<sup>6</sup>This is the case for so-called 'parametric tests', such as the  $t$ -test, ANOVA, variance ratio test, etcetera.

<sup>7</sup>Although Fisher developed most of these concepts, some of the notations were introduced by others and not necessarily used by Fisher.

<sup>8</sup>Well-chosen test statistics have desirable properties such as 'sufficiency', i.e. they encompass all essential information contained in the sample (for details compare, for example, Casella & Berger, 2002).

<sup>9</sup>Section 3.6

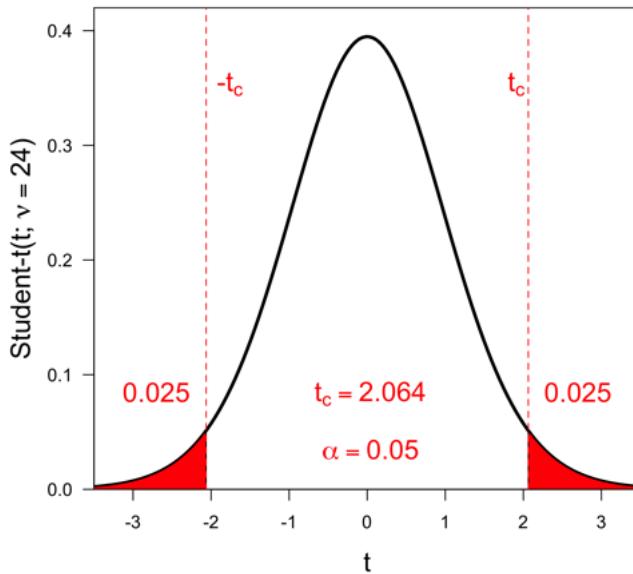


Figure 12.3: The  $t$ -distribution for  $\nu = 24$  (black line) and the rejection region (red area) for the level of significance  $\alpha = 0.05$ . In the two-side  $t$ -test the rejection region actually consists of two parts, one in each tail, with an area of  $\alpha/2 = 0.025$  each.  $t_c = t_{\alpha(2), \nu}$  is the critical  $t$  value for the two-sided  $t$ -test.

[NHST-t-test-RejectionRegion.R](#)

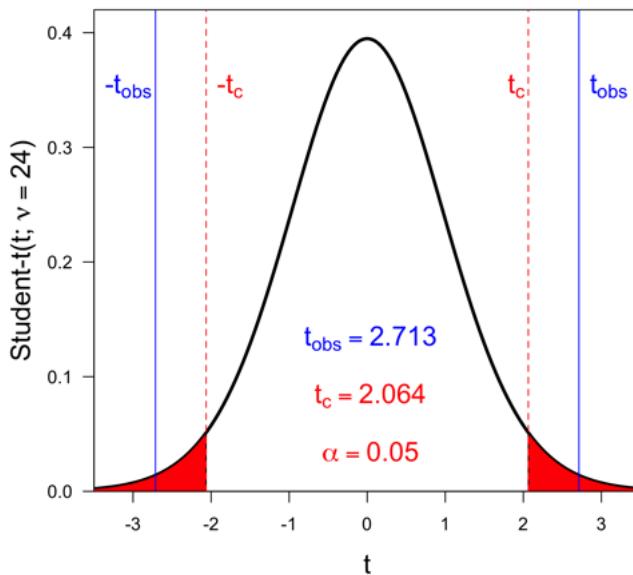


Figure 12.4: The  $t$ -distribution for  $\nu = 24$  (black line) and the observed  $t$ -value (blue vertical lines for  $\pm t_{\text{obs}} = \pm 2.713$ ). The observed  $t$ -value  $t_{\text{obs}} = 2.713$  falls into the right-tail rejection region which speaks against  $H_0$ .

[NHST-t-test-RRtObs.R](#)

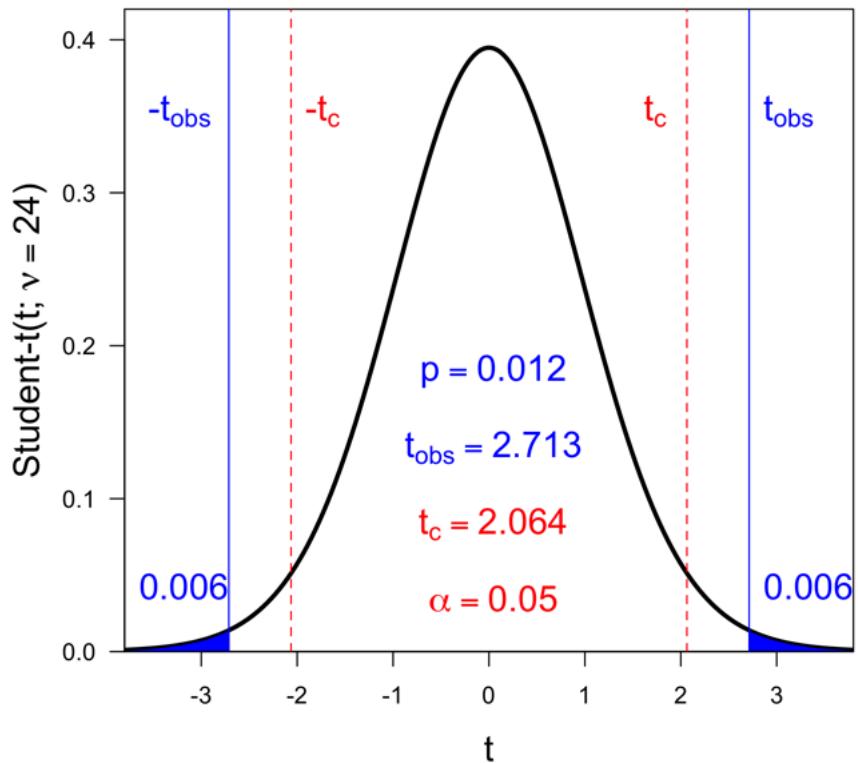


Figure 12.5: The  $p$ -value is the probability to observe  $t_{\text{obs}} = 2.713$  or more extreme values, i.e.  $t \geq t_{\text{obs}} = 2.713$  or  $t \leq -t_{\text{obs}} = -2.713$ . These are the two blue areas which each contribute a probability of about 0.006. Adding up these two probabilities yields the  $p$ -value 0.012. [NHST-t-test-pvalue.R](#)

**Making a decision:** The observed  $t$  value,  $t_{\text{obs}}$ , is located in the rejection region of the  $t$  distribution (red area) and the probability to observe such a  $t$  value or more extreme  $t$  values (blue regions) is  $p = 0.012$  ( $p$ -value, observed level of significance) and thus smaller than the chosen level of significance  $\alpha = 0.05$ . Based on this (and may be additional evidence) the null hypothesis is rejected.

### 12.1.4 Bayesian $t$ -test

In the Bayesian approach to testing one considers two competing hypotheses  $H_0$  (null hypothesis) and  $H_1$  (alternative or working hypothesis).<sup>10</sup> One has to specify priors for unknown (nuisance) parameters in the distribution functions describing the statistical populations from which the sample is supposed to stem. The discussion which priors to choose is still ongoing and thus other Bayesian tests for the same problem may be proposed in the near future. Here we will apply the Bayesian one-sample  $t$ -test developed by Rouder et al. (2009).

1. Formulation of  $H_0$  &  $H_1$ :

The null hypothesis,  $H_0$ , states that the sample  $x$  stems from a normal distribution with mean  $\mu = 0$  and unknown variance  $\sigma_0^2$ .

The alternative (or working) hypothesis,  $H_1$ , states that the sample stems from a normal distribution with unknown mean  $\mu$  and unknown variance  $\sigma^2$ .

2. Calculate likelihoods for  $H_0$  &  $H_1$  (compare Section H.11.1 for details). The likelihoods read for  $H_0$

$$f_0(x; \sigma_0^2) = (2\pi\sigma_0^2)^{-n/2} e^{-\frac{n}{2\sigma_0^2} [\bar{x}^2 + s'^2]} \quad (12.4)$$

and  $H_1$

$$f_1(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2} [(\bar{x} - \mu)^2 + s'^2]} \quad (12.5)$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{sample mean} \quad (12.6)$$

$$s'^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{modified sample variance} \quad (12.7)$$

are calculated from the sample (observed quantities or 'statistics').

3. Jeffreys (1961) proposed a reparametrization of  $H_1$  by introducing the effect size<sup>11</sup>

$$\delta = \frac{\mu}{\sigma} \quad \text{effect size} \quad (12.8)$$

Accordingly, the likelihood for  $H_1$  reads

$$f_1(x; \delta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n}{2} [(\bar{x}/\sigma - \delta)^2 + s'^2/\sigma^2]\right) \quad (12.9)$$

4. Choose priors:

$H_0$  contains a single unknown (nuisance) parameter, namely the standard deviation  $\sigma_0 > 0$ . Jeffreys (1961) suggests using

$$\pi_0(\sigma_0) = \frac{1}{\sigma_0} \quad (12.10)$$

as prior. Thus one puts less and less probability density on larger and larger standard deviations. This so-called **Jeffreys' prior** is not a PDF (it can not be normalized to 1 because of the singularity at  $\sigma_0 = 0$ ) and is therefore called an *improper prior*.

$H_1$  contains two unknown (nuisance) parameters, namely the standard deviation  $\sigma > 0$ , for which

<sup>10</sup>This is an essential difference to NHST which is exclusively based on a null hypothesis. Although an alternative hypothesis is often mentioned in the context of NHST, the alternative hypothesis is not required to calculate the  $p$  value used for making decisions.

<sup>11</sup>If the effect size is large, the effect is easily recognizable and one does not need to apply a test. It is more difficult to discern small from no (zero) effects.

we apply again Jeffreys' prior  $1/\sigma$ , and the effect size  $\delta$  for which Jeffreys motivates (Jeffreys, 1961, p. 269-270; compare also Ly et al., 2016) using the Cauchy prior (compare Section C.3.4)

$$\pi_{1a}(\delta) = \frac{1}{(1 + \delta^2) \pi} \quad (12.11)$$

Thus the prior for  $H_1$  reads

$$\pi_1(\delta, \sigma) = \frac{1}{\sigma (1 + \delta^2) \pi} \quad (12.12)$$

5. Calculate marginal likelihoods ('integrate out nuisance parameters'):

Marginal likelihoods are defined as integrals over the product of likelihood and prior(s):

$$m_0(\mathbf{y}) = \int_0^\infty f_0(\mathbf{x}; \sigma_0) \pi_0(\sigma_0) d\sigma_0 \quad (12.13)$$

$$= \int_0^\infty \underbrace{(2\pi\sigma_0^2)^{-n/2} e^{-\frac{n}{2\sigma_0^2} [\bar{x}^2 + s'^2]}}_{\text{likelihood}} \underbrace{\frac{1}{\sigma_0}}_{\text{prior}} d\sigma_0 \quad (12.14)$$

$$= (2\pi)^{-n/2} 2^{n/2-1} \frac{\Gamma(n/2)}{\left(n [\bar{x}^2 + s'^2]\right)^{n/2}} \quad \text{compare Jeffreys (1961, p. 273, Eq. 26)} \quad (12.15)$$

The marginal likelihood for  $H_1$

$$m_1(\mathbf{x}) = \int_{-\infty}^\infty d\delta \int_0^\infty d\sigma f_1(\mathbf{x}; \mu, \sigma) \pi_1(\delta, \sigma) \quad (12.16)$$

$$= \int_{-\infty}^\infty d\delta \int_0^\infty d\sigma \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n}{2} [(\bar{x}/\sigma - \delta)^2 + s'^2/\sigma^2]\right)}_{\text{likelihood}} \underbrace{\frac{1}{\sigma(1 + \delta^2)\pi}}_{\text{prior}} \quad (12.17)$$

is difficult to calculate (Jeffrey, 1961, has provided an approximate solution). Rouder et al. (2009) were able to reduce the problem to a single integral (they wrote "The derivation is straightforward and tedious and not particularly informative." and probably used a decomposition of the Cauchy prior that can be found, for example, in Liang et al., 2008).

6. Calculate the Bayes factor ( $m_0/m_1$  or  $m_1/m_0$  whatever is more convenient). Rouder et al. (2009) derived the following expression for  $B_{01}$

$$B_{01}(\mathbf{x}) = \frac{m_0}{m_1} = \frac{\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}}{\int_0^\infty (1 + ng)^{-1/2} \left(1 + \frac{t^2}{(1+ng)v}\right)^{-(v+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)} dg} \quad (12.18)$$

where  $v = n - 1$  are the degrees of freedom and  $t = \frac{\bar{x}}{\sqrt{s'^2/n}}$  is the test statistic for the (NHST)  $t$ -test. For the temperature data  $\mathbf{x}$  one obtains  $B_{01}(\mathbf{x}) = 0.289$  or  $B_{10}(\mathbf{x}) = 1/B_{01}(\mathbf{x}) = 3.45$ . How to interpret these values? Use the scales of evidence proposed by Jeffreys (1961)!

7. Scales of evidence (Jeffreys, 1961) for  $B_{01}$  ( $1/\sqrt{10} \approx 0.316$ ,  $\sqrt{10} \approx 3.16$ )

$B_{01} < 0.1$	strong evidence against $H_0$ (for $H_1$ )
$0.1 < B_{01} < 0.316$	substantial evidence against $H_0$ (for $H_1$ )
$0.316 < B_{01} < 1$	slight evidence against $H_0$ (for $H_1$ )
$1 < B_{01} < 3.16$	slight evidence against $H_1$ (for $H_0$ )
$3.16 < B_{01} < 10$	substantial evidence against $H_1$ (for $H_0$ )
$B_{01} > 10$	strong evidence against $H_1$ (for $H_0$ ).

or for  $B_{10}$

$B_{10} > 10$	strong evidence against $H_0$ (for $H_1$ )
$3.16 < B_{10} < 10$	substantial evidence against $H_0$ (for $H_1$ )
$1 < B_{10} < 3.16$	slight evidence against $H_0$ (for $H_1$ )
$0.316 < B_{10} < 1$	slight evidence against $H_1$ (for $H_0$ )
$0.1 < B_{10} < 0.316$	substantial evidence against $H_1$ (for $H_0$ )
$B_{10} < 0.1$	strong evidence against $H_1$ (for $H_0$ ).

The calculated Bayes factor  $B_{01}(x) = 0.289$  lies below 0.316 and thus can be considered as substantial evidence against  $H_0$  and for  $H_1$ .

### 12.1.5 Comparing $t$ -test and Bayesian $t$ -test

In this subsection we will compare results from the NHST one-sample two-sided  $t$ -test with the Bayesian  $t$ -test as derived by Rouder et al. (2009). For this purpose we will perform Monte Carlo simulations ( $M = 10^5$  runs) with three different effect sizes (zero:  $\delta = 0$ , small:  $\delta = 0.2$ , large:  $\delta = 1.2$ ) and two different sample sizes (small:  $n = 10$ , large:  $n = 100$ ).

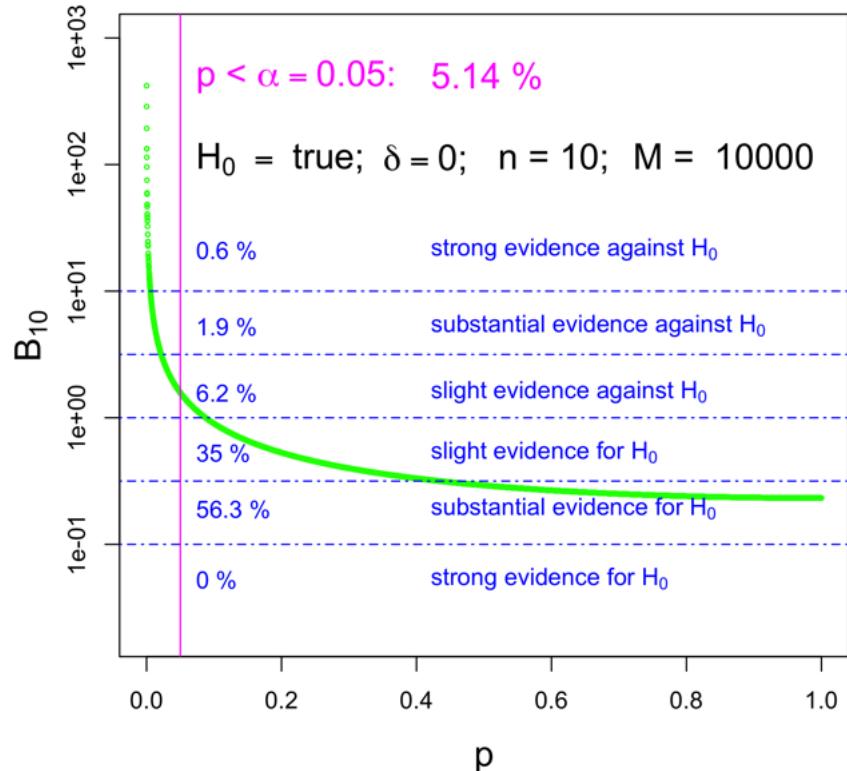


Figure 12.6: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 0$ , i.e.  $H_0$  is true and one would expect that  $H_0$  is not rejected. However, in about 5% of all samples  $H_0$  is (falsely) rejected based on  $p$ -values smaller than  $\alpha = 0.05$ . The Bayesian approach results in 0.6% plus 1.9% = 2.5% strong or substantial evidence against  $H_0$  and 56.3% substantial evidence for  $H_0$ . [BayesianNHST-t-test.R](#)

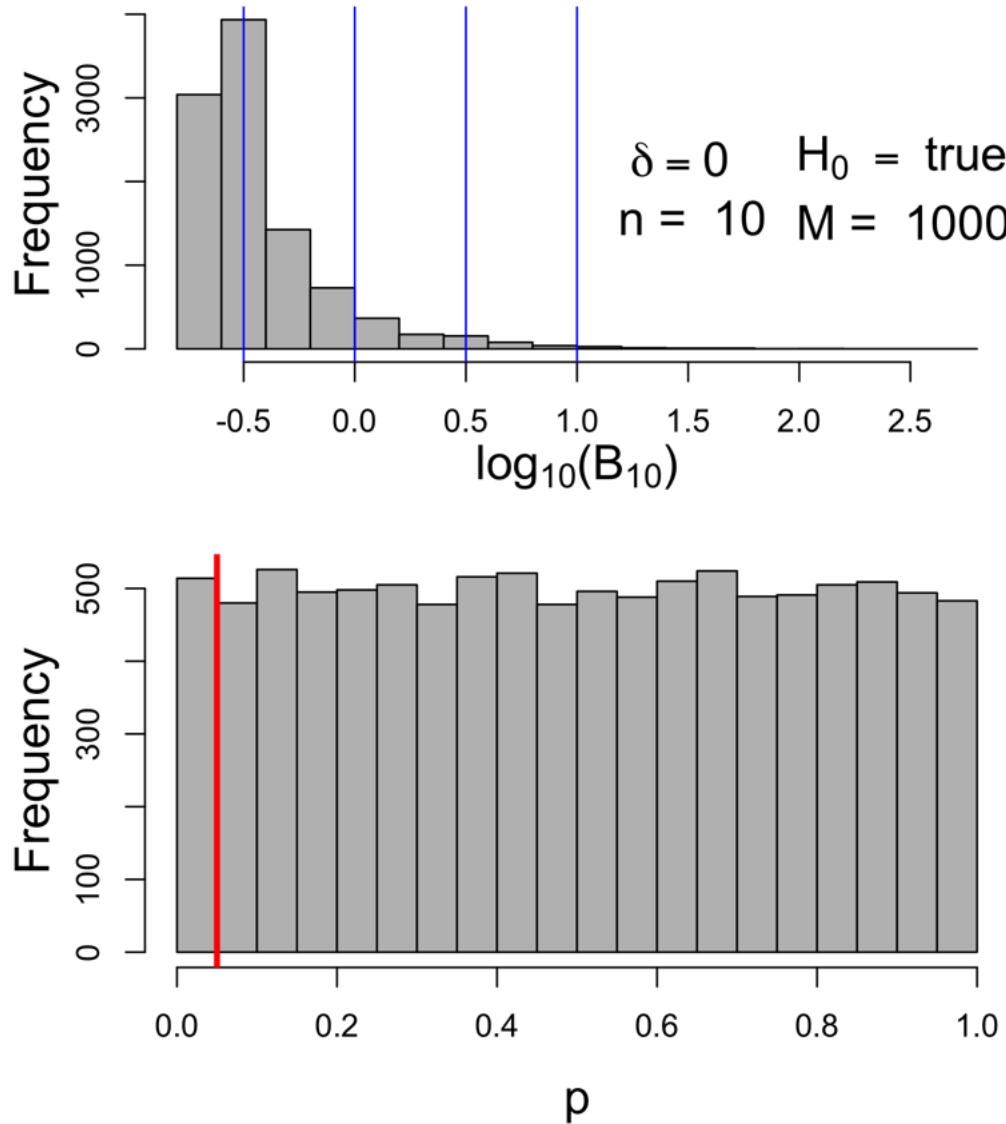


Figure 12.7: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 0$ , i.e.  $H_0$  is true and one would expect that  $H_0$  is not rejected. However, in about 5% of all samples  $H_0$  is (falsely) rejected based on  $p$ -values smaller than  $\alpha = 0.05$ . The Bayesian approach results in 0.6% plus 1.9% = 2.5% strong or substantial evidence against  $H_0$  and 56.3% substantial evidence for  $H_0$ . [BayesianNHST-t-testAll.R](#)

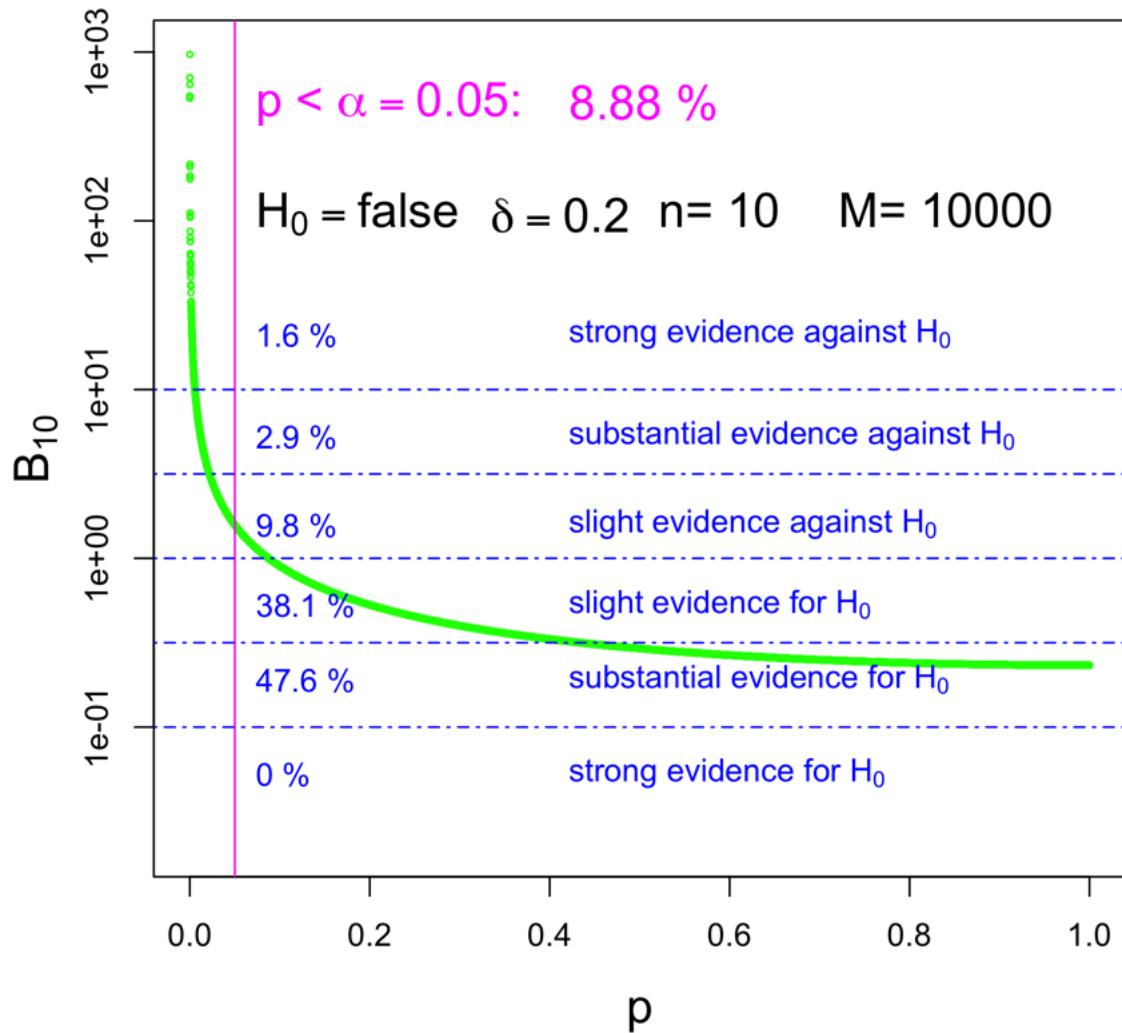


Figure 12.8: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 0.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. However, the true effect size  $\delta = 0.2$  is only slightly different from zero and, at least at small sample sizes, will often not stick out given the ‘noise level’  $\sigma = 1$ . The null hypothesis is only slightly more often rejected than the unavoidable level of  $100 \times \alpha = 5\%$ . The Bayesian results show also relative small frequencies for strong or substantial evidence against  $H_0$  and, compare to the  $\delta = 0$  case, a frequency reduction of substantial evidence for  $H_0$  (47.6% compared to 56.3%). [BayesianNHST-t-test.R](#) (line 9: change sflag to 2)

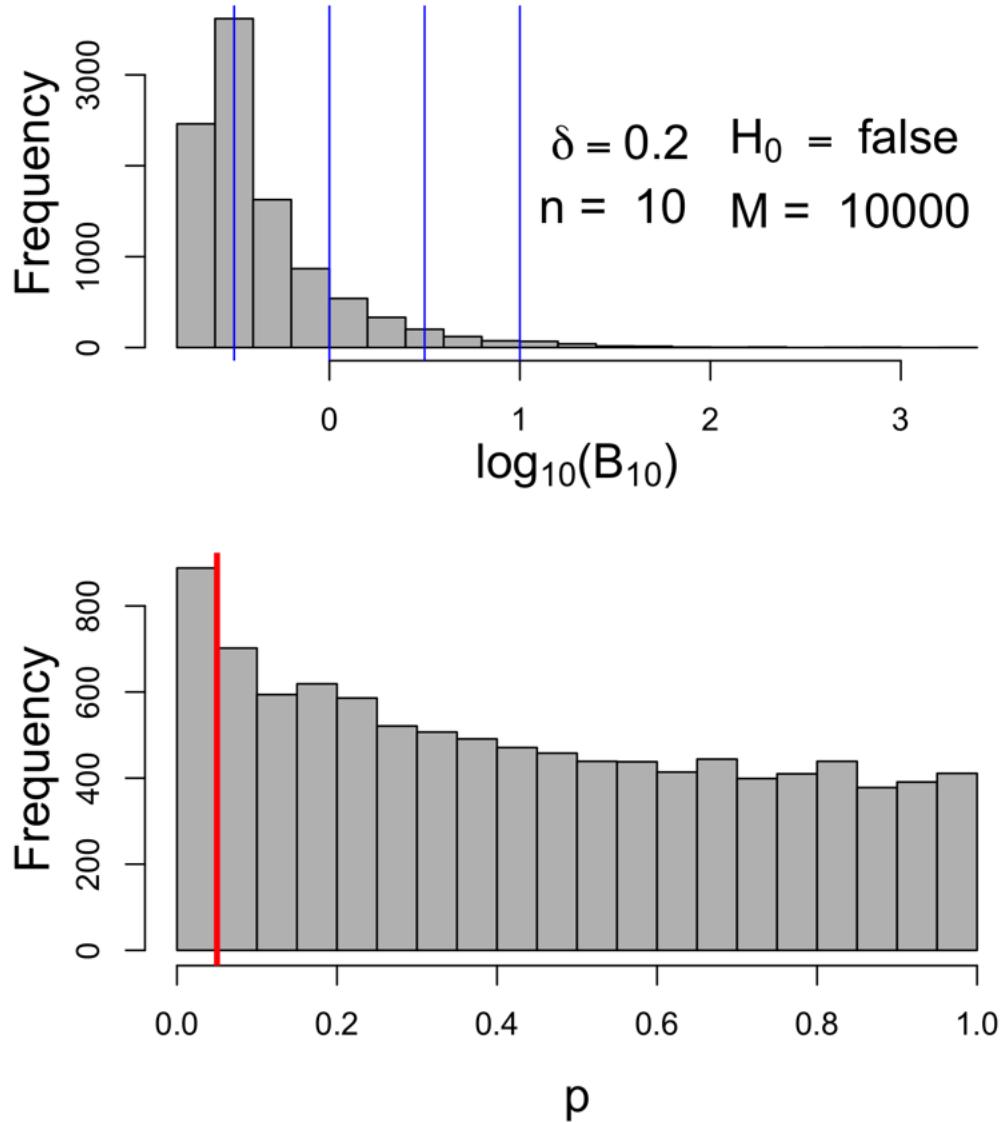


Figure 12.9: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 0.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. However, the true effect size  $\delta = 0.2$  is only slightly different from zero and, at least at small sample sizes, will often not stick out given the ‘noise level’  $\sigma = 1$ . The null hypothesis is only slightly more often rejected than the unavoidable level of  $100 \times \alpha = 5\%$ . The Bayesian results show also relative small frequencies for strong or substantial evidence against  $H_0$  and, compare to the  $\delta = 0$  case, a frequency reduction of substantial evidence for  $H_0$  (47.6% compared to 56.3%). [BayesianNHST-t-testAll.R](#) (line 8: change sflag to 2)

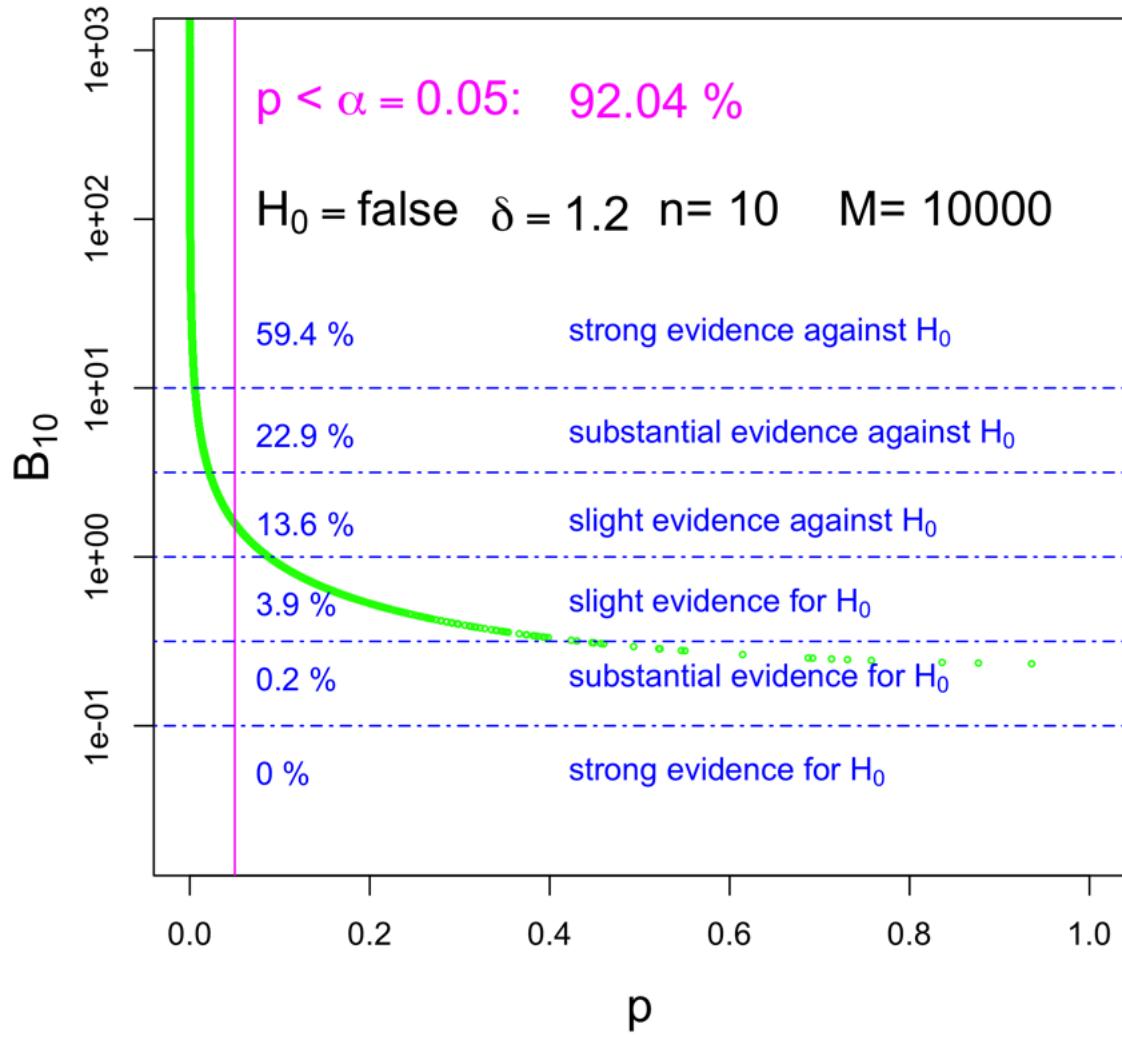


Figure 12.10: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 1.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. Indeed, the  $t$ -test suggests rejections based on  $p < \alpha = 0.05$  in 92% of all samples which is not bad for a small sample size. In the Bayesian approach there is strong evidence against  $H_0$  in almost 60% of all samples and substantial evidence for  $H_0$  in only 0.2%. [BayesianNHST-t-test.R](#) (line 9: change sflag to 3)

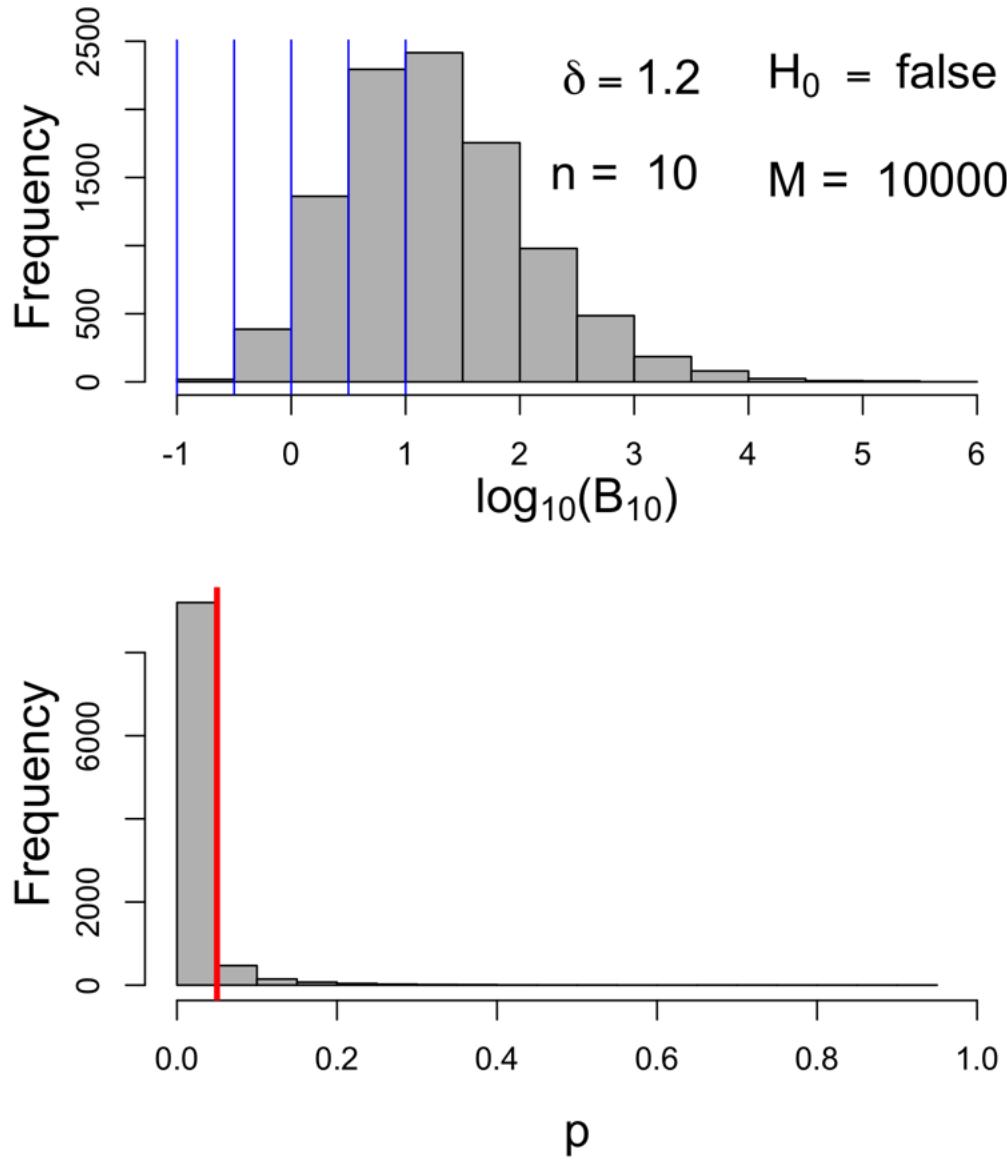


Figure 12.11: Results of Monte Carlo simulations for sample size  $n = 10$ ,  $\delta = 1.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. Indeed, the  $t$ -test suggests rejections based on  $p < \alpha = 0.05$  in 92% of all samples which is not bad for a small sample size. In the Bayesian approach there is strong evidence against  $H_0$  in almost 60% of all samples and substantial evidence for  $H_0$  in only 0.2%. [BayesianNHST-t-testAll.R](#) (line 8: change sflag to 3)

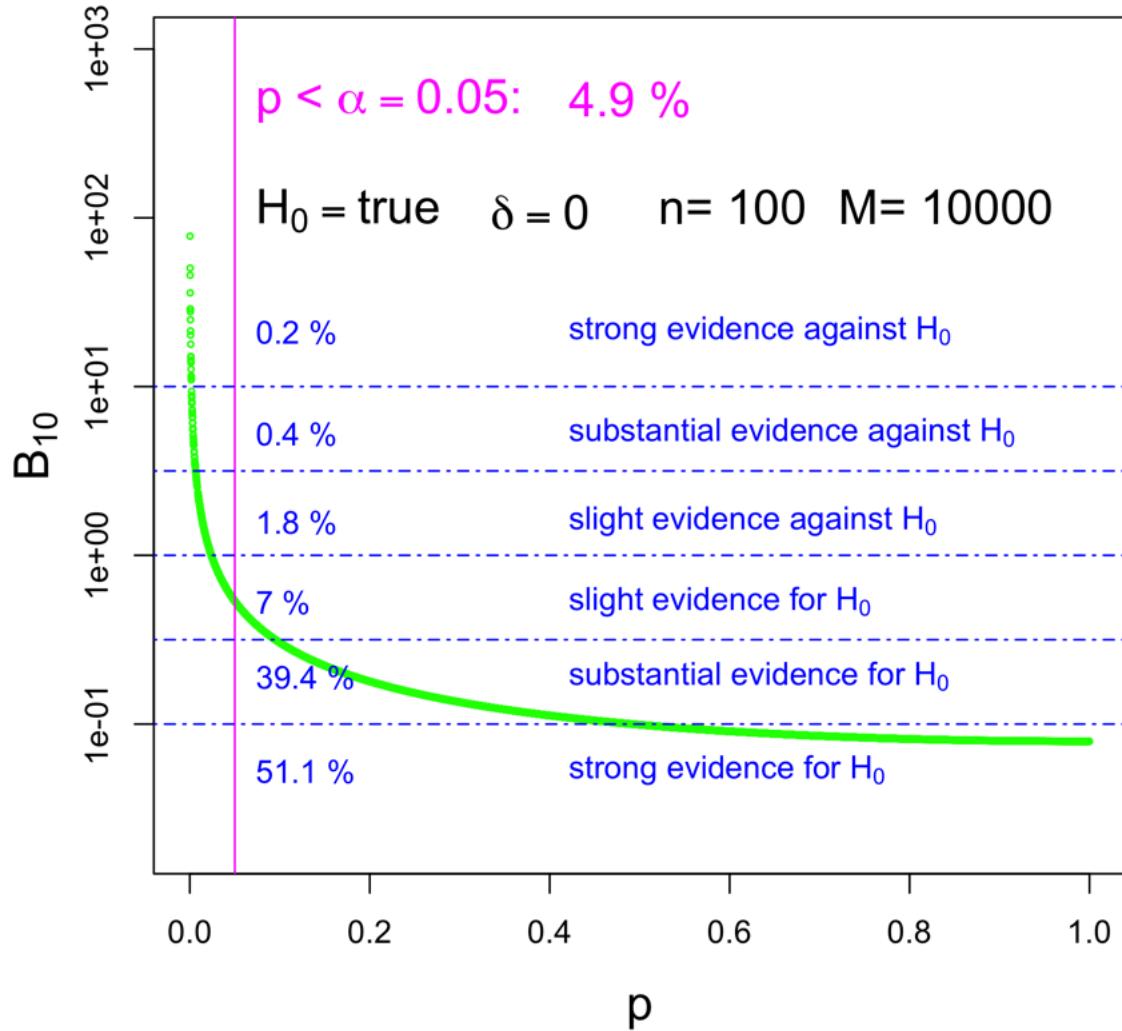


Figure 12.12: Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 0$ , i.e.  $H_0$  is true and one would expect that  $H_0$  is not rejected. However, in about 4.9% of all samples  $H_0$  is (falsely) rejected based on  $p$ -values smaller than  $\alpha = 0.05$ . I.e. even at higher sample size the Type I error ('reject  $H_0$  although it is true') stays at (about)  $100 \times \alpha = 5\%$ . In other words the  $t$ -test with fixed  $\alpha$  is inconsistent, i.e. it does not yield the correct results for large sample size (formally at  $n \rightarrow \infty$ ). The Bayesian approach results in 0.2% plus 0.4% = 0.6% strong or substantial evidence against  $H_0$  (smaller than for the  $n = 10$  case) and 39.4% and 51.1% substantial or strong evidence for  $H_0$  (larger than for the  $n = 10$  case). [BayesianNHST-t-test.R](#) (line 9: change sflag to 4)

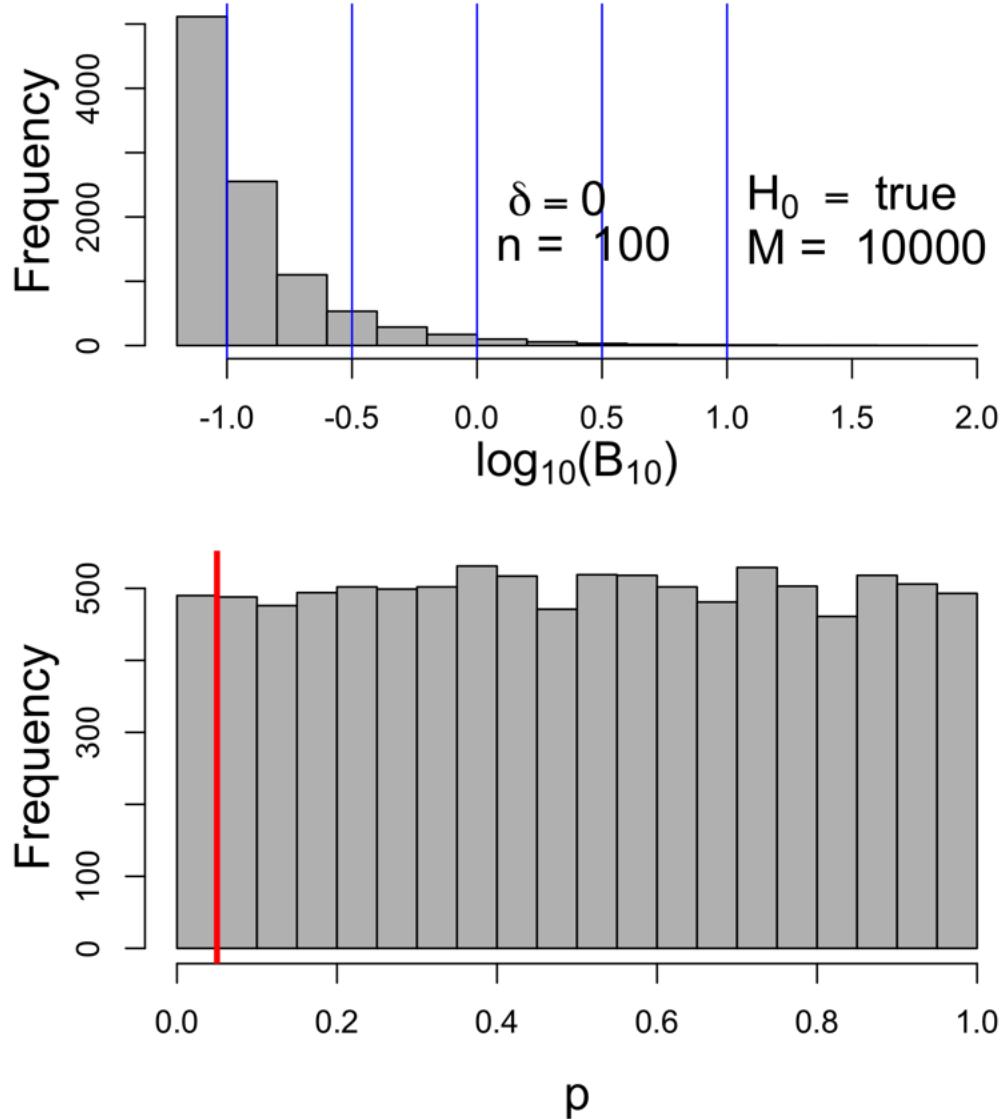


Figure 12.13: Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 0$ , i.e.  $H_0$  is true and one would expect that  $H_0$  is not rejected. However, in about 4.9% of all samples  $H_0$  is (falsely) rejected based on  $p$ -values smaller than  $\alpha = 0.05$ . I.e. even at higher sample size the Type I error ('reject  $H_0$  although it is true') stays at (about)  $100 \times \alpha = 5\%$ . In other words the  $t$ -test with fixed  $\alpha$  is inconsistent, i.e. it does not yields the correct results for large sample size (formally at  $n \rightarrow \infty$ ). The Bayesian approach results in 0.2% plus 0.4% = 0.6% strong or substantial evidence against  $H_0$  (smaller than for the  $n = 10$  case) and 39.4% and 51.1% substantial or strong evidence for  $H_0$  (larger than for the  $n = 10$  case). [BayesianNHST-t-testAll.R](#) (line 8: change sflag to 4)

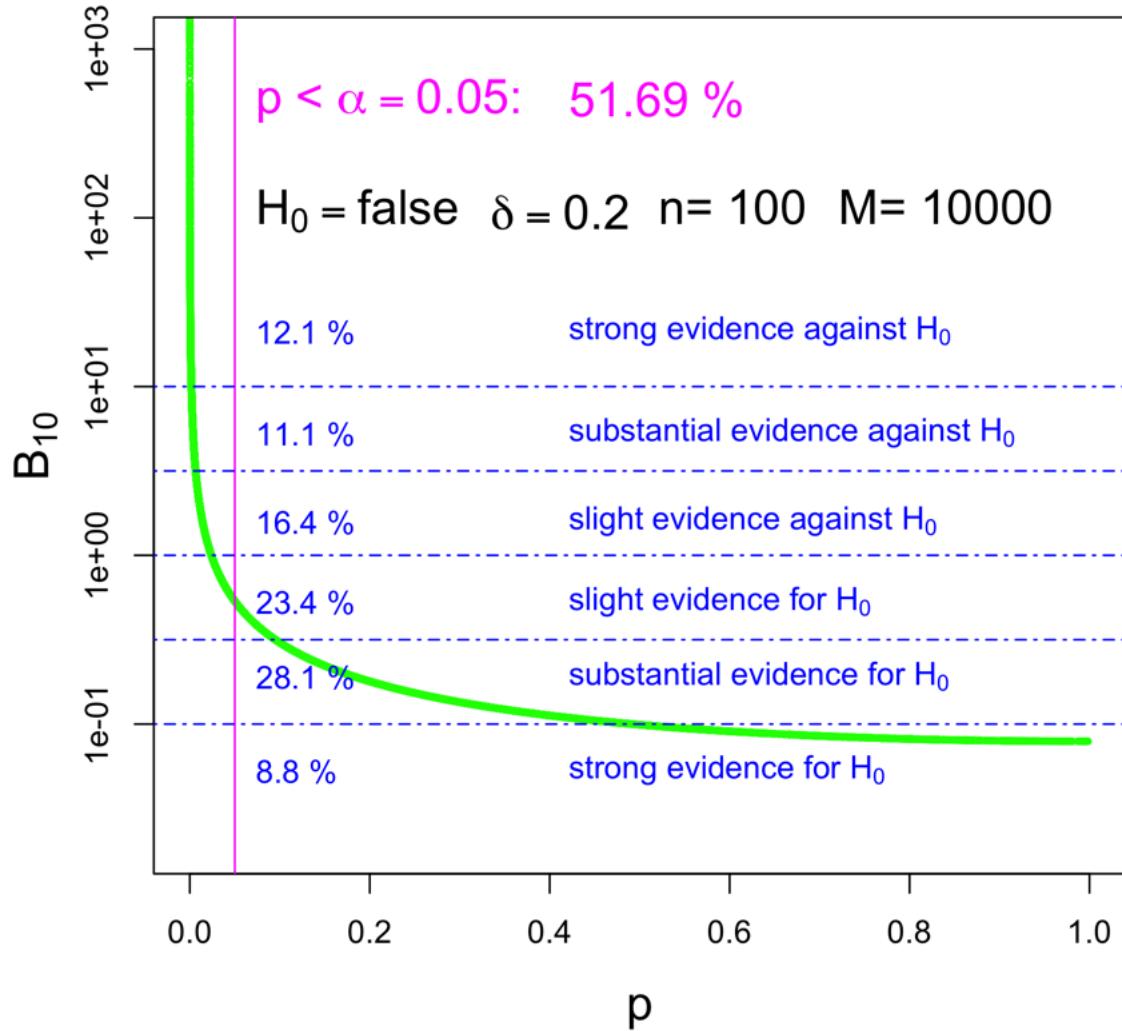


Figure 12.14: Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 0.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. However, the true effect size  $\delta = 0.2$  is only slightly different from zero and, at least at small sample sizes, will often not stick out given the 'noise level'  $\sigma = 1$ . The null hypothesis is rejected for about 52% of all samples for  $\alpha = 0.05$ . The Bayesian results show larger frequencies for strong or substantial evidence against  $H_0$  compared to the  $\delta = 0$  case, a large frequency reduction of strong evidence for  $H_0$  (8.8% compared to 51.1%). However, the strong evidence for  $H_0$  (8.8%) is in stark contrast to 0% in case of  $n = 10$ . [BayesianNHST-t-test.R](#) (line 9: change sflag to 5)

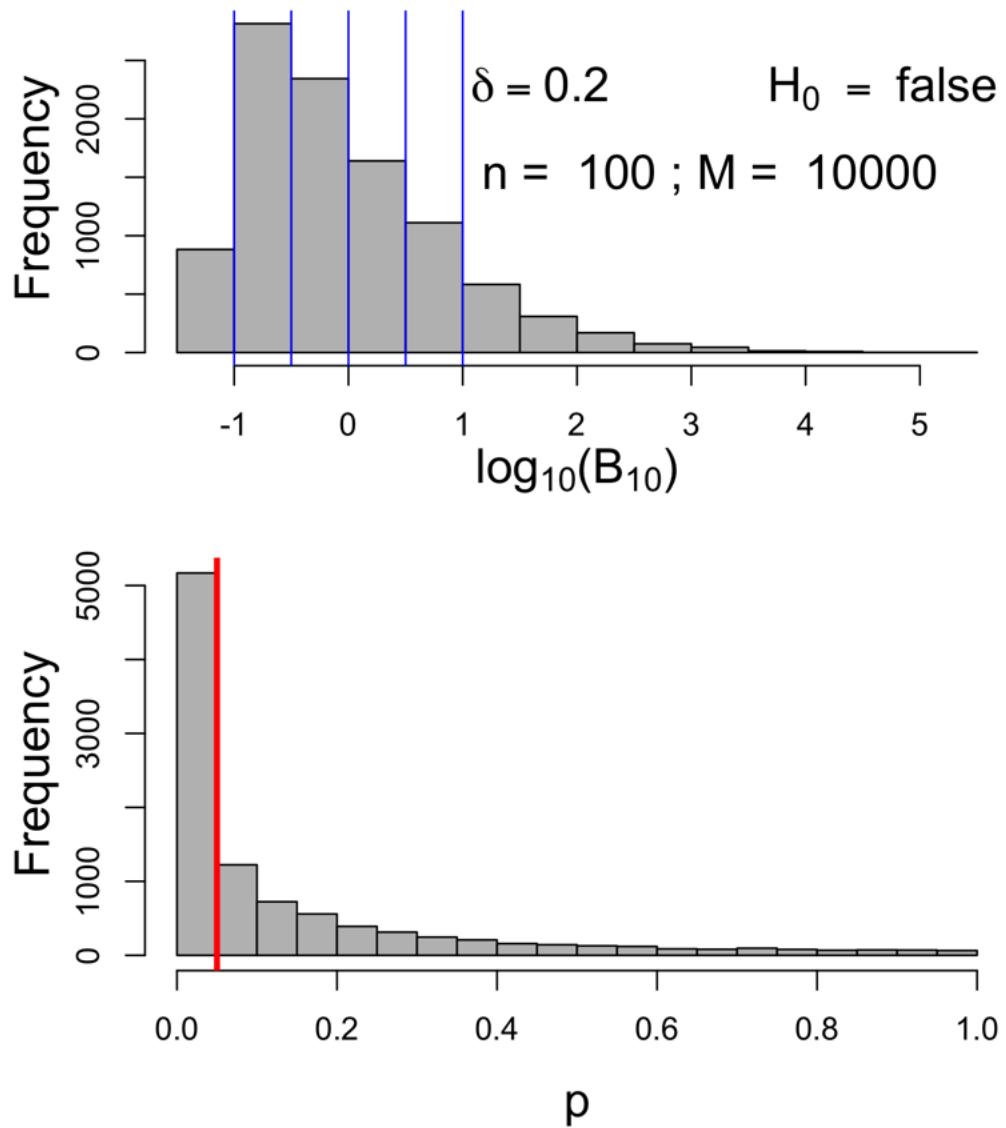


Figure 12.15: EEE Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 0.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. However, the true effect size  $\delta = 0.2$  is only slightly different from zero and, at least at small sample sizes, will often not stick out given the 'noise level'  $\sigma = 1$ . The null hypothesis is rejected for about 52% of all samples for  $\alpha = 0.05$ . The Bayesian results show larger frequencies for strong or substantial evidence against  $H_0$  compared to the  $\delta = 0$  case, a large frequency reduction of strong evidence for  $H_0$  (8.8% compared to 51.1%). However, the strong evidence for  $H_0$  (8.8%) is in stark contrast to 0% in case of  $n = 10$ . [BayesianNHST-t-testAll.R](#) (line 8: change sflag to 5)

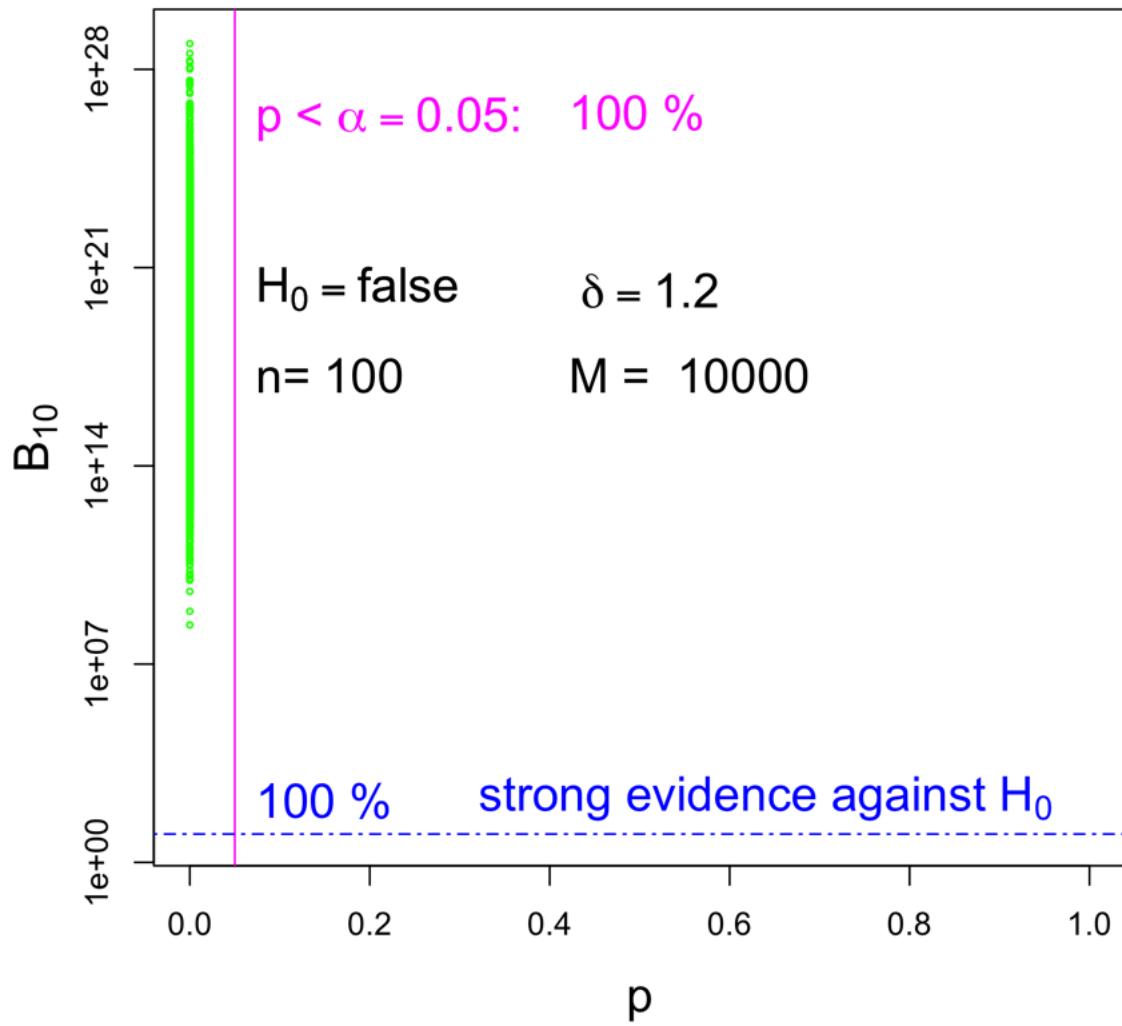


Figure 12.16: Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 1.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. Indeed, the  $t$ -test suggests rejections based on  $p < \alpha = 0.05$  in 100% of all samples (based on  $\alpha = 0.05$ ). In the Bayesian approach there is strong evidence against  $H_0$  in almost 100% of all samples. In summary: with a sample size  $n = 100$  a blind man (from either statistical school) can recognize that an effect size of 1.2 is different from zero. Jeffreys' quotation (1961, p. 393) "As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached." fits 100% to this case. [BayesianNHST-t-test.R](#) (line 9: change sflag to 6)

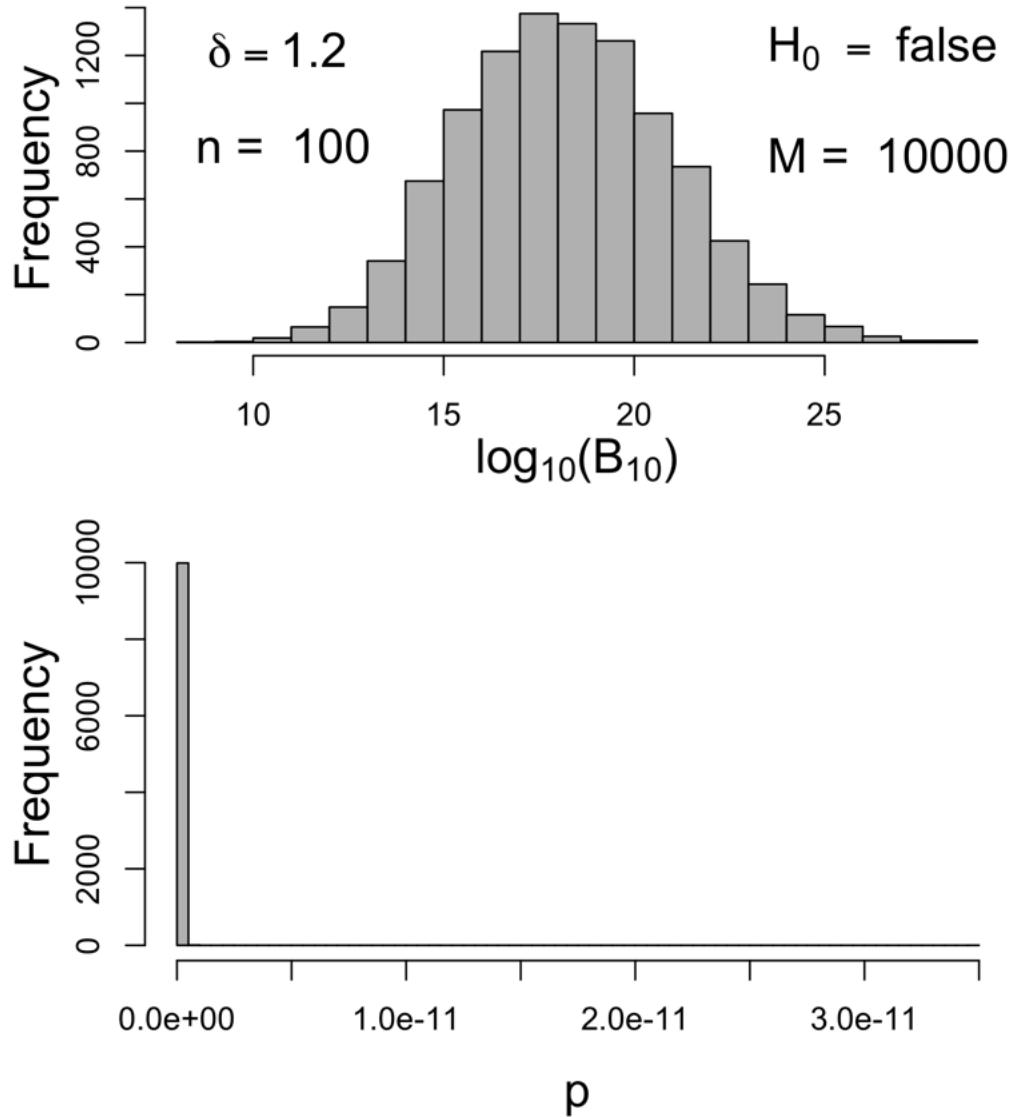


Figure 12.17: Results of Monte Carlo simulations for sample size  $n = 100$ ,  $\delta = 1.2$ , i.e.  $H_0$  is false and one would expect that  $H_0$  is rejected. Indeed, the  $t$ -test suggests rejections based on  $p < \alpha = 0.05$  in 100% of all samples (based on  $\alpha = 0.05$ ). In the Bayesian approach there is strong evidence against  $H_0$  in almost 100% of all samples. In summary: with a sample size  $n = 100$  a blind man (from either statistical school) can recognize that an effect size of 1.2 is different from zero. Jeffreys' quotation (1961, p. 393) "As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached." fits 100% to this case. [BayesianNHST-t-testAll.R](#) (line 8: change sflag to 6)

**In summary**, there is a lot of agreement when it comes to rejecting a false null hypothesis  $H_0$  (either by a significant  $p$ -value or a large Bayes factor). When the null hypothesis is true, the percentage of significant  $p$ -values is higher than the percentage of substantial plus strong evidence against  $H_0$ . A Monte Carlo simulation (Fig. 12.18,  $M = 10^5$  runs for each  $n$ ) shows that the percentage of significant  $p$ -values does not vary with sample size and thus, for  $\alpha = 0.05$ , even at very large sample sizes 5% of true null hypotheses will be rejected. The substantial or strong evidence against the true null hypothesis is smaller than 5% and decreases with sample size.

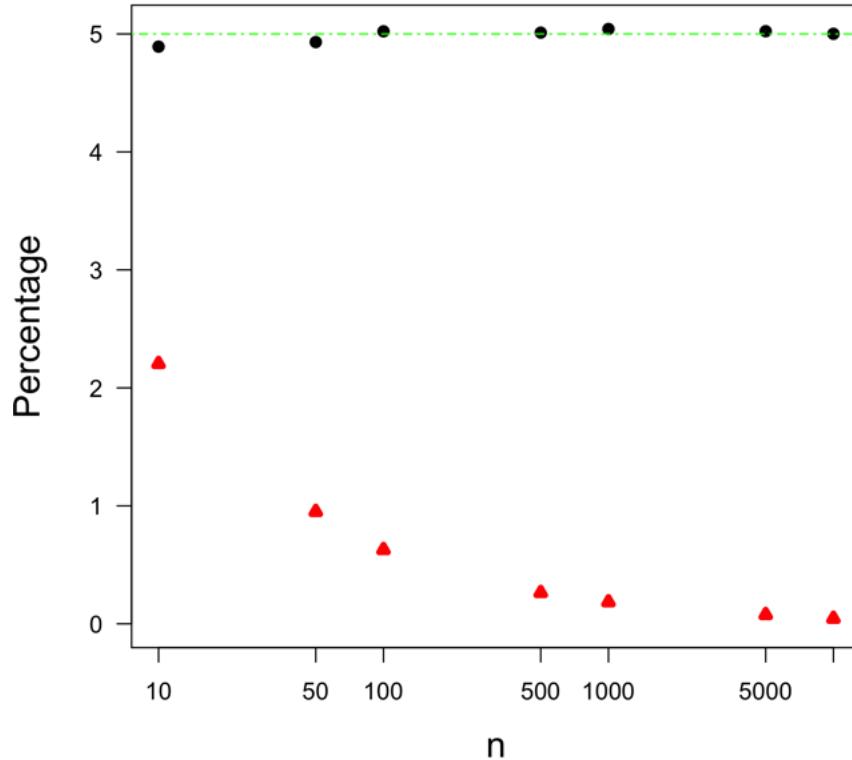


Figure 12.18: Percentages of significant  $p$ -values ( $p < \alpha = 0.05$ ; 1-sample 2-sided  $t$ -test; blue dots) or large Bayes factors ( $B_{10} > 3.16$ ; 1-sample 2-sided Bayesian  $t$ -test; red triangles) based on  $M = 10^5$  Monte Carlo simulations with random samples from the standard normal PDF with sample size  $n$ . The percentage of significant  $p$ -values does not vary with sample size  $n$  and thus even at very large sample sizes 5% of true null hypotheses will be rejected. The substantial or strong evidence against the true null hypothesis is smaller than 5% and decreases with sample size. [RejectTrueH0.R](#)

$p & B_{10}$	evidence, significant	$\delta = 0$	$\delta = 0.2$	$\delta = 1.2$	$\delta = 0$	$\delta = 0.2$	$\delta = 1.2$
		$n = 10$	$n = 10$	$n = 10$	$n = 100$	$n = 100$	$n = 100$
$p < 0.05$	significant ('reject $H_0$ ')	5.1%	8.9%	92.0%	4.9 %	51.7%	100%
$B_{10} > 10$	strong evidence against $H_0$	0.6%	1.6%	59.4%	0.2%	12.1%	100%
$3.16 < B_{10} < 10$	substantial evidence against $H_0$	1.9%	2.9%	22.9%	0.4%	11.1%	0.0%
$1 < B_{10} < 3.16$	slight evidence against $H_0$	6.2%	9.8%	13.6%	1.8%	16.4%	0.0%
$0.316 < B_{10} < 1$	slight evidence against $H_1$	35.0%	38.1%	3.9%	7.0%	23.4%	0.0%
$0.1 < B_{10} < 0.316$	substantial evidence against $H_1$	56.3%	47.6%	0.2%	39.4%	28.1%	0.0%
$B_{10} < 0.1$	strong evidence against $H_1$	0.0%	0.0%	0.0%	51.1%	8.8%	0.0%

Table 12.1: Results of Monte Carlo simulations ( $M = 10^4$  runs): we sample from normal populations with  $\sigma = 1$  and  $\mu = 0, 0.2$ , or  $1.2$ , i.e. the effect size  $\delta = \mu/\sigma$  is identical to the mean  $\mu$ .

## 12.2 Which test to apply for which question or hypothesis?

*A list of commonly asked questions is given. The choice of an appropriate test depends on the number of samples involved<sup>12</sup> and the assumptions about the statistical populations from which one samples.*

*Please note that 'in reality' (and also often in articles or even text books written by professional statisticians!) some tests are applied even if the prerequisites about the statistical populations are obviously violated (for example, the t-test is based on the assumption of normally distributed populations, however, it is often also applied to samples from discrete distributions like count data) and the decisions based on the resulting p-values are sensible or not, depending on the 'robustness' of the test and the actual distribution of the population from which the sample has been taken. The robustness can be tested for various distributions by Monte Carlo simulations.*

*Tests should be applied and results reported when it is not obvious which decision to take, i.e. when the p-value is close to the (chosen) level of significance  $\alpha$ . For the commonly used  $\alpha = 0.05$  this means that values  $p > 0.5$  (do not reject  $H_0$ ) or  $p < 10^{-4}$  (reject  $H_0$ ) indicate situations where 'a blind man can see' how to decide<sup>13</sup>.*

*If applying less commonly used tests please report the null hypothesis and give an appropriate reference to the test.<sup>14</sup> Bayesian tests for a limited number of questions have been already been developed by Jeffreys (1939, 1961). In recent time, Bayesian tests for more questions have been developed by various authors. For Bayesian t-tests, ANOVA etc. the R package **BayesFactor** became available in 2018.*

1. Equal means? Larger? Smaller?
  - (a) t-test: one or two samples from normal distributions with equal variances
  - (b) Welch t-test: two samples from normal distributions with different variances
  - (c) ANOVA: more than two samples from normal distributions with equal variances
  - (d) Tests for non-normal distributions can be devised using Monte Carlo simulations (an example is discussed in Section H.3.4).
2. Equal central tendency? For populations following non-normal distributions
  - (a) One-sample Wilcoxon test
  - (b) Two-sample Wilcoxon-Mann-Whitney test (Mann-Whitney test, U-test): two samples from non-normal (unknown) distributions
  - (c) Kruskall-Wallis test: more than two samples
3. Equal variances?
  - (a) Variance ratio test (F-test): two samples from normal distributions
  - (b) Bartlett test: more than two samples from normal distributions
  - (c) Levene's test: two samples from non-normal distributions
4. Slope of a straight line fitted to data different from zero? One-sample t-test!
5. Sample from a specified distribution? Goodness-of-fit tests
  - (a) Shapiro-Wilk test: sample from a normal distribution
  - (b) Kolmogorov-Smirnov test (KS-test): one sample from specified distribution (for example, standard normal distribution)
  - (c) Lilliefors test: one sample from normal distribution with unknown variance (Section 12.4.4)

<sup>12</sup>t-tests can be applied for one or two samples only, whereas ANOVA is required for more than two samples. An analog distinction is necessary for the variance ratio test (two samples) and the Bartlett test (more than two samples).

<sup>13</sup>I have seen an embarrassing small p-value of  $6 \times 10^{-93}$  in an article published in *Nature* (2016) shedding light on the attention to statistics by the authors, reviewers, and editors.

<sup>14</sup>There are hundreds (or thousands?) of tests described in the literature and one cannot expect that readers of your article are familiar with all of them.

6. Samples from different distributions?
  - (a) Kolmogorov-Smirnov test (KS-test): 2-sample KS-test
  - (b) Anderson-Darling test
  - (c) Chi-squared test ( $\chi^2$  test)

## 12.3 Equal means? More than two samples: ANOVA

When we like to compare more than two mean values, the t-test can not be applied anymore<sup>15</sup>. ANOVA (ANalysis Of VAriance), which is assuming random samples from normal populations, can be considered as a generalization of the t-test.

### 12.3.1 ANOVA: NHST

In this section we run [ANOVA-Zar10Ex10d1.R](#), producing a box plot and performing an ANOVA test on the data provided in Table 12.2.

Data #	Group 1	Group 2	Group 3	Group 4
1	60.8	68.7	69.6	61.9
2	67.0	67.7	77.1	64.2
3	65.0	75.0	75.2	63.1
4	68.6	73.3	71.5	66.7
5	61.7	71.8		60.3

Table 12.2: Zar (2010) Examples 10.1: weight of animals (kg).

Music: 4 your love (The Yardbirds) <https://www.youtube.com/watch?v=Z2LSSgQMc2E>

---

<sup>15</sup>Naively one could come up with the idea to apply the t-test for all combinations of the samples (i.e. for 3 samples X, Y, and Z one would t-test XY, YZ, and XZ) and reject  $H_0$  if one of the t-test indicates rejection. However, this approach is not working properly (see discussion in Zar, 2010, for a detailed explanation).

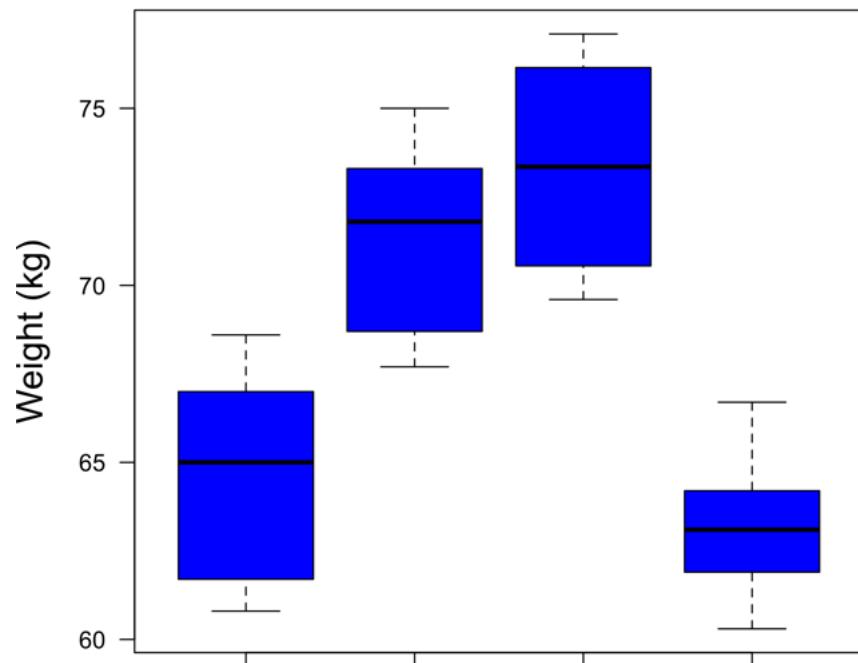


Figure 12.19: Look at 'your' data: box plot of the data of Zar (2010, Example 10.1). One could guess that the null hypothesis 'all true mean values are equal' is most probably false and thus should be rejected.

[ANOVA-Zar10Ex10d1.R](#)

### Interpretation of output<sup>16</sup> & decision

1. 'Df<sub>group</sub> = 3': groups degrees of freedom = number of groups (4) - 1 = 3
2. 'Df<sub>Residuals</sub> = 15': within-groups degrees of freedom = number of data (19) - number of groups (4) = 15
3. 'Sum Sq<sub>group</sub> = 338.9': groups sum of squares =  $\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$  where  $n_i$  is the number of data in group  $i$  (here: 5,5,4,5),  $\bar{X}$  is the mean over all data,  $\bar{X}_i$  is the sample mean for group  $i$ . Except for scaling (divide by  $N - 1$ ) the groups sum of squares is a **variance**.
4. 'Sum Sq<sub>Residuals</sub> = 140.8': within-groups sum of squares = total sum of squares - groups sum of squares
5. 'Mean Sq<sub>group</sub> = 112.98' = groups sum of squares / groups degrees of freedom
6. 'Mean Sq<sub>Residuals</sub> = 9.38' = within-groups sum of squares / within-groups degrees of freedom
7. 'F value = 12.04': test statistic  $F$  = groups mean squared deviation from the mean / error mean squared deviation from the mean
8. ' $\text{Pr}(>F) = 0.000283$ ' = observed level of significance (one-sided p-value) = probability to observe ' $F = 12.04$  or larger values'
9. '\*\*\*' = highly significant ( $p < 10^{-3}$ )
10. Null hypothesis = all true mean values are equal, i.e. here  $\mu_1 = \mu_2 = \mu_3 = \mu_4$
11.  $p = 0.000283 < 0.05 = \alpha \Rightarrow$  reject the null hypothesis
12. The 'observed' (i.e. calculated from the data)  $F$  value of 12.04 is deep in the rejection region for  $\alpha = 0.05$  (Fig. 12.20)  $\Rightarrow$  reject the null hypothesis
13. The 'observed' (i.e. calculated from the data)  $F$  value of 12.04 is (much!) larger than the critical value  $F_{\alpha=0.05, v_1=3, v_2=15} \approx 3.287 \Rightarrow$  reject the null hypothesis
14. The two ways for rejecting the null hypothesis are equivalent to each other.

<sup>16</sup>For the mathematical definition of various terms compare Zar (2010).

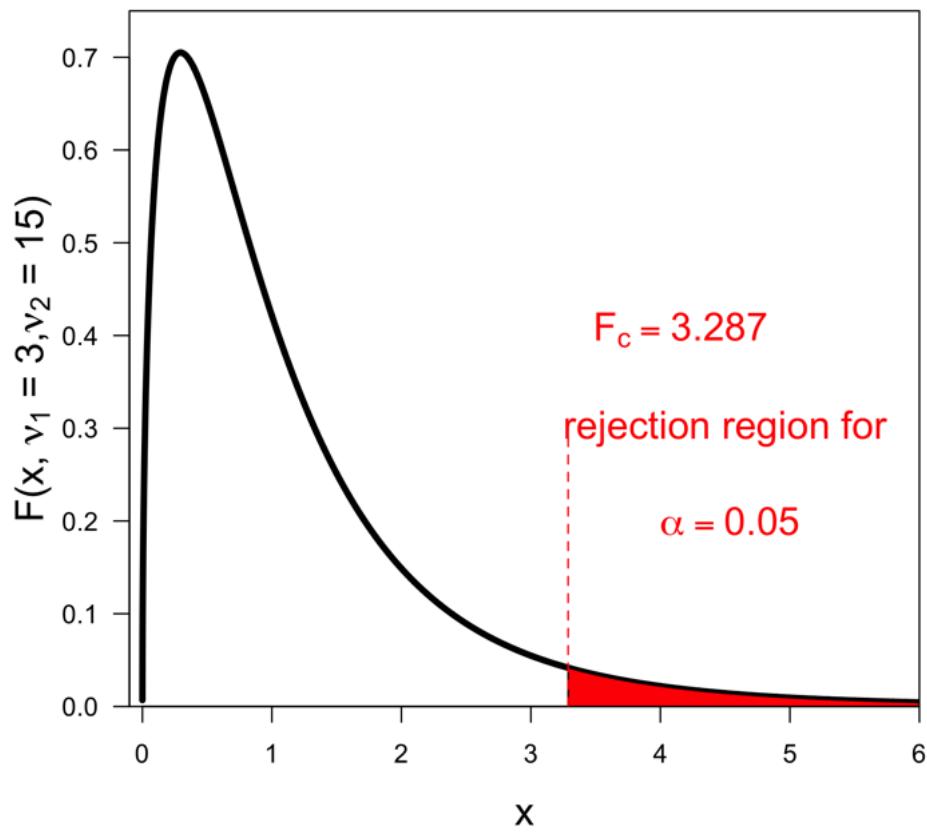


Figure 12.20: Rejection region (red area) for the  $F$ -distribution  $F(x, \nu_1 = 3, \nu_2 = 15)$  and chosen level of significance  $\alpha = 0.05$  = region beyond the critical value  $F_c = F_{\alpha=0.05, \nu_1=3, \nu_2=15} \approx 3.287$  [ANOVArejectionRegZarEx.R](#)

### 12.3.2 ANOVA: Bayesian

A Bayesian version of ANOVA has been developed by Rouder et al. (2012). The R routine `generalTestBF()`<sup>17</sup> for this test can be found in the package `BayesFactor`. The Bayes factor  $B_{10} = 90.5$  provides strong evidence against  $H_0$  and strong evidence for  $H_1$ . R code: [ANOVA-Bayesian.R](#)

**Exercise 38** Is  $F(x, \nu_1 = 3, \nu_2 = 15)$  equal to  $F(x, \nu_1 = 15, \nu_2 = 3)$ ?

Plot  $F(x, \nu_1 = 3, \nu_2 = 15)$  and  $F(x, \nu_1 = 15, \nu_2 = 3)$  for  $0 \leq x \leq 10$ .

---

<sup>17</sup>The output of `generalTestBF` provides the Bayes factor  $B_{10}$

## 12.4 Sample from a specified distribution? Normality tests etc.

Many statistical methods require normal distributions and thus it is of interest to test whether or not a sample is from a normally distributed population. For 'testing normality' the Shapiro test is recommended.

Various other tests have been developed for this and other purposes, including the Kolmogorov-Smirnov (KS for short) and the Lillifors test. The main difference between the KS and the Lillifors test is the requirement of knowledge of the true population variance  $\sigma^2$  for the KS test. Because  $\sigma^2$  is usually not known (except for textbook examples), it has to be estimated from the data. The KS and the Lillifors test can also be applied to other populations and to compare two samples.

### 12.4.1 Sample from normal distribution? The Shapiro-Wilk test

Shapiro and Wilk (1965) developed a test that can address the null hypothesis ‘the sample  $x$  stems from a normal distribution’. In contrast to the KS and the Lilliefors test it can not be applied to other distributions because its test statistic  $W$  uses specific properties of the standard normal distribution.<sup>18</sup> For the three samples  $y_1, y_2, y_3$  (given in the R script below) one obtains the  $p$ -values 0.485,  $6.2 \cdot 10^{-8}$ , 0.0121, respectively. The Shapiro-Wilk test can be applied already for small sample sizes  $n$  (implemented in R for  $n = 3$  to 5000).

Test	y1	y2	y3
KS	0.938	$4 \cdot 10^{-3}$	0.1355
Lilliefors	0.736	$3 \cdot 10^{-8}$	0.0022
Shapiro-Wilk	0.485	$6 \cdot 10^{-8}$	0.0121

Table 12.3: Observed levels of significance (‘p-values’) for the null hypothesis ‘the sample  $y_k$  stems from a normal distribution’ for the KS, the Lilliefors, and the Shapiro-Wilk test.

A few remarks are in order:

1. One parameter in the call of the R routine `ks.test()` is the CDF which for the normal distribution is `pnorm()`: when calling `ks.test(y1,'pnorm')` one assumes that the sample  $y_1$  is from the *standard* normal distribution.
2. If one would like to test  $H_0 = \text{‘sample } y_1 \text{ is from a normal distribution’}$  using the KS test one has to either standardize the sample or specify (estimates of) the parameters of the normal distribution:
  - a standardize  $y_1$ :  $y_{1s} = (y_1 - \text{mean}(y_1)) / \text{sd}(y_1)$ , then call `ks.test(y1s,'pnorm')` or
  - b call `ks.test(y1,'pnorm',mean(y1),sd(y1))`

The calls `ks.test(y1s,'pnorm')` and `ks.test(y1,'pnorm',mean(y1),sd(y1))` yield identical  $p$ -values.

3. For sample  $y_1$  the calls of `ks.test(y1,'pnorm')` and `ks.test(y1,'pnorm',mean(y1),sd(y1))` yield slightly different  $p$ -values. This kind of difference becomes even more obvious when we add 1.3 to all data in  $y_1$  (‘a shift’). Now the calls of `ks.test(y1,'pnorm')` and `ks.test(y1,'pnorm',mean(y1),sd(y1))` yield largely different  $p$ -values that would lead to different conclusions (reject  $H_0$  in the first case, despite the fact that the shifted  $y_1$  is from a normal PDF).
4. The R routine for the Lilliefors test is `lillieTest()`, is can be found in the package `fBasics`, it tests (by default)  $H_0 = \text{‘sample is from a normal distribution’}$ , and requires no additional information about the CDF. Thus the call is simply `lillieTest(y1)` (call of `lillieTest(y1s)` yields identical  $p$ -value).
5. Note that the output object of `lillieTest()` is `class S4` and thus one has to use the signs @ and \$ to access the  $p$ -value: `LF1@test$p.value`.
6. The R routine for the Shapiro-Wilk test is `shapiro.test()`. It tests  $H_0 = \text{‘sample is from a normal distribution’}$  and yields identical results for original and standardized data.
7. The  $p$ -values for single samples differ between tests. For a chosen level of significance  $\alpha = 0.05$  this would lead to the same decisions for samples  $y_1$  (do not reject  $H_0$ ) and  $y_2$  (reject  $H_0$ ), however, for sample  $y_3$  based on the Lilliefors and Shapiro-Wilk tests one would reject  $H_0$  whereas the results of the KS-test would suggest that  $H_0$  is not rejected at the significance level  $\alpha = 0.05$ . The result of the KS-test is less reliable because instead of the true variance (as required) we gave an estimate only.

<sup>18</sup>I do not know whether similar tests can be developed based on other distributions. The associated test statistics would look different from the one used here.

### 12.4.2 R codes: Shapiro-Wilk, KS, and Lilliefors test

R code: [Shapiro-WilkExamples.R](#)

### 12.4.3 Kolmogorov-Smirnov test

*The basic idea of the Kolmogorov-Smirnov test (KS test for short) is to compare cumulative probability distribution functions (CDFs), i.e. in the current application the CDF of the normal distribution and the estimated CDF from the sample (staircase, compare Fig. 12.21), and to use the maximal difference D between these two CDFs as test statistic. The observed level of significance, p, can then be calculated from the PDF of the test statistic D.*

First, let us consider several examples. We will generate random samples  $x$  of size  $n = 30$  from various statistical populations. The level of significance is chosen as  $\alpha = 0.05$ . The null hypothesis will always be  $H_0 = \text{'sample is from the standard normal distribution'}$ . The KS-test is performed by calling the R routine **ks.test**.

1. Take random sample of size  $n = 30$  from the standard normal distribution & apply KS-test (Fig. 12.21):  
`set.seed(1953); n = 30; x = rnorm(n); ks.test(x,'pnorm')`  $\Rightarrow D = 0.1473, p = 0.4882$ , do not reject  $H_0$  (no surprise!).
2. Take random sample of size  $n = 30$  from the non-standard normal distribution with  $\mu = 1$  and  $\sigma = 2$  & apply KS-test:  
`x = rnorm(n,1,2); ks.test(x,'pnorm')`  $\Rightarrow D = 0.3947, p = 10^{-4}$ . The tiny  $p$ -value would suggest rejecting  $H_0$ . Although  $x$  stems from a normal distribution, it's not from the standard normal distribution!
3. If one is interested in the modified  $H_0$  'x stems from a (any!) normal distribution', one has to modify the call of the KS routine: `ks.test(x,'pnorm',mean(x),sd(x))` by including the sample mean and the sample standard deviation (estimates of  $\mu$  and  $\sigma$ ) in the parameter list of `ks.test()`  $\Rightarrow D = 0.1149, p = 0.7812$ , do not reject the modified  $H_0$ . This is what is usually called a '**normality test**', i.e. one asks for 'normality' which means 'any normal distribution'.
4. Take random sample of size  $n = 30$  from the standard uniform PDF & apply KS-test assuming a normal PDF: one would expect that  $H_0$  'x stems from a normal PDF' is rejected because it is false. `x = runif(n); ks.test(x,'pnorm',mean(x),sd(x))`  $\Rightarrow D = 0.1043, p = 0.8663$ , do not reject  $H_0$  (surprise!). Why is the false  $H_0$  not rejected? The answer is given in the legend of Fig. 12.23.
5. Take the same random sample as before & apply KS-test assuming the standard uniform PDF: `ks.test(x,'punif')`  $\Rightarrow D = 0.1182, p = 0.7528$ , do not reject  $H_0$  (no surprise). Actually, the maximal absolute difference to the standard uniform CDF (0.1182) is slightly larger than the difference to the normal CDF considered before (0.1043) and thus the  $p$ -value for comparison with the standard uniform CDF (0.7528) is slightly smaller than that the normal CDF considered before (0.8663): this is caused by randomness; both  $p$ -values are much higher than  $\alpha = 0.05$  and thus the conclusions ('do not reject  $H_0$ ') are the same. This example shows, however, that not rejecting  $H_0$  does not mean that  $H_0$  is true!

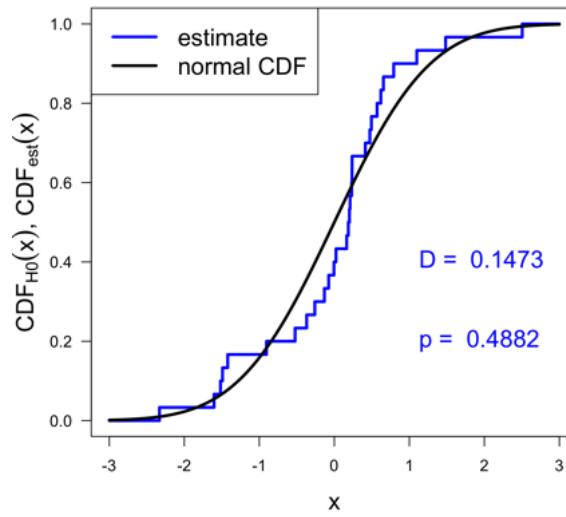


Figure 12.21: Estimated CDF for sample  $x$  from the standard normal distribution (blue 'staircase') and CDF of the standard normal distribution (black line). The maximal absolute difference between these two CDFs is  $D = 0.1473$ . Application of the KS-test yields  $p = 0.4882$ . The null hypothesis  $H_0$  ('sample stems from standard normal distribution') is not rejected ( $p > \alpha = 0.05$ ). This is no surprise!

[KStestNormal.R](#) [plotstaircaseKS.R](#)

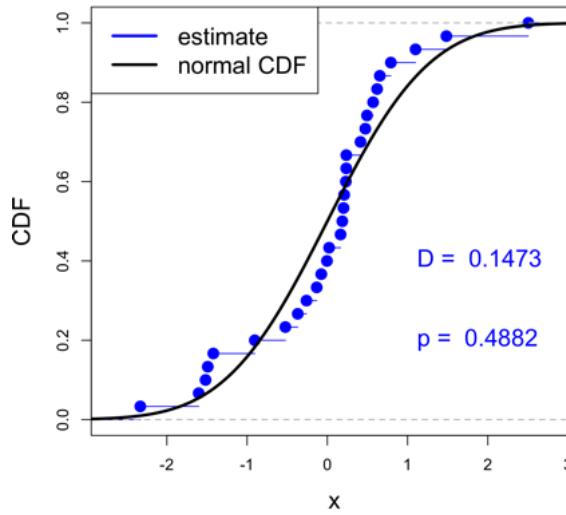


Figure 12.22: Estimated CDF for sample  $x$  from the standard normal distribution (blue 'staircase'; estimated using `ecdf()`) and CDF of the standard normal distribution (black line). The maximal absolute difference between these two CDFs is  $D = 0.1473$ . Application of the KS-test yields  $p = 0.4882$ . The null hypothesis  $H_0$  ('sample stems from standard normal distribution') is not rejected ( $p > \alpha = 0.05$ ). This is no surprise!

[KStestNormal.R](#) (line 19: set `sflag` to 3)

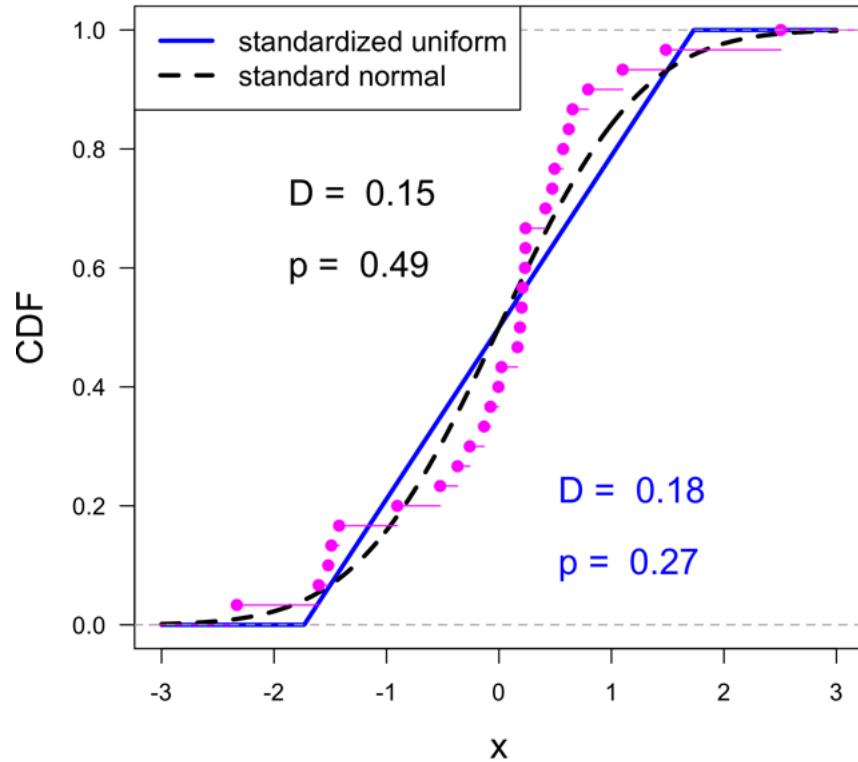


Figure 12.23: CDFs of the standardized uniform PDF (blue solid line; nonzero for  $-\sqrt{3} < x < \sqrt{3}$ ; Section 6.4.2) and the standard normal PDF (black broken line) with identical mean and standard deviation, i.e.  $\mu = 0$ ,  $\sigma = 1$ , and an estimated CDF for a sample of size  $n = 30$  from the standard normal PDF (magenta staircase). The maximal absolute difference between the two CDFs is  $D = 0.0572$  only and thus it is no surprise that both null hypotheses (1)  $H_0^N$  sample stems from the standard normal PDF (KS-test:  $D = 0.15$ ,  $p = 0.49$ ) and (2)  $H_0^U$  sample stems from the standardized uniform PDF (KS-test:  $D = 0.18$ ,  $p = 0.27$ ) can not be rejected on the chosen significance level of  $\alpha = 0.05$  ('too close to call'). [KSunifNormal.R](#)

**Exercise 39 Samples from normal distributions?**

Is any one of the three samples given below not from a normal distribution?

```
y1 = c(-0.68644781, -0.82379154, -0.98416919, -2.02230891, -0.43507791, -0.76655674,  
      1.22178443, 0.09767100, -0.93391714, -1.23458941, 0.09188711, 0.56736177,  
      -0.55276453, -0.07969400, 0.11767092, 2.07541230, 1.76443875, 0.60249792,  
      -1.29916116, -0.30322121, -0.77935252, -0.97190317, 0.84580262, 0.28698246,  
      1.15160104, 0.35533328, 0.32936546, 1.68584964, 0.18260973, 1.93600509)  
y2 = c(-0.43677762, -0.25606172, -0.05539436, -0.35955890, 2.52475885, -0.22159314,  
      -0.17360687, -0.24747086, -0.25016363, -0.39612793, -0.22807506, -0.35896958,  
      -0.23775192, 0.34039082, 1.31051033, -0.05988203, -0.15271093, 0.22672744,  
      -0.15967543, -0.13572103, -0.44215444, -0.40168018, 0.44666595, -0.53978945,  
      -0.45903502, -0.47862169, -0.22521386, -0.38695104, -0.44321571, -0.28632236)  
y3 = c(0.09188711, 0.56736177, -0.55276453, -0.07969400, 0.11767092, 2.07541230,  
      1.76443875, 0.60249792, -1.29916116, -0.30322121, -0.68644781, -0.82379154,  
      -0.98416919, -2.02230891, -0.43507791, -0.76655674, 1.22178443, 0.09767100,  
      -0.93391714, -1.23458941, -0.43677762, -0.25606172, -0.05539436, -0.35955890,  
      2.52475885, -0.22159314, -0.17360687, -0.24747086, -0.25016363, -0.39612793)
```

### 12.4.4 Sample from normal distribution? The Lilliefors test

"... testing whether a set of observations is from a normal population when the mean and variance are not specified but must be estimated from the sample." (Lilliefors, 1967)

*A prerequisite for the application of the KS-test is the knowledge of the variance of the population. In reality, the true variance is always never known and thus has to be estimated from the sample. The Lilliefors test<sup>19</sup> takes into account the additional uncertainty resulting from estimation of the variance, and thus its p-values are usually somewhat smaller than those of the KS-test (compare examples below). The test statistic D of the Lilliefors test is the same as that of the KS-test, namely the maximal absolute difference between the hypothesized CDF and the CDF estimated from the sample.*

We will apply the Lilliefors test to the three samples discussed in Exercise 39, reproduced here:

```
y1 = c(-0.68644781, -0.82379154, -0.98416919, -2.02230891, -0.43507791, -0.76655674,
      1.22178443, 0.09767100, -0.93391714, -1.23458941, 0.09188711, 0.56736177,
      -0.55276453, -0.07969400, 0.11767092, 2.07541230, 1.76443875, 0.60249792,
      -1.29916116, -0.30322121, -0.77935252, -0.97190317, 0.84580262, 0.28698246,
      1.15160104, 0.35533328, 0.32936546, 1.68584964, 0.18260973, 1.93600509)
y2 = c(-0.43677762, -0.25606172, -0.05539436, -0.35955890, 2.52475885, -0.22159314,
      -0.17360687, -0.24747086, -0.25016363, -0.39612793, -0.22807506, -0.35896958,
      -0.23775192, 0.34039082, 1.31051033, -0.05988203, -0.15271093, 0.22672744,
      -0.15967543, -0.13572103, -0.44215444, -0.40168018, 0.44666595, -0.53978945,
      -0.45903502, -0.47862169, -0.22521386, -0.38695104, -0.44321571, -0.28632236)
y3 = c(0.09188711, 0.56736177, -0.55276453, -0.07969400, 0.11767092, 2.07541230,
      1.76443875, 0.60249792, -1.29916116, -0.30322121, -0.68644781, -0.82379154,
      -0.98416919, -2.02230891, -0.43507791, -0.76655674, 1.22178443, 0.09767100,
      -0.93391714, -1.23458941, -0.43677762, -0.25606172, -0.05539436, -0.35955890,
      2.52475885, -0.22159314, -0.17360687, -0.24747086, -0.25016363, -0.39612793)
```

The results are shown in Figs. 12.24 to 12.25. The  $p$ -values for the Lilliefors test for y1 and y2 of 0.736 and  $3 \cdot 10^{-8}$ , respectively, are indeed smaller than the ones for the KS-test (0.938 and 0.004, respectively); the decisions based on an  $\alpha = 0.05$  are identical for both tests. For y3 the KS-test yields  $p = 0.136$  (Exercise 39) whereas the Lilliefors test gives  $p = 0.0022$ , i.e. one  $p$ -value is larger than  $\alpha = 0.05$  whereas the other is smaller and thus leading to different decisions:  $H_0$  is rejected (Lilliefors test) or not rejected (KS-test).

---

<sup>19</sup>The Lilliefors test for normality or Lilliefors test for short has been developed independently by Lilliefors (1967) and by van Soest (1967).

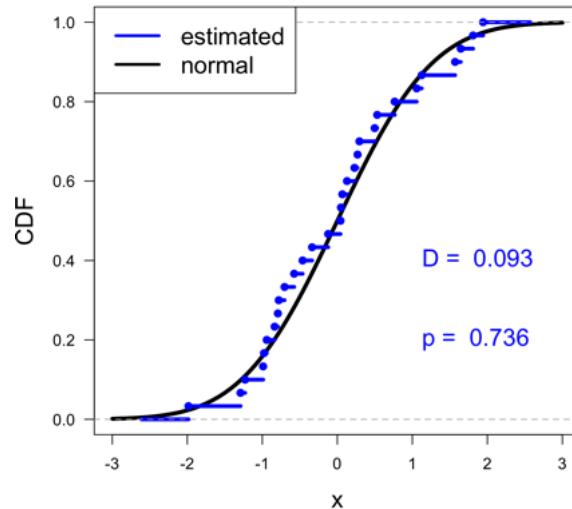


Figure 12.24: Lilliefors test applied to sample y1. [Lilliefors3Ex.R](#)

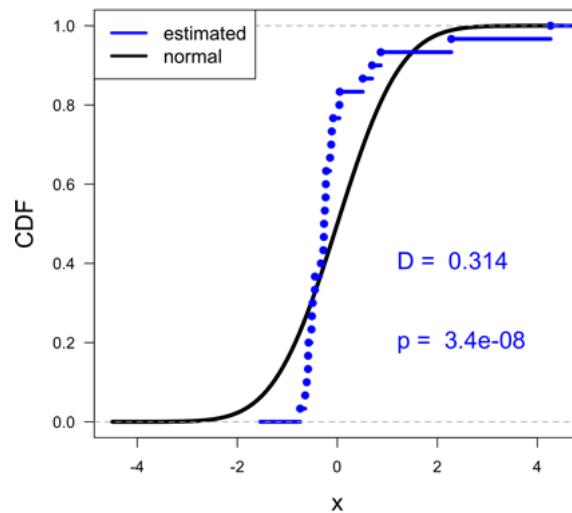


Figure 12.25: Lilliefors test applied to sample y2. [Lilliefors3Ex.R](#) (line 20: set sflag to 2).

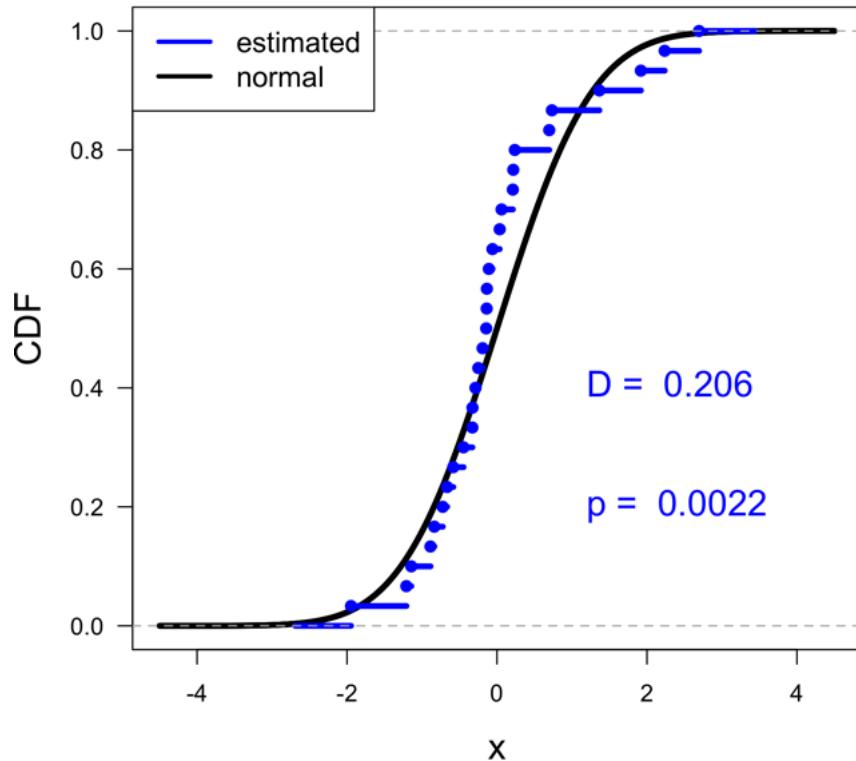


Figure 12.26: Lilliefors test applied to sample y3: For y3 the KS-test yields  $p = 0.136$  (Exercise 39) whereas the Lilliefors test gives  $p = 0.0022$ , i.e. one  $p$ -value is larger than  $\alpha = 0.05$  whereas the other is smaller and thus leading to different decisions:  $H_0$  is rejected (Lilliefors test) or not rejected (KS-test). [Lilliefors3Ex.R](#) (line 20: set sflag to 3).

### 12.4.5 Paul observed edelweiss

Paul, a professional statistician, is hiking in the Swiss Alps starting in S-charl (a village at 1810 m). Not far away from S-charl he recognizes the first edelweiss (*Leontopodium alpinum*; Fig. 12.27).



Figure 12.27: Edelweiss (*Leontopodium alpinum*). Source: Wikipedia

For a professional statistician it is difficult to 'just hike' – and Paul is no exception. While further climbing up, Paul records the observation of each edelweiss recognizable from the path between the meadows whereby his GPS provides the heights. Here are the  $L = 13$  data (in m):

$$x = \{1853, 1872, 1899, 1949, 1976, 1981, 2001, 2027, 2033, 2044, 2111, 2166, 2245\} \quad (12.19)$$

(compare Fig. 12.28).

Paul formulates the following null hypothesis  $H_0$ : 'Edelweiss are uniformly distributed between 1810 m (the height of S-charl where Paul started his hike) and 2270 m (the largest height reached by Paul)'.

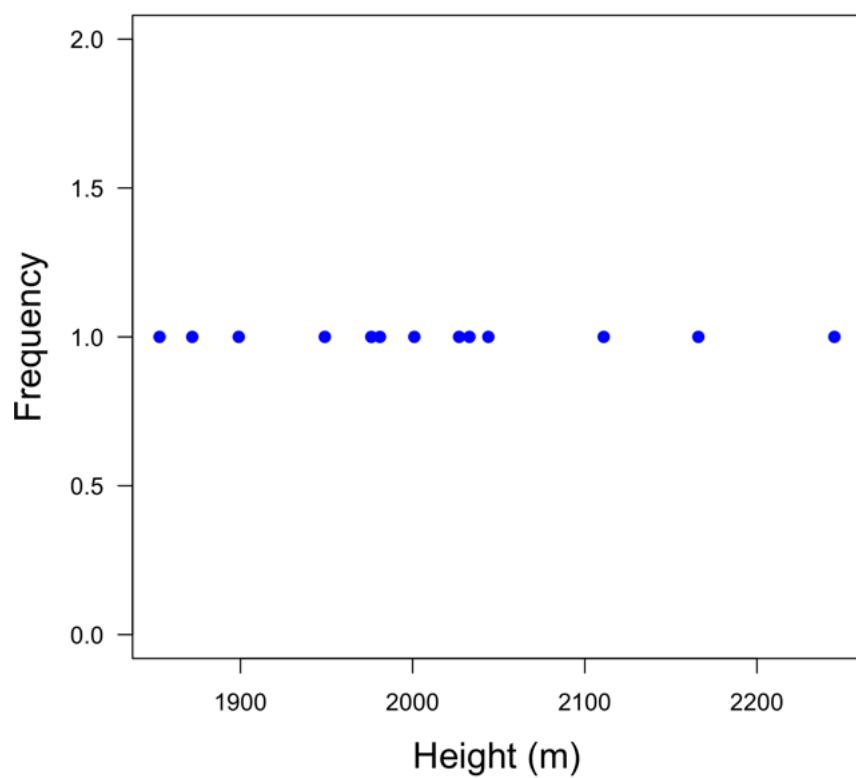


Figure 12.28: Edelweiss data collected by Paul. [EdelweissKS.R](#) (line 15: set sflag to 1)

The null hypothesis  $H_0$  can be tested using the KS-test in R by a '2-liner':

```
x = c(1853, 1872, 1899, 1949, 1976, 1981, 2001, 2027, 2033, 2044, 2111, 2166, 2245)
ks.test(x,'punif',min=1810,max=2270) # D = 0.26054, p-value = 0.2877
```

For the test statistic  $D = 0.2605$  and for  $L = 13$  data points one obtains a  $p$ -value of 0.2877. Thus for the chosen level of significance  $\alpha = 0.05$ , the null hypothesis is not rejected.

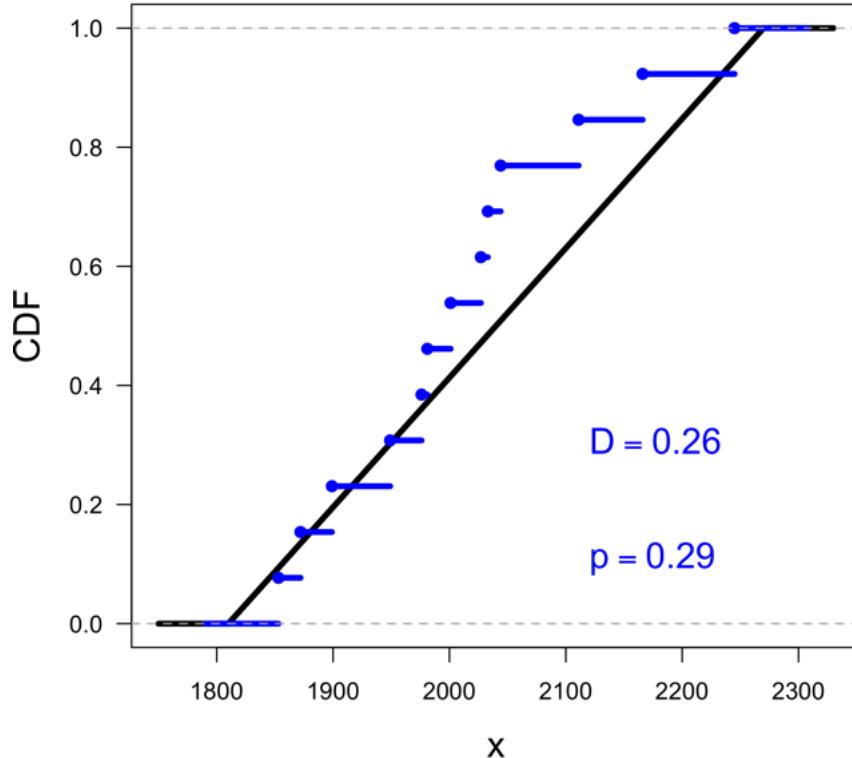


Figure 12.29: CDF estimated from the data. [EdelweissKS.R](#) (line 15: set sflag to 2)

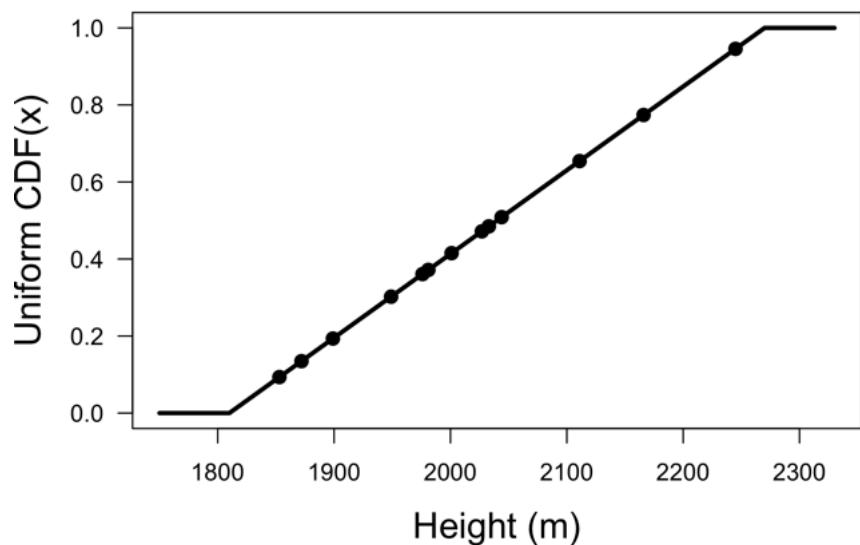


Figure 12.30: CDF corresponding to the PDF of the continuous uniform distribution in the range from 1810 to 2270 m

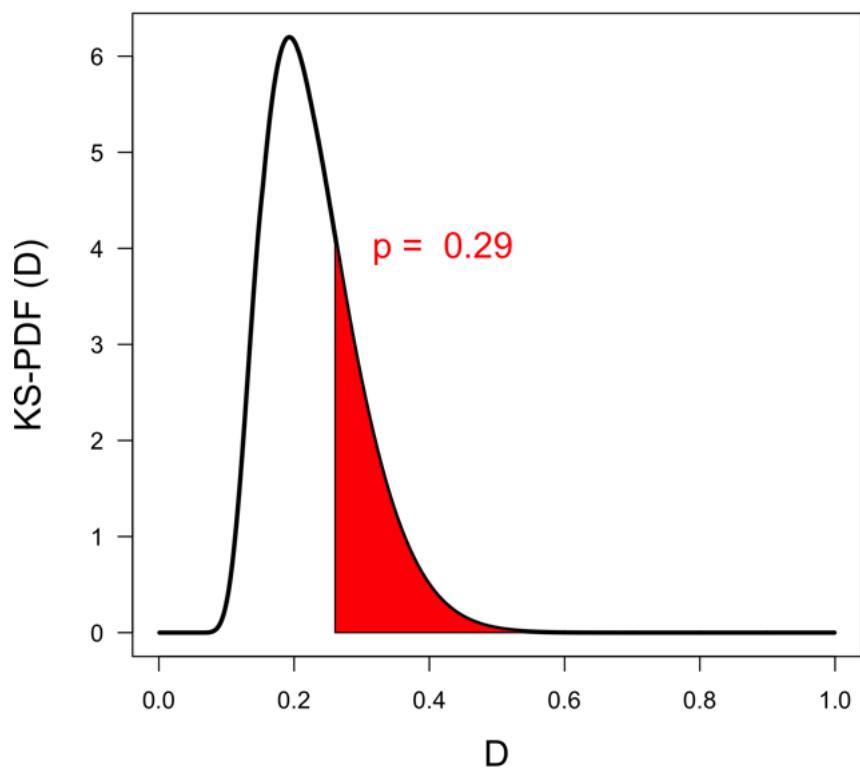


Figure 12.31: PDF of the KS-distribution (blue line) & observed level of evidence  $p = 0.2877$  (size of the red area). [EdelweissKS.R](#) (line 15: set sflag to 7)



## Chapter 13

# Beyond hypothesis testing: Bayesian inference

"Scientific conclusions and business or policy decisions should not be based only on whether a  $p$ -value passes a specific threshold."

Wasserstein (2016, ed.) on behalf of the American Statistical Association (ASA)

*In the current chapter an ecological problem will be considered: Is the abundance of vaquita<sup>1</sup> (*Phocoena sinus*) decreasing? In a series of articles Gerrodette and colleagues have analyzed data using various approaches (from significance test to likelihood and Bayesian inference) that lead to quite different conclusions.*

The content of this chapter is based on: Eguchi & Gerrodette (2009), Gerrodette (2011), Gerrodette et al. (2011)



Figure 13.1: A pair of vaquitas. Source: Paula Olson, NOAA

---

<sup>1</sup>The vaquita (*Phocoena sinus*); also called the Gulf of California or desert porpoise) with maximum size of about 1.5 m is one of the smallest whales worldwide.

## 13.1 Analysis based on summarized data

Gerrodette (2011) applies various inference approaches using only the following summarized data:  $\hat{N}_{97} = 409$  (estimated abundance of vaquita in 1997; also denoted by  $\hat{N}_{1997}$ ),  $\hat{s}_{e97} = 250$  (estimated standard error of  $N_{97}$ ),  $\hat{N}_{08} = 179$  (estimated abundance of vaquita in 2008; also denoted by  $N_{2008}$ ),  $\hat{s}_{e08} = 74$  (estimated standard error of  $N_{08}$ ):

1. Frequentistic approach, null hypothesis significance testing (NHST)
2. Frequentistic approach: lognormal confidence intervals
3. Frequentistic approach, confidence interval for the difference
4. Likelihood inference
5. Bayesian inference (specifying priors)

### 13.1.1 Frequentistic approach: null hypothesis significance test (NHST)

The Wald test<sup>2</sup> yields a  $p$ -value of 0.38 and thus the null hypothesis  $H_0$  'abundances in 1997 and 2008 are equal to each other' can *not* be rejected on the level of significance  $\alpha = 0.05$ .

### 13.1.2 Why using lognormal PDFs for abundances?

In the following analysis, lognormal PDFs will be used for the vaquita abundance. Buckland et al. (2001, reprinted 2009, p. 77) give the following justification: "However, the distribution of" [the estimate of the object density] " $\hat{D}$  is positively skewed, and an interval" [confidence interval] "with better coverage is obtained by assuming that  $\hat{D}$  is log-normally distributed." Actually, for abundance data one would use probability distributions (PDs, not PDFs) because the data are discrete, actually non-negative integer. The first PD candidate that comes to mind is the Poisson distribution. However, Poisson distributions are not very flexible because they depend only on a single parameter, namely the mean rate  $\lambda$ , and they possess the special property that the mean  $\mu = \lambda$  is equal to the variance  $\sigma^2$ . The (estimates of) variances of the vaquita data are much larger than the mean:  $\hat{s}_{e08}^2 = 62500 \gg 409 = \hat{N}_{97}$ . Thus a Poisson distribution is obviously ruled out. The next candidate is the negative binomial probability distribution (Fig. 13.2) which is defined for non-negative integers. Neither Buckland et al. (2001) nor Gerrodette (2011) give a hint why they do not use this distribution. Instead they use lognormal PDFs (continuous!) which might be considered as the continuous envelop of a discrete probability distribution which can provide good approximations for probabilities calculated over large abundance ranges. The use of lognormal PDFs and negative binomial PDs will be compared in Exercise 40.

<sup>2</sup>The Wald test is equal to the limit of the  $t$ -test for large number of degrees of freedom. The test statistic,  $z$ , is the same as for the  $t$ -test, namely (for the two-sample test) given by the difference of the sample means divided by the square root of the sum of squared standard errors:  $z = (\hat{N}_{08} - \hat{N}_{97}) / \sqrt{\hat{s}_{e97}^2 + \hat{s}_{e08}^2} = -230/260.7 = -0.882$ .

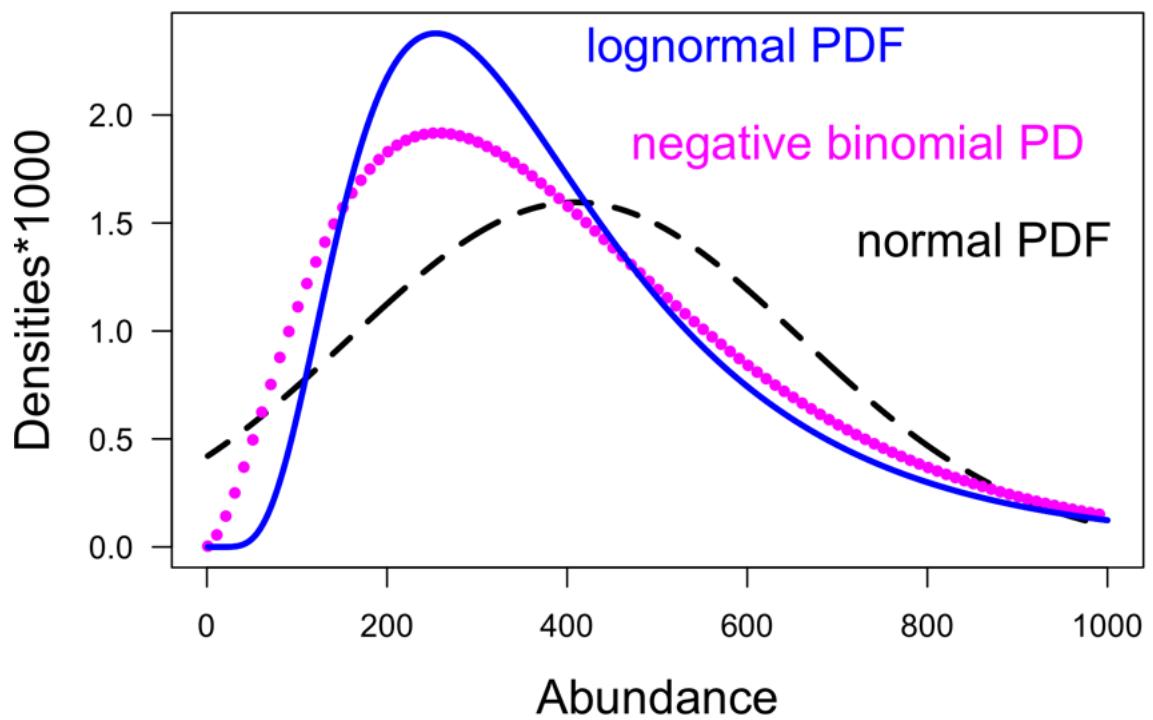


Figure 13.2: Plot of the negative binomial PD (magenta dots), the normal PDF (black broken line), and the lognormal PDF (blue solid line) all for mean  $\mu = 409$  and standard deviation  $\sigma = 250$ . [Gerrodette3Densities.R](#)

### 13.1.3 Frequentistic approach: lognormal confidence intervals

Given the abundance  $\hat{N}_{97}$  and the standard error  $\hat{s}_{e_{97}}$ , one can calculate the 95% confidence interval (Figs. 13.3 and 13.4) based on the assumption that the abundance stems from a lognormal PDF with mean  $\mu = \hat{N}_{97}$  and standard deviation  $\sigma = \hat{s}_{e_{97}}$ . **Exercise:** (i) calculate the distribution parameters  $\alpha$  and  $\beta$  from  $\mu$  and  $\sigma$ , (ii) find the lower and upper confidence interval boundaries from the CDF of the lognormal distribution.

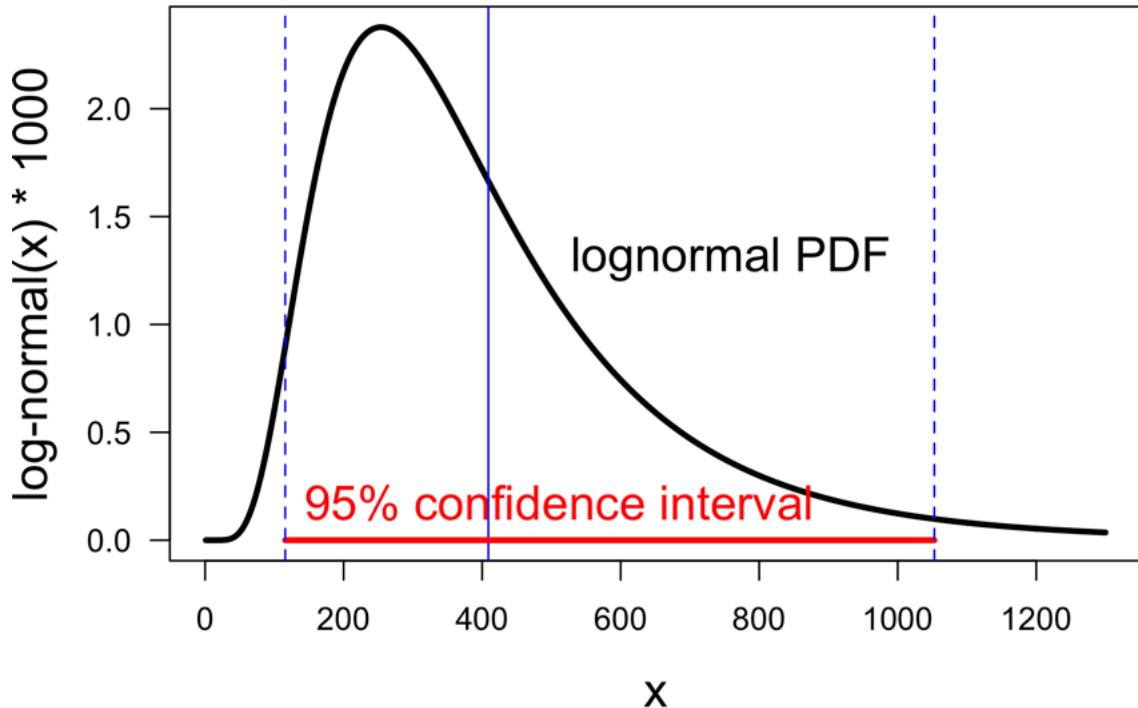


Figure 13.3: Lognormal PDF with mean  $\mu = \hat{N}_{97} = 409$ ,  $\sigma = \hat{s}_{e_{97}} = 250$  (black line), the observed abundance  $\hat{N}_{97}$  (vertical blue line). The 95% confidence interval ranges from 115.7 to 1052.9 (red line).

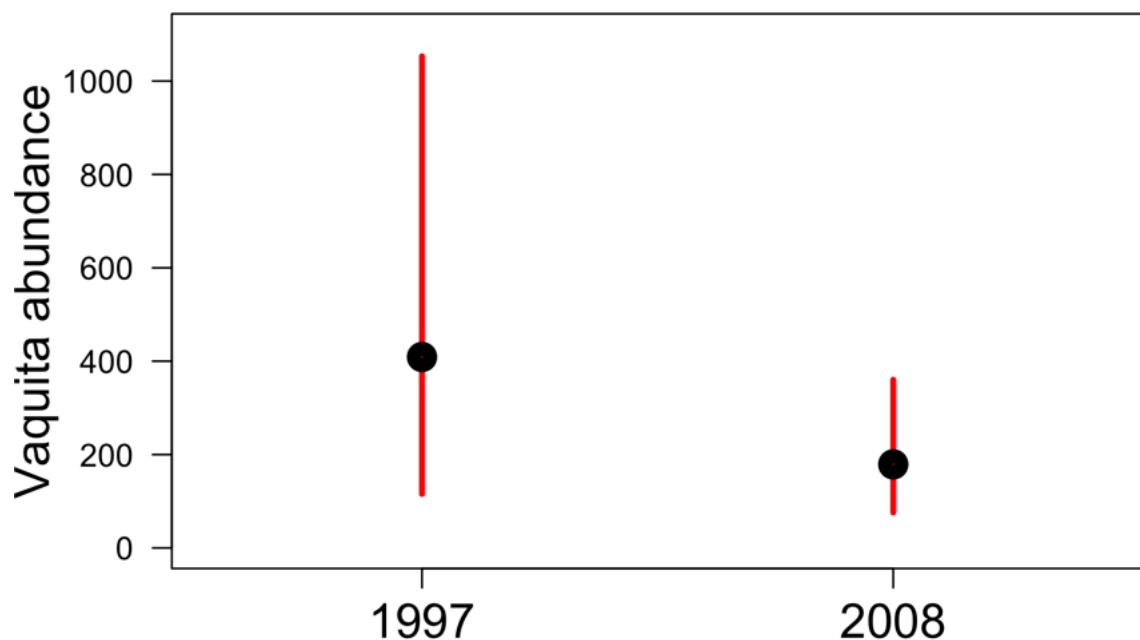


Figure 13.4: Gerrodette (2011, Fig. 2A, modified): "Frequentist inference for vaquita (*Phocoena sinus*) data. (A) Point estimates and 95% lognormal confidence intervals for vaquita abundance in 1997 and 2008 in the core area of the species' distribution."

### 13.1.4 Frequentistic approach: confidence interval for the difference

The 95% confidence interval given by the difference  $\hat{N}_{08} - \hat{N}_{97} = -230 \pm \sigma_{\hat{N}_{08} - \hat{N}_{97}}^3$ <sup>3</sup> extends from -741 to 281. It includes zero, meaning 'no change'.

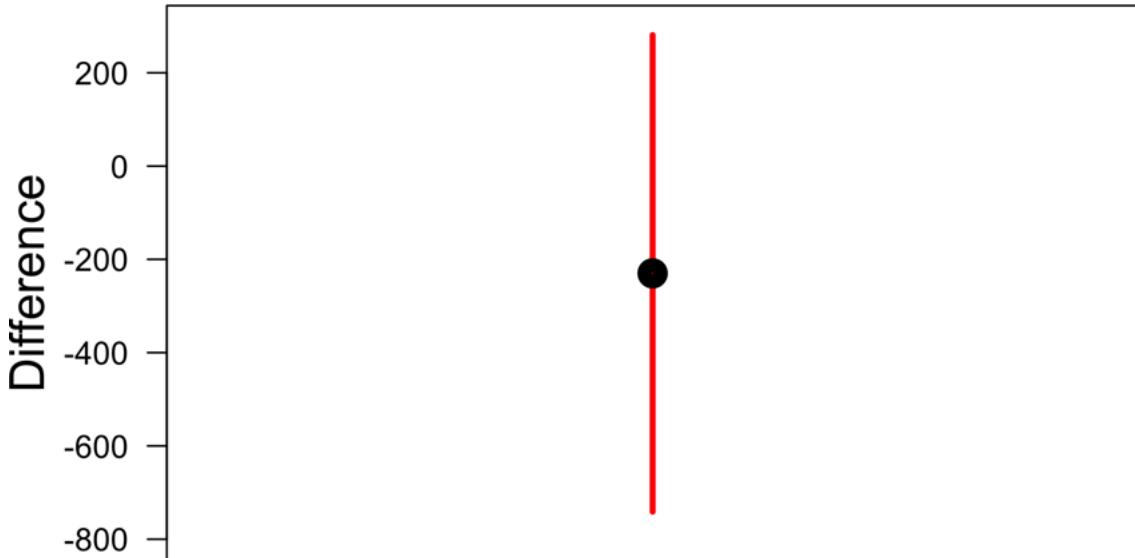


Figure 13.5: Gerrodette (2011, Fig. 2B, modified): "Frequentist inference for vaquita (*Phocoena sinus*) data. . . . (B) Point estimate and 95% normal confidence interval of  $d = N_{2008} - N_{1997}$ , the change in vaquita abundance between 1997 and 2008."

---

<sup>3</sup> $\sigma_{\hat{N}_{08} - \hat{N}_{97}} = \Phi(1 - \alpha/2) \sqrt{\hat{s}e_{97}^2 + \hat{s}e_{08}^2} = 511$  where  $\Phi(1 - \alpha/2) = 1.96$  ( $\Phi(z)$  is the cumulative distribution function for the standard normal PDF) and  $\sqrt{\hat{s}e_{97}^2 + \hat{s}e_{08}^2} = 260.7$  is the standard error of the difference.

### 13.1.5 Likelihood inference

The likelihood function for the true difference,  $d$ , and the true population in 1997,  $N_{97}$ , is given as the product<sup>4</sup> of two lognormal PDFs

$$L(d, N_{97} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08}) = \text{lnorm}(N_{97} | \hat{N}_{97}, \hat{s}_{e97}) \times \text{lnorm}(N_{97} + d | \hat{N}_{08}, \hat{s}_{e08}) \quad (13.1)$$

(Fig. 13.6). Note that  $d = N_{08} - N_{97}$  and thus  $N_{97} + d = N_{08}$ , i.e. the likelihood function for  $d$  and  $N_{97}$  is the product of the likelihood functions for  $N_{97}$  and  $N_{08}$ . The likelihood over the chosen sector of the  $d$ - $N_{97}$  plane is mostly small except for a narrow ridge. The maximum of the likelihood is located at the difference  $d = -113$  and the abundance  $N_{97} = 254$ .

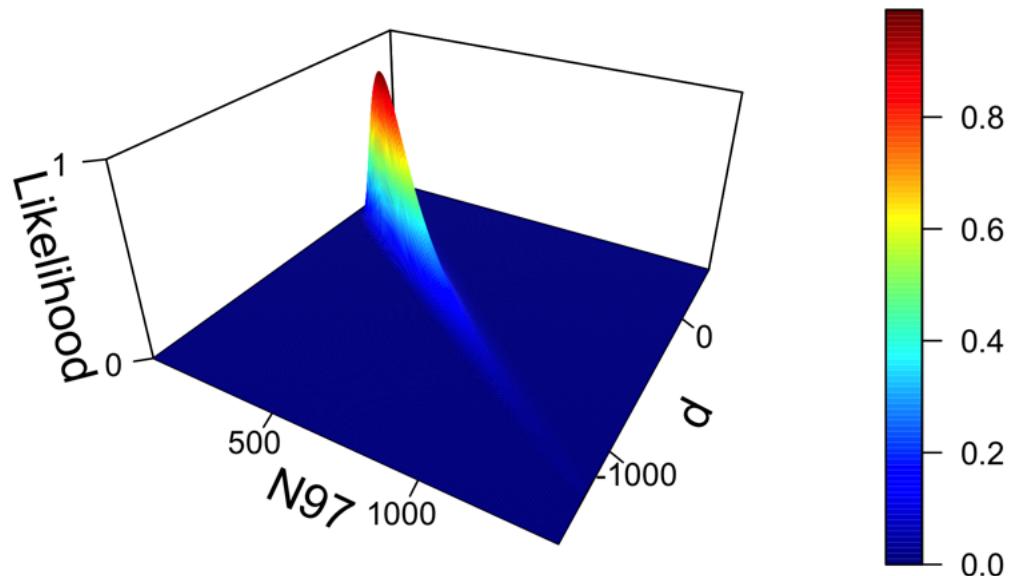


Figure 13.6: Gerrodette (2011, Fig. 3A, modified): "Likelihood inference for  $d = N_{2008} - N_{1997}$ , the change in vaquita abundance between 1997 and 2008, for summarized data. (A) Joint likelihood surface of  $d$  and  $N_{1997}$ ." The maximum of the likelihood is located at the difference  $d = -113$  and the abundance  $N_{97} = 254$ . Note: the likelihood has been scaled to a maximum of 1. [GerrodetteLikelihood.R](#)

<sup>4</sup>The use of the simplified product rule is allowed when the observations in 2008 are independent from those of 1997.

### 13.1.6 Likelihood inference: profile likelihood of the difference, $d$

There are various ways for obtaining a likelihood for  $d$ . By taking for each value of  $d$  the maximum likelihood computed over all  $N_{97}$  values, one obtains the [profile likelihood](#) (Fig. 13.7; projection on the  $N_{97} = 0$ -likelihood-plane).<sup>5</sup> The  $1/8$  likelihood is defined as the range between the two differences at which the profile likelihood is  $1/8$  of the maximum profile likelihood, here from -667 to +142. Although the  $1/8$  likelihood interval is narrower than the 95% confidence, it still contains zero difference and thus 'no change' is not excluded.

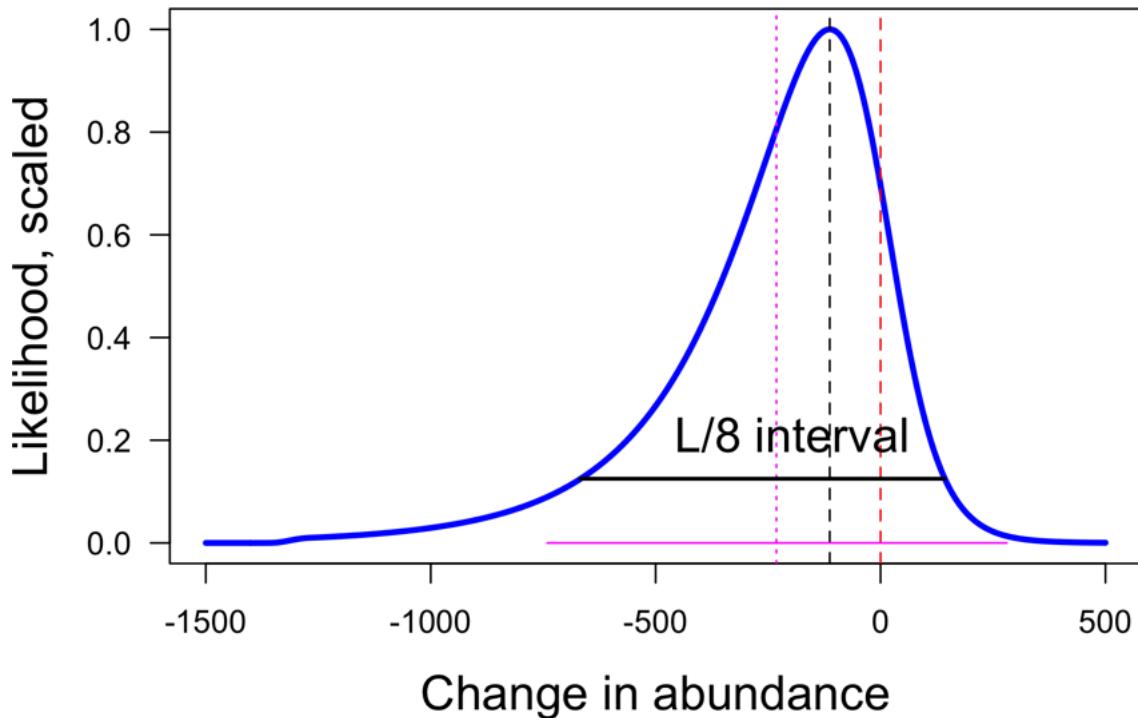


Figure 13.7: Gerrodette (2011, Fig. 3B, modified): "Likelihood inference for  $d = N_{2008} - N_{1997}$ , the change in vaquita abundance between 1997 and 2008, for summarized data. ··· (B) Profile likelihood of  $d$ ." The horizontal black line is the  $1/8$  likelihood interval from -667 to +142. The horizontal magenta line is the 95% confidence interval from -741 to +281. The vertical red line indicates the likelihood at  $d = 0$ . The dotted vertical black line indicates the maximum likelihood at  $d = -113$ . The dash-dotted vertical magenta line indicates the mean likelihood at  $d = -231$ .

<sup>5</sup>Another way to obtain a likelihood for  $d$  would be by integrating out the nuisance parameter  $N_{97}$ .

### 13.1.7 Bayesian inference: joint posterior for flat prior

The (full) Bayesian approach requires the specification of priors (probability densities for the model parameters without taking into account, i.e. prior to looking at, the data). If nothing is known about the model parameters (from previous observations or theory) one can specify noninformative priors. Here the prior distribution for  $N_{97}$  and  $N_{08}$  is replaced by a constant over a large, but finite rectangle (so-called 'flat' prior)<sup>6</sup> which means that it is assumed that all values of  $N_{97}$  and  $N_{08}$  are a priori equally probable and that there is no correlation between them. The joint posterior density  $P(N_{97}, N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08})$  (Fig. 13.8) is given by the product of lognormal likelihood functions for  $N_{97}$  and  $N_{08}$  times the constant prior.<sup>7</sup>

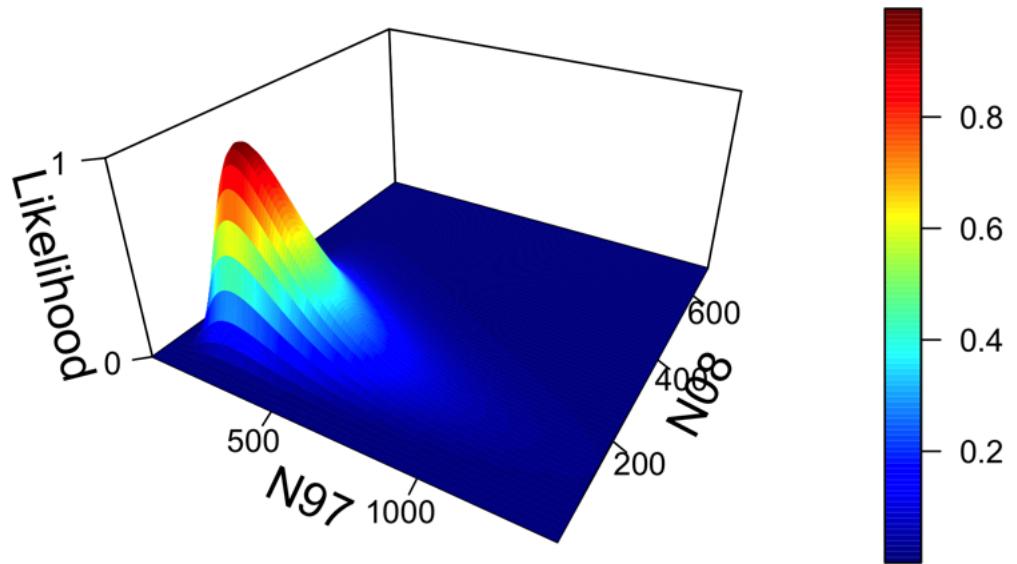


Figure 13.8: Gerrodette (2011, Fig. 4, modified): Joint posterior (not normalized; scaled to maximum = 1)  $P(N_{97}, N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08})$  for a flat prior. The maximum is located at  $N_{97} = 254, N_{08} = 141$ .

<sup>6</sup>Please note: (1) Assigning a flat prior with a constant  $\neq 0$  over an infinite interval may cause problems because it can not be normalized; one speaks of an 'improper' prior. (2) Not all noninformative priors are flat.

<sup>7</sup>The value of the prior constant does not play a role because it cancels out when the posterior density is normalized to 1.

### 13.1.8 Bayesian inference: marginal posteriors for flat prior

From the joint posterior density  $P(N_{97}, N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08})$  one can derive [marginal posteriors](#) for  $N_{97}$  and  $N_{08}$ , respectively, by integration over the respective other abundance:

$$P(N_{97} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08}) = \int_0^{\infty} P(N_{97}, N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08}) dN_{08} \quad (13.2)$$

$$P(N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08}) = \int_0^{\infty} P(N_{97}, N_{08} | \hat{N}_{97}, \hat{s}_{e97}, \hat{N}_{08}, \hat{s}_{e08}) dN_{97} \quad (13.3)$$

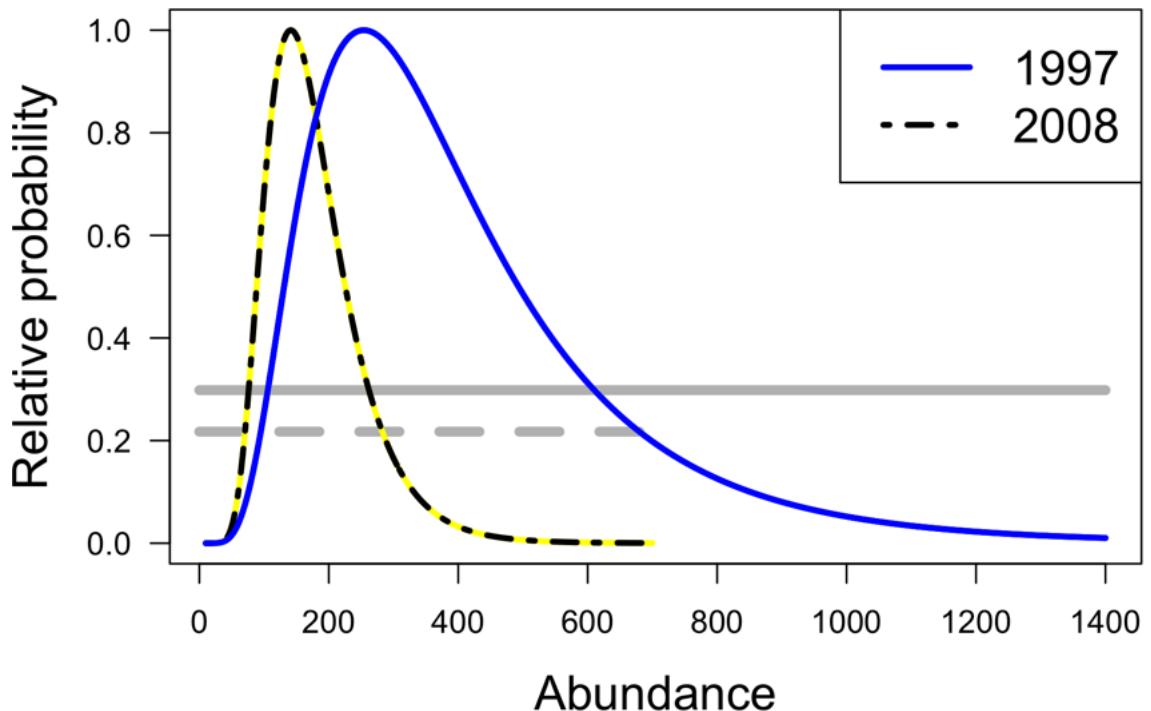


Figure 13.9: Gerrodette (2011, Fig. 5A, modified): "Bayesian inference for summarized vaquita data with non-informative priors. Distributions are scaled to the maximum value of the posteriors. (A) Prior and" marginal "posterior distributions of  $N_{1997}$  and  $N_{2008}$ . Priors are scaled to have the same area as the posterior for the same year."

### 13.1.9 Bayesian inference: change in abundance for flat prior

The change in abundance  $d = N_{08} - N_{97}$  (Fig. 13.10) can be obtained by sampling from the joint posterior density  $P(N_{97}, N_{08} | \hat{N}_{97}, \hat{N}_{08}, \hat{s}_{e97}, \hat{s}_{e08})$ .

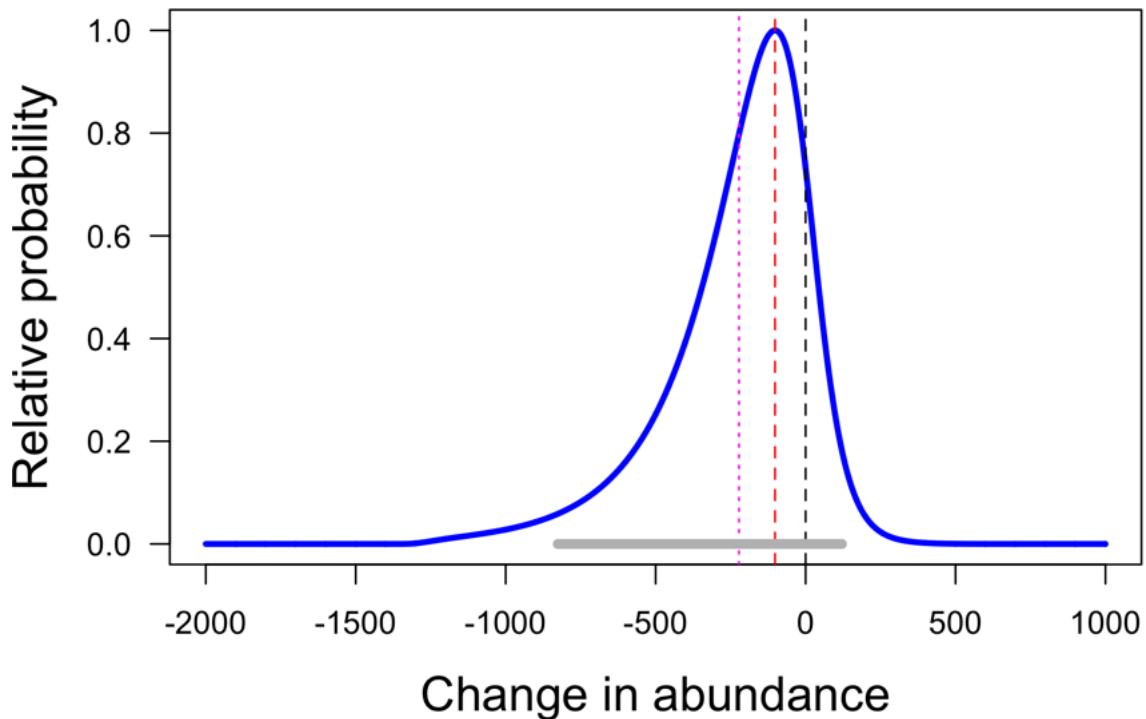


Figure 13.10: Gerrodette (2011, Fig. 5B, modified): "Bayesian inference for summarized vaquita data with noninformative priors. Distributions are scaled to the maximum value of the posteriors. (B) Posterior distribution of change in abundance  $d = N_{2008} - N_{1997}$ . The horizontal gray line is the central 95% probability interval, and the vertical dashed black line at  $d = 0$  indicates the fractions of the distribution above and below 0." The probability for obtaining negative differences (decrease of abundance) is estimated to be  $\hat{p}_{d < 0} = 0.86$  or 86%. The vertical dashed red line indicates the location of the maximum of the likelihood at  $d = -102$ . The vertical dash-dotted magenta line indicates the location of the mean difference at  $d = -222$ .

### 13.1.10 Bayesian inference: informative priors

Gerrodette (2011) mentions the following prior information: "Prior to the 1997 survey, there was some knowledge of vaquita abundance based on previous partial surveys in the area (Barlow et al. 1997). Likewise, prior to the 2008 survey, there were indications that the population could be as small as 150, based on increased fishing effort and the bycatch rate in gillnets (Jaramillo-Legorreta et al. 2007)."

Based on this prior information, Gerrodette (2011) assigns lognormal prior distributions with means of 600 and 150, respectively, for 1997 and 2008, and large uncertainties equal to the mean values.

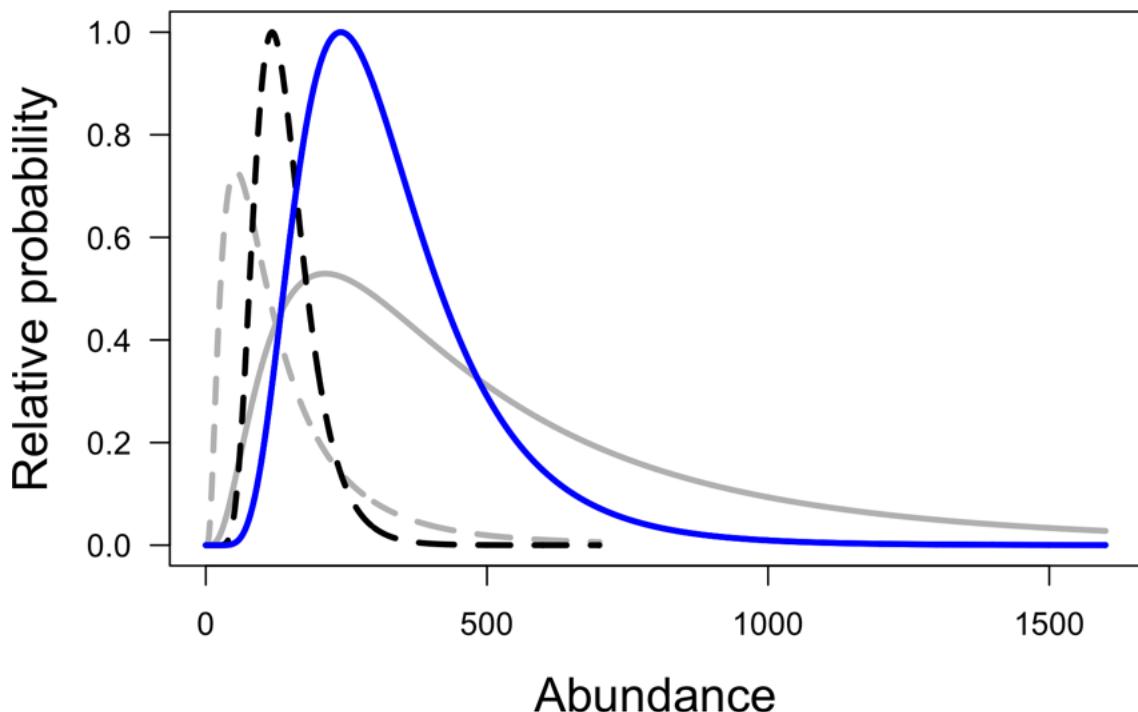


Figure 13.11: Gerrodette (2011, Fig. 6A, modified): "Bayesian inference for summarized vaquita data with informative priors. Distributions are scaled to the maximum value of the posteriors. (A) Prior and posterior distributions of  $N_{1997}$  and  $N_{2008}$ . Priors are scaled to have the same area as the posterior for the same year."

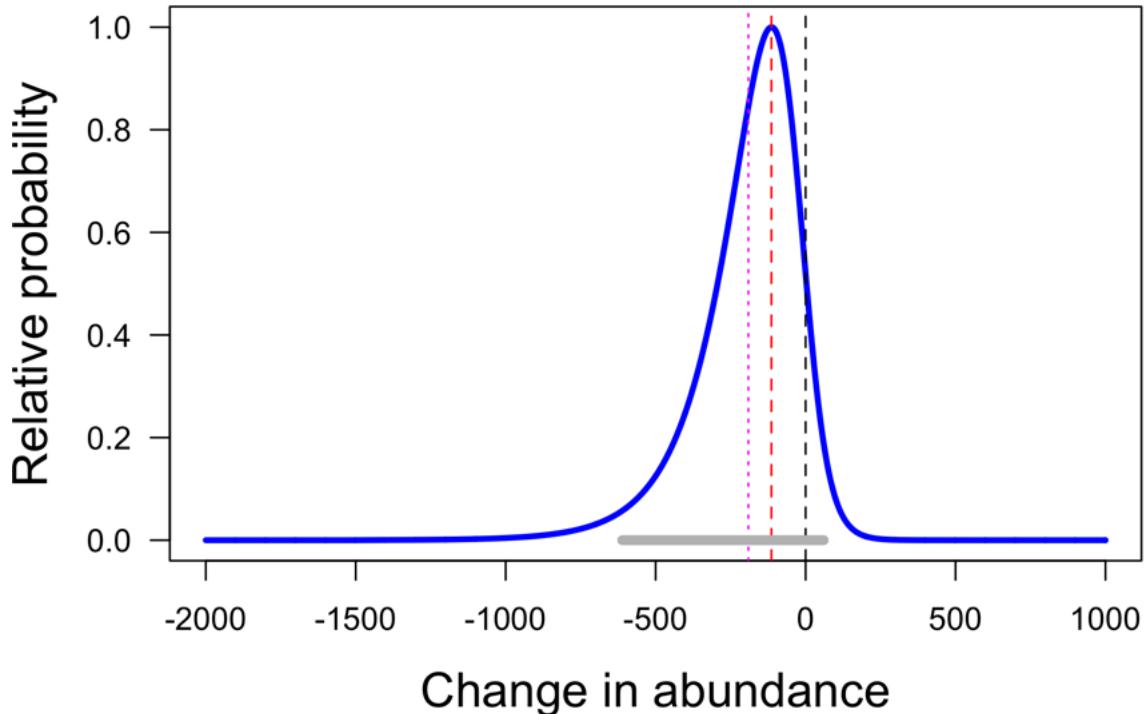


Figure 13.12: Gerrodette (2011, Fig. 6B, modified): "Bayesian inference for summarized vaquita data with informative priors. Distributions are scaled to the maximum value of the posteriors. . . . (B) Posterior distribution of change in abundance  $d = N_{2008} - N_{1997}$ . The horizontal gray line is the central 95% probability interval, and the vertical" dashed black "line at  $d = 0$  indicates the fractions of the distribution above and below 0." The probability for obtaining negative differences (decrease of abundance) is estimated to be  $\hat{p}_{d < 0} = 0.92$  or 92%. The vertical dashed red line indicates the location of the maximum of the likelihood at  $d = -114$ . The vertical dash-dotted magenta line indicates the location of the mean difference at  $d = -191$ .

**Summary:** The  $p$  value obtained from the Wald test (frequentist approach) is not small enough to reject the null hypothesis  $H_0$  ('no change in abundance') and also other frequentistic approaches (confidence intervals) could not exclude 'no change'. The posterior (Bayesian approach) based an a flat prior yields 86% probability for decrease of abundance. This value is further increased to 92% by applying an informed prior.

## 13.2 Inference based on original data

The original data might contain more information about vaquita abundances than the summarized data used above. The acquisition of data is described in Gerrodette et al. (2011). The theoretical background for analyzing line-transect data is based on Buckland et al. (2001) and Eguchi & Gerrodette (2009).

The original data consist of sightings of vaquita on various line-transects using research vessels or sailing boats where perpendicular distances to the transects were recorded (data listed below in R code).

The likelihood and the posterior depend on 5 variables. The ranges, likelihoods, and priors for these variables are explained in the appendix (Section I).

The calculation of the likelihood  $L$  requires some memory space and time because  $L$  depends on 5 variables (with the resolution in the R code provided by Gerrodette (2011), the likelihood matrix contains 12.9 million elements and computation takes less than 2 minutes on a PC).

The posterior for the change of abundance based on the profile likelihood (Fig. 13.13) and the marginal posterior (Fig. 13.14) possess their maximum at -246 (decrease in abundance) and the probability density at 0 (no change of abundance) is negligible. These results clearly support the conclusion that 'the abundance of vaquita has decreased between 1997 and 2008'.

**Further reading (vaquita):** Vance (2017)

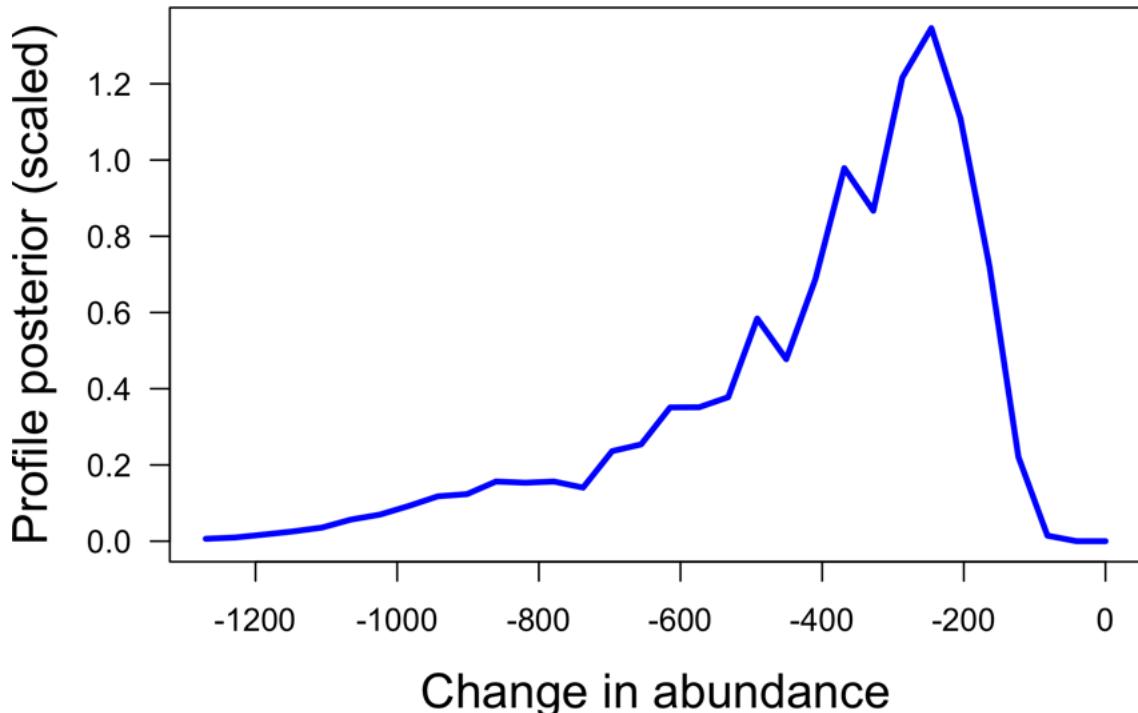


Figure 13.13: Profile posterior for the change of vaquita abundance between 1997 and 2008. The resulting distribution clearly supports the conclusion 'the abundance of vaquita has decreased between 1997 and 2008'.

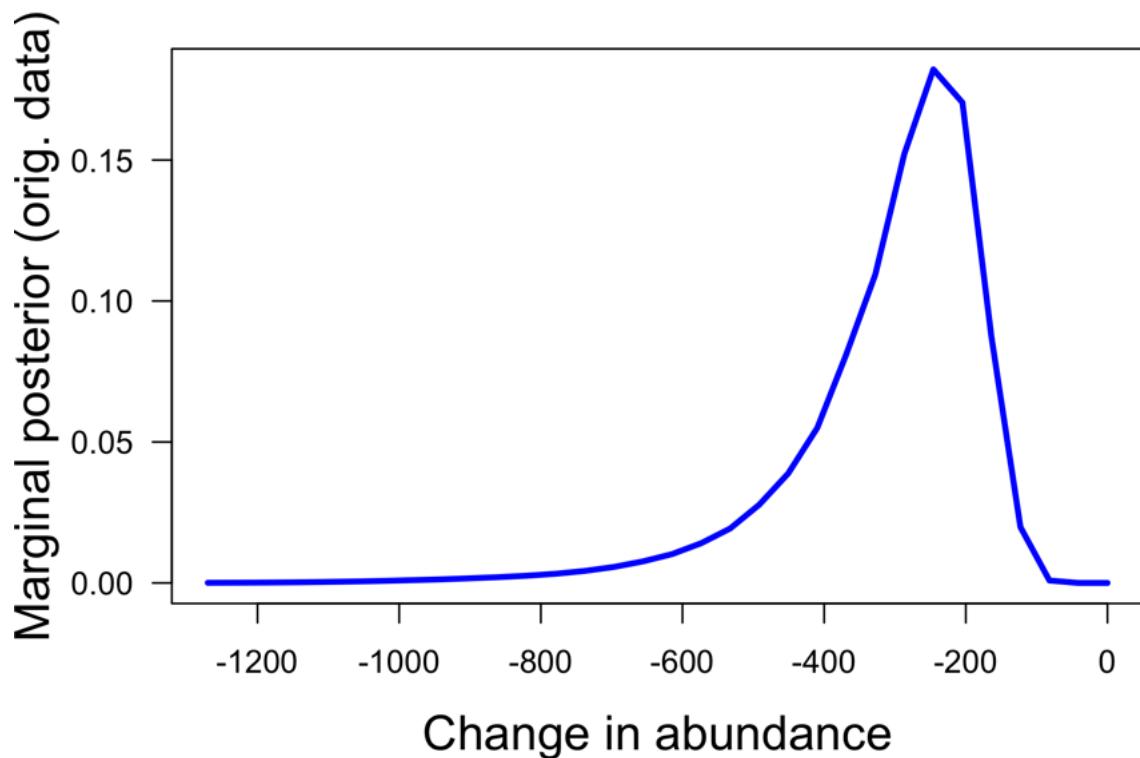


Figure 13.14: Marginal posterior for the change of vaquita abundance between 1997 and 2008. The resulting distribution clearly supports the conclusion 'the abundance of vaquita has decreased between 1997 and 2008'.

**Exercise 40 Negative binomial PD instead of lognormal PDF**

Gerrodette (2011) describes the abundance distributions by lognormal PDFs which can, however, be only an approximation for high abundances ( $N > 30$ ) because count data are discrete and thus their population should be described by a probability distribution (PD) and not by a PDF. A motivation for using the lognormal (instead of the normal) PDF is the asymmetry of the distribution with a finite left and a large right tail. The negative binomial PD fulfills the same requirements and is thus suggested as an alternative to the lognormal PDF.

1. Calculate the distribution parameters  $s$  and  $p$  for the negative binomial PD from the mean values and their uncertainties given in Gerrodette (2011).
2. Plot both the lognormal PDF and the negative binomial PD for  $\mu = 409$  and  $\sigma = 250$ .
3. Calculate the likelihood based on the negative binomial PD for the observed abundance data and its uncertainties.
4. Calculate and plot the profile likelihood and compare it with the profile likelihood given in Gerrodette (2011, Fig. 3B).
5. Calculate the mode, mean, and the likelihood ratio based on the profile likelihood and compare the values with those given by Gerrodette (2011).
6. Repeat the analysis with smaller observed abundances:  $\mu_1 = 41$ ,  $\sigma_1 = 25$ ,  $\mu_2 = 18$ ,  $\sigma_2 = 7$ . Does one obtain similar results based on the lognormal PDF or the negative binomial PD?

**Exercise 41 Jeffreys' prior for vaquita data**

Gerrodette (2011) assigns flat priors for  $N_{97}$  and  $N_{08}$  and calls them non-informative.

- (1) However, abundances can take only non-negative values and thus, analogue to the case of non-negative continuous model parameters investigated by Jeffreys (1961), one might assign Jeffreys' priors  $1/N_{97}$  and  $1/N_{08}$  instead of flat priors. Note that Jeffreys' priors are improper because of the singularity at zero abundance.
- (2) Apply lognormal priors with mean  $\mu = 600$  and standard deviation  $\sigma = 600$  for 1997 and mean  $\mu = 150$  and standard deviation  $\sigma = 150$  for 2008.
- (3) Calculate the marginal posteriors for  $N_{97}$ ,  $N_{08}$ , and from these the distribution for the difference  $d = N_{08} - N_{97}$ .
- (4) Compare the difference distributions with the ones for flat priors (plot).

# Chapter 14

## Linear models: straight lines

Observed pairs of data  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  are often considered as resulting from processes responsible for an *exact, often simple relationship* between the variables X and Y plus some noise leading to more or less deviations from the exact relationship. The processes leading to the underlying exact relationship might be known (for example, physical laws) or not. Several different goals of data analysis might be of interest here:

- (1) The form of the exact relationship (straight line, polynomial, exponential growth curve) is known from theory or a mechanism, however, the parameters (intercept, slope, polynomial coefficients, growth rate constant) are not known and shall be estimated from the data (together with their uncertainties).
- (2) The noisy data shall be replaced by simple relationships providing smooth approximations that can be used for further purposes (noise reduction, mean variations, management). Especially for complex systems (for example, organisms or ecosystems) exact laws comparable conservation of momentum, energy, or angular momentum in physics (all related to symmetries) are not expected to exist, non-the-less simple, however noisy, relationships may 'emerge' on certain system levels. Here it might be of interest to find the 'optimal' form of such relationships.
- (3) Variables X and Y could *covary without being in a causal relationship to each other* (for example, when they are both influenced by the same process). Here the aim of data analysis could be the separation between a simple relationship and noise.

The simple relationships between X and Y are called '**models**'. Arguably the simplest nontrivial model is a **straight line**  $Y = \beta_0 + \beta X$  where  $\beta_0$  is the intercept<sup>1</sup> and  $\beta$  is the slope.  $\beta_0$  and  $\beta$  are unknown model parameters that have to be estimated from the data, i.e. in contrast to the point estimation problems discussed in Chapter 10, one has to estimate several parameters at the same time. The straight line model in **linear in a two-fold sense**: Y is a linear function of X and Y is a linear function of the model parameters  $\beta_0$  and  $\beta$  whereby the second type of linearity is relevant in this chapter and in the following two chapters (Chapters 16 and 17).

Depending on our knowledge or on our assumptions about X, Y, their relationship, and the type of noise one can discern the following approaches to 'straight line fitting':

- (1) **Simple linear regression (SLR)**<sup>2</sup> X is a non-stochastic variable (i.e. it is known exactly or at least the uncertainties in measurements of X can be neglected in the current context; in statistical slang X is 'fixed'), Y is a given as a linear response to X according to  $Y = \beta_0 + \beta X$  plus additive normal noise with mean  $\mu = 0$  and (unknown) variance  $\sigma^2$ , i.e. the observations are given by  $y_j = \beta_0 + \beta x_j + \epsilon_j$  with  $\epsilon_j \in \mathcal{N}(\mu = 0, \sigma)$ . The additive normal noise makes Y a stochastic variable. X is called the *independent variable* or the *predictor*, Y is called the *dependent variable* or the *response*.
- (2) **Both X and Y are stochastic variables**, their relationship is non-causal and might be described by  $y_j = \beta_0 + \beta x_j + \epsilon_j$  or by  $x_j = \gamma_0 + \gamma y_j + \epsilon_j$  again with additive normal noise with mean  $\mu = 0$  and (unknown) variance  $\sigma^2$ . The estimation of  $\beta_0$  and  $\beta$  is more tricky than for SLR<sup>3</sup>; a **maximum likelihood approach** (also known as *Deming regression*) is the method of choice (Chapter 15).

<sup>1</sup>Y-intercept, i.e. Y value at X = 0

<sup>2</sup>'Regression to the mean' is a concept introduced by Francis Galton around 1880, i.e. at least more than 70 years after the invention of least squares which is most often used for straight line fitting. For an insightful explanation and discussion of the concept of 'regression to the mean' compare Stiegler (1999, Chapter 9) and Kahneman (2011, Chapter 17).

<sup>3</sup>Not all methods suggested in the literature are well justified and some are applicable only when X and Y are dimensionless or have identical units.

(3) *Various complications to SLR* might have to be considered as, for example, varying noise levels ('residuals with pattern') or non-normal noise or model parameters  $\beta_0$  and/or  $\beta$  containing 'fixed' (constant) and random contributions (so called '*mixed effects models*' discussed, for instance, in Zuur et al., 2009).

*Remark:* Do not ask how many papers about 'straight line fitting' have been published!

**Further reading:** Zellner (1971), Fahrmeir et al. (2013), Montgomery et al. (2015)

## 14.1 Fitting a straight line to an artificial data set: simple linear regression

Generate an artificial data set as follows:

1. The predictive variable  $X$  is considered a non-stochastic variable. For the sake of simplicity, one chooses 17 equidistant values between 1 and 5, namely

$$x = \{1, 1.25, 1.5, \dots, 4.75, 5\}. \quad (14.1)$$

These values are considered to be known exactly<sup>4</sup>, i.e.  $x \equiv x_{\text{exact}}$ .

2. One assumes that the response variable  $Y$  and the predictive variable  $X$  are exactly related to each other by a straight line:

$$y_{\text{exact}} = \beta_0 + \beta x_{\text{exact}} = 8 - 1.2 x_{\text{exact}} \quad (14.2)$$

where  $\beta_0 = 8$  for the intercept and  $\beta = -1.2$  for the slope.

3. Normally distributed noise with mean  $\mu = 0$  and (chosen) variance  $\sigma^2 = 1$  is added to  $y_{\text{exact}}$ .
4. Thus the 'recipe' for the generation of observation data  $y$  is

$$y = y_{\text{exact}} + \epsilon \mathcal{N}(\mu=0, \sigma=1) = 8 - 1.2 x_{\text{exact}} + \epsilon \mathcal{N}(\mu=0, \sigma=1). \quad (14.3)$$

For each of the 17  $x$  values one generates  $n = 10$  noise values and thus  $n = 10$   $y$  values; overall one generates  $17 \cdot 10 = 170$  response values  $y_j$  (Fig. 14.1).

Now the goal is to (i) estimate the model parameters  $\beta_0$ ,  $\beta$  and  $\sigma$  from the data if one already knows from theoretical considerations (mechanisms) or assumes the right form of relationship between  $X$  and  $Y$  or (ii) just describe the data by a simple relationship  $y = f(x)$ . In case (ii), it does not make sense to look for a relationship that fits all the data because (a) part of the variations are caused by noise ('Don't fit the noise!') and (b) a function  $y = f(x)$ , assigning a unique  $y$  value for each  $x$ , obviously does not exist here because several  $y$  values were generated for each  $x$ . In order to derive a simple relationship  $y = f(x)$  one first averages the data along the  $y$  direction (regression to the mean). Thus obtains for each  $x$  value a single mean  $y$  value (Fig. 14.2). The mean  $y$  values already lie close to the exact relationship (Eq. 14.2) because the normal noise with mean  $\mu = 0$  largely cancels out; for  $n = 10$  one expects deviations of the order of the standard error of the mean, i.e.  $\sigma / \sqrt{n} = 1 / \sqrt{10} \approx 0.316$  (Fig. 14.3).

---

<sup>4</sup>In reality the predictive variable  $X$  could be, for example, time or temperature that can be measured with small (in the current context negligible) uncertainty.

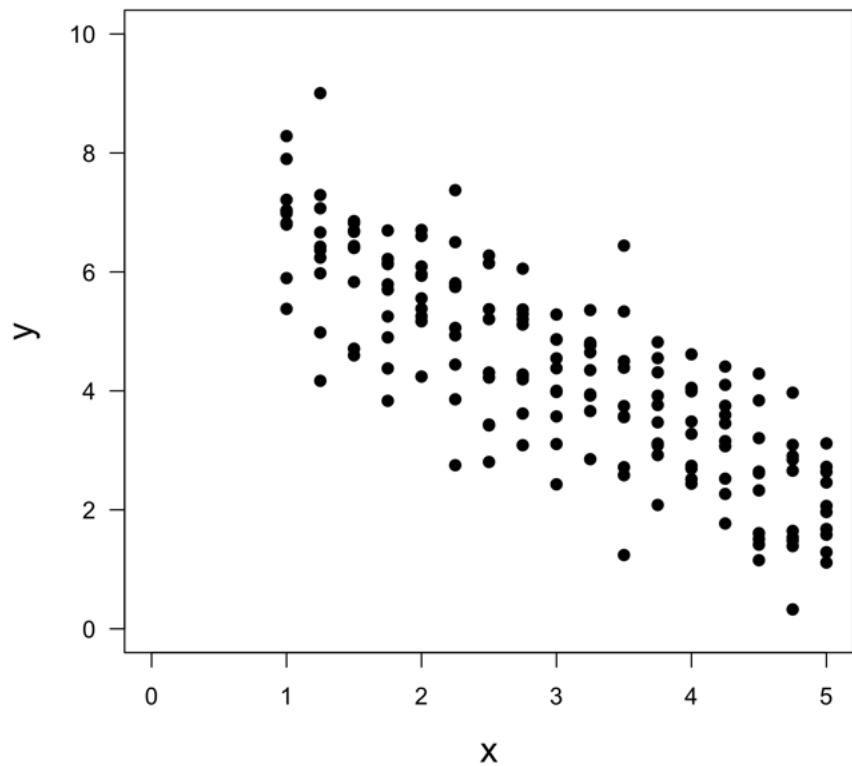


Figure 14.1: 170 artificial data generated using the recipe outlined in Eq. 14.3. This is a textbook example with several (here:  $n = 10$ )  $y$ -data for each  $x$  value. [LSartificialData.R](#)

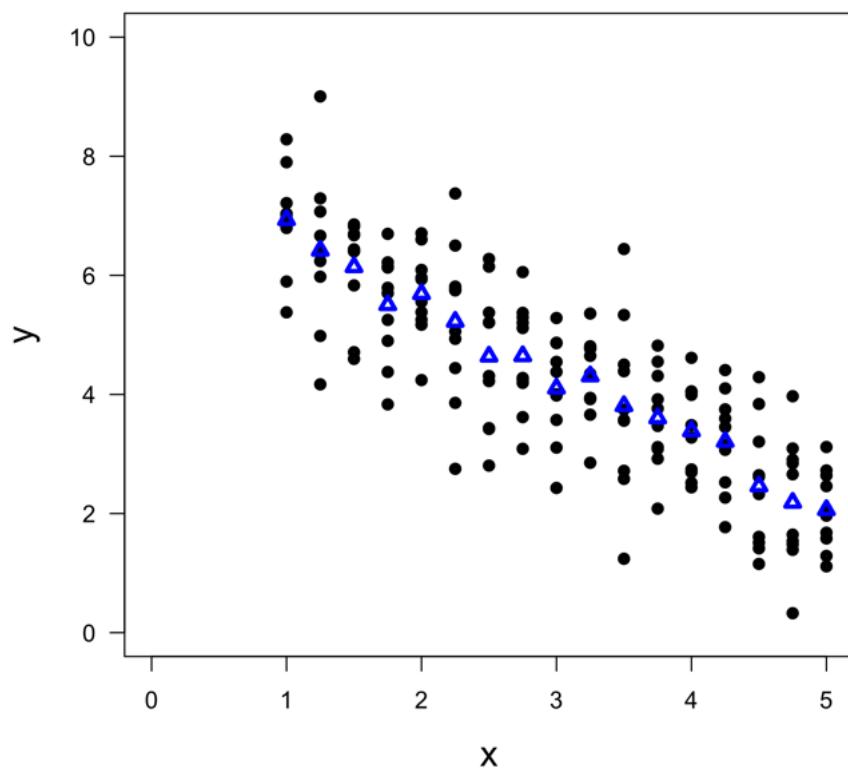


Figure 14.2: Regression to the mean: the data (black points) plus their mean values with respect to  $y$  (i.e. at constant  $x$ ; blue triangles). [LSdataMean.R](#)

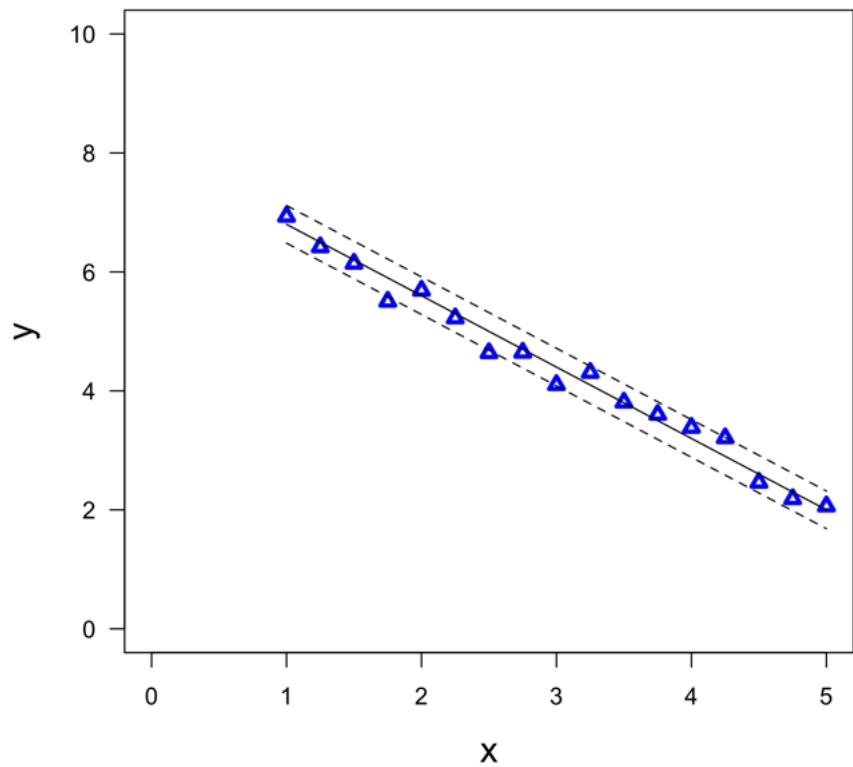


Figure 14.3: Regression to the mean: the mean values with respect to  $y$  (i.e. at constant  $x$ ; blue triangles) lie already close to the exact relationship (black solid line). The exact relationship  $\pm 1$  standard error of the mean are shown as black dashed lines. [LSdataSE.R](#)

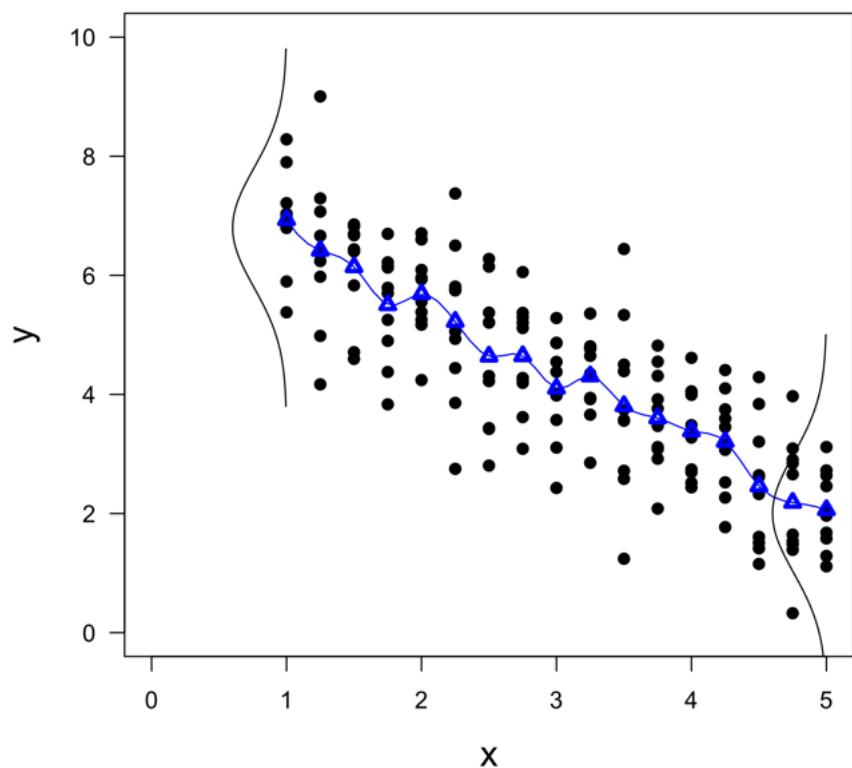


Figure 14.4: Data (black dots), mean values (blue triangles) connected by a cubic spline (blue line), and normal PDFs indicating additive normal noise (black lines). [LSartificialSpline.R](#)

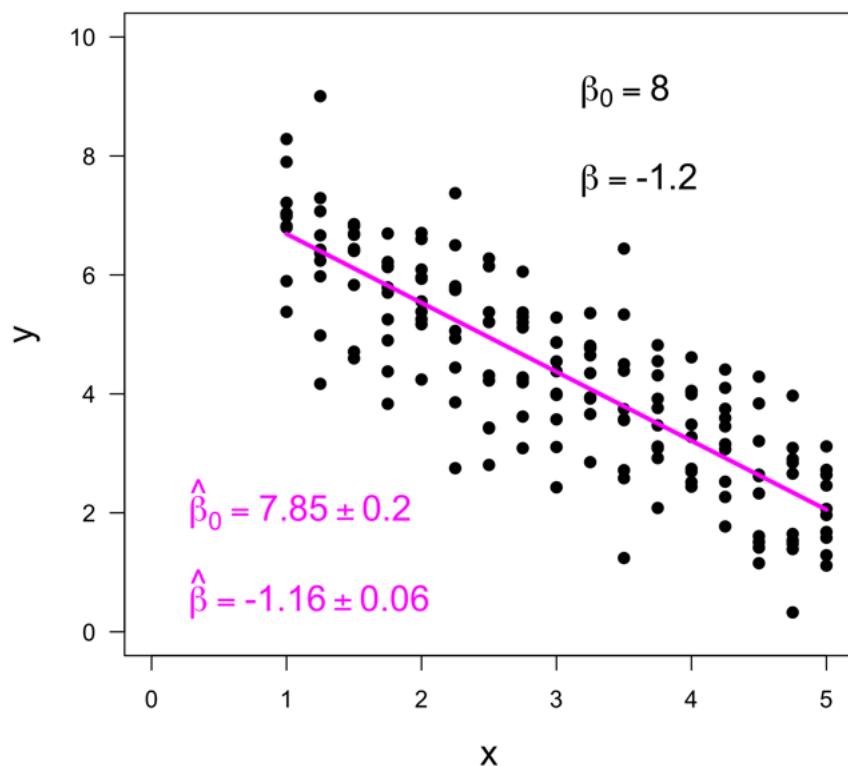


Figure 14.5: Mean values (blue triangles) and the fitted straight line (magenta solid line). The estimated values of the intercept  $\hat{\beta}_0 = b_0 = 7.85 \pm 0.20$  and slope  $\hat{\beta} = b = -1.16 \pm 0.06$  include the exact values  $\beta_0 = 8$  and  $\beta = -1.2$  in their uncertainty ranges. [LSdataEstimates.R](#)

## 14.2 How to estimate intercept & slope: least squares

'The method of least squares is the automobile of modern statistical analysis: despite its limitations, occasional accidents, and incidental pollution, this method and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all.' Stigler (2002, p. 320)

The estimate of intercept and slope provided by the R routine `lm()` (`lm` = linear modeling) is based on the so-called '[method of least squares](#)' (or '[least squares](#)' for short). Here least squares will be introduced in an ad-hoc way which has been done by several authors in the early 19th century<sup>5</sup>. In order to apply least squares, [four prerequisites](#) ('fixed X', independence of data, additive normal noise, homogeneous noise level, discussed in detail below) have to be fulfilled which is not obvious from the ad-hoc introduction. The least squares method can be derived under certain conditions from Bayes' Theorem (Bayesian approach; Section [J.3](#)). This derivation required the four prerequisites mentioned above and paves the way to more general estimation methods like, for example, [general least squares](#).

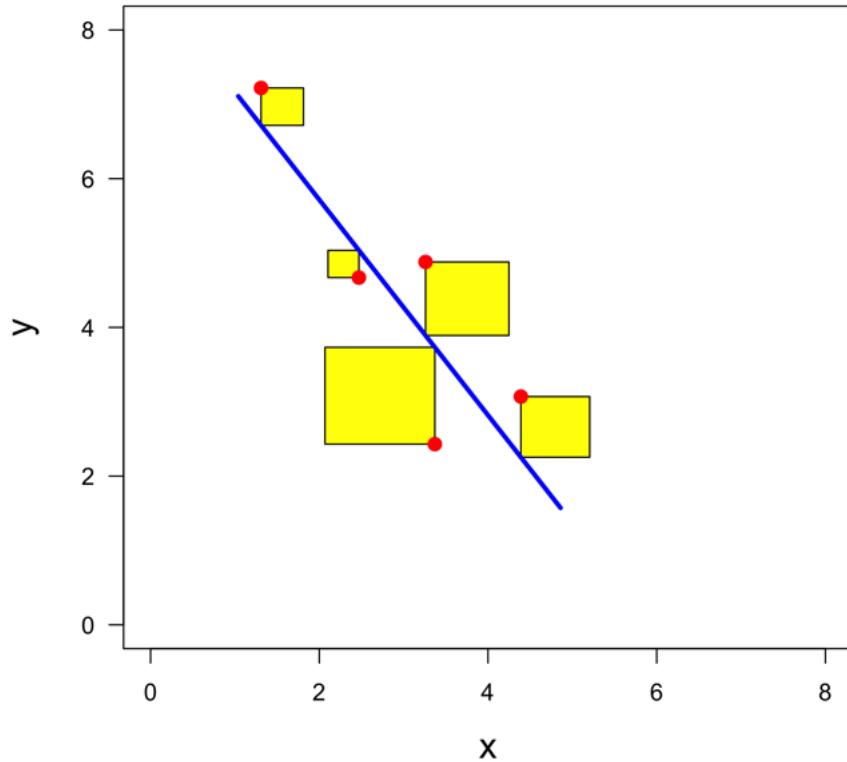


Figure 14.6: The least squares method minimizes the total area of the yellow squares (sum of squares) by varying the intercept and slope of the straight line. [LSconcept.R](#)

The method of least squares is easy to grasp (Fig. 14.6). Obviously, the data points do not lie on a single straight line. The reason for this is the noise included in the  $y$  values. One would like to find a straight line that is as 'close' (in a certain sense) as possible to the data points. How to measure the closeness between line and data?

<sup>5</sup>Least squares has been invented most probably independently by Adrien-Marie Legendre (1805), Robert Adrain (1808), and Carl-Friedrich Gauss (1809; claiming to have been in possession of the method since 1795; compare Stigler, 1999).

One can calculate the distances in  $y$  direction between the data points and the line<sup>6</sup>. For example, for the data point  $(x_k, y_k)$  and the line  $y_L(x_L) = a + b x_L$  this distance  $d_k$  is given by

$$d_k = y_k - y_L(x_k) = y_k - a - b x_k \quad (14.4)$$

where  $a$  and  $b$  are the (still unknown) intercept and slope, respectively. One could have the idea to do this for all data points and then try to minimize the sum over all distances by varying  $a$  and  $b$ . However, this does not work because the distances can be positive or negative and thus they would cancel each other when summing up. One requires that every deviation (independent of sign) counts! A simple way to do so is to square the distances before summing up, i.e. one sums up the area of the yellow squares shown in Fig. 14.6 and tries to find the minimum value for this sum by varying  $a$  and  $b$ . This procedure explains the name 'least squares'.

#### Further remarks:

- Is the square of the distance the only way to make every distance count? No! One could for example use the magnitude of the distance ( $|d_k| = \text{abs}(d_k)$ ), or any even power  $\geq 4$  of the distances ( $d_k^{2m}, m = 2, 3, 4, \dots$ ).
- If these alternatives are available, why then going for the sum of squares? For the sum of squares the resulting functions can be differentiated and it is relatively easy to find an analytical solution for the optimal intercept and slope (Section J.4). Higher powers of the distance would give more weight to the few largest distances. The abs-function is not differentiable at 0. **However, the main reason for using the sum of squares stems from the proper derivation of least-squares from Bayes' Theorem (Section J.3).**

### 14.2.1 Underlying assumptions for simple straight line fitting

According to the [Gauss-Markov theorem](#) the noise has to satisfy three conditions

1. the true mean of the noise is zero ( $\mu = 0$ )
2. no correlation of noise (independence)
3. homoscedasticity (no pattern in the noise; noise can be described by a single variance  $\sigma^2$ )

to prove that the OLS estimator is a [best linear unbiased estimator \(BLUE\)](#).<sup>7</sup>

Unfortunately, some confusion was caused by the article of Osborne & Waters (2002) claiming that four assumptions (including normal distribution of the noise) are required to justify OLS; this article has been cited more than 1600 times (Google Scholar, 2020). The famous four underlying assumptions have been further popularized by, for example, Zuur et al. (2007).<sup>8</sup> It took some time before Williams et al. (2013) corrected some misconceptions.

However, if the four underlying assumptions (see below) are fulfilled, one can show that the OLS estimator is a maximum likelihood estimator (Section J.3): OLS can be derived from the basic rules of probability using flat priors. For further benefits from obeying the four underlying assumptions compare Williams et al. (2013).

The four underlying assumptions read (compare, for example, Zuur et al., 2007, p. 52-53):

1. **Normality of additive noise:** the observations  $y_k, k = 1, 2, \dots, n$  contain noise from a normal distribution with mean  $\mu = 0$  and variance  $\sigma_k^2$
2. **Homogeneity** of the noise level: the variance of the noise is independent of  $x$ , i.e.  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2$ . This is called [homoscedasticity](#) (from Ancient Greek *skedasis* = 'dispersion') as opposed to heteroscedasticity (homoscedastic as opposed to heteroscedastic).

<sup>6</sup>The distances to the optimal straight line are called [residuals](#).

<sup>7</sup>A proof of the Gauss-Markov theorem is given in the Appendix (Section J.2).

<sup>8</sup>[In every course we ask the participants the question: 'Do you understand linear regression?' Three quarters will respond positively, but most will fail to identify the four underlying assumptions of linear regression ...'](#)  
Zuur et al. (2007)

3. **Independence** of the data: the observation of  $y_k$  should have no influence on the observation of  $y_m$  for  $m \neq k$ . This assumption is often violated for time series or spatial data (keyword: auto-correlation).
4. **Fixed X** is a slang term for 'the predictive variable  $X$  is not a stochastic variable'. The values of the predictive variable are assumed to be known exactly or, in reality, with uncertainties that are negligible in the respective context. In biological/ecological investigations, for example, time or temperature can be measured with very small uncertainties, whereas growth or grazing rates can usually be determined only with sizable uncertainties.

The first two underlying assumptions given by Zuur et al. (2007) can be written in compact form as  $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2)$  (read  $\sim$  here as 'stems from').

### 14.2.2 How to test the underlying assumptions?

*How to test the underlying assumptions of simple linear regression (ordinary least squares)? · · · plot residuals, plot residuals, plot residuals · · · and test for homo- versus heteroskedasticity*

Residuals are the differences between the observations  $y_k$  and the corresponding points  $\hat{\beta}_0 + \hat{\beta} x_k$  on the fitted straight line ( $\hat{\beta}_0$  is the estimate of the intercept,  $\hat{\beta}$  is the intercept of the slope). First one plots the residuals over  $x$  and looks for pattern in their distribution (Fig. 14.9). 'No pattern' is a good sign because it speaks for homogeneity of the variance of the noise (assumption 2).

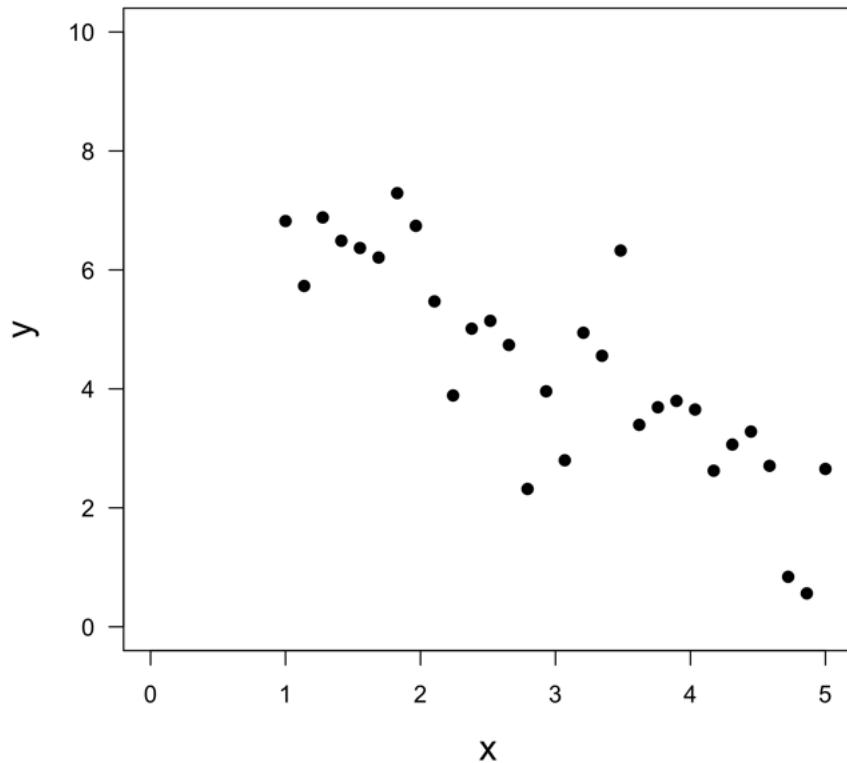


Figure 14.7: Plot of the  $n = 30$  data pairs  $(x_j, y_j)$ . [LS-example1.R](#) (line 21: set sflag to 1)

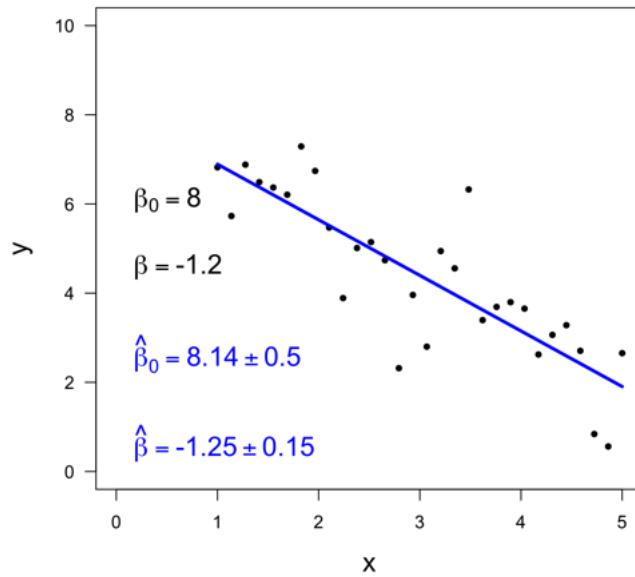


Figure 14.8: The data (black dots) and the fitted straight line (blue solid line). The estimates of the intercept,  $\hat{\beta}_0 = 8.14 \pm 0.50$ , and the slope,  $\hat{\beta} = -1.25 \pm 0.15$ , encompass the exact values  $\beta_0 = 8$  and  $\beta = -1.2$  in their uncertainty ranges. [LS-example1.R](#) (line 21: set sflag to 2)

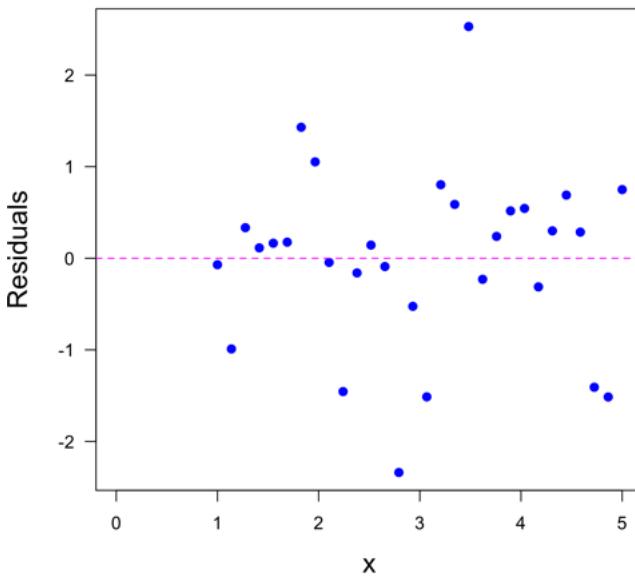


Figure 14.9: The residuals over  $x$  show no clear pattern and thus homogeneity of the noise seems to be fulfilled. A few residuals are quite large (about  $2\sigma$ ), however, still in an acceptable range. [LS-example1.R](#) (line 21: set sflag to 3)

If homogeneity is given, we can plot a histogram of all residuals which should look 'normal' if assumption 1 (normality of the noise) is fulfilled. If the data size is large enough (say  $n \geq 30$ ), one can apply a normality test (Kolmogorov-Smirnov) or even estimate the density (PDF) of the residuals.

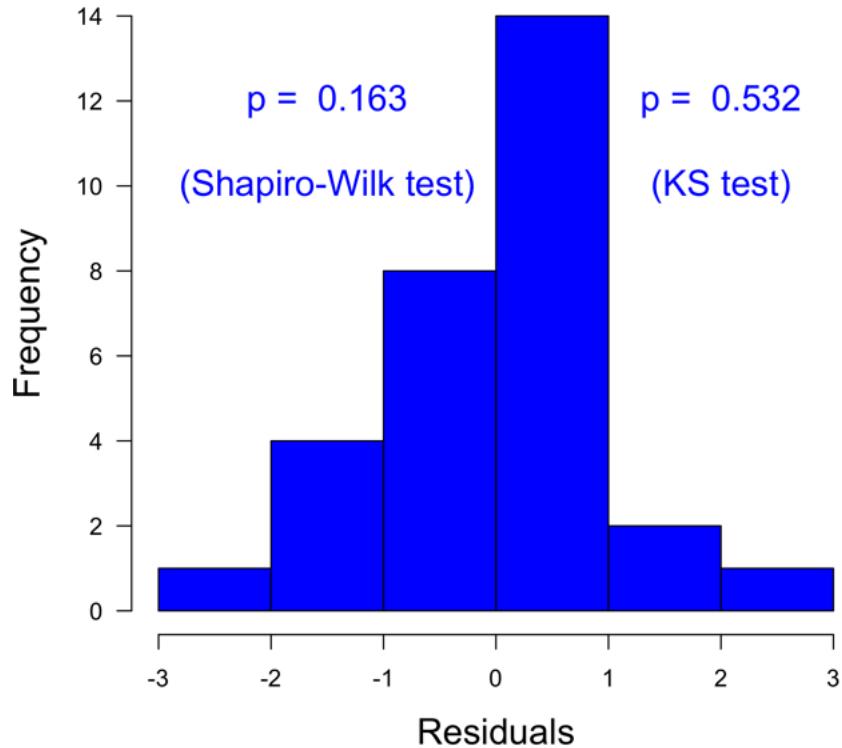


Figure 14.10: The histogram of the residuals shows a single maximum near zero (as expected for a normal distribution with  $\mu = 0$ ). The Shapiro-Wilk test yields a  $p$ -value of 0.163 and the KS-test yields a  $p$ -value of 0.532 and thus the null hypothesis 'standardized sample is from the standard normal distribution' is not rejected based on the usual  $\alpha = 0.05$ . [LS-example1.R](#) (line 21: set sflag to 4)

**Test for homo- versus heteroskedasticity** Several packages provide R routines for testing homo- versus heteroskedasticity. Here is the code using the routine `ncvTest()` from package `car`: [SkedasticityTest.R](#)

An alternative is the routine `breusch_pagan()` from package `skedastic` (Breusch-Pagan test).

**Exercise 42 Ordinary least squares with asymmetric noise (\*)**

According to the Gauss-Markov theorem it is not necessary that the noise stems from a normal distribution, i.e.  $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2)$  (read  $\sim$  here as 'stems from'); it is sufficient that the mean of the noise is zero, i.e.  $\mu = 0$ , no correlation of noise, and homoscedasticity. Thus the noise can even be asymmetrically distributed.

(1) Construct a PDF  $f(z; a, b)$  consisting of two parts: a uniform distribution over  $-2 \leq z \leq 0$  with density  $a$  and another uniform distribution over  $0 \leq z \leq 1$  with density  $b$ . The joint distribution should possess mean  $\mu = 0$ .

(2) Calculate the variance of the constructed PDF.

(3) Develop a simple procedure to take random samples from the constructed PDF.

(4) Test the procedure by a Monte Carlo simulation.

(5) Generate artificial data pairs  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  with noise from the constructed PDF, apply linear regression (ordinary least squares), and investigate how good slope and intercept can be estimated depending on the sample size  $n$ . To compare results it is suggested to use the true slope  $\beta_1 = -1.2$  and true intercept  $\beta_0 = 8$ ,  $1 \leq x \leq 5$ ;  $n = 10^m$ ,  $m = 1, 2, 3, \dots$

## 14.3 Confidence bands for simple linear regression

Although from the residuals of simple linear regression (SLM) one can estimate the variance of the noise and a standard error of the mean, these quantities give only rough estimates of more specific questions: What's the confidence interval (CI) for observed response ( $Y$ ) values at a given predictor ( $X$ ) value and chosen level of significance,  $\alpha$ ? How to construct confidence bands (boundaries around the estimated straight line for a whole interval) with a probability  $(1 - \alpha) \times 100\%$  that the estimated line lies within these bands? The first question has a simple answer (Subsection 14.3.1). The second question has more than one answer which has to do with the fact that one has the freedom to choose between different forms of the bands (mainly hyperbolic, Subsection 14.3.2, or straight, Subsection 14.3.3). Finally, we will compare these bands with an approximation based on the single point confidence interval.

**Further reading:** Scheffé (1953)

### 14.3.1 Confidence interval for a single point

The  $100(1 - \alpha)\%$  confidence interval for  $\beta_0 + \beta_1 x_0$  (at a single point,  $x = x_0$ ) is given by

$$\begin{aligned} & \hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \\ & \leq \beta_0 + \beta_1 x_0 \leq \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \end{aligned} \quad (14.5)$$

(Casella & Berger, 2002, p. 558) where  $n$  is the sample size,  $\alpha$  is the chosen level of significance (often 0.05 or 0.1 leading to 95% or 90% confidence intervals, respectively),  $\beta_0$  is the true intercept,  $\beta_1$  is the true slope,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the least squares estimates of  $\beta_0$  and  $\beta_1$ ,  $t_{n-2,\alpha/2}$  is the critical  $t$  value for  $p = \alpha/2$  and  $n - 2$  degrees of freedom,  $\bar{x}$  is the sample mean,  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  is the sum of squares for  $x - \bar{x}$ ; the **standard error of estimate**,  $S$ , (also called **standard error of the regression**; Zar, 2010, p. 340) is calculated from the residuals,  $r_i$ , of the linear regressions as

$$S = \frac{\sum_{i=1}^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 (x_i - \bar{x})]^2}{n - 2} = \frac{\sum_{i=1}^n r_i^2}{n - 2} \quad (14.6)$$

This answers our first question.

Note that the length of the confidence interval at a single point  $x = x_0$  depends on all data  $x_1, \dots, x_n$  via  $(x_0 - \bar{x})^2 / S_{xx}$  and also on  $y_1, \dots, y_n$  via  $S$ . It is minimal for  $x_0 = \bar{x}$  and increases from there towards smaller and larger  $x_0$ . An example is shown in Fig. 14.11 (red vertical line). If varying  $x_0$  from minimal  $x$  to maximal  $x$  values, the upper boundary of the confidence intervals would form a hyperbolic curve above the estimated regression line (compare Subsection 14.3.4).

### 14.3.2 Scheffé bands (hyperbolic)

One answer to the second question has been provided by Scheffé (1959). The probability is at least  $1 - \alpha$  that

$$\begin{aligned} & \hat{\alpha} + \hat{\beta} \textcolor{blue}{x} - M_\alpha S \sqrt{\frac{1}{n} + \frac{(\textcolor{blue}{x} - \bar{x})^2}{S_{xx}}} \\ & < \alpha + \beta \textcolor{blue}{x} < \hat{\alpha} + \hat{\beta} \textcolor{blue}{x} + M_\alpha S \sqrt{\frac{1}{n} + \frac{(\textcolor{blue}{x} - \bar{x})^2}{S_{xx}}}. \end{aligned} \quad (14.7)$$

simultaneously for all  $x$ , where  $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$  (Casella & Berger, 2002, p. 560).

The expression (14.7) looks very similar to (14.5) except that  $t_{n-2,\alpha/2}$  has been replaced by  $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$  and that (14.7) is valid for all  $x$  and not just for a single point  $x = x_0$ . The width of the confidence

band depends on all data  $x_1, \dots, x_n$  via  $(x_0 - \bar{x})^2 / S_{xx}$  and also on  $y_1, \dots, y_n$  via  $S$ . It is minimal for  $x_0 = \bar{x}$  and increases from there towards smaller and larger  $x_0$  forming two hyperbolic curves above and below the estimated straight line (Fig. 14.11).

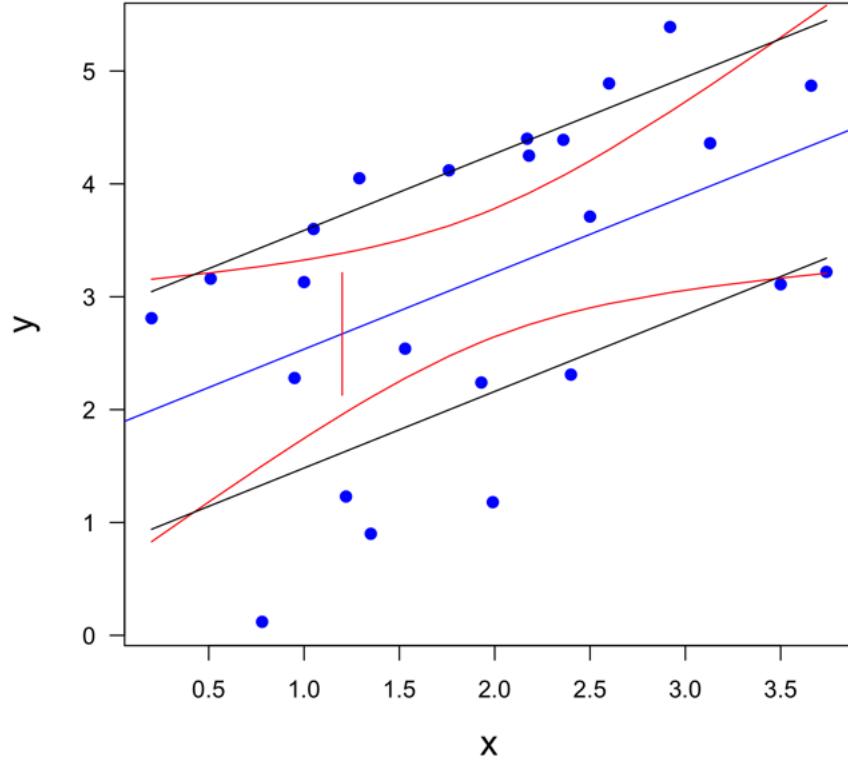


Figure 14.11: Simple linear regression (SLR): confidence interval and confidence bands for  $\alpha = 0.1$ . A straight line (blue solid line) has been fitted to data  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  (blue dots). A confidence interval for a single point  $x = x_0 = 1.2$  is plotted (vertical red line). The hyperbolic Scheffé bands (red lines) are valid for all  $x$ . The straight Gafarian bands (black lines) are valid for the range  $\min(x) \leq x \leq \max(x)$ . Please note that the confidence interval at  $x = x_0 = 1.2$  is smaller than the Scheffé band value at  $x_0$ . Data source: Casella & Berger (2002, p. 542, Table 11.3.1). [ConfidenceBands1.R](#) (line 208: set sflag to 2)

#### Exercise 43 Minimum width of Scheffé confidence band as function of sample size

The Scheffé confidence band is defined by

$$\hat{\alpha} + \hat{\beta} \textcolor{blue}{x} - \textcolor{blue}{M}_\alpha S \sqrt{\frac{1}{n} + \frac{(\textcolor{blue}{x} - \bar{x})^2}{S_{xx}}} \\ < \alpha + \beta \textcolor{blue}{x} < \hat{\alpha} + \hat{\beta} \textcolor{blue}{x} + \textcolor{blue}{M}_\alpha S \sqrt{\frac{1}{n} + \frac{(\textcolor{blue}{x} - \bar{x})^2}{S_{xx}}}.$$

- (1) Derive an expression for the minimum width of the band.
- (2) Discuss how the minimum width varies with  $n$ .

### 14.3.3 Gafarian bands (straight)

Gafarian (1964) has developed a method to construct straight confidence bands that are valid over a chosen interval  $[x_L, x_u]$ . Unfortunately, the constant band width  $2\delta s$  is defined only implicitly by

$$P \left\{ |(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x| \leq \delta \sqrt{S_{xx}}, \forall x \in [x_L, x_u] \right\} = 1 - \alpha \quad (14.8)$$

where  $s$  is the [standard deviation of the residuals](#) of the least squares fit:

$$s = \sqrt{\frac{\sum_{i=1}^n [y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \bar{x})]^2}{n - 2}} \quad (14.9)$$

Gafarian (1964, p. 187).  $\delta$  and  $\alpha$  are related via a two-dimensional integral over a parallelogram (a concise description is given in Miller, 1981).  $\delta$  for a desired  $\alpha$  can be calculated numerically by iteration (compare R code for Fig. 14.11). The straight Gafarian confidence bands for the example data from Casella & Berger (2002) is shown in Fig. 14.11 (black lines).

Bowden & Graybill (1966) derived expressions for calculation of straight confidence bands for arbitrary intervals (i.e., a generalization compared to Gafarian, 1964). In addition, the algorithm for calculating of the band width is somewhat easier to implement (still iteration, however, integration over a rectangular region; compare R code for Fig. 14.11).

### 14.3.4 Naive approximation

A somewhat naive approach consists of constructing a confidence band by using the single point confidence interval (14.5) at all  $x$ . As already discussed in Subsection 14.3.1, this will yield hyperbolic curves. How do they compare to the Scheffé bands? The only difference is the replacement of  $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$  (Scheffé) by  $t_{n-2,\alpha/2}$  ('single point CI approximation'). For  $\alpha = 0.1$   $t_{n-2,\alpha/2}$  is about 25% smaller than  $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$ ; for smaller  $\alpha$  the underestimation is smaller.

#### Exercise 44 Narrow Gafarian confidence bands

*What will happen when calculating Gafarian confidence bands over narrower and narrower intervals? One could guess that the band width should approach the size of the confidence interval for a single point  $x = x_0$ . In order to support (or reject) this hypothesis, construct Gafarian confidence bands for the Casella & Berger (2002) data (Fig. 14.11) for the interval  $1.1 \leq x \leq 1.3$ , i.e. around  $x_0 = 1.2$ .*

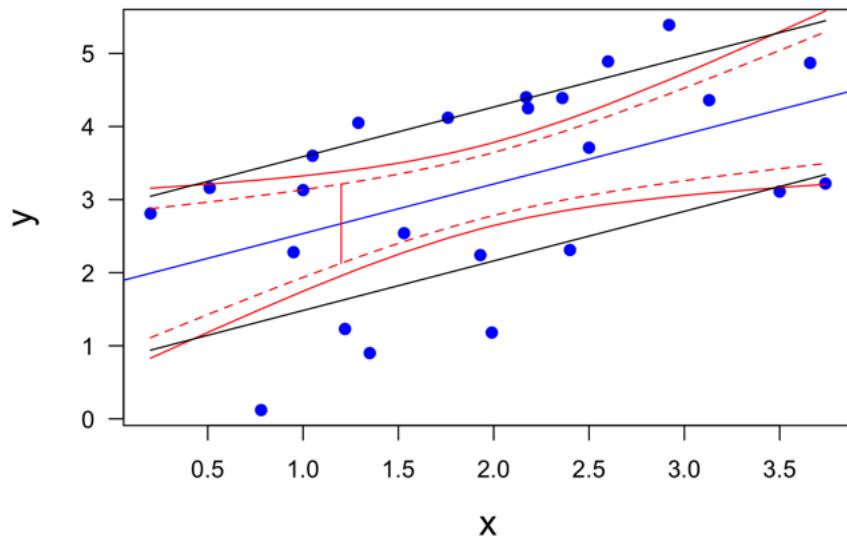


Figure 14.12: Simple linear regression (SLR): confidence interval and confidence bands for  $\alpha = 0.1$ . A straight line (blue solid line) has been fitted to data  $(x_i, y_i), i = 1, 2, \dots, n$  (blue dots). A confidence interval for a single point  $x = x_0 = 1.2$  is plotted (vertical red line). The hyperbolic Scheffé bands (red lines) are valid for all  $x$ . The straight Gafarian bands (black lines) are valid for the range  $\min(x) \leq x \leq \max(x)$ . Please note that the confidence interval at  $x = x_0 = 1.2$  is smaller than the Scheffé band width at  $x_0$ . The 'single point confidence interval approximation' (broken red line) underestimates the band widths of the Scheffé bands by about 25%. Data source: Casella & Berger (2002, p. 542, Table 11.3.1). [ConfidenceBands1.R](#) (line 208: set sflag to 1)

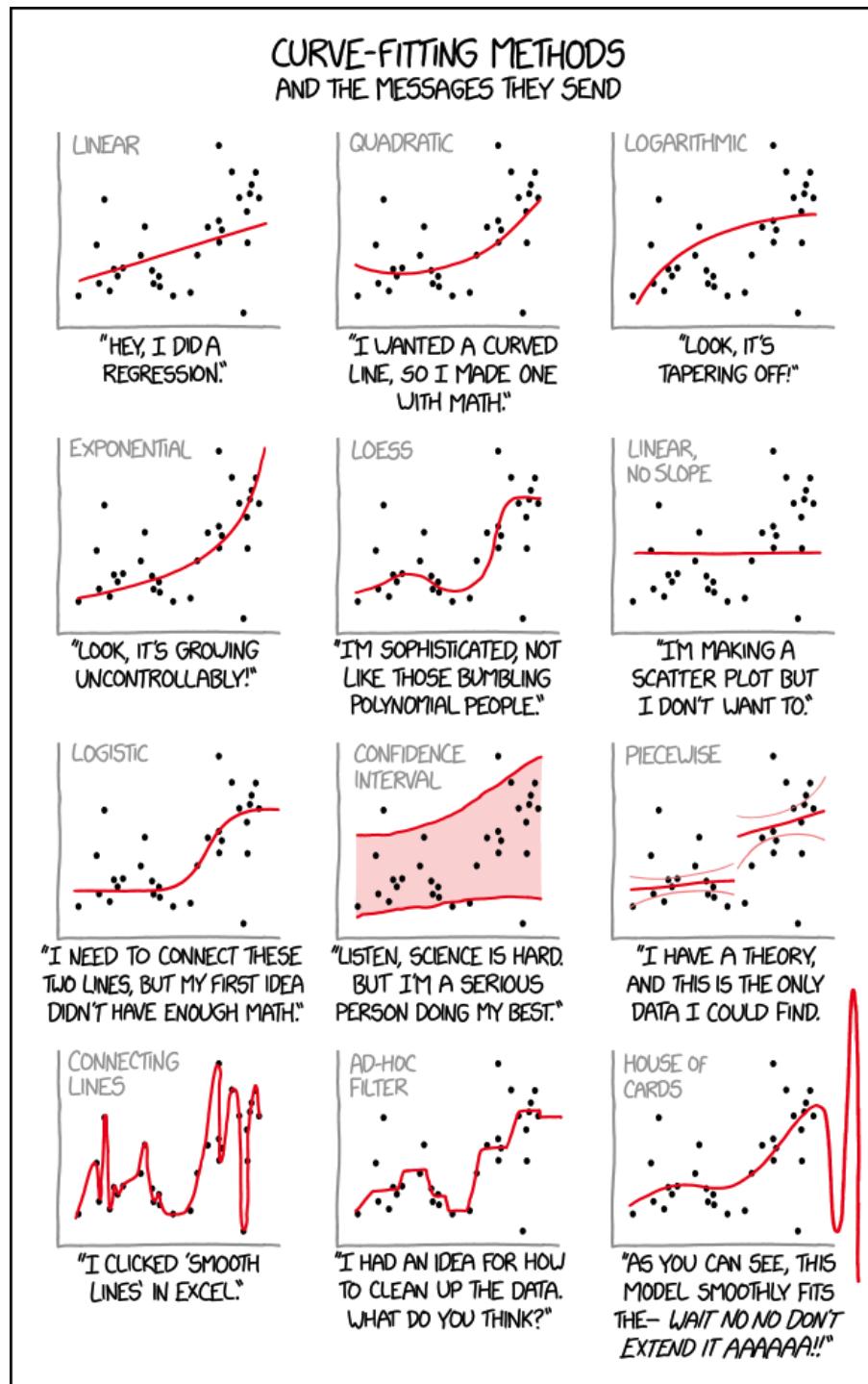


Figure 14.13: Curve fitting methods & messages. Source: <https://xkcd.com/2048/>

## 14.4 Straight line through origin

Sometimes the straight line that one wants to fit to the data has to go through the origin, i.e. the intercept is zero. How can one force the fitting routine `lm()` to fit such a line? Answer: instead of '`lm(y ~ x)`' one has to call '`lm(y ~ x - 1)`' where the '-1' has the meaning 'without intercept' (and not 'subtract 1 from x'). Results of linear fits to artificial data are shown in Fig. 14.14. The black broken line is the result of simple linear regression; the estimated intercept  $\hat{\beta}_0 = -2.02 \pm 1.14$  is significantly different from zero. The red solid line goes through the origin; its estimated slope of  $2.42 \pm 0.12$  is different from the slope of the unconstrained regression line ( $\hat{\beta}_{\text{unconstrained}} = 2.78 \pm 0.23$ ).

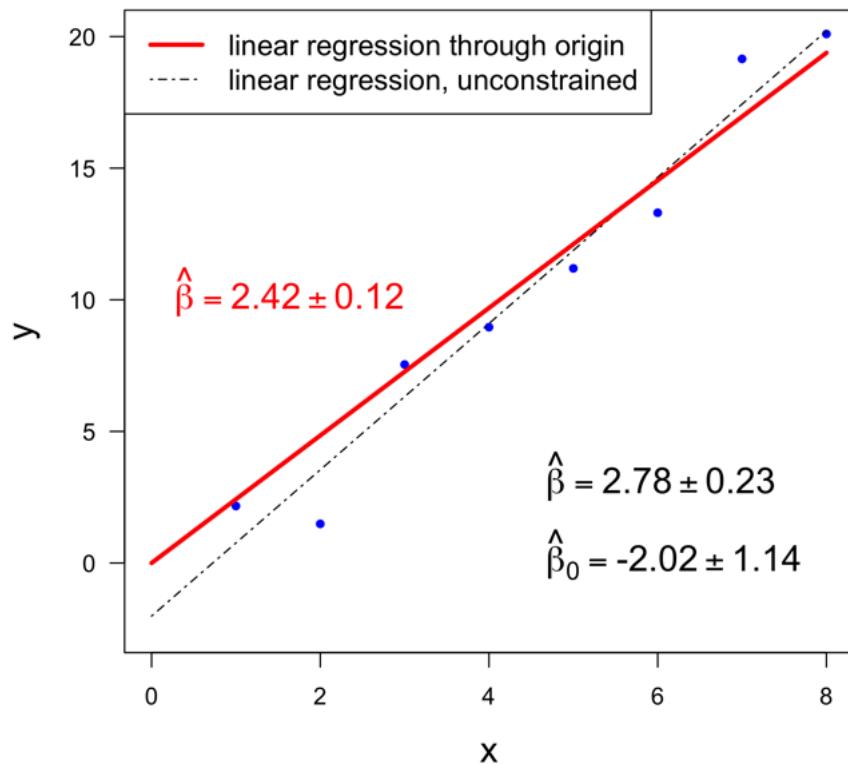


Figure 14.14: Results of linear fits to artificial data: The black broken line is the result of simple linear regression; the estimated intercept  $\hat{\beta}_0 = -2.02 \pm 1.14$  is significantly different from zero. The red solid line goes through the origin; its estimated slope of  $2.42 \pm 0.12$  is different from the slope of the unconstrained regression line. [LStthroughOrigin.R](#)

## 14.5 Which is the limiting nutrient? Redfield ratio

Marine microalgae require nitrate ( $\text{NO}_3$ ) and phosphate ( $\text{PO}_4$ ) to grow<sup>9</sup>. Alfred Redfield recognized already in the 1930ies (Redfield, 1934) that both the molar ratio of nitrate and phosphate in seawater and the molar ratio of nitrogen (N) and phosphorous (P) in phytoplankton are on average over larger volumes often close to 15:1 (Redfield 1958, 1963). In the surface ocean phosphate is often depleted to a limiting concentration before nitrate. There has been a long debate on the mechanisms keeping the  $\text{NO}_3:\text{PO}_4$  ratio in seawater around 15:1 (compare, for example, Tyrrell, 1999, and references therein) and on the question 'What is the limiting (macro)nutrient?'. In this section a modern compilation of nutrient data will be analyzed (and we will learn how to read netCDF files). Straight lines will be fitted to the data ( $\text{NO}_3$  over  $\text{PO}_4$ ) using different approaches (making different assumptions):

- (1) simple linear regression, i.e.  $\text{PO}_4$  is considered as 'fixed' variable (non-stochastic) and the 'response'  $\text{NO}_3$  is stochastic, or
- (2) Deming regression where both  $\text{PO}_4$  and  $\text{NO}_3$  are considered as stochastic variables. The second approach can be considered as more appropriate than the first.

### 14.5.1 World Ocean Atlas data

Global ocean annual mean nutrient data are available on a  $1^\circ \times 1^\circ$  grid on 33 vertical levels (World Ocean Atlas, 2009; Garcia et al., 2010):

<http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NODC/.WOA09/.Grid-1x1/.Annual/>

Please download the netCDF files 'nitrate\_annual\_1deg.nc' and 'phosphate\_annual\_1deg.nc' (60 MByte each).

The netCDF files can be opened in R using the routine `nc_open()` of the package `ncdf4`:

```
install.packages('ncdf4') # install package (apply only once on your PC)
library(ncdf4)
info.nc = nc_open('nitrate_annual_1deg.nc')
```

Look at the data! First, the surface concentrations of nitrate (Fig. 14.15) and phosphate (Fig. 14.16) are plotted.

---

<sup>9</sup>This is a somewhat simplified view because some of the microalgae (where cyanobacteria are included as well) can use also other sources of N and P like ammonia,  $\text{N}_2$  (nitrogen fixation!), or organic nutrients.

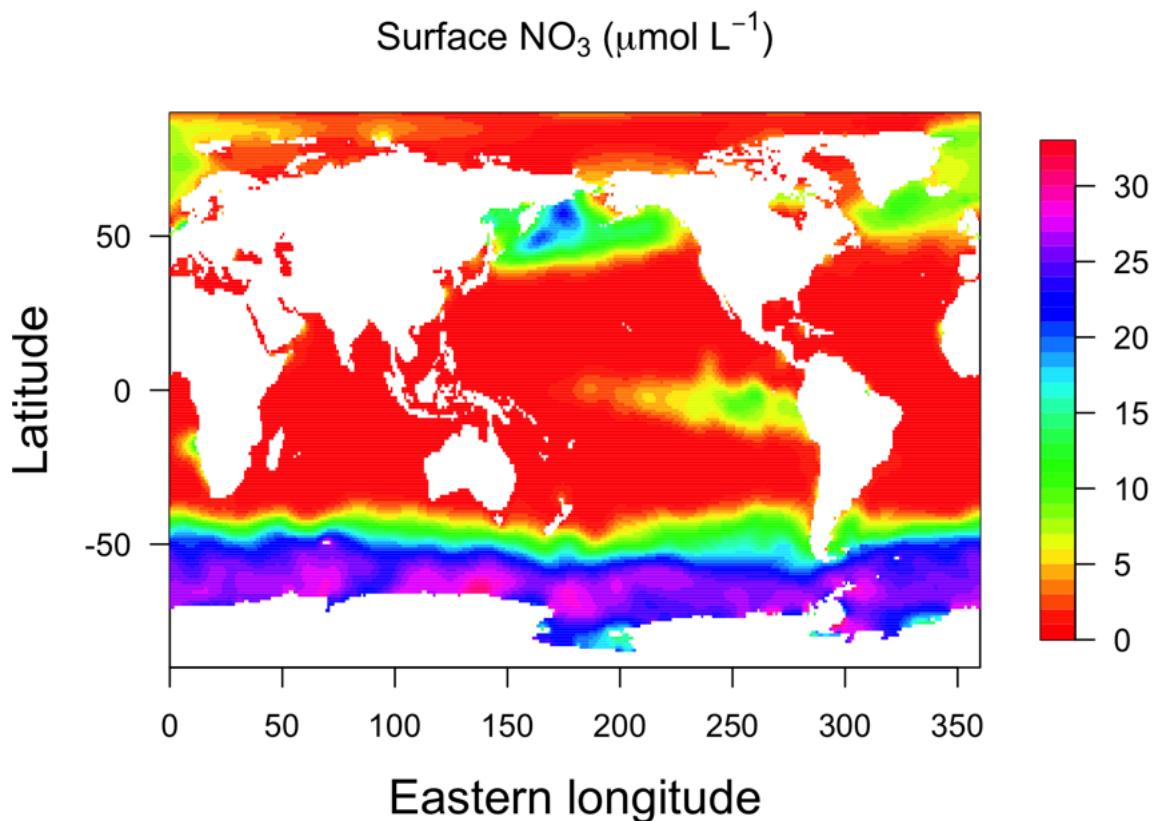


Figure 14.15: Annual mean nitrate (NO<sub>3</sub>) concentration (μmol L<sup>-1</sup>) in the surface ocean. Areas with high year-round nutrients (NO<sub>3</sub>, PO<sub>4</sub>) concentrations, namely the Southern Ocean, the northern North Pacific, and the equatorial Pacific, are called High-Nutrient Low-Chlorophyll (HNLC) regions. [NO3surfaceOcean.R](#)

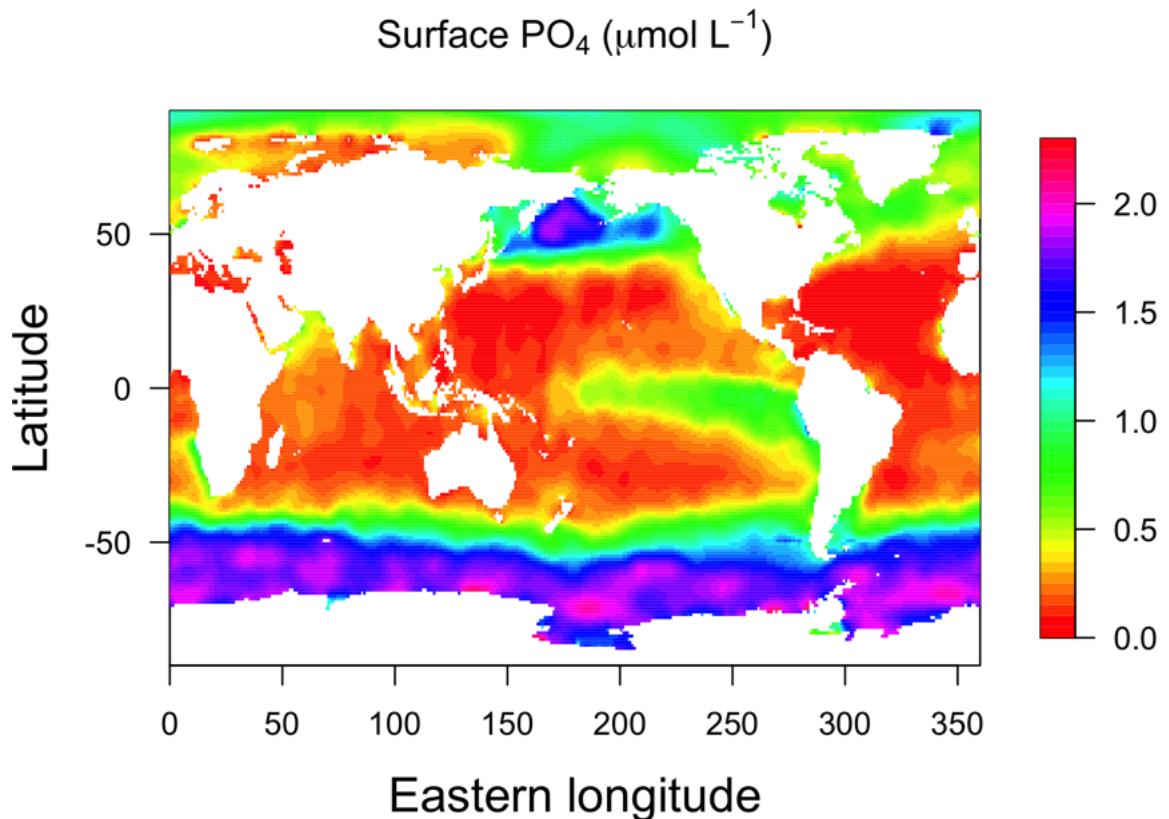


Figure 14.16: Annual mean phosphate (PO<sub>4</sub>) concentration (μmol L<sup>-1</sup>) in the surface ocean. Areas with high year-round nutrients (NO<sub>3</sub>, PO<sub>4</sub>) concentrations, namely the Southern Ocean, the northern North Pacific, and the equatorial Pacific, are called High-Nutrient Low-Chlorophyll (HNLC) regions. [PO4surfaceOcean.R](#)

### 14.5.2 Molar nitrate to phosphate ratio: simple linear regression

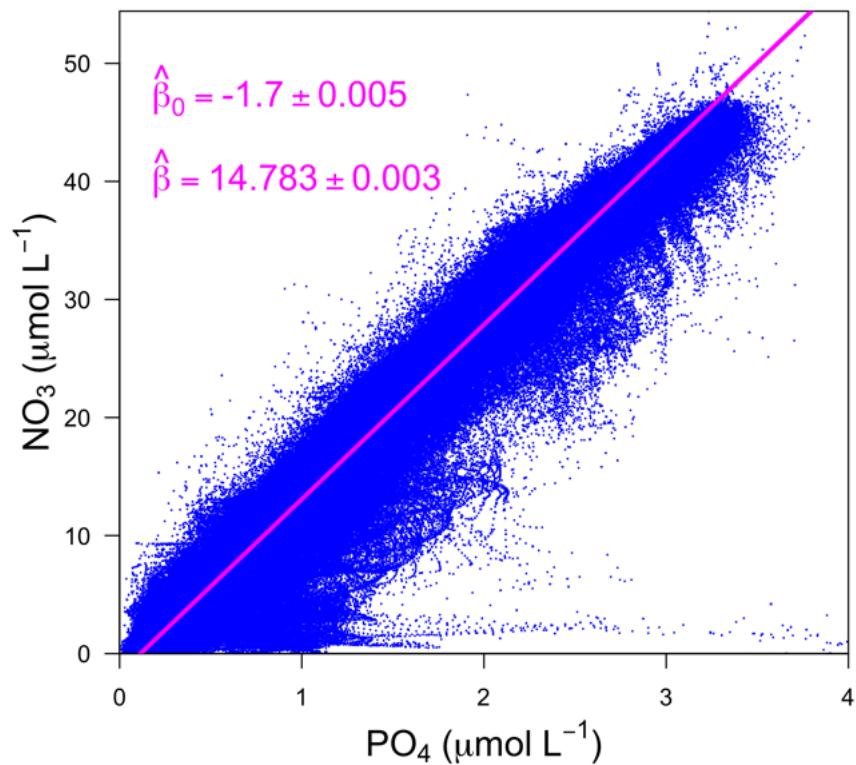


Figure 14.17: Nitrate ( $\text{NO}_3$ ) versus phosphate ( $\text{PO}_4$ ): plot of all data (blue dots) from the World Ocean Atlas (Garcia et al., 2010) and straight line fit by simple linear regression. The estimated uncertainties for the slope and the intercept (standard errors) are very small because of the large sample size ( $n > 40000$ ). The estimate of the intercept,  $\hat{\beta}_0$ , is negative indicating that most often nitrate is depleted before phosphate is used up. [SLR-NO3overPO4.R](#)

**Exercise 45 Residuals of  $\text{NO}_3:\text{PO}_4$  regression**

Inspect the residuals of the  $\text{NO}_3:\text{PO}_4$  simple linear regression shown in Fig. 14.17. Do you expect a symmetric density of the residuals? Are the residuals normally distributed? Do the residuals show a pattern over the range of phosphate values? Estimate the variance of the noise.

## 14.6 Inverse prediction in the context of paleoproxies

A quantitative analysis of any environment older than the instrumental record relies on proxies as, for example, isotopic compositions of biominerals or organic material archived in marine sediments. Calibration data as, for example, isotopic composition ( $\delta$  values) depending on environmental temperature can be obtained from laboratory experiments, i.e. the uncertainty in temperature is negligible and thus simple linear regression of  $\delta$  on temperature is appropriate. Non the less the inversion of such a relationship, i.e. the estimate of paleo-temperatures from sedimentary  $\delta$  values, is not trivial because the uncertainty in  $\delta$  is usually not only (normal) statistical noise (from calibration data) but may in part due to the influence of other variables not taken into account during calibration. For further discussion of inverse prediction see McClelland et al. (2021).

## 14.7 Fit exponential function

In this section an exponential function will be fitted to a given data set, i.e. the ideal model reads

$$y(x) = \alpha e^{\beta x} \quad (14.10)$$

where  $\alpha$  and  $\beta$  are the (unknown) model parameters. Why is this topic discussed in the chapter on linear regression given the fact that the exponential function is non-linear? Before applying fitting routines one transforms the  $y$  data by taking the natural logarithm. This leads to

$$\ln y(x) = \ln \alpha + \beta x = \beta_0 + \beta x \quad (14.11)$$

for which one can apply simple linear regression for  $\ln y(x)$  instead of  $y(x)$ .

However, this method has to be applied with care. If one assumes that  $y$  contains additive noise, i.e.

$$y(x) = \alpha e^{\beta x} + \text{noise}, \quad (14.12)$$

then by taking the logarithm of the observed  $y$  values one also transforms the additive noise and one has to check whether the transformed noise is distributed normally and with a homogeneous noise level (2 of the 4 prerequisites for applying simple linear regression).

Two examples using artificial data sets are discussed here. In the first example the  $x$ -values are chosen as  $x = \{1, 2, \dots, 30\}$  and the corresponding  $y$ -values (Fig. 14.18) are generated from the model

$$y(x) = \alpha e^{\beta x} + \text{noise} = 1.2 e^{0.03x} + \mathcal{N}(\mu = 0, \sigma = 0.5), \quad (14.13)$$

i.e.  $\alpha = 1.2$ ,  $\beta = 0.03$ . After log-transformation of the  $y$  values one fits a straight line to  $(x, \ln y)$  yielding an intercept  $\hat{\gamma} = 0.157 \pm 0.115$  and a slope  $\hat{\beta} = 0.0294 \pm 0.0065$  and thus  $\hat{\alpha} = 1.17 \pm 0.14$ . The true values  $\alpha$  and  $\beta$  are well within the  $\pm 1\sigma$  ranges of the estimates  $\hat{\alpha}$  and  $\hat{\beta}$ .

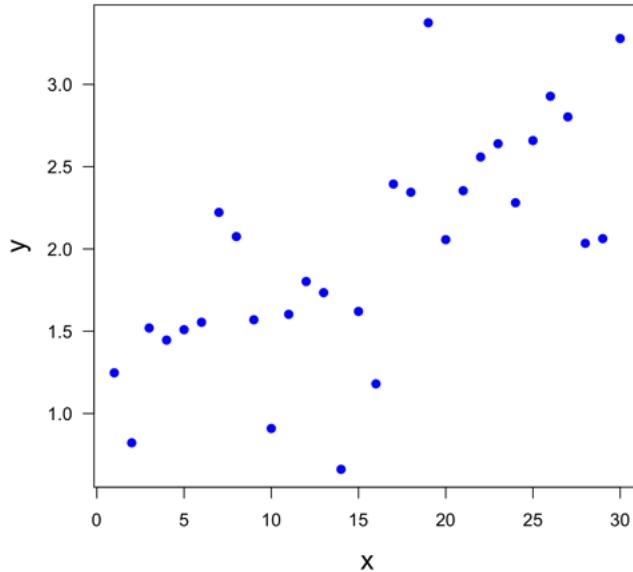


Figure 14.18: Data generated from the model  $y(x) = \alpha e^{\beta x} + \text{noise} = 1.2 e^{0.03x} + \mathcal{N}(\mu = 0, \sigma = 0.5)$ . [ExpFctFit.R](#)

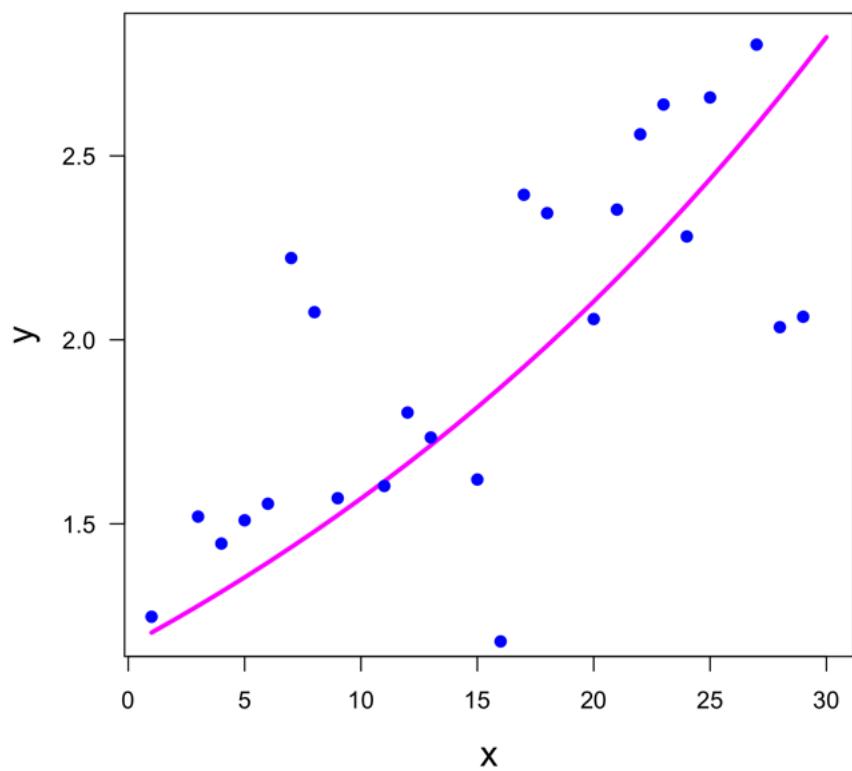


Figure 14.19: Data (blue dots) and exponential fit function (red solid line). [ExpFctFit.R](#) (line 10: set sflag to 2)

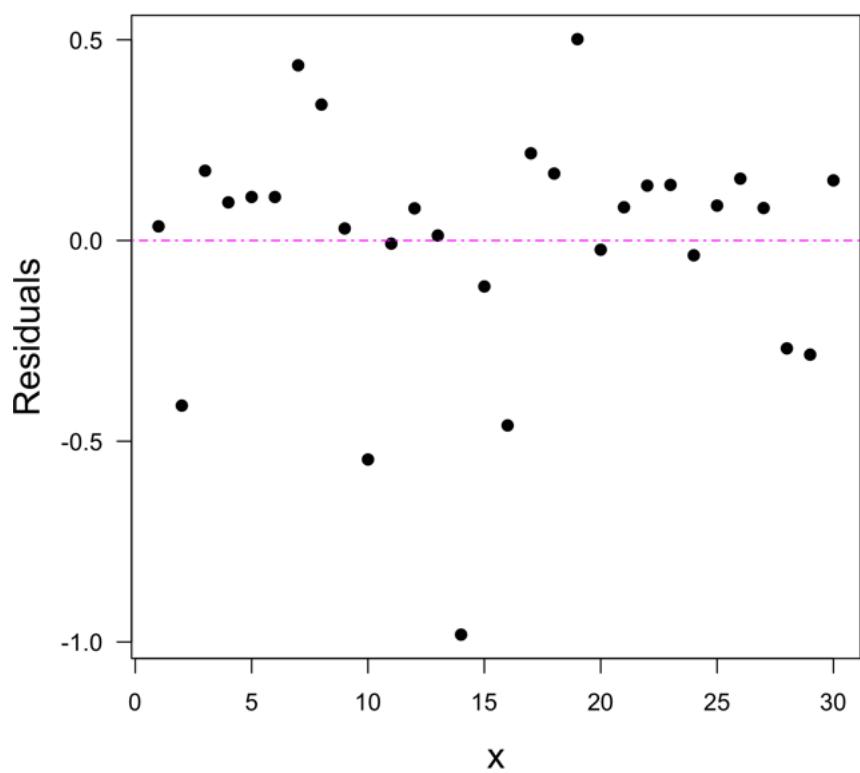


Figure 14.20: Residuals of simple linear regression for  $(x, \ln y)$ . [ExpFctFit.R](#) (line 10: set sflag to 3)

## 14.8 Fit polynomial to data

Polynomials

$$y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots = \sum_{k=0}^K \beta_k x^k \quad (14.14)$$

are also linear models in the sense that they are linear in their coefficients  $\beta_k$ . If we again assume that the predictor  $x$  is non-stochastic (or small uncertainties can be neglected in the current context), the noise in  $y$  is additive and stems from a normal population with mean  $\mu = 0$  and unknown variance  $\sigma^2$  that does not vary with  $x$  (homoscedasticity or 'no pattern in noise'), we can again apply ordinary least-squares by applying the R routine `lm()`. An example using artificial data is shown in Fig. J.5.

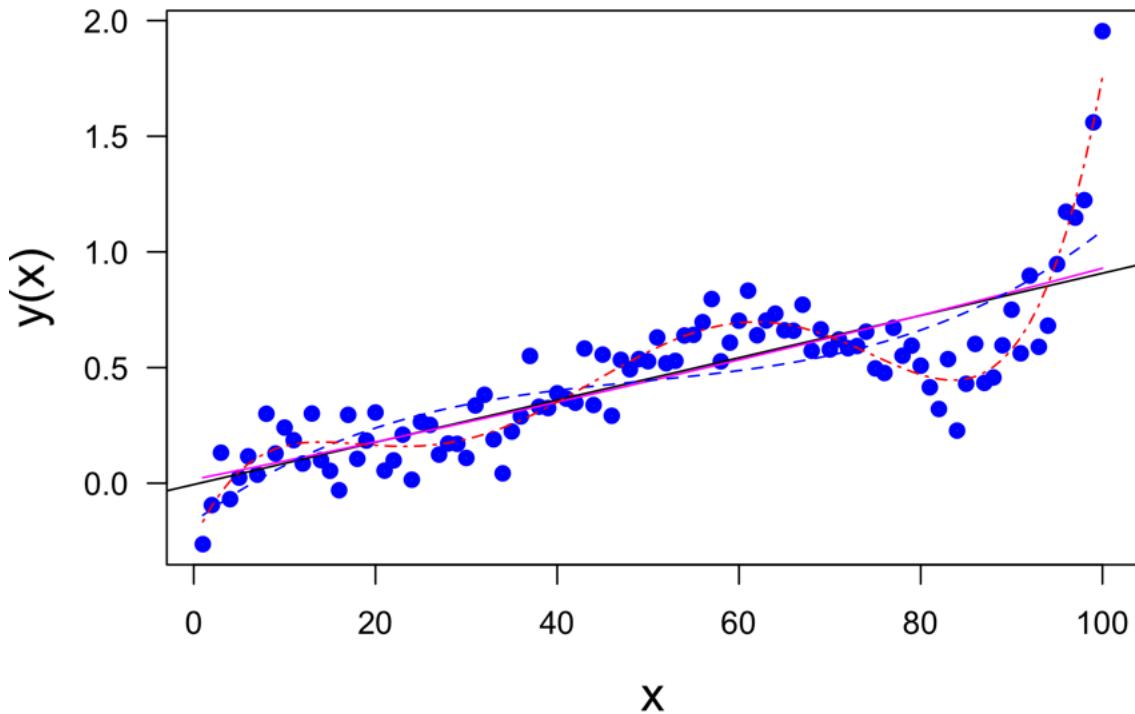


Figure 14.21: Ordinary least-squares fit to an artificial data set. 4 different polynomials have been fitted to the data: (1) up to linear in  $x$  (straight line, black), (2) up to second order in  $x$  (magenta curve), (3) up to third order in  $x$  (blue dashed line), (4) up to fifth order in  $x$  (red dash-dotted line). The fit by the fifth order polynomial looks best. A quantitative measure for comparing the 4 models can be obtained from the AIC values: -24.91, -23.12, -31.61, -161.32. The AIC values for the models differ by less than 2, and the quadratic model is not better than the linear model. The cubic model is clearly better than the first two models and the fifth order model is clearly the best. [PolynomialFit.R](#)



# Chapter 15

## Errors in $y$ and $x$ : errors in variables (EIV)

"Regression with *errors in variables* (EIV) ... is so fundamentally different from the simple linear regression ... that it is probably best thought of as a completely different topic."  
Casella & Berger (2002, p.577-578)

*In the regression analysis done so far we always assumed that there are no – or at least negligible – errors in  $x$  ('fixed'  $x$ ). When both  $y$  and  $x$  contain errors<sup>1</sup> one speaks of an errors in variables (EIV) problem. Although the errors in variables problem with two variables sounds like 'one step away' from simple linear regression it is much more difficult to solve even for normally distributed variables and errors. The problem has been investigated at least since Adcock (1878) and Kummell (1889). Draper and Smith, who wrote a whole book on 'Applied Regression Analysis' (1998), devoted only 7 out of 700 pages (p. 89–96) to EIV problems and wrote "The literature is too vast to discuss in full detail here, and we provide a selective, perhaps biased, discussion." (p. 90).*

*Despite a plethora of investigations and publications since the 19th century no satisfying solution seems to exist of this estimation problem based on data alone<sup>2</sup> – even for large sample sizes (compare the insightful discussion in Zellner, 1971). However, specific methods have been proposed for cases where additional information as, for example, estimates or guesses about the uncertainties in  $x$  and  $y$  ('weights' in the form of the inverse of variances for observations; compare: York, 1966) or at least the ratio of their variances,  $\lambda = \sigma_x^2 / \sigma_y^2$  (Deming, 1943) are available. We will discuss the following approaches: (1) Simple linear regressions of  $y$  on  $x$  and via  $x$  on  $y$  followed by taking the geometric mean of the two slopes estimated by simple linear regressions, called '[geometric line](#)' or '[reduced major-axis](#)'. (2) Simple linear regressions of  $y$  on  $x$  and via  $x$  on  $y$  followed by bisecting the slope angles, called '[bisection](#)', better called '[angle bisection](#)'. The parameter uncertainties (standard errors) for these two methods were derived by Isobe et al. (1990). We apply these two methods to a large ( $n > 10^6$ , 'Redfield data') and a small ( $n = 10$ ) data set. Other approaches including (3) splitting methods, (4) orthogonal regression, (5) Maximum Likelihood Estimation (MLE), (6) Markov Chain Monte Carlo (MCMC, Section [K.2](#)). will be discussed in the appendix (Section [K.2](#)). Application of various of these methods to artificial data sets can shed more light on the difficulty of errors in variables problems (Section [K.2](#)).*

**Further reading (errors in variables):** Zellner (1971), Riggs et al. (1978), Fuller (1987), Cheng & van Ness (1999), Gillard (2007, 2010, 2014), Legendre & Legendre (2012, Chapter 10.3); for resampling in the context of errors in variables compare Mudelsee (2023)

---

<sup>1</sup>Here 'error' refers to deviation of any kind due to a random process, not only measurement errors (Legendre & Legendre, 2012)

<sup>2</sup>Here, the data don't speak for themselves.

## 15.1 Redfield ratios

Stoichiometric data provide an example for data with errors in both variables. Alfred Redfield discovered in the 1930ies that marine phytoplankton (microalgae) and zooplankton (copepods, krill etc.) show on average *molar ratios*<sup>3</sup> of carbon (C), nitrogen (N), and phosphorous (P) C:N:P = 106:16:1 and that the nutrients nitrate ( $\text{NO}_3^-$ ) and phosphate ( $\text{PO}_4^{3-}$ ) in seawater show on average a ratio similar to N:P in plankton (Redfield, 1934, 1958, 1963).<sup>4</sup> The reason for the Redfield ratios in organisms are similarities in the ratios of proteins, lipids, and hydrocarbons (Geider & La Roche, 2002) and, if we leave out bones which contain a lot of phosphorous, even humans are close to Redfield ratios (Sterner & Elser, 2017). Fixed ratios between N:P in plankton and  $\text{NO}_3^-:\text{PO}_4^{3-}$  are still used quite often in marine biogeochemical models in order to save computational demand.<sup>5</sup> Thus one goal of fitting a straight line to the data would be to estimate the linear  $\text{NO}_3^-:\text{PO}_4^{3-}$  relationship in order to use it in biogeochemical models.

The data from the World Ocean Atlas (<https://www.ncei.noaa.gov/products/world-ocean-atlas>) are shown in Fig. 15.1 (black dots) together with a straight line fit (geometric line, blue) with estimates of slope  $\hat{\beta}_{\text{geom}} = 15.042 \pm 0.005 \text{ mol N (mol P)}^{-1}$  and intercept  $\hat{\beta}_0 = -2.130 \pm 0.005 \mu\text{mol N L}^{-1}$ . A few comments are in order:

1. Various methods have been applied to fit a straight line to data. Several of these methods will be discussed in sections below.
2. The estimated slopes and intercepts based on these methods are listed in Table 15.1.
3. The estimated slopes are all around 15 mol N (mol P)<sup>-1</sup> and confirm the ‘canonical’ Redfield ratio of N:P.
4. Note that the units of slopes, mol N (mol P)<sup>-1</sup> or mol  $\text{NO}_3^-$  (mol  $\text{PO}_4^{3-}$ )<sup>-1</sup>, are usually dropped (‘molar ratios’). Keep in mind that the N:P ratio in g N (g P)<sup>-1</sup> is different from the molar N:P value; giving N:P ratios in g N (g P)<sup>-1</sup> should be avoided.
5. The estimated intercept is negative (between -1.7 and -2.6  $\mu\text{mol N}$ ). Of course, negative concentrations are impossible. The estimated intercept is the  $y$ -intercept at  $x \equiv \text{PO}_4^{3-} = 0 \mu\text{mol}$ . However, the data show that  $\text{NO}_3^-$  is often already (almost) completely depleted when a bit of  $\text{PO}_4^{3-}$  is still available.
6. The uncertainties are standard errors and thus usually scale with  $1/\sqrt{n}$ . They are extremely small compared to the estimated values (for example, 0.005:15 = 1:3000). The main reason is the large sample size of  $n > 10^6$  and thus  $\sqrt{n} > 10^3$ . There is actually no ‘true’ Redfield ratio that could be estimated better and better by taking more and more data.<sup>6</sup> The stoichiometries of plankton organisms varies between species and even between individuals of the same species caused, for example, by varying growth conditions.
7. How would you explain the high  $\text{PO}_4^{3-}$  values at low  $\text{NO}_3^-$  (lower left region in Fig. 15.1) which clearly lie out of the main data cloud surrounding the fitted straight line? (Answer: at the end of this section).

---

<sup>3</sup>mol mol<sup>-1</sup>

<sup>4</sup>In seawater, the ratio of dissolved inorganic carbon (DIC) to  $\text{PO}_4^{3-}$  is much higher than 106:1. As a consequence, DIC is decreasing by less than 15% during algal blooms when  $\text{NO}_3^-$  and/or  $\text{PO}_4^{3-}$  are almost depleted.

<sup>5</sup>As a carbon cycle modeller from Hamburg used to say in the early 1990ies: “In marine biogeochemical models we need a nutrient. We call it phosphate.” Several nutrients were introduced in biogeochemical models starting at the end of the 1990ies.

<sup>6</sup>Tiny standard errors might be misleading: compare the size of the Emperor of China (Section 10.2.2).

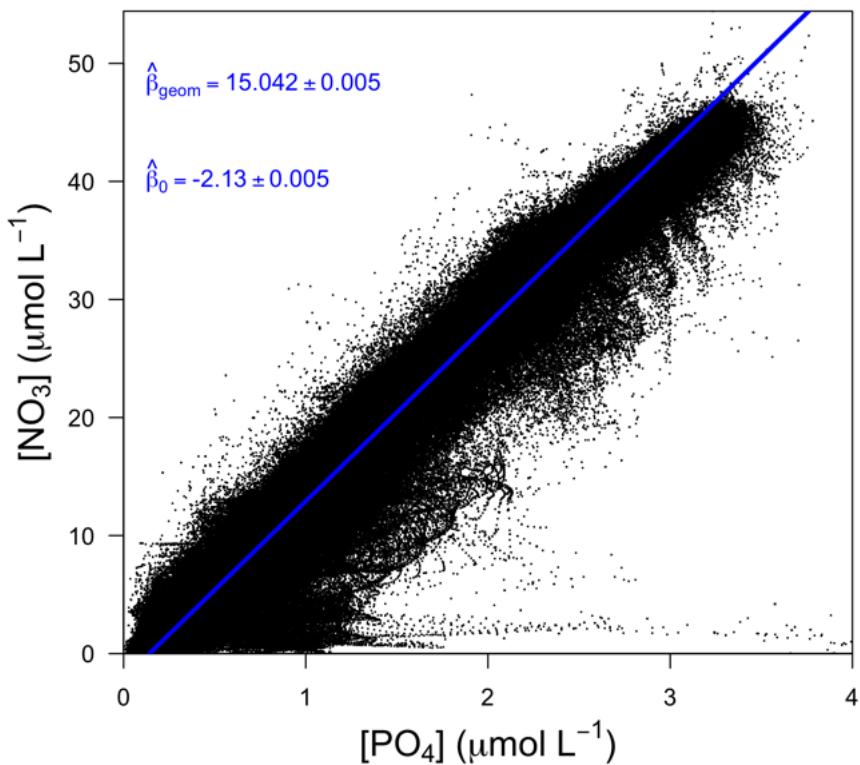


Figure 15.1: Stoichiometric data (blue dots,  $n = 1154926$ ; from World Ocean Atlas) and geometric line (black) with slope  $\hat{\beta}_{\text{geom}} = 15.042 \pm 0.005 \text{ mol N (mol P)}^{-1}$  and intercept  $\hat{\beta}_0 = -2.130 \pm 0.005 \mu\text{mol N L}^{-1}$ .  
Music The Rolling Stones: Memory Motel (from album 'Black and Blue')  
RedfieldEIV250404.R

Method	Slope	Intercept
$y$ on $x$	$14.783 \pm 0.003$	$-1.700 \pm 0.005$
$x$ on $y$	$15.306 \pm 0.003$	$-2.567 \pm 0.005$
Bisection	$15.040 \pm 0.002$	$-2.126 \pm 0.005$
Orthogonal	$15.304 \pm 0.003$	$-2.563 \pm 0.006$
Geometric	$15.042 \pm 0.002$	$-2.130 \pm 0.005$
MLE	15.043	-2.131

Table 15.1: Estimates of the Redfield ratio  $\text{NO}_3:\text{PO}_4 \pm$  standard error. Note that the units of slopes, mol N (mol P) $^{-1}$  or mol NO<sub>3</sub> (mol PO<sub>4</sub>) $^{-1}$ , are usually dropped ('molar ratios'). The intercept is given in mol N or mol NO<sub>3</sub>. R code: [RedfieldEIV250404.R](#)

How would you explain the high PO<sub>4</sub> values at low NO<sub>3</sub> (lower left region in Fig. 15.1) which clearly lie out of the main data cloud surrounding the fitted straight line? These data stem from anoxic/low oxygen regions where denitrification (conversion from NO<sub>3</sub> to N<sub>2</sub>) takes place. Exercise: Identify the oceanic regions where this takes place by searching locations with  $\text{PO}_4 > 1 \mu\text{mol L}^{-1}$  and unusual low N:P in the World Ocean Data.

## 15.2 Errors in variables model (EVM)

The errors in variables model (EVM) is defined as follows:

$$\gamma_i = \beta \xi_i + \beta_0 \quad i = 1, 2, \dots, n, \quad (15.1)$$

where  $\beta$  is the (true) slope,  $\beta_0$  is the (true) intercept,  $\xi_i$  and  $\gamma_i$  are the true but unknown variable values. Note that  $\xi_i$  and  $\gamma_i$  are on an equal footing with each other and thus one does not speak here of predictor and response variable.<sup>7</sup> The observed variables  $x_i$  and  $y_i$  are related to  $\xi_i$  and  $\gamma_i$ , respectively, by

$$x_i = \xi_i + u_{x,i} \quad u_{x,i} \sim \mathcal{N}(0, \sigma_x) \quad (15.2)$$

$$y_i = \gamma_i + u_{y,i} \quad u_{y,i} \sim \mathcal{N}(0, \sigma_y) \quad (15.3)$$

where the uncertainties (or errors)  $u_{x,i}$  and  $u_{y,i}$ , respectively, are normally distributed with zero means and variances  $\sigma_x^2$  and  $\sigma_y^2$ , respectively; i.e. all  $u_{x,i}$  come from identical populations with zero means and identical variances  $\sigma_x^2$  (homoscedasticity in  $x$ ; homoscedasticity in  $y$  with variance  $\sigma_y^2$  is assumed as well). It is further assumed (so called *structural* errors in variables model, Zellner, 1971) that the  $\xi_i$  stem from a normal distribution, i.e.  $\xi_i \sim \mathcal{N}(\mu_\xi, \sigma_\xi)$ . In vector notation the EVM reads:

$$\mathbf{x} = \boldsymbol{\xi} + \mathbf{u}_x \quad \mathbf{u}_x \sim \mathcal{N}(0, \sigma_x) \quad (15.4)$$

$$\mathbf{y} = \boldsymbol{\gamma} + \mathbf{u}_y \quad \mathbf{u}_y \sim \mathcal{N}(0, \sigma_y) \quad (15.5)$$

The list of unknowns is long  $\beta, \beta_0, \sigma_x^2, \sigma_y^2, \boldsymbol{\xi}, \boldsymbol{\gamma}$  and the number of unknowns,  $4 + 2n$ , is larger than the number of observations, namely  $2n$  from  $\mathbf{x}$  and  $\mathbf{y}$ . Zellner (1971, Chapter 5) has discussed this and similar problems (mean of a multivariate normal distribution) where the number of unknowns is larger than the number of observation.

## 15.3 Regression and regression-based methods

Jitjareonchai et al. (2006) have estimated the slope and intercept for an artificial data set given by ('raw data' in Table 2 of Jitjareonchai et al., 2006)

$$\mathbf{x} = \{18.6369, 25.8483, 25.0829, 30.2416, 14.7174, 14.2656, 17.2711, 14.0370, 28.1420, 12.2802\} \quad (15.6)$$

$$\mathbf{y} = \{70.5998, 96.9826, 89.3344, 99.3439, 55.9705, 57.3196, 65.9757, 55.7193, 87.9387, 54.2554\} \quad (15.7)$$

<sup>7</sup>Zellner (1971) makes this explicit in his notation by using  $y_1$  and  $y_2$  instead of  $x$  and  $y$ .

which is a random sample from a population<sup>8</sup> with slope  $\beta = 3$ , intercept  $\beta_0 = 10$ ,  $\sigma_x^2 = 4$ ,  $\sigma_y^2 = 9$ .<sup>9</sup> The sample size,  $n = 10$ , is small and thus based on our experience with simple<sup>10</sup> linear regression we do expect large uncertainties (standard errors) in the estimates of slope and of intercept.

In the following we will estimate the slope and intercept with or based on simple linear regressions methods. The Markov Chain Monte Carlo (MCMC) method used by Jitjareonchai et al. (2006) will be discussed in Subsection K.2.

The slopes and intercepts and their uncertainties can all be estimated from the mean values,  $\bar{x}$  and  $\bar{y}$ , of  $x$  and  $y$ , respectively, and the sum of squares  $S_{xx}$ ,  $S_{yy}$ , and  $S_{xy}$ :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 20.0523 \quad (15.8)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 73.3440 \quad (15.9)$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 396.4405 \quad (15.10)$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = 2998.9234 \quad (15.11)$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 1064.6057 \quad (15.12)$$

### 15.3.1 Simple linear regression: $y$ on $x$

Although knowing that an important prerequisite of simple linear regression (SLR)<sup>11</sup>, namely the 'fixed- $X$ ' condition, i.e. no or negligible errors in  $x$ -values, is violated, we will apply SLR, compare the resulting estimates with those from other methods and use them to construct the 'bisection' and 'geometric' lines (see below). For the Jitjareonchai data set  $(x, y)$  one obtains (using the R routine `lm()`)  $\hat{\beta}_{\text{SLR}} y$  on  $x = 2.69 \pm 0.21$  and  $\hat{\beta}_0 \text{SLR } y$  on  $x = 19.5 \pm 4.4$ . The true slope,  $\beta = 3$ , lies within the  $2\sigma$  range of its estimate and the true intercept  $\beta_0 = 10$  lies slightly outside the  $2\sigma$  range of its estimate. The results look better than expected.

The slope estimate  $\hat{\beta}_1 \equiv \hat{\beta}_{\text{SLR}} y$  on  $x$  is given by  $S_{xy}/S_{xx} = 1064.6057/396.4405 = 2.6854 \approx 2.69$ . The variance of this estimate can be estimated by (Isobe et al., 1990, Babu & Feigelson, 1992)<sup>12–13</sup>

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \hat{\beta}_1 x_i - \bar{y} + \hat{\beta}_1 \bar{x})^2 \right] = 0.04078 \quad (15.13)$$

and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = 0.20$ .

<sup>8</sup>The true values of the population are listed in Table 2 of Jitjareonchai et al. (2006). Note that my notation is different from Jitjareonchai et al. (2006) in order to be consistent with other parts of my script: the slope  $\beta$  and intercept  $\beta_0$  are denoted by  $\alpha^*$  and  $\beta^*$ , respectively, in Jitjareonchai et al. (2006).

<sup>9</sup>Given these large variances and thus standard deviations  $> 1$ , we do not know why Jitjareonchai et al. (2006) give true and raw data up to 4 decimals. Rounding to 2 or even 1 decimal most probably will lead to very similar numerical results and to the same conclusions (exercise left to the reader).

<sup>10</sup>'simple' in contrast to multivariate regression where more than two variables are considered

<sup>11</sup>Simple linear regression is based on least-squares and thus is often also called ordinary least-squares (OLR) although OLR also encompasses multivariate regressions.

<sup>12</sup>Note the typos in column 1 ('Method') of Table 1 of Isobe et al., 1990: in the 1. row 'OLS(X|Y)' should be corrected to 'OLS(Y|X)' (i.e. regression of  $Y$  on  $X$ ) and in the 2. row 'OLS(Y|X)' should be corrected to 'OLS(X|Y)' (compare Babu & Feigelson (1992, Table I) for the correct notation).

<sup>13</sup>The variances for all regression slope and intercept estimates "are calculated using the 'delta method' a technique that combines Taylor expansions, the central limit theorem, and Slutsky's theorem of probability theory (see Billingsley 1986, p. 380)." (Isobe et al., 1990)

The intercept estimate is given by

$$\alpha_1 \equiv \hat{\beta}_{0,\text{SLR}} y \text{ on } x = \bar{y} - \hat{\beta}_1 \bar{x} = 73.3440 - 2.6854 \cdot 20.0523 = 19.49555 \approx 19.5 \quad (15.14)$$

The variance of this estimate can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\alpha}_1) = \frac{1}{n^2} \sum_{i=1}^n \left\{ y_i^0 - \hat{\beta}_1 x_i^0 - n\bar{x} \cdot \left[ \frac{1}{S_{xx}} x_i^0 (y_i^0 - \hat{\beta}_1 x_i^0) \right] \right\}^2 = 10.2248 \quad (15.15)$$

where  $x_i^0 = x_i - \bar{x}$  and  $y_i^0 = y_i - \bar{y}$  and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\alpha}_1)} = 3.1976 \approx 3.2$ . Note that the standard errors based on the 'delta method' (Isobe et al., 1990) are slightly different from those provided by **lm()**.

### 15.3.2 Simple linear regression: via $x$ on $y$

A linear relationship between  $x$  and  $y$  can also be estimated by regressing  $x$  and  $y$  whereby again the 'fixed' condition (here: no errors in  $y$ ) is violated. One obtains  $\hat{\gamma}_{\text{SLR}} x \text{ on } y = 0.355 \pm 0.28$  and  $\hat{\delta}_{0,\text{SLR}} x \text{ on } y = -5.98 \pm 2.09$  for

$$x(y) = \gamma y + \delta \quad (15.16)$$

which can be 'inverted' to

$$y(x) = \frac{x}{\gamma} - \frac{\delta}{\gamma} \quad (15.17)$$

i.e. with slope  $1/\gamma$  and intercept  $-\delta/\gamma$ . Inserting our estimates for  $\gamma$  and  $\delta$  into the expressions for slope and intercept and propagating uncertainties (using Monte Carlo simulations combined with robust estimation) one obtains  $\hat{\beta}_{\text{SLR via } x \text{ on } y} = 2.82 \pm 0.22$  and  $\hat{\beta}_{0,\text{SLR via } x \text{ on } y} = 16.89 \pm 5.95$ . These estimates are closer to the true values as those from the SLR of  $y$  on  $x$ .<sup>14</sup>

The slope  $\hat{\beta}_2 \equiv \hat{\beta}_{\text{SLR via } x \text{ on } y}$  can be more directly estimated by

$$\hat{\beta}_2 \equiv \hat{\beta}_{\text{SLR via } x \text{ on } y} = S_{yy}/S_{xy} = 2998.9234/1064.6057 = 2.8169 \approx 2.82 \quad (15.18)$$

yielding a value that is indeed identical to the result given by **lm()**. The variance of this estimate can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\beta}_2) = \frac{1}{S_{xy}^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 (y_i - \hat{\beta}_2 x_i - \bar{y} + \hat{\beta}_2 \bar{x})^2 \right] = 0.04824 \quad (15.19)$$

and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\beta}_2)} = 0.2196 \approx 0.22$ .

The intercept estimate is given by

$$\alpha_2 \equiv \hat{\beta}_{0,\text{SLR via } x \text{ on } y} = \bar{y} - \hat{\beta}_2 \bar{x} = 73.3440 - 2.8169 \cdot 20.0523 = 16.8580 \approx 16.9 \quad (15.20)$$

The variance of this estimate can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\alpha}_2) = \frac{1}{n^2} \sum_{i=1}^n \left\{ y_i^0 - \hat{\beta}_2 x_i^0 - n\bar{x} \cdot \left[ \frac{1}{S_{xy}} y_i^0 (y_i^0 - \hat{\beta}_2 x_i^0) \right] \right\}^2 = 12.4811 \quad (15.21)$$

and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\alpha}_2)} = 3.5329 \approx 3.5$ .

<sup>14</sup>However, this should not indicate that estimates via the 'inverse' regression is always the better method.

### 15.3.3 The 'bisection' or 'double regression' method

The 'bisection' and the 'geometric' lines are regression-based methods in that they rely on both the regressions of  $y$  on  $x$  and  $x$  on  $y$ . They differ from each other by how the line slope is calculated from the two regression slopes  $\hat{\beta}_1$  and  $\hat{\beta}_2$ . In the bisection approach the angles of the lines with slopes  $\hat{\beta}_1$  and  $\hat{\beta}_2$  with the  $x$ -axis are calculated and their **arithmetic mean** is the slope  $\hat{\beta}_3 = 2.75$  of the bisection line. A formula for  $\hat{\beta}_3$  reads (Isobe et al., 1990)

$$\hat{\beta}_3 = \frac{1}{\hat{\beta}_1 + \hat{\beta}_2} \left[ \hat{\beta}_1 \hat{\beta}_2 - 1 + \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)} \right] \quad (15.22)$$

The estimate of the intercept for the bisection line is given by

$$\hat{\alpha}_3 = \bar{y} - \hat{\beta}_3 \bar{x} = 18.2 \quad (15.23)$$

The variance of  $\hat{\beta}_3$  can be estimated by (Isobe et al., 1990)

$$\begin{aligned} \widehat{\text{Var}}(\hat{\beta}_3) &= \frac{\hat{\beta}_3^2}{(\hat{\beta}_1 + \hat{\beta}_2)^2 (1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)} \\ &\times \left[ (1 + \hat{\beta}_2^2)^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2) \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) + (1 + \hat{\beta}_1^2)^2 \widehat{\text{Var}}(\hat{\beta}_2) \right] \\ &= 0.04325 \end{aligned} \quad (15.24)$$

where

$$\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = \frac{1}{\hat{\beta}_1 S_{xx}^2} \left\{ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) [y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})] [y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})] \right\} = 0.04224 \quad (15.25)$$

Thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\beta}_3)} = 0.2080 \approx 0.21$ .

The variance of  $\hat{\alpha}_3$  can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\alpha}_3) = \frac{1}{n^2} \sum_{i=1}^n \left\{ y_i^0 - \hat{\beta}_1 x_i^0 - n\bar{x} \cdot \left[ \frac{\gamma_{13}}{S_{xx}} x_i^0 (y_i^0 - \hat{\beta}_1 x_i^0) + \frac{\gamma_{23}}{S_{xy}} y_i^0 (y_i^0 - \hat{\beta}_2 x_i^0) \right] \right\}^2 = 10.7801 \quad (15.26)$$

where

$$\gamma_{13} = \gamma_1 (1 + \hat{\beta}_2^2) \quad (15.27)$$

$$\gamma_{23} = \gamma_1 (1 + \hat{\beta}_1^2) \quad (15.28)$$

$$\gamma_1 = \frac{\hat{\beta}_3}{(\hat{\beta}_1 + \hat{\beta}_2) \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)}} \quad (15.29)$$

and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\alpha}_3)} = 3.2833 \approx 3.3$ .

### 15.3.4 The 'geometric' line

The 'geometric' line is another regression-based method that relies on both the regressions of  $y$  on  $x$  and  $x$  on  $y$ . Its slope is calculated as the **geometric mean**<sup>15</sup> of the two regression slopes  $\hat{\beta}_1$  and  $\hat{\beta}_2$  (here for  $\hat{\beta}_1 > 0$  and  $\hat{\beta}_2 > 0$ ):

$$\hat{\beta}_5 \equiv \hat{\beta}_{\text{geometric}} = \sqrt{\hat{\beta}_1 \cdot \hat{\beta}_2} = \sqrt{2.69 \cdot 2.82} = 2.75 \quad (15.30)$$

<sup>15</sup>Note that the sign of the square root is given by the sign of  $S_{xy}$ . In order to obtain real values for the geometric mean of two values both have to be of the same sign. What if  $\hat{\beta}_1$  and  $\hat{\beta}_2$  have different signs? The answer is given at the end of Subsection 15.3.6

(Draper & Smith 2014; compare also Draper, 1992, or Legendre & Legendre, 2012; Isodore et al., 1990, following Kermack & Haldane 1950, call this method ‘reduced major axis’). A formula for  $\hat{\beta}_5$  taking into account the sign of  $S_{xy}$  reads (Isobe et al., 1990)

$$\hat{\beta}_5 = \text{Sign}(S_{xy}) (\hat{\beta}_1 \hat{\beta}_2)^{1/2} = \text{Sign}(S_{xy}) \left( \frac{S_{yy}}{S_{xx}} \right)^{1/2}. \quad (15.31)$$

The estimate of the intercept for the bisection line is given by

$$\hat{\alpha}_5 = \bar{y} - \hat{\beta}_5 \bar{x} = 18.1924 \approx 18.2 \quad (15.32)$$

The variance of  $\hat{\beta}_5$  can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\beta}_5) = \frac{1}{4} \left[ \frac{\hat{\beta}_2}{\hat{\beta}_1} \widehat{\text{Var}}(\hat{\beta}_1) + 2 \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) + \frac{\hat{\beta}_1}{\hat{\beta}_2} \widehat{\text{Var}}(\hat{\beta}_2) \right] = 0.04331 \quad (15.33)$$

where  $\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$  is given by Eq. (15.25) Thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\beta}_5)} = 0.2081 \approx 0.21$ .

The variance of  $\hat{\alpha}_5$  can be estimated by (Isobe et al., 1990)

$$\widehat{\text{Var}}(\hat{\alpha}_5) = \frac{1}{n^2} \sum_{i=1}^n \left\{ y_i^0 - \hat{\beta}_1 x_i^0 - n \bar{x} \cdot \left[ \frac{\gamma_{15}}{S_{xx}} x_i^0 (y_i^0 - \hat{\beta}_1 x_i^0) + \frac{\gamma_{25}}{S_{xy}} y_i^0 (y_i^0 - \hat{\beta}_2 x_i^0) \right] \right\}^2 = 10.7985 \quad (15.34)$$

where

$$\gamma_{15} = \gamma_1 (1 + \hat{\beta}_1^2) \quad (15.35)$$

$$\gamma_{25} = \frac{1}{2} \sqrt{\hat{\beta}_1 / \hat{\beta}_2} \quad (15.36)$$

$$(15.37)$$

( $\gamma_1$  given by Eq. (15.29)) and thus the standard error is  $\sqrt{\widehat{\text{Var}}(\hat{\alpha}_5)} = 3.2861 \approx 3.3$ .

This ‘geometric line’ is attractive because

1. it lies between the two regression lines of  $y$  and  $x$  and of those via  $x$  on  $y$ , respectively,
2. it treats  $x$  and  $y$  in a symmetric way, and
3. it minimizes the area of the sum of triangles formed by horizontal and vertical lines between the data points and the geometric line – defining two crossing points  $(x_{c1}, y_{c1})$  and  $(x_{c2}, y_{c2})$  – completed by the section on the geometric line (illustrated for a small artificial data set in Fig. 15.2 and for the Jitjareonchai et al. (2006) data in Fig. 15.3; Teissier, 1948; Barker, Soh, and Evans, 1988).

Estimations of straight lines including the geometric approach for temperature-salinity data in a sea of crumbling icebergs is given in Wolf-Gladrow et al. (2025).

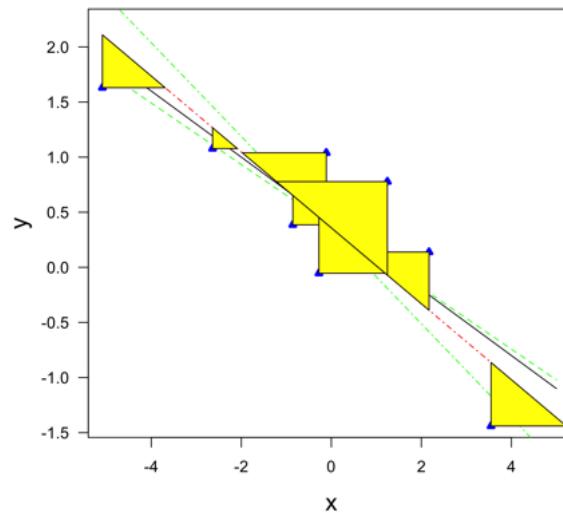


Figure 15.2: Artificial data (little blue triangles): The 'geometric' line minimizes the area of the sum of (yellow) triangles. Lines: true line (black solid), simple regression lines (green dashed lines), geometric line (magenta dash-dotted line).

R code: [EIVgeometricTriangles2504.R](#) (set: dflag = 0)

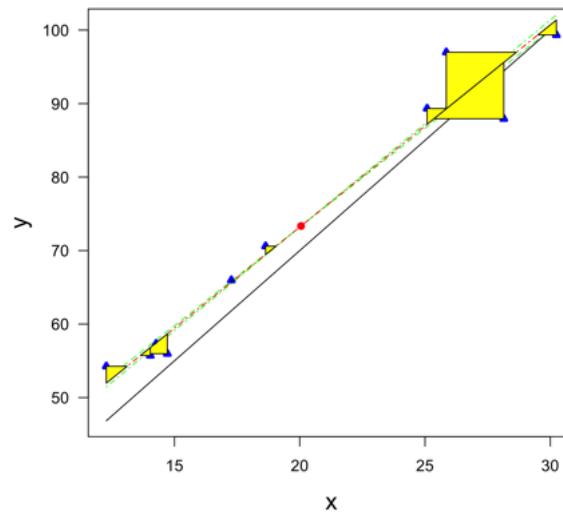


Figure 15.3: Jitjareonchai et al. (2006) data (little blue triangles): The 'geometric' line minimizes the area of the sum of (yellow) triangles. Lines: true line (black solid), simple regression lines (green dashed lines), geometric line (magenta dash-dotted line).

R code: [EIVgeometricTriangles2504.R](#) (set: dflag = 1)

### 15.3.5 Comparison of bisection and geometric lines

For given slopes  $\hat{\beta}_1$  and  $\hat{\beta}_2$ , bisection and geometric slopes are quite similar to each other, whereas a slope calculated as the **arithmetic mean of the two slopes** yields often very different slope values (Fig. 15.4). Bisection is the favoured method by Isobe et al. (1990) and Babu & Feigelson (1992) although their numerical results show that both bisection and geometric line yield good estimates; for the Jitjareonchai et al. (2006) data we obtained almost identical results (Table 15.2).

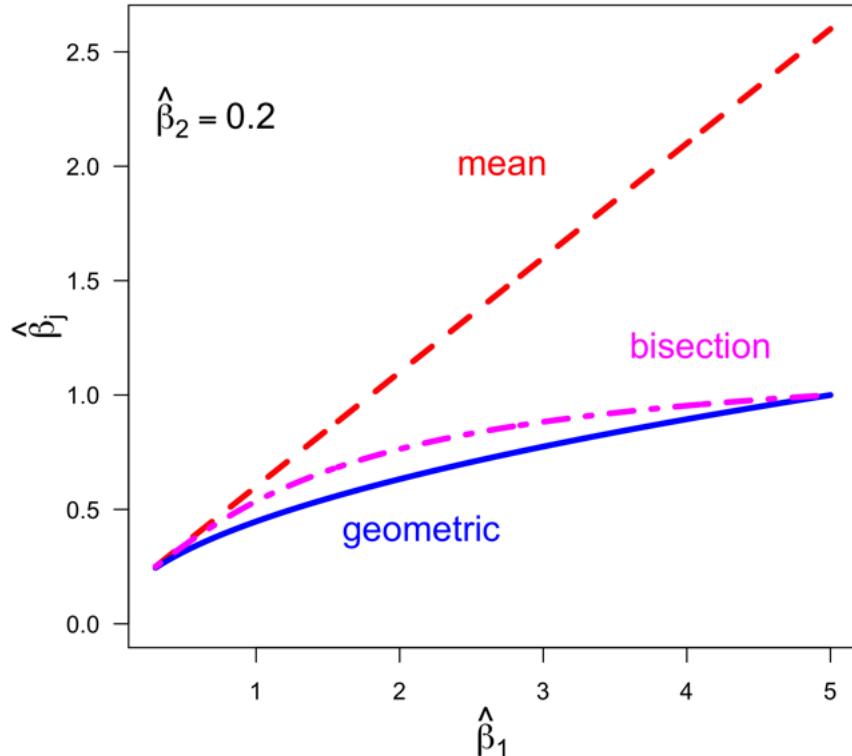


Figure 15.4: Slopes of the bisectional (magenta dash-dotted line), geometric (blue solid line), and mean line (red broken line) for variable slope  $\beta_1$  (from  $y$  on  $x$ ) and fixed slope  $\beta_2 = 0.2$  (from regression via  $x$  on  $y$ ).

R code: [Bisection2504.R](#)

### 15.3.6 Summary: regression and regression-based methods

The results of various various line estimates for Jitjareonchai et al. (2006) data are listed in Table and illustrated in Figs. 15.5 and 15.6. Given the small data set ( $n = 10$ ), the estimates of the slope are not too bad. The bisection and the geometric estimates are almost identical. The estimated intercept estimates come with large uncertainties and are quite far away from the true value. The large deviations from the true intercept value is related to the distribution of the  $x$  data: if  $\bar{x}$  is far away from  $x = 0$  (which is the case for the Jitjareonchai et al. (2006) data where all  $x_i$  are  $> 0$ ) small deviations in the slope estimate from the true value can ‘lead’ to large deviations in intercept estimates. A graph of line estimates restricted to the range of  $x$  data gives a bit more convincing impression of the success of our estimating exercise (15.6).

Method	Slope	Intercept	Remarks
$y$ on $x$	$2.69 \pm 0.21$	$19.5 \pm 4.4$	<b>lm()</b>
$y$ on $x$	$2.69 \pm 0.20$	$19.5 \pm 3.2$	based on Isobe et al. (1990)
via $x$ on $y$	$2.82 \pm 0.22$	$16.9 \pm 6.0$	<b>lm()</b> & Monte Carlo
via $x$ on $y$	$2.82 \pm 0.22$	$16.9 \pm 3.5$	based on Isobe et al. (1990)
Bisection	$2.75 \pm 0.21$	$18.2 \pm 3.3$	based on Isobe et al. (1990)
Geometric	$2.75 \pm 0.21$	$18.2 \pm 3.3$	based on Isobe et al. (1990)

Table 15.2: Estimates of slope and intercept and their uncertainties for the Jitjareonchai et al. (2006) data.

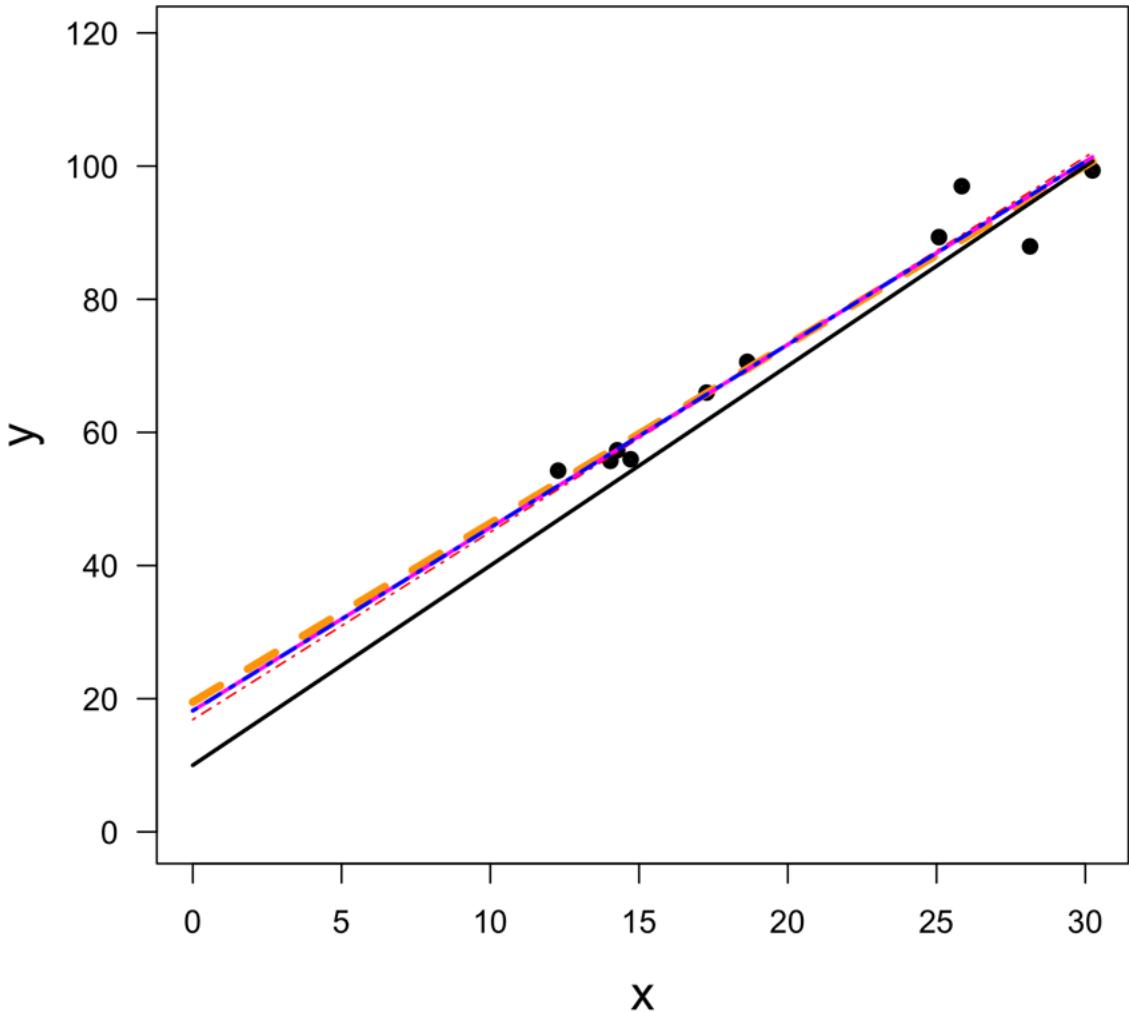


Figure 15.5: Straight lines based on various estimates of the slope and intercept using the data (Eqs. 15.6 – 15.7) of Jitjareonchai et al. (2006): true line (black solid), SLR  $y$  on  $x$  (orange broken line), SLR via  $x$  on  $y$  (red dash-dotted line), bisection line (magenta solid line), geometric line (blue dashed line). Small uncertainties in slope estimates can ‘lead’ to large deviations of intercept estimates from their true values.  
**R code:** [EIVregressions250406.R](#) (set: `sflag = 1`)

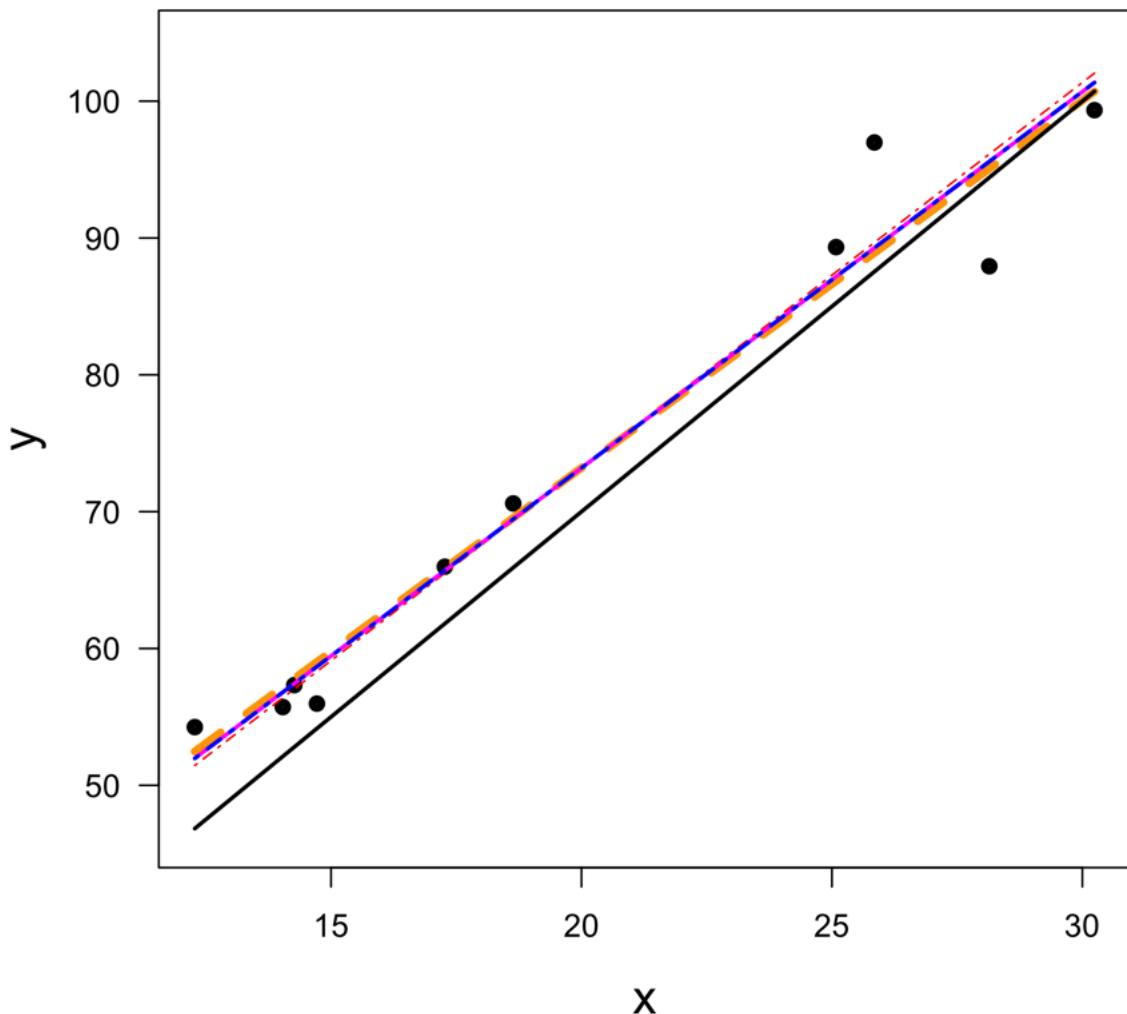


Figure 15.6: Straight lines based on various estimates of the slope and intercept using the data (Eqs. 15.6 – 15.7) of Jitjareonchai et al. (2006): true line (black solid), SLR  $y$  on  $x$  (orange broken line), SLR via  $x$  on  $y$  (red dash-dotted line), bisection line (magenta solid line), geometric line (blue dashed line). Plot restricted to the range of  $x$  data.

R code: [EIVregressions250406.R](#) (set: `sflag = 2`)

Method	Slope estimate	Variance of slope estimate ( $\widehat{\text{Var}}(\hat{\beta}_i)$ )
$y$ on $x$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \hat{\beta}_1 x_i - \bar{y} + \hat{\beta}_1 \bar{x})}{S_{xx}^2}$
via $x$ on $y$	$\hat{\beta}_2 = \frac{S_{yy}}{S_{xy}}$	$\frac{\sum_{i=1}^n (y_i - \bar{y})^2 (y_i - \hat{\beta}_2 x_i - \bar{y} + \hat{\beta}_2 \bar{x})}{S_{xy}^2}$
Bisection	$\hat{\beta}_4 = \frac{\hat{\beta}_1 \hat{\beta}_2 - 1 + \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)}}{\hat{\beta}_1 + \hat{\beta}_2}$	$\begin{aligned} & \frac{\hat{\beta}_3^2}{(\hat{\beta}_1 + \hat{\beta}_2)^2 (1 + \hat{\beta}_1^2) (1 + \hat{\beta}_2^2)} \\ & \times \left[ \left(1 + \hat{\beta}_2^2\right)^2 \widehat{\text{Var}}(\hat{\beta}_1) + \left(1 + \hat{\beta}_1^2\right)^2 \widehat{\text{Var}}(\hat{\beta}_2) \right. \\ & \quad \left. + 2 \left(1 + \hat{\beta}_1^2\right) \left(1 + \hat{\beta}_2^2\right) \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \right] \end{aligned}$
Geometric	$\hat{\beta}_5 = \text{sign}(S_{xy}) \sqrt{\hat{\beta}_1 \hat{\beta}_2}$	$\frac{1}{4} \left[ \frac{\hat{\beta}_2}{\hat{\beta}_1} \widehat{\text{Var}}(\hat{\beta}_1) + 2 \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) + \frac{\hat{\beta}_1}{\hat{\beta}_2} \widehat{\text{Var}}(\hat{\beta}_2) \right]$

Table 15.3: Slope estimates for four different regression or regression-based methods. The uncertainties (standard error!) of slope estimates can be calculated by taking the square root of the variance estimates. The covariance can be estimated as follows (Isobe et al., 1990):

$$\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = \frac{1}{\hat{\beta}_1 S_{xx}^2} \left\{ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) [y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})] [y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})] \right\}.$$

It is remarkable that analytical expressions for variances are given for all four methods (Isobe et al., 1990). Isobe et al. (1990) note that their variance estimates for regression of  $y$  on  $x$  are different from that used by other authors and applied in the R routine `lm()` (compare Table 15.2).

In order to obtain real values for the geometric mean of two values both have to be of the same sign. What if  $\hat{\beta}_1$  and  $\hat{\beta}_2$  have different signs? Different signs of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  may be obtained when the magnitude of the slope is small (near zero) and the uncertainties are large enough to lead to one positive and one negative slope estimate. This would indicate that based on the data we cannot make a decision whether the slope is positive or negative. A simple estimate of the correlation between  $x$  and  $y$  would lead to the same conclusion.

Other methods for treating errors-in-variable problems are discussed in the appendix. Some of these methods are more sophisticated (MLE, Monte Carlo Markov Chain) whereas others are more of historical interest (not applied anymore, without uncertainty estimates, or even can yield misleading results because not scale-invariant).



# Chapter 16

## Multiple linear regression (MLR)

If appropriate assumptions ('fixed  $x$ ', additive normal noise, independence, homogeneous noise level) are assumed, simple linear regression is easy to apply (Chapter 14). In multiple linear regression (MLR) the response variable depends on several predictor variables. The correlation between predictor variables (co-linearity problem) can make it difficult to attribute variations of the response variable to specific predictor variables (Chapter 17).

### 16.1 Multiple linear regression: a simple example

Artificial data are used in order to judge the results of MLR in the light of exact results. For this purpose samples  $x_1$ ,  $x_2$ , and  $x_3$  are generated with low correlations between each other (3 samples of size  $n = 100$  from the standard uniform PDF; a constant value of 4 is added to one of the samples in order to obtain larger values for one of the predictors; the correlation coefficients read  $r(x_1, x_2) = -0.14$ ,  $r(x_1, x_3) = -0.11$ ,  $r(x_2, x_3) = -0.04$ ). The exact model reads

$$Y_{\text{exact}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (16.1)$$

with intercept  $\beta_0 = 1.5$  and slopes  $\beta_1 = -0.3$ ,  $\beta_2 = 0.8$ ,  $\beta_3 = 1.0$ . The observed values of the response variable  $Y$  contain additive noise from the standard normal distribution with a noise level of  $z_{\text{NL}} = 0.2 [\max(Y_{\text{exact}}) - \min(Y_{\text{exact}})] \approx 0.34$  (the noise level is scaled with range of variation in  $Y_{\text{exact}}$  in order to obtain values for the noise in a range typical for applications, i.e. not too small and not too large), i.e.  $+z_{\text{NL}} \epsilon_i$  with  $\epsilon_i \in \mathcal{N}(\mu = 0, \sigma = 1)$ . Multilinear regression yields the following estimates for the model parameters:  $\hat{\beta}_0 = 1.166 \pm 0.485$ ,  $\hat{\beta}_1 = -0.231 \pm 0.104$ ,  $\hat{\beta}_2 = 0.745 \pm 0.101$ ,  $\hat{\beta}_3 = 0.994 \pm 0.095$ , which are all (according to  $t$ -tests) significantly ( $p < \alpha = 0.05$ ) different from zero. The exact values fall into the  $\pm 1 - \sigma$  uncertainty ranges of the estimates (Fig. 16.1).

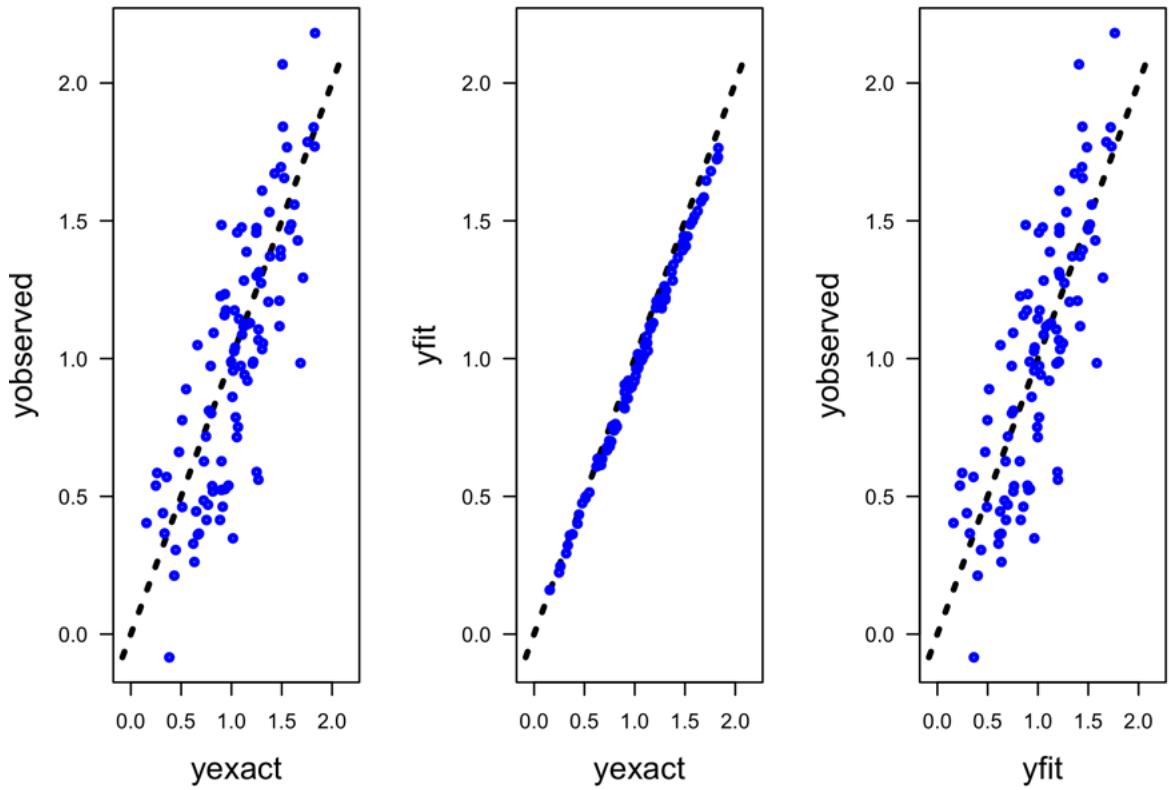


Figure 16.1: Left panel: observed over exact values of the response variable (blue dots) and the 1-to-1 line through the origin (black broken line in all panels); the plot gives an impression of the noise level. Middle panel: fitted over exact values of the response variable (blue dots); the fitted values lie much closer to the exact values than the observations, i.e. ‘regression to the mean’ has been successful. Right panel: observed over fitted values of the response variable (blue dots); the graph looks very similar to that of the left panel because the fitted values are very close to the exact values. [MLRlowNoiseLowCor.R](#)

**Exercise 46 MLR with artificial data at higher noise level**

*Redo the MLR with artificial data, however, at higher noise level, say by doubling the noise level. What do you expect? How do the estimates of intercept and slopes change?*

## 16.2 Multiple linear regression: a more difficult example

A common issue in multiple linear regression is the so-called **co-linearity problem**, i.e. the high correlation or anti-correlation between two or more predictor variables. In case of co-linearity, although a good fit to the data can still be obtained, the attribution of the variation of the response variable to specific predictor variables can be difficult, numerical results may be misleading, and estimated slopes may change by large amounts if a small percentage of new data are added. This is easy to understand in case of two highly correlated predictor variables  $X_1$  and  $X_2$ . For the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (16.2)$$

and in the limit  $X_2 = -\alpha X_1$  (where  $\alpha > 0$  is a constant which also takes care of the in general different units of  $X_1$  and  $X_2$ ; perfect anti-correlation) one could write

$$Y = \beta_0 + \beta_1 X_1 - \alpha \beta_2 X_1 = \beta_0 + (\beta_1 - \alpha \beta_2) X_1 + 0 X_2 \quad (16.3)$$

or

$$Y = \beta_0 - \frac{\beta_1}{\alpha} X_2 + \beta_2 X_2 = \beta_0 + \left( \beta_2 - \frac{\beta_1}{\alpha} \right) X_2 + 0 X_1 \quad (16.4)$$

or other linear combinations with arbitrary values  $\gamma$

$$Y = \beta_0 + \beta_1 X_1 - \alpha \beta_2 X_1 = \beta_0 + (\beta_1 - \alpha \beta_2 - \gamma) X_1 - \frac{\gamma}{\alpha} X_2, \quad (16.5)$$

i.e. the slopes (coefficients of  $X_1$  and  $X_2$ ) are not uniquely fixed; the same applies in the case of perfect correlation ( $X_2 = \alpha X_1$ ,  $\alpha > 0$ ). In other words, the variation of the response variable can be explained by a single predictor or by many different linear-combinations of two predictors.

In applications, the predictor variables are usually not perfectly co-linear. However, high correlation in combination with noise can lead to problems in the estimation and interpretation of slopes. In order to get a feeling for the co-linearity problem, an example problem will be constructed with artificial data similar to the one in the previous section; the major difference is the high correlation between the first two predictor variables.

A method described by Howell (2013) is used to create two predictors  $x_1$  and  $x_2$  of size  $n = 100$  with a high anti-correlation ( $r = -0.95$ ) to each other (Fig. 16.2). The third predictor  $x_3$  is again a random sample from the standard uniform PDF. The same exact linear model as before (Eq. 16.1) is used and normal noise is added with a similar level as before.

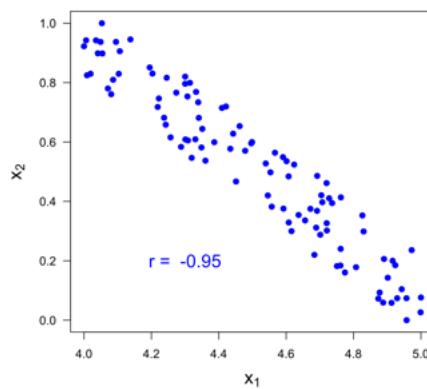


Figure 16.2: Plot of the two highly anti-correlated predictors  $x_1$  and  $x_2$ . [MLRhighCor.R](#) (line 27: set sflag to 1) [sampleUcorFct.R](#)

Application of MLR to the artificial data set (Fig. 16.3) yields the following estimates:  $\hat{\beta}_0 = -0.292 \pm 1.892$ ,  $\hat{\beta}_1 = 0.065 \pm 0.374$ ,  $\hat{\beta}_2 = 1.001 \pm 0.406$ ,  $\hat{\beta}_3 = 1.136 \pm 0.118$ . The estimate of the intercept is far away from the exact value ( $\beta_0 = 1.5$ ), most uncertainty ranges are relatively large (except for  $\hat{\beta}_3$ ), and the estimates for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are not significantly ( $\alpha = 0.05$ ) different from zero despite the moderate noise level.

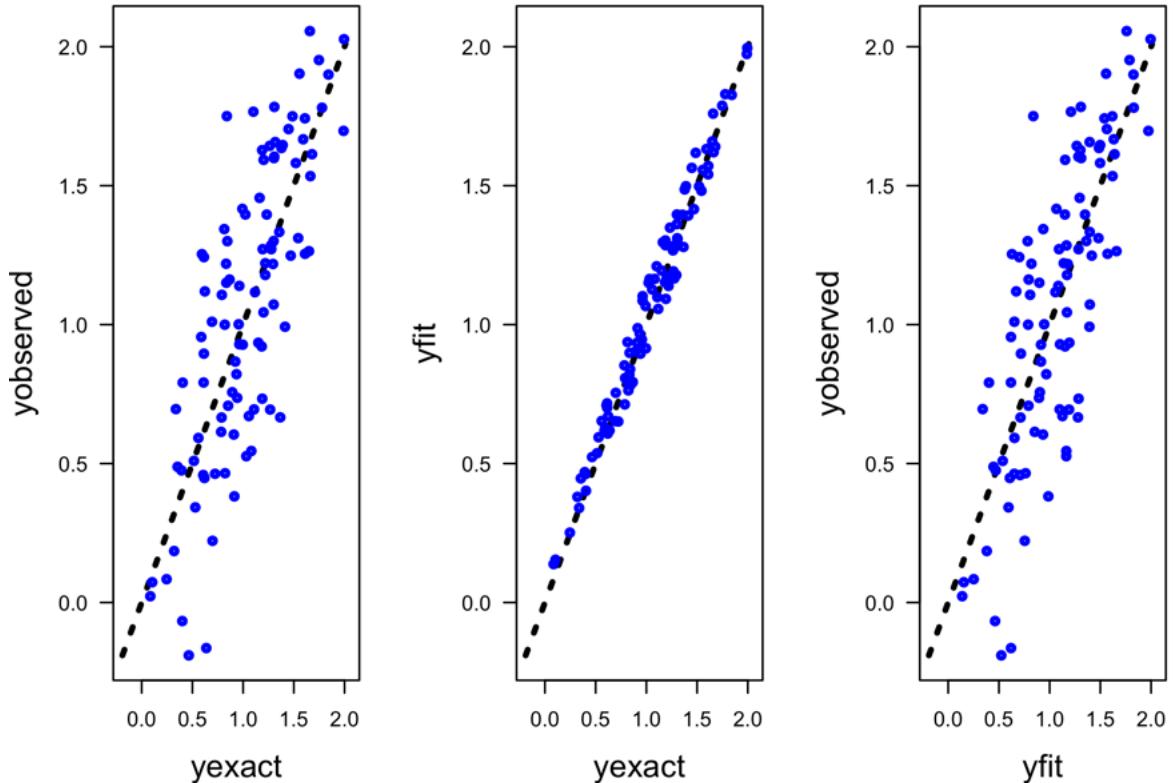


Figure 16.3: Left panel: observed over exact values of the response variable (blue dots) and the 1-to-1 line through the origin (black broken line in all panels); the plot gives an impression of the noise level. Middle panel: fitted over exact values of the response variable (blue dots); the fitted values lie much closer to the exact values than the observations, i.e. ‘regression to the mean’ has been successful. Right panel: observed over fitted values of the response variable (blue dots); the graph looks very similar to that of the left panel because the fitted values are very close to the exact values. [MLRhightCor.R](#) [sampleUcorFct.R](#)

### 16.2.1 Is less more?

The estimated values of intercept and slopes do not agree very well with the true values used to generate the data set. One may ask if 'less is more', reduce the model complexity by dropping one of the predictor variables and compare the full model ( $M_0$ ) with the three models of reduced complexity using the information criterion AICc (Table 16.1).

$\beta_k$	$\beta_0 = 1.5$	$\beta_1 = -0.3$	$\beta_2 = 0.8$	$\beta_3 = 1.0$	AIC	AICc	$\Delta$	$w$
Model:	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$				
$M_0 : y(x_1, x_2, x_3)$	$2.11 \pm 0.54$	$-1.12 \pm 0.52$	$1.07 \pm 0.22$	$-0.35 \pm 0.97$	24.6	39.6	8.8	1.15%
$M_1 : y(x_2, x_3)$	$2.09 \pm 0.66$		$0.64 \pm 0.12$	$-0.30 \pm 1.19$	28.3	36.3	5.5	5.99%
$M_2 : y(x_1, x_3)$	$2.44 \pm 1.10$	$1.15 \pm 0.47$		$-0.99 \pm 1.97$	38.5	46.5	15.7	0.04%
$M_3 : y(x_1, x_2)$	$1.93 \pm 0.19$	$-1.12 \pm 0.49$	$1.08 \pm 0.20$		22.8	30.8	0	92.83%

Table 16.1: Results of multiple linear regression of example data set (sample size  $n = 10$ ): The estimated intercepts ( $\hat{\beta}_0$ ) come with relative high uncertainties (especially in model  $M_2$ , where  $x_2$  has been dropped) and are one or even two ( $M_3$ )  $\sigma$  above the true value ( $\beta_0$ ). The estimated slopes for the third predictor  $\hat{\beta}_3 \equiv b_3$  are far away (actually with opposite sign) from the true value ( $\beta_3$ ). The estimated slopes  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are quite a bit away from the true slopes  $\beta_1$  and  $\beta_2$  (for  $M_2$   $\hat{\beta}_1$  has opposite sign with respect to  $\beta_1$ ), which is not surprising given the fact that one predictor has been dropped and the other highly correlated predictor has to carry the 'extra burden'. According to the information criterion AICc, the reduced complexity model  $M_3$  ( $y(x_1, x_2)$ ) is much better than the full model  $M_0$  ( $y(x_1, x_2, x_3)$ ). The rather poor estimates of  $\beta_1$  in models  $M_0$  and  $M_3$  might be caused by the high correlation between predictors  $x_1$  and  $x_2$  and/or by the impact of randomness at a (too) small sample size.

[MLRn10AICc.R](#)

### 16.2.2 Fitting the noise?

Models with larger number of model parameters (higher complexity) yield better goodness-of-fit. However, that does not necessarily provide the best solution as it can lead to overfitting (i.e. 'fitting noise'). As an example one may consider a data set  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  that can be perfectly fitted by a polynomial of order  $n - 1$ <sup>1</sup> (Fig. 16.4); however, this often leads to unacceptable response values when used for prediction. Alternatively, a model that is too simple (too parsimonious) misses important factors/processes/dependencies.

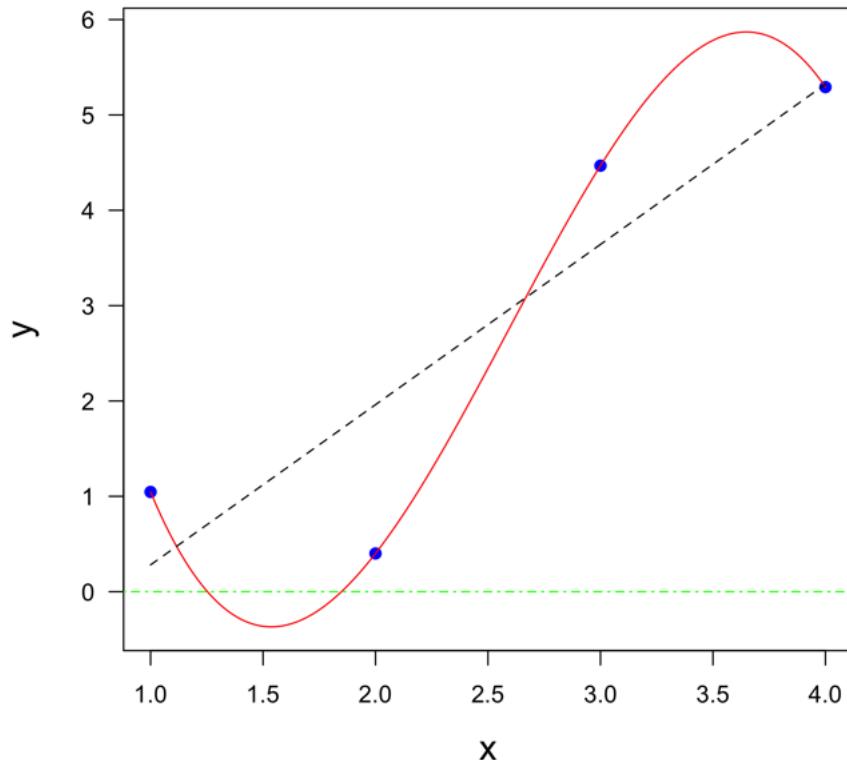


Figure 16.4: Perfect polynomial fit (red solid line) and simple linear regression (black broken line) to 4 data points (blue dots). For non-negative quantities (for example, concentrations are non-negative by definition) the perfect polynomial fit would yield absurd (negative) values when used for prediction. The green dash-dotted line indicates zero y-values. [FitTheNoise.R](#)

<sup>1</sup>... except for pathological cases, for example, when two pairs  $(x_i, y_i)$  and  $(x_j, y_j)$  are identical to each other ...

### 16.2.3 Akaike information criterion (AIC, AICc) and other information-theoretic approaches

The fit of a function  $f()$  to data can be motivated by a well established theory suggesting the form of  $f()$ , however, without specified model parameters (intercept, slopes, variance of noise). These parameters have to be estimated from the data. Often such a theoretical background does not exist (yet). In those cases one can try to find a model that is optimal in the sense that it provides a good fit to the data without requiring too many model parameters.<sup>2</sup> How can this optimality be measured? Akaike (1974) proposed the following quantity as a measure of the trade-off between model complexity and goodness-of-fit

$$\text{AIC} := 2k - 2\log\text{Lik} \quad (16.6)$$

where  $k$  is the number of model parameters and  $\log\text{Lik}$  is the natural logarithm of the maximum likelihood of the fitted model. Akaike called this quantity 'an information criterion'; nowadays it is called 'Akaike Information Criterion' (AIC). The goodness-of-fit is measured by the deviance ( $-2\log\text{Lik}$ ) and model complexity by twice the number of model parameters whereby the variance of the additive noise is also seen as a model parameter. For small sample sizes,  $n$ , the slightly modified criterion

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \quad (16.7)$$

should be used (Hurvich and Tsai, 1989).<sup>3</sup>

The full model (**Model 0**)

$$y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon, \quad (16.8)$$

where  $\epsilon$  is from a normal distribution with zero mean and variance  $\sigma^2$ , has  $k = 5$  model parameters, namely 1 intercept ( $\beta_0$ ), 3 slopes ( $\beta_1, \beta_2, \beta_3$ ), and the variance of the noise ( $\sigma^2$ ). The goodness-of-fit is measured by twice the natural logarithm of the maximum likelihood of the fitted model. The maximum likelihood is given by

$$\mathcal{L}(r|\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma}^2) = \left( \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \right)^n e^{-\frac{\sum_{j=1}^n r_j^2}{2\hat{\sigma}^2}} \quad (16.9)$$

where the  $r_j$  are the residuals resulting from linear regression (based on least squares; R routine **lm()**). In order to avoid calculating tiny values of the likelihood, it is recommended to calculate  $\log\text{Lik}$  directly:

$$\log\text{Lik} = \log \left( \mathcal{L}(r|\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma}^2) \right) = -\frac{\sum_{j=1}^n r_j^2}{2\hat{\sigma}^2} - \frac{n}{2} \log(2\pi\hat{\sigma}^2) \quad (16.10)$$

where  $\log$  is the natural logarithm. The variance of the noise can be estimated from the residuals whereby one divides by the sample size  $n$ .<sup>4</sup> One obtains the following values:  $\hat{\sigma}^2 = 0.2519650, 6.775276 \cdot 10^{-4}$ ,  $\log\text{Lik} = -7.297060$ , and  $\text{AIC} = 24.59412$ ; this result is identical to the AIC value given by the R routine **AIC()**.

The values of AICc for a single model has almost no meaning because a scale is missing. AICc values can be used to compare the optimality of different models fitted to identical response data<sup>5</sup>. Lower AICc values mean more optimal models and thus the goal is to find models with low values. In the example discussed in Subsection 16.2.1, the model  $M_3$  ( $y(x_1, x_2)$ ) with reduced complexity (2 instead of 3 predictors) leads to the

<sup>2</sup>A deeper justification in terms of the Kullback-Leibler information can be found, for example, in Burnham et al. (2011) or Burnham and Anderson (2002).

<sup>3</sup>For  $n \rightarrow \infty$  the factor  $\frac{2k(k+1)}{n-k-1} \rightarrow 0$  and thus AICc approaches AIC.

<sup>4</sup>The estimator with division by  $n$  is a maximum likelihood estimator (Zuur et al., 2009, p. 118). Although it is biased (Zuur et al., 2009, p. 118; Cassela & Berger, 2002, p. 551), it is applied in the R routine **AIC()**. One obtains this estimator by the simple argumentation, that one knows the mean of the residuals (it's zero) and thus should divide by  $n$ . An unbiased estimator for the variance of the residuals in simple linear regression is given by division by  $(n-2)$  (Cassela & Berger (2002, p. 552), taking into account the estimates of intercept and slope (= 2 constraints), or, in case of multilinear regression, by division by  $(n-p)$  where  $p$  is the number of estimated model parameters (= number of slopes or predictors plus one for the intercept, however, here not counting the variance of the noise as an additional model parameter).

<sup>5</sup>Comparison of information criteria for models fitted to different data make no sense.

lowest AICc (30.8 versus 39.6 for the full model  $M_0$ ) value, i.e. it is better in the sense of trade-off between model complexity and goodness-of-fit. Thus 'less is more' in this case. The difference in AICc between these two models is  $39.6 - 30.8 = 8.8$ . If differences are smaller than 2, models are commonly considered as equivalent and it is recommended to select the simpler model with less model parameters (Burnham & Anderson, 2002; Zuur et al., 2007). The different models can be ranked according to their *AICc differences* with respect to the minimum of AICc values found for models  $i = 1, 2, \dots, R$ :

$$\Delta_i = \text{AICc}_i - \text{AICc}_{\min}, \quad \text{for } i = 1, 2, \dots, R \quad (16.11)$$

(here:  $R = 4$ ,  $\text{AICc}_{\min} = 30.8$  for model  $M_3$ ). The (relative) likelihood of each model  $i$  are

$$\text{RL}_i = \exp(-\Delta_i/2) \quad (16.12)$$

(Burnham et al., 2011). For the data given and models considered in our example one obtains  $\Delta_1 = 8.8$ ,  $\Delta_2 = 5.5$ ,  $\Delta_3 = 15.7$  and thus  $\text{RL}_1 = 0.01236$ ,  $\text{RL}_2 = 0.06453$ ,  $\text{RL}_3 = 0.00038$ ,  $\text{RL}_4 = 1$ . From the (relative) likelihoods one can calculate *likelihood ratios*, for example,  $\text{RL}_4/\text{RL}_2 = 15.5$ . This ratios can be interpreted as, for example, "model  $M_3$  is 15.5 times better supported by the data than model  $M_1$ ". The probability  $w_i$  for each model  $M_i$  given the data and the  $R$  models considered, can be calculated by

$$w_i = \frac{\text{RL}_i}{\sum_{j=1}^R \text{RL}_j} \quad (16.13)$$

(Burnham et al., 2011). For our example, one obtains  $w(M_0) = 0.011$  (or 1.1%),  $w(M_1) = 0.060$ ,  $w(M_2) = 0.0004$ ,  $w(M_3) = 0.928$  (or 92.8%). These probabilities can be interpreted as a measure of *strength of evidence*.

Burnham et al. (2011) write: "The quantitative evidence is represented by the model likelihoods, model probabilities, and evidence ratios; these are the science results. Then a value judgment is made as the results are interpreted and qualified. Such value judgments attempt to explain the science result. The word 'significant' is to be avoided as it implies the older approaches and implies a dichotomy (reject or not) that is not appropriate."

Of the four models ( $M_0, \dots, M_3$ ) model  $M_3$  shows the smallest AICc and the largest probability  $w$ . However, even though these results seem to suggest a clear cut decision which model to choose, one has to be keep in mind that the inference is based on a small data set (sample size  $n = 10$ ). By comparing model estimates for intercept and slopes with true values (this is only possible because we use artificial data) one recognizes that the estimates are not reliable (large quantitative differences, even wrong signs) which could be caused by the high correlation of the first two predictors and by the large impact of randomness at small sample sizes. Therefore we repeated the investigation based on a ten-fold larger sample size ( $n = 100$ ; Table 16.2).

**Further reading (AIC, AICc):** Burnham & Anderson (2002), Yang (2005), Vrieze (2012)

$\beta_k$	$\beta_0 = 1.5$	$\beta_1 = -0.3$	$\beta_2 = 0.8$	$\beta_3 = 1.0$	AIC	AICc	$\Delta$	$w$
Model:	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$				
$M_0 : y(x_1, x_2, x_3)$	$1.50 \pm 0.16$	$-0.44 \pm 0.17$	$0.88 \pm 0.08$	$1.17 \pm 0.32$	241.2	241.8	0	89.9 %
$M_1 : y(x_2, x_3)$	$1.56 \pm 0.17$		$0.70 \pm 0.04$	$1.03 \pm 0.32$	245.8	246.3	4.5	9.7 %
$M_2 : y(x_1, x_3)$	$1.64 \pm 0.24$	$1.19 \pm 0.12$		$0.85 \pm 0.47$	319.1	319.5	77.7	< 0.1 %
$M_3 : y(x_1, x_2)$	$2.03 \pm 0.08$	$-0.33 \pm 0.18$	$0.86 \pm 0.09$		252.3	252.7	10.9	0.4 %

Table 16.2: Results of multiple linear regression of example data set (sample size  $n = 100$ ): The estimated intercepts ( $\hat{\beta}_0$ ) are close to the true value (except for  $M_3$ ) and the uncertainties are relatively small. The estimates of the slopes for the first two predictors are reasonable when both predictor are included in the model ( $M_0, M_3$ ), however, for off in model  $M_2$  where the second predictor is dropped. Note that for  $n = 100$  (and  $k \leq 5$ ) the differences between AIC and AICc are already less than one and can be neglected here. The  $\Delta$ s and the probabilities  $w_i$  now support model  $M_0$  and the model  $M_3$  (which was supported based on the small sample set) ends up only at third-best with a probability of less than 1%. [MLRn100AICc.R](#)

### 16.2.4 Dredging another example

Zar (2010, Example 20.1) provides data for multi-linear regression, namely four predictors,  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_4$ , and one response variable,  $Y$  (compare R code below). A multi-linear model using all four predictors ('full model') as well as all 'sub-models', that use only subsets of the predictors, will be compared to each other. AIC is used to pick the 'most optimal' model.

Fitting all sub-models one after the other would require quite a bit of coding. Fortunately, the R package **MuMIn** contains the routine **dredge()** that can do the work for us: it provides a list of all sub-models (model 1 = constant, model 2 = intercept + slope  $\times X_1$ , ..., model 16 = full model) ordered according to the size of the chosen information criterion (here: AIC). The R code and output can be found here [DredgingZar20d1.R](#) and is listed below:

```
print('file: DredgingZar20d1.R')
# dredging of Zar (2010) Example 20.1; MuMIn
# install.packages('MuMIn')
library(MuMIn)
X1 = c(6,1,-2,11,-1,2,5,1,1,3,11,9,5,-3,1,8,-2,3,6,10,4,5,5,3,8,8,6,6,3,5,1,8,10)
X2 = c(99,93,94,91,69,93,79,74,73,88,98,105,91,101,72,117,87,76,
      86,109,76,73,92,70,72,70,88,101,121,77,78,115,104)/10
X3 = c(57,64,57,61,60,57,59,62,55,52,57,61,64,55,55,60,55,62,
      59,56,58,58,52,60,55,64,62,54,54,62,68,62,64)/10
X4 = c(16,30,34,34,30,44,22,22,19,2,42,24,34,30,2,39,22,44,2,24,
      24,44,16,19,16,41,19,22,41,16,24,19,22)/10
Y = c(212,339,361,172,180,321,259,325,286,232,157,150,269,406,
      198,229,355,331,183,169,242,298,184,248,283,241,178,
      222,272,236,281,164,182)/100
Mfull = lm(Y ~ X1+X2+X3+X4)
options(na.action = 'na.fail') # change the default 'na.omit' to prevent models
# from being fitted to different datasets in case of missing values.
Mdfull = dredge(Mfull)
MdfullAIC = dredge(Mfull,rank='AIC')

# ----- results:
# > MdfullAIC
# Global model call: lm(formula = Y ~ X1 + X2 + X3 + X4)
# ---
#   Model selection table
# (Intrc)     X1      X2      X3      X4      df logLik  AIC delta weight
# 10  2.552 -0.1324           0.2013  4 -15.863 39.7  0.00  0.516
# 12  2.673 -0.1305 -0.01542  0.2045  5 -15.816 41.6  1.90  0.199
# 14  2.707 -0.1319           -0.02768 0.2035  5 -15.852 41.7  1.98  0.192
# 16  2.958 -0.1293 -0.01878 -0.04621 0.2088  6 -15.787 43.6  3.85  0.075
# 2   3.050 -0.1292           3 -20.933 47.9  8.14  0.009
# 6   2.355 -0.1312  0.11970  4 -20.768 49.5  9.81  0.004
# 4   2.929 -0.1310  0.01444  4 -20.902 49.8 10.08  0.003
# 8   2.064 -0.1344  0.02259  0.13740  5 -20.694 51.4 11.66  0.002
# 11  3.073 -0.12680  0.2077  4 -30.416 68.8 29.11  0.000
# 15  5.217 -0.14490 -0.35110 0.2398  5 -29.659 69.3 29.59  0.000
# 9   2.019           0.1791  3 -31.838 69.7 29.95  0.000
# 13  3.387 -0.24090  0.1983  4 -31.502 71.0 31.28  0.000
# 1   2.474           2 -33.535 71.1 31.34  0.000
# 3   3.335 -0.09689           3 -32.776 71.6 31.83  0.000
# 5   3.039 -0.09604           3 -33.484 73.0 33.24  0.000
# 7   4.286 -0.10280 -0.15270  4 -32.644 73.3 33.56  0.000
```

```
# Models ranked by AIC(x)
```

Discussion (dredging, AIC):

1. Lowest AIC = 39.7 for model 10:  $Y = 2.552 - 0.1324 X_1 + 0.2013 X_4$
2. Model 12 with only slightly higher AIC = 41.6;  $AIC_{12} - AIC_{10} = 1.9 < 2$ ; AIC differences of less than 2 are considered as 'not significant'.
3. Model 14 with only slightly higher AIC = 41.7;  $AIC_{14} - AIC_{10} = 2.0$ .
4. If no other information is available to choose between models with non-significant differences in AIC, Zuur et al. (2007) suggest to pick the model with the smallest number of model parameters. This would be model 10 here.

## 16.3 Wilkinson notation

Wilkinson & Rogers (1973) proposed a short symbolic notation for specifying models of linear regression. Examples are given in Table 16.3.

Model	Remarks
$\text{lm}(y \sim x)$	straight line: $y = \beta_0 + \beta \cdot x + \text{noise}$
$\text{lm}(y \sim x - 1)$	straight line through origin ('-1' stands for 'no intercept'): $y = \beta \cdot x + \text{noise}$
$\text{lm}(y \sim x_1 + x_2)$	MLR with predictors $x_1$ and $x_2$ ; it does not mean addition of $x_1$ and $x_2$ which would make no sense because in general $x_1$ and $x_2$ have different dimensions/units $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \text{noise}$
$\text{lm}(y \sim x_1 : x_2)$	':' specifies an interaction, i.e. the model is defined by $y = \beta_0 + \beta \cdot x_1 \cdot x_2 + \text{noise}$
$\text{lm}(y \sim x_1 * x_2)$	crossing operator '*' defines an interaction and all lower-order terms (linear, intercept), i.e. the long version is $\text{lm}(y \sim x_1 + x_2 + x_1 \cdot x_2)$ or $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \text{noise}$
$\text{lm}(y \sim x_1 / x_2)$	nesting operator '/': long version is $\text{lm}(y \sim x_1 + x_1 \cdot x_2)$
$\text{lm}(y \sim x_1 * x_2 * x_3 - x_1 : x_2 : x_3)$	model with all interactions among $x_1$ , $x_2$ , and $x_3$ , except the three-way interaction $\Rightarrow$ $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_1 \cdot x_2 + \beta_5 \cdot x_1 \cdot x_3 + \beta_6 \cdot x_2 \cdot x_3 + \text{noise}$
$\text{lm}(y \sim x_1 * x_2^2)$	'^' raises the predictor to a power (exactly as in '*' repeated) so '^' includes lower order terms as well; it is equivalent to $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \beta_4 \cdot x_2^2 + \beta_5 \cdot x_1 \cdot x_2^2 + \text{noise}$
$\text{lm}(y \sim x_1 + x_2 + x_1 : x_2 + x_2 : x_1 + x_1 : x_2 : x_2)$	

Table 16.3: Wilkinson notation:  $y$  is the response variable,  $x$ ,  $x_i$  are predictors, 'noise' stands for normal noise with zero mean, '-' stands for 'without', '1' stands for 'intercept', '+' stands for 'and' (however, not addition of values), ':' stands for 'interaction' (products of predictors), '\*' implies taking into account all terms linear in the predictors and all interactions, '^' is the power function.

### "Random-Effects and Mixed-Effects Models

For random-effects and mixed-effects models, the formula specification includes the names of the predictor variables and the grouping variables. For example, if the predictor variable  $x_1$  is a random effect grouped by the variable  $g$ , then represent this in Wilkinson notation as follows:  $(x_1 | g)$

Remark: a grouping variable in MATLAB is the same as a factor in R.

### Examples:

$$x_1 * x_2 \hat{=} x_1 + x_2 + x_1 \cdot x_2 \quad (16.14)$$

$$x_1 * x_2 * x_3 - x_1 \cdot x_2 \cdot x_3 \hat{=} x_1 + x_2 + x_3 + x_1 \cdot x_2 + x_1 \cdot x_3 + x_2 \cdot x_3 \quad (16.15)$$

$$(x_1 + x_2) * x_3 \hat{=} x_1 + x_2 + x_3 + x_1 \cdot x_3 + x_2 \cdot x_3 \quad (16.16)$$

# Chapter 17

## Collinearity

A common problem in multiple linear regression (MLR) is the occurrence of strong correlation or strong anti-correlation between two or more predictors. If two predictors  $x_1$  and  $x_2$  are highly correlated, it will be difficult to say whether the observed response is 'caused'<sup>1</sup> by  $x_1$  or  $x_2$  or by a linear combination of  $x_1$  and  $x_2$  (attribution problem). Noise in the data can lead to estimates of the regression slopes that are largely off from the true values. Prediction based on the estimated slopes can easily lead to large errors or nonsense (for example, negative concentrations of chemical substances). Ridge regression is a method to deal with the collinearity problem.

### 17.1 Effects of collinearity

For an [ideal model](#) without noise and two identical predictors  $x_1 \equiv x_2$ , i.e. implying perfect correlation, one can write

$$\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 = (\beta_1 + \beta_2) x_1 = (\beta_1 + \beta_2) x_2 = (\beta_1 + c) x_1 + (\beta_2 - c) x_2 \quad (17.1)$$

where  $\mathbf{y}$  is the response and  $c$  is an arbitrary constant. Real world data contain noise and the correlation between predictors is not perfect for various reasons. Consequently, multiple linear regression yields

$$\hat{\mathbf{y}} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = (\beta_1 + c_1) x_1 + (\beta_2 - c_2) x_2 \quad (17.2)$$

where  $\hat{\mathbf{y}}$  is the estimated or predicted response,  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are the estimated slopes, and  $c_1$  and  $c_2$  are the deviations of the estimated from the true slopes. When the predictors  $x_1$  and  $x_2$  are highly correlated, the magnitude of  $c_1$  and  $c_2$  can be large and actually be larger than the magnitude of the true slopes (see numerical example below).

In order to illustrate the effects of collinearity, we consider a [toy regression model](#) with two predictors (Montgomery & Peck, 1982, Section 8.3). The least-squares solution  $\hat{\boldsymbol{\beta}}$  of the regression model

$$\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 + \epsilon = \mathbf{X}\boldsymbol{\beta} + \epsilon \quad (17.3)$$

is given by the solution of the linear system

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad (17.4)$$

which, for appropriate scaling of the variables (Section 17.2.3), reads

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \text{rhs}_1 \\ \text{rhs}_2 \end{pmatrix} \quad (17.5)$$

<sup>1</sup>'Caused' is put here in quotation marks because in linear regression one uses *correlations* between predictor and response variables. The correlations can reflect causalities or predictors can be just proxies for the true causality.

where  $-1 \leq r_{12} \leq +1$ , i.e. the matrix  $\mathbf{X}'\mathbf{X}$  has the form of a correlation matrix (symmetric, ones on the diagonal, off-diagonal elements with magnitude  $\leq 1$ ). The inverse of  $\mathbf{X}'\mathbf{X}$  reads

$$\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \quad (17.6)$$

Finally, the solution  $\hat{\beta}$  reads

$$\hat{\beta}_1 = \frac{\text{rhs}_1 - r_{12} \cdot \text{rhs}_2}{1 - r_{12}^2} \quad (17.7)$$

$$\hat{\beta}_2 = \frac{\text{rhs}_2 - r_{12} \cdot \text{rhs}_1}{1 - r_{12}^2} \quad (17.8)$$

which shows that for high squared correlation coefficients ( $r_{12}^2$  close to 1) one divides by small numbers which can lead to large magnitudes of the slope estimates.

The considerations given above can be illustrated by a [numerical regression example](#) (Fig. 17.1):  $n = 10$  random data  $X_1$  from the standard uniform PDF are generated and the [unit length scaling](#)

$$x_1 = \frac{X_1 - \bar{X}_1}{\sqrt{\sum_i (X_{1,i} - \bar{X}_1)^2}} \quad (17.9)$$

is applied to obtain the scaled predictor vector  $x_1$  which possesses unit length. From now on in this chapter unit scaled vectors are denoted by lowercase letters. The second predictor  $x_2$  is obtained by adding normal noise to  $x_1$  followed by unit length scaling. The two predictors are highly correlated to each other ( $r_{12} = 0.987$ ). The true slopes are set to  $\beta_1 = 1$  and  $\beta_2 = -0.5$ . The response  $Y$  is calculated by

$$Y = \beta_1 x_1 + \beta_2 x_2 + \text{normal noise}. \quad (17.10)$$

Application of unit length scaling to  $Y$  yields  $y$ . The regression problem is solved using the equations given above (Eqs. 17.7 – 17.8; Montgomery & Peck, 1982) or by calling the `lm()` routine in R. The resulting estimates of the slopes are the same for both approaches. In addition, `lm()` provides estimates of the uncertainties of the slopes. One obtains

$$\hat{\beta}_1 = 2.144 \pm 0.387 \quad (17.11)$$

$$\hat{\beta}_2 = -1.190 \pm 0.387 \quad (17.12)$$

which are quite a bit off from the true slopes by  $c_1 = \hat{\beta}_1 - \beta_1 = 1.144$  and  $c_2 = \beta_2 - \hat{\beta}_2 = 0.690$ . Although the response predicted based on the model and these estimates is good (as expected for least squares; Fig. 17.1), the estimated slopes, which have particular meaning in the associated context, are not reliable and should be interpreted with caution (compare below the analysis of the acetylene data).

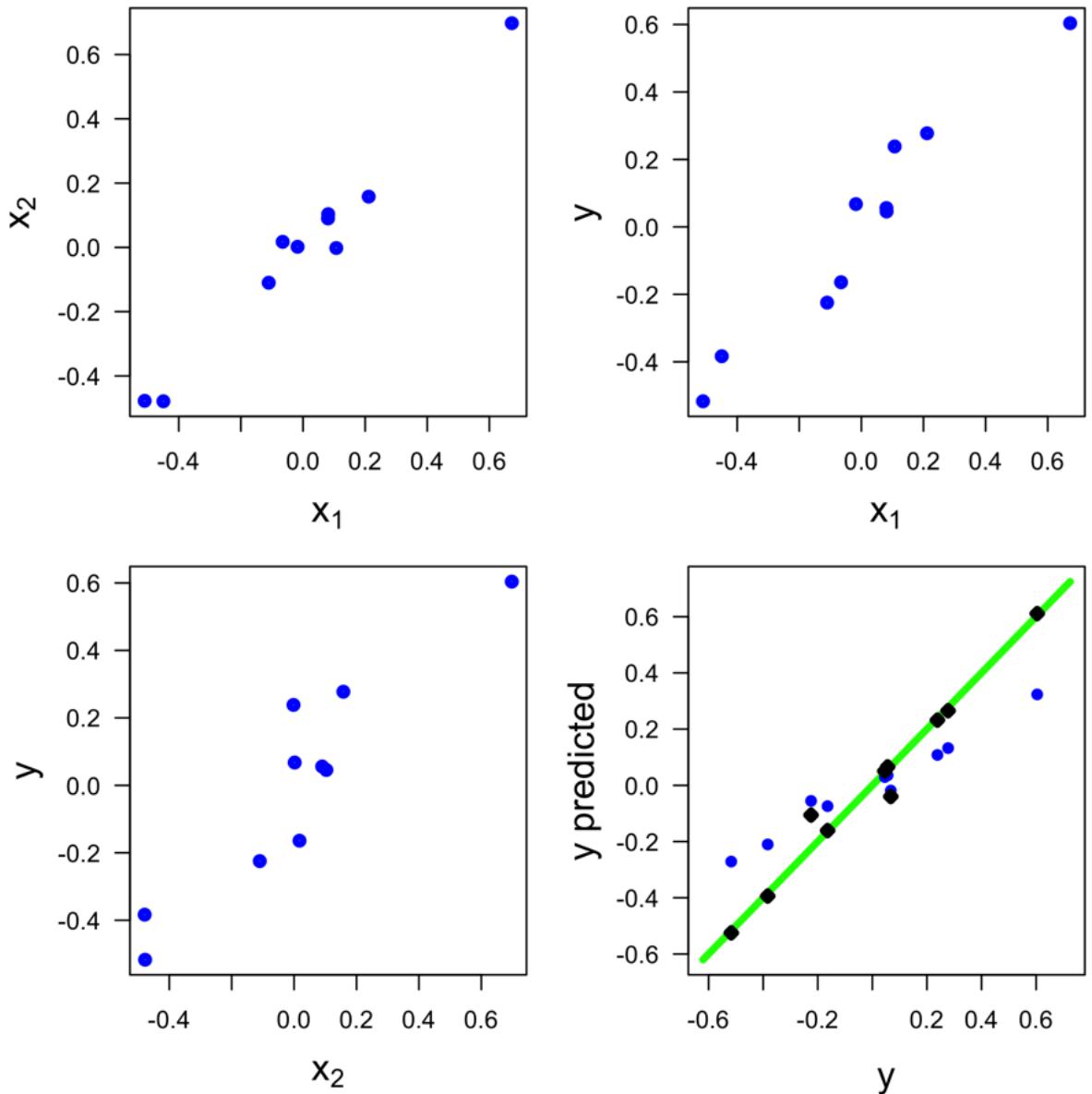


Figure 17.1: Collinearity problem. Toy regression model with two highly ( $r = 0.987$ ) correlated predictors  $x_1$  and  $x_2$  (upper left panel). The relationships of the response  $y$  with the two predictors are shown in the upper right and the lower left panel. The response predicted by the true slopes  $\beta_1 = 1$ ,  $\beta_2 = -0.5$  (blue dots) and by the estimated slopes  $\hat{\beta}_1 = 2.14$ ,  $\hat{\beta}_2 = -1.19$  (red diamonds) compares quite well with each other (despite the large differences between estimated and true slopes) and with the 'observed' response (lower right panel; the green line indicates the 1-to-1 relationship between observed and predicted response). [CollinearityExample.R](#)

## 17.2 Acetylene data, choice of predictors, and unit length scaling

### 17.2.1 The acetylene data

In this section the following acetylene data are analyzed (Kunugi et al., 1961; reprinted in various articles, for example, Marquardt & Snee, 1975)

$$\begin{aligned} X_1 &= \{1300, 1300, 1300, 1300, 1300, 1300, 1200, 1200, 1200, 1200, 1200, 1200, 1100, 1100, 1100, 1100\} \\ X_2 &= \{7.5, 9.0, 11.0, 13.5, 17.0, 23.0, 5.3, 7.5, 11.0, 13.5, 17.0, 23.0, 5.3, 7.5, 11.0, 17.0\} \end{aligned} \quad (17.13)$$

$$\begin{aligned} X_3 &= \{12, 12, 11.5, 13, 13.5, 12, 40, 38, 32, 26, 34, 41, 84, 98, 92, 86\} / 1000 \\ Y &= \{49.0, 50.2, 50.5, 48.5, 47.5, 44.5, 28.0, 31.5, 34.5, 35.0, 38.0, 38.5, 15.0, 17.0, 20.5, 29.5\} \end{aligned} \quad (17.14)$$

where  $X_1$  ( $^{\circ}\text{C}$ ) are reactor temperatures,  $X_2$  ( $\text{mol mol}^{-1}$ )  $\text{H}_2$  to n-heptane ratios,  $X_3$  (s) contact times, and  $Y$  (%) fractions of conversion of n-heptane ( $\text{C}_7\text{H}_{16}$ ) to acetylene ( $\text{C}_2\text{H}_2$ ).

Observation	Conversion of n-heptane to acetylene (%)	Reactor temperature ( $^{\circ}\text{C}$ )	Ratio of $\text{H}_2$ to n-heptane ( $\text{mol mol}^{-1}$ )	Contact time (s)
1	49.0	1300	7.5	0.0120
2	50.2	1300	9.0	0.0120
3	50.5	1300	11.0	0.0115
4	48.5	1300	13.5	0.0130
5	47.5	1300	17.0	0.0135
6	44.5	1300	23.0	0.0120
7	28.0	1200	5.3	0.0400
8	31.5	1200	7.5	0.0380
9	34.5	1200	11.0	0.0320
10	35.0	1200	13.5	0.0260
11	38.0	1200	17.0	0.0340
12	38.5	1200	23.0	0.0410
13	15.0	1100	5.3	0.0840
14	17.0	1100	7.5	0.0980
15	20.5	1100	11.0	0.0920
16	29.5	1100	17.0	0.0860

Table 17.1: Acetylene data (Kunugi et al., 1961; reprinted in various articles, for example, Marquardt & Snee, 1975; data can be loaded in MATLAB by typing 'load acetylene').

A plot of  $X_3$  versus  $X_1$  shows the hight anti-correlation ( $r = -0.958$ ) between these two predictors.

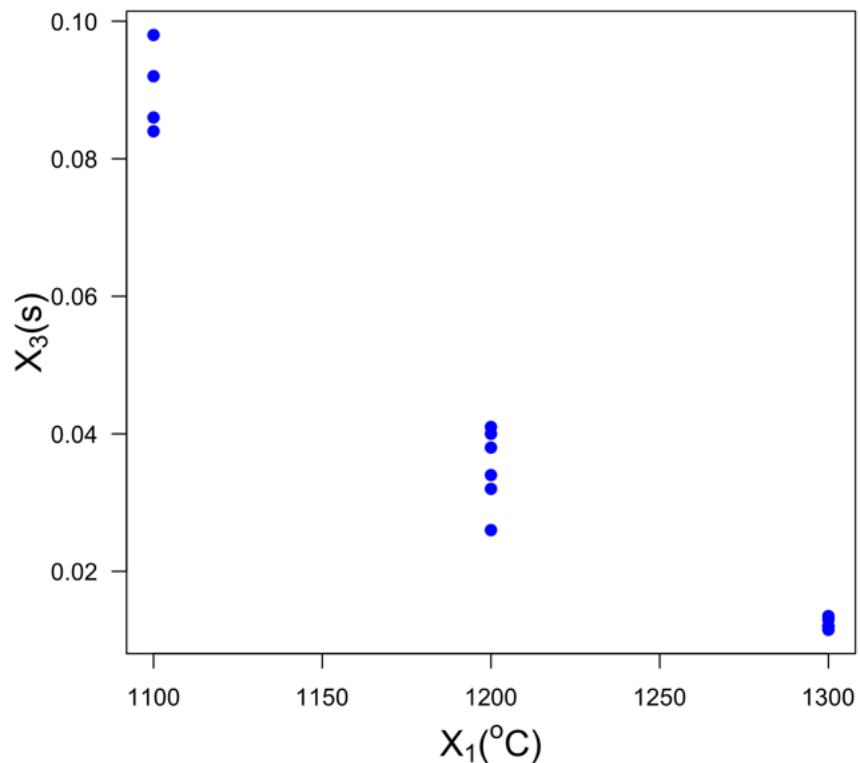


Figure 17.2:  $X_3$  versus  $X_1$ : these two predictors are highly anti-correlated ( $r = -0.958$ ).  
[AcetyleneX1X3antiCor.R](#)

### 17.2.2 Extended set of predictors: interaction and quadratic terms

Although only three predictors have been measured (Table 17.1), more predictors can be constructed by products of different predictors (called 'interactions') or by powers of single predictors. A regression model that takes into all linear and quadratic terms encompasses nine predictors (Table 17.2).

$X_1$	$X_2$	$X_3$	$X_4$ $= X_1 \cdot X_2$	$X_5$ $= X_1 \cdot X_3$	$X_6$ $= X_2 \cdot X_3$	$X_7$ $= X_1^2$	$X_8$ $= X_2^2$	$X_9$ $= X_3^2$
°C	mol mol <sup>-1</sup>	s	°C mol mol <sup>-1</sup>	°C s	mol mol <sup>-1</sup> s	(°C) <sup>2</sup>	(mol mol <sup>-1</sup> ) <sup>2</sup>	s <sup>2</sup>
1300	7.5	0.0120	9750	15.60	0.0900	1690000	56.25	0.00014
1300	9.0	0.0120	11700	15.60	0.1080	1690000	81.00	0.00014
1300	11.0	0.0115	14300	14.95	0.1265	1690000	121.00	0.00013
1300	13.5	0.0130	17550	16.90	0.1755	1690000	182.25	0.00017
1300	17.0	0.0135	22100	17.55	0.2295	1690000	289.00	0.00018
1300	23.0	0.0120	29900	15.60	0.2760	1690000	529.00	0.00014
1200	5.3	0.0400	6360	48.00	0.2120	1440000	28.09	0.00160
1200	7.5	0.0380	9000	45.60	0.2850	1440000	56.25	0.00144
1200	11.0	0.0320	13200	38.40	0.3520	1440000	121.00	0.00102
1200	13.5	0.0260	16200	31.20	0.3510	1440000	182.25	0.00068
1200	17.0	0.0340	20400	40.80	0.5780	1440000	289.00	0.00116
1200	23.0	0.0410	27600	49.20	0.9430	1440000	529.00	0.00168
1100	5.3	0.0840	5830	92.40	0.4452	1210000	28.09	0.00706
1100	7.5	0.0980	8250	107.80	0.7350	1210000	56.25	0.00960
1100	11.0	0.0920	12100	101.20	1.0120	1210000	121.00	0.00846
1100	17.0	0.0860	18700	94.60	1.4620	1210000	289.00	0.00740

Table 17.2: Nine predictors based on the acetylene data (Table 17.1): note the uncommon ('strange') and hard to interpret units of the interaction terms. Note that  $X_1 \cdot X_2$  is meant here as element-wise multiplication yielding the vector  $X_4$  and not the scalar product.

#### Exercise 47 Correlation matrix

Calculate the correlation matrix for the 9 predictors of the acetylene data, plot a heatmap of the correlation matrix, and discuss the results.

### 17.2.3 Unit length scaling

Unit length scaling of samples is described, for example, in Montgomery & Peck (1982). In the current context (multiple linear regression analysis), the goal is to use scaling in order to obtain dimensionless predictors  $x_j$ , lump them together into a matrix  $\mathbf{X}$  with the property that  $\mathbf{X}'\mathbf{X}$  is the correlation matrix. The unit length scaling encompasses subtraction of the sample mean (centering) followed by division by the square root of the sample sum of squares:

$$x_j = \frac{\mathbf{X}_j - \bar{\mathbf{X}}_j}{S_{\mathbf{X}_j}} \quad (17.15)$$

where

$$S_{\mathbf{X}_j} = \sqrt{\sum_i (X_{j,i} - \bar{X}_j)^2}. \quad (17.16)$$

$\mathbf{X}'\mathbf{X}$  (Table 17.4) is the correlation matrix.

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$
0.28022	-0.22544	-0.23106	-0.19552	-0.23706	-0.24550	0.28422	-0.20569	-0.18360
0.28022	-0.15704	-0.23106	-0.12536	-0.23706	-0.23360	0.28422	-0.16612	-0.18360
0.28022	-0.06584	-0.23514	-0.03180	-0.24203	-0.22137	0.28422	-0.10217	-0.18449
0.28022	0.04817	-0.22289	0.08515	-0.22711	-0.18897	0.28422	-0.00424	-0.18170
0.28022	0.20777	-0.21881	0.24887	-0.22214	-0.15326	0.28422	0.16643	-0.18070
0.28022	0.48138	-0.23106	0.52954	-0.23706	-0.12252	0.28422	0.55015	-0.18360
-0.04003	-0.32577	-0.00255	-0.31751	0.01081	-0.16483	-0.04820	-0.25071	-0.07311
-0.04003	-0.22544	-0.01887	-0.22251	-0.00755	-0.11657	-0.04820	-0.20569	-0.08495
-0.04003	-0.06584	-0.06784	-0.07138	-0.06264	-0.07226	-0.04820	-0.10217	-0.11682
-0.04003	0.04817	-0.11680	0.03657	-0.11772	-0.07293	-0.04820	-0.00424	-0.14323
-0.04003	0.20777	-0.05152	0.18770	-0.04428	0.07717	-0.04820	0.16643	-0.10681
-0.04003	0.48138	0.00561	0.44678	0.01999	0.31850	-0.04820	0.55015	-0.06697
-0.36029	-0.32577	0.35653	-0.33658	0.35047	-0.01064	-0.35403	-0.25071	0.34090
-0.36029	-0.22544	0.47078	-0.24950	0.46829	0.18097	-0.35403	-0.20569	0.53424
-0.36029	-0.06584	0.42182	-0.11096	0.41779	0.36413	-0.35403	-0.10217	0.44774
-0.36029	0.20777	0.37285	0.12653	0.36730	0.66167	-0.35403	0.16643	0.36670

Table 17.3: Matrix  $\mathbf{X}$  of the unit length scaled predictors where  $x_4 = \text{scaling}(X_1 \cdot X_2)$  etcetera.

1.00000	0.22363	-0.95820	0.34632	-0.96279	-0.75623	<b>0.99967</b>	0.20192	-0.89263
0.22363	1.00000	-0.24023	0.99020	-0.23848	0.33343	0.22124	0.98172	-0.24776
-0.95820	-0.24023	1.00000	-0.35284	<b>0.99957</b>	0.76227	-0.95148	-0.21346	0.98089
0.34632	0.99020	-0.35284	1.00000	-0.35204	0.20450	0.34431	0.97431	-0.34895
-0.96279	-0.23848	<b>0.99957</b>	-0.35204	1.00000	0.76450	-0.95673	-0.20993	0.97545
-0.75623	0.33343	0.76227	0.20450	0.76450	1.00000	-0.75316	0.31923	0.72331
<b>0.99967</b>	0.22124	-0.95148	0.34431	-0.95673	-0.75316	1.00000	0.19917	-0.88162
0.20192	0.98172	-0.21346	0.97431	-0.20993	0.31923	0.19917	1.00000	-0.22954
-0.89263	-0.24776	0.98089	-0.34895	0.97545	0.72331	-0.88162	-0.22954	1.00000

Table 17.4: The correlation matrix  $\mathbf{X}'\mathbf{X}$ : the magnitude of several off-diagonal elements is quite high (highlighted in blue). Compare also Fig. 17.3.

Note that Montgomery & Peck (1982) apply unit length scaling to quadratic terms formed from unit length scaled predictors  $x_1, x_2$  and  $x_3$ , i.e. for example

$$x_{4,\text{MP82}} = \text{scaled} [x_1 \cdot x_2] = \text{scaled} [\text{scaled} (\mathbf{X}_1) \cdot \text{scaled} (\mathbf{X}_2)]. \quad (17.17)$$

In general,

$$x_{4,\text{MP82}} = \text{scaled} [\text{scaled} (\mathbf{X}_1) \cdot \text{scaled} (\mathbf{X}_2)] \neq \text{scaled} [\mathbf{X}_1 \cdot \mathbf{X}_2] = x_4 \quad (17.18)$$

and thus the unit length scaled predictors  $x_{j,\text{MP82}}, j = 4, 5, \dots, 9$  look different from those in Table 17.3. Calculating the estimated slopes for the original predictors as, for example,  $\mathbf{X}_1 \cdot \mathbf{X}_2$  is tedious for the Montgomery & Peck scaling (compare Section L.2 in the Appendix), however, yields the same results as the scaling applied here.

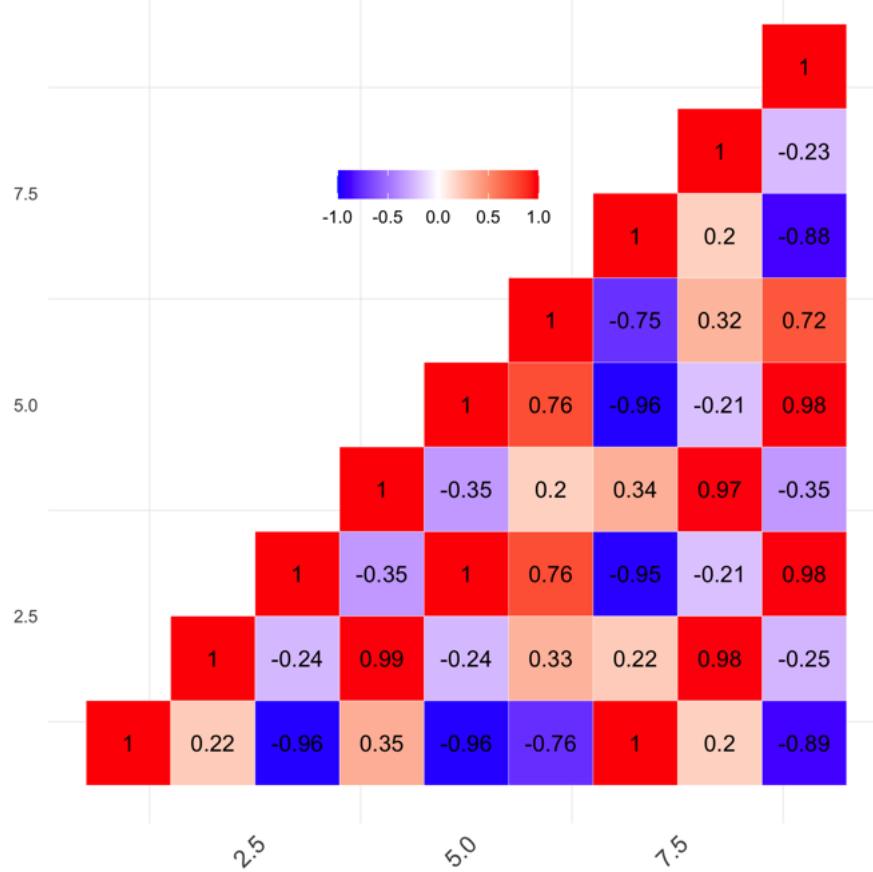


Figure 17.3: Plot of the symmetric matrix  $\mathbf{X}'\mathbf{X}$  (Table 17.4). [AcetyleneULScorMat.R](#), [plotCorMatrix.R](#)

### 17.3 The Webster et al. (1974) data

Webster et al. (1974) generated the artificial response ( $Y$ ) and predictor ( $X_j, j = 1, \dots, 6$ ) data (Table 17.5)

Observation	$Y$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
1	10.006	8.000	1.000	1.000	1.000	0.541	-0.099
2	9.737	8.000	1.000	1.000	0.000	0.130	0.070
3	15.087	8.000	1.000	1.000	0.000	2.116	0.115
4	8.422	0.000	0.000	9.000	1.000	-2.397	0.252
5	8.625	0.000	0.000	9.000	1.000	-0.046	0.017
6	16.289	0.000	0.000	9.000	1.000	0.365	1.504
7	5.958	2.000	7.000	0.000	1.000	1.996	-0.865
8	9.313	2.000	7.000	0.000	1.000	0.228	-0.055
9	12.960	2.000	7.000	0.000	1.000	1.380	0.502
10	5.541	0.000	0.000	0.000	10.000	-0.798	-0.399
11	8.756	0.000	0.000	0.000	10.000	0.257	0.101
12	10.937	0.000	0.000	0.000	10.000	0.440	0.432

Table 17.5: Artificial response ( $Y$ ) and predictor ( $X_j, j = 1, \dots, 6$ ) data (Webster et al., 1974)

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
0.4842	-0.0990	-0.1147	-0.150	0.0475	-0.1230
0.4842	-0.0990	-0.1147	-0.222	-0.0552	-0.0327
0.4842	-0.0990	-0.1147	-0.222	0.4409	-0.0087
-0.2201	-0.1980	0.4971	-0.150	-0.6864	0.0645
-0.2201	-0.1980	0.4971	-0.150	-0.0992	-0.0610
-0.2201	-0.1980	0.4971	-0.150	0.0035	0.7331
-0.0440	0.4951	-0.1912	-0.150	0.4109	-0.5320
-0.0440	0.4951	-0.1912	-0.150	-0.0307	-0.0995
-0.0440	0.4951	-0.1912	-0.150	0.2570	0.1980
-0.2201	-0.1980	-0.1912	0.498	-0.2870	-0.2832
-0.2201	-0.1980	-0.1912	0.498	-0.0235	-0.0162
-0.2201	-0.1980	-0.1912	0.498	0.0222	0.1606

Table 17.6: Matrix  $\mathbf{X}$  of the unit length scaled predictors of the Webster et al. (1974) data.

1.0000	0.0523	-0.3434	-0.4976	0.4173	-0.1921
0.0523	1.0000	-0.4316	-0.3707	0.4845	-0.3167
-0.3434	-0.4316	1.0000	-0.3551	<b>-0.5052</b>	0.4944
-0.4976	-0.3707	-0.3551	1.0000	-0.2146	-0.0869
0.4173	0.4845	<b>-0.5052</b>	-0.2146	1.0000	-0.1230
-0.1921	-0.3167	0.4944	-0.0869	-0.1230	1.0000

Table 17.7: The correlation matrix  $\mathbf{X}'\mathbf{X}$  for the Webster et al. (1974) data where  $\mathbf{X}$  is the matrix of unit length scaled predictors (Table 17.6).

## 17.4 Diagnostic of collinearity and multicollinearity

Several techniques have been proposed to detect collinearity or multicollinearity (the near linear dependence of more than two predictors). Dependencies between two predictors can be detected by inspection of the correlation matrix. This diagnostic is, however, not sufficient in the case of multicollinearity. Eigenvalue analysis of the correlation matrix, singular value decomposition of the predictor matrix, and variance inflation factors can be better options.

### 17.4.1 Inspection of the correlation matrix

Inspection of the correlation matrix of the acetylene data (Table 17.4) yields several extremely high correlation values: the correlations between  $x_1$  and  $x_7$  and between  $x_3$  and  $x_5$  are above 0.999. Thus it is obvious that one has to deal with collinearity problems.

The maximum magnitude of the correlations  $r_{i,j}$ ,  $i \neq j$  of the Webster data (Table 17.7) is much smaller:  $x_3$  and  $x_5$  are anti-correlated with a correlation of  $r_{3,5} = -0.51$ . The small magnitude of this correlation speaks against collinearity. However, it will be seen in a moment that this data set is plagued by multicollinearity. Inspection of the correlation matrix alone is often not sufficient to detect multicollinearity.

### 17.4.2 Variance inflation factors

The diagonal elements of the inverse of the correlation matrix  $\mathbf{X}'\mathbf{X}$  are called variance inflation factors (VIFs, Marquardt, 1970). 'One or more large VIFs indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.' Montgomery & Peck (1982, p. 300). The VIFs for the acetylene data

$$\text{VIFs}_{\text{acetylene}} = \{2856749, 10956, 2017163, 2501945, 65.7, 12667, 9803, 1428092, 240.4\} \quad (17.19)$$

are all larger than 10 and four are larger than  $10^6$ .

The VIFs for the Webster data read

$$\text{VIFs}_{\text{Webster}} = \{182.1, 161.4, 266.3, 297.7, 1.92, 1.46\}. \quad (17.20)$$

Four of these VIFs are larger than 100. This is indicating multicollinearity which was not obvious from the inspection of the correlation matrix. Thus with VIFs one can detect multicollinearity that is not recognizable by inspection of the correlation matrix.

### 17.4.3 Eigensystem analysis of the correlation matrix

The eigenvalues of the correlation matrix  $\mathbf{X}'\mathbf{X}$  can be used to measure the extend of multicollinearity in the predictor data. The ratio of the largest and the smallest eigenvalue is called the condition number of  $\mathbf{X}'\mathbf{X}$

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \quad (17.21)$$

If  $\kappa < 100$ , there is no serious problem with multicollinearity. For  $100 < \kappa < 1000$  moderate to strong multicollinearity is occurring. Severe multicollinearity is indicated by  $\kappa > 1000$ . This is obviously the case for the 9-predictor acetylene model with  $\kappa_{\text{acetylene}} = 50202670$  and also for the Webster data with  $\kappa_{\text{Webster}} = 2195.9$ .

### 17.4.4 Singular value decomposition of the predictor matrix

The matrix  $\mathbf{X}$  of unit length scaled predictors is in general not a square matrix and thus the standard eigensystem analysis is not applicable. However, one can perform a singular value decomposition of  $\mathbf{X}$ . This yields

the singular values  $\mu_j$ . The **condition number** of  $\mathbf{X}$  is defined as the ratio of the maximum to the minimum singular value

$$\eta = \frac{\mu_{\max}}{\mu_{\min}}. \quad (17.22)$$

Condition numbers  $\eta > 30$  indicate multicollinearity. This is the case for the acetylene data with  $\eta_{\text{acetylene}} = 7085.4$  as well as for the Webster data with  $\eta_{\text{Webster}} = 46.86$ .

### 17.4.5 Summary

*Inspection of the correlation matrix is a valid technique to detect collinearity. Multicollinearity can be detected using variance inflation factors (VIFs), eigensystem analysis of the correlation matrix  $\mathbf{X}'\mathbf{X}$ , or singular value decomposition of the matrix of unit scaled predictors  $\mathbf{X}$ . Further detection techniques for multicollinearity are discussed in Montgomery & Peck (1982).*

## 17.5 Multiple linear regression of the acetylene data

*Application of multiple linear regression (MLR) to data plagued by (multi)collinearity can yield results that fit the data almost perfectly (high value of the coefficient of determination), however, give unsatisfactory results when applied to new data or used in extrapolation. In this section, MLR is applied to the original acetylene data (not scaled) as well as to the unit length scaled data. It is shown here that the two approaches are equivalent, i.e. the slope and intercept estimates for the original data can be derived from the slope and intercept estimates for the scaled data. What may look like a detour, will be soon a path to more satisfactory parameter estimates via ridge regression.*

### 17.5.1 Multiple linear regression of the acetylene data (not scaled)

The nine predictors  $X_j$  (merged into the matrix  $X$ ; Table 17.2) will be used in setting up a multiple linear regression model

$$Y = \beta_{0,o} + X \cdot \beta_o + \text{normal noise} \quad (17.23)$$

where  $Y$  is the response vector,  $\beta_o$  is the vector of 9 slopes,  $\beta_{0,o}$  is the intercept, and the noise is from a normal distribution with mean  $\mu = 0$  and (unknown) variance  $\sigma^2$ . The slopes  $\beta_{j,o}$  and the intercept  $\beta_{0,o}$  can be estimated by the least squares method (call the R routine `lm()`). The results and their uncertainties are listed in Table 17.8. Uncertainties of the same size as the magnitude of the estimated model parameters already indicate that these parameters are not significantly different from zero; this applies to the slopes  $\beta_{1,o}, \beta_{3,o}, \beta_{5,o}, \beta_{7,o}$  and the intercept  $\beta_{0,o}$ . Two-sided one-sample *t*-tests confirm this statement. The magnitude of the estimated slopes can not be compared to each other because they possess different units and the range of variations of the corresponding predictors varies. Therefore the MLR exercise will be repeated using unit length scaled data.

Slope number $j$	$\hat{\beta}_{j,o}$	$\hat{\beta}_{j,s}$
1	$5.32 \pm 4.88$	$36.08 \pm 33.06$
2	$19.24 \pm 4.30$	$9.16 \pm 2.05$
3	$13770 \pm 10450$	$36.60 \pm 27.78$
4	$-0.0141 \pm 0.0032$	$-8.53 \pm 1.94$
5	$-10.58 \pm 8.24$	$-30.00 \pm 23.38$
6	$-21.03 \pm 9.24$	$-0.69 \pm 0.30$
7	$-0.00193 \pm 0.00190$	$-31.44 \pm 30.94$
8	$-0.00303 \pm 0.00117$	$-0.412 \pm 0.16$
9	$-11580 \pm 7699$	$-3.31 \pm 2.20$
0	$-3617 \pm 3136$	

Table 17.8: Estimates of slopes  $\beta_j$  and their uncertainties for the original acetylene data,  $\hat{\beta}_{j,o}$  and for the unit length scaled data,  $\hat{\beta}_{j,s}$ . The magnitudes of  $\hat{\beta}_{1,o/s}$ ,  $\hat{\beta}_{3,o/s}$ ,  $\hat{\beta}_{5,o/s}$ , and  $\hat{\beta}_{7,o/s}$  and their uncertainties are large indicating that the corresponding slopes are not significantly different from zero ( $\alpha = 0.05$ , two-sided one-sample *t*-test); only  $\beta_{2,o/s}$ ,  $\beta_{4,o/s}$ , and  $\beta_{8,o/s}$  are significantly different from zero. The estimated intercept  $\hat{\beta}_{0,o}$  is negative with a large magnitude, however, with a large uncertainty and thus  $\beta_{0,o}$  is not significantly different from zero. Note that the intercept  $\beta_{0,s}$  for the scaled data is not defined (or 'zero by definition').

### 17.5.2 Multiple linear regression of the scaled acetylene data

Now MLR is performed using the nine unit length scaled predictors  $x_j$  (Table 17.3) and the model

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta}_s + \text{normal noise} \quad (17.24)$$

where  $\mathbf{y}$  is the unit length scaled response

$$\begin{aligned} \mathbf{y} = & \{0.2798, 0.3058, 0.3123, 0.2689, 0.2472, 0.1821, -0.1759, -0.1000, -0.0349, \\ & -0.0240, 0.0411, 0.0519, -0.4580, -0.4146, -0.3386, -0.1434\} \end{aligned} \quad (17.25)$$

The vector of slopes  $\boldsymbol{\beta}_s$  can be estimated by calling the linear modeling routine `lm()` in R or or solving the linear system

$$\mathbf{X}' \mathbf{X} \boldsymbol{\beta}_s = \mathbf{X}' \mathbf{y}. \quad (17.26)$$

Both ways are based on least squares and yield identical results whereby `lm()` provides also estimates of the uncertainties of the slopes (Table 17.8). Uncertainties of the same size as the magnitude of the estimated model parameters already indicate that these parameters are not significantly different from zero; this applies to the slopes  $\beta_{1,s}$ ,  $\beta_{3,s}$ ,  $\beta_{5,s}$ , and  $\beta_{7,s}$ . Two-sided one-sample  $t$ -tests confirm this statement. The magnitude of the estimated slopes can be compared to each other because they are all dimensionless and the range of variations of the corresponding predictors are all the same. The magnitudes of the values  $\hat{\beta}_{1,s}$ ,  $\hat{\beta}_{3,s}$ ,  $\hat{\beta}_{5,s}$ , and  $\hat{\beta}_{7,s}$  are the largest of all slope estimates, i.e. the problem with large and uncertain slope estimates does not vanish with scaling.

### 17.5.3 Predicted response values

The estimated slopes  $\hat{\beta}_{j,s}$  can be used to predict the response values

$$\mathbf{Y}_{predicted} = \mathbf{X} \hat{\boldsymbol{\beta}}_s \cdot S_Y + \bar{Y}. \quad (17.27)$$

The predicted values are very close to the observed values (Fig. 17.4); the coefficient of determination is  $r^2 = 0.9977$ . How is this possible given the large magnitude and large uncertainty of 4 out of nine estimated slopes? Of course `lm()` generates a fit that is optimal in the least squares sense. If some of the estimated slopes are far away from their true values, the products with their corresponding predictors can still compensate each other and yield reasonable response values. Consider the following four vectors

$$\begin{aligned} \mathbf{v}_1 &= \hat{\beta}_{1,s} \mathbf{x}_1 \\ &= \{10.1, 10.1, 10.1, 10.1, 10.1, 10.1, -1.4, -1.4, -1.4, -1.4, -1.4, -1.4, -1.4, -13, -13, -13, -13\} \\ \mathbf{v}_3 &= \hat{\beta}_{3,s} \mathbf{x}_3 \\ &= \{-8.5, -8.5, -8.6, -8.2, -8, -8.5, -0.1, -0.7, -2.5, -4.3, -1.9, 0.2, 13.1, 17.2, 15.4, 13.6\} \\ \mathbf{v}_5 &= \hat{\beta}_{5,s} \mathbf{x}_5 \\ &= \{7.1, 7.1, 7.3, 6.8, 6.7, 7.1, -0.3, 0.2, 1.9, 3.5, 1.3, -0.6, -10.5, -14, -12.5, -11\} \\ \mathbf{v}_7 &= \hat{\beta}_{7,s} \mathbf{x}_7 \\ &= \{-8.9, -8.9, -8.9, -8.9, -8.9, -8.9, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 11.1, 11.1, 11.1, 11.1\} \end{aligned} \quad (17.28)$$

Most elements of these vectors have much larger magnitudes compared to the magnitudes of the elements of the scaled response vector (Eq. 17.25). Combinations of these vectors as, for example,  $\mathbf{v}_1 + \mathbf{v}_3$ ,  $\mathbf{v}_5 + \mathbf{v}_7$ , or  $\mathbf{v}_1 + \mathbf{v}_3 + \mathbf{v}_5 + \mathbf{v}_7$ , have elements of much smaller magnitudes (Exercise 48). The **compensation effect** can be illustrated by drawing the contributions of each predictor to the first element of the predicted scaled response (Fig. 17.5). However, this compensation effect usually breaks down when new predictor data are used (see next subsection).

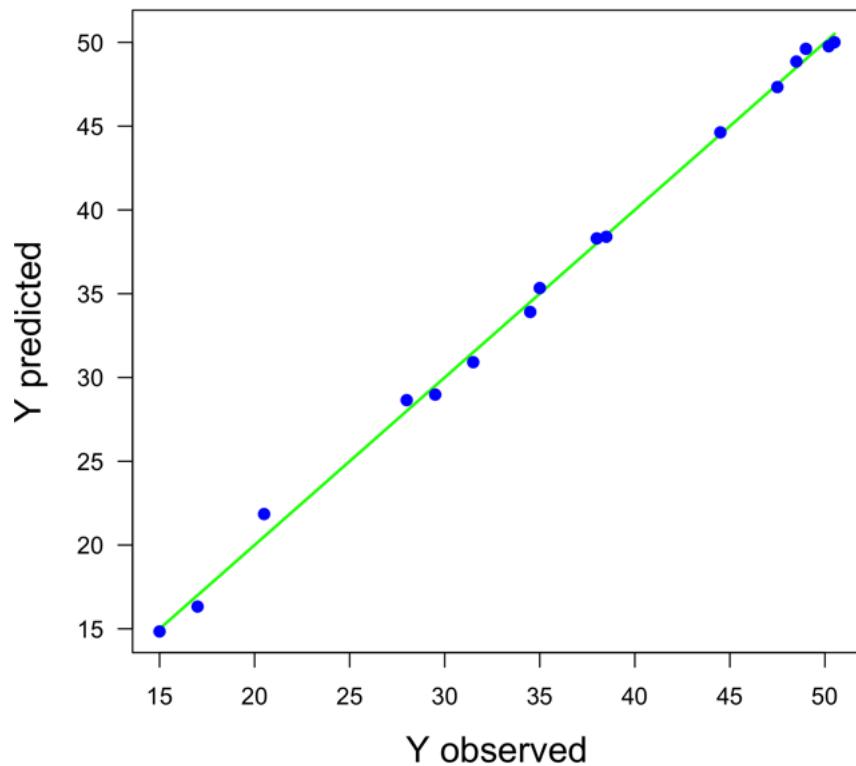


Figure 17.4: Predicted response  $Y_{predicted}$  (based on MLR) over  $Y_{observed}$  (blue dots) and the 1-to-1 relationship (green line): the predicted values are very close to the observed values; coefficient of determination  $r^2 = 0.9977$ .  
[AcetylenePredictionYY.R](#)

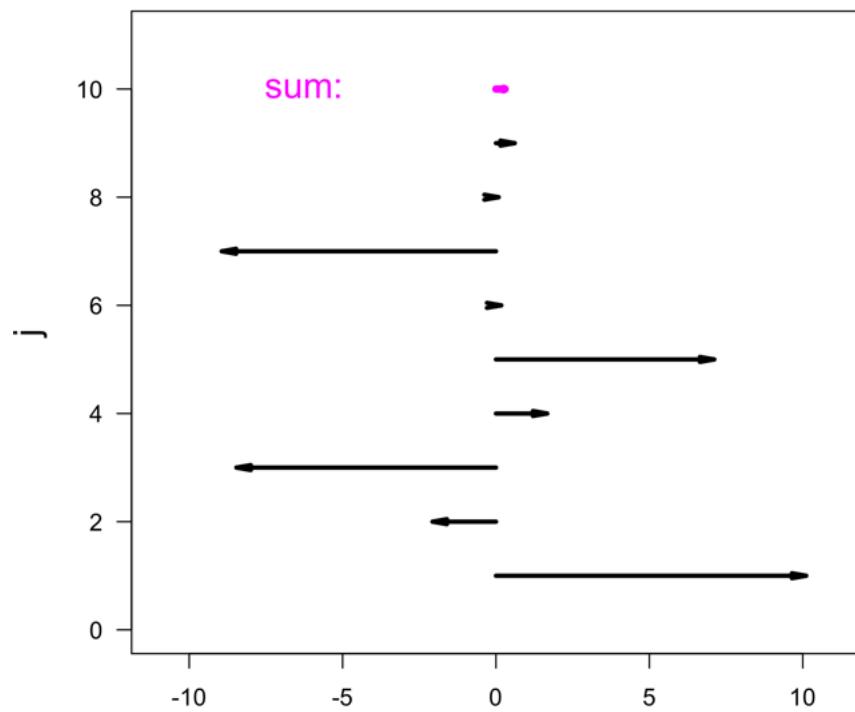


Figure 17.5: The compensation effect can be illustrated by drawing the contributions of each predictor to the first element of the predicted scaled response: the black arrows correspond to the contribution  $\hat{\beta}_{j,s} \cdot x_{j,1}$  to the sum  $y_1 = \sum_{j=1}^9 \hat{\beta}_{j,s} \cdot x_{j,1}$  (magenta). The magnitude of the sum is much smaller than the magnitude of the contributions from  $j = 1, 3, 5$ , and  $7$ . [AcetyleneCompensation.R](#)

### 17.5.4 Extrapolation

The results of MLR are often used to predict response values for predictor values that have not been observed. For the current acetylene example relatively mild extrapolation is taken by choosing for  $X_2$  the mean observed value (rounded to one decimal) and for  $X_1$  and  $X_3$  values between their observed minimal and maximal values; the predictors  $X_4$  to  $X_9$  are calculated from  $X_1$  to  $X_3$  as before. Negative response values  $Y$  are predicted for large regions of combinations of  $X_1$  and  $X_3$  (Fig. 17.6). These values make no sense because 'negative conversion' (from acetylene to n-heptane) is not possible.

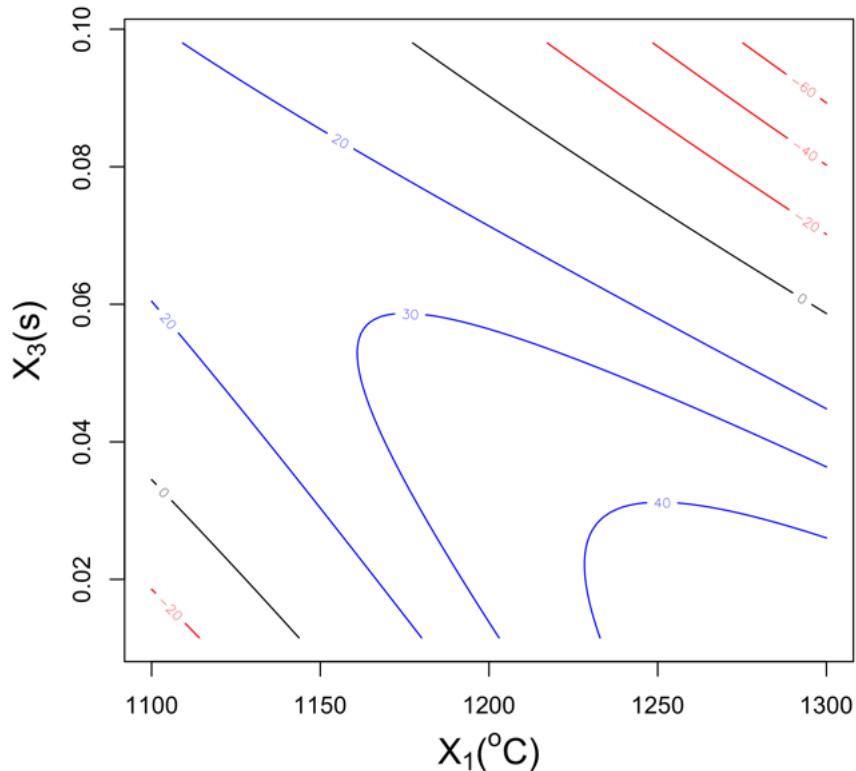


Figure 17.6: Iso-contour plot for the predicted response  $Y$  (conversion of n-heptane to acetylene) for  $X_2 = 12.4 \text{ mol mol}^{-1} \text{ H}_2$  to n-heptane ratio.  $X_1$  and  $X_3$  values vary between their observed minimal and maximal values. Negative response values (regions with red contour lines) make no sense because 'negative conversion' (from acetylene to n-heptane) is not possible. [AcetyleneRidgeCon.R](#) [ridgeDWG1706s.R](#)

### 17.5.5 Summary

MLR of data plagued by (multi)collinearity yield slope estimates with large magnitudes and large uncertainties. Although the fit to the response values can look very good (high coefficient of determination), even mild extrapolation can yield unreasonable response values (for example, negative concentrations). Thus the results of MLR for (multi)collinearity are not satisfactory for many purposes and thus one should look for alternative approaches. In the next section, we will discuss ridge regression which is a powerful method that can deal with (multi)collinearity problems.

#### Exercise 48 Compensation effect

Calculate the combinations  $v_1 + v_3$ ,  $v_5 + v_7$ , or  $v_1 + v_3 + v_5 + v_7$  of the vectors  $v_1, v_3, v_5, v_7$  given in Eq. 17.28 and use the lengths of the vectors as a measure of the compensation effect mentioned in Section 17.5.3.

## 17.6 Ridge regression

Ridge regression is a method to deal with (multi)collinearity problems. The algorithm for ridge regression is relatively simple: it consists essentially of solving a slightly perturbed linear system. Before justifying this method (including the Bayesian approach), it is applied to the acetylene data and the results are compared to classical MLR discussed before.

### Remarks:

(1) Ridge regression has been invented several times and thus is known under various names as, for example, Tikhonov regularization (after Tikhonov, 1943), Tikhonov-Phillips regularization, weight decay, Tikhonov-Miller method, Phillips-Twomey method, constrained linear inversion method, or method of linear regularization.

(2) Ridge regression of the acetylene data has been discussed by various authors (including Marquardt & Snee, 1975; Snee, 1977; Smith & Campbell, 1980; Montgomery & Peck, 1982; Charnes et al., 1986). Unfortunately, most of them use a particular choice of predictors (they first scale the three observed predictors, then generate an extended set of predictors by forming quadratic terms from the three scaled predictors, and finally scale the quadratic terms) which makes 'un-scaling' of the slopes a bit tedious. Several numerical values in some of the articles differ from values obtained by using the R codes listed below; the reasons for these deviations are not known (rounding errors, typos?). We will take a somewhat different choice of predictors in that we first form quadratic terms from the original data and then scale all predictors.

### 17.6.1 Ridge regression of the acetylene data: choice of predictors and algorithm

#### Step 1 choose the predictor variables.

In Model 1 we will consider in addition to the observed predictors  $X_1$ ,  $X_2$ , and  $X_3$  the following quadratic combinations as predictors:

$$X_4 = X_1 X_2, \quad X_5 = X_1 X_3, \quad X_6 = X_2 X_3, \quad X_7 = X_1^2, \quad X_8 = X_2^2, \quad X_9 = X_3^2 \quad (17.29)$$

where  $X_4$ ,  $X_5$ , and  $X_6$  are called the 'interaction' terms.

#### Step 2 Apply unit length scaling to all predictors $X_k$ and the response $Y$

$$x_k = (X_k - \bar{X}_k) / S_{X_k} \quad (17.30)$$

where  $\bar{X}_k$  is the sample mean and  $S_{X_k}$  is the square root of the sample sum of squares

$$S_{X_k} = \sqrt{\sum_i (X_{k,i} - \bar{X}_k)^2}. \quad (17.31)$$

Consequently the matrix  $\mathbf{X}'\mathbf{X}$  (see below) will look like a correlation matrix, i.e. it is symmetric, has ones on the diagonal, and the off-diagonal elements are in the range  $-1$  to  $+1$ .

#### Step 3 Set up the model for the scaled variables:

The scaled response variable  $y$  is related to the scaled predictors  $x_k$  by

$$y = \beta_{1,s} x_1 + \beta_{2,s} x_2 + \cdots + \beta_{9,s} x_9 + \epsilon \quad (17.32)$$

where  $\beta_{j,s}$  are the model parameters (in the current context called 'slopes') and  $\epsilon$  is additive normal noise with zero mean and (unknown) variance  $\sigma^2$ . This can be written in more compact matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_s + \epsilon \quad (17.33)$$

where the columns of the matrix  $\mathbf{X}$  consists of the scaled predictors  $x_1, \dots, x_9$ .

**Reminder:** MLR least squares solution of the unperturbed system:

Multiplication of Eq. 17.33 from left by the transpose of  $\mathbf{X}$ ,  $\mathbf{X}'$ , yields

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}_s + \mathbf{X}'\epsilon \quad (17.34)$$

It can be shown that

$$\hat{\beta}_s = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (17.35)$$

yields a solution in the least squares sense.  $\mathbf{X}'\mathbf{X}$  is a correlation matrix, i.e. it is symmetric, consists of ones on the diagonal, and off-diagonal elements are in the range  $-1$  to  $+1$ . It is difficult to invert when the magnitude of some off-diagonal elements becomes large, i.e.  $|XX_{i,j}|$  close to 1 for  $i \neq j$ . This is the case when some of the predictors are strongly correlated or anti-correlated to each other (collinearity problem). Even if the numerical evaluation yield values for  $\hat{\beta}_s$  that lead to predictions fitting the observed data quite well, the solution is often not satisfactory at all because the length of the vector  $\hat{\beta}_s$  is large, leading to large terms in Eq. 17.33 that cancel each other in order to yield a good fit. Addition of a few new data will often change  $\hat{\beta}_s$  dramatically and thus the interpretation of single  $\hat{\beta}_{j,s}$  values as changes due to variations in the corresponding predictor becomes meaningless (attribution of response to certain causes not possible; attribution problem).

**Step 4** Perturb the MLR system (Eq. 17.36) by adding a small perturbation to the correlation matrix  $\mathbf{X}'\mathbf{X}$ :

$$\hat{\beta}_s = (\mathbf{X}'\mathbf{X} + k \mathbf{I})^{-1} \mathbf{X}' \mathbf{y} \quad (17.36)$$

will be solved for the **biasing parameter**  $0 \leq k \leq 1$ ;  $\mathbf{I}$  is the identity matrix (ones in the diagonal, all off-diagonal elements are zero; dimension: same as  $\mathbf{X}'\mathbf{X}$ , i.e.  $9 \times 9$  for the current model).

**Step 5** Converting solution values  $\hat{\beta}_{j,s}$  derived for scaled predictors and response to coefficients applicable to original data (Section L.1) yields the intercept  $\hat{\beta}_{0,o}$  and the slopes  $\hat{\beta}_{j,o}$

$$\hat{\beta}_{0,o} = \bar{Y} + S_Y \sum_k c_j = \bar{Y} + \sum_j \frac{-S_Y \hat{\beta}_{j,s} \bar{X}_j}{S_{X_j}} \quad (17.37)$$

$$\hat{\beta}_{j,o} = \frac{S_Y \hat{\beta}_{j,s}}{S_{X_j}}. \quad (17.38)$$

### 17.6.2 How do the slopes change with $k$ ? Ridge trace

The ridge regression estimates of the regression slopes for the scaled data,  $\hat{\beta}_{j,s}$ , are largely different from the MLR values (equivalent to  $k = 0$ ) already for small (say,  $k = 0.001$ ) values of the biasing parameter  $k$ , vary strongly for somewhat larger  $k$  values, before the variations level off at higher values (Fig. 17.7). The plot of the  $\hat{\beta}_{j,s}$  versus  $k$  is called a **ridge trace**. It can be used to chose a 'good' value of the biasing parameter  $k$ :  $k$  should be in the range where the  $\hat{\beta}_j$  do not vary much anymore (this is speaking for large  $k$  values), however, it should not be too large, because the coefficient of determination is decreasing with increasing  $k$  (trade-off; Fig. 17.8).

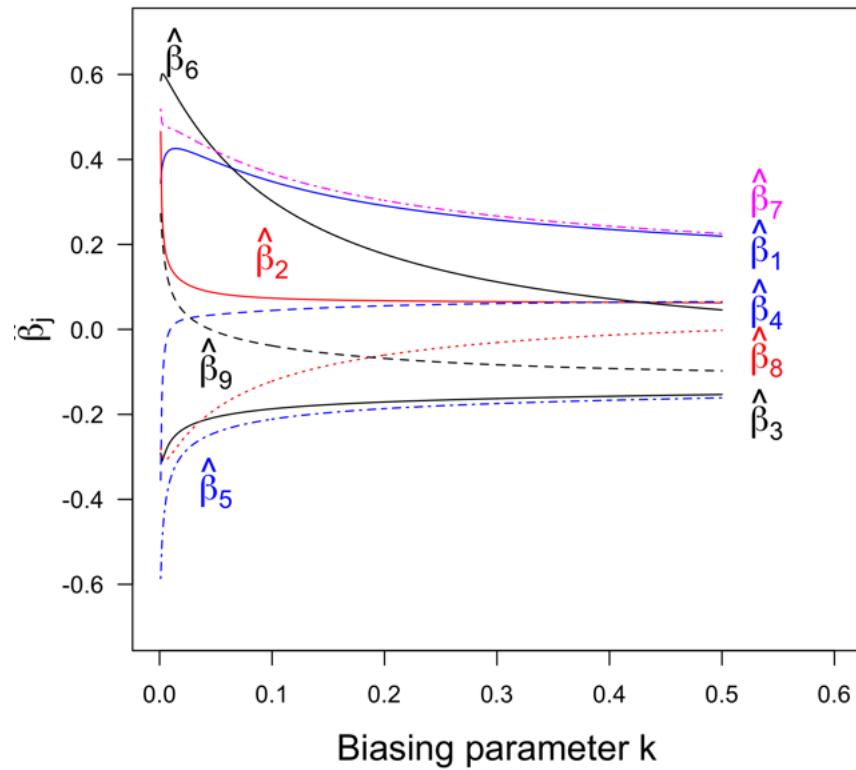


Figure 17.7: Ridge trace of the slope estimates  $\hat{\beta}_{j,s}$  for the 9-predictor model for the acetylene data.  
[AcetyleneRidgeTrace.R](#) [ridgeDWG1706s.R](#)

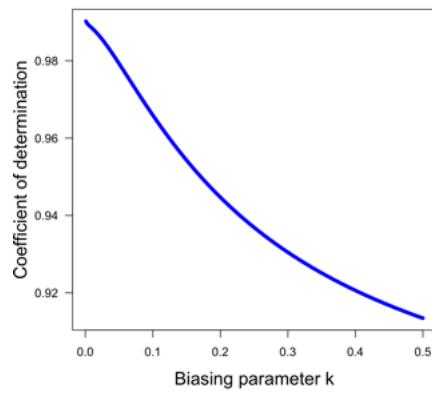


Figure 17.8: The coefficient of determination  $r^2$  for the fit of the original response versus the biasing parameter:  $r^2$  decreases with increasing biasing parameter  $k$ . [AcetyleneRidgeTrace.R](#) [ridgeDWG1706s.R](#)

### 17.6.3 Choice of $k$ , fit of predicted response data, and extrapolation

Based on the ridge trace a value of  $k = 0.01$  is chosen for the biasing parameter. Although the perturbation is relatively small, the resulting estimates for the intercept and the slopes differ dramatically from the classical MLR values (Table 17.9)

Slope number $j$	$\hat{\beta}_{j,o}$	$\hat{\beta}_{j,o}$	$\hat{\beta}_{j,s}$	$\hat{\beta}_{j,s}$
	$k = 0.01$	$k = 0$ (MLR)	$k = 0.01$	$k = 0$ (MLR)
1	0.0625	$5.32 \pm 4.88$	0.424	$36.08 \pm 33.06$
2	0.303	$19.24 \pm 4.30$	0.144	$9.16 \pm 2.05$
3	-98.87	$13770 \pm 10450$	-0.263	$36.60 \pm 27.78$
4	$3.472 \cdot 10^{-6}$	$-0.0141 \pm 0.0032$	0.00209	$-8.53 \pm 1.94$
5	-0.123	$-10.58 \pm 8.24$	-0.349	$-30.00 \pm 23.38$
6	17.42	$-21.03 \pm 9.24$	0.572	$-0.69 \pm 0.30$
7	$2.897 \cdot 10^{-5}$	$-0.00193 \pm 0.00190$	0.473	$-31.44 \pm 30.94$
8	-0.0218	$-0.00303 \pm 0.00117$	-0.296	$-0.412 \pm 0.16$
9	329.9	$-11580 \pm 7699$	0.0943	$-3.31 \pm 2.20$
0	-81.41	$-3617 \pm 3136$		

Table 17.9: Estimates of slopes  $\beta_j$  and their uncertainties for the original acetylene data,  $\hat{\beta}_{j,o}$  and for the unit length scaled data,  $\hat{\beta}_{j,s}$  for ridge regression with  $k = 0.01$  and for classical MLR (equivalent to  $k = 0$ ). For MLR, the magnitudes of  $\hat{\beta}_{1,o/s}$ ,  $\hat{\beta}_{3,o/s}$ ,  $\hat{\beta}_{5,o/s}$ , and  $\hat{\beta}_{7,o/s}$ ; their uncertainties are large and thus  $\beta_{1,o/s}$ ,  $\beta_{3,o/s}$ ,  $\beta_{5,o/s}$ , and  $\beta_{7,o/s}$  are not significantly different from zero ( $\alpha = 0.05$ , two-sided one-sample  $t$ -test). Only  $\beta_{2,o/s}$ ,  $\beta_{4,o/s}$ , and  $\beta_{8,o/s}$  are significantly different from zero. The estimated intercept  $\hat{\beta}_{0,o}$  is negative with a large magnitude; its uncertainty is large and thus  $\beta_{0,o}$  is not significantly different from zero. Note that the intercept  $\beta_{0,s}$  for the scaled data is not defined (or 'zero by definition'). **The corresponding estimates based on ridge regression are dramatically smaller in magnitude. This will have consequences for extrapolations.**

The decrease of the magnitude of the model parameters is a good sign, but not more. Now we first ask whether the fit of the response data is still good. This is the case indeed because the coefficient of determination  $r^2$  decreased only slightly for  $k = 0.01$  (Fig. 17.8) and thus it is not surprising that the predicted versus the observed response data are close to the 1-to-1 line (Fig. 17.9).

**The true litmus test for the quality of the parameter estimates is the extrapolation for reasonable predictor values (Fig. 17.10).** In contrast to MLR-based estimates, one does not see any negative response values (which would make no sense in the current context). In other words, ridge regression applied to the acetylene data seems to work out fine given reasonable values also for new combinations of predictor values (extrapolation).

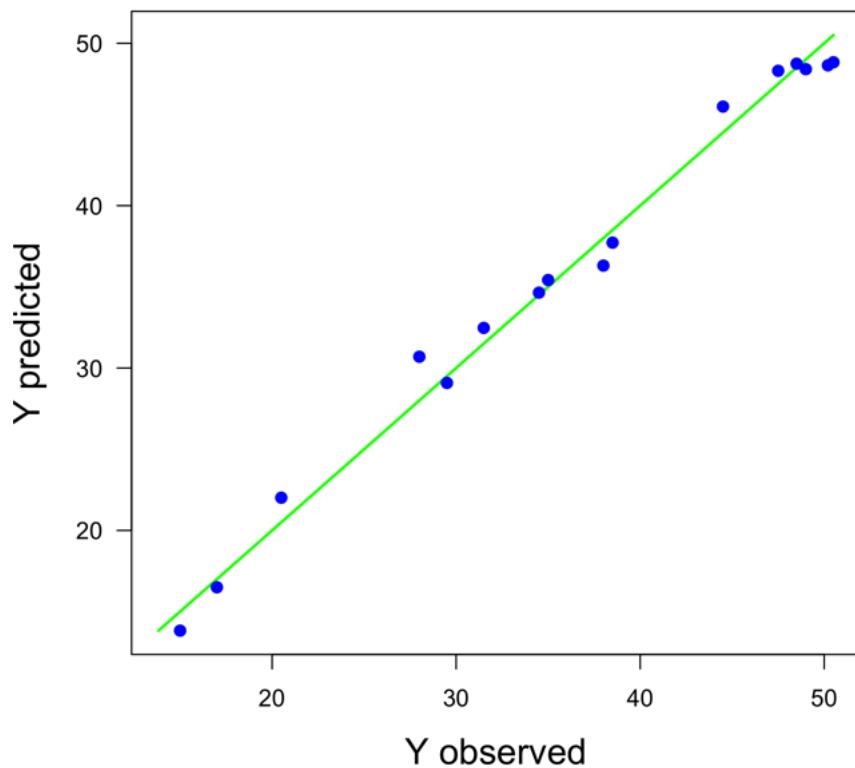


Figure 17.9: Predicted versus observed response data (blue dots) based on ridge regression estimates with  $k = 0.01$  are close to the 1-to-1 line (green solid line). [AcetyleneYpredOrig.R](#)

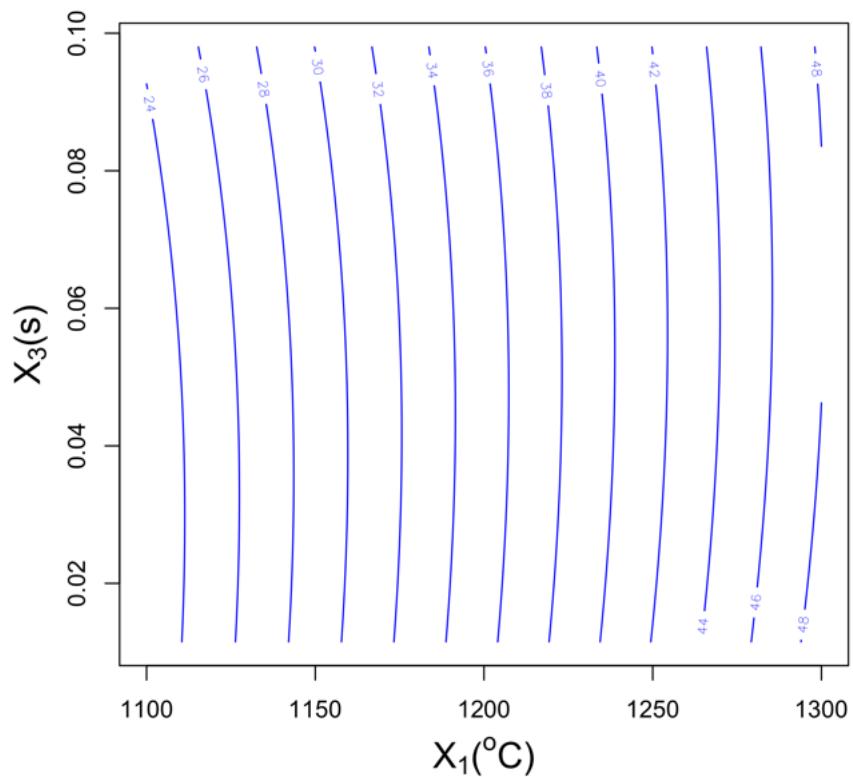


Figure 17.10: Iso-contour plot for the predicted response  $Y$  (conversion of n-heptane to acetylene) for  $X_2 = 12.4 \text{ mol mol}^{-1} \text{ H}_2$  to n-heptane ratio.  $X_1$  and  $X_3$  values vary between their observed minimal and maximal values. Based on the ridge regression estimates one does not obtain negative response values (in contrast to predictions based on MLR estimates) which would make no sense because 'negative conversion' (from acetylene to n-heptane) is not possible. [AcetyleneRidgeCon2.R](#)

### 17.6.4 Justification for ridge regression

As the acetylene example shows, ridge regression seems to work very well. Although this is not a convincing justification for ridge regression, the argument can be turned around by asking 'Why can MLR be failing?'. The reason is given in Montgomery & Peck (1982, p. 311) – and it sounds paradoxically. The estimator used in MLR for the slopes  $\beta_j$  is unbiased (usually a most desired property) and the Gauss-Markov property (Montgomery & Peck, 1982, Section 4.2.3) 'assures us that the least squares estimator has minimum variance in the class of unbiased linear estimators, **but there is no guarantee that this variance will be small**'. And in general, as the acetylene data demonstrate, the variance of the estimates can be large. Thus it seems worthwhile seeking biased estimators that possess smaller variances. Ridge regression is one answer to this question (for more mathematical details compare Montgomery & Peck, 1982).

Another justification for ridge regression comes from a Bayesian approach (Leamer, 1973, 1978; Zellner, 1971; summarized in Montgomery & Peck, 1982, p. 319-320; further discussed in Efron & Hastie, 2021). It is based on a prior for  $\beta$  consisting of a  $n$ -variate ( $n$  = number of predictors) normal distribution with mean  $\beta_0 = 0$  and covariance matrix  $V_0 = \sigma_0^2 I$  where  $I$  is the identity matrix. Based on these assumptions one obtains the linear system for ridge regression

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad (17.39)$$

where

$$k = \frac{\sigma^2}{\sigma_0^2} \quad (17.40)$$

is the usual biasing parameter. 'In effect, the method of least squares can be viewed as a Bayes estimator using an unbounded uniform prior distribution for  $\beta'$ , i.e.  $\sigma_0^2 \rightarrow \infty \Rightarrow k \rightarrow 0$ . The ridge estimator results from a prior distribution that places weak boundedness conditions on  $\beta'$ , i.e.  $\sigma_0^2$  large, but finite  $\Rightarrow k > 0$ .

### 17.6.5 Summary

*Ridge regression looks like a valid method for dealing with (multi)collinearity problems. The algorithm is simple and consists of a small perturbation of the linear system for the least squares method. The choice of an appropriate biasing parameter  $k$  can be based on the inspection of the ridge trace. Techniques for an automatic choice of  $k$  are discussed in Montgomery & Peck (1982). Ridge regression is based on a biased estimator that 'works better' (in a certain sense) than the unbiased least squares estimator. Ridge regression can be derived in the Bayesian approach.*

#### Exercise 49 Ridge regression of diabetes data

Efron & Hastie (2021, Sections 7.3 & 7.5) discuss ridge regression based on the diabetes data. The data consist of  $p = 10$  predictors – age, sex, bmi (body mass index), map (mean arterial blood pressure), and six serum measurements (tc, ldl, hdl, tch, ltg, glu) including ldl (low-density lipoprotein) and hdl (high-density lipoprotein) – and one response variable called 'prog' (disease progress) for  $n = 442$  patients.<sup>2</sup>

- (1) Calculate the correlation matrix of the predictors. Which predictor show large correlations or anti-correlation?
- (2) Calculate the coefficients of a linear model using all predictors. Discuss the the t- and p-values for the null hypotheses  $H_0$  slope  $\beta_j = 0$  for  $j = 0, 1, 2, \dots, 10$ .
- (3) Apply ridge regression using standardized predictors using (a) scaling using standard deviations (Efron & Hastie, 2021, Eq. 7.58) and (b) unit length scaling for  $\lambda = 0.1$  and compare the results with the values given in Efron & Hastie (2021, Table 7.3).
- (4) Produce a trace plot, i.e. plot all coefficients over  $\lambda$  in the range  $0 \leq \lambda \leq 0.25$ .
- (5) Which of the coefficients change sign with increasing  $\lambda$  between 0 and 0.25? What are your conclusions from this change of signs?

<sup>2</sup>Unfortunately, the variables come without units, although they could be guessed: age (yr), sex (dimensionless), bmi ( $\text{m kg}^{-2}$ ), map (mg Hg), and serum measurements ( $\text{mg dL}^{-1}$ ).

# Chapter 18

## Linear modeling including factors

In the chapters about simple (Chapter 14) and multiple linear regression (Chapter 16) we always used predictors that were numerical variables. In a sense each predictor was considered as without any relevant substructure or, in other words, 'on the same level'. In the current chapter, we consider more complex data sets with a substructure that can be described by nominal data (like, for example, 'male' or 'female' etc.). These nominal data are called 'factors' in R. Factors can be included in linear models in order to obtain better – in the sense of an information criterion like AIC – statistical models for given data sets.

### 18.1 Example 1: Parties and drinks

The first example is based on a survey on the behavior of teenies in the US (file available on GitHub: [stat100\\_2013fall\\_survey02.csv](#)). The discussion is inspired by 'Regression With Factor Variables' (<http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/RProgramming/RegressionFactors.html>)<sup>1</sup>. From the extensive data set we will use three variables only:

1. 'drinks': About how many alcoholic drinks do you consume per week on the average?
2. 'partyHr': About how many hours do you party per week on the average?
3. 'gender': When you were born, what was your biological sex?

We want to investigate the relationship between number of drinks (response) and hours of party (predictor).<sup>2</sup> We assume that the relationship between hours of party and number of drinks is (approximately) linear, however, with large scatters due to people who like to join parties, however, consume no or rarely alcohol, and extensive drinkers. We will consider the following statistical models:

1. Simple linear regression between 'drinks' and 'partyHr' not taking into account the gender (model LM0). This model yields a straight line with estimated intercept  $\hat{\beta}_0 = 0.62 \pm 0.34$  drinks, i.e. less than one drink in the limit of zero hours of party. significance level  $\alpha = 0.05$  (t-test,  $p = 0.068$ ). The estimate of the slope is  $\hat{\beta} = 1.112 \pm 0.035$  drinks per hour.
2. The variable 'gender' can take values 'Female' or 'Male'. As given in the file, 'gender' is of class 'character'. In contrast to the numeric variables 'partyHr' and 'drinks', 'gender' is a nominal variable. It can be converted to a 'factor' by calling the R routine `factor()`. In R, factors can be taken into account in various ways when fitting linear models to numerical data. In the model LMf1 we 'add' the gender factor to the predictor 'partyHr' and obtain a model with separate straight lines for 'Female' and

---

<sup>1</sup>The author(s) of 'Regression With Factor Variables' they used a larger data set, that was not available to me, and used number of drinks as predictor and hours of party as response whereas we chose the reverse.

<sup>2</sup>The choice of drinks as response and hours of party as predictor ('more drinks during longer parties') is not obvious because the other way around ('more drinks impacting the length of parties') might also make sense.

'Male' that possess the same slope, however, different intercepts:  $\hat{\beta} = 1.090 \pm 0.034$  drinks per hour,  $\hat{\beta}_{0,\text{Female}} = -0.46 \pm 0.36$  drinks,  $\hat{\beta}_{0,\text{Male}} = 3.09 \pm 0.49$  drinks. The estimated number of drinks per hour is almost identical to the estimate from simple linear regression (model LM0; 1.090 versus 1.112 drinks per hour). The estimated intercepts for females or males are quite different from each other.

3. Instead of making the intercept dependent on gender, one can vary the slope. This procedure is called an 'interaction' and coded by combining the numerical predictor (here: 'partyHr') by colon (:) with the gender factor (model LMf2). The estimate of the intercept is  $\hat{\beta}_0 = 1.06 \pm 0.34$ . The estimates of the slopes for females,  $0.912 \pm 0.044$ , and males,  $1.265 \pm 0.048$  differ by more than their estimated uncertainties.
4. Finally, one can fit a model where both intercept and slope depend on gender (LMf3; Fig. 18.1). The estimates read:  $\hat{\beta}_{0,\text{Female}} = 0.24 \pm 0.43$  drinks,  $\hat{\beta}_{0,\text{Male}} = 2.33 \pm 0.69$  drinks,  $\hat{\beta}_{\text{Female}} = 0.983 \pm 0.049$  drinks per hour,  $\hat{\beta}_{\text{Male}} = 1.188 \pm 0.068$  drinks per hour

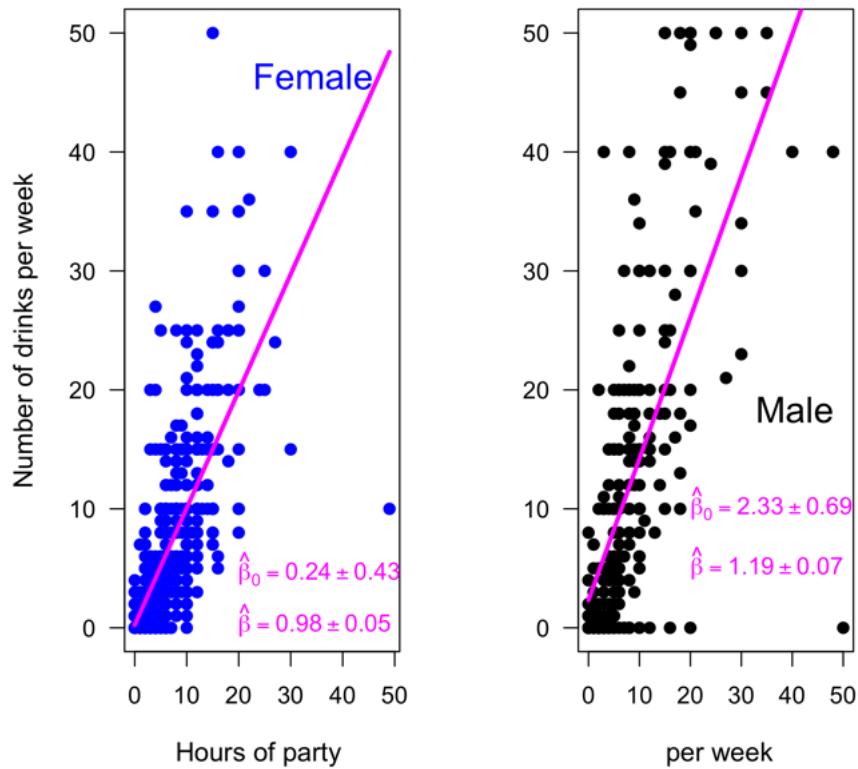


Figure 18.1: Data and model (LMf3; see text).

Which of these four models is best in the sense of the information criterion AIC<sup>3</sup> (tradeoff between model complexity as measured by number of model parameters and goodness of fit measured by logarithm of maximum likelihood)? All models including the gender factor are much better than the simple linear regression model (Table ). Models LMf1 and LMf2 perform almost equally well (difference of information criteria smaller than 1). The most complex model, LMf3, is the best because of a much better goodness-of-fit.

<sup>3</sup>The sample size  $n = 893$  is large enough to neglect the small differences between AIC and AICc (compare Section O.1).

Model	AIC
LM0	6039.86
LMf1	5990.00
LMf2	5990.14
LMf3	5982.97

Table 18.1: **AIC for Survey2 models.** All models including the gender factor perform much better – in the sense of AIC – than the simple linear regression model. Models LMf1 and LMf2 perform almost equally well (difference of AIC smaller than 1). The most complex model, LMf3, is the best because of a much better goodness-of-fit.

## 18.2 Example 2: Sleep deprivation

In the previous example the inclusion of a factor with only two levels did not very much increase the model complexity, as measured by number of model parameters. In the current example, the number of levels (18) is larger and thus increases the model complexity quite a bit. The question arises whether this increase of complexity can be 'compensated' by an improved goodness-of-fit such that Akaike's information criterion (AIC) yields smaller values. The example discussed in this section is based on the impact of several days of sleep deprivation on the reaction time of humans (Belenky et al., 2003).<sup>4</sup> The individuals who volunteered for this investigation were truck drivers (or more exactly "held valid Commercial Motor Vehicle (CMV) drivers' licenses"). The data set available in R contains reaction times of 18 individuals (subjects) over 10 days (Fig. 19.1). As expected sleep deprivation usually leads to an increase in reaction time, however, the initial reaction time (before sleep deprivation) and the amount of increase varies from individual to individual. Thus a model based on a simple linear regression of all reaction time data might not be the best option. In the following we will discuss how to take the variable 'Subject' into account as a 'factor'.

The average reaction times versus days of sleep deprivation for each of the 18 individuals are shown in Fig. 19.1. As expected, the reaction time usually increases with duration of sleep deprivation, however, the initial reaction time (before sleep deprivation) and the amount of increase varies from individual to individual.

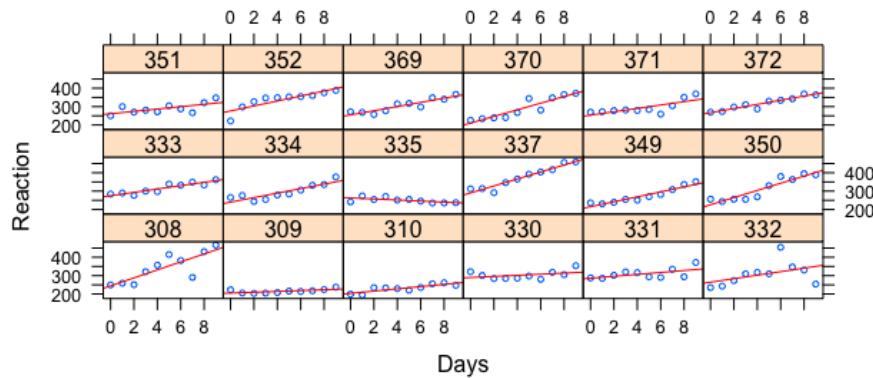


Figure 18.2: Look at the data! Average reaction time versus days of sleep deprivation by subject.

We will consider the following statistical models:

1. Simple linear regression between 'Reaction' (response) and 'Days' (predictor) not taking into account 'Subject' as factor (model LM0). This model yields a straight line with estimated intercept  $\hat{\beta}_0 = 251.4 \pm 6.6$  ms and slope  $\hat{\beta} = 10.5 \pm 1.2$  ms per day.
2. The variable 'Subject' is a factor with 18 levels. Its values '308', '309' etc. might suggest that it could be a numerical predictor, however, it is not! Each number stands for an individual as, for example, Kieran White<sup>5</sup> and could be replaced by the real name of the person. Asking `class(sleepstudy$Subject)` yields 'factor', i.e. the values have been converted already to a factor in R and don't have to be converted from character into factor as in the party example discussed above. In R, factors can be taken into account in various ways when fitting linear models to numerical data. In the model LMf1 we 'add' the 'Subject' factor to the predictor 'Days' and obtain a model with separate straight lines for each individual that possess the same slope ( $\hat{\beta} = 10.47 \pm 0.80$  ms per day), however, different intercepts (Fig. 18.3) which vary by almost a factor of 2 between 168 and 329 ms.

<sup>4</sup>If you think 'This sounds like there may be interest by the military' look at the affiliation of the authors. However, the results are of interest as well for various professions where sleep deprivation is a consequence of working conditions.

<sup>5</sup>Kieran White sang and played guitar in the British blues-rock band Steamhammer (1968-1971). Later he settled in Oregon and became a truck driver.

3. Instead of making the intercept dependent on 'Subject', one can vary the slopes. This procedure is called an 'interaction' and coded by combining the numerical predictor (here: 'Days') by colon (:) with the 'Subject' factor (model LMf2). The estimate of the intercept is  $\hat{\beta}_0 = 251.4 \pm 4.0$  ms. The estimated values of the slopes (Fig. 18.4) vary between 15.6 and 45.8 ms per day.
4. Finally, one can fit a model where both intercept and slope depend on gender (LMf3; Fig. 18.5). The intercepts vary between 203 and 290 ms and the slopes between -2.9 and 36.8 ms per day.

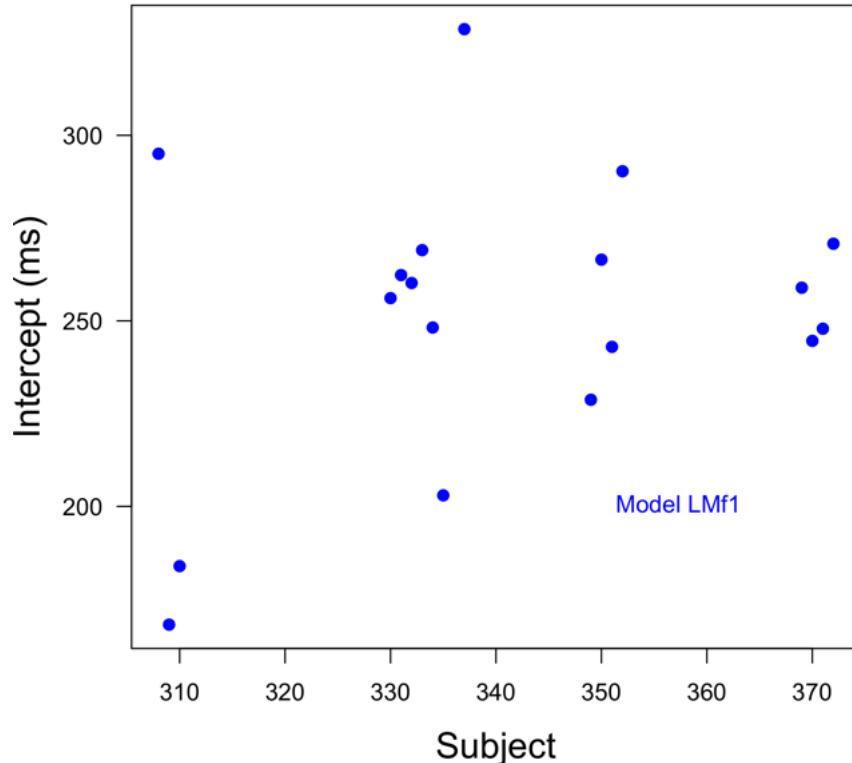


Figure 18.3: Estimated intercepts for model LMf1 (see text).

Which of these four models is best in the sense of the information criterion AICc (tradeoff between model complexity as measured by number of model parameters and goodness of fit measured by logarithm of maximum likelihood)? All models including the 'Subject' factor are much better than the simple linear regression model (Table ). According to AIC, the most complex model (LMf3) performs best (minimum AIC) In Chapter 19 we will see how to obtain a good fit while drastically reduce the number of model parameters.

#### Music

"Water (Part One)" and "Junior's Wailing" from Steamhammer (1969)

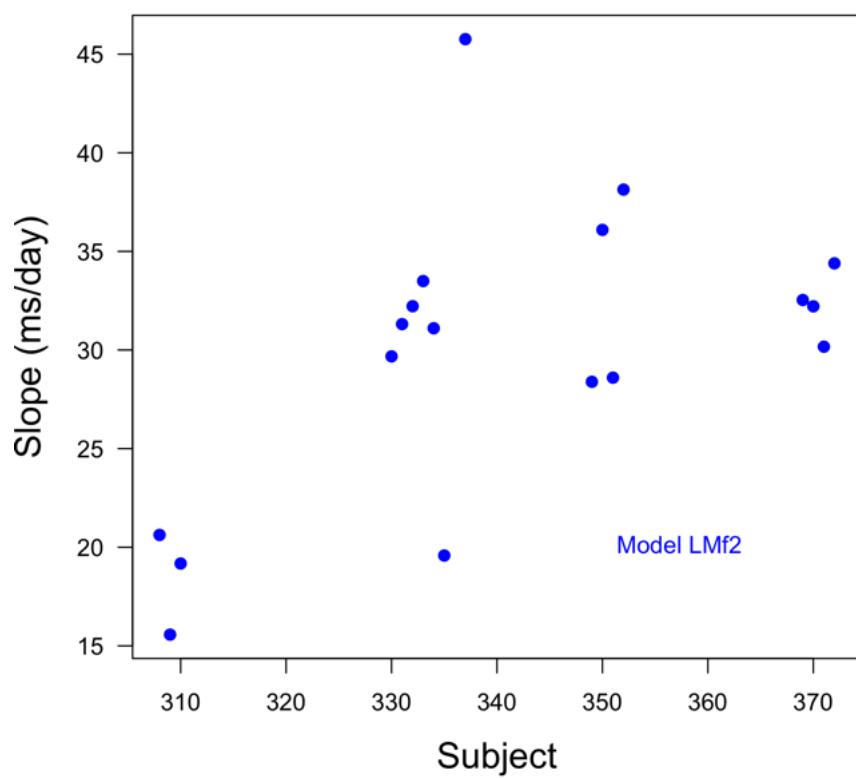


Figure 18.4: Estimated slopes for model LMf2 (see text).

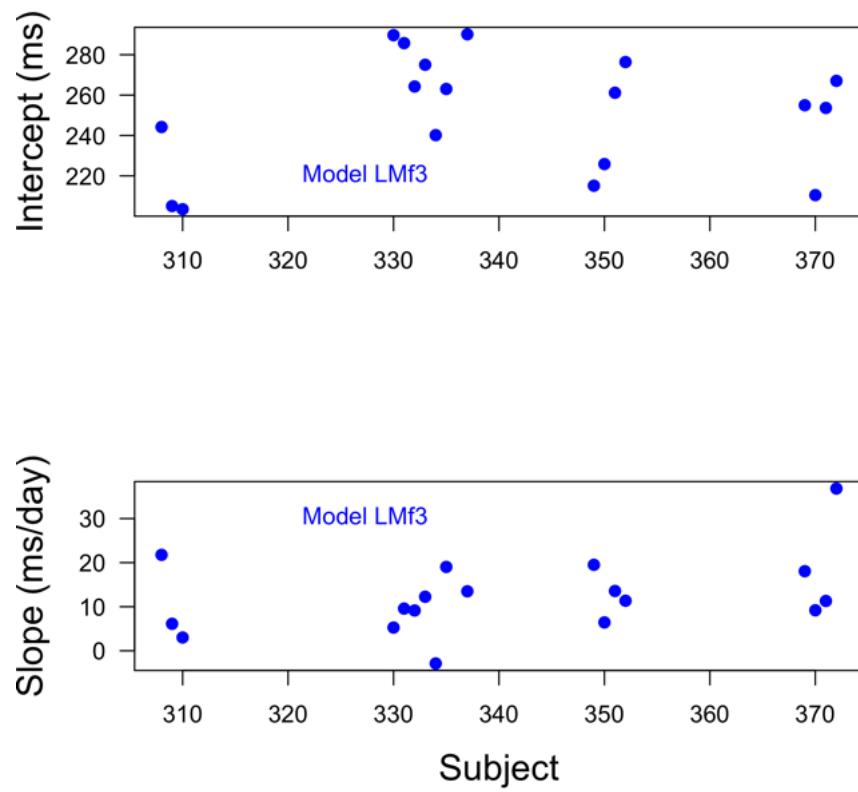


Figure 18.5: Estimated intercepts for model LMf3 (see text).

Model	AIC	logLik	k
LM0	1906.29	-950.15	3
LMf1	1766.87	-863.44	20
LMf2	1743.19	-851.59	20
LMf3	1711.87	-818.94	37

Table 18.2: **AIC, natural logarithm of the maximum likelihood (logLik), and number of model parameters (k) for sleep deprivation models.** All models including the subject factor perform much better in the sense of AIC than the simple linear regression model. Models LMf1 and LMf2 perform almost equally well (difference of AIC smaller than 1). The most complex model, LMf3, is the best because of a much better goodness-of-fit.

# Chapter 19

## Mixed effects models

So far we have fitted models to data where the true model parameters are considered as constants ('fixed') like intercepts and slopes (uni- and multivariate linear regression, generalized linear modeling). A new type of model considers (part of) model parameters as representing random effects.<sup>1</sup> These random effects are described by probability distributions (PDs) or probability density function (PDFs) and the goal is to estimate the values of parameters that characterize these PDs or PDFs. When a model takes into account fixed as well as random effects, one speaks of 'mixed effects models' or 'mixed models' for short.

### 19.1 Sleep deprivation

The example is based on the impact of several days of sleep deprivation on the reaction time of humans (Belenky et al., 2003)<sup>2</sup>. The individuals who volunteered for this investigation were truck drivers (or more exactly 'held valid Commercial Motor Vehicle (CMV) drivers' licenses'). The data set available in R contains reaction times of 18 individuals (subjects) over 10 days (Fig. 19.1). As expected sleep deprivation usually leads to an increase in reaction time, however, the initial reaction time (before sleep deprivation) and the amount of increase varies from individual to individual. The goal is to find a statistical model that describes the general tendency (average behavior; initial reaction time and daily increase = fixed effect) as well as the inter-subject variations (random effect).

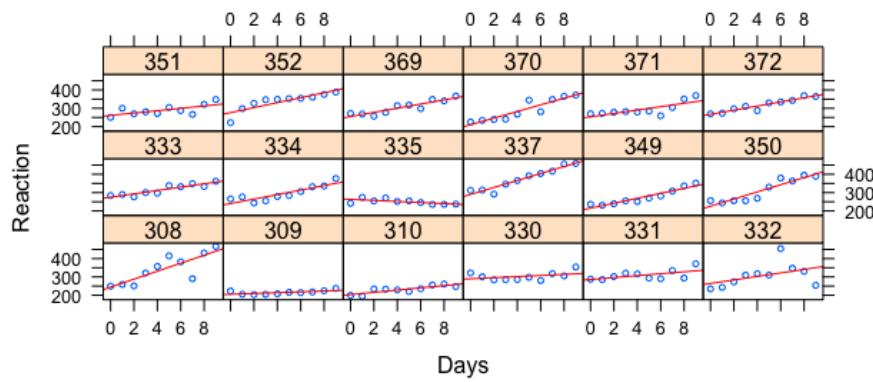


Figure 19.1: Look at the data! Average reaction time versus days of sleep deprivation by subject.  
[MixedEffectsSleep.R](#)

<sup>1</sup>In reality, the effects are usually not random, however, when we have only limited information about the processes involved, the effects appear as random to us.

<sup>2</sup>The sleep deprivation data have been discussed and modeled including the 'Subject' factor in Section 18.2.

The data (Fig. 19.1) show that reaction time usually increases with length of sleep deprivation, however, for one individual (subject 335) after a first rise the reaction time decreases over the following days. The initial (before sleep deprivation) reaction times vary from individual to individual in the range from 199 to 322 ms. The change of reaction time over time for each individual can approximately be described by straight lines, however, with different slopes. I.e. linear modeling seems appropriate.

Simple linear regression for data from each individual would lead to 18 different intercepts and slopes, i.e. to  $36 + 1 = 37$  model parameters (the additional parameter is for the standard deviation of the noise). An alternative is to set up a mixed effects model with

1. fixed effect: straight line based on all data, with a single (fixed!) intercept and slope
2. random effect: a random contribution to the overall intercept and slope that depends on the subject; these random contributions can be modeled by normal distributions with zero means and (unknown) standard deviations, i.e. one has to estimate one standard deviation (or variance) for the random contribution to the intercept and one standard deviation (or variance) for the random contribution to the slope;
3. finally, there is a noise term  $\epsilon_i$  for all data, that can also be modeled by a normal distribution with zero mean and (unknown) standard deviation.

In summary, 6 instead of 37 model parameters have to be estimated. Application of the routine **lmer()** from package **lme4** yields estimates for model parameter densities and optimal values (Fig. 19.2).

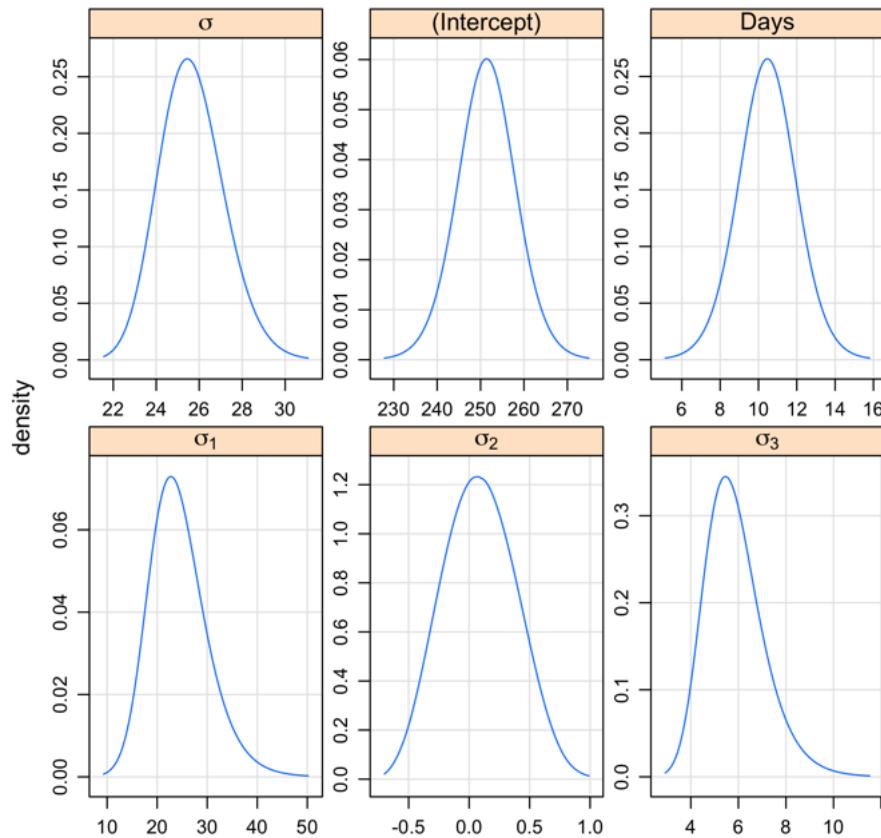


Figure 19.2: Estimated densities for standard deviation of the overall noise ( $\hat{\sigma} = 25.59$ , upper left panel), the (fixed) intercept ( $\hat{\beta}_0 = 251.4 \pm 6.8$  ms), upper middle panel), the (fixed) slope ( $\hat{\beta} = 10.5 \pm 1.5$  ms per day, upper right panel), the standard deviation for the random effect on intercept ( $\hat{\sigma}_1 = 24.7$  ms, lower left panel), the correlation between random effects on intercept and slope ( $\hat{\sigma}_2 = 0.07$ , lower middle panel), and the standard deviation for the random effect on slope ( $\hat{\sigma}_3 = 5.9$  ms per day, lower right panel). [MixedEffectsSleep.R](#)

Model	AIC	logLik	<i>k</i>
LM0	1906.29	-950.15	3
LMf1	1766.87	-863.44	20
LMf2	1743.19	-851.59	20
LMf3	1711.87	-818.94	37
MEM	1755.63	-871.81	6

Table 19.1: AIC, natural logarithm of the maximum likelihood (logLik), and number of model parameters (*k*) for sleep deprivation models. All models including the subject factor (LMf1, LMf2, LMf3, Section 18.2) perform much better – in the sense of AIC – than the simple linear regression model; The mixed effects model MEM does not perform as good as LMf2 or LMf3, however, it has much less model parameters (6 instead of 20 or 37) and thus may be easier to interpret.

*Mixed effects models are easy to formulate and the routine `lmer()` from package `lme4` yields reliable estimates of model parameters for fixed and random effects. Because of the (much) lower number of model parameters compared to models with factors, the goodness-of-fit is worse and the AIC values are usually higher than for models with factors (Table 19.1). However, the lower number of model parameters might be an advantage when interpreting model results.*

**Further reading (mixed effects models):** Pinheiro & Bates (2000), Zuur et al. (2009), Bates et al. (2014; also available as offprint: Bates et al., 2015) are the developers of the `lme4` package; they give a detailed description of the machinery behind the R routine `lmer()`. The package `nlme` (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2015) is an alternative to `lme4`.

# Chapter 20

## Least squares for non-linear models

In the derivation of the least squares method in Section J.3 we never use the assumption that the model should be linear in the model parameters or the predictor variable. In this section we will discuss an example of a non-linear – namely the Michaelis-Menten or Monod – model which is very popular in biology. This model is special in that it can be transformed into a linear model. We will exploit this special feature to discuss the impact of data transformations on noise pattern and on parameter estimation.

Michaelis & Menten (1913; written in German; translated into English by Johnson & Goody, 2011) studied the kinetic of the enzyme invertase<sup>1</sup> and described it by the now called Michaelis-Menten equation (Fig. 20.1)

$$V(S) = V_{\max} \frac{S}{S + K} \quad (20.1)$$

where  $S$  is the substrate concentration,  $V(S)$  is the (enzymatic/conversion) rate,  $V_{\max}$  is the maximum rate (at large substrate concentration), and  $K$  is the half-saturation or Michaelis-Menten constant (at  $S = K$  the rate  $V$  is half the maximum rate:  $V(K) = V_{\max}/2$ ). In the 1940ies Monod (1942, 1949) used the same equation, however, with a different interpretation,<sup>2</sup> to describe the growth of bacteria. Nowadays it is often used to describe the growth of more complex organisms (eukaryotes as, for example, microalgae) depending on certain growth-limiting nutrients.

In the following we will use the notation

$$y(x) = \alpha \frac{x}{x + \beta} + \epsilon \quad (20.2)$$

i.e. Greek letters for the (unknown) model parameters  $\alpha = V_{\max}$  and  $\beta = K$ ,  $x$  for the predictor (independent) variable (assumed to be non-stochastic),  $y$  for the response (dependent) variable, and  $\epsilon$  for the noise. We will discuss two examples. The data read:

**Example 1:**

$x = c(0.40, 1.05, 1.62, 2.07, 2.32, 2.46, 2.55, 3.21, 3.87, 4.17, 4.64, 4.64, 4.86, 4.93, 5.10, 5.13, 5.23, 5.78, 5.86, 5.87, 6.04, 6.05, 6.25, 6.70, 7.46, 8.51, 8.68, 9.33, 9.44, 9.66)$

$y = c(0.36, 1.54, 1.90, 2.54, 2.10, 2.25, 2.34, 2.56, 2.56, 2.75, 2.92, 2.84, 2.54, 2.52, 3.00, 2.86, 2.76, 3.13, 3.05, 3.03, 2.70, 2.97, 2.62, 3.13, 2.95, 3.31, 3.32, 3.02, 3.21, 3.24)$

**Example 2:**

$x = c(0.31, 0.32, 0.32, 0.35, 0.36, 0.41, 0.45, 0.48, 0.50, 0.50, 0.52, 0.52, 0.52, 0.58, 0.59, 0.59, 0.62, 0.62, 0.65, 0.65, 0.73, 0.78, 0.94, 1.19, 1.23, 1.30, 1.46, 1.87, 2.87, 7.58)$

$y = c(0.57, 0.54, 0.55, 0.55, 0.61, 0.67, 0.72, 0.77, 0.81, 0.79, 0.80, 0.81, 0.88, 0.97, 0.88, 0.92, 0.97, 0.92, 0.97, 0.97, 1.16, 1.14, 1.48, 1.46, 1.65, 1.54, 1.64, 2.22, 2.48, 3.34)$

<sup>1</sup>'... which was so named because its reaction results in the inversion of optical rotation from positive for sucrose to a net negative for the sum of fructose plus glucose.' Johnson & Goody, 2011

<sup>2</sup>i.e. growth of whole organisms instead of kinetics of a single enzyme

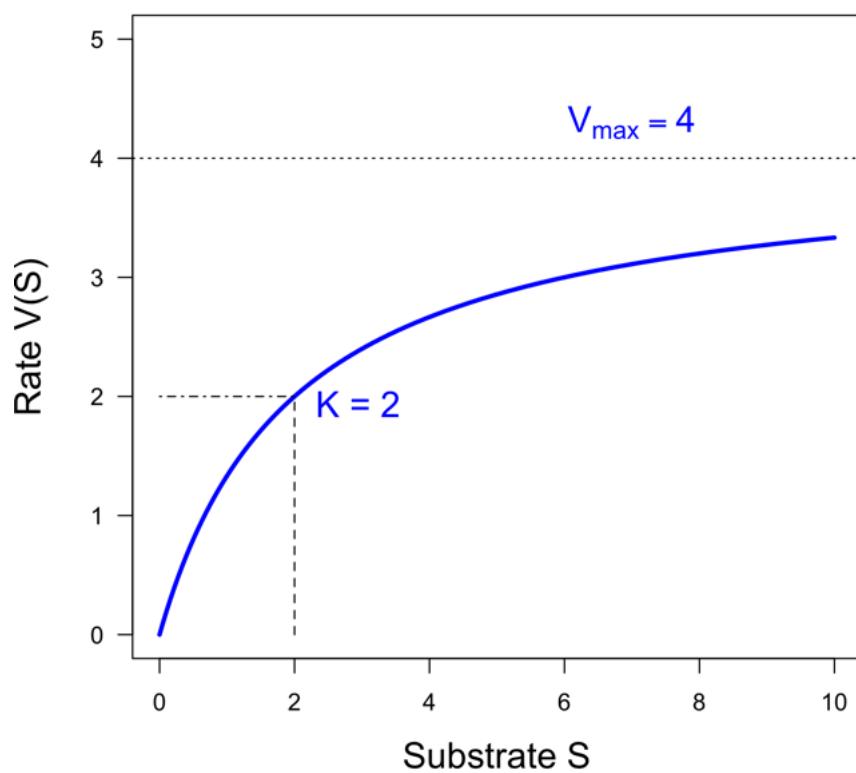


Figure 20.1: Michaelis-Menten kinetics [MMkinetics.R](#)

## 20.1 Non-linear regression/least squares

A Michaelis-Menten curve can be fitted to the data using the R routine `nls()`. In contrast to the linear case, one has to provide guess values for the model parameters  $\alpha$  and  $\beta$  because the solution (optimal values) has to be found by iteration. I used  $\alpha = 5$  and  $\beta = 3$  as guess values for  $\alpha = V_{\max}$  and  $\beta = K$ , respectively, which seem reasonable when looking at the data (Figs. 20.2 – 20.12)

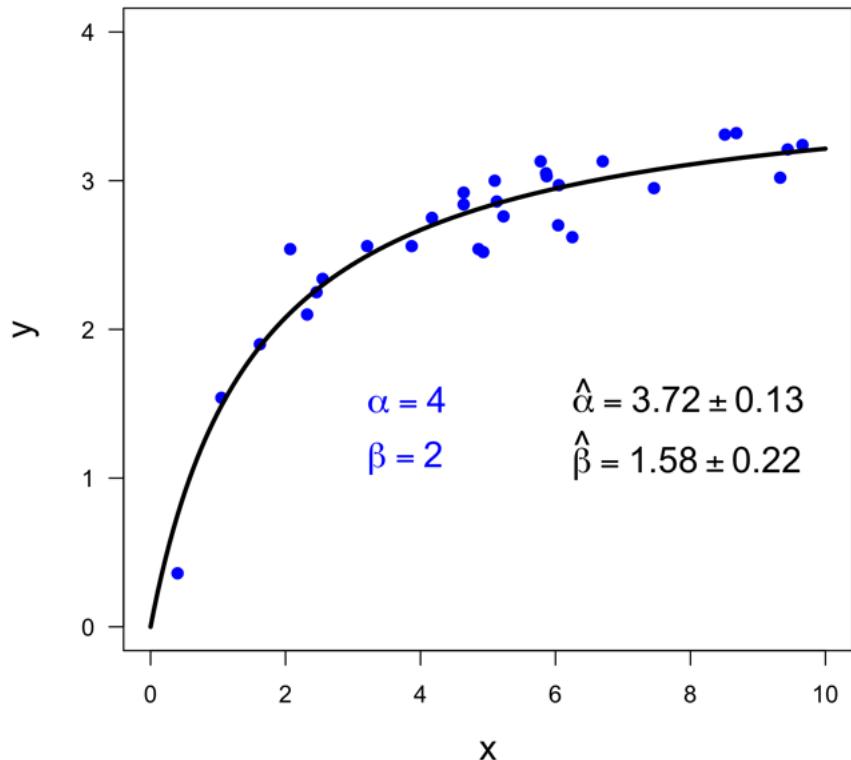


Figure 20.2: Non-linear regression of first data set to a Michaelis-Menten model. The estimates of  $V_{\max}$ ,  $\hat{\alpha} = 3.72 \pm 0.13$ , and of  $K$ ,  $\hat{\beta} = 1.58 \pm 0.22$ , are both close to the true values  $\alpha = 4$  and  $\beta = 2$  used to generate the data; the differences are about  $2\sigma$ . [MMnonlinearFit.R](#)

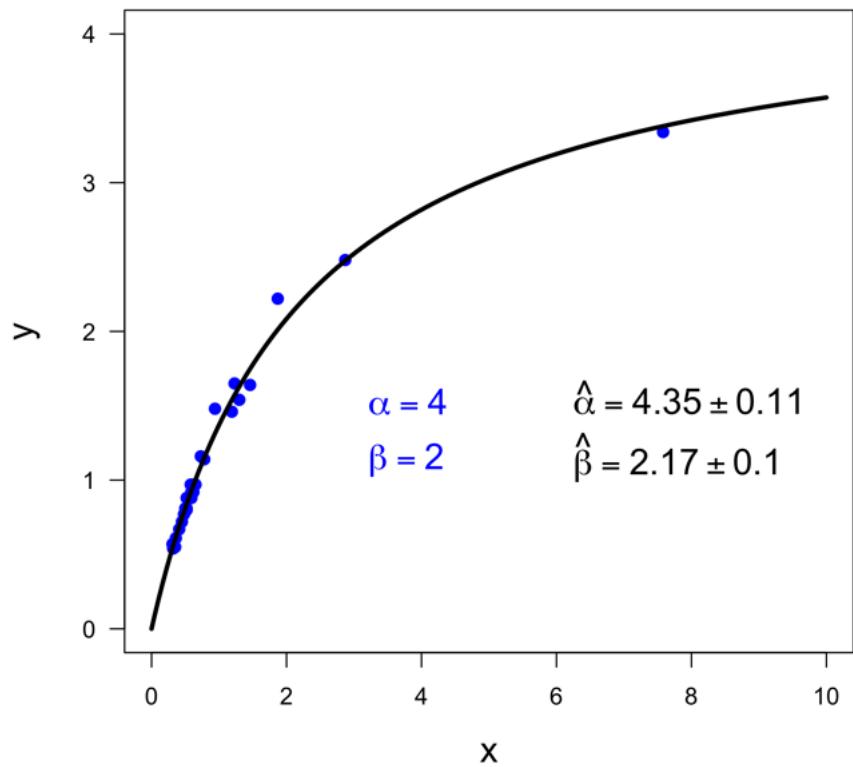


Figure 20.3: Non-linear regression of second data set to a Michaelis-Menten model. The estimates of  $V_{\max}$ ,  $\hat{\alpha} = 4.35 \pm 0.11$ , and of  $K$ ,  $\hat{\beta} = 2.17 \pm 0.10$  are both larger than the true values  $\alpha = 4$  and  $\beta = 2$  used to generate the data; the differences are 2 to 3  $\sigma$ . [MMnonlinearFit.R](#)

**Discussion:** Although both data set look ‘Michaelis-Menten-like’ the estimates of the model parameters are different from the true values by  $2\sigma$  (first data set) and up to  $3\sigma$  (second data set), respectively. Differences between true and estimated model parameters are expected to depend on noise level, distribution of the predictor variable, and sample size ( $n = 30$  for both data sets). A look at the residuals (Figs. 20.4 – 20.5) shows that the residuals for the non-linear regression of the second data set clearly violate a prerequisite of least squares, namely, the homogeneity of the noise (no pattern). This fact might explain the larger difference between true and estimated model parameters for the second data set.

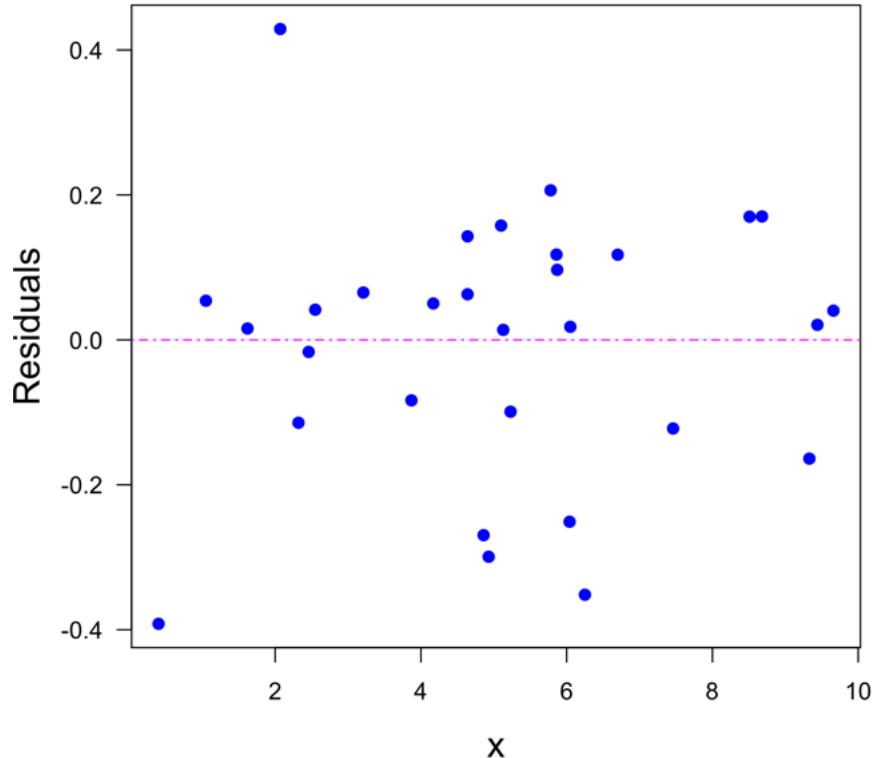


Figure 20.4: The residuals of the non-linear regression of the first data set look o.k., i.e. show no pattern and application of the Shapiro-Wilk test for normality yields  $p = 0.21$  ( $H_0$  = ‘sample from normal distribution’ not rejected on the common level of significance 0.05). [MMnonlinearFit1Res.R](#)

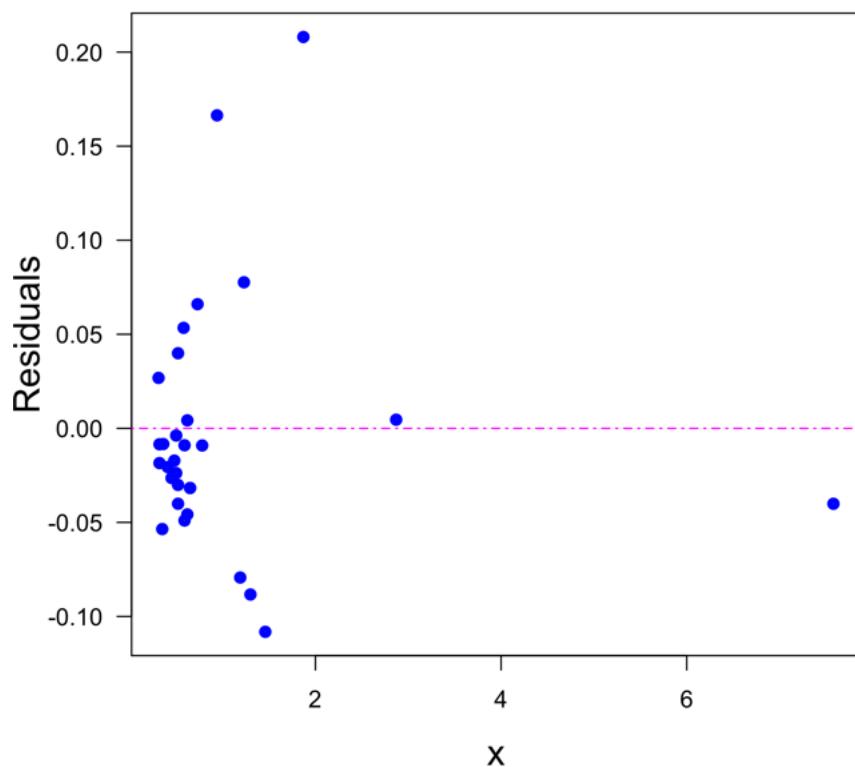


Figure 20.5: The residuals of the non-linear regression of the second data set look show a clear pattern and thus violate one of the prerequisites of least squares. Application of the Shapiro-Wilk test for normality yields  $p = 0.001$  ( $H_0 = \text{'sample from normal distribution'}$  is rejected on the level of significance 0.05).

`MMnonlinearFit2Res.R`

## 20.2 Lineweaver-Burk transformation

Lineweaver and Burk (1934) applied the transformation ( $x_{LB} = 1/x = 1/S$ ,  $y_{LB} = 1/y = 1/V$ ) to the Michealis-Menten equation and thereby obtained the linear equation

$$\frac{1}{V} = \frac{K}{V_{\max}} \frac{1}{S} + \frac{1}{V_{\max}} \quad (20.3)$$

or in different notation

$$y_{LB} = \delta x_{LB} + \gamma. \quad (20.4)$$

I.e. by applying this transformation one can estimate the parameters  $\gamma$  and  $\delta$  from simple linear regression (which was a major advantage in 1934!) and then calculate estimates of model parameters  $\alpha = V_{\max}$  and  $\beta = K$  from  $\hat{\gamma}$  and  $\hat{\delta}$ .

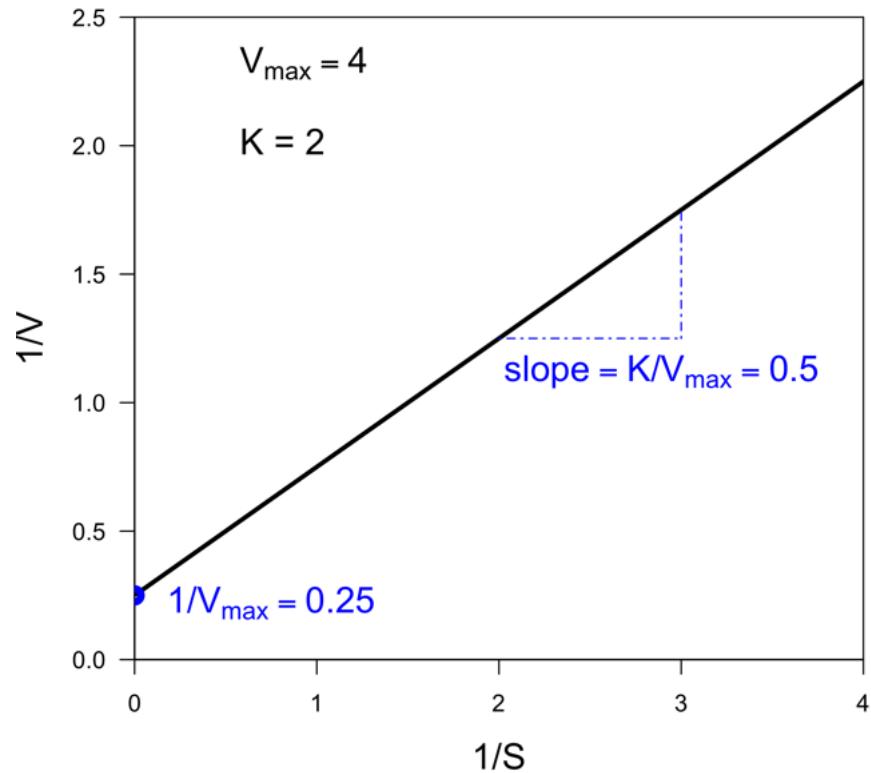


Figure 20.6: The Lineweaver and Burk (1934) transformation ( $x_{LB} = 1/S$ ,  $y_{LB} = 1/V$ ) results in a straight line with  $y$ -intercept  $1/V_{\max}$  and slope  $\frac{K}{V_{\max}}$ . MM-Lineweaver-Burk-Trans.R

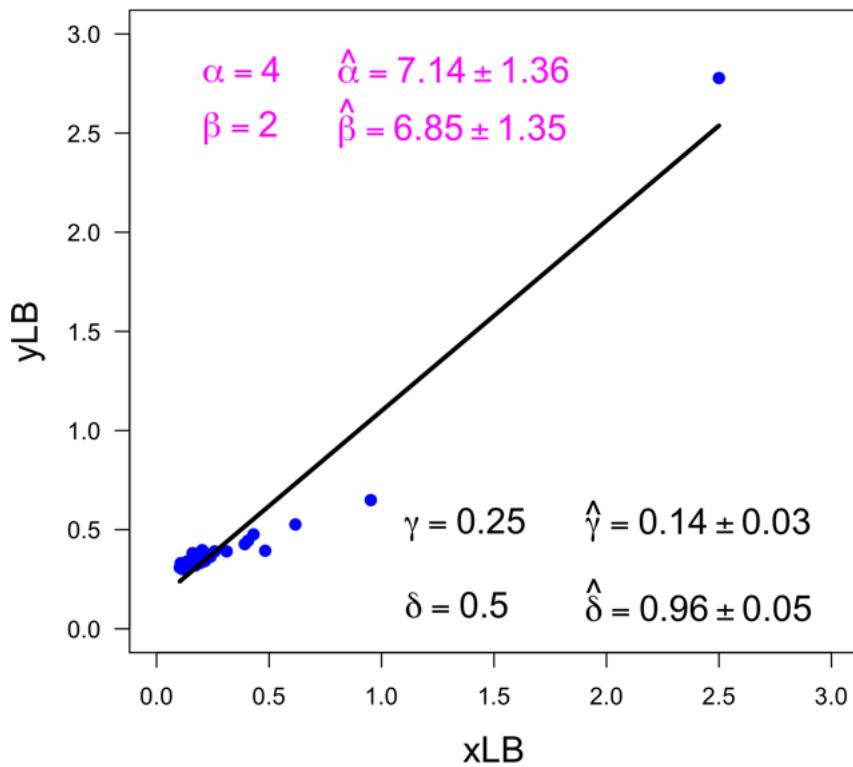


Figure 20.7: Simple linear regression of first data set after Lineweaver-Burk transformation. The straight line does not look like a good fit to the data and the residuals show a strong pattern (most data above the line for small  $x_{LB}$ , most data below the line for larger  $x_{LB}$ ; Fig. 20.9). This violation of the prerequisite of homoscedasticity (no pattern in noise) leads to estimates that are far from the true values. The estimates  $\hat{\gamma} = 0.140 \pm 0.027$  and  $\hat{\delta} = 0.959 \pm 0.049$  are both more than  $4\sigma$  away from the true values. The resulting estimates  $\hat{\alpha} = 7.14 \pm 1.36$  and  $\hat{\beta} = 6.85 \pm 1.35$  are both more than  $3\sigma$  away from the true values. [MM-LB1.R](#)

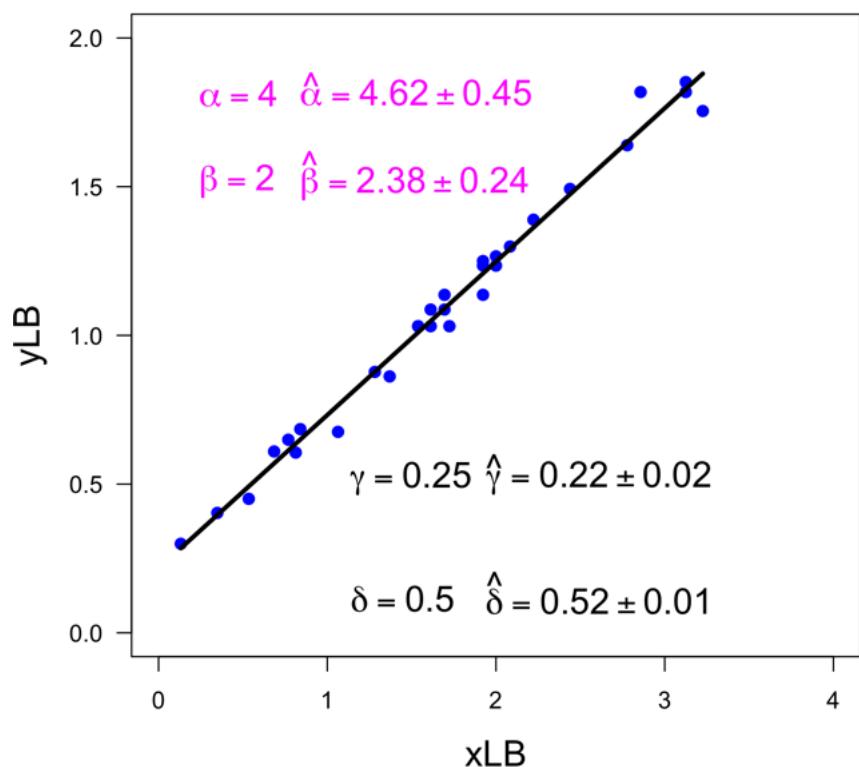


Figure 20.8: Simple linear regression of second data set after Lineweaver-Burk transformation. The estimates  $\hat{\gamma} = 0.213 \pm 0.020$  and  $\hat{\delta} = 0.518 \pm 0.010$  are less than  $1\sigma$  and  $2\sigma$ , respectively, above the true values. The resulting estimates  $\hat{\alpha} = 4.69 \pm 0.43$  and  $\hat{\beta} = 2.43 \pm 0.23$  are less than  $2\sigma$  above the true values and thus much closer than the estimates from the non-linear regression (Fig. 20.12). [MM-LB2.R](#)

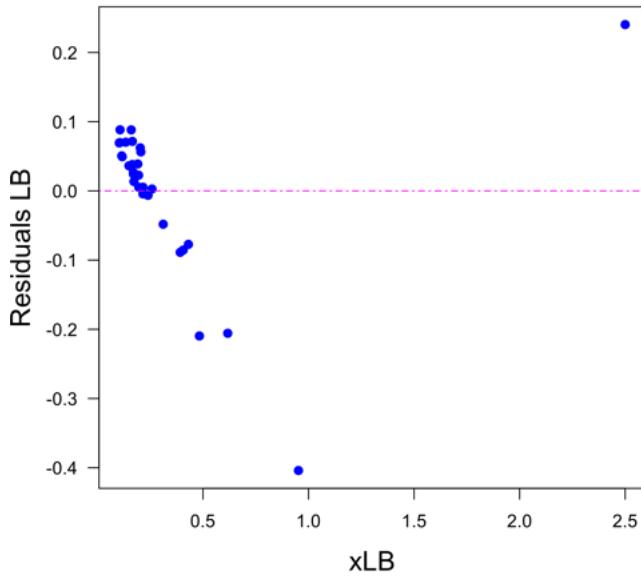


Figure 20.9: Residuals for simple linear regression of Lineweaver-Burk transformed first data set. The prerequisite of homoscedasticity (no pattern in noise) is obviously violated ( $p = 5 \cdot 10^{-8}$ ). [MM-LB1res.R](#)

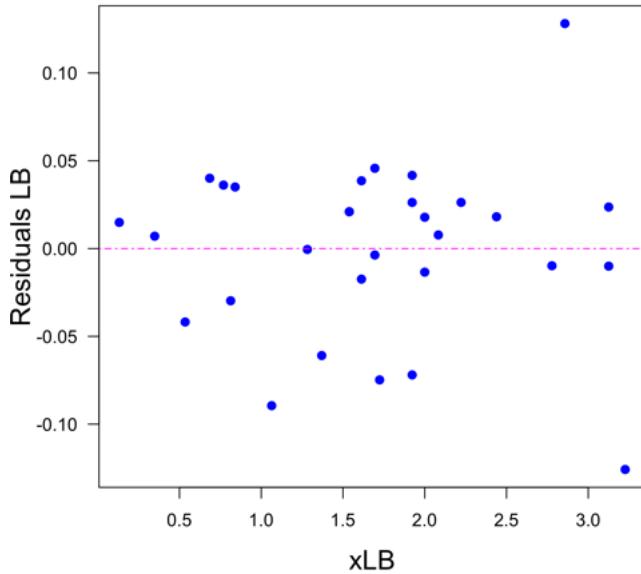


Figure 20.10: Residuals for simple linear regression of Lineweaver-Burk transformed second data set. The prerequisite of homoscedasticity (no pattern in noise) is slightly violated ( $p = 0.027$ ). [MM-LB2res.R](#)

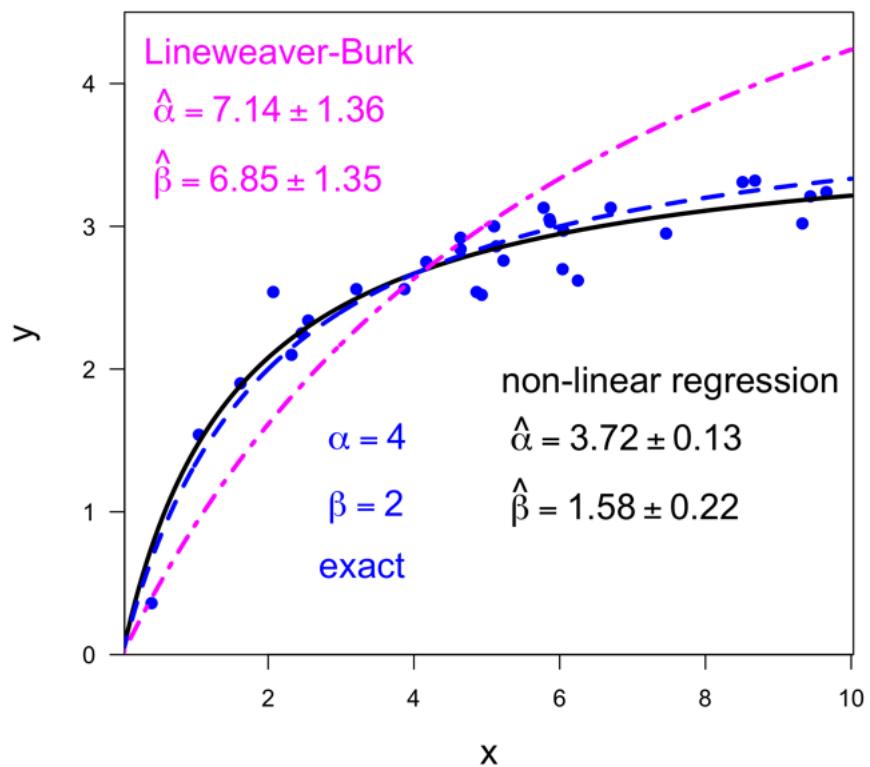


Figure 20.11: Example I: data (blue dots), exact model (blue dashed line), fit from non-linear regression (black solid line) and from Lineweaver-Burk transformation followed by simple linear regression (magenta dash-dotted line). Estimate of model parameters based on non-linear regression,  $\hat{\alpha} = 3.72 \pm 0.13$ ,  $\hat{\beta} = 1.58 \pm 0.22$ , come with relative large uncertainties and are  $2\sigma$  away from the true values; those based on Lineweaver-Burk transformation followed by simple linear regression,  $\hat{\alpha} = 7.14 \pm 1.36$ ,  $\hat{\beta} = 6.85 \pm 1.35$ , are much worse. [MM-LB1compare.R](#)

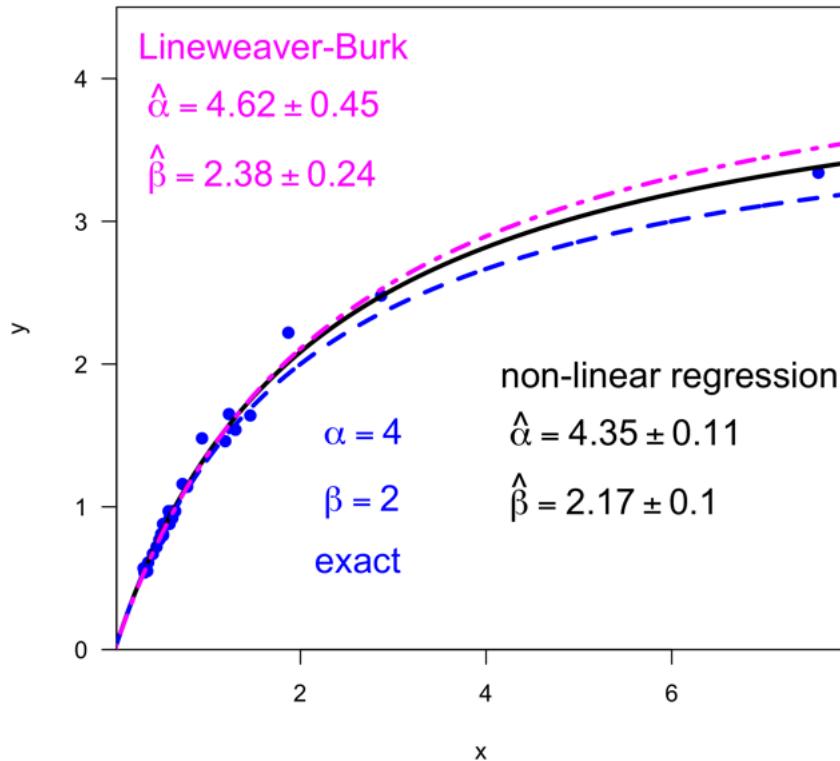


Figure 20.12: Example II: data (blue dots), exact model (blue dashed line), fit from non-linear regression (black solid line) and from Lineweaver-Burk transformation followed by simple linear regression (magenta dash-dotted line). Estimate of model parameters based on non-linear regression  $\hat{\alpha} = 4.35 \pm 0.11$ ,  $\hat{\beta} = 2.17 \pm 0.1$  come with relative large uncertainties and are 2 to 3  $\sigma$ , respectively, away from the true values; those based on Lineweaver-Burk transformation followed by simple linear regression  $\hat{\alpha} = 4.62 \pm 0.45$ ,  $\hat{\beta} = 2.38 \pm 0.24$  are a bit worse. [MM-LB2compare.R](#)

### Take-Home Message

1. Least squares can be applied also for non-linear models.
2. In contrast to the linear case (simple linear regression) start values have to be provided for the model parameters. The start values usually stem from guessing based on previous knowledge or a look at the data.
3. The example of the Michaelis-Menten equation clearly demonstrates the importance of the homoscedasticity (no pattern in noise) prerequisite.
  - The first data set was generated by adding normal noise to exact Michaelis-Menten pairs  $(x_i, y_i)$  for  $\alpha = 4$  and  $\beta = 2$ . Accordingly, the non-linear fit gives relative good estimates for the model parameters.
  - The estimates from the Lineweaver-Burk transformed first data set are worse because the homoscedastic noise becomes heteroscedastic by this transformation.
  - The second data set was generated by first adding normal noise to an exact straight line. These data were then converted into Michaelis-Menten-like data by the inverse Lineweaver-Burk transformation. The homoscedastic noise becomes heteroscedastic by this inverse transformation and thus the non-linear regression does not yield good estimates of the model parameters.
  - When the second data set is Lineweaver-Burk transformed, the homoscedasticity is regained.
4. Data are often transformed in various ways (taking, for example, the square root, the logarithm, or the inverse) in order to get rid of patterns in residuals. The choice of appropriate transformations (in combination with the choice of models) requires some experience. Examples can be found, for example, in Zuur et al. (2007, p. 39-42).

### Exercise 50 Michaelis-Menten kinetics

Another kinetic data set has been 'collected': **Example 3:**

$x = c(0.40, 1.05, 1.62, 2.07, 2.32, 2.46, 2.55, 3.21, 3.87, 4.17, 4.64, 4.64, 4.86, 4.93, 5.10, 5.13, 5.23, 5.78, 5.86, 5.87, 6.04, 6.05, 6.25, 6.70, 7.46, 8.51, 8.68, 9.33, 9.44, 9.66)$

$y = c(0.70, 1.31, 1.70, 1.62, 2.21, 2.16, 2.13, 2.32, 2.77, 2.62, 2.57, 2.72, 3.60, 3.69, 2.62, 2.92, 3.19, 2.66, 2.85, 2.88, 3.89, 3.10, 4.41, 2.97, 3.75, 3.07, 4.24, 3.55, 3.51)$

- (A) Fit a Michaelis-Menten equation to the data by (1) non-linear regression and (2) via Lineweaver-Burk transformation followed by simple linear regression.  
 (B) Discuss the estimates in the light of the residuals.

### Exercise 51 Propagation of uncertainty

The simple linear regression of Lineweaver-Burk transformed data yields estimates for the intercept  $\gamma$ , the slope  $\delta$  and their uncertainties. The parameters  $\gamma$  and  $\delta$  are related to the parameter  $\alpha$  and  $\beta$  of the Michaelis-Menten equation by  $\alpha = 1/\gamma$  and  $\beta = \delta/\gamma$ . Calculate the uncertainties of  $\alpha$  and  $\beta$  by propagation of uncertainty.



# Chapter 21

## Stochastic count data: Poisson models

Count data stem from counting<sup>1</sup> numbers and thus consist of non-negative integers  $\{0, 1, 2, 3, \dots\}$ . Examples are the detected neutrinos (Section 1.2), fatal horse kicks (Section 21.1), or the incidences during the COVID-19 pandemic (Section ...). Models of stochastic count data often involve Poisson distribution or its variants such as, for example, the zero inflated Poisson distribution (ZIP, Section C.2.2) or the zero truncated Poisson distribution (Section C.2.3), but also negative binomial distributions (Section C.2.7, Exercise 40) are used. In this chapter a few examples will be discussed where Poisson distributions are involved. Poisson regression is presented in the follow-up Chapter 22.

The Poisson distribution can be described by a single parameter, namely the mean rate  $\lambda$ . The sample mean  $\bar{x}$  is an unbiased estimator of  $\lambda$ . In the current section we will use this estimator to fit Poisson distributions to various data sets and discuss the results. The Bayesian approach to the estimate of  $\lambda$  has been discussed already in Section 1.4 yielding the estimate (Eq. 1.22)

$$\hat{\lambda} = \frac{s+1}{n} \pm \frac{\sqrt{s+1}}{n} \quad (21.1)$$

where  $n = \sum_{i=0}^n f_i$  is the sum of the number of observed frequencies  $f_i$  and  $s = \sum_{i=1}^n i \cdot f_i$ .

### 21.1 Fatal horse kicks

A classical example for Poisson statistics is the set of figures on the numbers of Prussian soldiers kicked to death by horses (Table 21.2, Barrow, 1999).

NOD $j$	0	1	2	3	4
Frequencies $f_j$	109	65	22	3	1

Table 21.1: Death due to fatal horse kicks (Barlow, 1999): NOD = number of deaths per corps per year, Frequencies = actual number of such cases (for example: in  $f_2 = 22$  corps  $j = 2$  soldiers died).

From the data one calculates the total number of death

$$s = \sum_{j=0}^4 j \cdot f_j = 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 = 122 \quad (21.2)$$

and the total number of cases

$$n = \sum_{j=0}^4 f_j = 109 + 65 + 22 + 3 + 1 = 200 \quad (21.3)$$

<sup>1</sup>'Counting' in contrast to 'ranking' which also leads to integer values.

The mean number of deaths gives an estimate of the true mean rate  $\lambda$

$$\bar{x} = \frac{\sum_{j=0}^4 j \cdot f_j}{\sum_{j=0}^4 f_j} = \frac{122}{200} = 0.61 \quad (21.4)$$

The Bayesian approach allows us to estimate  $\lambda$  and its uncertainty:

$$\hat{\lambda} = \frac{s+1}{n} \pm \frac{\sqrt{s+1}}{n} = \frac{123}{200} \pm \frac{\sqrt{123}}{200} = 0.615 \pm 0.055 \quad (21.5)$$

Using  $\hat{\lambda} = 0.615$  one obtains predictions from the Poisson distribution that compare very well with the observed number of deaths (Table 21.2, Fig. 21.1). **This close similarity can be interpreted as an indication that the variation of deaths between corps is purely random and not caused, for example, by different training of the soldiers.**

NOD $j$	0	1	2	3	4
Frequencies $f_j$	109	65	22	3	1
Poisson	108.7	66.3	20.2	4.1	0.6

Table 21.2: Death due to fatal horse kicks (Barlow, 1999): NOD = number of death per corps per year, Frequency = actual number of such cases, Poisson = Poisson prediction (= probability from Poisson distribution with  $\hat{\lambda} = 0.61$  times the total number of cases  $N_c = 200$ ).

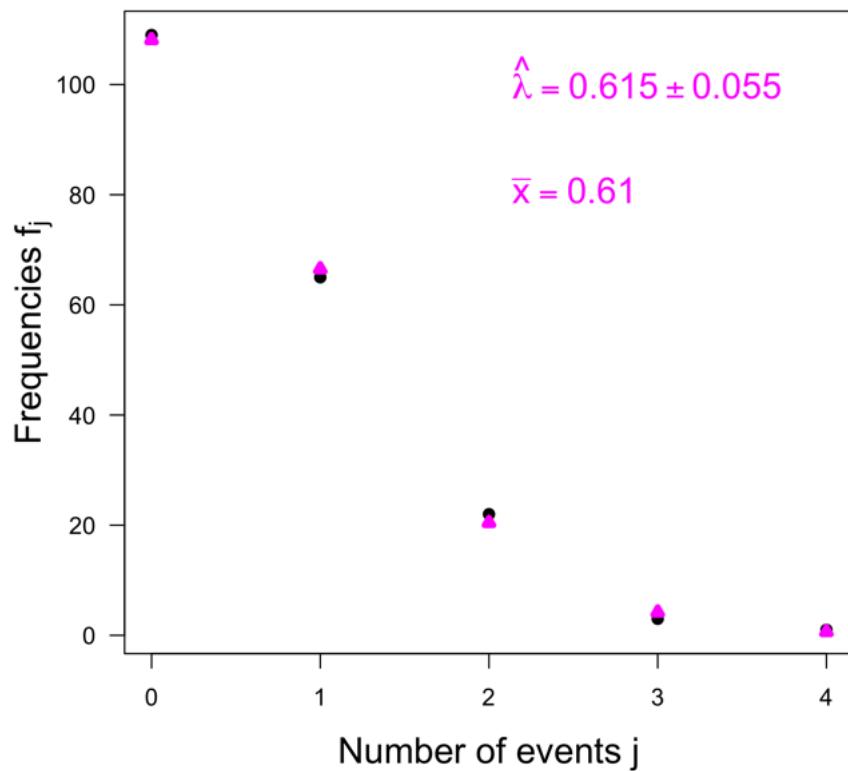


Figure 21.1: Observed (black dots) and predicted frequencies (magenta triangles) – based on a Poisson distribution with mean rate  $\hat{\lambda} = 0.615$  (Bayesian estimate) – are very close to each other. This close similarity can be interpreted as an indication that the variation of deaths between corps is purely random and not caused, for example, by different training of the soldiers. [PointEstFatalHorseKicks.R](#)

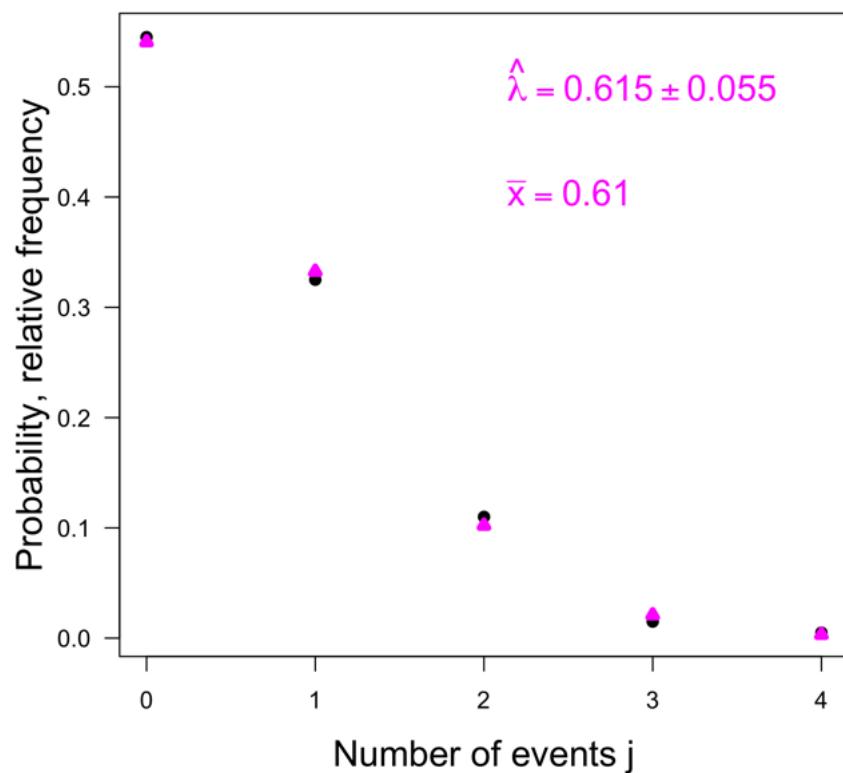


Figure 21.2: Observed relative frequencies (black dots) and predicted probabilities (magenta triangles, based on a Poisson distribution with mean rate  $\hat{\lambda} = 0.61$ ) are very close to each other. [PointEstFatalHorseKicks.R](#)

## 21.2 Estimating the R-value of the COVID-19 pandemic

An interesting parameter for the description of the development of a pandemic is the effective reproduction number  $R$  which gives the mean number of persons infected by a single infected person. If  $R > 1$ , the number of newly infected persons (incidence) is increasing, whereas for  $R < 1$  the incidences are decreasing. Usually  $R$  can not be directly derived from observations or questioning of involved persons. Thus  $R$  has to be derived from incidence data.

For the COVID-19 pandemic in Germany, the Robert-Koch-Institut (RKI) uses a very simple – and as we will see a quite robust and reliable – method:  $R$  is estimated by the ratio of the 7-day-incidences at time 0 and at 4 days before

$$\hat{R} = \frac{\text{7-Tage-Inzidenz}(t=0)}{\text{7-Tage-Inzidenz}(t=-4)} = \frac{I_0 + I_{-1} + I_{-2} + I_{-3} + I_{-4} + I_{-5} + I_{-6}}{I_{-4} + I_{-5} + I_{-6} + I_{-7} + I_{-8} + I_{-9} + I_{-10}} \quad (21.6)$$

where  $I_0$  is the present-day incidence,  $I_{-1}$  is yesterdays incidence, etcetera. The averaging over 7 days is necessary because of the variation of reported incidences with the day of the week – not reflecting the true development of the pandemic but caused by varying number of tests and delay in information transfer.<sup>2</sup> The time difference of 4 days is due to the delay between infection of person A and infection by person A; it is characteristic for the COVID-19 variants before Omicron<sup>3</sup> and is different for other variants, pandemics or epidemics.

However, this simple estimation method does not yield an estimate of the statistical uncertainty of  $\hat{R}$ . This uncertainty is of interest especially when  $\hat{R}$  is close to the critical value  $R = 1$ . How sure can we be, that the pandemic is declining when, for example,  $\hat{R}$  is 0.97?

Cori et al. (2013) proposed a Bayesian approach to estimate  $R$  and its uncertainty. Their approach will be presented here and applied to the COVID-19 pandemic in Germany. The estimation is based on the following assumptions:

1. The proliferation of the infection can be described by a **Poisson process**, i.e. the infected person A can infect  $k = 0, 1, 2, 3, 4, 5, \dots$  other persons whereby the mean number  $\lambda$  of persons infected by a single infected person is fixed, but the actual number  $k$  for person A is a random number from a Poisson distribution (Fig. 21.3)

$$p(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (21.7)$$

where  $k! = 1 \cdot 2 \cdot \dots \cdot k$  is the factorial of  $k$  and  $0! = 1$  by definition. Of course, the mean rate  $\lambda$  is not known and thus has to be estimated from data (see below).

### 2. The serial distribution

The serial interval of is defined as the time duration between a primary case-patient (infector) having symptom onset and a secondary case-patient (infectee) having symptom onset. The probability distribution for the interval length is difficult to estimate because the exact time points of symptom onsets are usually not reported, it is often not clear who infected whom, and because of other factors. Non the less, several attempts have been made to estimate serial distributions for COVID-19 (for example, Du et al., 2020, Fig. 21.4). The serial distributions for COVID-19 typically increase quite fast until a maximum at about 4 days and then fall off slowly over about 2 weeks and thus they are asymmetric with respect to the maximum at 4 days. Instead of using a serial distribution estimated from data we will later on (following Cori et al., 2013) use a ‘discrete approximation’ to a shifted gamma distribution characterized by its mean and standard deviation (Fig. 21.5). It turns out, that for large incidences the estimates of  $R$  and its uncertainty are relatively insensitive to details of the serial distribution and thus other distributions, for example based on the negative binomial distribution, would also do the job.

<sup>2</sup>German: Meldeverzug

<sup>3</sup>The faster spread of the Omicron variant indicates 2 to 3 instead of 4 days.

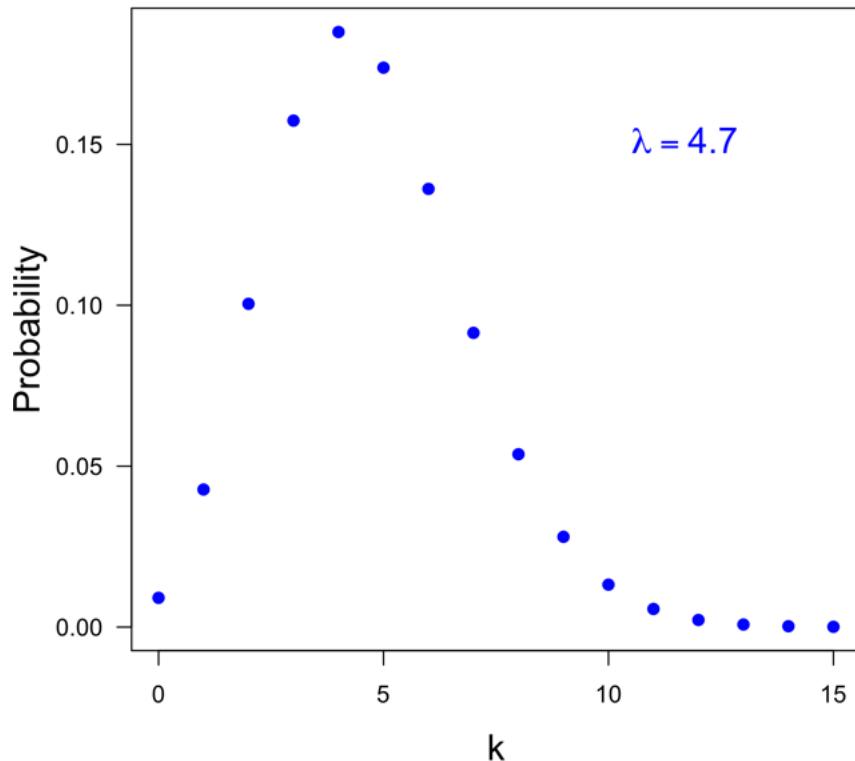


Figure 21.3: Poisson distribution for  $\lambda = 4.7$ . The probability distribution is asymmetric with respect to the location of the maximum at  $k = 4$ : steep increase for  $k < 4$  and long tail for  $k > 4$ .

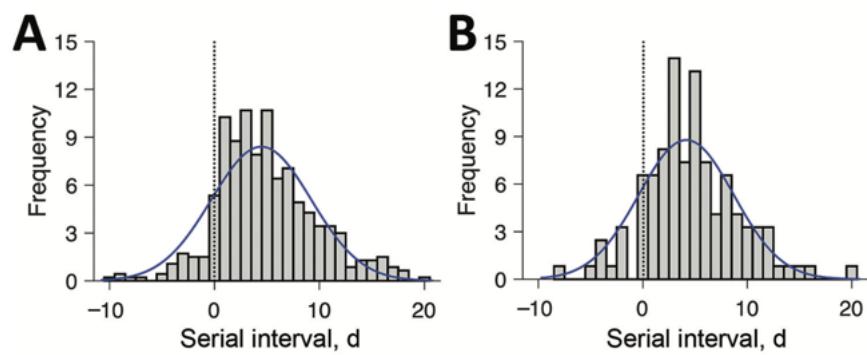


Figure 21.4: Du et al. (2020): Estimated serial interval distribution for coronavirus disease (COVID-19) based on 468 reported transmission events, China, January 21–February 8, 2020. A) All infection events ( $N = 468$ ) reported across 93 cities of mainland China as of February 8, 2020; B) the subset infection events ( $n = 122$ ) in which both the infector and infectee were infected in the reporting city (i.e., the index patient's case was not an importation from another city). Gray bars indicate the number of infection events with specified serial interval, and blue lines indicate fitted normal distributions. Negative serial intervals (left of the vertical dotted lines) suggest the possibility of COVID-19 transmission from asymptomatic or mildly symptomatic case-patients.

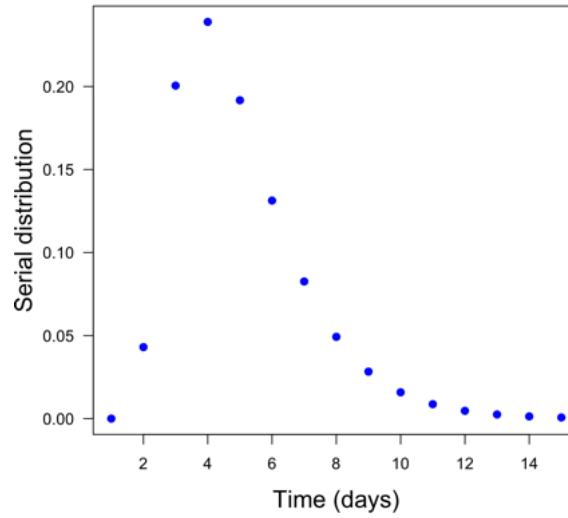


Figure 21.5: Serial distribution: a ‘discrete approximation’ to a shifted gamma distribution with mean  $\mu = 4$  and standard deviation  $\sigma = 2$ .

### 3. Likelihood distribution

The mean rate  $\lambda_{t,s}$  with which persons infected at time  $t - s$  transmit the infection to other persons is equal to the product of the reproduction rate at time  $t$ ,  $R_t$ , and the serial probability  $w_s$ :

$$\lambda_{t,s} = R_t w_s. \quad (21.8)$$

Here is a numerical example: for  $R_t = 2$  (reproduction rate at time  $t$ ) and a (serial) probability  $w_s = 0, 1$  (10%) one obtains a mean rate  $\lambda_{t,s} = 0.2$ , i.e.  $I = 100$  infected people would infect (on average)  $I\lambda_{t,s} = 20$ . Taking into account the whole serial interval one obtains

$$\lambda_t = R_t \sum_{s=1}^t I_{t-s} w_s.$$

In order to get a feeling about the order of  $\lambda_t$ , we assume for the moment that all  $I_{t-s}$  are equal to  $I$ . Based on this assumption one obtains

$$\lambda_t = R_t I \sum_{s=1}^t w_s = 2 \cdot 100 \cdot 1 = 200,$$

i.e. 100 infectors infect another 200 persons (infectees), which of course fit with an R-value of 2. The likelihood for  $I_t$  is given by the following Poisson distribution

$$P(I_t | I_0, \dots, I_{t-1}, \mathbf{w}, R_t) = \frac{\lambda_t^{I_t} e^{-\lambda_t}}{I_t!} = \frac{(R_t \Lambda_t)^{I_t} e^{-R_t \Lambda_t}}{I_t!}. \quad (21.9)$$

In order to iron out the incidence variations, one calculates the likelihood for the incidences of  $\tau = 7$  consecutive days, whereby it is assumed that the R-value does not vary much over this period and thus can be described by a mean value  $R_{t,\tau}$ . Independence of data leads to the product

$$L(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \mathbf{w}, \mathbf{R}_{t,\tau}) = \prod_{s=t-\tau+1}^t \frac{(\mathbf{R}_{t,\tau} \Lambda_s)^{I_s} e^{-\mathbf{R}_{t,\tau} \Lambda_s}}{I_s!} \quad (21.10)$$

4. The **likelihood function LF** results from a switch of perspective: in the argument list of the likelihood, the **model parameter  $R_{t,\tau}$**  is interchanged with the data (observed incidences):

$$\text{LF}(R_{t,\tau} | I_0, \dots, I_t, w) = \prod_{s=t-\tau+1}^t \frac{(R_{t,\tau} \Lambda_s)^{I_s} e^{-R_{t,\tau} \Lambda_s}}{I_s!} \quad (21.11)$$

5. As prior for  $R_{t,\tau}$  the gamma distribution

$$\mathcal{U}\mathcal{G}(R_{t,\tau}; \alpha, \theta) = \frac{R_{t,\tau}^{\alpha-1} e^{-R_{t,\tau}/\theta}}{\theta^\alpha \Gamma(\alpha)} \quad (21.12)$$

with  $\alpha = 1$  and  $\theta = 5$  is applied; this distribution has a mean  $\mu = 5$  and standard deviation  $\sigma = 5$  (Fig. 21.6). This prior is quite diffuse with high values in the most interesting range (around  $R_{t,\tau} = 1$ ) and low values for very large  $R_{t,\tau}$ . For incidences above 100 the influence of the prior is negligible.

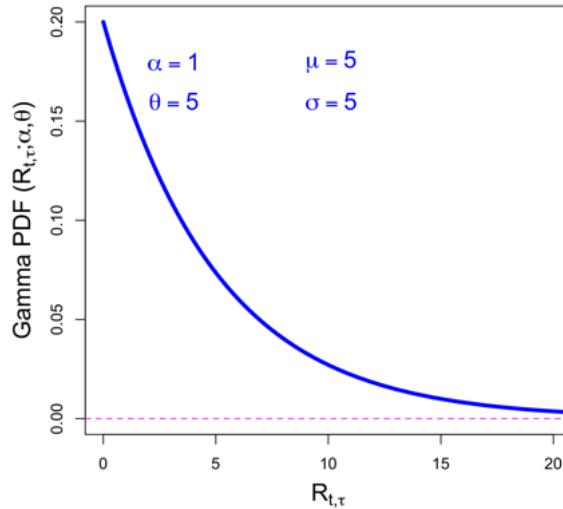


Figure 21.6: Prior for  $R_{t,\tau}$ : gamma distribution with mean  $\mu = 5$  and standard deviation  $\sigma = 5$ .

6. The **posterior** is proportional to the likelihood function times the prior:

$$\begin{aligned} & \text{Post}(R_{t,\tau} | I_0, \dots, I_t, w) \propto \text{LF}(R_{t,\tau} | I_0, \dots, I_t, w) \mathcal{U}\mathcal{G}(R_{t,\tau}; \alpha, \theta) \\ &= \prod_{s=t-\tau+1}^t \frac{(R_{t,\tau} \Lambda_s)^{I_s} e^{-R_{t,\tau} \Lambda_s}}{I_s!} \frac{R_{t,\tau}^{\alpha-1} e^{-R_{t,\tau}/\theta}}{\theta^\alpha \Gamma(\alpha)} \\ &= R_{t,\tau}^{\alpha-1 + \sum_{s=t-\tau+1}^t I_s} e^{-R_{t,\tau} \left( \frac{1}{\theta} + \sum_{s=t-\tau+1}^t \Lambda_s \right)} \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s!} \frac{1}{\theta^\alpha \Gamma(\alpha)} \\ &\propto R_{t,\tau}^{\alpha-1 + \sum_{s=t-\tau+1}^t I_s} e^{-R_{t,\tau} \left( \frac{1}{\theta} + \sum_{s=t-\tau+1}^t \Lambda_s \right)} \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s!}. \end{aligned} \quad (21.13)$$

where we dropped the factor  $1 / (\theta^\alpha \Gamma(\alpha))$  because it's independent of  $R_{t,\tau}$ . The expression (21.13) shows

that the posterior (normalized!) is a gamma distribution

$$\text{Post}(R_{t,\tau}; \alpha_{\text{Post}}, \theta_{\text{Post}}) = \frac{R_{t,\tau}^{\alpha_{\text{Post}}-1} e^{-R_{t,\tau}/\theta_{\text{Post}}}}{\theta_{\text{Post}}^{\alpha_{\text{Post}}} \Gamma(\alpha_{\text{Post}})} \quad (21.14)$$

with parameters

$$\alpha_{\text{Post}} = \alpha + \sum_{s=t-\tau+1}^t I_s \quad (21.15)$$

and

$$\theta_{\text{Post}} = \left( \frac{1}{\theta} + \sum_{s=t-\tau+1}^t \Lambda_s \right)^{-1}. \quad (21.16)$$

This result allows estimating the R-value and its uncertainty by the mean

$$\hat{R} = \alpha_{\text{Post}} \theta_{\text{Post}} = \frac{\alpha + \sum_{s=t-\tau+1}^t I_s}{\frac{1}{\theta} + \sum_{s=t-\tau+1}^t \Lambda_s} \quad (21.17)$$

and the standard deviation

$$\hat{\sigma}_R = \theta_{\text{Post}} \sqrt{\alpha_{\text{Post}}} = \frac{\sqrt{\alpha + \sum_{s=t-\tau+1}^t I_s}}{\frac{1}{\theta} + \sum_{s=t-\tau+1}^t \Lambda_s}. \quad (21.18)$$

For incidences  $I > 100$ , the impact of the chosen prior on  $\hat{R}$  and  $\hat{\sigma}_R$  is negligible and the standard deviation  $\hat{\sigma}_R$  is approximately equal to  $1/\sqrt{n}$  where  $n$  is the sum over the incidences of 7 consecutive days. For incidences  $I = 1500$ , the uncertainty is already smaller than 0.01.

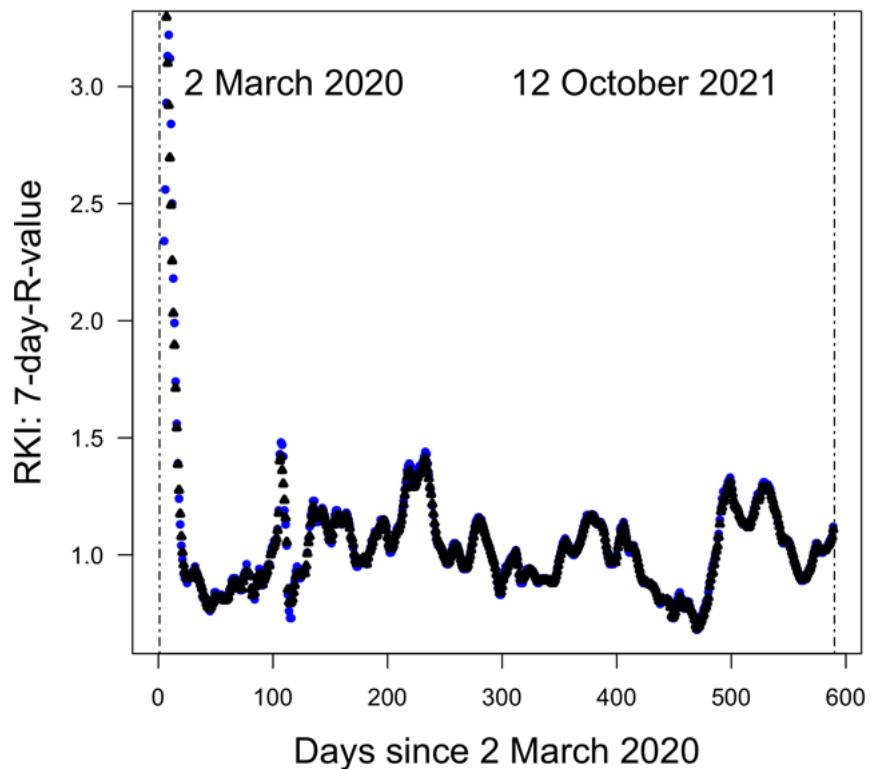


Figure 21.7: 7-day-R-values estimated by RKI (blue dots) and Bayesian estimates (black triangles; based on the method developed by Cori et al., 2013, calculated using the **R** package **EpiEstim**) compare very well with few exceptions especially at the beginning of the pandemic when the incidences were very small.

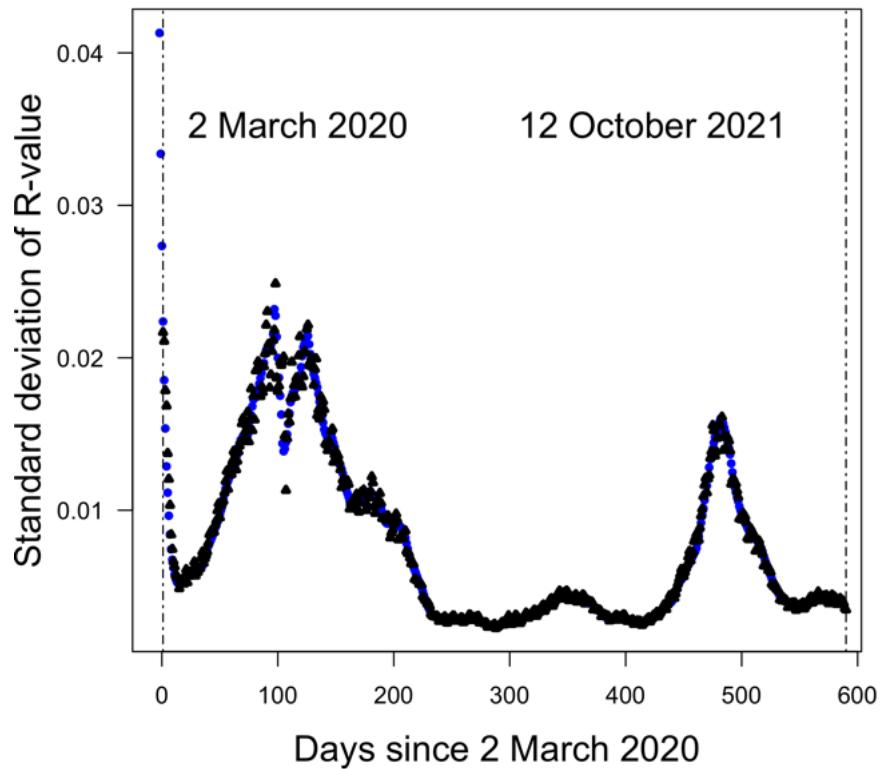


Figure 21.8: Estimates of the uncertainty based on the method developed by Cori et al., 2013, calculated using the R package **EpiEstim** (blue points) and by the  $1/\sqrt{n}$  approximation valid for large incidences (black triangles). For incidences  $I = 1500$ , the uncertainty is already smaller than 0.01.

**Exercise 52 Aitkin dust counter**

Data from an Aitkin dust counter are listed in Table 21.3. Do the data follow (approximately) a Poisson distribution?

$j$	0	1	2	3	4	5	6	7	8
Frequencies $f_j$	23	56	88	95	73	40	17	5	3

Table 21.3: Data from an Aitkin dust counter: frequencies  $f_j$  for number of particles  $j$ .

**Exercise 53 German V2 bombs on London during World War II**

Michael Lewis (2017, *The Undoing Project*) writes on p. 183: 'Londoners in the Second World War thought that German bombs where targeted, because some parts of the city were hit repeatedly while others were not hit at all.' Is this true? Clarke (1946) selected an area 'comprising 144 square kilometres of south London over which the basic probability function of the distribution [of bomb hits] was very nearly constant ... The selected area was divided into 576 squares of 1/4 square kilometre each, and a count was made of the numbers of squares containing 0, 1, 2, 3, ..., etc. flying bombs. Over the period considered the total number of bombs within the area involved was 537.' The data are given in Table 21.4. Do the data (approximately) follow a Poisson distribution?

Hint: First answer the question: How many bombs hit the square with  $\geq 5$  bombs?

No. of flying bombs per square	Actual no. of squares
0	229
1	211
2	93
3	35
4	7
$\geq 5$	1

Table 21.4: Data by Clarke (1946)

Thomas Pynchon (1973):

"Couldn't there be an equation for us too, something to help us find a safer place?

'Why am I surrounded,' his usual understanding self today, 'by statistical illiterates? There's no way, love, not as long as the mean density of strikes is constant. Pointsman doesn't even understand that.'

The rockets are distributing about London just as Poisson's equation in the textbooks predicts. As the data keep coming in, Roger looks more and more like a prophet. Psi Section people stare after him in the hallways. It's not precognition, he wants to make an announcement in the cafeteria or something ... have I ever pretended to be anything I'm not? all I'm doing is plugging numbers into a well-known equation, you can look it up in the book and do it yourself ...

His little bureau is dominated now by a glimmering map, a window into another landscape than winter Sussex, written names and spidering streets, an ink ghost of London, ruled off into 576 squares, a quarter square kilometer each. Rocket strikes are represented by red circles. The Poisson equation will tell, for a number of total hits arbitrarily chosen, how many squares will get none, how many one, two, three, and so on."

Pynchon: *Gravity's Rainbow* (1973; cited after: 1975, p. 54-55)

## Chapter 22

# Generalized Linear Modeling (GLM)

In Generalized Linear Modeling (GLM) one does regression not to the mean  $\mu$  but to a function of the mean, i.e. to  $f(\mu)$ , where  $f()$  is called the link function. The regressed values are then (like in linear regression) fitted to a straight line (or a linear expression if there are more than one predictor variables).

**Further reading (GLM):** Generalized Linear Models (GLMs) were first proposed by Nelder and Wedderburn (1972). McCullagh and Nelder (1989); Dobson & Barnett (2008); Aitkin, Francis, & Hinde (2005); Firth (1991); Guisan et al. (2002); Zuur et al. (2009); Hilbe (2011).

GLMs consist of three components (compare Box 9): (1) specification of the population of the response variable, (2) a linear predictor, and (3) a link function. Specific models are often named after (1) and (3) is not reflected in the name, although, for example, a Poisson regression always comes with the logarithm as link function.

### 9: Generalized Linear Model (GLM)

A generalized linear model (GLM) consists of three components:

1. Specification of the population (probability distribution or probability density function) from which the depending variable  $Y$  (response variable) is assumed to come from.
2. A linear predictor:

$$\eta_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_N X_{1,N} \quad (22.1)$$

where the  $X_k, k = 1, 2, \dots, N$  are explanatory (independent) variables, and  $\beta_k, k = 0, 1, 2, \dots, N$  are  $N+1$  model parameters that have to be estimated from the data.

3. A link function  $g()$  which transforms the expectation (= mean value) of the response variable,  $\mu_i = E(Y_i)$ , to the linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_N X_{1,N} \quad (22.2)$$

$g()$  should be invertible, i.e. we can write  $\mu_i = g^{-1}(\eta_i)$ , and 'thus, the GLM may be thought of as a linear model for a transformation of the expected response ...'.

Example:  $\eta = \beta_0 + \beta X = g(\mu)$ ,  $g^{-1}(\eta) = \exp(\eta) = \exp(\beta_0 + \beta X)$ ,  $\eta = \log \mu$ , where  $\log$  is the natural logarithm.

## 22.1 Poisson regression

In simple linear regression one assumes that the observations consist of values from an exact relationship (for example a straight line) plus additive normally distributed noise. These assumptions obviously do not fit when the observations are discrete entities (for example, count data or non-negative integers for species richness). Nonetheless one can 'regress to the mean' in a certain sense. In the current section, it is assumed that the response variable  $Y$  follows Poisson distributions and that the logarithm of the mean values of  $Y$  can be modeled by a linear relationship.

Three ingredients are required for the definition of a GLM: (1) the population of response variable, (2) the linear predictor, and (3) the link function. In Poisson regression one assumes that the data are from Poisson distributions with mean (average number of events)  $\mu(x)$  and that  $\log(\mu(x))$  is a straight line, i.e.

$$\log(\mu(x)) = \beta_0 + \beta x \quad (22.3)$$

where  $\log()$  is the link function and  $\eta(x) = \beta_0 + \beta x$  is the linear predictor. The inverse of the link function  $\log()$  is the exponential function  $\exp()$  and thus the model for  $\mu(x)$

$$\mu = \exp(\beta_0 + \beta x) \quad (22.4)$$

is a non-linear function of the model parameters. The goal of Poisson regression is to estimate the two model parameters  $\beta_0$  and  $\beta$  from the data. This can be done by calling the R routine **glm()**.

### 22.1.1 Poisson regression: an example

Apply a Poisson regression to the artificial data<sup>1</sup> (Fig. 22.1)

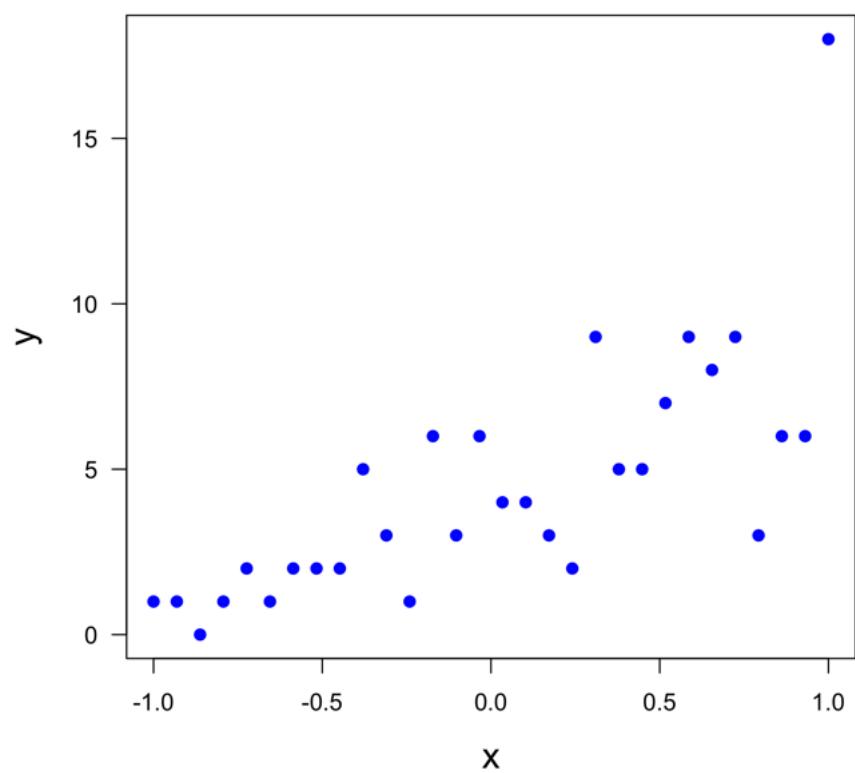
$$\begin{aligned} x &= \{-1.000, -0.931, -0.862, -0.793, -0.724, -0.655, -0.586, -0.517, -0.448, -0.379 \\ &\quad -0.310, -0.241, -0.172, -0.103, -0.034, 0.034, 0.103, 0.172, 0.241, 0.310, \\ &\quad 0.379, 0.448, 0.517, 0.586, 0.655, 0.724, 0.793, 0.862, 0.931, 1.000\} \end{aligned} \quad (22.5)$$

$$y = \{1, 1, 0, 1, 2, 1, 2, 2, 2, 5, 3, 1, 6, 3, 6, 4, 4, 3, 2, 9, 5, 5, 7, 9, 8, 9, 3, 6, 6, 18\} \quad (22.6)$$

The application of **glm()** to the data is described in the R code listed below. Application of the Poisson regression yields the estimates  $\hat{\beta}_0 = 1.297 \pm 0.104$  and  $\hat{\beta} = 1.08 \pm 0.16$  which encompass the exact population values  $\beta_0 = 1.3$  and  $\beta = 1.2$  in the  $\pm 1 \sigma$  uncertainty ranges.

---

<sup>1</sup>The artificial data  $y$  is a random sample from Poisson distributions with mean rate  $\mu = \exp(\beta_0 + \beta x)$ ,  $\hat{\beta}_0 = 1.3$  and  $\beta = 1.2$ .

Figure 22.1: Artificial data for Poisson regression [PoissonRegrData.R](#)

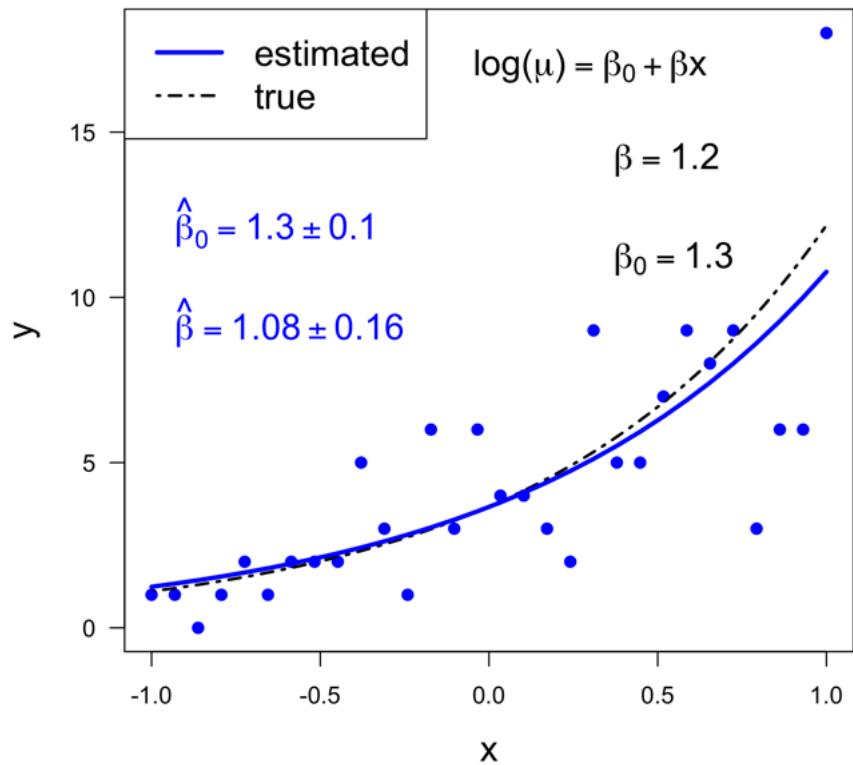


Figure 22.2: Poisson regression: the artificial data (blue dots) were generated as a random sample based on the exact model  $\mu = e^{\beta_0 + \beta x} = e^{1.3 + 1.2x}$  (black broken line). Application of the Poisson regression yields the estimates  $\hat{\beta}_0 = 1.297 \pm 0.104$  and  $\hat{\beta} = 1.08 \pm 0.16$  (red solid line). [PoissonRegr1Exam.R](#)

### 22.1.2 Poisson regression: Zuur et al. (2007) species richness data

As an example for Poisson regression we discuss an example given by Zuur et al. (2007). The species richness data over NAP (height above mean sea level; units: m) are shown in Fig. 22.3. In Poisson regression one assumes that (1) the data at each fixed NAP value follow a Poisson distribution with mean rate  $\lambda = \mu(\text{NAP})$  and (2) the natural logarithm of  $\mu(\text{NAP})$  is a linear function of NAP, i.e.

$$\ln \mu(\text{NAP}) = \beta_0 + \beta \cdot \text{NAP} \quad (22.7)$$

The unknown model parameters  $\beta_0$  (intercept) and  $\beta$  (slope) have to be estimated from the data by a method resembling least squares. The quantity that is minimized is call [deviance](#).

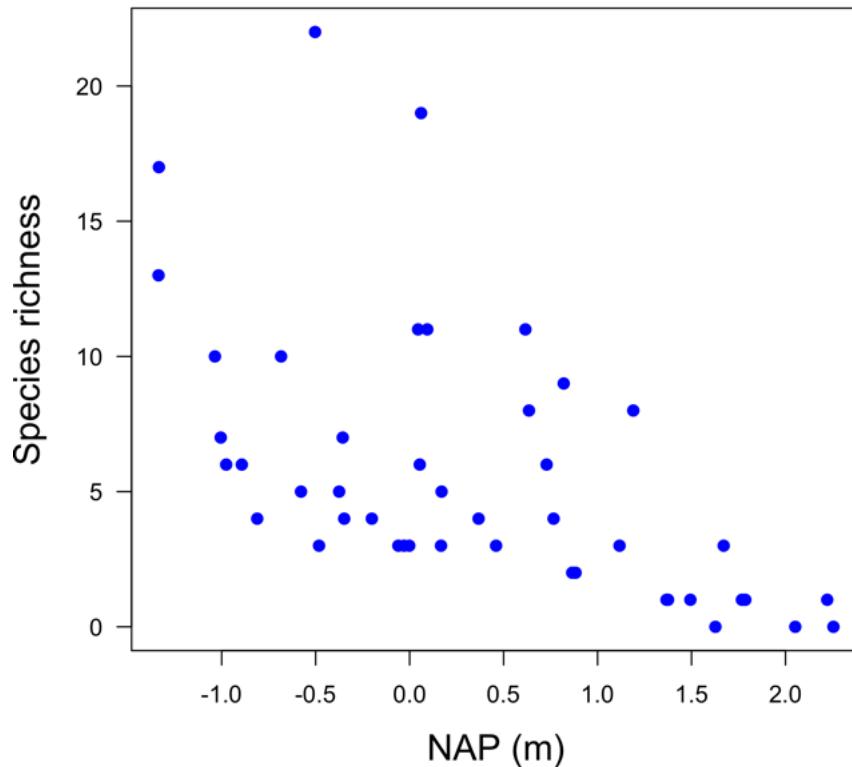


Figure 22.3: Species richness data as over NAP (Normaal Amsterdams Peil; roughly: height above mean sea level); RIKZ = Rijksinstituut voor Kust en Zee = National Institute for Coastal and Marine Management, The Netherlands; data from Zuur et al. (2007). [PoissonRegrRIKZ.R](#)

Application of the generalized linear modeling routine `glm()` yields the following estimates and uncertainties ( $\pm 1\sigma$ ) for  $\beta_0$  (intercept) and  $\beta$  (slope)

$$\hat{\beta}_0 = +1.79 \pm 0.06 \quad (22.8)$$

$$\hat{\beta} = -0.56 \pm 0.07 \quad (22.9)$$

#### Remark on the application of Poisson regression in ecology: overdispersion

"In reality, Poisson regression hardly ever works for ecological count data due to its underlying assumption that the variance equals the mean of the data. For most ecological data sets, the variance is larger than the mean;

this phenomenon is called overdispersion. Negative binomial GLMs and GAMs have become increasingly popular to deal with overdispersion." (Zuur et al., 2009, p. 3)

GAMs = Generalised Additive Models

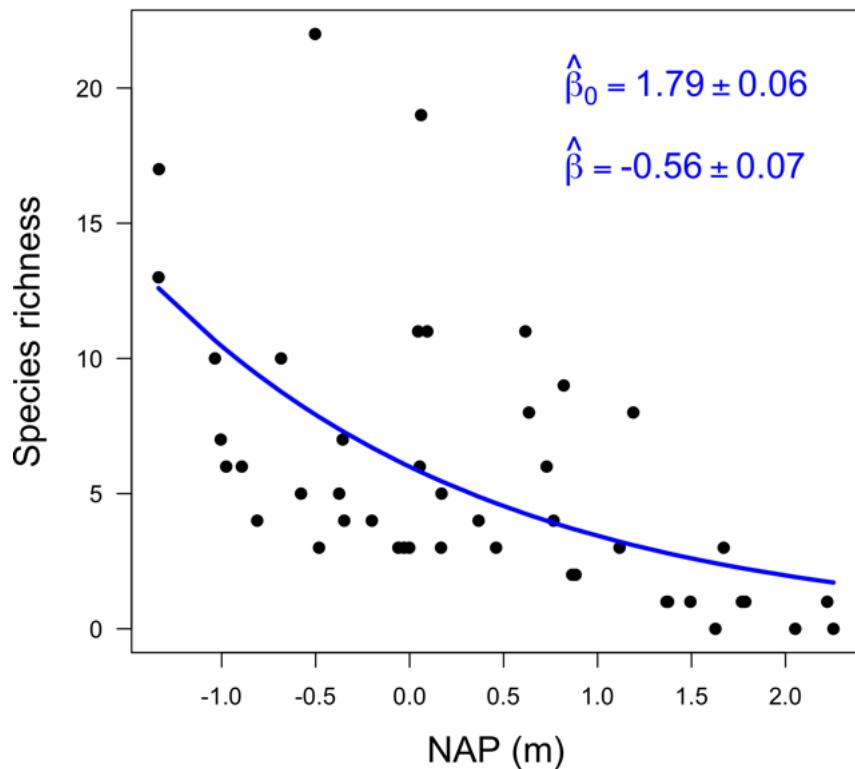


Figure 22.4: Poisson regression of the species richness data (black dots). Imagine, associated with each point on the regression line (blue line), a Poisson distribution with mean rate  $\mu$  equal to  $\exp(\hat{\beta}_0 + \hat{\beta} \text{NAP})$ .  
[PoissonRegrRIKZest.R](#)

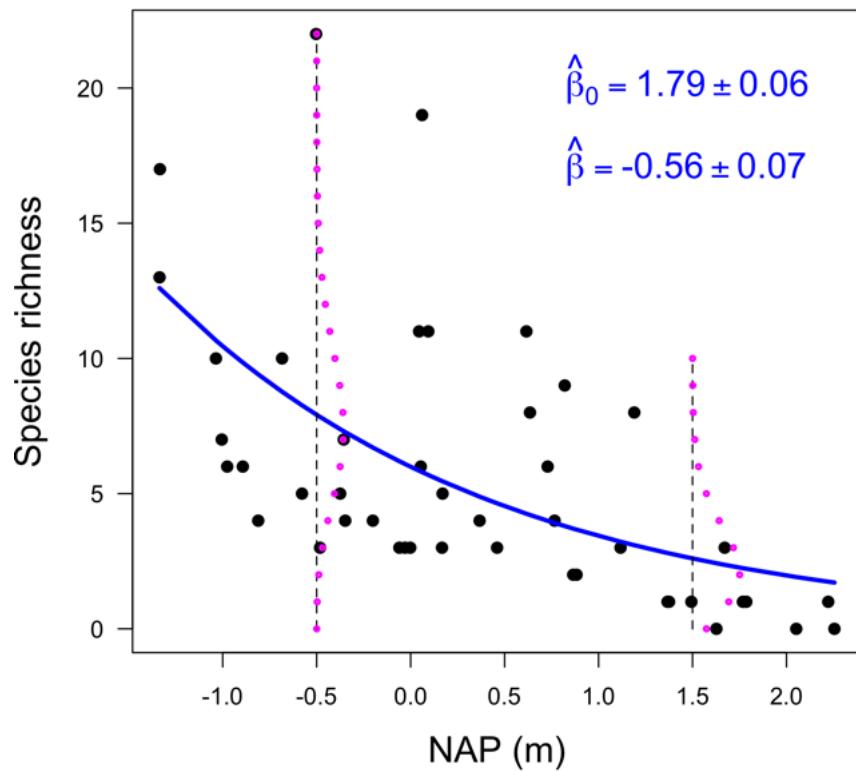


Figure 22.5: Poisson regression of the species richness data (black dots). Associated with each point on the regression line (blue line) is a Poisson distribution (magenta dots; shown for two points only; the black broken lines indicate zero probability for the Poisson distributions) with mean rate  $\mu$  equal to  $\exp(\hat{\beta}_0 + \hat{\beta} NAP)$ .  
[PoissonRegrRIKZplus.R](#)

## 22.2 Logistic regression

Given binary data (0 or 1, failure or success) varying with a quantity  $x$ , one may try to model the probability of success  $p(x)$  of the Bernoulli (also called binary) distribution function. In logistic regression 'logit' is used as link function. This link function is appropriate for cases where the probability  $p(x)$  changes from low to high values following a sigmoidal curve, i.e. with a more or less sharp transition at a critical value of  $x$ .

Logistic regression is a type of generalized linear modeling (GLM). One has to specify (1) the population of the response variable, (2) a linear predictor, and (3) a link function. The population is the Bernoulli distribution, i.e. probability  $p(x)$  for success and  $1 - p(x)$  for failure. The link function

$$\log \frac{p(x)}{1 - p(x)} \quad (22.10)$$

is called 'logit'. The argument  $p(x)/[1 - p(x)]$  is called the 'odds of  $p(x)$ '. The linear predictor  $\eta(x)$  is given by

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta x = \eta(x) \quad (22.11)$$

The inverse of the link function reads

$$p(x) = \frac{1}{1 + e^{-\eta(x)}} \quad (22.12)$$

Without graphical illustration it is not obvious why such a 'strange' link function has been chosen here. Let us assume that  $x$  can vary between -1 and +1 and that  $\beta_0 = 0.2$  and  $\beta = 3$  (just an example). The linear predictor  $\eta(x)$  is a straight line (Fig. 22.6). The corresponding probability of success  $p(x)$  is a sigmoidal function (Fig. 22.7). Now one can understand the choice of the link function:  $\log \{p(x)/[1 - p(x)]\}$  transforms the sigmoidal curve into a straight line which is linear in the model parameters  $\beta_0$  and  $\beta$ .

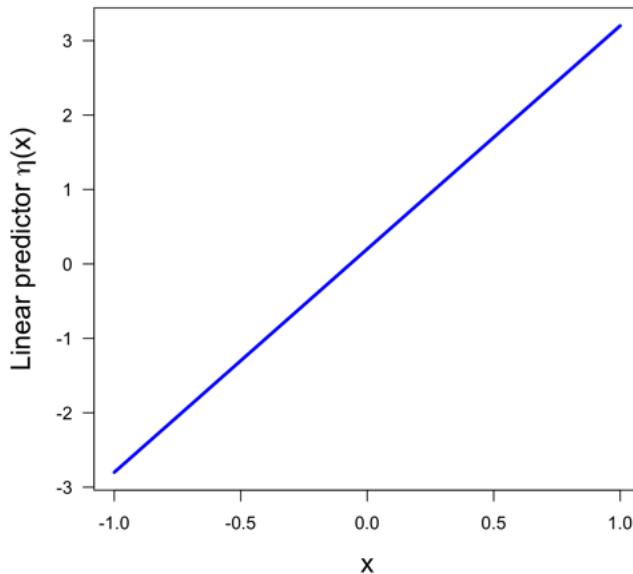


Figure 22.6: The linear predictor  $\eta(x)$  is a straight line. Here: intercept  $\beta_0 = 0.2$ , slope  $\beta = 3$ . [LogisticRegr-Prob.R](#)

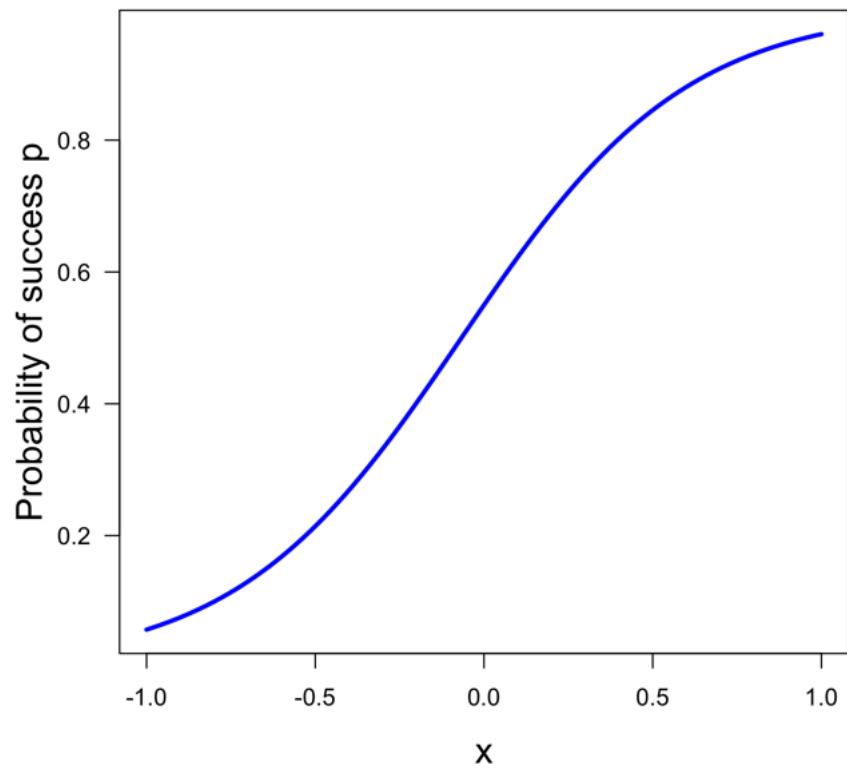


Figure 22.7: The probability of success  $p(x)$  is a sigmoidal function. [LogisticRegrProb.R](#)

### 22.2.1 Logistic regression: an example using artificial data

Apply logistic regression to the following data set

$$\begin{aligned} x &= \{-1.000, -0.931, -0.862, -0.793, -0.724, -0.655, -0.586, -0.517, -0.448, -0.379, \\ &\quad -0.310, -0.241, -0.172, -0.103, -0.034, 0.034, 0.103, 0.172, 0.241, 0.310, \\ &\quad 0.379, 0.448, 0.517, 0.586, 0.655, 0.724, 0.793, 0.862, 0.931, 1.000\} \end{aligned} \quad (22.13)$$

$$y = \{0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\} \quad (22.14)$$

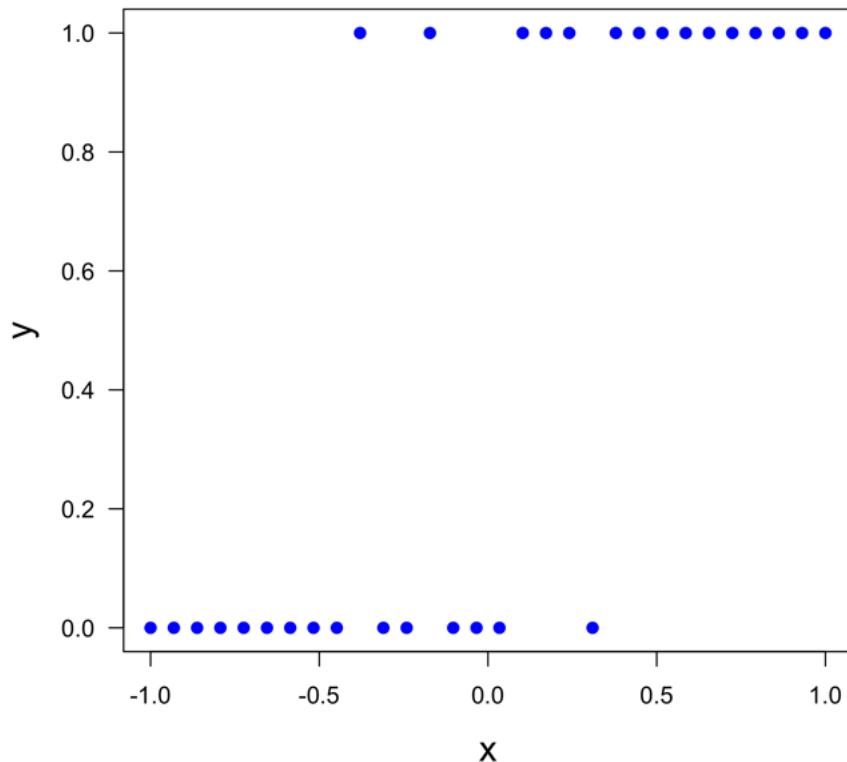


Figure 22.8: Data for logistic regression (Eqs. 22.13 – 22.14). LogRegrEx1n30.R (line 31: set sflag to 1)

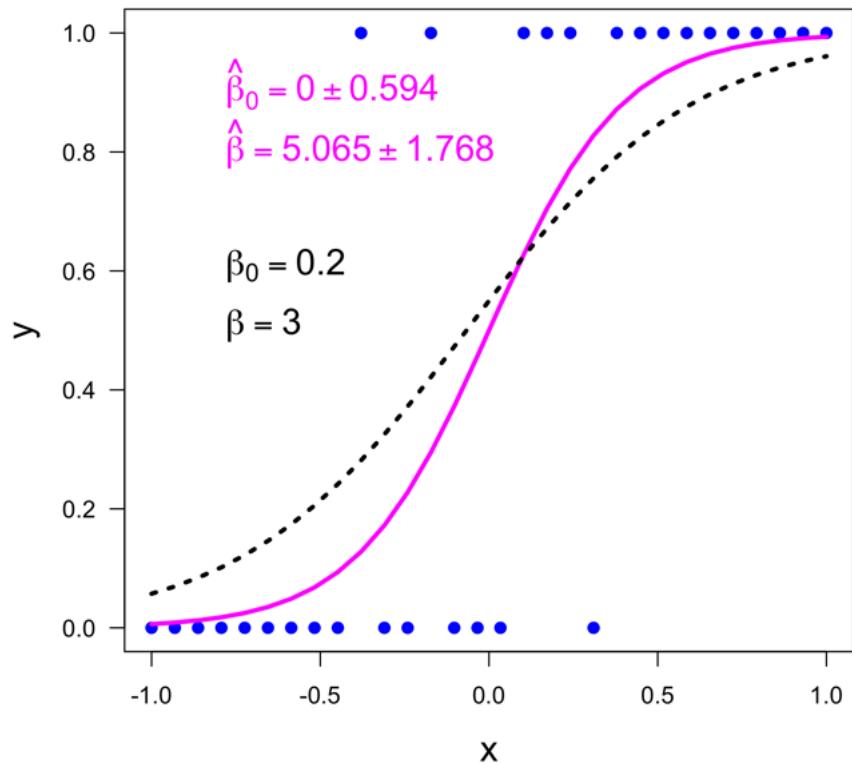


Figure 22.9: Logistic regression (sample size  $n = 30$ ): data (blue dots),  $p(x)$  estimated from the data (magenta solid line), and exact model  $p_{\text{exact}}(x)$  (black broken line) used to generate the data. The estimated intercept  $\hat{\beta}_0 = 0.00 \pm 0.59$  includes the exact value  $\beta_0 = 0.2$  in its  $\pm 1\sigma$  uncertainty range (which is actually quite large!). The estimated slope  $\hat{\beta} = 5.06 \pm 1.77$  does not include the exact slope  $\beta = 3.00$  in its  $\pm 1\sigma$  uncertainty range. [LogRegrEx1n30.R](#) (line 31: set `sflag` to 2)

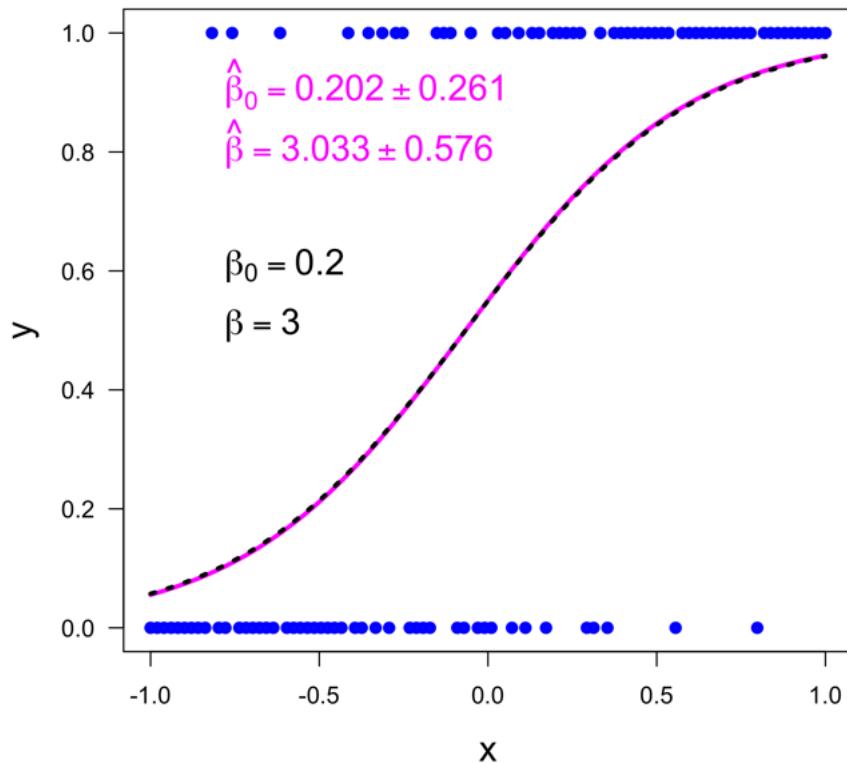


Figure 22.10: Logistic regression (sample size  $n = 100$ ): data (blue dots),  $p(x)$  estimated from the data (red solid line), and exact model  $p_{\text{exact}}(x)$  (black broken line) used to generate the artificial data. The estimated intercept  $\hat{\beta}_0 = 0.202 \pm 0.261$  is very close to the exact value  $\beta_0 = 0.2$ . The estimated slope  $\hat{\beta} = 3.033 \pm 0.576$  is close to the exact value  $\beta = 3$ . [LogRegrEx1n100.R](#)

**Result of the logistic regression:** For a sample size of  $n = 30$ , the estimated curve  $p(x)$  is sigmoidal like the exact curve, however, the deviations in intercept and slope of the linear predictor are quite large and the uncertainties of intercept and slope are large. One might guess that a sample size of  $n = 30$  is not large enough to obtain a reliable estimate of the model parameters and, consequently, of  $p(x)$ . To test this hypothesis, the exercise is repeated with sample size  $n = 100$ . Indeed, for  $n = 100$  one obtains an estimate of  $p(x)$  that is almost identical to the exact curve (Fig. 22.10).

Link	Link function $g$ $\eta_i = g(\mu_i)$	Inverse of link function $\mu_i = g^{-1}(\eta_i)$	Y population	Name of GLM
Identity	$\mu_i$	$\eta_i$	normal	linear regression
Log	$\log(\mu_i)$	$e^{\eta_i}$	Poisson	Poisson regression
Inverse	$\mu_i^{-1}$	$\eta_i^{-1}$		
Inverse-square	$\mu_i^{-2}$	$\eta_i^{-1/2}$		
Square-root	$\sqrt{\mu_i}$	$\eta_i^2$		
Logit	$\log \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$	binomial	logistic regression
Probit	$\Phi(\mu_i)$	$\Phi(\eta_i)$	binomial	
Log-log	$-\log[-\log(\mu_i)]$	$\exp[\exp(-\eta_i)]$	binomial	
Complementary log-log	$-\log[-\log(1-\mu_i)]$	$1 - \exp[\exp(-\eta_i)]$	binomial	

Table 22.1: Some common link functions and their inverses.  $\Phi()$  is the CDF of the standard normal distribution.  $\log$  is the natural logarithm.



# Appendix A

## Probabilities

**Music:** The Rolling Stones: On with the show (from the album Their Satanic Majesties Request, 1967)

### A.1 Frequencies & probabilities: the law of large numbers

*What's the relationship between frequencies (observations) and probabilities (theory)? Can we infer something about probabilities from frequencies? Common sense tells us that relative frequencies are often close to probabilities and this is more often the case for larger and larger sample sizes. A quantitative investigation of the relationship started with Bernoulli (1713) and lead to refinements of the 'law of large numbers' with contributions from Markov, Borel, Cantelli and Kolmogorov and Khinchin in the 20th century. In the following, we will discuss a special case of the law of large numbers, namely a so-called Bernoulli process where only two outcomes (success or failure) are possible in a single trial.*

#### A.1.1 Rolling a fair (unbiased) die

The probability to obtain a '6' when rolling a fair die is  $1/6$  (principle of indifference, common sense). Common sense tells us that on average one can expect one times '6' in 6 trials, or 10 times '6' in 60 trials. However, we also know that it can take a long time (much more than 6 trials) to obtain a '6', that is the probability of  $1/6$  is not a guarantee for at least one times '6' in 6 trials. Rolling the die is a so-called **Bernoulli process** where only two outcomes (success or failure) are possible in a single trial. Here '6' is considered a success, obtaining any other number is a failure.

We will ask the following questions:

1. What's the probability to obtain a relative frequency,  $f$ , that is exactly equal to the probability  $1/6$ ? We expect that the answer depends on sample size,  $n$ .
2. What is the probability to find a relative frequency,  $f$ , that is close to the probability  $1/6$ ? Let's assume in the range of  $\pm 10\%$  around  $1/6$ . We expect that the answer depends on sample size,  $n$ .
3. By switching the point of view: What is the probability that the true probability  $p$  is in the range  $f \pm \delta$  for observed relative frequency  $f$  and chosen uncertainty  $\delta$ ?

In order to answer these question we have to know: What are the probabilities for the various possible observations (= number of successes or, after division by the number of trials  $n$ , relative frequencies)? The answer is given by the binomial distribution (Subsection , Eq. 6.32):

$$\mathcal{B}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (\text{A.1})$$

with  $k$  number of successes (here: number of sixes),  $n$  number of trials,  $p$  probability for success in a single trial. For  $n = 6$  and  $p = 1/6$  the probabilities for the different number of successes ( $k$ ) are shown in Fig. A.1. The probability to obtain exactly one '6' in 6 trials is  $\approx 0.40$ , however, there is also a large probability ( $\approx 0.33$ ) for no '6' in 6 trials. The probability to obtain at least one '6' is  $1 - 0.33 = 0.67$ .

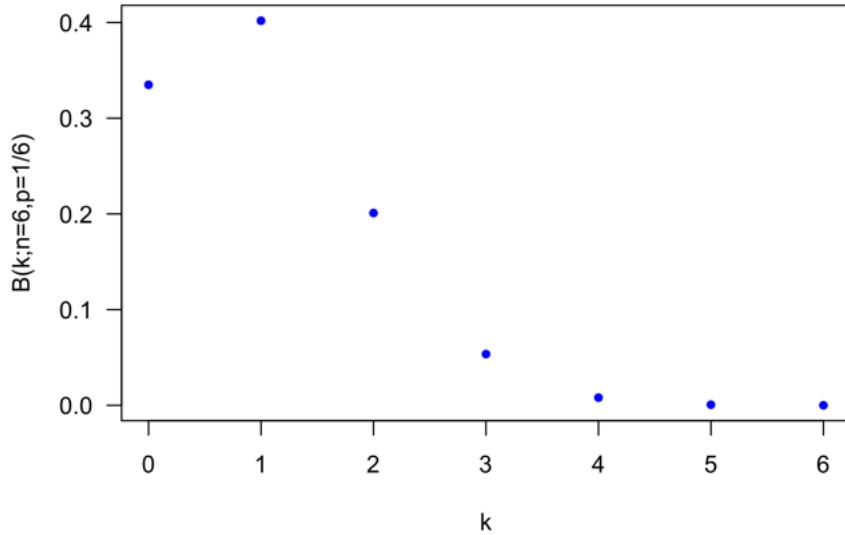


Figure A.1: The binomial probability distribution for  $n = 6$  trials and probability for success in a single trial  $p = 1/6$  (fair die). The probability to obtain exactly one '6' in 6 trials is  $\approx 0.40$  (or 40%), however, there is also a large probability ( $\approx 0.33$  or 33%) for no '6' in 6 trials. The probability to obtain at least one '6' is  $1 - 0.33 = 0.67$  or 67%. The probability for '6' in all 6 trials is  $\approx 2.1 \cdot 10^{-5} = 0.000021$  or in 1 of 46656 samples of size  $n = 6$ . [ProbBernoulliDice6n.R](#)

We will now address the first question: What's the probability to obtain a relative frequency,  $f$ , that is exactly equal to the probability  $1/6$ ? In order to allow relative frequencies that are exactly equal to  $p = 1/6$ , the number of trials,  $n$ , have to be a multiple of  $1/p = 6$ , i.e.  $n = 6, 12, 18, \dots$  The probability for obtaining a relative frequency,  $f$ , that is exactly equal to the probability  $p = 1/6$  for number of trials  $n = 6, 12, 18, \dots, 6000$  is shown in Fig. A.2. With increasing  $n$  this probability decreases and it looks as if it approaches 0 for  $n \rightarrow \infty$ . This is may be not what we expected – or did we had certain expectations? Anyway, we would like to understand, why this probability is decreasing with increasing  $n$  (question 1a).

To answer this question, let us look at the binomial probability distribution for  $n = 600$  (Fig. A.3). It peaks at  $k = n \cdot p = 600/6 = 100$ . A closer look in the vicinity of  $k = 100$  (Fig. A.4) allows us to answer question 1a: although the probability for  $f = k/n = 100/600 = p$  is maximal, the probabilities for many frequencies in the vicinity of  $f = p$  are also quite large and their sum is actually much higher than that of the single maximum.

The normal distribution with the mean  $\mu = n p = 100$  and variance  $\sigma^2 = n p(1 - p) \approx 83.3$  ( $\sigma \approx 9.1$ ) of the binomial distribution forms an envelope to the discrete distribution (Fig. A.5). It looks like as if  $1/\sqrt{2\pi} \approx 0.3989$  (the maximum of the standard normal distribution) divided by the standard deviation of the binomial distribution provides an upper limit of the probability for  $f = p$  that is approached for  $n \rightarrow \infty$ :

$$\text{probability}(f = p; n) \leq \frac{1}{\sqrt{2\pi n p(1 - p)}} \quad (\text{A.2})$$

(Fig. A.6).

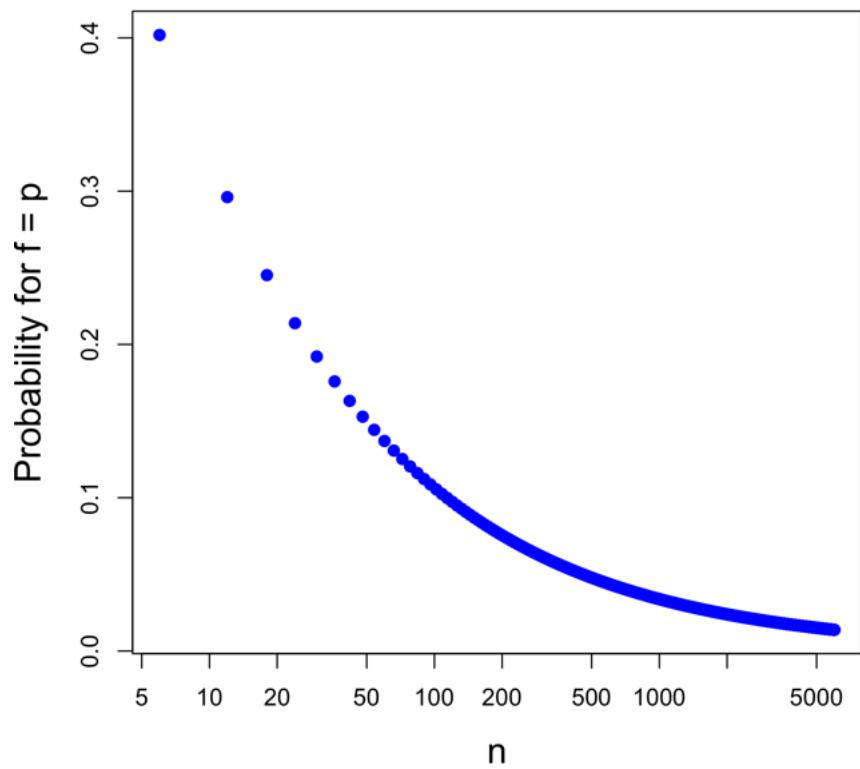


Figure A.2: The probability for obtaining a relative frequency,  $f$ , that is exactly equal to the probability  $p = 1/6$  for number of trials  $n = 6, 12, 18, \dots, 6000$ . [ProbBernoulliDieEx.R](#)

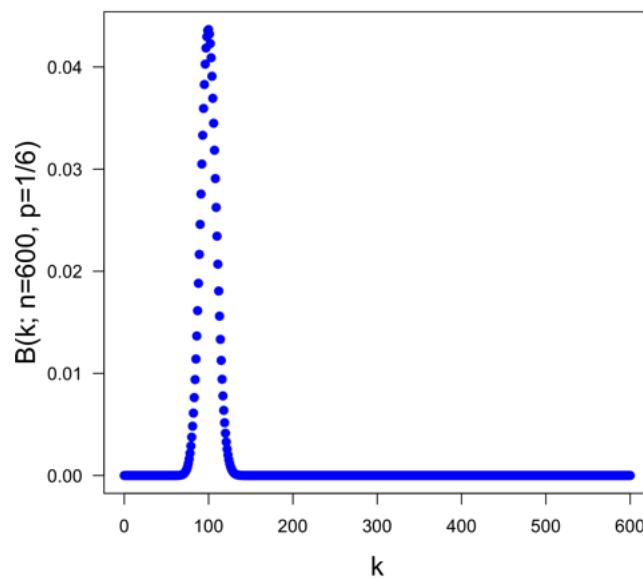


Figure A.3: The binomial probability distribution for  $n = 600$  and  $p = 1/6$  peaks at  $k = n \cdot p = 600/6 = 100$ .  
[ProbBinomPDn600p1o6.R](#)

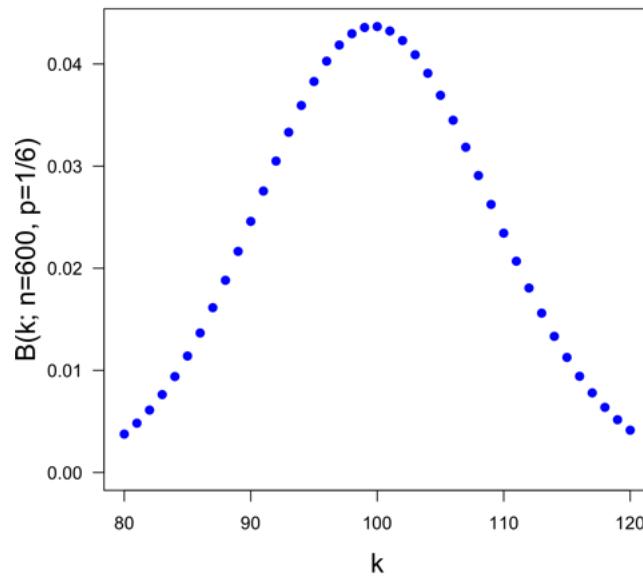


Figure A.4: The binomial probability distribution for  $n = 600$  and  $p = 1/6$  in the vicinity of  $k = n \cdot p = 600/6 = 100$ .  
[ProbBernoulliDice6n.R](#)

In summary, we have seen that the probability for  $f = p$  decreases with increasing  $n$ . The plot of the binomial distribution for  $n = 600$  helped us to understand why this is the case. In addition, we found an upper limit for the probability for  $f = p$  that gives a good approximation already at  $n > 50$ .

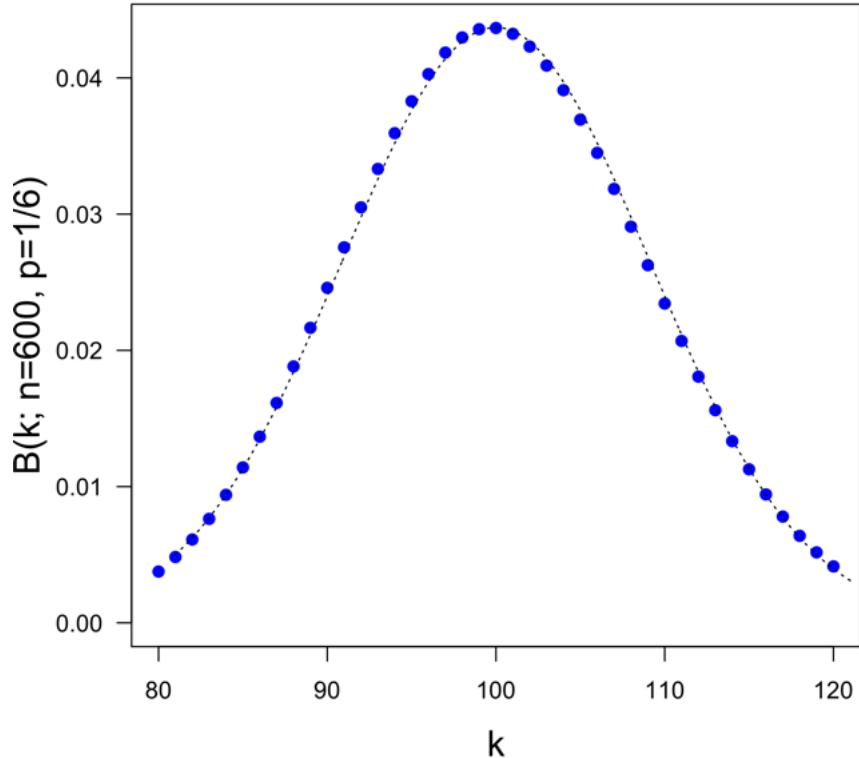


Figure A.5: The binomial probability distribution (blue dots) for  $n = 600$  and  $p = 1/6$  in the vicinity of  $k = n \cdot p = 600/6 = 100$ . The normal distribution with the mean  $\mu = n p = 100$  and variance  $\sigma^2 = n p(1 - p) \approx 83.3$  ( $\sigma \approx 9.1$ ) of the binomial distribution forms an envelope (black dashed curve) to the discrete distribution.  
[ProbBernoulliDie600nDnormal.R](#)

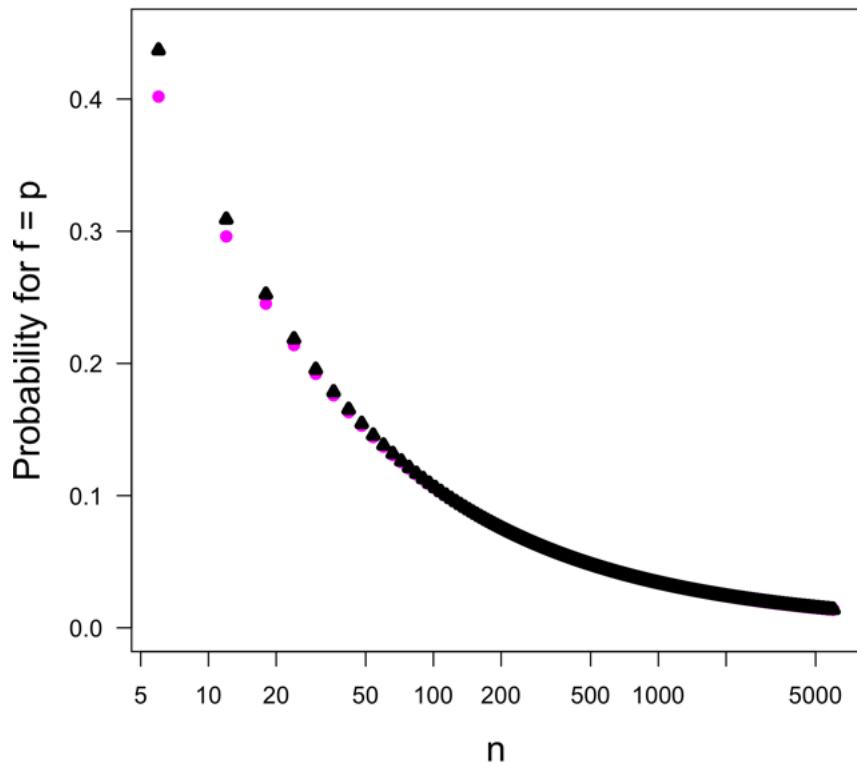


Figure A.6: The probability for obtaining a relative frequency,  $f$ , that is exactly equal to the probability  $p = 1/6$  for number of trials  $n = 6, 12, 18, \dots, 6000$  (magenta dots). The ratio of  $1/\sqrt{2\pi} \approx 0.3989$  (the maximum of the standard normal distribution) and the standard deviation of the binomial distribution (black triangles) seems to provide an upper limit of the probability for  $f = p$  that is approached for  $n \rightarrow \infty$ .

[ProbBernoulliDieExSigma.R](#)

We will now address the second question: What is the probability to find a relative frequency,  $f$ , that is close to the the probability  $1/6$ ? Let's assume in the range of  $\pm 10\%$  around  $1/6$ . The probability for relative frequency in the specified range ( $p = 1/6 \pm 1/60$  or  $n/6 - n/60 \leq k \leq n/6 + n/60$ ) is shown in Fig. A.7: it is approaching 1 for  $n \rightarrow \infty$ . The explanation of the up-and-down variations at small  $n$  values is left as an exercise for the reader (it's not that difficult to find out).

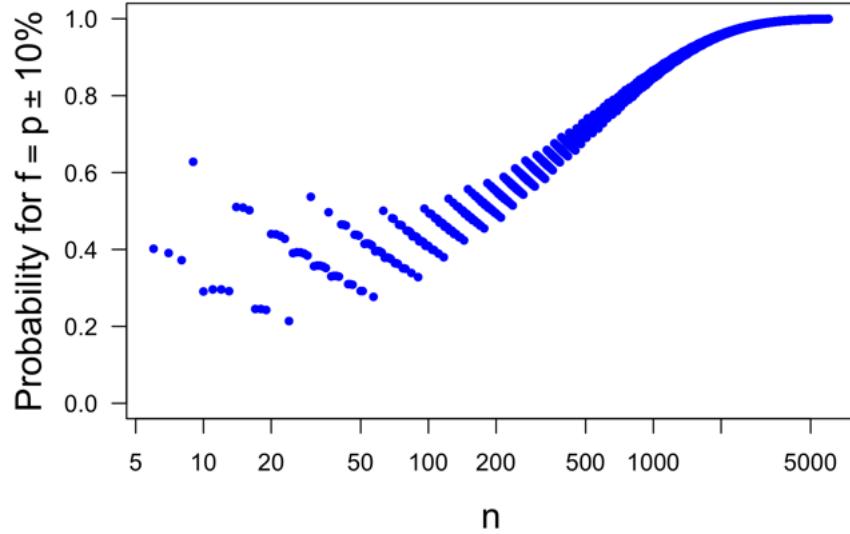


Figure A.7: The probability for relative frequency in the specified range:  $p = 1/6 \pm 1/60$  or  $n/6 - n/60 \leq k \leq n/6 + n/60$ . [ProbBernoulliDie10percent.R](#)

The normal envelope to the binomial distribution is a most valid information. We can express our chosen uncertainty of 10% – which is equivalent to  $\Delta p = p/10$  or  $\Delta k = n p/10$  – in terms of the standard deviation of the binomial distribution,  $\sigma = \sqrt{n p(1 - p)}$ , which is used as the standard deviation of the normal envelope:

$$x(n) = \frac{\Delta k}{\sigma} = \frac{n p}{10 \sqrt{n p(1 - p)}} \propto \sqrt{n}, \quad (\text{A.3})$$

i.e. the uncertainty range is  $\pm x(n)$  sigmas. How much probability is in the range  $\pm x(n)$  sigmas around the location of the peak (mean) of a normal distribution? This can be easily calculated from the CDF of the standard normal distribution:  $p_{\text{normal approximation}} = \text{CDF}(x(n)) - \text{CDF}(-x(n))$ . The resulting values from this approximation provide (almost everywhere) a lower limit to the exact values.

**In summary**, the probability that relative frequencies,  $f$ , fall into a specified uncertainty range,  $p \pm \Delta p$ , around the (true) probability,  $p$ , increases with sample size (number of trials),  $n$ , and approaches 1 for  $n \rightarrow \infty$ . The probability for the specified range can be calculated as a sum over probabilities from the binomial distribution. The sum

$$\Pr(n, p, \Delta k) = \sum_{k=n p-\Delta k}^{n p+\Delta k} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (\text{A.4})$$

can be approximated by the integral

$$\frac{2}{\sqrt{\pi}} \int_0^q e^{-\xi^2} d\xi \quad \text{where} \quad q = \frac{\Delta k + 1/2}{\sqrt{2 n p(1 - p)}} \quad (\text{A.5})$$

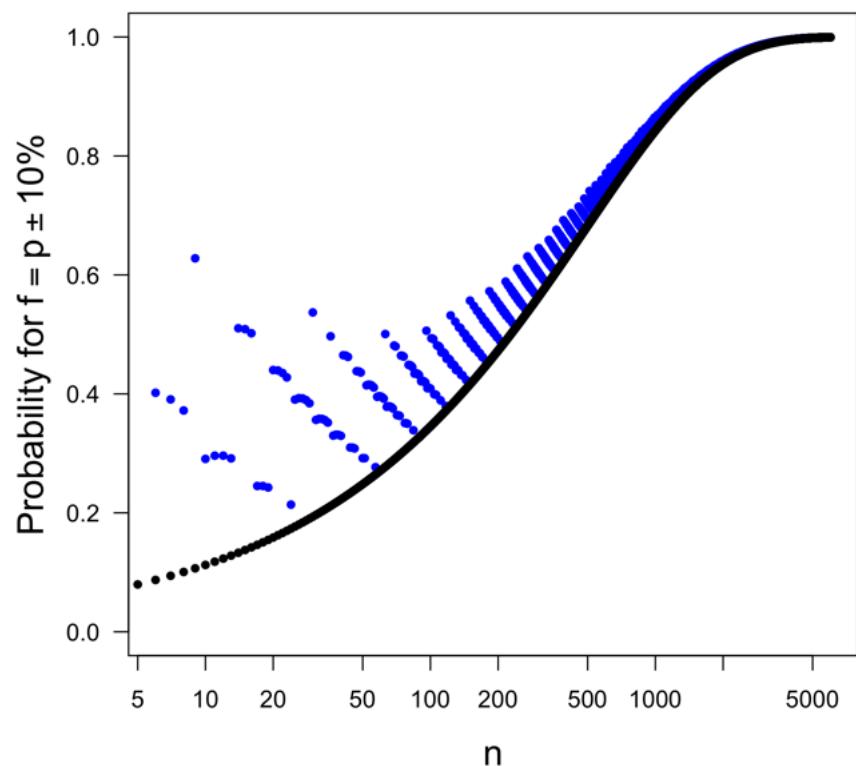


Figure A.8: The probability for relative frequency (blue dots) in the specified range:  $p = 1/6 \pm 1/60$  or  $n/6 - n/60 \leq k \leq n/6 + n/60$ . The approximation based on a normal envelope to the binomial distributions yields (almost everywhere) a lower limit (black dots). [ProbBernoulliDie10percentNormal.R](#)

(Haussner, 1899, p. 158).

Finally, we will address the third question: What is the probability that the true probability  $p$  is in the range  $f \pm \delta$  for observed relative frequency  $f$  and chosen uncertainty  $\delta$ ? The answer is given, for example, in Haussner (1899, p. 159):

**Theorem of Bernoulli:** Given  $a$  successes in  $n$  trials. The probability  $p$  for success in a single trial lies with probability

$$\Pr = \frac{2}{\sqrt{\pi}} \int_0^q e^{-\xi^2} d\xi \quad (\text{A.6})$$

between the boundaries

$$\frac{a}{n} \pm q \sqrt{\frac{2a(n-a)}{n^2}} \quad (\text{A.7})$$

Example:  $n = 1000, a = 130 \Rightarrow p$  lies (1) with probability  $\Pr = 0.986$  (98.6%) in the range  $0.104 \leq p \leq 0.156$  ( $\hat{p} = 0.13 \pm 0.026$ ) or (2) with probability  $\Pr = 0.778$  (77.8%) in the range  $0.117 \leq p \leq 0.143$  ( $\hat{p} = 0.13 \pm 0.013$ ) or ... It's your choice, however, there is a trade-off: if you choose a small uncertainty the probability for the narrow range is small and vice versa.

The Theorem of Bernoulli is a special case of the '**law of large numbers**' which states (under more general conditions, i.e. other than Bernoulli processes) that 'the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed' (Wikipedia, law of large numbers, 7 May 2020).

#### Exercise 54 Theorem of Bernoulli

Generate a random sample of size  $n = 1000$  from a Bernoulli process with probability  $p = 1/6$  for success in a single trial, estimate the probability by the relative frequency of success,  $p_{\text{est}}$ , and its uncertainty by the standard error of the mean,  $\Delta p_{\text{est}}$ . Use the Theorem of Bernoulli to calculate the probability for  $p$  lies in the range  $[p_{\text{est}} - \Delta p_{\text{est}}, p_{\text{est}} + \Delta p_{\text{est}}]$  (also denoted as  $\hat{p} = p_{\text{est}} \pm \Delta p_{\text{est}}$ ). Is the result surprising?

Repeat the analysis with sample of size  $n = 10000$ .

## A.2 Application of Bayes' Theorem: heroin addiction

Weber et al. (2018) formulate and solve the following 'heroin addiction problem': "The probability of being addicted to heroin is 0.01% for a person randomly picked from a population (base rate). If a randomly picked person from this population is addicted to heroin, the probability is 100% that he or she will have fresh needle pricks (sensitivity). If a randomly picked person from this population is not addicted to heroin, the probability is 0.19% that he or she will still have fresh needle pricks (false alarm rate). What is the probability that a randomly picked person from this population who has fresh needle pricks is addicted to heroin (posterior probability)?"

Here is the solution:  $H$  denotes the proposition 'person is addicted to heroin' and  $N$  denotes the background information 'person has fresh needle pricks'. We want to know the posterior probability  $P(H|N)$  = 'person is addicted to heroin given that he/she has fresh needle pricks'. Bayes' Theorem reads

$$P(H|N) = \frac{\overbrace{P(N|H)}^{=100\%} \overbrace{P(H)}^{=0.01\%}}{\underbrace{P(N)}_{=?}} \quad (\text{A.8})$$

The probability to observe a person with needle pricks,  $P(N)$ , is not directly given, however, it is equal to

$$P(N|H \text{ or not } H) = \underbrace{P(N|H)}_{=100\%} \underbrace{P(H)}_{=0.01\%} + \underbrace{P(N|\text{not } H)}_{=0.19\%} \underbrace{P(\text{not } H)}_{=(100-0.19)\% \approx 100\%} \quad (\text{A.9})$$

where we applied marginalization and the generalized sum rule to mutually exclusive propositions ( $H$  and not  $H$  are mutually exclusive and thus  $P(H \text{ and not } H) = 0$ ). Inserting (Eq. A.9) into (Eq. A.8) yields

$$P(H|N) = \frac{\overbrace{P(N|H)}^{=100\%} \overbrace{P(H)}^{=0.01\%}}{\underbrace{P(N|H)P(H) + P(N|\text{not } H)}_{=100\% \cdot 0.01\% + 0.19\%} \underbrace{P(\text{not } H)}_{=(100-0.19)\% \approx 100\%}} \quad (\text{A.10})$$

$$\approx \frac{100 \cdot 0.01}{100 \cdot 0.01 + 0.19 \cdot 100} = \frac{1}{20} \quad (\text{A.11})$$

or a 5% chance.

## A.3 Calculation of the Lagrange multipliers for the loaded die

### A.3.1 Method 1: derive and solve the 'z-equation'

The Lagrange multipliers  $\lambda_1$  and  $\lambda_2$  can be calculated from the constraints

$$1 = \sum_{j=1}^6 p_j = \sum_{j=1}^6 e^{-1+\lambda_1+j\lambda_2} = e^{-1+\lambda_1} \sum_{j=1}^6 z^j = e^{-1+\lambda_1} \left( \frac{1-z^7}{1-z} - 1 \right) \quad (\text{A.12})$$

$$\mu = \sum_{j=1}^6 j \cdot p_j = \sum_{j=1}^6 j \cdot e^{-1+\lambda_1+j\lambda_2} = e^{-1+\lambda_1} \sum_{j=1}^6 j z^j = e^{-1+\lambda_1} z \frac{1-7z^6+6z^7}{(1-z)^2} \quad (\text{A.13})$$

where  $z = e^{\lambda_2}$ . Elimination of  $e^{-1+\lambda_1}$  between these equations leads to an equation for  $z$

$$f(z) := \frac{1-z^7}{1-z} - 1 - \frac{z}{\mu} \frac{1-7z^6+6z^7}{(1-z)^2} = 0 \quad (\text{A.14})$$

that has to be solved numerically. For  $\mu = 3.8$  one obtains the solution (root)  $z = 1.1091$  and thus  $\lambda_2 = 0.1035$  and  $\lambda_1 = -1.1697$ . The probabilities read

$$p = \{0.1267, 0.1405, 0.1558, 0.1728, 0.1917, 0.2126\} \quad (\text{A.15})$$

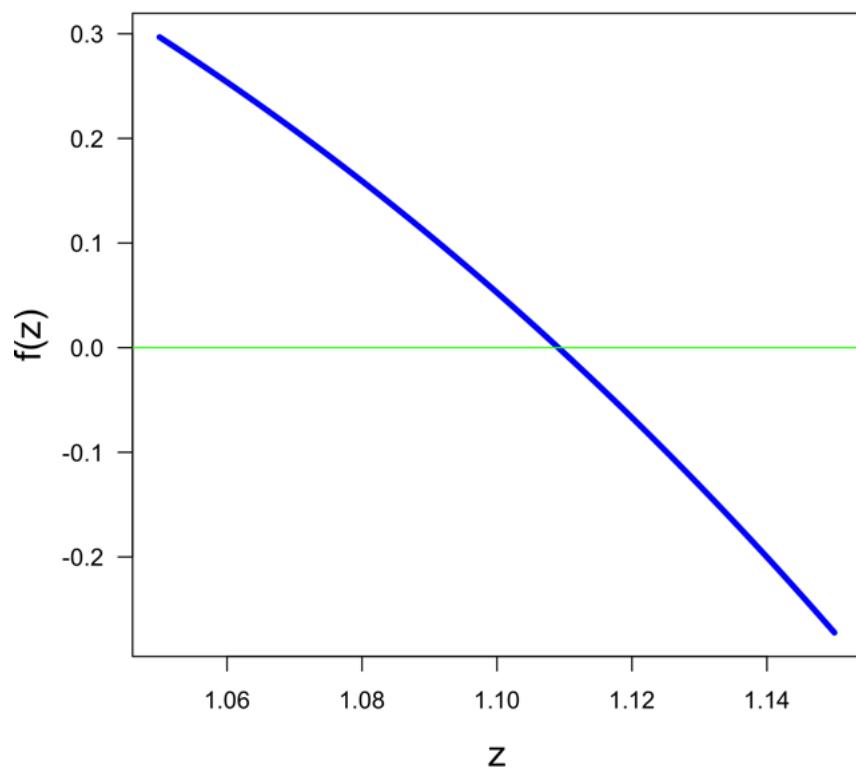


Figure A.9: The function  $f(z)$  (Eq. A.14) in an interval about the root  $z = 1.1091$ .  
[ProbLagrangeMultipliersLoadedDie.R](#)

### A.3.2 Method 2: brute force numerical solution

Instead of searching for the solution of a system of nonlinear equations one may apply a dirty trick: write each constraint in the form 'constraint = 0', square it, than add up the squares, and finally find the minimum with a value equal to zero.<sup>1</sup> In the case of the loaded die we define the function  $h()$  by

$$h(\lambda_1, \lambda_2) := \left( -1 + \sum_{j=1}^6 e^{-1+\lambda_1+j\lambda_2} \right)^2 + \left( -3.8 + \sum_{j=1}^6 j \cdot e^{-1+\lambda_1+j\lambda_2} \right)^2 \quad (\text{A.16})$$

Because  $\mu = 3.8$  is only slightly different from the value of 3.5 for the unbiased dice, we expect the searched for Lagrange multipliers  $\lambda_1$  and  $\lambda_2$  to be close to the Lagrange multipliers for the unbiased die, namely,  $\lambda_1^{(u)} = 1 - \ln 6 = -0.7918$  (calculated above, Eq. 4.52) and  $\lambda_2^{(u)} = 0$  (all  $p_j$  independent from  $j$ , Eq. 4.59). We will evaluate  $h(\lambda_1, \lambda_2)$  for its arguments first on a grid  $(-1.2 < \lambda_1 < -0.7) \times (-0.1 < \lambda_2 < 0.2)$ , search for the minimum, and then refine the grid around the minimum. The minimum of  $h(\lambda_1, \lambda_2)$  is located at  $\lambda_1^{(1)} \approx -1.1700$  and  $\lambda_2^{(1)} \approx 0.1036$ . Repetition of the described procedure on a refinement grid around this first approximation values yields  $\lambda_1^{(2)} \approx -1.16969$  and  $\lambda_2^{(1)} \approx 0.10353$ . R code: [LagrangeLDieBrute.R](#)

The 'brute force' method can be applied in case of one or two unknown Lagrange multipliers, however, becomes more and more inefficient for more complicated problems.

### A.3.3 Method 3: iterative solution

An iterative solution method for the Lagrange multipliers of the loaded die can be derived by expanding all probabilities  $p_j(\lambda_1, \lambda_2)$  around an initial guess into a Taylor series and to drop quadratic and higher order terms. This linear approximation is then inserted into the two constraints. The resulting system of linear equations can be solved analytically. The solution can be used as initial guess for the next iteration. As initial guess  $p_j^{(in)}$  we will use the probabilities for the unbiased die which are all equal to 1/6.

The expression for the probabilities  $p_j$  can be rewritten as

$$p_j = e^{-1+\lambda_1+j\lambda_2} = e^{-1+\lambda_1^{(u)}+j\lambda_2^{(u)}+x+jy} = p_j^{(in)} e^{x+jy} \approx \frac{1}{6} (1 + x + jy) \quad (\text{A.17})$$

where we expanded the exponential function and dropped all terms beyond the linear one. Inserting this approximation into the constraints yields

$$1 = \sum_{j=1}^6 p_j^{(in)} (1 + x + jy) = c_1 + c_1 x + c_2 y \quad (\text{A.18})$$

$$3.8 = \sum_{j=1}^6 j p_j^{(in)} (1 + x + jy) = c_2 + c_2 x + c_3 y \quad (\text{A.19})$$

where

$$c_1 = \sum_{j=1}^6 p_j^{(in)} \quad (\text{A.20})$$

$$c_2 = \sum_{j=1}^6 j p_j^{(in)} \quad (\text{A.21})$$

$$c_3 = \sum_{j=1}^6 j^2 p_j^{(in)} \quad (\text{A.22})$$

---

<sup>1</sup>Here we assume that only one such minimum exists.

and thus

$$y = \frac{3.8 - c_2 - c_2(1 - c_1)/c_1}{c_3 - c_2^2/c_1} \quad (\text{A.23})$$

$$x = \frac{1 - c_1}{c_1} - \frac{c_2}{c_1} y \quad (\text{A.24})$$

The results of the iterations are shown in Figs. A.10 to A.11.

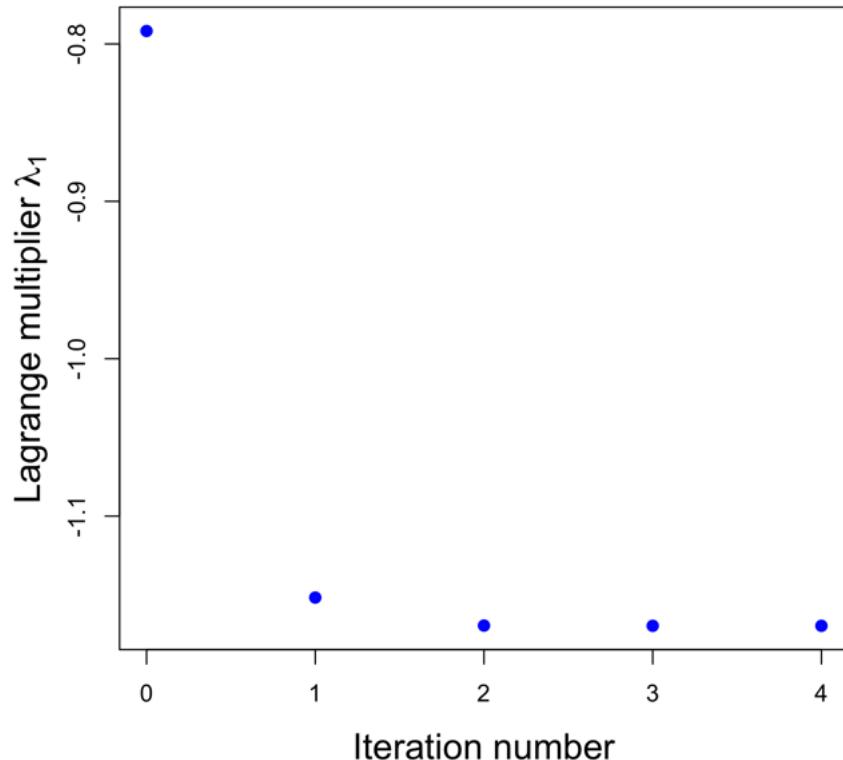


Figure A.10: Lagrange multiplier  $\lambda_1$  for the loaded die with  $\mu = 3.8$ : the iteration starts with  $\lambda_1^{(u)} = 1 - \ln 6 \approx -0.791759$  and approaches extremely fast (2 to 3 iterations)  $\lambda_1 \approx -1.169714$ . [LagrangeLDielter.R](#)

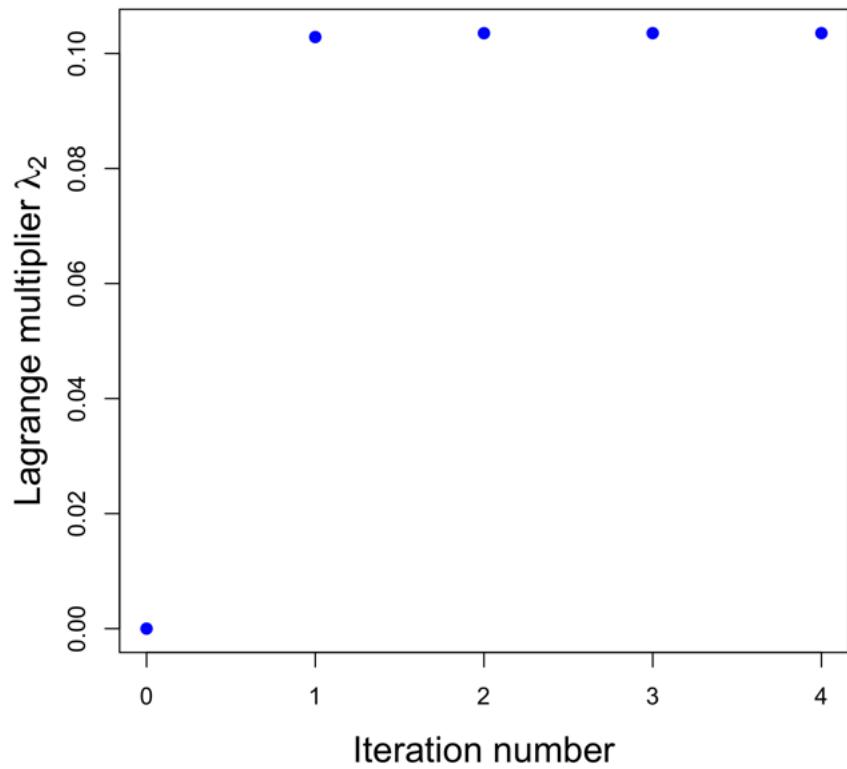


Figure A.11: Lagrange multiplier  $\lambda_2$  for the loaded die with  $\mu = 3.8$ : the iteration starts with  $\lambda_2^{(u)} = 0$  and approaches extremely fast (2 to 3 iterations)  $\lambda_2 \approx 0.103535$ . [LagrangeLDieIter.R](#)

## A.4 The loaded die once again

### Exercise 55 Loaded die

A loaded die shows a mean outcome  $\mu = 3.8$  that is different from the value 3.5 for the unbiased die. Modify the probability distribution for the unbiased die (all  $p_j = 1/6$ ) in such a way that two opposite faces (1 and 6, 2 and 5, or 3 and 4, respectively) possess probabilities  $p_j = 1/6 \pm x$  ( $x > 0$ ) while all other probabilities remain unchanged. Show that  $p_3$  can become negative using this procedure.

## A.5 Lagrange multipliers for the discrete exponential function

The Lagrange multipliers are calculated from the constraints for normalization and mean value:

$$1 = \sum_{j=0}^{\infty} e^{\lambda_0-1+\lambda_1 j} = e^{\lambda_0-1} \sum_{j=0}^{\infty} q^j = e^{\lambda_0-1} \frac{1}{1-q} = \frac{e^{\lambda_0-1}}{1-e^{\lambda_1}} \quad (\text{A.25})$$

$$\mu = \sum_{j=0}^{\infty} j e^{\lambda_0+\lambda_1 j-1} = e^{\lambda_0-1} \sum_{j=0}^{\infty} j q^j e^{\lambda_0-1} = e^{\lambda_0-1} q \frac{d}{dq} q^j = e^{\lambda_0-1} q \frac{d}{dq} \sum_{j=0}^{\infty} q^j \quad (\text{A.26})$$

$$= e^{\lambda_0-1} q \frac{d}{dq} \frac{1}{1-q} = e^{\lambda_0-1} \frac{q}{(1-q)^2} = e^{\lambda_0-1} \frac{e^{\lambda_1}}{(1-e^{\lambda_1})^2} = \underbrace{\frac{e^{\lambda_0-1}}{1-e^{\lambda_1}}}_{=1} \frac{e^{\lambda_1}}{1-e^{\lambda_1}} \quad (\text{A.27})$$

where the magnitude of  $q = e^{\lambda_1}$  has to be smaller than 1 and thus  $\lambda_1 < 0$ . The combination of the two constraints leads to an equation for  $e^{\lambda_1}$

$$\mu = \frac{e^{\lambda_1}}{1-e^{\lambda_1}} \quad (\text{A.28})$$

which can be readily solved yielding

$$e^{\lambda_1} = \frac{\mu}{1+\mu}. \quad (\text{A.29})$$

Inserting this expression for  $e^{\lambda_1}$

$$e^{\lambda_0-1} = 1 - \frac{\mu}{1+\mu} = \frac{1}{1+\mu} \quad (\text{A.30})$$

and finally

$$p_j = \frac{1}{1+\mu} \left( \frac{\mu}{1+\mu} \right)^j \quad (\text{A.31})$$

## A.6 More MaxEnt distributions

MaxEnt distributions for various constraints are given in Lisman & Van Zuylen (1972) and Park & Bera (2009). These include the discrete uniform and the geometric probability distribution. Unfortunately, the binomial and the Poisson distribution can not be derived from MaxEnt (Lisman & Van Zuylen, 1972) although they maximize entropy under appropriate constraints (Harremoës, 2001). In the following, the derivations of various probability distributions are asked for in exercises.

**Exercise 56 Derive the MaxEnt probability distribution for  $k = a, a+1, \dots, b$  with normalization as the only constraint**

**Exercise 57 Derive the MaxEnt probability distribution for  $k = 0, 1$  and the constraint mean  $\mu = p$**

**Exercise 58 Derive the MaxEnt probability distribution  $p_k, k = 1, 2, 3, \dots$  when the mean  $\mu = 1/p$  is constrained**

## A.7 Mean kinetic energy for particles with three possible speeds (MaxEnt)

Suppose a system contains particles that can take on one of three different speeds, namely 1, 2, 3. The corresponding kinetic energy is given by the square of the speeds (we will leave out the factor  $m/2$  where  $m$  is the mass of a single particle; all particles have the same mass), i.e. the kinetic energy can take the values 1, 4, 9, respectively. Calculate the probability distribution for a mean energy per particle  $E_{\text{mean}}$  of 4.

Guess how the probability distribution  $\{p_1, p_2, p_3\}$  (index = speed) will look like: The mean energy per particle of 4 is close to the energy of particles with speed = 2. Therefore one might expect high values for  $p_2$ . Few speed 3 particles contribute a lot to the mean energy and therefore one might expect very small values for  $p_3$ . In order to balance the energy contribution of speed 3 particles, the probability for speed 1 particles has to be quite large. Overall we expect  $p_3 < p_1 < p_2$  or even  $p_3 < p_2 < p_1$  and we hope that MaxEnt can give us a quantitative answer.

Constraints:

$$\sum_{j=1}^3 p_j = 1 \quad (\text{normalization}) \quad (\text{A.32})$$

$$\sum_{j=1}^3 j^2 p_j = E_{\text{mean}} = 4 \quad (\text{mean energy per particle}) \quad (\text{A.33})$$

$\Rightarrow$  maximize

$$\mathcal{L}(p_j; \lambda_0, \lambda_1) = \sum_{j=1}^3 \left( -p_j \ln p_j + \lambda_0 p_j + \lambda_1 j^2 p_j \right) \quad (\text{A.34})$$

where  $\mathcal{L}$  is the Lagrange function (Shannon entropy coupled with constraints). A necessary condition for a maximum is the vanishing of the partial derivatives of  $\mathcal{L}$  with respect to the probabilities  $p_j$ :

$$\frac{\partial \mathcal{L}}{\partial p_j} = \underbrace{-p_j \frac{1}{p_j}}_{=-1} - \ln p_j + \lambda_0 + \lambda_1 j^2 = 0 \quad \text{for } j = 1, 2, 3 \quad (\text{A.35})$$

and thus

$$p_j = e^{-1+\lambda_0+\lambda_1 j^2} = e^{-1+\lambda_0} e^{\lambda_1 j^2} = e^{-1+\lambda_0} z^{j^2} \quad (\text{A.36})$$

where  $z = e^{\lambda_1}$ . The Lagrange multipliers are calculated from the constraints

$$1 = \sum_{j=1}^3 p_j = e^{-1+\lambda_0} (z + z^4 + z^9) \quad (\text{A.37})$$

$$E_{\text{mean}} = \sum_{j=1}^3 j^2 p_j = e^{-1+\lambda_0} (z + 4z^4 + 9z^9) \quad (\text{A.38})$$

Elimination of  $e^{-1+\lambda_0}$  between the two constraints yields an equation for  $z$

$$z + z^4 + z^9 - \frac{z + 4z^4 + 9z^9}{E_{\text{mean}}} = 0 \quad (\text{A.39})$$

The trivial solution  $z = 0$  is useless in the current context. Division by  $z \neq 0$  leads to

$$1 + z^3 + z^8 - \frac{1 + 4z^3 + 9z^8}{E_{\text{mean}}} = 0 \quad (\text{A.40})$$

that has to be solved numerically. For  $E = 4$  one obtains  $z = 0.9381$ ,  $\lambda_0 = 0.1778$ ,  $\lambda_1 = -0.0639$ , and  $p_1 = 0.4123$ ,  $p_2 = 0.3404$ ,  $p_3 = 0.2473$ . R code: [MaxEnt3kinEnergy.R](#)

**Exercise 59 Probabilities as function of mean energy**

The mean energy  $E_{\text{mean}}$  can vary between 1 and 9. Calculate and plot the probabilities  $p_j$ ,  $j = 1, 2, 3$  over this energy range.

## A.8 Derivation of the normal PDF by MaxEnt

The normal PDF can be derived as follows. One has to take into account three constraints, namely normalization, known mean value, and known variance:

$$\int_{-\infty}^{+\infty} p(x) dx = 1 \quad (\text{normalization}) \quad (\text{A.41})$$

$$\int_{-\infty}^{+\infty} x p(x) dx = \mu \quad (\text{mean}) \quad (\text{A.42})$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx = \sigma^2 \quad (\text{variance}) \quad (\text{A.43})$$

According to the Principle of Maximum Entropy, the PDF  $p(x)$  reads

$$p(x) = \exp \left[ \lambda_0 - \lambda_1 x - \lambda_2 (x - \mu)^2 \right] \quad (\text{A.44})$$

where  $\lambda_0$ ,  $\lambda_1$ , and  $\lambda_2$  are Lagrange multipliers (unknown constants at the moment) which are uniquely determined by the constraints given above. The Lagrange multipliers  $\lambda_0$ ,  $\lambda_1$ , and  $\lambda_2$  have to be calculated from the constraints. Insertion of  $p(x)$  (Eq. A.44) into Eqs. (A.41–A.43) leads to

$$\int_{-\infty}^{+\infty} \exp \left[ \lambda_0 - \lambda_1 x - \lambda_2 (x - \mu)^2 \right] dx = 1 \quad (\text{A.45})$$

$$\int_{-\infty}^{+\infty} x \exp \left[ \lambda_0 - \lambda_1 x - \lambda_2 (x - \mu)^2 \right] dx = \mu \quad (\text{A.46})$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 \exp \left[ \lambda_0 - \lambda_1 x - \lambda_2 (x - \mu)^2 \right] dx = \sigma^2 \quad (\text{A.47})$$

We will transform the integrands such that we can apply the following definite integral

$$\int_0^\infty x^n e^{-ax^2} dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2a^{(n+1)/2}} \quad \text{for } a > 0, \quad n > -1 \quad (\text{A.48})$$

From this integral we immediately derive

$$\int_{-\infty}^{+\infty} x^n e^{-ax^2} dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{a^{(n+1)/2}} \quad \text{for } a > 0, \quad n > -1, \quad n \text{ even}; \quad = 0 \text{ otherwise} \quad (\text{A.49})$$

One can simplify the calculation by shifting the distribution such that the mean vanishes:  $\mu = 0$ .

We will start with the second constraint

$$\mu = 0 = \int_{-\infty}^{+\infty} x \exp \left[ \lambda_0 - \lambda_1 x - \lambda_2 x^2 \right] dx \quad (\text{A.50})$$

$$= \exp \left[ \lambda_0 + \frac{\lambda_1^2}{4\lambda_2} \right] \int_{-\infty}^{+\infty} \left( y + \frac{\lambda_1}{2\lambda_2} \right) \exp \left[ -\lambda_2 y^2 \right] dy \quad (\text{A.51})$$

$$= \exp \left[ \lambda_0 + \frac{\lambda_1^2}{4\lambda_2} \right] \frac{\lambda_1}{2\lambda_2} \frac{\pi^{1/2}}{\lambda_2^{1/2}} \quad (\text{A.52})$$

where we applied the substitution

$$y = x - \frac{\lambda_1}{2\lambda_2}, \quad dx = dy \quad (\text{A.53})$$

which will be used also below. The second constraint can be fulfilled only when  $\lambda_1 = 0$ . This fact will simplify further calculations.

First constraint (already using  $\lambda_1 = 0$ ):

$$1 = \int_{-\infty}^{+\infty} \exp [\lambda_0 - \lambda_2 x^2] dx = e^{\lambda_0} \int_{-\infty}^{+\infty} \exp [-\lambda_2 x^2] dx \quad (\text{A.54})$$

$$= e^{\lambda_0} \frac{\Gamma(\frac{1}{2})}{\lambda_2^{1/2}} = e^{\lambda_0} \frac{\pi^{1/2}}{\lambda_2^{1/2}} \quad (\text{A.55})$$

Third constraint (already using  $\lambda_1 = 0$ ):

$$\sigma^2 = \int_{-\infty}^{+\infty} x^2 \exp [\lambda_0 - \lambda_2 x^2] dx = e^{\lambda_0} \int_{-\infty}^{+\infty} x^2 \exp [-\lambda_2 x^2] dx \quad (\text{A.56})$$

$$= e^{\lambda_0} \frac{\Gamma(\frac{3}{2})}{\lambda_2^{3/2}} = e^{\lambda_0} \frac{\pi^{1/2}}{2\lambda_2^{3/2}} \quad (\text{A.57})$$

Elimination of  $e^{\lambda_0}$  between the 1. and 3. constraint leads to

$$\frac{1}{\sigma^2} = \frac{e^{\lambda_0} \frac{\pi^{1/2}}{\lambda_2^{1/2}}}{e^{\lambda_0} \frac{\pi^{1/2}}{2\lambda_2^{3/2}}} = 2\lambda_2 \quad (\text{A.58})$$

or

$$\lambda_2 = \frac{1}{2\sigma^2} \quad (\text{A.59})$$

as it should for the normal distribution. Substituting  $\lambda_2$  in the 1. constraint leads to

$$e^{\lambda_0} = \frac{1}{\sqrt{2\pi\sigma^2}} \quad (\text{A.60})$$

Finally, one obtains the normal distribution for  $\mu = 0$

$$p(x) = \exp [\lambda_0 - \lambda_1 x - \lambda_2 x^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{x^2}{2\sigma^2} \right] \quad (\text{A.61})$$

**Exercise 60** Derive the MaxEnt distribution over the range  $[a, b]$  whereby the normalization is the only constraint.

**Exercise 61** MaxEnt distribution with fixed mean and expectation of  $\ln x$

Derive the maximum entropy distribution over the range  $[0, \infty)$  under the constraints of given expectations for the mean and for  $\ln x$ , i.e.  $E(x) = \mu = \alpha_1$   $E(\ln(x)) = \alpha_2$ .

Hints: (1) You can apply the functional equation for the  $\Gamma$ -function:

$$\Gamma(x+1) = x \Gamma(x) \quad (\text{A.62})$$

(2) Gradshteyn & Ryzhik (2014, integral 4.352.1):

$$\int_0^\infty x^{\nu-1} e^{-\mu x} \ln x dx = \frac{\Gamma(\nu)}{\mu^\nu} [\psi(\nu) - \ln \mu] \quad \operatorname{Re} \mu > 0, \quad \operatorname{Re} \nu > 0 \quad (\text{A.63})$$

where  $\psi(\nu) = \frac{d}{dx} \ln \Gamma(x)$  is the logarithmic factorial function.

**Exercise 62** MaxEnt distribution with constraint  $E(\ln(x)) = \frac{1}{\alpha} + \ln(x_m) = \beta$

Derive the maximum entropy distribution over the range  $[x_m, \infty)$  under the constraint  $E(\ln(x)) = \frac{1}{\alpha} + \ln(x_m)$ .

## A.9 Relative entropy, cross-entropy, directed divergence (\*)

The relative entropy<sup>2</sup> is defined as

$$S_r(q, p) = - \int q(x) \log \left( \frac{q(x)}{p(x)} \right) dx. \quad (\text{A.64})$$

where  $q(x)$  and  $p(x)$  are PDFs and  $p(x)$  is often a prior. It was proposed by Kullback & Leibler (1951; compare also Kullback, 1959) under the name 'directed divergence' with a different sign. Other names of the same concept are cross-entropy (Good, 1963) and weight of evidence. We use here the definition with the negative sign because the probability density  $q(x)$  will be derived by *maximizing*  $S_r(q, p)$  for a given prior distribution  $p(x)$  and constraints for  $q(x)$  (the constraints will be coupled to  $S_r$  by Lagrange multipliers). This procedure is called the [Principle of Maximum Relative Entropy \(MaxRelEnt\)](#); for the definition with the positive sign one has to *minimize* the directed divergence. Please note that  $S_r(q, p)$  is not symmetric in its arguments, i.e. it matters what's the prior and what's the searched for probability distribution. For (discrete) probability distributions the relative entropy reads

$$S_r(q, p) = - \sum_i q_i \log \left( \frac{q_i}{p_i} \right). \quad (\text{A.65})$$

### Music

[Stand by me](#)

written by Leibler-Stoller-King, performed by Ben E. King (1961), John Lennon (1975)

<https://www.youtube.com/watch?v=hwZNL7QVjE>

<https://www.youtube.com/watch?v=ubg7oAYP5aQ>

### A.9.1 Equilibrium distributions for the D2Q9 lattice Boltzmann model (\*)

The following example is based on Section 5.2.2 in Wolf-Gladrow (2000). The goal is to calculate the equilibrium distributions  $F_i^{(0)}$  as functions of mass density  $\rho$  and momentum  $j$  for the lattice Boltzmann model D2Q9 (Fig. A.12). From other sources (for details compare Wolf-Gladrow, 2000) one knows a good prior, namely the equilibrium distribution for a fluid at rest:

$$\mathbf{p} = \{p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8\} \quad (\text{A.66})$$

$$= \{W_0, W_1, W_1, W_1, W_1, W_2, W_2, W_2, W_2\} \quad (\text{A.67})$$

$$= \left\{ \frac{4}{9}\rho_0, \frac{1}{9}\rho_0, \frac{1}{9}\rho_0, \frac{1}{9}\rho_0, \frac{1}{9}\rho_0, \frac{1}{36}\rho_0, \frac{1}{36}\rho_0, \frac{1}{36}\rho_0, \frac{1}{36}\rho_0 \right\} \quad (\text{A.68})$$

What we are looking for are the equilibrium distributions of a non-resting/moving fluid. The constraints (normalization, conservation of mass and momentum) are, however, not sufficient to assign all 9 probabilities. Therefore we will apply MaxRelEnt to derive the probability distribution for given momentum  $j$ .

Koelman (1991) defines the *relative entropy* density by

$$S(\rho, j) := - \frac{k}{m} \sum_i F_i^{(0)}(\rho, j) \ln \frac{F_i^{(0)}(\rho, j)}{W_i}. \quad (\text{A.69})$$

The weighting by  $1/W_i$  in the logarithmic factor will lead to equilibrium distributions of the form  $F_i^{(0)} = W_i e^{h(\rho, j, c_i)}$  and implies  $S = 0$  for  $F_i^{(0)} = W_i$ . The  $W_i$  are given by

$$W_0/\rho_0 = \frac{4}{9} \quad (\text{A.70})$$

---

<sup>2</sup>The notation 'relative entropy' was introduced by Wehrl (1978).

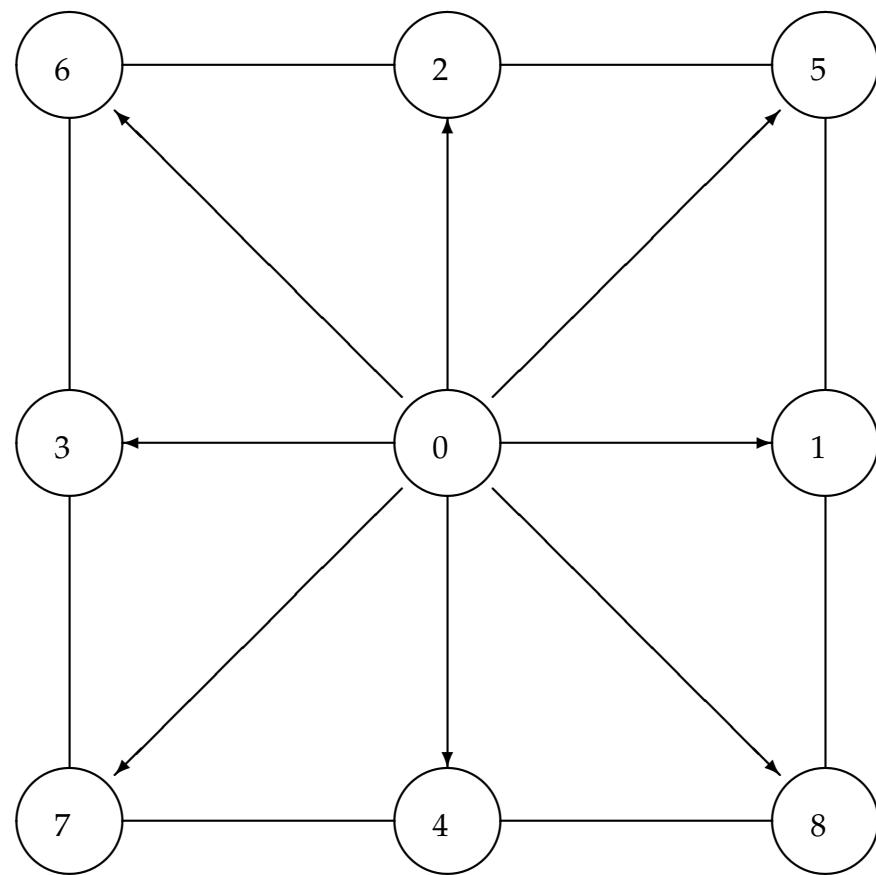


Figure A.12: D2Q9 lattice

$$W_1/\rho_0 = \frac{1}{9} \quad (A.71)$$

$$W_2/\rho_0 = \frac{1}{36} \quad (A.72)$$

The equilibrium distributions  $F_i^{(0)}$  will be calculated by maximizing the relative entropy for given constraints which for the case under consideration are the mass and momentum density

$$\begin{aligned} \rho(\rho, j) &= \sum_i F_i^{(0)}(\rho, j) \\ j(\rho, j) &= \sum_i c_i F_i^{(0)}(\rho, j) \end{aligned}$$

where the  $c_i$  are the grid velocities

$$\begin{aligned} c_0 &= (0, 0), & c_1 &= (1, 0), & c_2 &= (0, 1), & c_3 &= (-1, 0), & c_4 &= (0, -1), \\ c_5 &= (1, 1), & c_6 &= (-1, 1), & c_7 &= (-1, -1), & c_8 &= (1, -1). \end{aligned} \quad (A.73)$$

The functional

$$\begin{aligned} \hat{S} &:= S + \tilde{A}\rho + \tilde{\mathbf{B}} \cdot \mathbf{j} \\ &= -\frac{k}{m} \sum_i F_i^{(0)}(\rho, j) \ln \frac{F_i^{(0)}(\rho, j)}{W_i} + \tilde{A} \sum_i F_i^{(0)}(\rho, j) + \tilde{\mathbf{B}} \sum_i c_i F_i^{(0)}(\rho, j) \end{aligned}$$

encompasses the constraints coupled by Lagrange multipliers  $\tilde{A}$  und  $\tilde{\mathbf{B}}$ . The necessary conditions for an extremum of  $\hat{S}$  read

$$\forall i : \frac{\partial \hat{S}}{\partial F_i^{(0)}} = -\frac{k}{m} \left[ \ln \frac{F_i^{(0)}}{W_i} + 1 \right] + \tilde{A} + \tilde{\mathbf{B}} \cdot \mathbf{c}_i = 0. \quad (A.74)$$

The solutions of (A.74) are of the form

$$F_i^{(0)} = W_i e^A(\rho, j) + \mathbf{B}(\rho, j) \cdot \mathbf{c}_i$$

with

$$A = \frac{m}{k} \tilde{A} - 1, \quad \text{and} \quad \mathbf{B} = \frac{m}{k} \tilde{\mathbf{B}}.$$

$A$  and  $\mathbf{B}$  can be determined by Taylor series expansions of  $F_i^{(0)}$  around  $j = 0$ . Because of the symmetry of the D2Q9 lattice the ansatz

$$\begin{aligned} A(\rho, j) &= A_0(\rho) + A_2(\rho) j^2 + \mathcal{O}(j^4) \\ \mathbf{B}(\rho, j) &= B_1(\rho) j + \mathcal{O}(j^3). \end{aligned}$$

is sufficient. The expansion of  $F_i^{(0)}$  around  $j = 0$  reads

$$\begin{aligned} \frac{\partial F_i^{(0)}}{\partial j_\alpha} &= (2A_2 j_\alpha + B_1 c_{i\alpha}) F_i^{(0)} \\ &\rightarrow B_1 c_{i\alpha} W_i e^{A_0} \quad \text{at } j = 0 \\ \frac{\partial^2 F_i^{(0)}}{\partial j_\alpha^2} &= [(2A_2 j_\alpha + B_1 c_{i\alpha})^2 + 2A_2] F_i^{(0)} \\ &\rightarrow (B_1^2 c_{i\alpha}^2 + 2A_2) W_i e^{A_0} \quad \text{at } j = 0 \\ \frac{\partial^2 F_i^{(0)}}{\partial j_\alpha \partial j_\beta} &= (2A_2 j_\alpha + B_1 c_{i\alpha})(2A_2 j_\beta + B_1 c_{i\beta}) F_i^{(0)} \\ &\rightarrow B_1^2 c_{i\alpha} c_{i\beta} W_i e^{A_0} \quad \text{at } j = 0 \end{aligned}$$

and finally up to terms of second order in  $j$

$$F_i^{(0)} = W_i e^{A_0} \left\{ 1 + B_1 c_i \cdot j + \frac{B_1^2}{2} (c_i \cdot j)^2 + A_2 j^2 \right\}. \quad (\text{A.75})$$

Now the definitions of  $\rho$  and  $j$  will be exploited to calculate the unknown coefficients  $A_0(\rho)$ ,  $A_2(\rho)$  and  $B_1(\rho)$ :

$$\sum_i F_i^{(0)} = \rho = e^{A_0} \left\{ \rho_0 + \frac{B_1^2}{2} \rho_0 \frac{k_B T}{m} j^2 + \rho_0 A_2 j^2 \right\} \quad (\text{A.76})$$

$$\sum_i c_i F_i^{(0)} = j = e^{A_0} B_1 \rho_0 \frac{k_B T}{m} j. \quad (\text{A.77})$$

The vector equation (A.77) reduces to a scalar constraint

$$B_1 = \frac{1}{e^{A_0}} \frac{m}{\rho_0 k_B T},$$

i.e. an auxiliary condition is required to solve for all three unknowns ( $A_0$ ,  $A_2$ ,  $B_1$ ). Solving Eq. (A.76) for  $A_2$  yields

$$A_2 = -\frac{B_1^2}{2} \frac{k_B T}{m} + \underbrace{\frac{1}{j^2} \left( \frac{\rho}{\rho_0 e^{A_0}} - 1 \right)}_{(*)}.$$

To obtain  $A_2$  independent of  $j$  (as implied by the ansatz) the expression  $(*)$  must vanish. This is the third constraint looked for. It immediately follows that

$$e^{A_0} = \frac{\rho}{\rho_0}, \quad B_1 = \frac{m}{\rho k_B T}, \quad \text{and} \quad A_2 = -\frac{1}{2\rho^2} \frac{m}{k_B T}.$$

Insertion into (A.75) finally yields the equilibrium distributions

$$F_i^{(0)}(\rho, j) = \frac{W_i}{\rho_0} \left\{ \rho + \frac{m}{k_B T} c_i \cdot j + \frac{m}{2\rho k_B T} \left[ \frac{m}{k_B T} (c_i \cdot j)^2 - j^2 \right] \right\}. \quad (\text{A.78})$$

Koelman (1991) used  $1/\rho_0$  instead of  $1/\rho$  in the coefficient of the third term which is a good approximation for small Mach numbers.

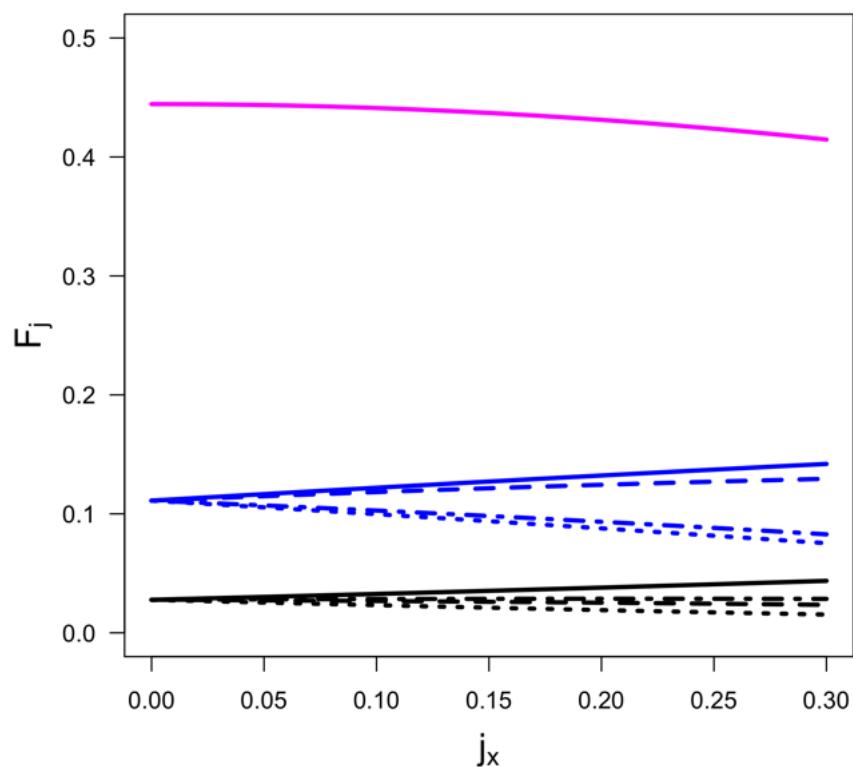


Figure A.13: The equilibrium distributions for the D2Q9 lattice Boltzmann model as a function of the momentum  $j_x$  ( $j_y$  varies with  $j_x$  according to  $j_y = 0.7 j_x$ ) at  $\rho = \rho_0 = 1$  and  $\frac{m}{k_B T} = 1$ . [RelEntropyD2Q9.R](#)

## A.10 Probability for '4' in next throw of a die (\*)

Cox (1946) posed the following problem: "... let it be supposed that there are two dice, both dynamically symmetric, but one of them defectively marked, having two faces instead of one stamped with four dots. Then for either of these dice there is a stable probability of throwing a four, equal to  $\frac{1}{6}$  if it is the true die and to  $\frac{1}{3}$  if it is the defective one. Suppose one die of the pair is picked up at random and, without being examined, is tossed  $N$  times. If a four turns up on  $n$  of these throws, what is the probability of a four on the next throw?" As solution to his problem Cox (1946) gave a formula, however, he did not derive it ("The theorems available are enough to give the result . . .") and he did not discuss any numerical examples.

In this section we will derive the analytical solution from the basic rules of probability and by application of marginalization, then discuss a numerical example, and, finally, confirm the analytical solution by a Monte-Carlo simulation.

### A.10.1 Derive analytical solution

The analytical solution reads (notation<sup>3</sup>:  $p('4' \text{ in next throw} | k \text{ times '4' in } N \text{ throws}) \equiv p('4' | k(N))$ ):

$$p('4' | k(N)) = \frac{p('4' \text{ and } k(N))}{p(k(N))} \quad (\text{A.79})$$

$$= \frac{\frac{1}{3} \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6} \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}}{\frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}} \quad (\text{A.80})$$

$$= \frac{\frac{1}{3^{k+1}} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6^{k+1}} \left(1 - \frac{1}{6}\right)^{N-k}}{\frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}} \quad (\text{A.81})$$

Derivation of formula (in order to keep the notation as simple as possible, I will leave out the background information  $I$ ):

What we know:

(1) Probability for using die #1 and die #2 are equal to each other and thus both = 1/2

$$p(\text{die } \#1) = 0.5 = p(\text{die } \#2) \quad (\text{A.82})$$

(2) Probability to obtain  $k$  times '4' in  $N$  trials (short notation ' $k(N)$ ') using die #1 is given by the binomial distribution with  $p_1 = 1/3$  for success in a single trial:

$$p(k(N) | \text{die } \#1) = \frac{N!}{k!(N-k)!} p_1^k (1-p_1)^{N-k} = \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} \quad (\text{A.83})$$

(3) Probability to obtain  $k$  times '4' in  $N$  trials using die #2 is given by the binomial distribution with  $p_2 = 1/6$  for success in a single trial:

$$p(k(N) | \text{die } \#1) = \frac{N!}{k!(N-k)!} p_2^k (1-p_2)^{N-k} = \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k} \quad (\text{A.84})$$

Application of the product rule  $p(A | B) \cdot p(B) = p(A \text{ and } B)$  yields

$$p(k(N) \text{ and die } \#1) = \frac{1}{2} \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} \quad (\text{A.85})$$

---

<sup>3</sup>We will use  $k$  (instead of  $n$  used by Cox) for the number of successes in order to avoid confusion with  $N$  = number of throws..

and

$$p(k(N) \text{ and die } \#2) = \frac{1}{2} \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k} \quad (\text{A.86})$$

The probability for  $k$  times '4' in  $N$  trials is given by

$$p(k(N)) = p(k(N) \text{ and (die } \#1 \text{ or die } \#2)) \quad (\text{A.87})$$

$$= p(k(N) \text{ and die } \#1) + p(k(N) \text{ and die } \#2) \quad (\text{A.88})$$

$$= \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} \quad (\text{A.89})$$

$$+ \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k} \quad (\text{A.90})$$

where we first applied **marginalization**, i.e. we combined (by 'and') a proposition (here: '( $k(N)$ )') with something that is true for sure (here: die #1 or die #2), and then used the generalized sum rule for two propositions (here: die #1, die #2) that are mutually exclusive.

Finally, we like to know the probability for '4' in the next trial given ' $k(N)$ ', i.e.  $p('4' | k(N))$ . The product rule tells us:

$$p('4' \text{ and } k(N)) = p('4' | k(N)) \cdot p(kN) \quad (\text{A.91})$$

or

$$p('4' | k(N)) = p('4' \text{ and } k(N)) / p(kN) \quad (\text{A.92})$$

The probability  $p(kN)$  has been assigned already and thus what remains is  $p('4' \text{ and } k(N))$ . We will apply again marginalization and the product and generalized sum rules:

$$p('4' \text{ and } k(N)) = p((('4' \text{ and (die } \#1 \text{ or die } \#2)) \text{ and } (k(N) \text{ and (die } \#1 \text{ or die } \#2))) \quad (\text{A.93})$$

$$= p \left( ('4' \text{ and die } \#1 \text{ and } kN \text{ and die } \#1) \text{ or } \underbrace{('4' \text{ and die } \#1 \text{ and } kN \text{ and die } \#2)}_{\text{impossible}} \right. \\ \left. \text{or } \underbrace{('4' \text{ and die } \#2 \text{ and } kN \text{ and die } \#1)}_{\text{impossible}} \text{ or } ('4' \text{ and die } \#2 \text{ and } kN \text{ and die } \#2) \right)$$

$$= p('4' \text{ and die } \#1 \text{ and } kN) + p('4' \text{ and die } \#2 \text{ and } kN) \quad (\text{A.94})$$

$$= p_1 \frac{N!}{k!(N-k)!} p_1^k (1-p_1)^{N-k} + p_2 \frac{N!}{k!(N-k)!} p_2^k (1-p_2)^{N-k} \quad (\text{A.95})$$

$$= \frac{1}{3} \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6} \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k} \quad (\text{A.96})$$

and thus

$$p('4' | k(N)) = \frac{p('4' \text{ and } k(N))}{p(kN)} \quad (\text{A.97})$$

$$= \frac{\frac{1}{3} \frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6} \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}}{\frac{N!}{k!(N-k)!} \frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{N!}{k!(N-k)!} \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}} \quad (\text{A.98})$$

$$= \frac{\frac{1}{3^{k+1}} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6^{k+1}} \left(1 - \frac{1}{6}\right)^{N-k}}{\frac{1}{3^k} \left(1 - \frac{1}{3}\right)^{N-k} + \frac{1}{6^k} \left(1 - \frac{1}{6}\right)^{N-k}} \quad (\text{A.99})$$

**q.e.d**

### A.10.2 A numerical example

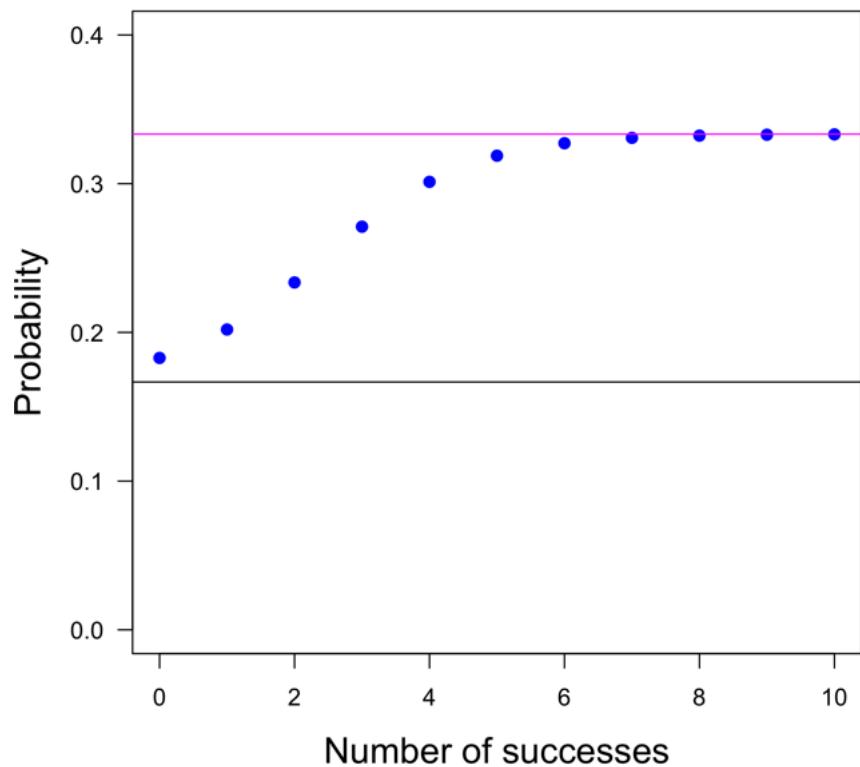


Figure A.14: Probabilities (blue dots) for '4' in the next trial given  $k = n$  successes in  $N$  trials with die #1 or die #2. The resulting probabilities lie between the ones for die #1 ( $p_1 = 1/3$ ; magenta line) and die #2 ( $p_2 = 1/6$ ; black line) [Cox46Next4.R](#)

**Exercise 63 Bags filled with both red poker chips and white poker ships**

Michael Lewis (*The Undoing Project*, 2017, p. 142) writes: 'The specific study Amos [Tversky] described was about how people, in their decision making, responded to new information. As Amos told it, the psychologists had brought people in and presented them with two book bags filled with poker chips. Each bag contained both red poker chips and white poker chips. In one of the bags, 75 percent of the chips were white and 25 percent were red, in the other bag, 75 percent of the chips were red and 25 percent were white. The subject picked one of the bags at random and, without glancing inside the bag, began to pull chips out of it, one at a time. After extracting each chip, he'd give the psychologists his best guess of the odds that the bag he was holding was filled with mostly red, or mostly white, chips.'

The odds for proposition A,  $O(A)$ , is given by the ratio of the probability for A divided by the probability for not A,  $\bar{A}$ :

$$O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)} \quad (\text{A.100})$$

where for the last conversion we applied the sum rule of probabilities (Eq. 4.3).

'What is the probability that I am holding the bag of mostly red chips?' given the following observations (draws):

- (1) red in first draw (r);
- (2) red in first two draws (rr);
- (3) red in first three draws (rrr);
- (4) 2 times red and 1 times white in first three draws (rrw).

Calculate the odds from the probabilities.

**Music** Frank Zappa: *Fifty-Fifty* ('... I figure the odds be fifty-fifty I just might have some thing to say ...')

<https://www.youtube.com/watch?v=25ThICK0Fbw>

# Appendix B

## Random numbers

### B.1 Generation of random numbers from other distributions

#### B.1.1 Example: random numbers from the tent distribution

Goal: Generate random numbers that are distributed according to a given PDF  $q(y)$ . Example:  $q(y) = 4y$  for  $0 \leq y \leq 0.5$ ,  $q(y) = 4 - 4y$  for  $0.5 \leq y \leq 1$ , and zero otherwise (symmetric triangular or 'tent' distribution, Fig. B.1).

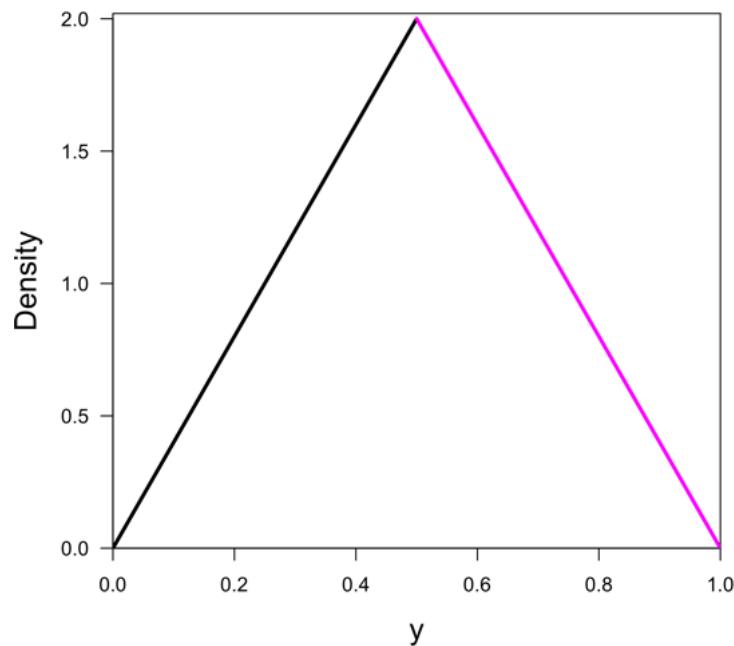


Figure B.1: The triangle or tent PDF [RandomTentPDF.R](#)

A coordinate transformation  $y(x)$  usually causes a change in the probability density function, i.e. the PDF  $q(y)$  expressed in the new coordinates  $y$  is different from the PDF  $r(x)$  expressed in the initial coordinates  $x$ , whereas the (infinitesimal) probabilities expressed in different coordinates are equal. Thus the **fundamental transformation law of probabilities** reads

$$|q(y) dy| = |r(x) dx| \quad (\text{B.1})$$

or (because probability densities are non-negative)

$$q(y) = r(x) \left| \frac{dx}{dy} \right| \quad (\text{B.2})$$

Let us assume that a method for generation of random numbers distributed uniformly between 0 and 1 is already available (function `runif()` in R), i.e. we know how to generate random numbers that follow the uniform PDF  $u(x) = 1$  for  $0 \leq x \leq 1$  and zero otherwise.  $u(x)$  is of course normalized, i.e.

$$\int_{-\infty}^{+\infty} u(x) dx = \int_0^1 u(x) dx = 1 \quad (\text{B.3})$$

Substitution of the uniform PDF  $u(x)$  for  $r(x)$  in Eq. (B.2) leads to

$$q(y) = \left| \frac{dx}{dy} \right|, \quad (\text{B.4})$$

which relates the PDF  $q(y)$  in the new coordinates  $y$  to the first derivative of the initial coordinates  $x$  with respect to the new coordinates  $y$ .

The back transformation  $x(y)$  can be easily calculated by separation of variables followed by integration (assuming  $dx > 0, dy > 0$ ):

$$\int dx = x = \int q(y') dy' = F(y) \quad (\text{B.5})$$

where  $F(y)$  is the cumulative distribution function (CDF) of  $q(y)$ . Thus

$$y(x) = F^{-1}(x), \quad (\text{B.6})$$

i.e. the transformation  $y(x)$  from  $x$  to  $y$  is the inverse function of the cumulative probability distribution of  $q(y)$ . If we know the transformation  $y(x)$  we can generate random numbers  $x$  that follow the uniform PDF  $u(x)$  and transform them to random numbers  $y(x)$  that follow the PDF  $q(y)$ .

Our example ('tent'): 1.) for  $0 \leq y \leq 0.5$

$$x = F(y) = 4 \int_0^y y' dy' = 2y^2 \quad (\text{B.7})$$

$\Rightarrow$

$$y(x) = \sqrt{\frac{x}{2}} \quad \text{for } 0 \leq x \leq \frac{1}{2} \quad (\text{B.8})$$

The formal solution  $y(x) = -\sqrt{x/2}$  yields negative values and is therefore discarded.

2.) for  $0.5 \leq y \leq 1$

$$F(y) = \frac{1}{2} + 4 \int_{1/2}^y (1 - y') dy' = \frac{1}{2} + 4y - 2 - 2y^2 + \frac{1}{2} = 4y - 1 - 2y^2 \quad (\text{B.9})$$

Inverse function: solve quadratic equations  $x = F(y) = \alpha y^2 + \beta y + \gamma$ :

$$\begin{aligned} y^2 - 2y &= -\frac{1}{2} - \frac{x}{2} \\ y^2 - 2y + 1 &= 1 - \frac{1}{2} - \frac{x}{2} \\ y_{1,2} &= 1 \pm \sqrt{\frac{1-x}{2}} \end{aligned}$$

$\Rightarrow$

$$y(x) = 1 - \sqrt{\frac{1-x}{2}} \quad \text{for} \quad \frac{1}{2} \leq x \leq 1 \quad (\text{B.10})$$

The formal solution  $y_1(x) = 1 + \sqrt{\frac{1-x}{2}}$  yields values above 1 and is therefore discarded.

$$y(x) = 0 \quad \text{for} \quad x < 0 \quad \text{or} \quad x > 1 \quad (\text{B.11})$$

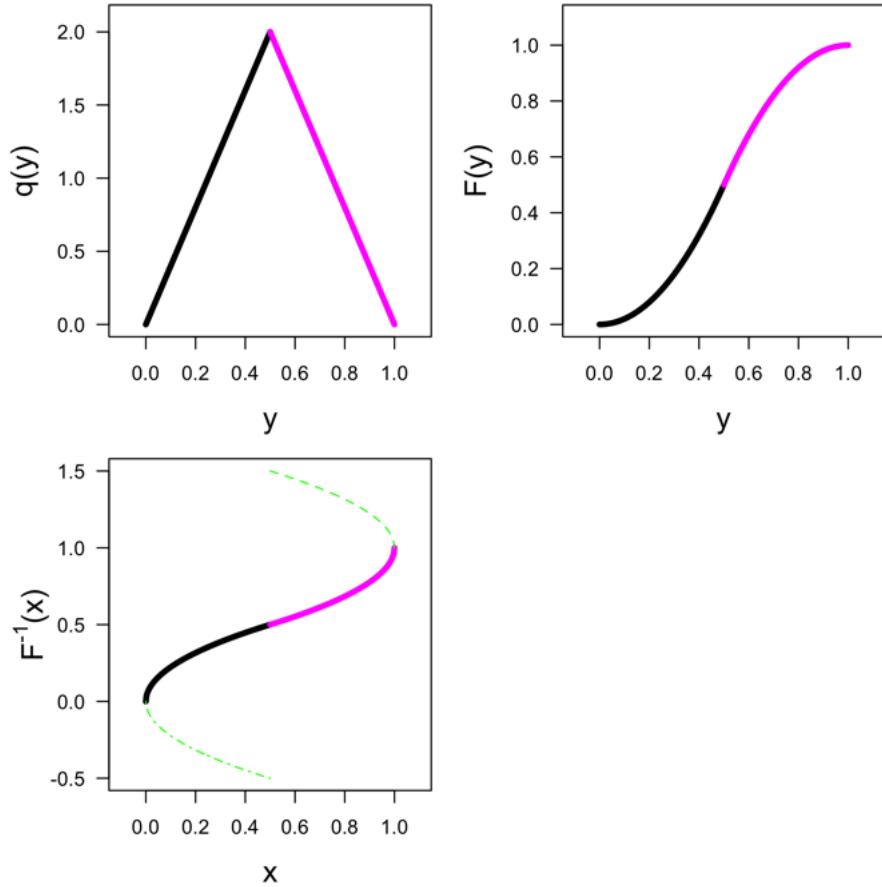


Figure B.2: The 'tent' distribution  $q(y)$  (upper left panel). The integral  $F(y)$  over the distribution  $q(y)$  is easy to calculate (upper right panel). The transformation  $y(x)$  is given by the inverse of  $F(y)$ :  $y(x) = F^{-1}(x)$ :  $y = (x/2)^{1/2}$  for  $0 \leq x \leq 0.5$  and  $y = 1 - [(1-x)/2]^{1/2}$  for  $0.5 \leq x \leq 1$ . Note that the formal solutions with negative values (dash-dotted line) or with values above 1 (dashed line) are discarded.

[RandomRNfromTentPDF.R](#)

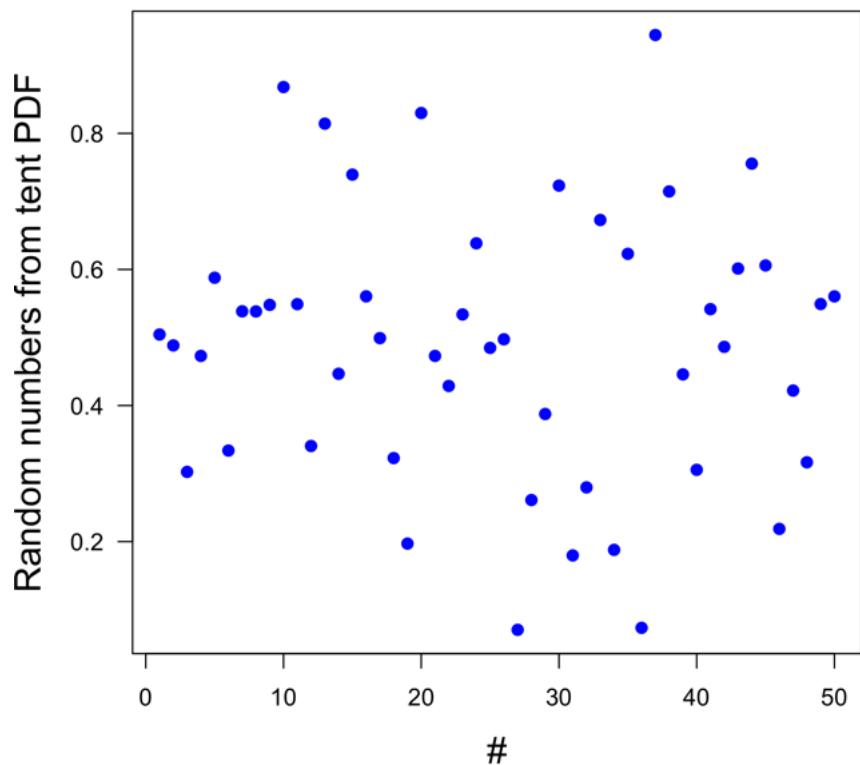


Figure B.3: 50 random numbers from the tent PDF. [Random50RNfromTentPDF.R](#)

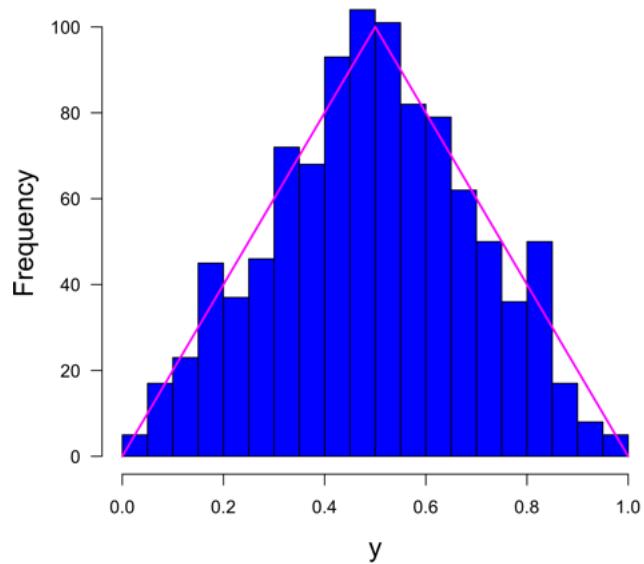


Figure B.4: A histogram of 1000 random numbers (blue bins) that follow the tent PDF (scaled, magenta).  
[RandomRNfromTentHist.R](#)

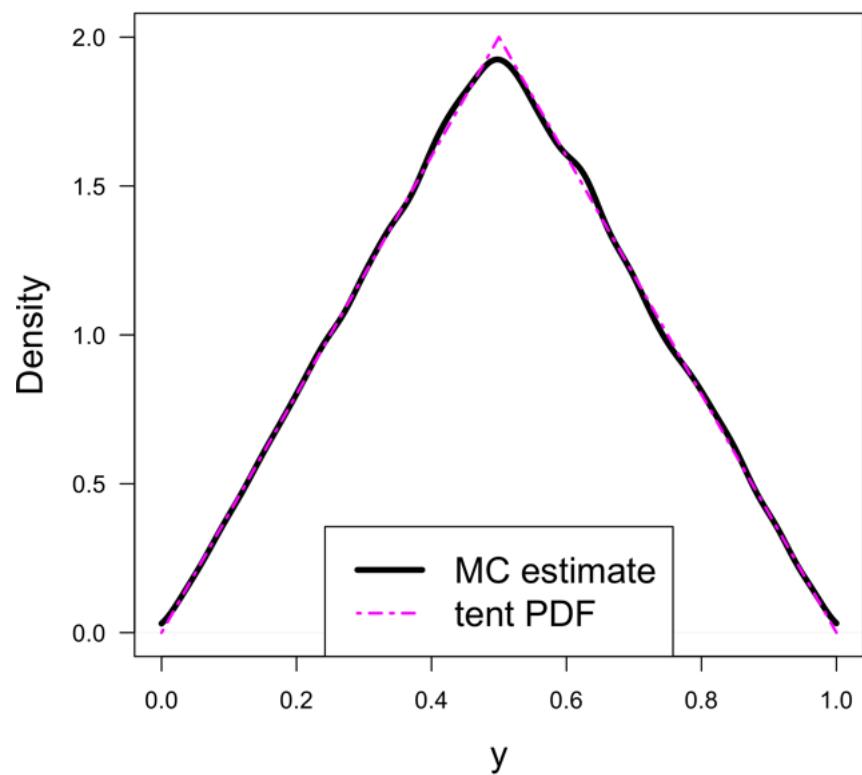


Figure B.5: Estimate of the tent density from  $10^5$  random samples (black solid line) and exact tent density (magenta dashed line). [RandomTentMCdensityEst.R](#)

## B.1.2 Random numbers from discrete distributions

In Monte Carlo simulations (marginal) PDFs often result from numerical integrations and are approximated by discrete distributions (compare, for example, Gelman et al., 2020, p. 76). Random numbers from discrete distributions can be derived as follows:

1. Calculate (an estimate of) the CDF by summing up the discrete values.
2. Generate random numbers from the uniform distribution on the interval [0,1] by calling the R routine `runif()`.
3. Invert the CDF numerically, i.e. find the CDF value that is closest to the random number from the uniform distribution.

**Example** Assume that we know the  $\beta$  PDF  $\text{Beta}(x; \alpha = 2, \beta = 5)$  (Fig. B.6, upper left panel) only on an equidistant grid covering the domain of definition<sup>1</sup> (here: [0,1]). In our example, the grid spacing is 0.01. The corresponding CDF can be approximated by summing up the discrete values  $y_i = \text{Beta}(x_i; \alpha = 2, \beta = 5)$  (Fig. B.6, upper right panel). The approximate CDF can be graphically (Fig. B.6, lower left panel) and numerically inverted. Finally, we generate  $M = 10^4$  random numbers from the uniform PDF and numerically invert the approximate CDF to obtain random numbers that follow the  $\beta$  distribution  $\text{Beta}(x; \alpha = 2, \beta = 5)$ . The density estimated from these random numbers is very close to the exact distribution  $\text{Beta}(x; \alpha = 2, \beta = 5)$  (Fig. B.6, lower right panel).

---

<sup>1</sup>For infinite domains, this is not possible. The grid should at least cover a large part of the domain of definition such that the integral/sum of the distribution is close to 1.

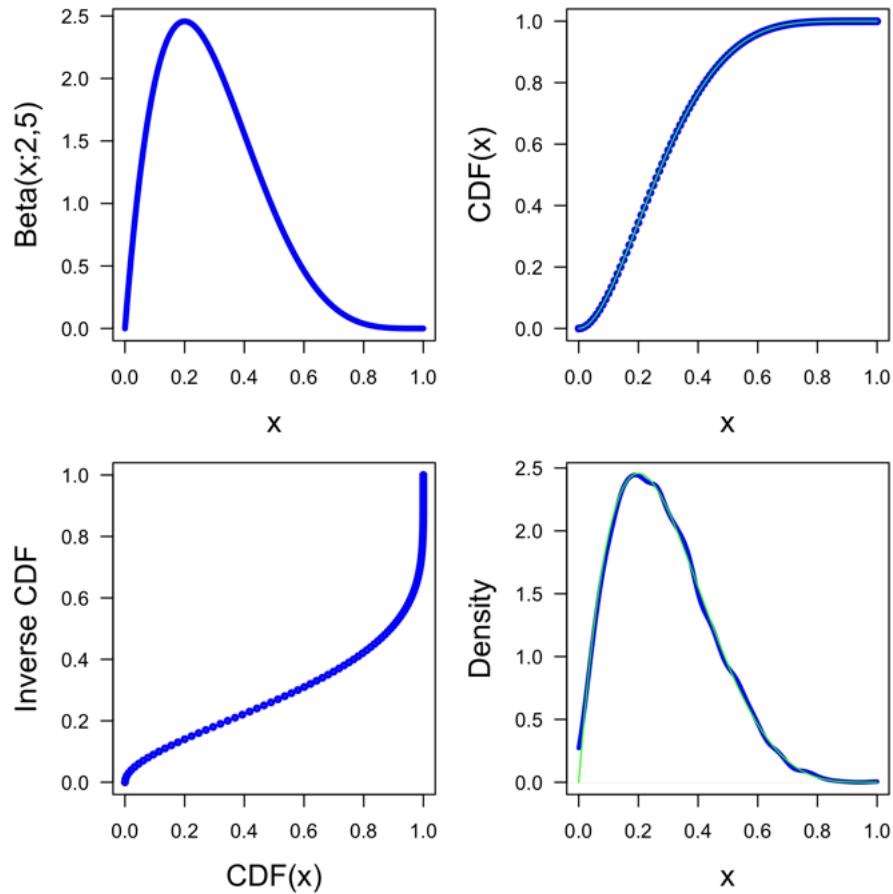


Figure B.6: Upper left panel: Density of the  $\beta$  PDF  $\text{Beta}(x; \alpha = 2, \beta = 5)$ . Upper right panel: CDF for  $\text{Beta}(x; \alpha = 2, \beta = 5)$ : for a grid spacing of  $\Delta x = 0.01$  the approximation (blue dots) compares quite well with the exact CDF (green line). Lower left panel: Inverse CDF for  $\text{Beta}(x; \alpha = 2, \beta = 5)$ : for a grid spacing of  $\Delta x = 0.01$  the approximation (blue dots) compares quite well with the exact CDF (green line). Lower right panel: The density estimated from random numbers (blue thick line) is very close the the exact distribution  $\text{Beta}(x; \alpha = 2, \beta = 5)$  (green thin line). [RandomBeta2x5.R](#)

# Appendix C

## PDs & PDFs (ref)

*The appendix for PDs & PDFs is largely meant for reference. In addition, some details are given to more mathematical inclined readers.*

### C.1 Most common univariate distributions and some of their relationships

Leemis (1986) wrote a nice little paper showing the most common univariate distributions and some of their relationships. Some of these relationships are shown in Figs. C.1 and C.2.

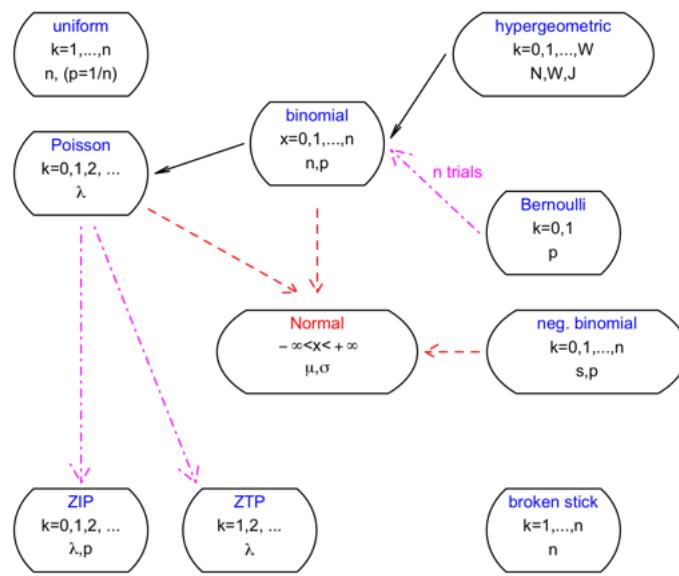


Figure C.1: Relationships between various PDs (Leemis, 1986, redrawn and modified). Relationships resulting from change of parameters are indicated by black arrows. Poisson distributions, for example, can be derived from binomial distributions in the limit  $n \rightarrow \infty$  and  $p \rightarrow 0$  while  $\lambda = n \cdot p$  stays constant. A change of parameters can not lead from a PD (with discrete variable  $k$ ) to a PDF (with continuous variable  $x$ ). However, normal distributions can form an envelop of PDs for certain parameter ranges (indicated by red arrows; for example, Poisson distributions with large mean rates  $\lambda$ ) and can even be used as approximations for PDs. Other relationships are indicated by magenta arrows. An example is the binomial distribution that results from  $n$  Bernoulli trials. R code: [PDsRelationships.R](#)

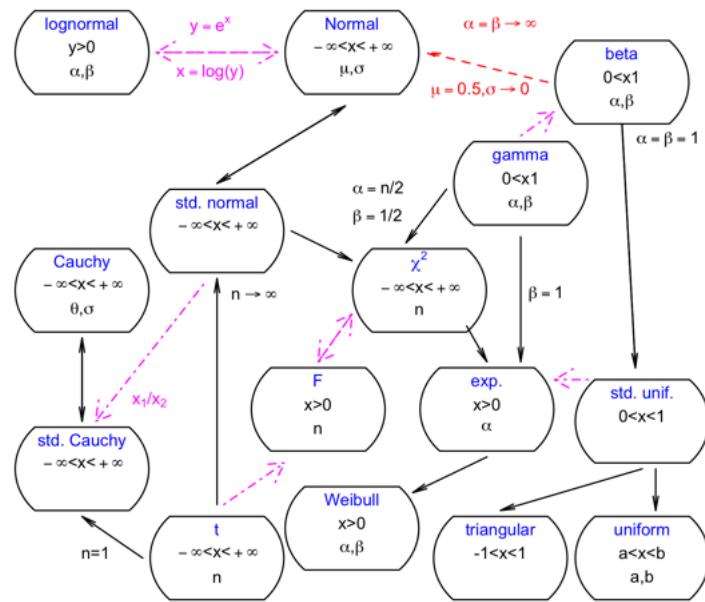


Figure C.2: Relationships between various PDFs (Leemis, 1986, redrawn and modified). The boxes contain the names (in blue), the variable range, and the parameters. The black errors indicate relations by change of parameters (note that the relations  $\alpha = n/2, \beta = 1/2$  for the relationship between the gamma and the  $\chi^2$  distributions is different from Leemis (1986) because Leemis used a different parameterization of the gamma distribution). The red arrows indicate approximate relationships (the beta distribution for large and equal parameters can be approximated by a normal distribution, the range of  $x$  is, however, still restricted by 0 and 1). The magenta arrows indicate more indirect relationships where variables from one distribution have to be transformed in order to follow a second distribution and sometimes vice versa. R code: [PDFsRelationships.R](#)

## C.2 Probability distributions (PDs)

This section presents more probability distributions (PDs) which are used in applications. It is meant for reference and not for reading in the first run.

### C.2.1 From the binomial to the Poisson distribution

The Poisson distribution can be derived from the binomial distribution in the limit of a small probability of success,  $p$ , and a large number of trials,  $n$ , where  $p \cdot n = \lambda$  is kept constant. Starting point is the binomial distribution (Eq. 6.32)

$$\text{Binom}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

Under the assumption that  $p \cdot n = \lambda$  for small  $p$  and large  $n$ ,  $k$  is usually small and one can derive the following identities or approximations:

(1)

$$\frac{n!}{k!(n-k)!} = \frac{n(n-1)\dots(n-k+1)}{k!} \approx \frac{n^k}{k!} \quad (\text{C.1})$$

(2)  $k$ th power of assumption:

$$p^k = \frac{\lambda^k}{n^k} \quad (\text{C.2})$$

(3)

$$(1-p)^{n-k} = \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \approx 1 \cdot e^{-\lambda} \quad (\text{C.3})$$

and thus

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \approx \frac{n^k}{k!} \frac{\lambda^k}{n^k} e^{-\lambda} = \frac{\lambda^k}{k!} e^{-\lambda} \quad (\text{C.4})$$

### C.2.2 Zero Inflated Poisson distribution

The Zero Inflated Poisson (ZIP) distribution is defined by

$$\text{ZIP}(k; \lambda, p) = p I(k=0) + (1-p) \text{Poisson}(k; \lambda) \quad k = 0, 1, 2, \dots \quad (\text{C.5})$$

where  $0 \leq p \leq 1$  is the zero-inflation parameter<sup>1</sup> and  $I()$  is the indicator or characteristic function with  $I(x_i = 0) = 1$  if  $x_i = 0$ ,  $I(x_i = 0) = 0$  if  $x_i \neq 0$ . The probability for  $k = 0$  is equal to  $p + (1-p)e^{-\lambda}$ ; the probabilities for  $k > 0$  are  $(1-p) \frac{\lambda^k}{k!} e^{-\lambda}$ .

ZIP has been proposed by Diane Lambert in 1992. ZIPs can be applied in ecology: "A common feature of ecological data sets is their tendency to contain many zero values." (Martin et al., 2005).

---

<sup>1</sup>For  $p = 0$  one obtains the Poisson distribution.

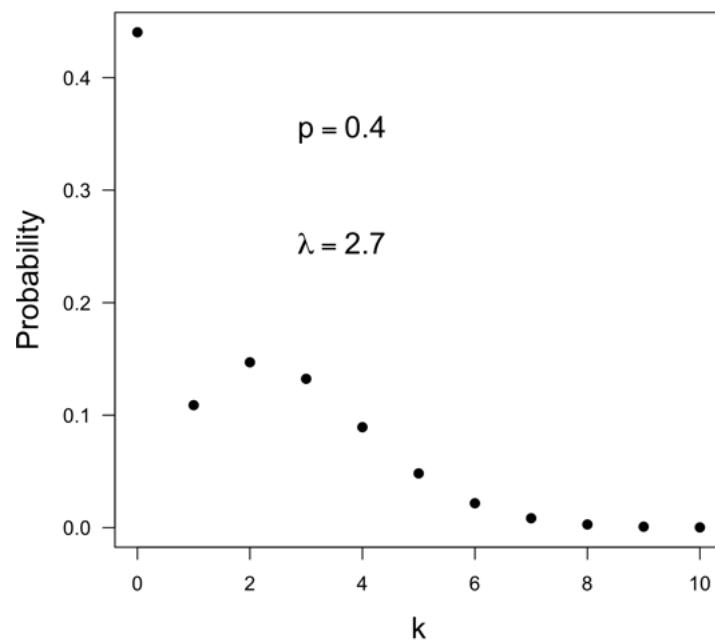


Figure C.3: The Zero Inflated Poisson (ZIP) distribution for the zero-inflation parameter  $p = 0.4$  and  $\lambda = 2.7$ .  
PDsPDFsPoissonZeroInflated.R

### C.2.3 Zero truncated Poisson distribution

The zero truncated Poisson distribution

$$\mathcal{P}(k; \lambda | k > 0) = \frac{\lambda^k}{k! (1 - e^{-\lambda})} e^{-\lambda} \quad k = 1, 2, \dots \quad (\text{C.6})$$

is applied, for example, in the analysis of count data when the case  $k = 0$  is excluded by experimental design or other circumstances (example: length of stay of patients in a hospital; Zuur et al., 2009). It is derived from the Poisson distribution (Section 6.3.3) by dropping the probability  $e^{-\lambda}$  for  $k = 0$ . Accordingly, the remaining probabilities sum up to  $1 - e^{-\lambda}$ . By dividing all remaining probabilities by  $1 - e^{-\lambda}$  one obtains again a proper (normalized to 1) probability distribution.

**Further reading:** Conway & Maxwell (1962) proposed other modifications of the Poisson distribution. These so-called COM-Poisson distributions can be used in generalized linear modeling (GLM) to analyze data when underdispersion occurs while applying Poisson distributions (Lord et al., 2010).

#### Exercise 64 Mean and variance of zero truncated Poisson distribution

*Derive formulas for the mean and variance of the zero truncated Poisson distribution.*

### C.2.4 Hypergeometric distribution: sampling without replacement

The hypergeometric distribution is a sibling of the binomial distribution. While the binomial distribution applies for the case of 'sampling with replacement' (a term derived from drawing balls from an urn and putting each back directly after drawing, i.e. the system – here: the amount and composition of balls in the urn – does not change during sampling), the hypergeometric distribution applies to cases of '[sampling without replacement](#)', i.e. the change of the system during sampling – here: reduced number of balls and changing composition with respect to color – is taken into account, which is important especially for small systems. It is not surprising that the hypergeometric distribution requires more parameters, namely 3, than the binomial distribution. Consider an urn that contains  $N$  balls that are identical in every aspect except that  $W$  are white and  $N - W$  are black, i.e. the probability to obtain a white ball in the first draw (called 'success') is  $p = W/N$ , however, the probability for the draw of a white ball will change depending on how many of what color have been already drawn. The third parameter is the total number of balls  $J$  taken out. Despite these many parameters and the changes in the system during sampling, the formula for hypergeometric distribution (Fig. C.7) looks relatively simple (for example, Jaynes, 2003):

$$\mathcal{H}(k; N, W, J) = \frac{\binom{W}{k} \binom{N-W}{J-k}}{\binom{N}{J}} = \frac{W! (N-W)! (N-J)! J!}{(W-k)! k! (N-W-J+k)! / J-k)! N!}. \quad (\text{C.7})$$

**Some properties of the hypergeometric distribution:**

**Mean:**  $\mu = \frac{JW}{N}$ ; **Variance:**  $\sigma^2 = \frac{JW}{N} \frac{(N-W)(N-J)}{N(N-1)}$

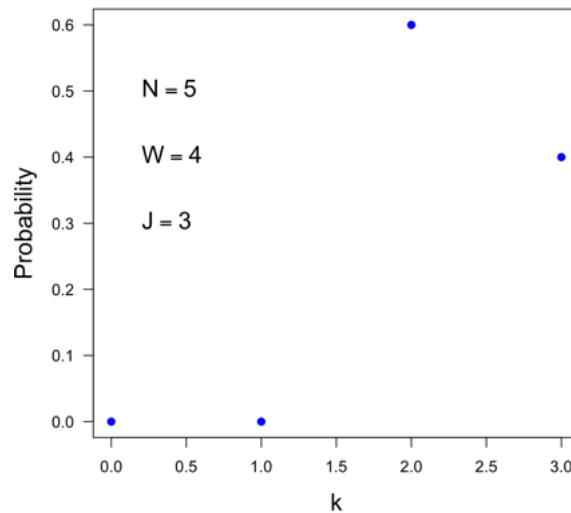


Figure C.4: The hypergeometric probability distribution (Eq. C.7);  $k$  is the number of white balls drawn from the urn. Initially, the urn contains 4 white balls and 1 black ball. We take out 3 balls, which could be (a) 2 white and 1 black or (b) 3 white and no black. Other combinations (2 black and 1 white, or 3 black and no white) are impossible and their corresponding probabilities are zero. The probabilities for (a) and (b) are 0.6 and 0.4, respectively. [B2-PDsPDFsHypergeoPDeexample.R](#)

### C.2.5 Geometric probability distribution

The distribution

$$\text{Geometric}(k; h) = (1 - h)^{k-1} h \quad k = 1, 2, 3, \dots \quad (\text{C.8})$$

is called the geometric probability distribution (Fig. C.5);  $0 < h < 1$  is the probability of success in a single trial and  $\text{Geometric}(k; h)$  is the probability for the first success occurring after  $k$  trials.

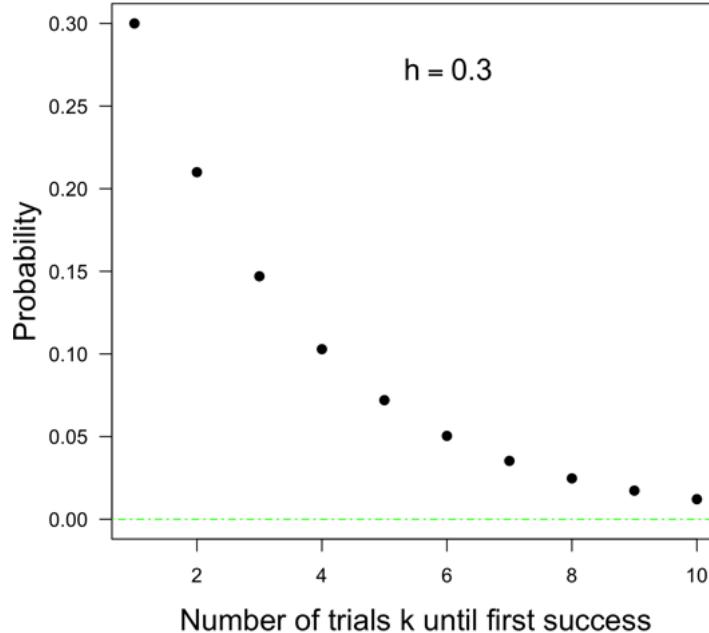


Figure C.5: Geometric probability distribution for the probability of success in a single trial  $h = 0.3$ .  
[B2-PDsPDFsGeometricPDexample.R](#)

#### Exercise 65 Prove normalization of geometric PD (\*)

Prove that the geometric probability distribution is normalized, i.e.

$$\sum_{k=1}^{\infty} \text{Geometric}(k; h) = 1 \quad (\text{C.9})$$

#### Exercise 66 CDF for geometric PD

Derive the cumulative distribution function (CDF) for the geometric probability distribution (PD). Plot the CDF for  $h = 0.3$ .

Hint: the sum of the first  $k$  terms of the geometric series is given by

$$h \sum_{j=0}^{k-1} q^j = h \frac{1 - q^k}{1 - q} \quad \text{for } q \neq 1 \quad (\text{C.10})$$

#### Exercise 67 How many students will pass their exams?

The test can be taken only twice. What percentage of students will pass for a given probability of success in a single trial  $h = 0.4$  ('success rate')? Do you expect a constant success rate?

### C.2.6 Broken stick distribution

The probability distribution (Fig. C.6)

$$\text{BrokenStick}(k; n) = \frac{1}{n} \sum_{j=k}^n \frac{1}{j} \quad (\text{C.11})$$

was termed 'broken stick' by MacArthur (1957). It has been applied in ecology (for example, Bennett, 1996) and for the interpretation of eigenvalues in principal component analysis (Legendre & Legendre, 1983).

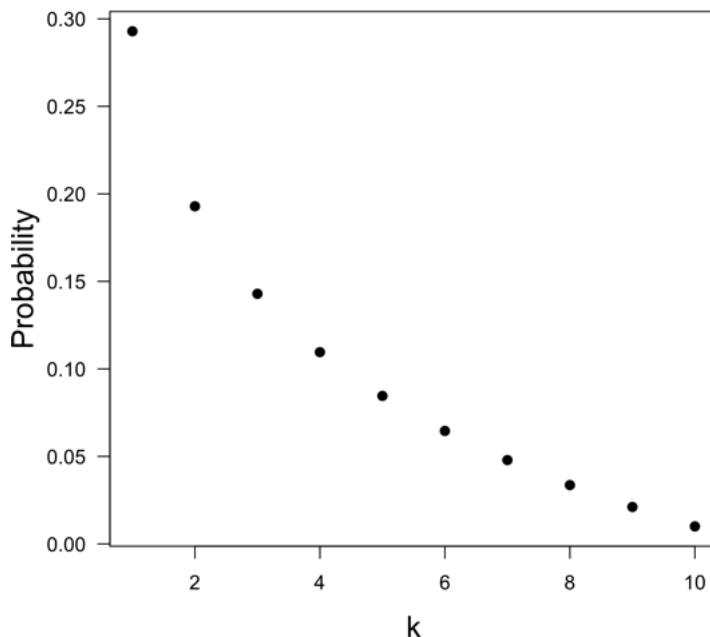


Figure C.6: The broken stick distribution for  $n = 10$ . [B2-PDsPDFsBrokenStickPDXexample.R](#)

#### Exercise 68 Broken stick versus geometric PD

Plot the broken stick PD for  $n = 10$ , the geometric PD for probability of success in single trial  $h = 0.2929$  for  $k = 1, 2, \dots, n$  trials, and the differences between these PDs.

### C.2.7 Negative binomial distribution

Instead of counting the successes in a fixed number of Bernoulli trials (= binomial distribution) we now ask 'What is the probability for  $k$  failures before the  $s$ th success?'. The negative binomial distribution (Casella & Berger, 2002, Eq. 3.2.10) is given by

$$\text{NegBino}(k; s, p) = \binom{k+s-1}{k} p^s (1-p)^k \quad (\text{C.12})$$

where  $p$  is the probability for success in a single Bernoulli trial,  $s$  is the (chosen) number of successes (an integer  $> 0$ ), and  $k$  is the number of failures ( $k = 0, 1, 2, 3, \dots$ ). It possesses the mean  $\mu = s(1-p)/p$  and variance  $\sigma^2 = s(1-p)/p^2$ .

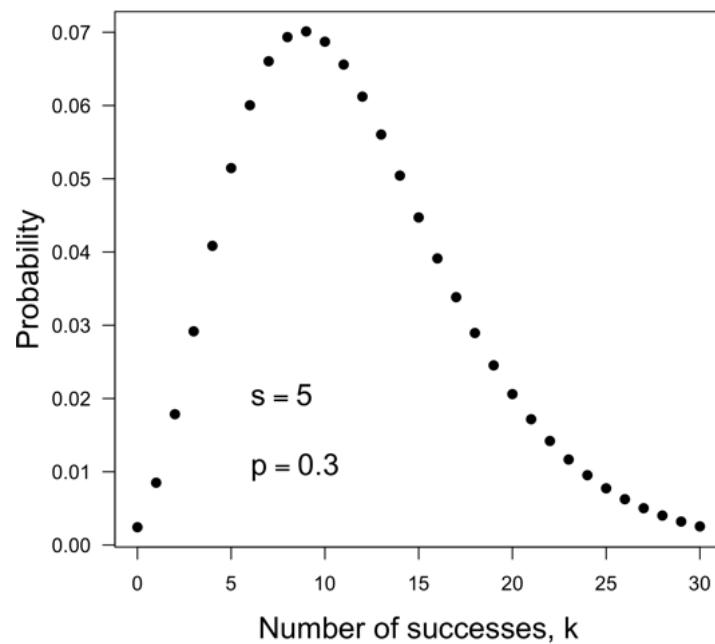


Figure C.7: The negative binomial distribution (Eq. C.12) for  $s = 5$ ,  $p = 0.3$ ,  $k = 0, 1, 2, \dots, 30$  (containing  $> 99\%$  of the probability). [https://github.com/WolfGladrow/TalksDA22\\_B2-PDsPDFsNegativeBinomialPDexample.R](https://github.com/WolfGladrow/TalksDA22_B2-PDsPDFsNegativeBinomialPDexample.R)

This definition of the negative binomial distribution is coded in **R** under the name '**dnbino(k,s,p)**'. The explanation in **R** is (as often!) a little bit confusing because they use the Gamma function (which is not required for integer values of  $k$  and  $s$ )<sup>2</sup>:

$$\text{NegBino}(k; s, p) = \frac{\Gamma(k+s)}{\Gamma(s) k!} p^s (1-p)^k \quad (\text{C.13})$$

$$\underbrace{\quad}_{k, s \text{ integers}} \frac{(k+s-1)!}{(s-1)! k!} p^s (1-p)^k \quad (\text{C.14})$$

$$= \binom{k+s-1}{k} p^s (1-p)^k \quad (\text{C.15})$$

The mean and the variance are given by

$$\mu = \frac{s(1-p)}{p} \quad (\text{C.16})$$

$$\sigma^2 = \mu + \frac{\mu^2}{s} \quad (\text{C.17})$$

and  $s$  is called the **dispersion parameter**.  $s$  and  $p$  can be calculated from  $\mu$  and  $\sigma$  as follows:

$$s = \frac{\mu^2}{\sigma^2 - \mu} \quad (\text{C.18})$$

$$p = \frac{s}{s + \mu} \quad (\text{C.19})$$

For  $s = 1$  one obtains the **geometric distribution** with

$$\mu = \frac{1-p}{p} \quad (\text{C.20})$$

$$\sigma^2 = \mu + \mu^2 \quad (\text{C.21})$$

In the context of generalized linear modeling (GLM), the dispersion parameter is estimated from data and thus usually not an integer. For a non-integer dispersion parameter  $s$  the definition of the negative binomial distribution using the gamma functions has to be applied.

Unfortunately, several (related) distributions go under the same name. Their definitions are given here for reference. Casella & Berger (2002, Eq. 3.2.9) define the probability distribution for  $n$  trials required before the  $s$ th success occurs

$$p_{\text{NB}'}(n; s, p) = \binom{n-1}{s-1} p^s (1-p)^{n-s} \quad (\text{C.22})$$

where  $p$  is the probability for success in a single Bernoulli trial,  $s$  is the (chosen) number of successes (an integer  $> 0$ ), and  $n$  is the number of Bernoulli trials ( $k = s, s+1, s+2, \dots$ ). The distributions (Eq. C.12) and (Eq. C.22) are related to each other since  $n = k + s$ :

$$p_{\text{NB}'}(n; s, p) = \binom{n-1}{s-1} p^s (1-p)^{n-s} \quad (\text{C.23})$$

$$= \underbrace{\binom{k+s-1}{s-1}}_{= \binom{k+s-1}{k}} p^s (1-p)^k \quad (\text{C.24})$$

$$= \binom{k+s-1}{k} p^s (1-p)^k = \text{NegBino}(k; s, p). \quad (\text{C.25})$$

---

<sup>2</sup>The **R** routine **dnbino(k,s,p)** is more general than the negative binomial distribution as defined in Eq. C.12 in that it can handle cases where  $s$  (however, not  $k$ ) is non-integer.

Wikipedia (Negative binomial distribution, accessed 16 July 2018) gives another definition:

$$p_{\text{NB}''}(m; f, q) = \binom{m+f-1}{m} q^m (1-q)^f \quad (\text{C.26})$$

where  $m = 0, 1, 2, \dots$  is the number of successes before the  $f$ th failure ( $f$  is an integer  $> 0$ ) occurs and  $q$  is the probability for failure. The substitutions  $m \rightarrow k$ ,  $f \rightarrow s$ ,  $q \rightarrow 1 - p$  lead to  $\text{NegBino}(k; s, p)$ .

For  $s \rightarrow \infty$  and  $p \rightarrow 1$  such that  $s(1-p) \rightarrow \lambda$  with  $0 < \lambda < \infty$ , the negative binomial distribution (Eq. C.12) becomes the Poisson distribution (Casella & Berger, 2002, p. 96).

**In summary,** Eq. (C.12) is the most commonly used definition of the negative binomial distribution. In order to be on the safe side, you should carefully check the meaning of the variable  $k$  and the model parameters  $s$  and  $p$ . The negative binomial distribution can be applied, for example, in the analysis of gene expressions (Robinson et al., 2010), especially when Poisson distributions do not work because of overdispersion (Huang et al., 2015). Further reading (negative binomial regression): Hilbe (2011).

### Exercise 69 Siblings of the normal distribution for count data

Binomial distributions and negative binomial distributions can be considered as siblings of normal distributions in that mean  $\mu$  and variance  $\sigma^2$  can be chosen separately from each other (whereas for Poisson distributions  $\mu = \sigma^2$ ).

- (1) Calculate the model parameters for a binomial distribution  $\text{Binomial}(k; n, p_b)$  and a negative binomial distribution  $\text{NegBino}(k; s, p)$  with  $\mu = 6$  and  $\sigma^2 = 3.2308$  (round  $n$  and  $s$  to the nearest integer). Plot both distributions in the same graph.
- (2) Plot the CDFs for the binomial distribution, the negative binomial distribution, and the normal distribution with these values of the mean and variance over the range from 0 to  $\mu + 5\sigma$ .

## C.3 Probability density functions: give me more

### C.3.1 Half-normal distribution

The half-normal distribution  $\mathcal{H}(x; \mu = 0, \sigma^2)$  is derived by taking the ‘positive half’ ( $x \geq 0$ ) of the normal distribution  $\mathcal{N}(x; \mu = 0, \sigma^2)$  and multiplying it by 2 to obtain normalization to 1. Formally it is defined by

$$\mathcal{H}(x; \mu = 0, \sigma) = \text{HalfNormal}(x; \mu = 0, \sigma) = \begin{cases} 2\mathcal{N}(x; \mu = 0, \sigma^2) = \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \quad (\text{C.27})$$

Please note that the true mean,  $\mu_{\text{HN}}$ , and true standard deviation,  $\sigma_{\text{HN}}$ , of the half-normal distribution with  $\mu = 0$  are (obviously!) different from  $\mu, \sigma$ :  $\mu_{\text{HN}} = \sigma\sqrt{2}/\sqrt{\pi}$ ,  $\sigma_{\text{HN}} = \sigma\sqrt{1 - 2/\pi}$ .

Application of half-normal distributions: Eguchi & Gerrodet (2009)

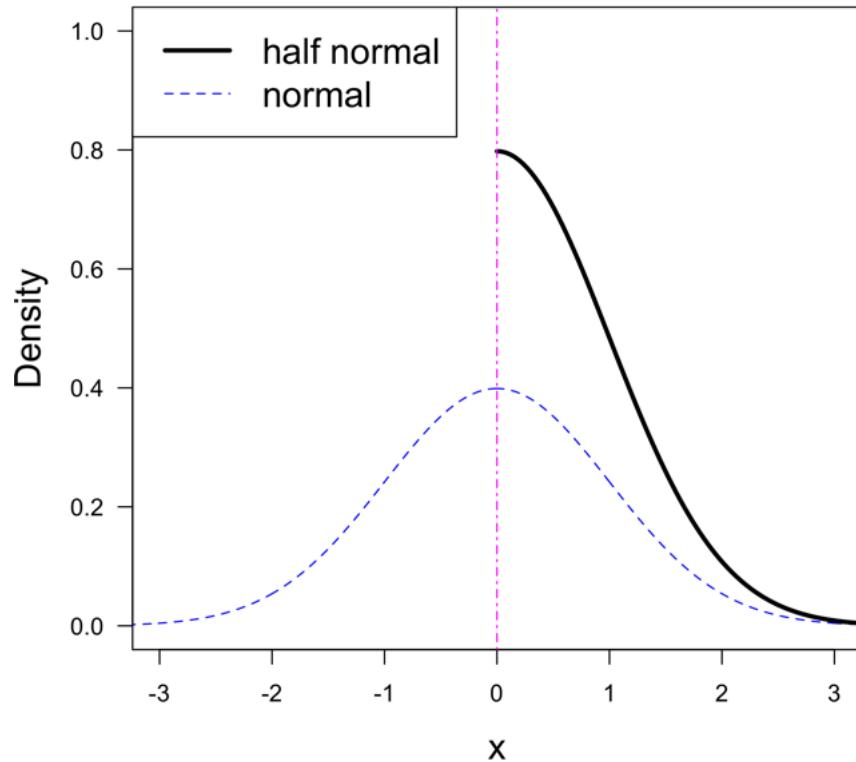


Figure C.8: PDF of the half-normal distribution (thick black solid line); standard normal PDF (thin green broken line) for comparison. [PDsPDFsHalfNormalPDFR](#)

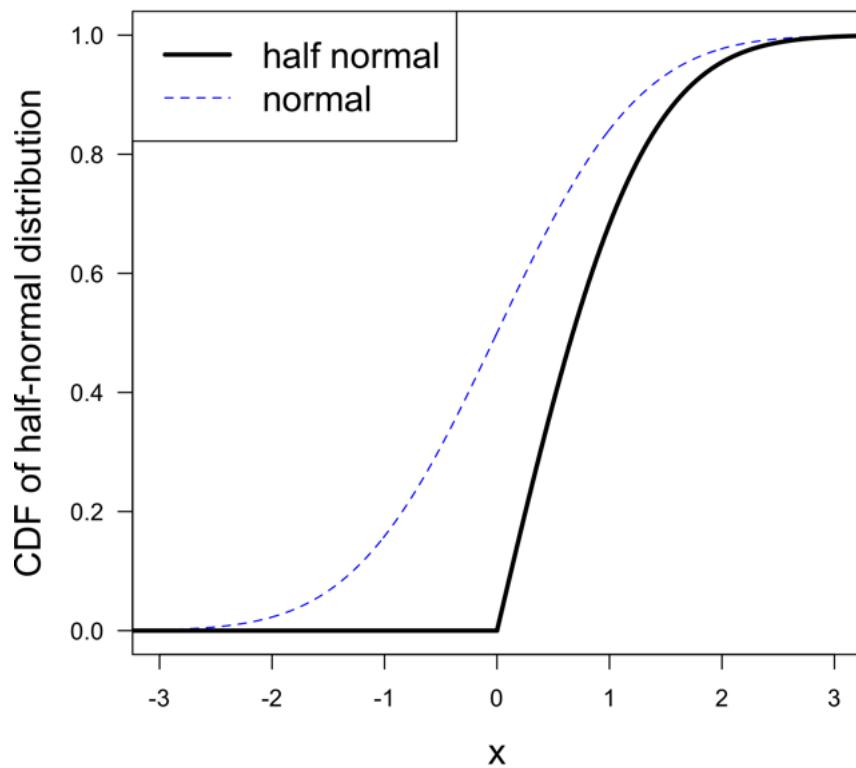


Figure C.9: CDF of the half-normal distribution (thick black solid line); CDF of standard normal PDF (thin green broken line) for comparison. [PDsPDFsHalfNormalCDF.R](#)

### C.3.2 The non-standardized Student's $t$ -distribution

The non-standardized Student's  $t$ -distribution

$$p(x; \nu, \mu, \beta^2) \equiv p_\nu(x; \mu, \beta^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu\beta^2}} \left(1 + \frac{1}{\nu} \frac{(x-\mu)^2}{\beta^2}\right)^{-\frac{\nu+1}{2}} \quad (\text{C.28})$$

can be considered as a  $t$ -distribution that is shifted by  $\mu$  and squeezed or stretched, i.e. it's variance is changed from  $\nu/(\nu-2)$  for the standard  $t$ -distribution to  $\sigma^2 = \beta^2\nu/(\nu-2)$  (both for degrees of freedom  $\nu > 2$ ).

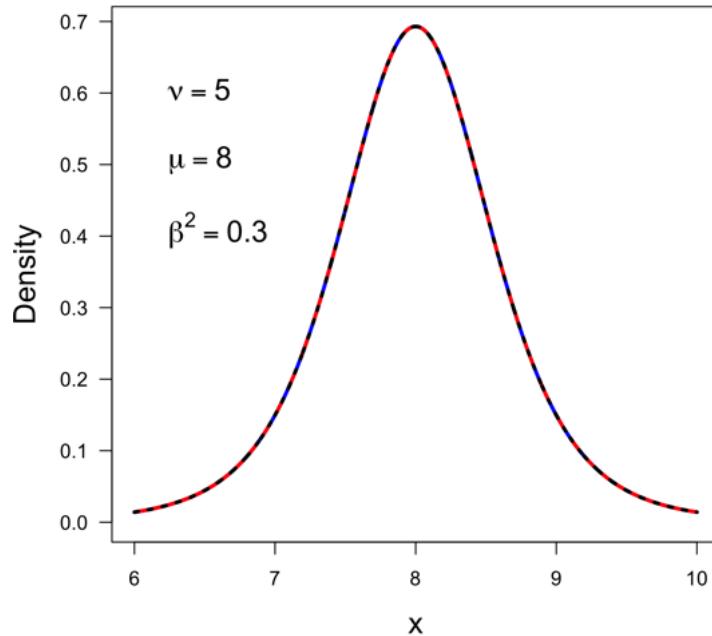


Figure C.10: The non-standardized Student's  $t$ -distribution for  $\nu = 5$ ,  $\mu = 8$ , and  $\beta^2 = 0.3$ . The distribution is symmetric about the mean at  $\mu = 8$ . The variance of the distribution  $\sigma^2 = \beta^2 \frac{\nu}{\nu-2} = 0.5$  is smaller than the variance  $\frac{\nu}{\nu-2} = 5/3$  of the  $t$ -distribution, i.e. the non-standardized Student's  $t$ -distribution shown in the graph has a smaller spread and a higher maximum than the  $t$ -distribution. [NonStandardized-t-PDF.R](#)

**Some properties of non-standardized Student's  $t$ -distributions:**

**Mean & mode** =  $\mu$  for  $\nu > 1$ .

**Variance:**  $\sigma^2 = \beta^2 \frac{\nu}{\nu - 2}$  for  $\nu > 2$ .

The alternative expression (Zellner, 1971, Eq. A.14)

$$p(x; \nu, \mu, h) = \frac{\Gamma[(\nu + 1)/2]}{\Gamma(1/2) \Gamma(\nu/2)} \left(\frac{h}{\nu}\right)^{1/2} \left[1 + \frac{h}{\nu} (x - \mu)^2\right]^{-(\nu+1)/2} \quad (\text{C.29})$$

looks slightly different, however, note that with  $h = \beta^2$  and  $\Gamma(1/2) = \sqrt{\pi}$  it's equivalent to (Eq. C.28).

We can obtain the standardized form of the  $t$  distribution (Eq. 6.54) by the following change of variable (Zellner, 1971)

$$t = \sqrt{h}(x - \mu) = \frac{x - \mu}{\beta} \quad (\text{C.30})$$

leading to

$$p_\nu(t) = \frac{\Gamma[(\nu + 1)/2]}{\sqrt{\nu} \Gamma(1/2) \Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty. \quad (\text{C.31})$$

The important conclusion from this change of variable is: [when the variable  \$x\$  obeys the non-standardized  \$t\$  distribution \(Eq. C.28\) with parameters  \$\mu\$ ,  \$\beta^2\$ , and  \$\nu\$ , the variable  \$t = \(x - \mu\)/\beta\$  obeys the standard  \$t\$  distribution \(Eq. 6.54\)](#). This result will be used in Section 11.5 when calculating the marginal posterior distribution for  $\mu$  for samples from normal populations with both mean and variance unknown. It can also be used to construct R routines for the density of non-standardized  $t$  distributions or for drawing random numbers from non-standardized  $t$  distributions by reducing it to the R routines for standard  $t$  distributions: [mydnst.R](#); [myrnst.R](#)

### C.3.3 Chi-squared ( $\chi^2$ ) distribution

**C.3.3** The  $\chi^2$  (chi-squared) distribution with  $\nu$  degrees of freedom is the distribution of the sum of the squares of  $\nu$  independent standard normal random variables.

$$\chi^2(x; \nu) = \frac{x^{(\nu/2 - 1)} e^{-x/2}}{2^{\nu/2} \Gamma\left(\frac{\nu}{2}\right)} \quad \text{for } x > 0 \quad (\text{C.32})$$

**Some properties of  $\chi^2$  distributions:**

**Mean:**  $\mu = \nu$

**Variance:**  $\sigma^2 = 2\nu$

**Mode:**  $\max(\nu - 2, 0)$

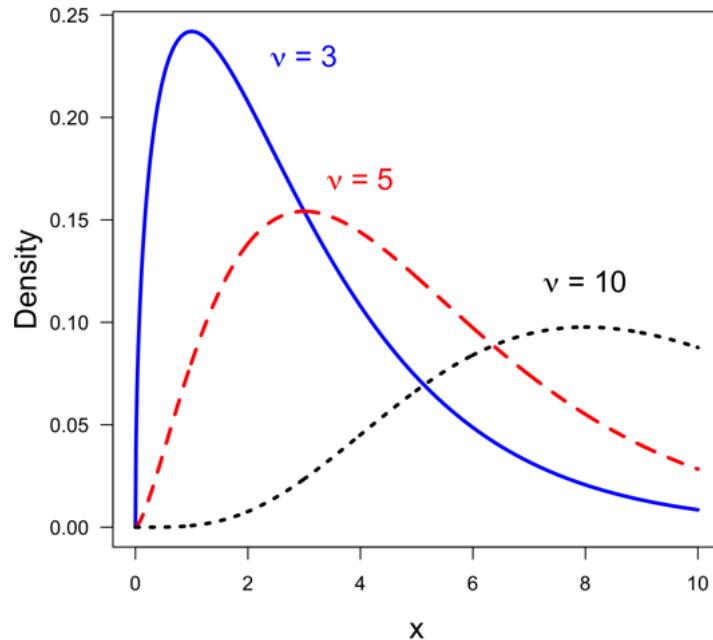


Figure C.11: Chi-squared distributions for  $\nu = 3, 5$ , and  $10$ . [B2-PDsPDFs-chi-squared-PDFs.R](#)

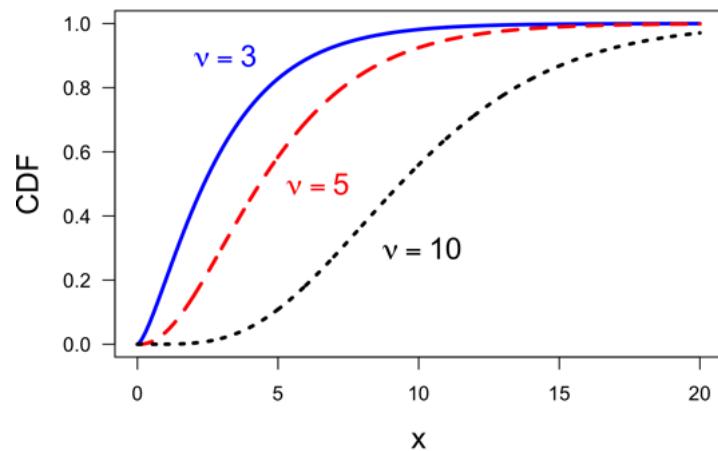


Figure C.12: CDFs for chi-squared distributions for  $\nu = 3, 5, \text{ and } 10$ . [B2-PDsPDFs-chi-squared-CDFs.R](#)

**Exercise 70 Density of sum of squares**

Estimate the density of sum of squares divided by  $\sigma^2$  using a Monte Carlo simulation – sampling from a normal distribution with variance  $\sigma^2$ , sample size  $n = 10$ , number of Monte Carlo run  $M = 10^4$  – and compare it with the  $\chi^2$  density for  $v = n$  degrees of freedom.

### C.3.4 Cauchy distribution

The Cauchy distribution<sup>3</sup>

$$\text{Cauchy}(x; \theta, \sigma) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x - \theta)^2} \quad (\text{C.33})$$

possesses no mean or standard deviation (the corresponding integrals diverge; Casella & Berger, 2002, p. 108), however, the median is equal to  $\theta$ . For  $\theta = 0$  one obtains the form of the Cauchy distribution most often used

$$\text{Cauchy}(x; \sigma) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + x^2} \quad (\text{C.34})$$

For  $\theta = 0, \sigma = 1$  one obtains the [standard Cauchy distribution](#) (Fig. C.13)

$$\text{Cauchy}(x; \theta = 0, \sigma = 1) = \text{Cauchy}(x) = \frac{1}{\pi(1 + x^2)}. \quad (\text{C.35})$$

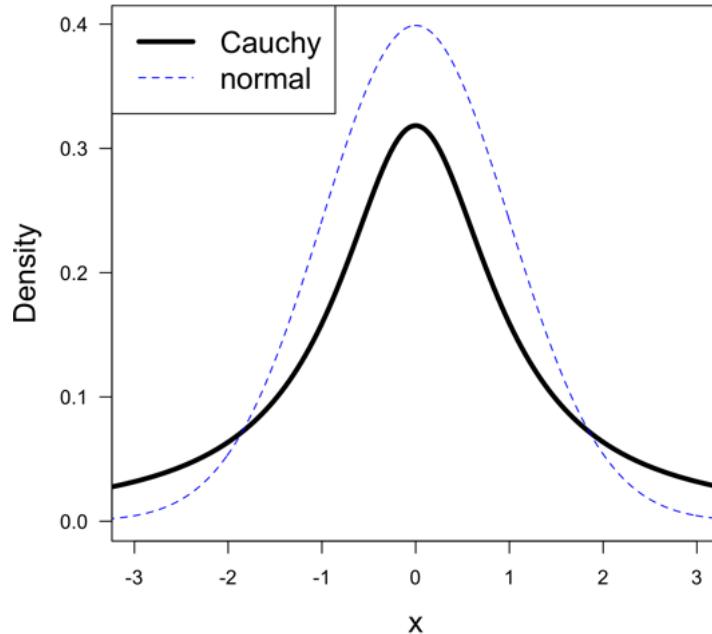


Figure C.13: Standard Cauchy distribution (Eq. C.35) [B2-PDsPDFsCauchy.R](#)

The ratio of two independent standard normal variables follows the standard Cauchy distribution.

The Cauchy distribution has been applied, for example, as a prior in hypothesis testing (Jeffreys, 1961); compare Section [12.1.4](#).

---

<sup>3</sup>'The Cauchy distribution, named after Augustin Cauchy' (1789-1857)', is a continuous probability distribution. It is also known, especially among physicists, as the Lorentz distribution (after Hendrik Antoon Lorentz', 1853-1928'), Cauchy-Lorentz distribution, Lorentz(ian) function, or Breit-Wigner distribution.' (Wikipedia)

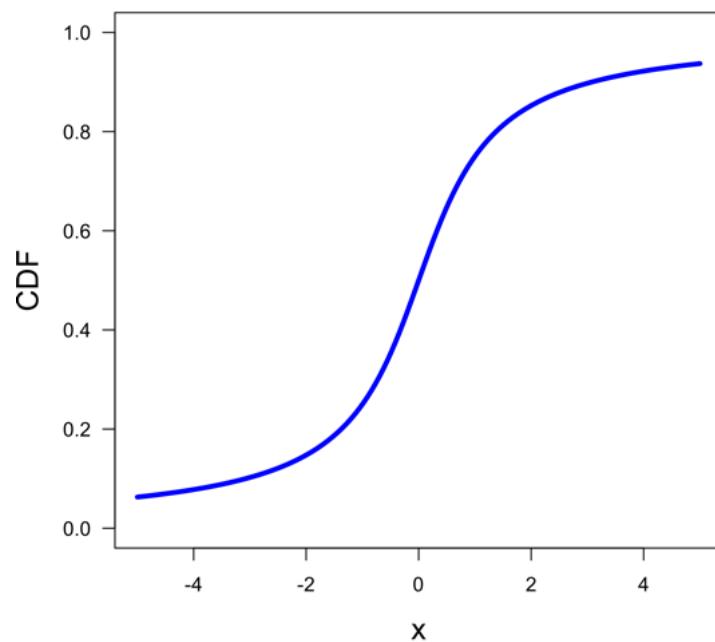


Figure C.14: CDF of standard Cauchy distribution [CauchyCDF.R](#)

### C.3.5 Scaled inverse $\chi^2$ & inverse-gamma distribution

The scaled inverse- $\chi^2$  and the inverse-gamma distribution are identical PDFs with different parameterizations. The inverse-gamma distribution occurs, for example, in the Bayesian approach to linear regression (Zellner, 1971). The distributions read:

$$\text{Inv-}\chi^2(x; \nu, \tau^2) = \frac{(\tau^2 \nu / 2)^{\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2+1)} \exp(-\nu \tau^2 / (2x)) \quad (\text{C.36})$$

with degrees of freedom  $\nu$ , scale parameter  $\tau^2$  and

$$\text{Inv-Gamma}(x; \phi, \psi) = \frac{\psi^\phi}{\Gamma(\phi)} x^{-\phi-1} \exp(-\psi/x) \quad (\text{C.37})$$

where  $\phi > 0$  is the shape parameter and  $\psi > 0$  is the scale parameter. The parameters of the distributions are related by  $\phi = \nu/2$ ,  $\psi = \tau^2 \nu / 2$  and  $\nu = 2\phi$ ,  $\tau^2 = \psi/\phi$ .

**Some properties of inverse- $\chi^2$  distributions:**

**Mean:**  $\mu = \frac{\nu \tau^2}{\nu - 2}$  for  $\nu > 2$ ; **Mode:**  $\frac{\nu \tau^2}{\nu + 2}$ ; **Variance:**  $\sigma^2 = \frac{2 \nu^2 \tau^4}{(\nu - 2)^2 (\nu - 4)}$  for  $\nu > 4$

**Some properties of inverse-gamma distributions:**

**Mean:**  $\mu = \frac{\psi}{\phi - 1}$  for  $\phi > 1$ ; **Mode:**  $\frac{\psi}{\phi + 1}$ ; **Variance:**  $\sigma^2 = \frac{\psi^2}{(\phi - 1)^2(\phi - 2)}$  for  $\phi > 2$

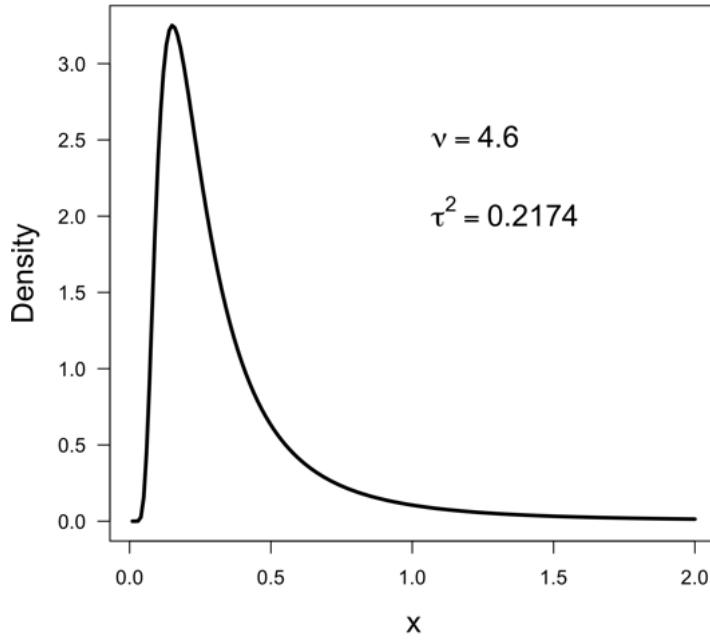


Figure C.15: The scaled inverse- $\chi^2$  PDF (Eq. C.36) for  $\nu = 4.6$ ,  $\tau^2 = 0.2174$  (these parameter values are consistent with  $\phi = 2.3$ ,  $\psi = 0.5$  used in Fig. C.16) [InvChisqPDF.R](#)

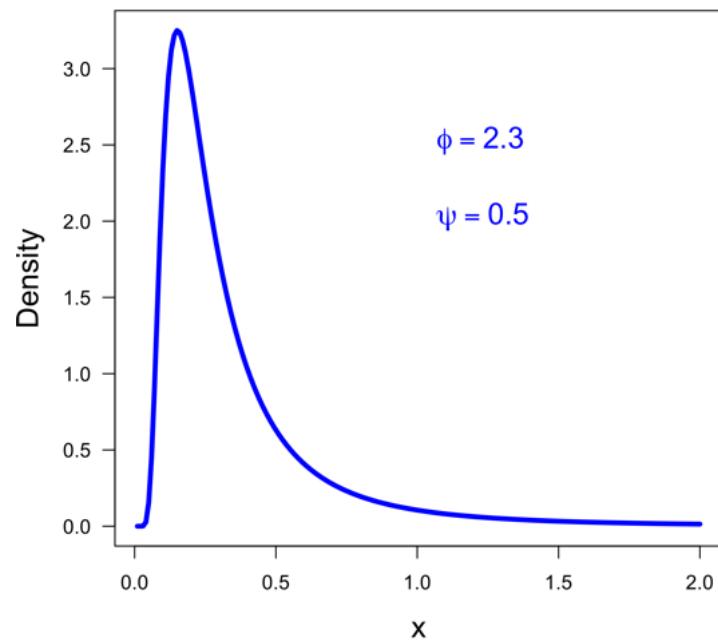


Figure C.16: The inverse-gamma PDF (Eq. C.37) for  $\phi = 2.3$ ,  $\psi = 0.5$  (these parameter values are consistent with  $\nu = 4.6$ ,  $\tau^2 = 0.2174$  used in Fig. C.15). `PDsPDFsInvGammaPDF.R`

### C.3.6 Kolmogorov-Smirnov CDF & PDF

The CDF for the test statistic  $D$  (Fig. C.17) is called the Kolmogorov-Smirnov distribution,  $f_{KS}(D; n)$ , where  $n$  is the sample size. For a long time it was difficult to calculate  $f_{KS}(D; n)$ . In 2003, Wang et al. proposed an efficient algorithm that was further improved by Carvalho (2015a) and implemented in R by Carvalho (2015b). The corresponding PDF (Fig. C.18) can be calculated by numerical differentiation of the CDF.

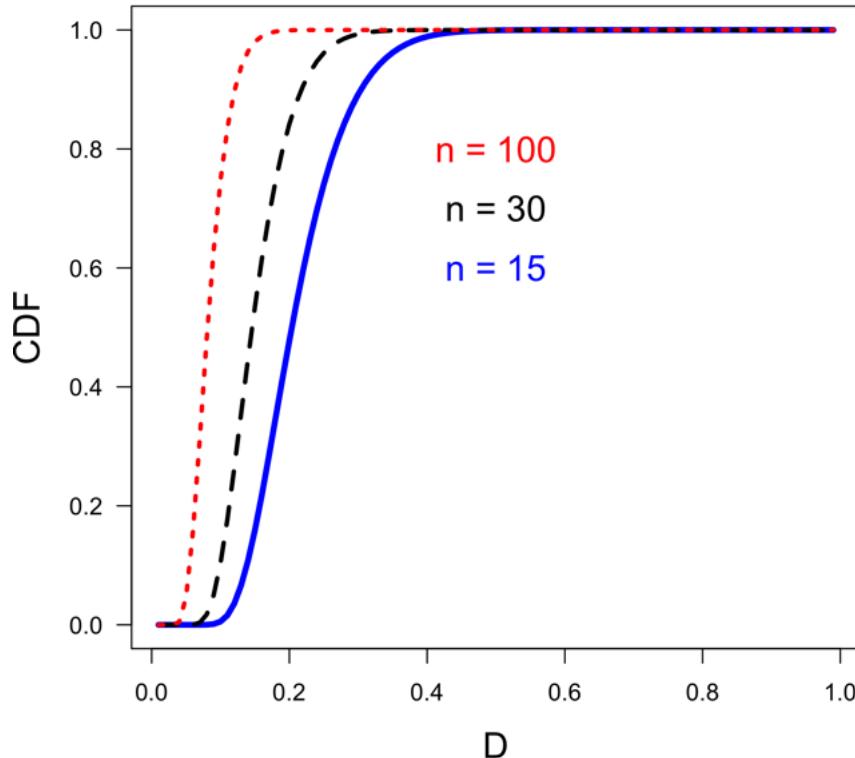


Figure C.17: Kolmogorov-Smirnov CDF for the test statistic  $D$  using the R function `pkolm()` of the package **kolmim** (Carvalho, 2015a,b):  $n = 15$  (blue line),  $n = 30$  (black line),  $n = 100$  (red line). [PDsPDFsKS-CDF.R](#)

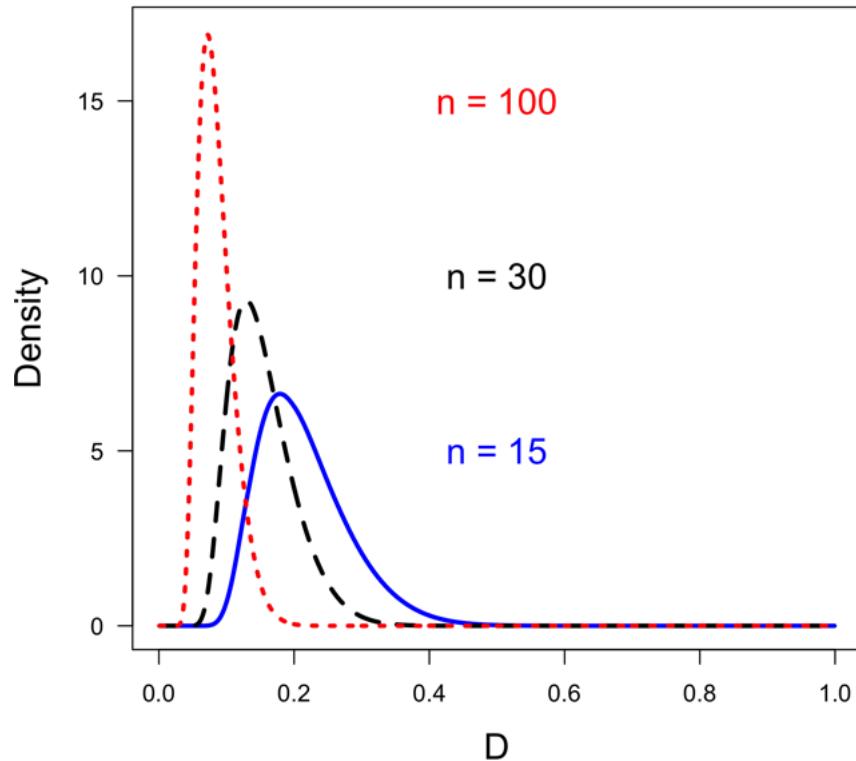


Figure C.18: The Kolmogorov-Smirnov PDF for the test statistic  $D$  has been calculated from the CDF (Fig. C.17) by numerical differentiation:  $n = 15$  (blue solid line),  $n = 30$  (black dashed line),  $n = 100$  (red dashed line). For small sample sizes ( $n = 15$ , blue solid line) one can expect (under the condition that the null hypothesis is true!) relative large maximum differences  $D$  which is reflected in the location of the PDF maximum at a relative large value (around  $D = 0.2$ ). When  $n$  is larger ( $n = 100$ , red dashed line) the expected  $D$  values are smaller which is consistent with a location of the maximum of the PDF at about  $D = 0.072$ . [PDsPDFsKS-PDF.R](#)

### C.3.7 Inverse chi-squared ( $\chi^2$ ) distribution

The inverse chi-squared distribution is defined as follows:

$$\text{Inv } \chi^2(x; \nu) = \frac{2^{-\nu/2}}{\Gamma(\frac{\nu}{2})} x^{-\nu/2-1} e^{-1/(2x)} \quad \text{for } x > 0 \quad (\text{C.38})$$

and 0 otherwise.

**Some properties of inverse  $\chi^2$  distributions:**

**Mean:**  $\mu = \frac{1}{\nu - 2}$  for  $\nu > 2$ .

**Variance:**  $\sigma^2 = \frac{2}{(\nu - 2)^2 (\nu - 4)}$  for  $\nu > 4$ .

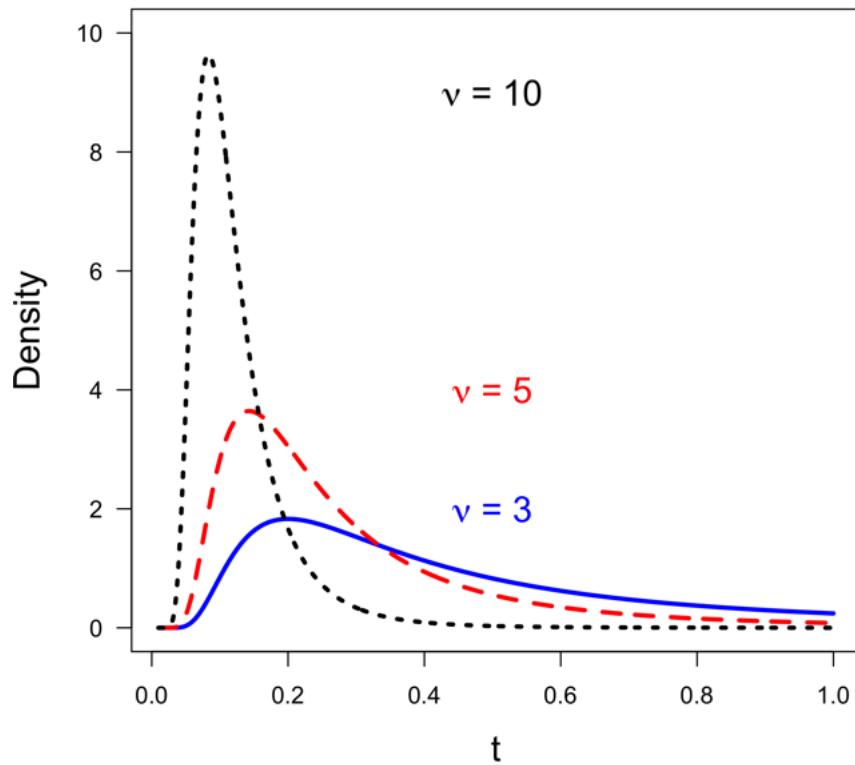


Figure C.19: Inverse chi-squared distributions for  $\nu = 3, 5$ , and  $10$ . [PDsPDFsInvChiSquaredPDF.R](#)

The cumulative distribution function (CDF) of the inverse chi-squared distribution is given by

$$\text{Inv } \chi^2_{\text{CDF}}(x; \nu) = \frac{\Gamma\left(\frac{\nu}{2}, \frac{1}{2x}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \quad (\text{C.39})$$

where  $\Gamma(a, b)$  (2 arguments) is the incomplete gamma function.

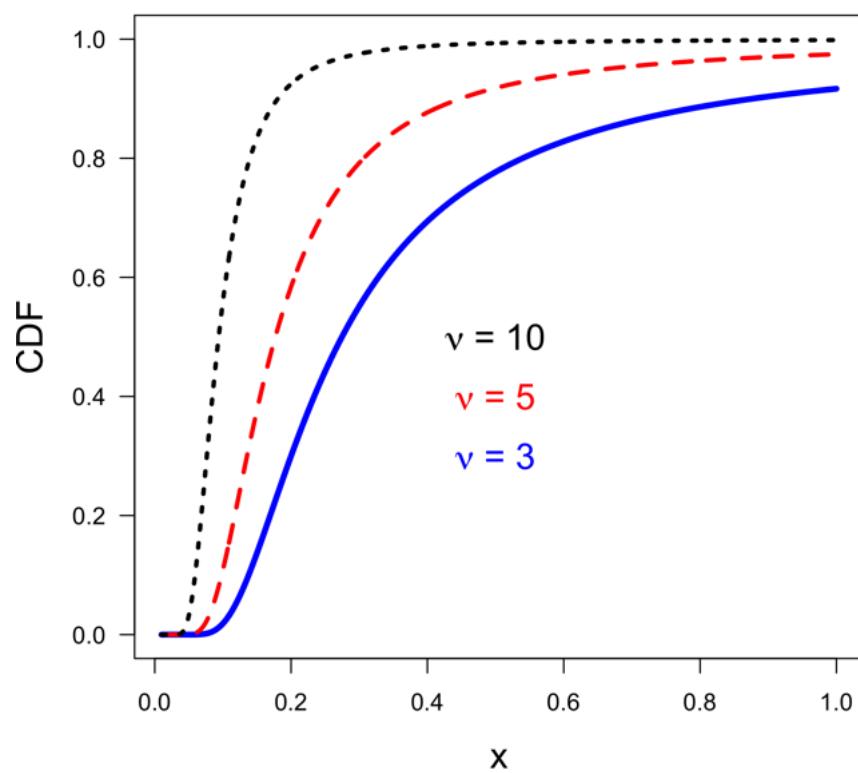


Figure C.20: CDFs of the inverse chi-squared distributions for  $\nu = 3, 5$ , and  $10$ . [PDsPDFsInvChiSquaredCDF.R](#)

### C.3.8 Lognormal PDF

The lognormal PDF (Fig. C.21) is defined by

$$\text{LogNormal}(x; \alpha, \beta) = \frac{1}{x \beta \sqrt{2 \pi}} e^{-\frac{(\log x - \alpha)^2}{2 \beta^2}} \quad (\text{C.40})$$

**Some properties of lognormal distributions:**

**Mean:**  $\mu = e^{\alpha + \beta^2/2}$ .

**Variance:**  $\sigma^2 = (e^{\beta^2} - 1) e^{2\alpha + \beta^2}$ .

Vice versa, the distribution parameters  $\alpha$  and  $\beta$  can be calculated from  $\mu$  and  $\sigma$  as follows:

$$\alpha = \log(\mu) - \log((\sigma/\mu)^2 + 1)/2 \quad (\text{C.41})$$

$$\beta = \sqrt{\log((\sigma/\mu)^2 + 1)} \quad (\text{C.42})$$

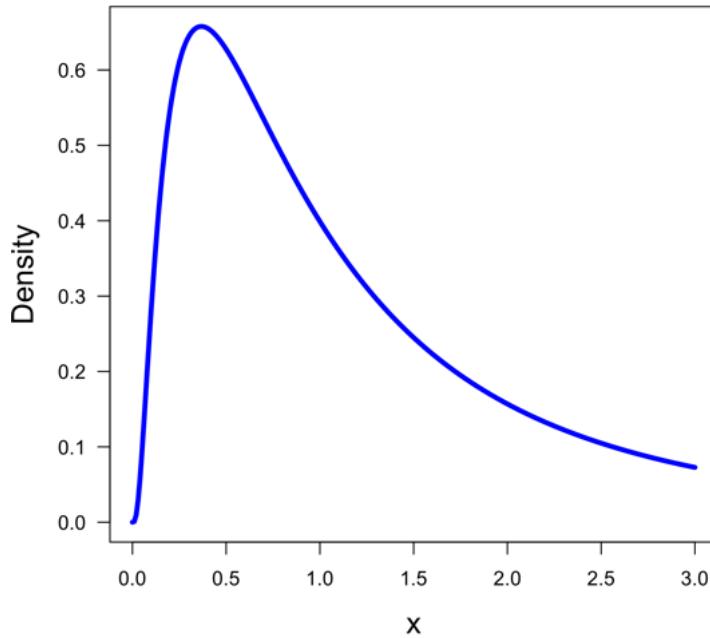


Figure C.21: The standard ( $\alpha = 0, \beta^2 = 1$ ) log-normal PDF. [PDsPDFsLogNormalPDF.R](#)

The CDF of the log-normal distribution (Fig. C.22) reads

$$\text{CDF}_{\text{LN}}(x; \alpha, \beta^2) = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left[ \frac{\log x - \alpha}{\beta \sqrt{2}} \right] \quad (\text{C.43})$$

**Relation to the Central Limit Theorem** (Chapter 7): 'A log-normal process is the statistical realization of the multiplicative product of many independent random variables, each of which is positive. This is justified by

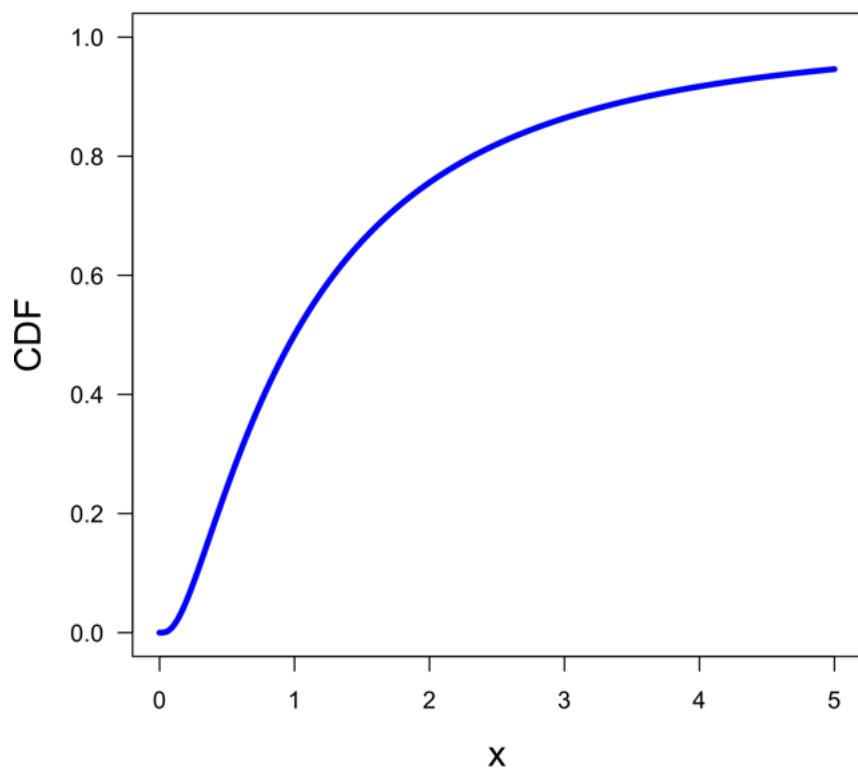


Figure C.22: The CDF of the standard log-normal distribution. [PDsPDFsLogNormalCDF.R](#)

considering the central limit theorem in the log domain.' (Wikipedia) Or, in other words und using equations, if  $X$  is the product of many stochastic (random) factors  $F_k$ , i.e.

$$X = \prod_k F_k, \quad (\text{C.44})$$

then

$$\log X = \sum_k \log F_k \quad (\text{C.45})$$

and thus according to the Central Limit Theorem  $\log X$  is normally distributed for  $k \rightarrow \infty$ .

"The log-normal distribution is the maximum entropy probability distribution for a random variate  $X$  for which the mean and variance of  $\ln(X)$  are specified." (Wikipedia, Log-normal distribution)

The log-normal distribution occurs in many different contexts in nature, society or technical systems. One reason is Gibrat's law stating that relative growth rate is independent of size (Robert Gibrat, 1904-1980).

Occurrence & application in oceanography, ecology, etcetera: Campbell (1995), Limpert et al. (2001), Gerrodette (2011)

### **Exercise 71 Plot lognormal PDF for given mean & standard deviation**

*Plot the lognormal PDF for mean  $\mu = 409$  and standard deviation  $\sigma = 250$  and compare it with the corresponding normal PDF.*

**Exercise 72 Plot lognormal PDF and negative binomial PD**

Plot the lognormal PDF and the negative binomial PD with mean  $\mu = 409$  and standard deviation  $\sigma = 250$  for  $x = 1$  to 1000.

### C.3.9 $\beta$ distribution

$$\text{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}, \quad 0 \leq p \leq 1, \quad \alpha > 0, \quad \beta > 0. \quad (\text{C.46})$$

**Mean:**  $\mu = \frac{\alpha}{\alpha + \beta}$ ; **Variance:**  $\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

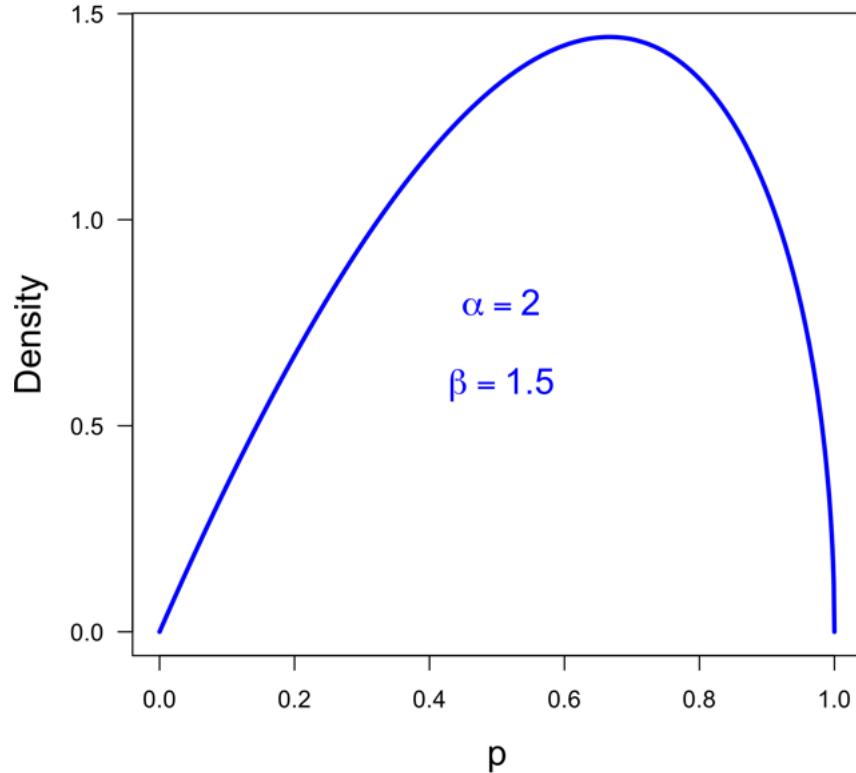


Figure C.23: The beta PDF for  $\alpha = 2, \beta = 1.5$ . [PDsPDFsBetaPDF.R](#)

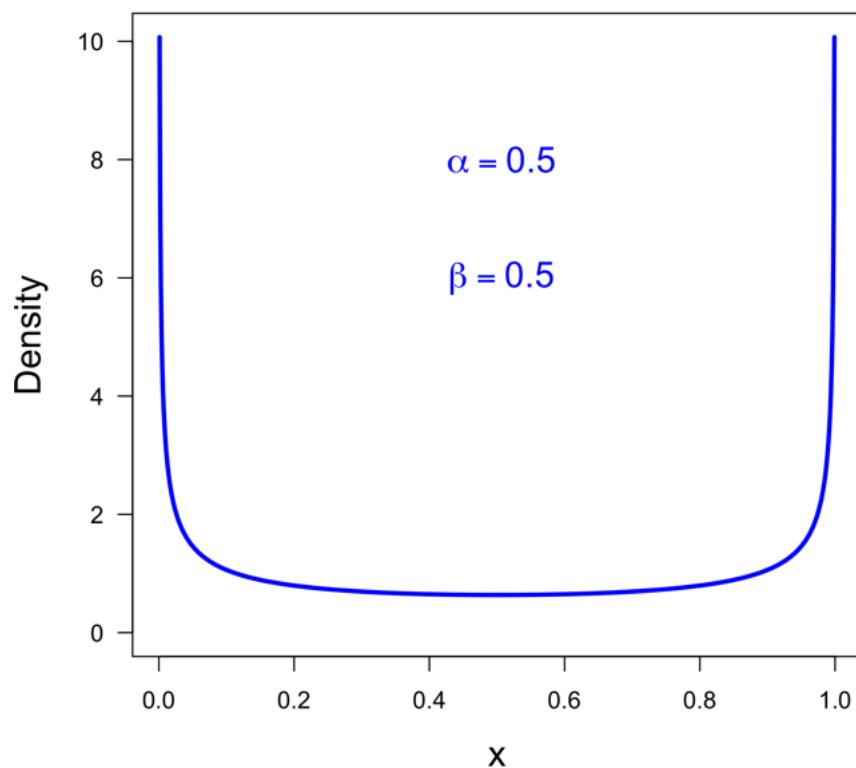


Figure C.24: The beta PDF for  $\alpha = 0.5$ ,  $\beta = 0.5$  is used as a reference prior for the mode of the nonsymmetric triangular PDF (Section M.2.4). [PDsPDFsBetaPDF0505.R](#)

### C.3.10 Exponential PDF

The exponential PDF (Fig. C.25) is defined by

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0 \quad (\text{C.47})$$

where  $\lambda > 0$  is the rate.

**Some properties of exponential PDFs:**

**Mean:**  $\mu = \frac{1}{\lambda}$

**Variance:**  $\sigma^2 = \frac{1}{\lambda^2}$

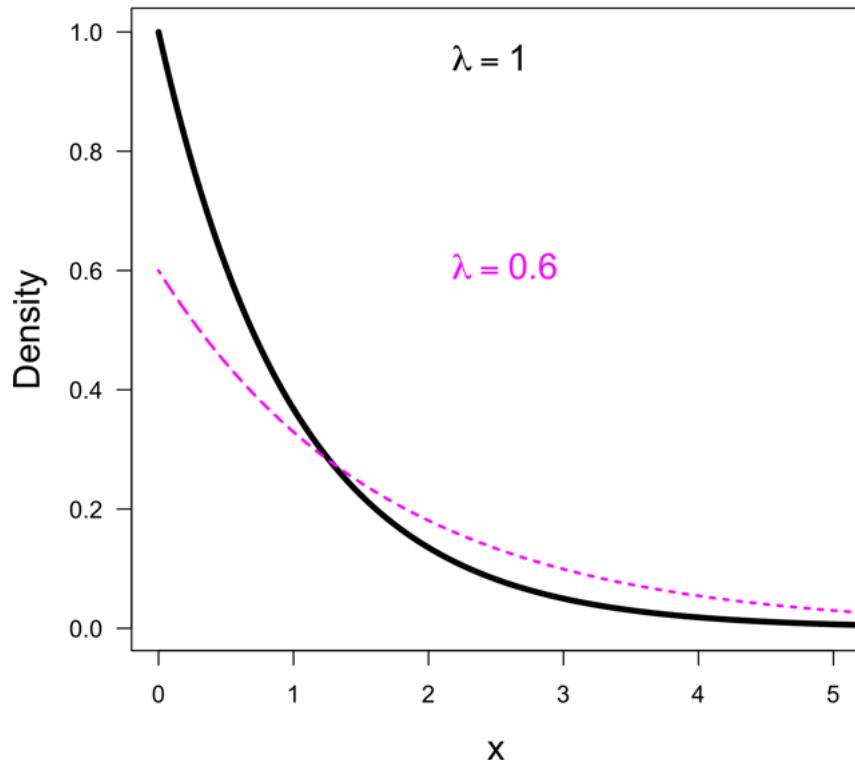


Figure C.25: The exponential PDF  $f(x; \lambda) = \lambda e^{-\lambda x}$  for  $\lambda = 1$  (blue solid line) and  $\lambda = 0.6$  (magenta dash-dotted line). [PDsPDFsExpPDF.R](#)

### C.3.11 Gamma PDF: $\text{Gamma}(x; \alpha, \beta)$

The gamma PDF<sup>4</sup> is defined by (Robert, 2007, p. 519)<sup>5</sup>

$$\text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad (\text{C.48})$$

for  $0 < x < \infty, \alpha > 0, \beta > 0$ . Notation:  $\alpha$  is usually called the 'shape' and  $\beta$  the 'rate'.

**Some properties of gamma PDFs:**

**Mean:**  $\mu = \frac{\alpha}{\beta}$

**Variance:**  $\sigma^2 = \frac{\alpha}{\beta^2}$

The gamma PDF is conjugate to the Poisson distribution. It is applied in the analysis of the neutrino data in Section 1.4.

For  $\beta = 1$  one obtains the standard gamma PDF

$$\text{Gamma}(x; \alpha) = \frac{x^{\alpha-1}}{\Gamma(\alpha)} e^{-x} \quad (\text{C.49})$$

for  $0 < x < \infty, \alpha > 0$  with mean  $\mu = \alpha$  and variance  $\sigma^2 = \alpha$ .

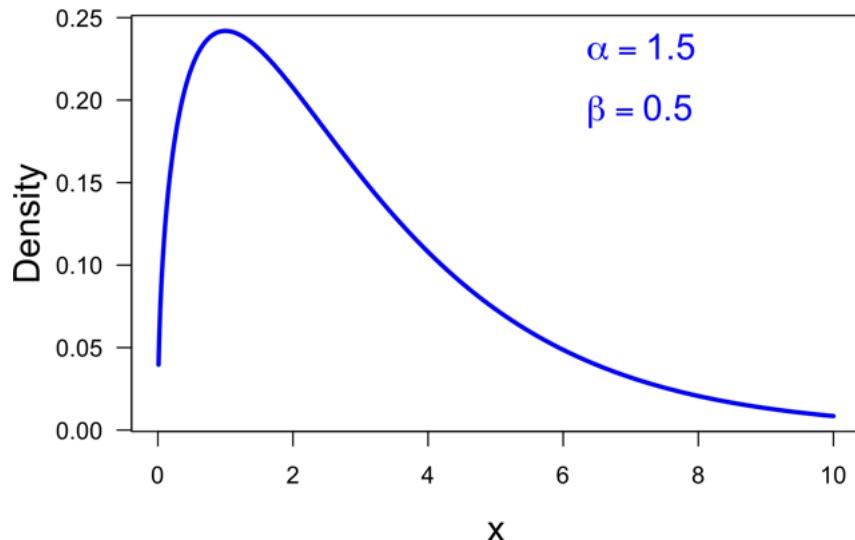


Figure C.26: The gamma PDF for  $\alpha = 2$  and  $\beta = 0.5$ . [GammaPDF.R](#)

<sup>4</sup>Please don't confuse with the upper or lower incomplete gamma function  $\Gamma(s, x)$  or  $\gamma(s, x)$ , respectively, which are defined by  $\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$  and  $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$

<sup>5</sup>Casella & Berger (2002, p. 99) use a different parameterization with  $\theta = 1/\beta$  leading to  $\text{Gamma}(x; \alpha, \theta) = \frac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}$ .

### C.3.12 Upper & lower incomplete $\Gamma$ function (\*)

The lower incomplete  $\Gamma$  function is defined by

$$\gamma(x; s) = \frac{1}{\Gamma(s)} \int_0^x t^{s-1} e^{-t} dt \quad (\text{including normalization factor}) \quad (\text{C.50})$$

for  $s > 0$  and  $x \geq 0$ .  $\gamma(x; s)$  is a cumulative probability distribution function (CDF), i.e.  $\gamma(x; s) \geq 0$  increasing with  $x$ ,  $\gamma(0; s) = 0$ ,  $\gamma(x \rightarrow \infty; s) = 1$  (Fig. C.27).

The upper incomplete  $\Gamma$  function is defined by

$$\Gamma(x; s) = \frac{1}{\Gamma(s)} \int_x^\infty t^{s-1} e^{-t} dt = 1 - \gamma(x; s). \quad (\text{C.51})$$

Remark: In the literature one can find an alternative definition of  $\gamma(x; s)$  without the normalization factor  $1/\Gamma(s)$ ; this alternative form is not a CDF and thus not discussed here.

#### Incomplete gamma functions in R

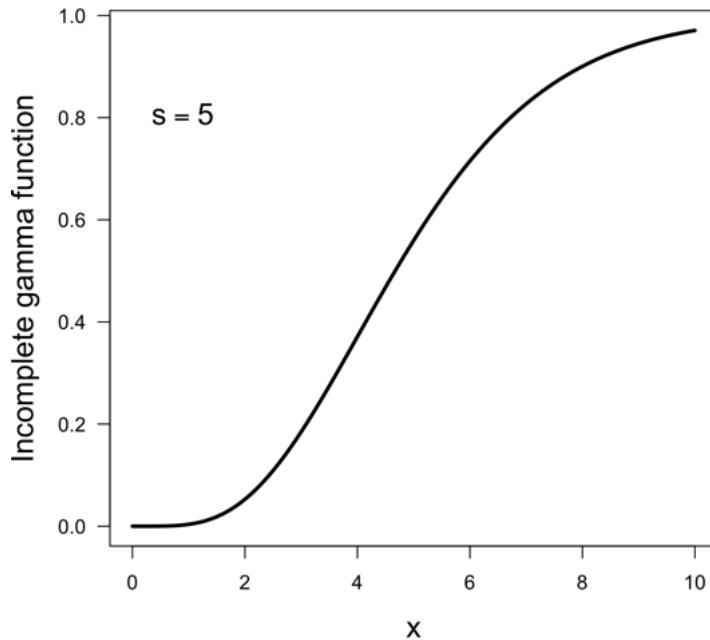


Figure C.27: Lower incomplete gamma function  $\gamma(x; s) = \frac{1}{\Gamma(s)} \int_0^x t^{s-1} e^{-t} dt$  for  $s = 5$ .  
[LowerIncompleteGamma.R](#)

### C.3.13 Family of Student PDFs

The family of Student PDFs is defined by (Maronna et al., 2006, p. 20)

$$f_\nu = c_\nu \left( 1 + \frac{x^2}{\nu} \right)^{-(\nu+1)/2} \quad (\text{C.52})$$

where

$$c_\nu = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \quad (\text{C.53})$$

is the normalization constant and  $\nu > 0$  the degrees of freedom. Special cases: Cauchy distribution for  $\nu = 1$ , normal distribution for  $\nu \rightarrow \infty$ . Family members with  $\nu < \infty$  posses tails that are 'fatter' than the normal distribution. An extreme case is the Cauchy distribution: here the tails are so fat that the variance does not exist.

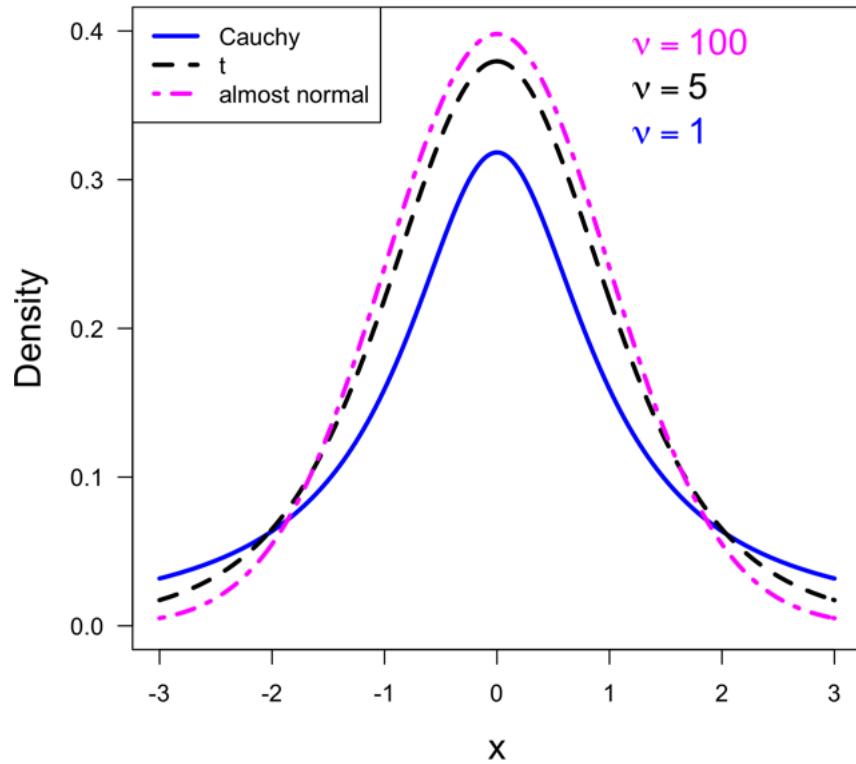


Figure C.28: Members of the Student family of PDFs: Cauchy ( $\nu = 1$ , blue solid line), Student-t PDF with  $\nu = 5$  (black dashed line; 'in between' Cauchy and normal), and Student-t PDF with  $\nu = 100$  (magenta dash-dotted line; almost normal). [PDsPDFs-t-Family.R](#)

## C.4 Joint PDFs: polynomial example (\*)

In the following we will construct a polynomial joint PDF  $f(x, y)$  over the unit square ( $(x, y) \in [0, 1] \times [0, 1]$ ).

$$f(x, y; \alpha, \beta) = c (\alpha x^2 y + \beta x y^3) \quad (\text{C.54})$$

where  $\alpha$  and  $\beta$  are the parameters of the distribution and  $c$  is a normalisation constant that can be calculated from the normalisation constraint

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y; \alpha, \beta) dx dy \\ &= c \int_0^1 \int_0^1 (\alpha x^2 y + \beta x y^3) dx dy = c \int_0^1 \left( \frac{\alpha}{3} x^3 y + \frac{\beta}{2} x^2 y^3 \right) \Big|_0^1 dy \\ &= c \int_0^1 \left( \frac{\alpha}{3} y + \frac{\beta}{2} y^3 \right) dy = c \left( \frac{\alpha}{6} y^2 + \frac{\beta}{8} y^4 \right) \Big|_0^1 \\ &= c \left( \frac{\alpha}{6} + \frac{\beta}{8} \right) = c \frac{4\alpha + 3\beta}{24} \\ \Rightarrow & c = \frac{24}{4\alpha + 3\beta} \end{aligned} \quad (\text{C.55})$$

and thus finally

$$f(x, y; \alpha, \beta) = \frac{24}{4\alpha + 3\beta} (\alpha x^2 y + \beta x y^3) \quad (\text{C.56})$$

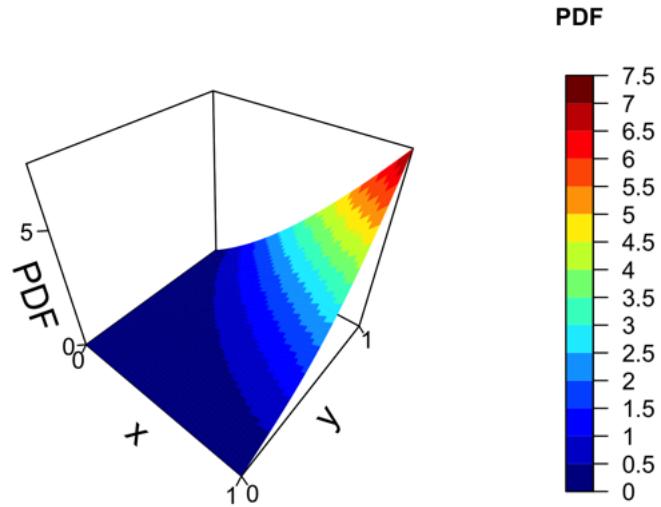


Figure C.29: The joint PDF  $f(x, y; \alpha, \beta)$  (Eq. C.56) for  $\alpha = 2$  and  $\beta = 3$ . [PDFsPDFsJointPolyPDF.R](#)

The means of  $x$  and  $y$  with respect to  $f()$  are defined by

$$\begin{aligned}
 \mu_x &= E_f[x] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f(x, y; \alpha, \beta) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \int_0^1 x (\alpha x^2 y + \beta x y^3) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \int_0^1 (\alpha x^3 y + \beta x^2 y^3) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \left( \frac{\alpha}{4} x^4 y + \frac{\beta}{3} x^3 y^3 \right) \Big|_0^1 dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \left( \frac{\alpha}{4} y + \frac{\beta}{3} y^3 \right) dy \\
 &= \frac{24}{4\alpha + 3\beta} \left( \frac{\alpha}{8} y^2 + \frac{\beta}{12} y^4 \right) \Big|_0^1 \\
 &= \frac{24}{4\alpha + 3\beta} \left( \frac{\alpha}{8} + \frac{\beta}{12} \right) = \frac{24}{4\alpha + 3\beta} \frac{3\alpha + 2\beta}{24} \\
 &= \frac{3\alpha + 2\beta}{4\alpha + 3\beta}
 \end{aligned}$$

and

$$\begin{aligned}
 \mu_y &= E_f[y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x, y; \alpha, \beta) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \int_0^1 y (\alpha x^2 y + \beta x y^3) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \int_0^1 (\alpha x^2 y^2 + \beta x y^4) dx dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \left( \frac{\alpha}{3} x^3 y^2 + \frac{\beta}{2} x^2 y^4 \right) \Big|_0^1 dy \\
 &= \frac{24}{4\alpha + 3\beta} \int_0^1 \left( \frac{\alpha}{3} y^2 + \frac{\beta}{2} y^4 \right) dy \\
 &= \frac{24}{4\alpha + 3\beta} \left( \frac{\alpha}{9} y^3 + \frac{\beta}{10} y^5 \right) \Big|_0^1 \\
 &= \frac{24}{4\alpha + 3\beta} \left( \frac{\alpha}{9} + \frac{\beta}{10} \right) = \frac{24}{4\alpha + 3\beta} \frac{10\alpha + 9\beta}{90} \\
 &= \frac{4}{15} \frac{10\alpha + 9\beta}{4\alpha + 3\beta}
 \end{aligned}$$

The mean values are functions of the model parameters  $\alpha$  and  $\beta$ . The coefficients of these rational functions depend on the exponents of  $x$  and  $y$  in the distribution  $f()$ .

The variances and covariance are defined by

$$\begin{aligned}
 \sigma_x^2 &= E_f[(x - \mu_x)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x, y; \alpha, \beta) dx dy \\
 \sigma_y^2 &= E_f[(y - \mu_y)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (y - \mu_y)^2 f(x, y; \alpha, \beta) dx dy \\
 \text{cov}(x, y) &= E_f[(x - \mu_x)(y - \mu_y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) f(x, y; \alpha, \beta) dx dy
 \end{aligned}$$

Although these integrals can be solved analytically for our polynomial distribution, it is not done here because this is a tedious procedure<sup>6</sup>.

<sup>6</sup>Could be done, for example, by using MATHEMATICA (Wolfram Research).



# Appendix D

## Uncertainty

'There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. These are things we do not know we don't know.'

United States Secretary of Defense Donald Rumsfeld (press briefing February 12, 2002)

### D.1 Paradigm shift

'The question of accuracy of a measurement result has permanently accompanied the measuring techniques of the modern age since its beginning more than 300 years ago. When searching for an answer during the last decades a change in view has occurred that can be called a **paradigm shift**.

Previously the measurements aimed at detecting the 'true' value of a measurand or the characteristics of a measured object. Systematic or random errors in measurements were identified as the reason for the more or less accurate determination. These measurement errors which in principle cannot be determined completely prevent the detection of the 'true' value.

The new view carries on adhering to the statement that the measureable quantity already exists prior to the measurement (the time and the extend are in the world, even if we do not measure them). **However, a 'true value' independent from any measurement does not exist.** In fact, the aim of a measurement is to determine the value of the measured quantity. Only by measurement a value regarding to a conventional system of units is attributed to a measurand. (A stretch gets a length value only by measuring. The measured value is the only value we are able to know. It comes into the world only by measuring) . . .

The change from the former traditional view to the new operational view is affected by the fact that both views use techniques from **probability theory**. The formulas look equal but their meaning is different.'

Kunzman, H. and W. Kessel, GUM-konforme Auswertung von Messungen, tm - Technisches Messen, 68(1), 3-4, 2001, doi.10.1524/terme.2001.68.1.3

(translated from German to English by Dörte Rosenbaum & DWG; emphasized by DWG).

## D.2 Taylor & Kuyatt (1994) definitions

If not otherwise stated, all quotations in this section are from Taylor & Kuyatt (1994).

### Repeatability

**Repeatability is the 'closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement.'**

These conditions are called repeatability conditions. Repeatability conditions include:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time.

Repeatability may be expressed quantitatively in terms of the dispersion characteristics of the results.'

**See also:** reproducibility

### Reproducibility

**Reproducibility (of results of measurements) is the 'closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement.'** A valid statement of reproducibility requires specification of the conditions changed. The changed conditions may include:

- principle of measurement
- method of measurement
- observer
- measuring instrument
- reference standard
- location
- conditions of use
- time.

### Examples:

- (1) Yubin measures the pH (= the measurand here) of a water sample using 5 different electrodes (change of measuring instrument).
- (2) Clara measures the pH (= the measurand here) of a water sample using two different methods (a) using a pH electrode and (b) by a spectrophotometric method (Bellerby et al., 2002).
- (3) Paul measures salinity (= the measurand here) of a sample from a large water batch. The next day, Paula measures salinity of a second sample from the same batch (change of observer and time).

**See also:** repeatability.

### Accuracy

**Accuracy of measurement is defined as the 'closeness of the agreement between the result of a measurement and the value of the measurand'** (NIST TN1297, 1994).

Remarks:

- The 'measurand' is the 'specific quantity subject to measurement' or the quantity we would like measure as, for example, the mass of an electron or the mean growth rate of the algal species *Skeletonema costatum* (a diatom).
- The 'value of the measurand' was formerly called the 'true value of the measurand' or the 'true value' for short..
- 'Accuracy' is a **qualitative concept** and thus one should not use it quantitatively, that is, associate numbers with it.
- **Do not use:** 'The accuracy is  $4 \mu\text{mol kg}^{-1}$ '.
- **Use:** 'The standard uncertainty is  $4 \mu\text{mol kg}^{-1}$ '.

### Error

**Error of results of measurement is defined as the 'the results of a measurement minus the value of the measurand'** (NIST TN1297, 1994).

Remarks:

- The 'measurand' is the 'specific quantity subject to measurement' or the quantity we would like to measure as, for example, the mass of an electron or the mean growth rate of the algal species *Skeletonema costatum* (a diatom).
- The 'value of the measurand' was **formerly** called the 'true value of the measurand' or 'true value' for short..
- The 'error' could be considered as what was **formerly** known as 'accuracy', which now is not a quantitative concept anymore. It is difficult to determine the error. That is why 'uncertainty' is usually the preferred quantity.



# Appendix E

## Monte Carlo simulations

"In a recent conversation with [Emilio] Segre, I learned that Fermi took great delight in astonishing his Roman colleagues with his remarkable accurate, 'too-good-to-believe' predictions of experimental results. After indulging himself, he revealed that his 'guesses' were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck in the wee morning hours! And so it was that nearly fifteen years earlier [than 1987], Fermi had independently developed the Monte Carlo method."

Metropolis (1987)

The term 'Monte Carlo method' for statistical sampling was invented in 1947 by N. Metropolis (Metropolis, 1987).

**References:** Metropolis et al. (1953), Metropolis (1987), Robert & Casella (2009, 2013)

### E.1 Monte Carlo integration (\*)

Starting point of Monte Carlo integration is the definition of the **expectation** (= mean over the PDF  $f(X)$ ) of the function  $h()$  by

$$E_f [(h(X))] = \int h(x) f(x) dx. \quad (\text{E.1})$$

The expectation of  $h(X)$  can be estimated by generation of a random sample  $\{x_1, \dots, x_n\}$  from the PDF  $f()$ , calculate the values of  $h()$  at these sample points, i.e.  $\{h(x_1), \dots, h(x_n)\}$ , and finally calculate the sample mean of the latter values:

$$\hat{E}_f [(h(X))] = \frac{1}{n} \sum_{k=1}^n h(x_k). \quad (\text{E.2})$$

This is the basic principle of classic Monte Carlo integration. One can show that this estimate 'converges almost surely' (i.e. for almost every generated sequence) to  $E_f [(h(X))]$  by the Strong Law of Large Numbers' (Robert & Casella, 2009).

**Example:**  $h(x) = \cos(x)$ ,  $f(x)$  is the standard normal distribution, integral from  $-\infty$  to  $\infty$ :

$$I_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos x e^{-x^2/2} dx \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \cos x e^{-x^2/2} dx \frac{2}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2/\sqrt{2}} e^{-1/2} = e^{-1/2} \quad (\text{E.3})$$

where we used the fact that the integrand is an even function and the definite integral

$$\int_0^{\infty} e^{-a^2 x^2} \cos b x dx = \frac{\sqrt{\pi}}{2a} e^{-b^2/4a^2} \quad (\text{E.4})$$

with  $a = 1/\sqrt{2} > 0$  and  $b = 1$ .

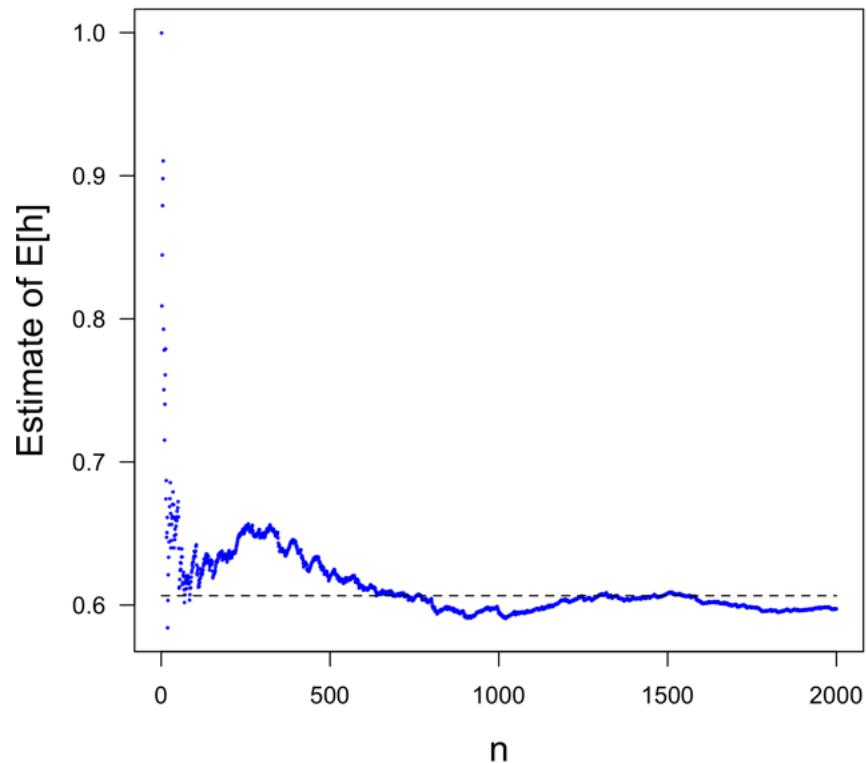


Figure E.1: Monte Carlo estimates of the integral  $I_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos x e^{-x^2/2} dx = e^{-1/2} \approx 0.6065$  (blue dots) and the analytical solution (black broken line). [MonteCarloIntegrationEx1.R](#)

The Monte Carlo method can be applied to (almost) any integral over finite intervals

$$I = \int_a^b h(x) dx \quad (\text{E.5})$$

by multiplying the integrand by the uniform PDF over the interval  $[a, b]$   $\mathcal{U}_{a,b}$  times  $(b - a)$

$$I = \int_a^b h(x) dx = (b - a) \int_a^b h(x) \mathcal{U}_{a,b}(x) dx. \quad (\text{E.6})$$

The integral can be estimated by the expectation of  $h$  over  $\mathcal{U}_{a,b}$  times  $(b - a)$ , i.e.

$$\hat{I} = (b - a) \frac{1}{n} \sum_{k=1}^n h(x_k) \quad (\text{E.7})$$

where the  $x_k$  are random sample points from  $\mathcal{U}_{a,b}$ .

**Example:**

$$I_2 = \int_0^2 \cos x dx = \sin x|_0^2 = \sin 2 \quad (\text{E.8})$$

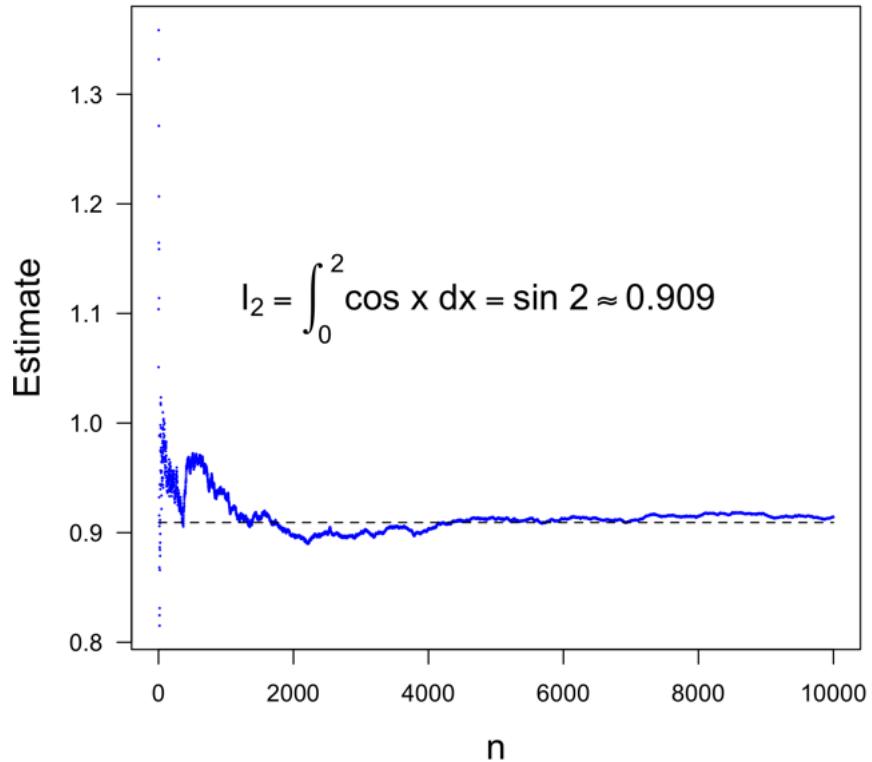


Figure E.2: Monte Carlo estimates of the integral  $I_2 = \int_0^2 \cos x dx = \sin 2 \approx 0.909$  (blue dots) and the analytical solution (black broken line). [MonteCarloIntegrationEx2.R](#)

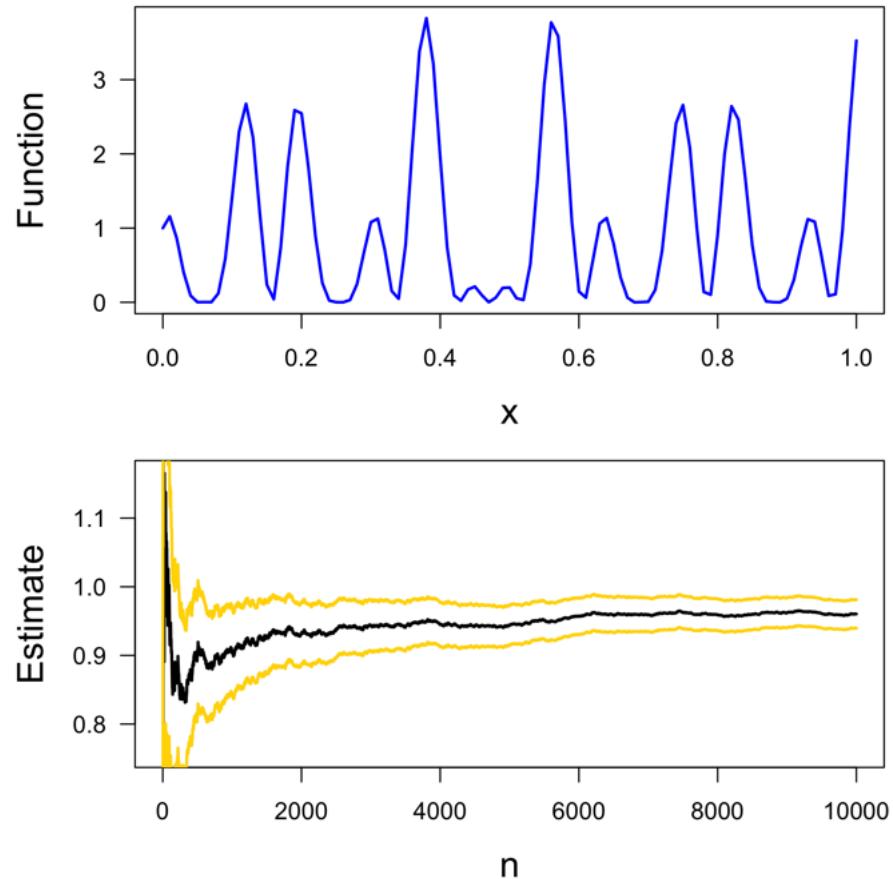


Figure E.3: Evaluation of the integral  $\int_0^1 [\cos(50x) + \sin(20x)]^2$  using Monte Carlo simulations (Robert & Casella, 2009, Fig. 3.3, modified). [MonteCarloIntegrationEx3.R](#)

### E.1.1 Importance sampling

Classical Monte Carlo integration can be generalized by multiplication of the integrand of Eq. E.1 by  $g(x)/g(x)$ , where  $g(x)$  is a PDF, and rearrangement of terms yielding

$$\mathbb{E}_f [(h(X))] = \int \frac{h(x) f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[ \left( \frac{h(x) f(x)}{g(x)} \right) \right]. \quad (\text{E.9})$$

This trick gives us the freedom to choose any PDF  $g(x)$  that could speed up the convergence of the estimate

$$\hat{\mathbb{E}}_g \left[ \left( \frac{h(x) f(x)}{g(x)} \right) \right] = \frac{1}{n} \sum_{k=1}^n \frac{h(x_k) f(x_k)}{g(x_k)} \quad (\text{E.10})$$

**Example:** Robert & Casella (2009)

#### Exercise 73 Monte Carlo integration

*Evaluate the integral*

$$\int_0^1 [\sin(38x) + \cos(23x)]^2 dx \quad (\text{E.11})$$

*using Monte Carlo simulations and compare the results with numerical integration using integrate() and with the analytical solution.*

## E.2 Monte Carlo: CDF of normal PDF

Robert & Casella (2009, p.67-68, Table 1)

The cumulative distribution function (CDF) is given by the integral

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \quad (\text{E.12})$$

Given the sample of size  $n$ ,  $(x_1, \dots, x_n)$  the integral can be approximated by Monte Carlo method

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^n \chi_{x_i \leq t} \quad (\text{E.13})$$

where  $\chi_{x_i \leq t}$  is the indicator function, i.e.  $\chi_{x_i \leq t} = 1$  when  $x_i \leq t$  and zero otherwise.

A few comments to the R code are in order:

1.  $10^8$  random data points are generated from the standard normal PDF
2.  $c(.5,.75,.8,.9,.95,.99,.999,.9999)$  is a list of CDF values (a CDF of 0.5 corresponds to  $x = 0$  of the standard normal PDF  $\mathcal{N}(x; \mu = 0, \sigma = 1)$ )
3.  $qnorm(t)$  calculates the corresponding  $x$  for the standard normal PDF ( $t = 0.5$  corresponds to  $x = 0$ ,  $t \rightarrow 1$  corresponds to  $x \rightarrow \infty$ ).
4. for  
 $t = \{0.5, 0.75, 0.8, 0.9, 0.95, 0.99, 0.999, 0.9999\}$  one obtains  $x = \{0.0, 0.67, 0.84, 1.28, 1.64, 2.33, 3.09, 3.72\}$  (stored in array 'bound')
5. 'res' is a matrix with 8 columns for the 8 different  $x$  (or corresponding  $t$  values) and 7 rows for the estimates based on 7 different sample sizes ( $n = 10^i$ ,  $i = 2, 3, \dots, 8$ )
6. Matrix 'res' is filled by means of the double loop with index  $i$  for sample size and index  $j$  for choice of  $x$  value.

7. The heart of the double loop is the statement

```
mean(x[1:10^i]<bound[j])
```

which stands for: take the first  $10^i$  random numbers (from  $10^8$  stored in 'x'), ask whether these values are smaller than the  $x$  value bound[j]; the answer to this question is either TRUE or FALSE whereby TRUE equivalent to 1, FALSE equivalent to 0; finally the mean of the sequence of 1's and 0's is calculated. In other words: we ask about the percentage of data points in the sample that are smaller than the  $x$  value bound[j]. Example: for  $x = 0$  the CDF is 0.5 and thus about half of the random data points should lie below  $x = 0$  and vice versa, i.e. if half the data points lie below  $x = 0$  the estimate for the CDF value is 0.5.

8. Here is an example:

```
set.seed(1953) # set seed for random number generators
x = rnorm(10)
# [1] 0.021924601 -0.904325473 0.413237769 0.186621223 0.230817687
# [6] 0.235680176 1.483738895 1.099462672 -0.004685482 -1.421197846
mean(x[1:5]<0.1) # 0.4
x[1:5] # 0.0219246 -0.9043255 0.4132378 0.1866212 0.2308177
x[1:5]<0.1 # TRUE TRUE FALSE FALSE FALSE
# i.e. 2 TRUEs out of 5 -> 2/5 = 0.4
```

9. Please note that in Eq. E.13 the condition is ' $x_i$  smaller than or equal to  $t'$  whereas in the R code it is ' $x_i$  smaller than  $t'$ . However, for random real number as sample this difference has practically no impact on the final results.

### E.3 Estimate $t$ distribution by Monte Carlo simulation (alternative code)

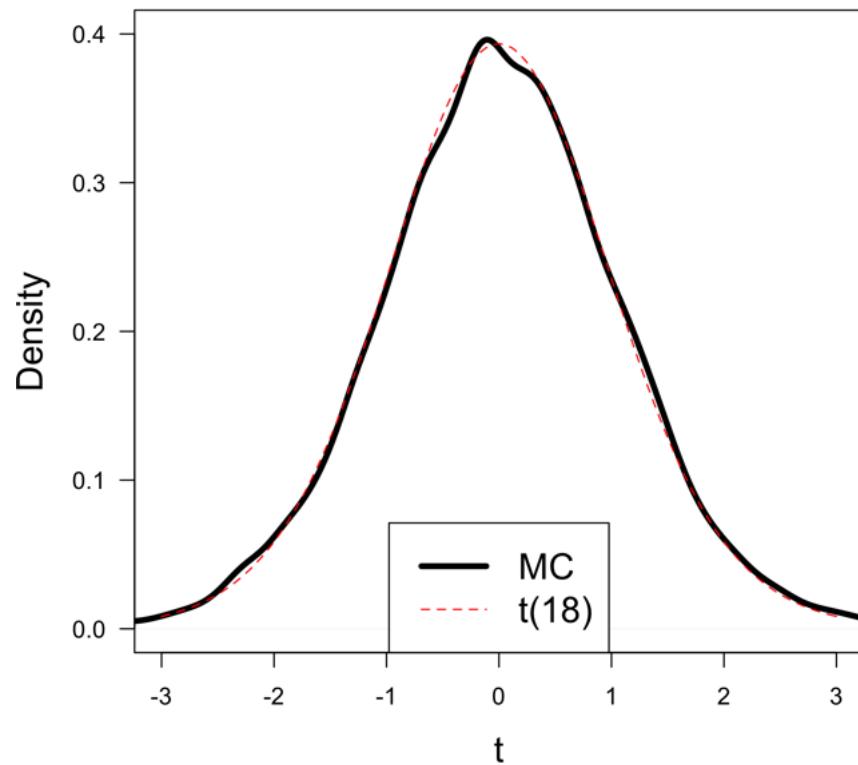


Figure E.4: Monte Carlo estimate of the  $t$  PDF. [Est-t-PDF2.R](#); [tstatisticFct.R](#)



# Appendix F

## Point estimators

### F.1 Estimate mean & variance (analytical results)

#### F.1.1 Unbiased estimators for $\mu$ & $\sigma^2$

Casella & Berger (2002, p. 213-214, Theorem 5.2.6):

**Theorem** "Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Then

- a.  $E\bar{X} = \mu$ ,
- b.  $\text{Var } \bar{X} = \frac{\sigma^2}{n}$ ,
- c.  $ES^2 = \sigma^2$ .

Here,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is the sample mean and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  is the sample variance.

**Remarks:**

1. This theorem is most remarkable because it is independent of the PDs or PDFs involved.
2. a. is telling us: The sample mean is an unbiased estimator for the population mean.
3. b. is telling us: The variance of the estimate of the population mean by the sample mean decreases by  $1/n$ , or, if we call in this context the standard deviation  $\sigma$  the 'uncertainty', the uncertainty of the estimate of the population mean decreases by  $1/\sqrt{n}$ .
4. c. is telling us: The sample variance  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , i.e. dividing the sum of squares  $\sum_{i=1}^n (X_i - \bar{X})^2$  by  $(n-1)$ , is an unbiased estimator of the population variance  $\sigma^2$ .

#### Application of the theorem

1. Normal PDFs: the average estimate of the mean should approach  $\hat{\mu}_a \rightarrow \mu \pm \sigma/\sqrt{n}$  and thus for the standard normal PDF ( $\mu = 0, \sigma = 1$ ) and sample size  $n = 5$   $\hat{\mu}_a \rightarrow 0 \pm 0.447$ .
2. Poisson PD with mean rate  $\lambda = 2.7$  and sample size  $n = 5$ : the average estimate of the mean should approach  $\hat{\mu}_a = \hat{\lambda} \rightarrow \lambda \pm \sqrt{\lambda}/\sqrt{n} = 2.7 \pm 0.7348$

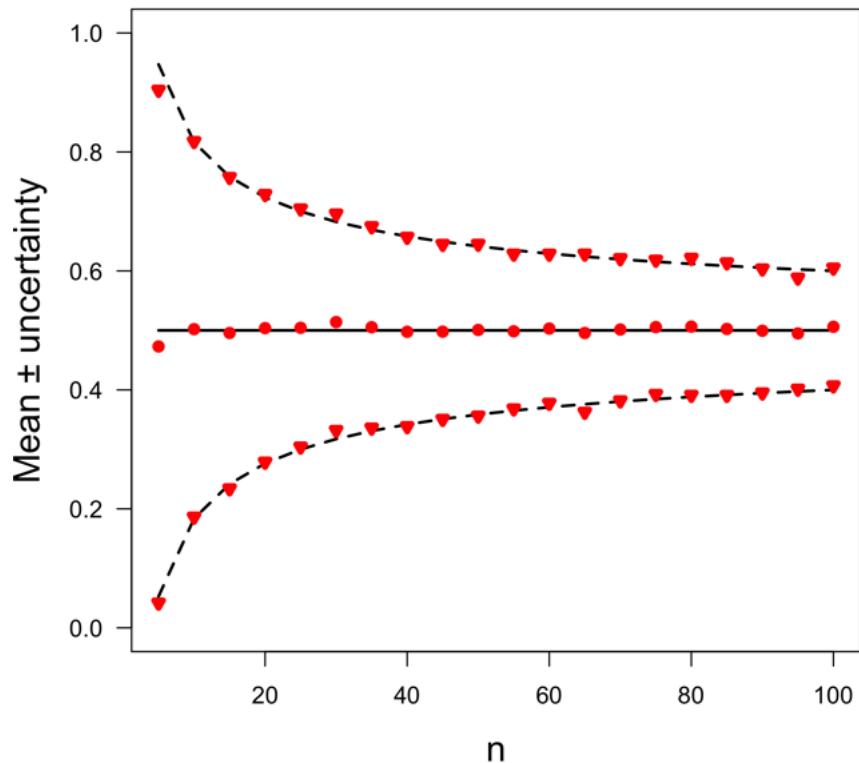


Figure F.1: Illustration of Casella & Berger (2002, p. 213-214, Theorem 5.2.6) using Monte Carlo simulations with  $M = 10^3$  runs each; sampling from a normal distribution with mean  $\mu = 0.5$  and standard deviation  $\sigma = 1$  should yield estimates of the mean approaching  $\hat{\mu}_a = \mu \pm \sigma/\sqrt{n} = 0.5 \pm 1/\sqrt{n}$  (solid black line  $\pm$  broken black lines); the Monte Carlo simulations yield estimates (red dots  $\pm$  red triangles) that are very close to the expected results. [PointEstMeanUncertaintyTheorem.R](#)

## F.2 Estimators for the standard deviation (\*)

'It is well known that unbiasedness of an estimator is not generally invariant under parameter transformations' (Gurland & Tripathi, 1971). One can show that

$$\hat{\sigma}_b = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2} \quad (\text{F.1})$$

is a biased estimator of the standard deviation. Holzman (1950) gave the following correction factor

$$C_n = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\frac{n-1}{2}} \quad (\text{F.2})$$

resulting in an unbiased estimator of the standard deviation

$$\hat{\sigma} = \sqrt{\frac{C_n}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}. \quad (\text{F.3})$$

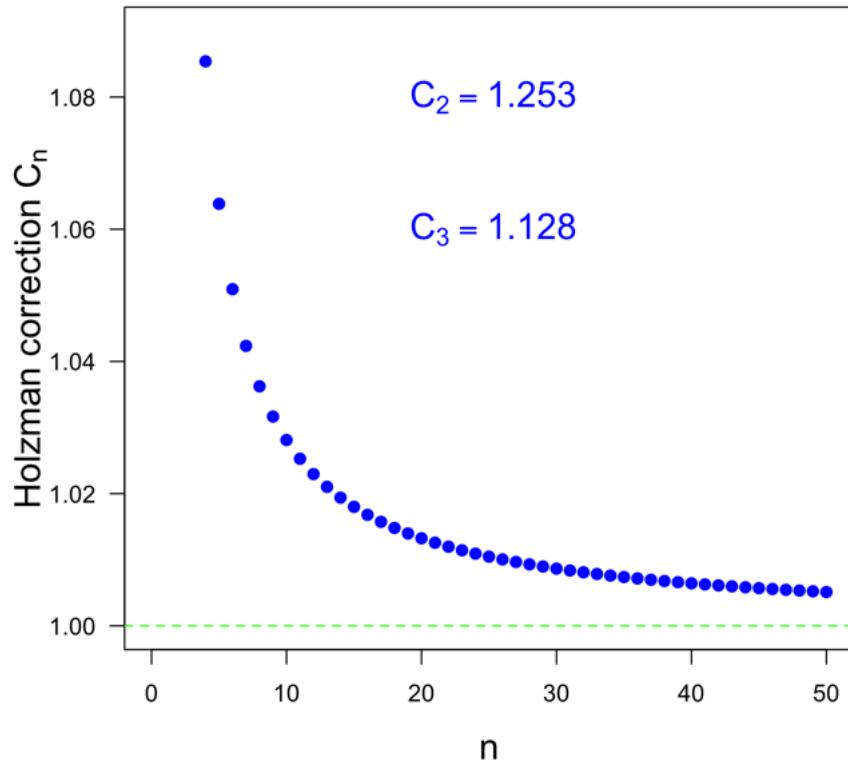


Figure F.2: The correction factor  $C_n$  as a function of  $n$  (Eq. F.2; Holzman, 1950) [PointEstSdHolzman.R](#)

## F.3 How to find point estimators? Examples

As mentioned already in Chapter 10, point estimators can be found by (1) intuition, (2) the method of moments, (3) the maximum likelihood approach, (4) the Bayesian approach, or by (5) the expectation-maximization (EM) algorithm. Here, we will give a few examples which are mainly based on Casella & Berger (2003).

### F.3.1 Methods of moments

The basic idea of the methods of moments is to equate the first  $j$  moments over the population with those over the sample and to solve for the  $j$  unknown parameters of the population. We will discuss two examples: (1) normal PDFs and (2) binomial PDs.

#### Example 1: Mean $\mu$ & variance $\sigma^2$ of normal distributions

The normal distribution

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\text{F.4})$$

possesses two parameters, namely the mean  $\mu$  and the variance  $\sigma$ . Thus we have to calculate the first and second moments over the distribution

$$\mu_1 = \int_{-\infty}^{+\infty} x^1 \mathcal{N}(x; \mu, \sigma) dx = \mu \quad (\text{F.5})$$

$$\mu_2 = \int_{-\infty}^{+\infty} x^2 \mathcal{N}(x; \mu, \sigma) dx = \mu^2 + \sigma^2 \quad (\text{F.6})$$

and over the sample

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i^1 = \bar{x} \quad (\text{F.7})$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n x_i^2. \quad (\text{F.8})$$

Equating of  $m_1$  with  $\mu_1$  and  $m_2$  with  $\mu_2$  leads to

$$\bar{x} = \mu, \quad (\text{F.9})$$

$$\frac{1}{n} \sum_{i=1}^n x_i^2 = \mu^2 + \sigma^2. \quad (\text{F.10})$$

Solving these two equations for  $\mu$  and  $\sigma^2$  leads to the following estimators

$$\hat{\mu} = \bar{x}, \quad (\text{F.11})$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (\text{F.12})$$

These estimators coincides with our intuition. The sample mean is a good estimator (unbiased) and is the commonly used estimator for the mean also for other distributions. However, the estimator for the variance is biased and should not be used for small sample sizes. It can be improved by division by  $n - 1$  instead of by  $n$ .

**Example 2: Number of trials  $k$  & probability of success  $p$  in single trials for binomial distributions**

The binomial distribution

$$\text{Binomial}(x; k, p) = \binom{k}{x} p^x (1-p)^{k-x} \quad (\text{F.13})$$

possesses two parameters, namely the number of trials  $k$  and the probability of success  $p$  in single trials. Thus we have to calculate the first and second moments over the distribution

$$\mu_1 = \sum_{x=0}^k x \text{Binomial}(x; k, p) = k p \quad (\text{F.14})$$

$$\mu_2 = \sum_{x=0}^k x^2 \text{Binomial}(x; k, p) = k p (1-p) + k^2 p^2 \quad (\text{F.15})$$

and over the sample

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i^1 = \bar{x} \quad (\text{F.16})$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n x_i^2. \quad (\text{F.17})$$

Equating of  $m_1$  with  $\mu_1$  and  $m_2$  with  $\mu_2$  leads to

$$\bar{x} = k p, \quad (\text{F.18})$$

$$\frac{1}{n} \sum_{i=1}^n x_i^2 = k p (1-p) + k^2 p^2. \quad (\text{F.19})$$

Solving these two equations for  $k$  and  $p$  leads to the following estimators

$$\hat{k} = \frac{\bar{x}^2}{\bar{x} - \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (\text{F.20})$$

$$\hat{p} = \frac{\bar{x}}{\hat{k}}. \quad (\text{F.21})$$

One problem with these estimators is that  $\hat{k}$  and thus also  $\hat{p}$  can yield negative values which, of course, makes no sense at all. This is the case (compare denominator in Eq. F.20) when the variance of the data, calculated by  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ , is larger than the sample mean.

**Remarks:** (Casella & Berger, 2003, p. 312)

- (1) The method of moments is 'quite simple to use and almost always yields some sort of estimate'.
- (2) The resulting estimators are often not very good and thus need further improvements.

**Exercise 74 Binomial estimator: method of moments**

The following binomial estimators have been derived by the method of moments:

$$\hat{k} = \frac{\bar{x}^2}{\bar{x} - \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (\text{F.22})$$

$$\hat{p} = \frac{\bar{x}}{\hat{p}} \quad (\text{F.23})$$

where  $\bar{x}$  is the sample mean. Investigate the quality of these estimators by a Monte Carlo simulation with  $M = 10^3$  runs using samples from a binomial distribution with  $k = 10$ , probability  $p = 0.4$  for success in single trials, and sample size  $n = 12$ . What's the percentage of negative values? What's the percentage of estimates  $\hat{k} > 20$ ? Remove the negative values and the values  $\hat{k}$  larger than 20, then calculate the mean values of the estimates  $\hat{k}$  and  $\hat{p}$ .

### F.3.2 Maximum likelihood estimators (MLEs) (\*)

The basic idea of maximum likelihood estimation (MLE) is to derive the likelihood function and to maximize this function or its logarithm by varying the unknown population parameters. Construction of the likelihood function is easy for independent data because it is just the product of the likelihood functions for single data points. For continuous parameters differentiation is the method of choice for finding the maximum. Finding the maximum in case of discrete parameters can be more difficult, especially when in combination with continuous parameters. Maximum likelihood estimators possess the nice invariance property, i.e., informally speaking, if  $\hat{\theta}$  is the MLE of  $\theta$  than  $\tau(\hat{\theta})$  is the MLE of  $\tau(\theta)$  where  $\tau(\theta)$  can be any function of  $\tau$  (Casella & Berger, 2003, p. 319-321). Unfortunately, MLE are often biased (this can be proven for all MLEs based on likelihood functions where the location of the global maximum is identical to the mean, i.e., for instance, for one-dimensional functions symmetric about the maximum.) We will discuss a few examples (based on Casella & Berger (2003)).

**Example 1: normal distributions** Let  $x = \{x_1, \dots, x_n\}$  be  $n$  independent data from the normal distribution

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\text{F.24})$$

with unknown mean,  $\mu$ , and unknown variance,  $\sigma^2$ . The likelihood to observe the single data point  $x_i$  is

$$\mathcal{L}(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (\text{F.25})$$

If the data are independent of each other, the likelihood for observing  $x = \{x_1, \dots, x_n\}$  is the product of the single data point likelihoods:

$$\mathcal{L}(x; \mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \prod_{i=1}^n e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2} \quad (\text{F.26})$$

A switch of perspective yields the corresponding likelihood function

$$L(\mu, \sigma^2; x) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2} \quad (\text{F.27})$$

The logarithm of the likelihood function is the log likelihood

$$\log L(\mu, \sigma^2 | x) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2. \quad (\text{F.28})$$

Because  $\log$  is a monotonically increasing function, the maxima of  $L()$  and  $\log L()$  are at the same location. We will now find the maximum of  $\log L(\mu, \sigma^2; x)$  by varying  $\mu, \sigma^2$ . The vanishing of the partial derivatives of  $\log L(\mu, \sigma^2; x)$  with respect to  $\mu$  and  $\sigma^2$  is a necessary condition for a maximum<sup>1</sup>

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2 | x) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i-\mu) = \frac{n}{\sigma^2} (\bar{x}-\mu) = 0 \quad (\text{F.29})$$

and

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 | x) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i-\mu)^2 = 0 \quad (\text{F.30})$$

---

<sup>1</sup>This could be also a minimum or an inflection point. Even if it is a maximum, one has to show that it is a global maximum.

Because the variance is an always positive quantity, the partial derivative with respect to  $\mu$  can only vanish when  $\hat{\mu}$  is equal to the sample mean  $\bar{x}$  (no surprise!). The second equation yields

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (\text{F.31})$$

i.e. the 'intuitive' estimator that is, unfortunately, biased. The proof that this is actually the global maximum can be found in Casella & Berger (2003, p. 321-322).

**Example 2: Poisson distribution** Let  $x = \{x_1, \dots, x_n\}$  be  $n$  independent data from the Poisson distribution

$$\mathcal{P}(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (\text{F.32})$$

with a single parameter, namely the mean rate  $\lambda$ . The likelihood function for independent data is derived by multiplication of single data likelihoods and switch of perspective:

$$L(\lambda|x) = e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \quad (\text{F.33})$$

The log likelihood is given by

$$\log L(p|x) = -n\lambda + \sum_{i=1}^n [x_i \log \lambda - \log x_i!] \quad (\text{F.34})$$

The vanishing of the first derivative of the log likelihood with respect to  $\lambda$  is a necessary condition for a maximum:

$$\frac{d}{d\lambda} \log L(p|x) = -n + \frac{1}{\lambda} \underbrace{\sum_{i=1}^n x_i}_{=n\bar{x}} = 0 \quad (\text{F.35})$$

Solving this equation leads to the estimator

$$\hat{\lambda} = \bar{x} \quad (\text{F.36})$$

which is a **best unbiased estimator** (Casella & Berger, 2003, p. 339).

# Appendix G

## Parameter estimation: the Bayesian approach (Appendix)

### G.1 Variance of a normal population for known mean value: conjugate prior (\*)

The estimation of the variance  $\sigma^2$  of a normal population when the mean value is known, i.e.  $\mu = \mu_0$  has been discussed already in Section 11.3 and Exercise 36) using the Jeffreys prior  $1/\sigma^2$ . Here we will follow Gelman et al. (2020) in using the scaled inverse- $\chi^2$  PDF as conjugate prior.

The likelihood can be derived like in Section 11.2 (compare (Eq. 11.10) with slight modifications):

$$L(\mathbf{y}|\mu_0, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{k=1}^n \frac{(y_k - \mu_0)^2}{2\sigma^2}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \quad (\text{G.1})$$

with sample variance  $s^2 = \frac{1}{\nu} \sum_{k=1}^n (y_k - \mu_0)^2$  and  $\nu = n - 1$ . The corresponding likelihood function reads

$$LF(\sigma^2|\mathbf{y}, \mu_0) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \quad (\text{G.2})$$

$$\propto \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{\nu s^2/2}{\sigma^2}\right) \quad (\text{G.3})$$

where the dependency on  $\nu = \sigma^2$  is the same as for inverse-gamma distributions. The product of two inverse-gamma distributions yields after normalization again to an inverse-gamma distribution. Thus if we choose an inverse-gamma distribution as prior the posterior will be an inverse-gamma distribution as well, i.e. obtain a simple analytic result and solve the normalization problem at the same time. The prior with hyperparameters  $(\alpha, \beta)$

$$p(\sigma^2) \propto \left(\sigma^2\right)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \quad (\text{G.4})$$

in combination with the likelihood function leads to the posterior

$$p(\sigma^2|\mathbf{y}, \mu_0) \propto LF(\sigma^2|\mathbf{y}, \mu_0) p(\sigma^2) \quad (\text{G.5})$$

$$= \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{\nu s^2/2}{\sigma^2}\right) \left(\sigma^2\right)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \quad (\text{G.6})$$

$$= \left(\sigma^2\right)^{-(\alpha+n/2+1)} \exp\left(-\frac{\nu s^2/2 + \beta}{\sigma^2}\right) \quad (\text{G.7})$$

which leads to the inverse-gamma PDF

$$p(\sigma^2 | \gamma, \delta) = \frac{\delta^\gamma}{\Gamma(\gamma)} (\sigma^2)^{-(\gamma+1)} \exp\left(-\frac{\delta}{\sigma^2}\right) \quad (\text{G.8})$$

with  $\gamma = \alpha + n/2$  and  $\delta = \nu s^2 / 2 + \beta$ .

# Appendix H

## Hypothesis testing (Appendix)

### H.1 Fisher's scale of evidence against the null hypothesis

- $\alpha = 0.05 < p < \alpha = 0.1$  borderline
- $\alpha = 0.025 < p < \alpha = 0.05$  moderate
- $\alpha = 0.01 < p < \alpha = 0.025$  substantial
- $\alpha = 0.005 < p < \alpha = 0.01$  strong
- $\alpha = 0.001 < p < \alpha = 0.005$  very strong
- $p < \alpha = 0.001$  overwhelming

(Efron & Gous, 2001)

### H.2 Two-sample $t$ test

One of the most common questions is whether or not two samples stem from distributions with the same true mean values, i.e.  $\mu_1 = \mu_2$ . If the data of both samples are independent and identically distributed (i.i.d., no serial correlation) and stem from populations with equal variances one can apply the Bayesian two sample  $t$  test (Rouder et al., 2009; R routine `ttestBF()` from package **BayesFactor**).

**Example 1:** We generate two random samples of size  $n_1 = 30$  and  $n_2 = 32$ , respectively, from the normally distributed population with true mean  $\mu = 1.8$  and true standard deviation  $\sigma = 1.2$ , i.e. the null hypothesis  $H_0$ : ' $\mu_2 - \mu_1 = 0$ , normal distributions with equal variances' is true. We apply 3 different tests: conventional  $t$ -test, Welch  $t$ -test, and Bayesian  $t$ -test (2 different routines from R package 'BayesFactor' and one 'pedestrian' code based on Rouder et al., 2009):

1. Conventional  $t$ -test:  $p = 0.7788$ ,  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$  (no surprise!)
2. Welch  $t$ -test:  $p = 0.7787$  is only slightly different from the  $p$ -value of the conventional  $t$ -test (this is no surprise given that the two samples stem from the same population); again:  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$ .
3. Applying the routine `ttestBF()` from package **BayesFactor** results in a Bayes factor  $B_{10} = 0.1987$  (or  $B_{01} = 1/B_{10} = 5.03$ ) which yields, according to Jeffreys' scales of evidence (Subsection 12.1.4), 'substantial evidence against  $H_1$ ' and 'substantial evidence for  $H_0$ '.

4. The 'pedestrian way' of calculating the Bayes factor is based on Eq. (1) of Rouder et al., 2009. For this 2-sample  $t$ -test one has to provide (i) the  $t$  statistic of the conventional  $t$ -test, (ii) the effective sample size  $n = n_1 \cdot n_2 / (n_1 + n_2)$ , and (iii) the degrees of freedom  $\nu = n_1 + n_2 - 2$ . The Bayes factor  $B_{10} = 0.1987$  is identical to the one provided by the routine **ttestBF()**.
5. The package **BayesFactor** contains a second routine for a Bayesian  $t$ -test, namely **meta.ttestBF** which requests (i) the  $t$  statistic of the conventional  $t$ -test and (ii) the two sample sizes  $n_1$  and  $n_2$ . The Bayes factor  $B_{10} = 0.1987$  is identical to the one provided by the routine **ttestBF()**.

**Example 2:** This is a modification of Example 1 where we will violate one the prerequisites of the conventional  $t$  test by specifying two different true variances. We generate two random samples of size  $n_1 = 30$  and  $n_2 = 32$ , respectively, from the normally distributed populations with true mean  $\mu = 1.8$  and true standard deviations  $\sigma_1 = 1.2$  and  $\sigma_2 = 2.4$ , respectively. The goal is to see whether the tests are robust against the violation of the equal variances prerequisite.

1. Conventional  $t$ -test:  $p = 0.724$  is a little bit smaller than in Example 1, however,  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$ .
2. Welch  $t$ -test:  $p = 0.719$  is still only slightly different from the  $p$ -value of the conventional  $t$ -test; again:  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$ .
3. Applying the routine **ttestBF()** from package **BayesFactor** results in a Bayes factor  $B_{10} = 0.203$  (or  $B_{01} = 1/B_{10} = 4.93$ ) which yields, according to Jeffreys' scales of evidence (Subsection 12.1.4), 'substantial evidence against  $H_1$ ' and 'substantial evidence for  $H_0$ '.

Thus the changes of  $p$ -values and Bayes factor compared to Example 1 are relative small and the decision of not rejecting  $H_0$  has not changed although the true variances differ by a factor of 4. The tests seem to be quite robust against such a difference in variances; this may change with sample size and variance ratio (could be studied as an exercise).

**Example 3:** We will now look at what happens when  $H_0$  is false. We generate two random samples of size  $n_1 = 30$  and  $n_2 = 32$ , respectively, from the normally distributed population with true means  $\mu_1 = 1.8$  and  $\mu_2 = 1.2$ , respectively, and true standard deviation  $\sigma = 1.2$ , i.e. the null hypothesis  $H_0$ : ' $\mu_2 - \mu_1 = 0$ , normal distributions with equal variances' is false.

1. Conventional  $t$ -test:  $p = 0.0586$ ,  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$ , however, the difference between  $p$  and  $\alpha$  is small.
2. Welch  $t$ -test:  $p = 0.0584$  is only slightly different from the  $p$ -value of the conventional  $t$ -test (this is no surprise given that the two samples stem from populations with identical variances); again:  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$ .
3. Applying the routine **ttestBF()** from package **BayesFactor** results in a Bayes factor  $B_{10} = 1.005$  (or  $B_{01} = 1/B_{10} = 0.995$ ). This Bayes factor very close to 1 indicates that no decision can be made based on the data.

[BayesianHyp-2-sided-t-testApp.R](#)

## H.3 Equal variances? (ref)

Several tests are available for addressing the 'Equal variances?' question. The most often applied two-sample test is the variance ratio test where the variance ratio is the test statistic denoted by  $F$  (sometimes called F-test for short). The presentation of other variance tests fills quite some pages and is meant largely for reference.

### H.3.1 Equal variances? The two-sample two-tailed variance ratio test

Are the true variance  $\sigma_1^2$  and  $\sigma_2^2$  equal to each other? The null hypothesis  $H_0$  reads:

The samples stem from normal distributions with equal variances ( $\sigma_1^2 = \sigma_2^2$ ) whereby their mean values  $\mu_1$  and  $\mu_2$  can be different from each other ( $\mu_1 \neq \mu_2$ ).

The test statistic  $F$  is chosen as the ratio of the estimated variances of the samples

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{s_1^2}{s_2^2} = \frac{\frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (x_{1,k} - \bar{x}_1)^2}{\frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (x_{2,k} - \bar{x}_2)^2}. \quad (\text{H.1})$$

The test statistic  $F$  follows the  $F$ -distribution  $\mathcal{F}(x; \nu_1, \nu_2)$  (Section 6.4.4) where  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  are the degrees of freedom for the two samples with sample sizes  $n_1$  and  $n_2$ , respectively.<sup>1</sup>

The test statistic  $F$  is always non-negative and take on very large values ( $0 < F < \infty$ ). As a consequence, the  $F$ -distribution is not symmetric about the position of its maximum (in contrast to the normal or the  $t$ -distribution) and the definition of 'two-tailed' is a bit more tricky than for the  $t$ -test.

First, we will use artificial data from normal distributions in order to show how to apply the test and how to interpret the results. The variance ratio test is often applied although prerequisites of the null hypothesis ('sample from normal distributions') are obviously violated as, for example, when the observations can take on integer values only (implying a discrete probability distribution instead of a continuous probability density function). In Section H.3.4 we will discuss the consequences of this violation and suggest a better approach.

#### An example using artificial data

In the first example for the variance ratio test we will use artificial data generated by random sampling from two normal distributions with unequal means and equal variances, i.e. the null hypothesis is true:

$$x_1 = \{2.033, 0.644, 2.620, 2.280, 2.346, 2.354, 4.226\} \quad (\text{H.2})$$

$$x_2 = \{4.649, 2.993, 0.868, 2.801, 3.247\} \quad (\text{H.3})$$

with sample sizes  $n_1 = 7$  and  $n_2 = 5$ , respectively. Application of the variance ratio test in R is easy (1 line of code!): `var.test(x1,x2)`.

Discussion of output from `var.test(x1,x2)`:

1.  $F = 0.60129$  is the value of the test statistic.
2. 'num df = 6, denom df = 4' are the degrees of freedom  $\nu_1 = n_1 - 1 = 7 - 1 = 6$  for the numerator and  $\nu_2 = n_2 - 1 = 5 - 1 = 4$  for the denominator, respectively.

---

<sup>1</sup>Tests with test statistics that follow the  $F$ -distribution are called **F-tests**. Originally, the  $F$ -distribution was developed for the variance ratio test in the 1920ies by Fisher and thus the variance ratio test was the first F-test. One-way (one-factor) ANOVA (Section 12.3) is another example of an F-test.

3.  $p = 0.5497$  is the observed level of significance ('p-value'). How is  $p$  calculated? The test statistic  $F$  follows the  $F$ -distribution  $\mathcal{F}(x; \nu_1 = 6, \nu_2 = 4)$  (red line in the left panel of Fig. H.1). The  $p$ -value is usually calculated as the probability in the tail(s) of the distribution of the test statistic. In the case of the variance ratio test this depends on the observed value of the test statistic  $F$  which divides the area under the  $F$ -distribution into two parts ('left' and 'right'; left panel of Fig. H.1)). It could be either

$$p_{\text{left}} = \int_0^F \mathcal{F}(x; \nu_1 = 6, \nu_2 = 4) dx = 0.2749 \quad (\text{H.4})$$

or

$$p_{\text{right}} = \int_F^\infty \mathcal{F}(x; \nu_1 = 6, \nu_2 = 4) dx = 0.7251 = 1 - p_{\text{left}}. \quad (\text{H.5})$$

The observed  $F$  value of 0.60129 is actually smaller than the hypothesized value of  $F = 1$  (= equal variances) and thus it makes sense to consider values of 0.60129 and smaller (black area in left panel of Fig. H.1) as defining the tail yielding  $p_{\text{left}} = 0.2749$ . Two times  $p_{\text{left}}$  gives  $p = 0.5497$  which is identical to the p-value from `var.test()`.

**However**, for the null hypothesis of equal variances it does not matter which sample we call the first and which one the second. What happens when we change the order of the sample in the call of `var.test()`? **The call of `var.test(x2,x1)` yields a different value for the observed test statistic  $F$ , however, the same p-value. How is this possible?**

Discussion of output from `var.test(x2,x1)`:

1.  $F = 1.6631$  is the value of the test statistic.
2. 'num df = 4, denom df = 6' are the degrees of freedom  $\nu_2 = n_2 - 1 = 5 - 1 = 4$  for the numerator and  $\nu_1 = n_1 - 1 = 7 - 1 = 6$  for the denominator, respectively.
3.  $p = 0.5497$  is the observed level of significance ('p-value'). How is  $p$  calculated? The test statistic  $F$  follows the  $F$ -distribution  $\mathcal{F}(x; 4, 6)$  (blue line in the right panel of Fig. H.1). The  $p$ -value is usually calculated as the probability in the tail(s) of the distribution of the test statistic. It could be either

$$p_{\text{left}} = \int_0^F \mathcal{F}(x; 4, 6) dx = 0.7251 \quad (\text{H.6})$$

or

$$p_{\text{right}} = \int_F^\infty \mathcal{F}(x; 4, 6) dx = 0.2749 = 1 - p_{\text{left}}. \quad (\text{H.7})$$

The observed  $F$  value of 1.6631 is actually larger than the hypothesized value of  $F = 1$  (= equal variances) and thus it makes sense to consider values of 1.6631 and larger (black area in right panel of Fig. H.1) as defining the tail yielding  $p_{\text{right}} = 0.2749$ . Two times  $p_{\text{right}}$  gives  $p = 0.5497$  which is identical to the p-value from `var.test()`.

It is remarkable that we obtain the same  $p$ -value although we use different values for the observed test statistic  $F$  and different distribution functions (red and blue lines, respectively, in Fig. H.1; the order of the degrees of freedom makes a difference!). The discussion above would suggest using  $p_{\text{left}}$  when  $F < 1$  and  $p_{\text{right}}$  when  $F \geq 1$ . However, this can lead after multiplication by 2 to  $p$ -values larger than one which makes no sense. This can happen when the observed  $F$  is close to 1 and thus in cases where one expects large  $p$ -values and no rejection of the null hypothesis. **In order to avoid  $p$ -values above 1, one uses (and so does `var.test()`) always double the minimum of  $p_{\text{left}}$  and  $p_{\text{right}}$  as  $p$ -value. Consequences of this procedure for calculating the  $p$ -value are for samples of different sizes: (1) for  $F = 1$  the  $p$ -value is smaller than one and (2) the maximum  $p$ -value is obtained for  $F \neq 1$ .**

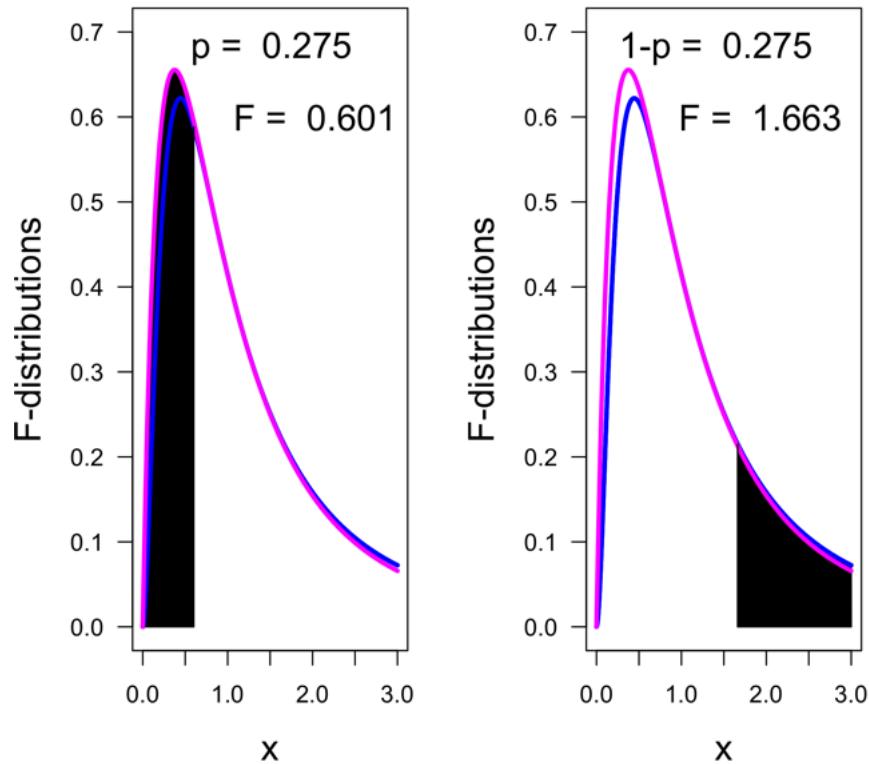


Figure H.1: Variance ratio test for the example with artificial data (see text). The F-distributions  $\mathcal{F}(x; \nu_1 = 6, \nu_2 = 4)$  (magenta solid line) and  $\mathcal{F}(x; 4, 6)$  (blue solid line) are different from each other (different order of degrees of freedom makes all the difference!). Left panel: calculation of the observed level of significance (p-value) in the case of a small (compared to the hypothesized value of 1) observed  $F$  value (here:  $F = 0.6013$ );  $p = \text{two times the area of the left tail}$  (black area under the magenta curve). Right panel: calculation of the observed level of significance (p-value) in the case of a large (compared to the hypothesized value of 1) observed  $F$  value (here:  $F = 1.6631$ );  $p = \text{two times the area of the right tail}$  (black area under the blue curve). The routine `var.test()` actually always uses double the minimum of  $p_{\text{left}}$  (left tail) and  $p_{\text{right}}$  (right tail) as p-value. [VarianceRatioTestExample.R](#)

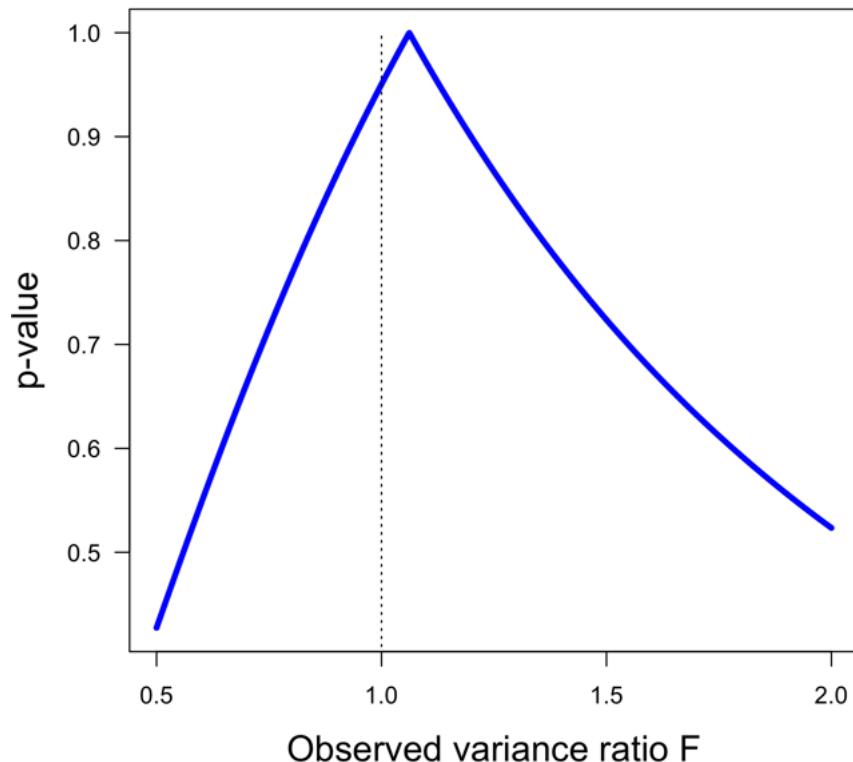


Figure H.2: The observed level of significance  $p$  (p-value) versus test statistic  $F$  for two samples of size  $n_1 = 7$  and  $n_2 = 5$ . For  $F = 1$  the  $p$ -value is smaller than one and the maximum  $p$ -value is obtained for  $F \neq 1$  (here: at  $\approx 1.062$ ). [VarianceRatioTestpOverF.R](#)

### H.3.2 Equal variances? Count data (\*)

Zar (2010, Example 8.7) Number of moths were caught repeatedly by two traps (1 & 2):

$$X_1 = \{41, 35, 33, 36, 40, 46, 31, 37, 34, 30, 38\} \quad (\text{H.8})$$

$$X_2 = \{52, 57, 62, 55, 64, 57, 56, 55, 60, 59\} \quad (\text{H.9})$$

We would like to know whether the variances are equal to each other. Zar formulates the following null and alternative (working) hypotheses:

'The two-tailed variance ratio test for the [null] hypothesis  $H_0: \sigma_1^2 = \sigma_2^2$  and [the alternative hypothesis]  $H_A: \sigma_1^2 \neq \sigma_2^2$ .' Note that in NHST only the null hypothesis is tested. This is the R code: [VarRatioTestCountData.R](#)

The variance ratio test assumes that the samples stem from normal distributions whereby their true mean values can be different from each other (compare Section H.3.3). Obviously, this is not the case here because single observations yield only integer values (count data; the number of moths in traps) and thus samples are from an unknown (discrete!) probability distribution. One can hope that the test is robust enough to deal with such a case; this can be checked by Monte Carlo simulations (Section H.3.4).

#### Interpretation of test results & decision

1. 'F test to compare two variances': the test statistic,  $F$ , of the variance ratio test is given by the ratio of the variances estimated from the samples:  $F = \hat{\sigma}_1^2 / \hat{\sigma}_2^2$ . The test statistic follows the  $F$ -distribution (Section 6.4.4) with appropriate degrees of freedom (see below).
2. 'F = 1.6956': the test statistic is the ratio of the two estimated variances, here:  $F = \hat{\sigma}_1^2 / \hat{\sigma}_2^2 = 21.87 / 12.90 = 1.696$
3. 'num df = 10': 10 (degrees of freedom of numerator) = 11 (number of data in 1. sample) - 1 (constraint: sample mean is necessary for estimating the variance  $\hat{\sigma}_1^2$ )
4. 'denom df = 9': 9 (degrees of freedom of denominator) = 10 (number of data in 2. sample) - 1 (constraint: sample mean is necessary for estimating the variance  $\hat{\sigma}_2^2$ )
5. 'p-value = 0.4401': the observed level of significance  $p = 0.4401$  is larger than the commonly chosen level of significance  $\alpha = 0.05 \Rightarrow$  the null hypothesis can not be rejected. The  $F$  value of 1.696, although quite different from 1, is not atypical given the null hypothesis is true. The  $p$ -value is given by (note the factor 2 for the 'two-sided' test):

$$p = 2 \int_{F=1.6956}^{\infty} F(x; \nu_1 = 10, \nu_2 = 9) dx \quad (\text{H.10})$$

6. 'alternative hypothesis: true ratio of variances is not equal to 1': or in other words the alternative (working) hypothesis states that 'the true variances are different from each other'; the negation of this hypothesis (complementary hypothesis) is the null hypothesis 'the true variances are equal to each other'.
7. '95 percent confidence interval: 0.4277543 6.4074588': if we choose  $\alpha = 0.05$  (corresponding to 5%) to define the rejection region, then the other  $100 - 5 = 95\%$  of the total probability correspond to a variance ratios between 0.4278 and 6.4074; the 'observed' variance ratio of 1.696 lies well within this range and thus  $H_0$  is not rejected.
8. 'sample estimates: ratio of variances 1.69556': this is identical to the test statistic  $F$ ;
9. Decision:  $p = 0.44 > 0.05 = \alpha \Rightarrow$  do not reject null hypothesis, i.e. data contain no evidence for different true variances.

### H.3.3 Monte Carlo estimate of the density for the test statistic $F$ (\*)

The variance ratio test is based on the assumption that the samples stem from normal distributions. The test statistic  $F$  of this test is the ratio of the estimated sample variances:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{s_1^2}{s_2^2} = \frac{\frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (x_{1,k} - \bar{x}_1)^2}{\frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (x_{2,k} - \bar{x}_2)^2}. \quad (\text{H.11})$$

The test statistic  $F$  follows the  $F$ -distribution  $\mathcal{F}(x; \nu_1, \nu_2)$  where  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  are the degrees of freedom for the two samples with sample sizes  $n_1$  and  $n_2$ , respectively.

The following Monte Carlo simulation will support the statement given above. Although this will yield no new insight, the R code can later be modified in order to estimate the distributions of the test statistic  $F$  under different assumptions about the populations (for example, discrete uniform instead of normal distributions, Section H.3.4). A few comments about the Monte Carlo simulation are in order.

1. Set a seed for the random number generators: `set.seed(1953)`
2. Create a R-function (denoted `my.Fsimulation()`) that generates two samples from normal PDFs with different mean values, however, equal variances as required by the null hypothesis. The sample sizes are  $n_1 = 11$  and  $n_2 = 10$  (identical to the sample sizes of moths data discussed above). The test statistic  $F$  is calculated from the artificial samples and returned by the function.
3. The function `my.Fsimulation()` is called  $10^6$  times and the results are stored in the array 'Fstat.vector'. The corresponding R code reads: `Fstat.vector=replicate(1e6,my.Fsimulation())`
4. Estimate the density from the data array Fstat.vector by applying the routine `density()` over the interesting range from 0 to 3 (range explicitly specified in the argument list of `density()`).
5. Calculate the difference between the estimated density values ( $y$ ) and the corresponding values of the  $F$ -distribution
6. Finally, plot the results, i.e. the Monte Carlo estimate, the  $F$ -distribution, and the difference between these two densities.

[The Monte Carlo estimates of the density for  \$F\$  are almost identical to the analytical values \(Fig. H.3\).](#)

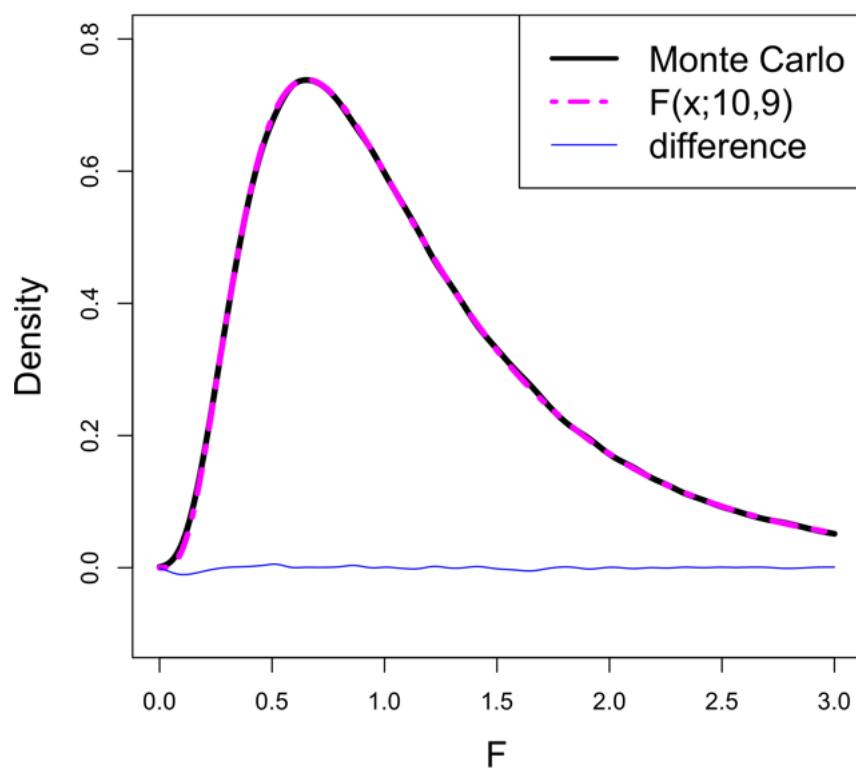


Figure H.3: Estimate of the density for the test statistic  $F$  using Monte Carlo simulations for artificial samples from normal distributions with equal variances (black line) and the  $F$ -distribution for  $\nu_1 = n_1 - 1 = 10$  and  $\nu_2 = n_2 - 1 = 9$  (dash-dotted magenta line). The difference between these two densities (blue line) is small.  
[MC-F-PDFs-VRnormal.R](#)

### H.3.4 Monte Carlo simulations: sampling from discrete uniform populations (\*)

Now we will modify Step 2 of the Monte Carlo simulations described in the previous subsection. Instead of normal distributions we will sample from (discrete!) uniform probability distributions for integer values. The observations of the first sample vary between 30 and 46 (range length = 46-30 = 16), and those of second sample between 52 and 64 (range length = 64-52 = 12). In order to estimate the density of the test statistic  $F$  under the null hypothesis we can use populations with different mean values, however, the variances have to be equal to each other.<sup>2</sup> Thus we should not use uniform probability distributions with differing length of ranges: the uniform PD of integers from 30 to 46 has a true variance that is different from the uniform PD of integers from 52 to 64. We can use, for example, uniform PD of integers from 30 to 46 and from 50 to 66, i.e. PDs with different mean values, however, same range length and thus same true variances. The result of the Monte Carlo simulation based on these assumptions yields an estimate of the density for  $F$  that is quite different from the  $F$ -distribution (Fig. H.4) and thus the naive application of `var.test()` is not advised because it will not yield correct  $p$ -values and thus can lead to wrong decisions (conclusions). In the case of an observed  $F$  value of 0.6 (moths trap example discussed above), the  $p$ -values for the two-sided variance ratio test based on the  $F$ -distribution ( $p = 0.55$ ) or based on the more appropriate density estimate from Monte Carlo simulations  $p_{MC} = 0.27$  differ largely, namely by a factor of 2. (Fig. H.4) Because  $p_{MC}$  is still larger than  $\alpha = 0.05$  we still come to the same conclusion: do not reject the null hypothesis. However, for other data sets and corresponding  $F$  values decisions based on  $p$  and  $p_{MC}$  can lead to different conclusions.

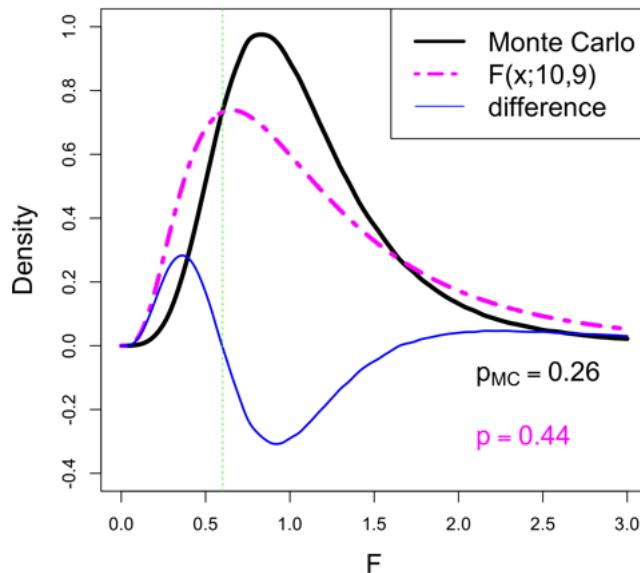


Figure H.4: Monte Carlo-based estimate of the density of  $F$  from samples from a discrete uniform distributions of integers between 30 and 46 and between 50 and 66, respectively (blue line). This density is quite different from the  $F$ -distribution (red line) that applies for sampling from normal distributions with equal variances. The difference between these two densities (black line) is large. An observed  $F$  value of 0.6 (green vertical line) will result in quite different  $p$ -values for the two-sided variance ratio test:  $p = 0.55$  based on the  $F$ -distribution and  $p_{MC} = 0.27$  based on the more appropriate density estimate from Monte Carlo simulations.

[MC-F-PDFs-VRTcount.R](#)

<sup>2</sup>The consequences of a violation of this variance prerequisite could be studied as well by using Monte Carlo simulations.

**Exercise 75 Variance ratio test for count data**

(1) Estimate the PDF of the test statistic  $F$  for sampling from two discrete uniform distributions  $\mathcal{P}_1$  with  $k = 10, 11, \dots, 26$ , sample size  $m = 11$  and  $\mathcal{P}_2$  with  $k = 33, 34, \dots, 49$ , sample size  $n = 10$  using Monte Carlo simulations.

(2) Calculate the p-values for observed  $F$ -values over the range  $0.1 \leq F \leq 0.5$  using (a) the estimated PDF from (1) and (b) the  $F$ -distribution with degrees of freedom  $v_1 = m - 1 = 10$ ,  $v_2 = n - 1 = 9$ .

(3) For which  $F$ -values are the p-values equal to  $\alpha = 0.05$ ?

### H.3.5 Equal variances? Two or more samples: Levene test

Levene (1960) proposed a robust procedure for testing the null hypothesis 'all variances are equal'. The test statistic called  $L$  (or  $W$  or  $F$ ) is defined by

$$L = \frac{n-k}{k-1} \frac{\sum_{i=1}^k n_i (y_i - \bar{y})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2} \quad (\text{H.12})$$

where  $n$  is the total number of data,  $k$  is the number of samples,  $x_{i,j}$  are the observations in sample  $i$ ,  $j = 1, 2, \dots, n_i$ ,  $n_i$  is the size of sample  $i$ ,

$$y_{i,j} = |x_{i,j} - \bar{x}_i| \quad \text{absolute deviation from sample mean} \quad (\text{H.13})$$

or

$$y_{i,j} = |x_{i,j} - \text{median}(x_i)| \quad \text{absolute deviation from sample median} \quad (\text{H.14})$$

and

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j} \quad (\text{H.15})$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j} \quad (\text{H.16})$$

The Levene test has been applied to the moths data set by Zar (2010, Example 8.9, p. 155-156) using the sample means. The resulting two-sided two-sided p-value is 0.49 (Zar gives a slightly different value of 0.48) and thus the null hypotheses ('no difference of variances') is not rejected on the significance level  $\alpha = 0.05$ . When using the sample medians instead of the sample means one obtains a slightly smaller p-value of 0.47 which leads to the same conclusion. The R package **car** contains the routine **leveneTest()**<sup>3</sup> which yields a p-value that is identical to the value obtained based on sample medians.

R code: [LeveneTest3.R](#)

The Levene test has been applied also to the data set of Zar (2010, Example 10.1) that has been discussed in the context of ANOVA (Section 12.3.1, Fig. 12.19). Application of the Levene test using **leveneTest()** yields  $p = 0.72$  and thus the null hypothesis ('all variances are equal') will not be rejected on the level  $\alpha = 0.05$ . This is no surprise because from looking at the boxplot (Fig. 12.19) one could not find strong indications of large differences between variances.

R code: [LeveneTestZar10d1Ex.R](#)

---

<sup>3</sup>The call of **leveneTest()** is not obvious and not well explained by the R help page; see the R code available via hyperlink for proper explanation.

### H.3.6 Equal variances? More than two samples: Bartlett test

Example: animal weights (Zar, 2010, Example 10.1)

$$\begin{aligned}x_1 &= \{60.8, 67.0, 65.0, 68.6, 61.7\} \\x_2 &= \{68.7, 67.7, 75.0, 73.3, 71.8\} \\x_3 &= \{69.6, 77.1, 75.2, 71.5\} \\x_4 &= \{61.9, 64.2, 63.1, 66.7, 60.3\}\end{aligned}$$

The Bartlett test can be called in **R** by providing the data as a 'list', which is appropriate because the samples can be of different size, or by loading the data using `read.table()` and calling `bartlett.test()` with arguments referencing to this special data format. The **R** codes below show both approaches. The null hypothesis ('all variances are equal') is not rejected because  $p = 0.92$  is much larger than the chosen level of significance  $\alpha = 0.05$ . Note that the  $p$ -value of the Bartlett test ( $p = 0.92$ ) is higher than the one from the Levene test ( $p = 0.72$ ).

**R** code: [BartlettTestZar10d1.R](#)

### H.3.7 Equal variances? More than two samples: Fligner-Killeen test (\*)

The Fligner-Killeen test is a non-parametric alternative to the Bartlett test. We will apply the Fligner-Killeen test to the animal weight data of Zar (2010, Example 10.1) used already for the illustration of the Bartlett test. The observed level of significance  $p = 0.57$  is smaller than that of the Bartlett test ( $p = 0.92$ ), however, still much larger than the chosen level of significance  $\alpha = 0.05$  and thus the null hypothesis ('all variances are equal') is not rejected.

**R** code: [Fligner-KilleenZar10d1.R](#)

## H.4 Goodness-of-fit test: Mendelian factors?

In our discussion of the one-sample test of the mean ( $\mu = \mu_0?$ , Section 12.1) we saw that, at least for large sample size  $n$ , false hypotheses are rejected based on the estimation approach of physicists as well as by the conventional  $t$ -test. And if  $H_0$  is rejected, it is tempting to accept the alternative hypothesis  $H_1 : \mu \neq \mu_0$  (depending on the context, often the discovery or at least confirmation of an 'effect') because together  $H_0$  and  $H_1$  cover the whole range of possibilities ( $H_0$  and  $H_1$  are complementary with respect to the hypothesized parameter  $\mu_0$ ). However, we have seen the problems with significance tests (they are not consistent, i.e. rejection rate for true  $H_0$  does not approach zero for large sample size).

The next example, where instead of a single parameter ( $\mu_0$ ) a distribution  $\mathcal{P}$  is hypothesized, shows a further complication for significance tests. The complementary hypothesis  $H_1 : \text{'the distribution is not } \mathcal{P}$ ' is unspecific and it is not obvious which quantity to estimate in order to get more insight. At least, rejected null hypotheses can indicate the existence of altered or additional processes, however, not in a specific way.

The example discussed below is taken from Fisher (1925), who never accepted a null hypothesis. And, guess what will happen with  $H_0$  here.

In a cross involving two Mendelian factors we expect by interbreeding the hybrid ( $F_1$ ) generation to obtain four classes in the ratio 9:3:3:1; the hypothesis in this case is that the two factors segregate independently, and that the four classes of offspring are equally viable. Are the following observations on Primula (de Winton and Bateson) in accordance with this hypothesis? The number of observed individuals are: 328 (flat leaves, normal eye), 122 (flat leaves, Primrose Queen eye), 77 (crimped leaves, Lee's eye), 33 (crimped leaves, Primrose Queen eye).

The analysis proceeds as follows:

1. The hypothesized ratios (9:3:3:1) are converted to a probability distribution

$$\mathcal{M}_2 = \{9, 3, 3, 1\} / (9 + 3 + 3 + 1) = \left\{ \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right\}.$$

2. The total number of individuals is  $328+122+77+33 = 560$ .
3. The expected number of individuals (expected frequencies  $f_{\text{exp}}$ ) are given by the probability distribution times the total number of individuals:

$$\left\{ \frac{9 \cdot 560}{16}, \frac{3 \cdot 560}{16}, \frac{3 \cdot 560}{16}, \frac{560}{16} \right\} = \{315, 105, 105, 35\} = f_{\text{exp}}$$

4. The observed (blue dots) and expected (red triangles) frequencies are shown in Fig. H.5. The observed and expected frequencies show a similar pattern (decreasing frequencies), however, the observed frequencies for cases 2 and 3 are quite different from each other although they are expected to be equal.

5. The test statistic is defined by

$$\chi^2 = \sum_{j=1} \frac{(f_{\text{obs}} - f_{\text{exp}})^2}{f_{\text{exp}}}$$

where  $f_{\text{obs}}$  and  $f_{\text{exp}}$  are the observed and expected frequencies, respectively. For the current example one obtains  $\chi^2 = 10.87$ .

6. The degrees of freedom are given by the number of cases (4) minus the number of constraints (1, namely the total number of individuals):  $\nu = 4 - 1 = 3$ . The  $p$ -value is calculated as the area of the right tail of the  $\chi^2$  PDF:  $p = (1-\text{pchisq}(q=\text{totalq}, df=\text{nu})) = 0.0125$ .

The  $p$ -value is smaller than the magic limit of  $\alpha = 0.05$  and thus we would reject the hypothesis. No clear alternative hypothesis has been formulated and the complementary hypothesis 'the ratios do not follow the Mendelian rule' is not very enlightening. Instead one should try finding a better hypothesis (Exercise 76).

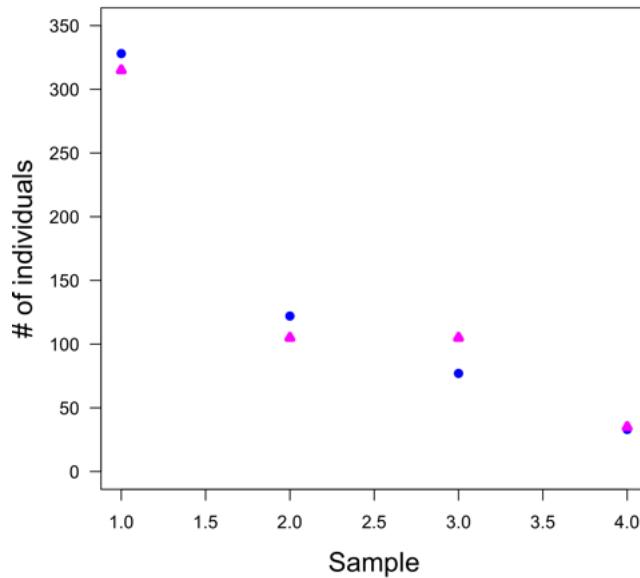


Figure H.5: The observed (blue dots) and expected (magenta triangles) frequencies: the observed and expected frequencies show a similar pattern (decreasing frequencies), however, the observed frequencies for cases 2 and 3 are quite different from each other although they are expected to be equal. [NHST-Mendel-Fisher.R](#)

**Exercise 76 Mendelian factors: modified hypothesis**

Let us consider a second hypothesis in relation to the same data as in the example discussed above (Section H.4), differing from the first in that we suppose that the plants with crimped leaves are to some extent less viable than those with flat leaves. Such a hypothesis could of course be tested by means of additional data, however, here we are only concerned with the question whether or no it accords with the values before us. The hypothesis tells us nothing of what degree of relative viability to expect; we therefore take the totals of flat and crimped leaves observed, and divide each class in the ratio 3:1.

(1) Calculate the expected number of individuals based on the total number of observed and the hypothesis just formulated.  
 (2) Compare the observed and expected values (plot). Do you believe the hypothesis is consistent with the data or would you reject the hypothesis?

(3) Calculate the test statistics  $\chi^2 = \sum_{j=1} \frac{(f_{\text{obs}} - f_{\text{exp}})^2}{f_{\text{exp}}}$  where  $f_{\text{obs}}$  and  $f_{\text{exp}}$  are the observed and expected frequencies, respectively.

(4) Calculate the p-value for this test statistic and the appropriate number of degrees of freedom.

(5) Make a decision based on the p-value and a chosen rejection limit of  $\alpha = 0.05$  ('Fisher's suggestion').

**Exercise 77 Sample from Poisson distribution?**

For the outcomes 0, 1, 2, ..., 12 the following frequencies have been observed:  
 $f_{\text{obs},0} = \{0, 20, 43, 53, 86, 70, 54, 37, 18, 10, 5, 2, 2\}$ .

(1) Estimate the average rate  $\lambda$ .

(2) Calculate the expected frequencies and compare them with the observed frequencies (plot). What is your guess concerning the hypothesis 'sample stems from a Poisson distribution'?

In order to apply the goodness-of-fit ( $\chi^2$ ) test the minimum frequency per case should be at least 5. Therefore both the first two (0 and 20) and the last three (5,2,2) values have been lumped together yielding  $f_{\text{obs},1} = \{20, 43, 53, 86, 70, 54, 37, 18, 10, 9\}$ . The lumped expected values are given by  $f_{\text{exp},1} = \{21.08, 40.65, 63.41, 74.19, 69.44, 54.16, 36.21, 21.18, 11.02, 8.66\}$ .

(3) How to derive the lumped expected values, especially the first and last value?

(4) Calculate the test statistic

$$\chi^2 = \sum_j \frac{(f_{\text{obs},1} - f_{\text{exp},1})^2}{f_{\text{exp},1}}$$

(5) Calculate the p-value and make a decision based on the magic  $\alpha = 0.05$  limit.

## H.5 Correlation significantly different from zero?

*Pearson's correlation coefficient is easy to calculate, however, 'Is the correlation coefficient significantly different from zero?'.*

The null hypothesis  $H_0$  reads 'the correlation between the two stochastic variables X and Y from normal distributions is zero'; its negation is the working hypothesis 'the correlation between the two stochastic variables X and Y is (significantly) different from zero'. As test statistic one uses

$$t_{\text{cor}} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad (\text{H.17})$$

(Fig. H.6) where  $r$  is the correlation coefficient (Pearson) between the samples  $x$  and  $y$ , and  $n$  is the sample size. For  $n \geq 6$  and samples from normal distributions,  $t_{\text{cor}}$  is approximately distributed according to Student's  $t$  distribution with  $v = n - 2$  degrees of freedom.

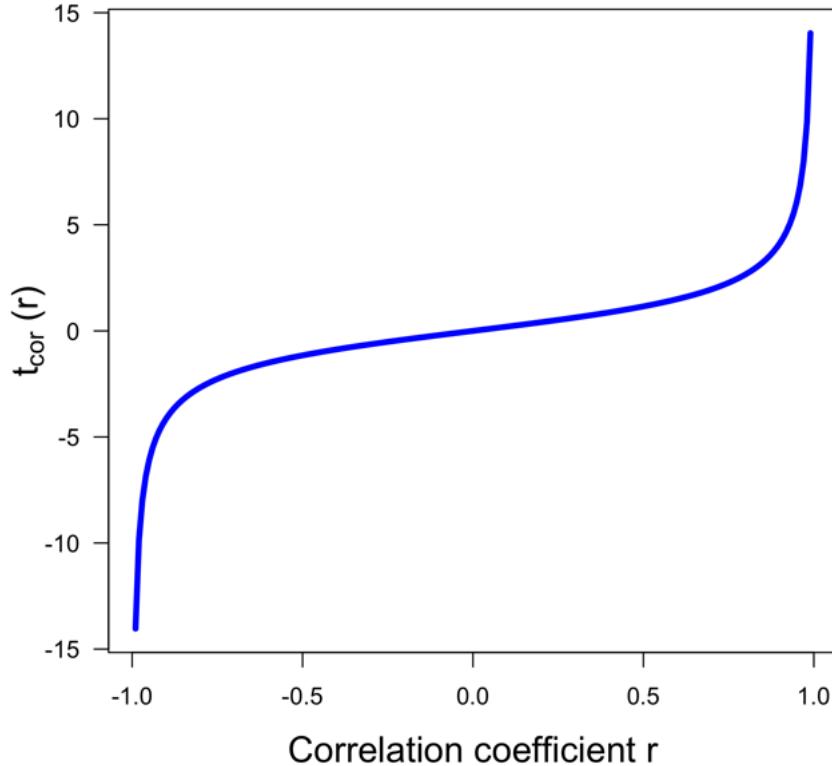


Figure H.6: Test statistic  $t_{\text{cor}}$  as a function of the correlation coefficient  $r$  (for sample size  $n = 6$ ): large values of  $|r|$  result in large values of  $|t_{\text{cor}}|$ . [CorTestStatistic.R](#)

For two data sets of size  $n = 8$  and correlation  $r = 0.6$  one obtains  $t_{\text{cor}} = 1.8371$  and  $p = 0.1158$ . Thus one would not reject the null hypothesis 'correlation is zero' although the observed value of  $r$  is 0.6. This result is typical for small sample sizes (Fig. H.7). For correlations to be significantly different from zero, larger sample sizes are required (Fig. H.8).

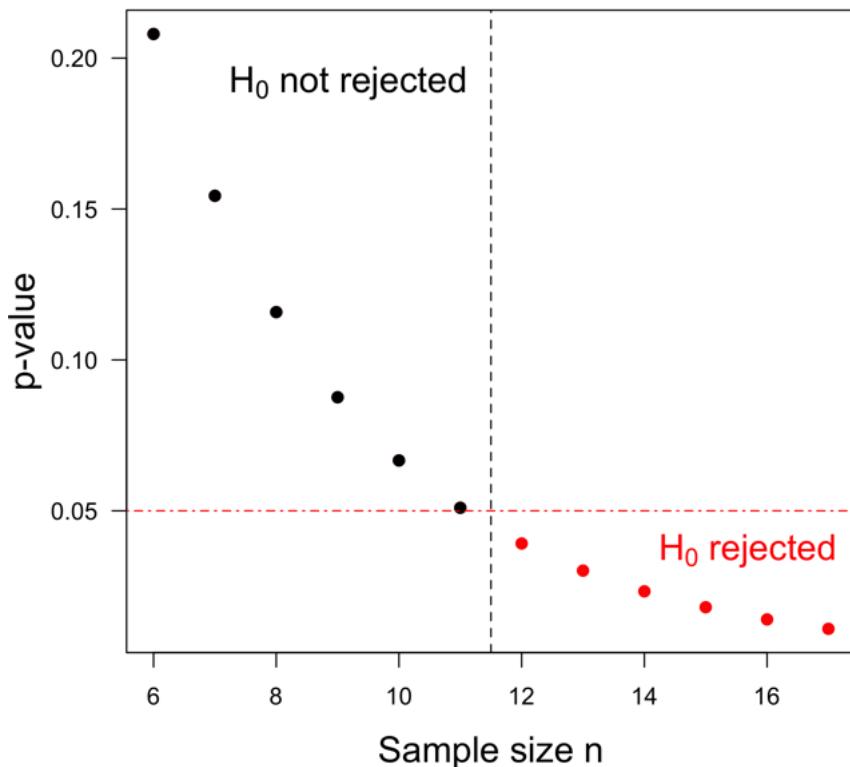


Figure H.7: Observed levels of significance ( $p$ -values) for the null hypothesis 'correlation is not different from zero' and an observed correlation of  $r = 0.6$  (positive correlation;  $r^2 = 0.36$ ) as function of sample size  $n$ : the null hypothesis can only be rejected on the level of significance  $\alpha = 0.05$  when the sample size  $n$  is larger than 11 (blue dots); whereas for smaller sample sizes  $H_0$  can not be rejected on the significance level  $\alpha = 0.05$  (red dots). [CorTestFct.R](#)

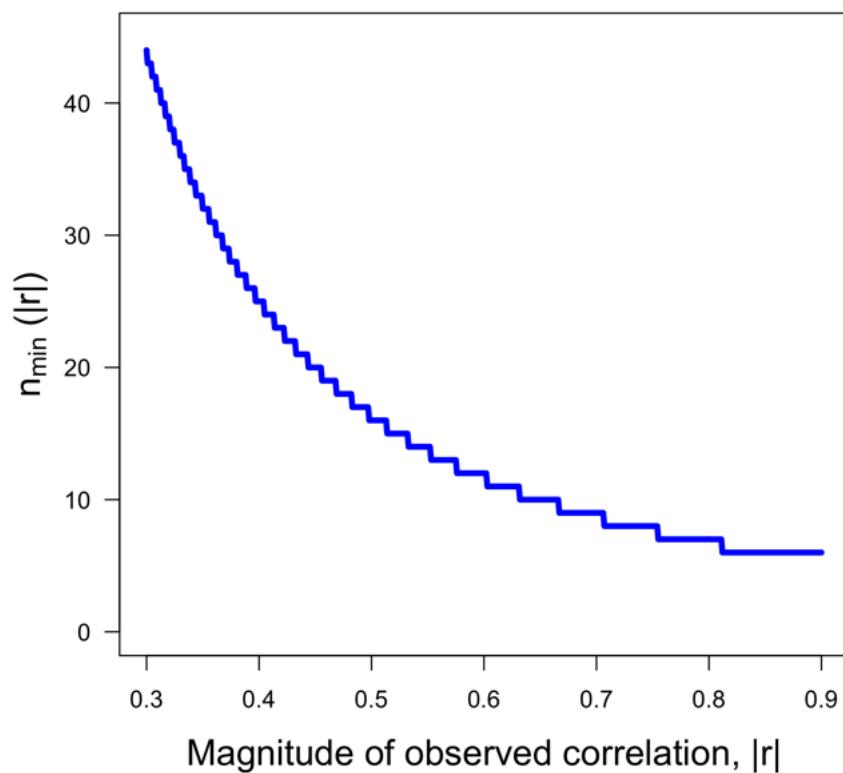


Figure H.8: Minimum sample size  $n_{\min}$  required to reject the null hypothesis 'correlation is zero' at  $\alpha = 0.05$  for observed values of correlation  $r$ : for small observed correlations large sample sizes are required to reject the null hypothesis (example:  $n_{\min} = 44$  for  $r = 0.3$ ). [CorTestMinN.R](#)

### H.5.1 Does the test statistic $t_{\text{cor}}$ follow the $t$ -distribution? (\*)

A Monte Carlo simulation is straight forward:

- Choose the sample size  $n$ ; here  $n = 6$ .
- Generate many (here  $M = 10^6$ ) sample pairs  $x, y$  from normal distributions.
- Calculate the correlation coefficient  $r$  between  $x$  and  $y$ .
- Calculate the test statistic  $t_{\text{cor}}$ .
- Estimate the density (= probability density function) from the  $M t_{\text{cor}}$  values and compare with the appropriate (degrees of freedom  $\nu = n - 2$ )  $t$ -distribution.

Monte Carlo simulations yield a density estimate that is very close to the  $t$ -distribution (Fig. H.9).

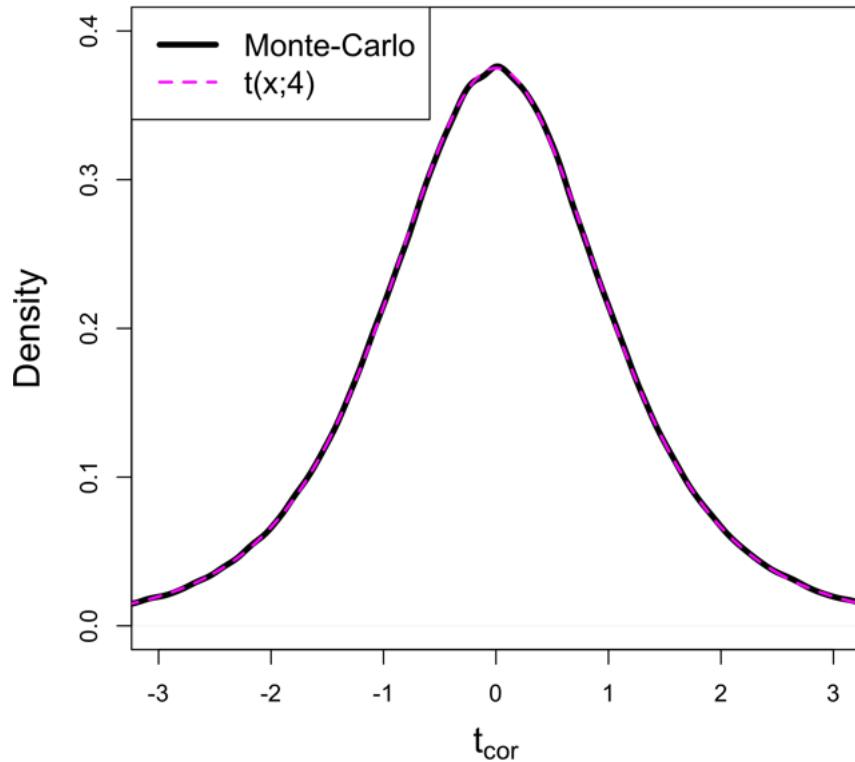


Figure H.9: Estimate of the density of  $t_{\text{cor}}$  from Monte Carlo simulations with  $M = 10^6$  runs for sample size  $n = 6$  (black solid line) and the  $t$ -distribution for  $\nu = n - 2 = 4$  degrees of freedom (magenta broken line). [CorTestMCTPDF.R](#)

**Further reading:** Best & Roberts (1975)

**Exercise 78 O<sub>2</sub> versus CO<sub>2</sub>**

Ralf Keeling (1988) measured atmospheric O<sub>2</sub> and CO<sub>2</sub> mole fractions and reports the following mole fraction differences (ppm = parts per million) against a reference gas:

CO<sub>2</sub> = c(12.5,17.0,13.5,18.3,60.0,55.8,62.2,56.2,49.3,42.1,27.6,21.4,26.2,28.8, 33.1,24.3,17.0,4.4,7.0,3.9, 9.3,14.1,2.8,4.9,4.9)

O<sub>2</sub> = c(-5.6,-8.4,-6.3,-9.6,-62.9,-51.8,-60.6,-53.2,-48.6,-37.0, -18.6,-12.2,-14.8,-19.6,-29.9,-16.9,-7.9,1.2,-1.3,6.8,-1.0,-7.2,6.5,5.0,7.3)

(a) Calculate Pearson's correlation coefficient  $r$ .

(b) Test whether  $r$  is different from zero by applying the R routines `cor.test()`, `correlationTest()` (library `fBasics`), and `pearsonTest()`. Do they yield the same results?

## H.6 Permutation test on correlation coefficients

Greenacre & Primicerio (2013) discuss a significance test for correlation coefficients. The null hypothesis  $H_0$  states that there is no correlation between two samples  $x$  and  $y$ . Under the assumption that  $H_0$  is true, one can use any permutation of the data in  $x$  and  $y$  and obtain small values (around zero) of the Pearson correlation coefficient  $r$ . One can perform a Monte Carlo simulation where a large number ( $M$ ) of permutations are generated in a pseudo-random way; the resulting  $r$  values provide an estimate of the  $r$  distribution. The  $p$ -value is calculated as all generated  $r$  values that are as or more extreme than the observed  $r$  value divided by  $M$  (relative frequency as an approximation for probability). For the common choice of  $\alpha = 0.05$  as level of significance, it is sufficient to use  $M = 10^4$  ( $M$  has to be larger for smaller  $\alpha$ ).

Remark: The permutation test on correlation coefficients requires no assumptions about the distribution from which data have been collected. Thus it is a **non-parametric** or **distribution-free** test.

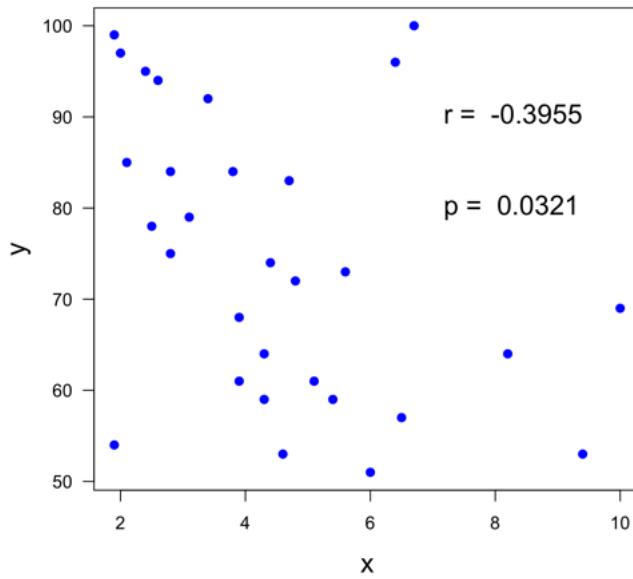


Figure H.10: Pearson correlation coefficient  $r = -0.3955$  between  $x$  and  $y$ . The null hypothesis  $H_0: r = 0$  can be rejected on the level of significance  $\alpha = 0.05$  based of the (estimated)  $p$ -value of 0.032.

[PermutationTestCor.R](#) [CorTestPerm.R](#)

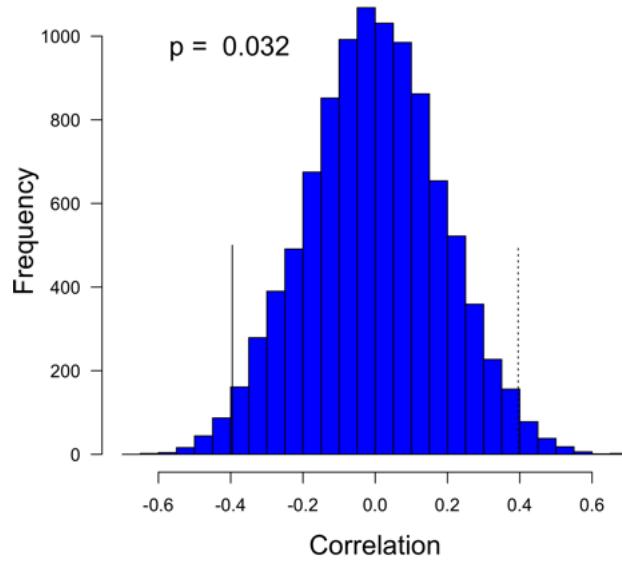


Figure H.11: Histogram of  $r$  values generated from  $M = 10^4$  random permutations of the data. The  $p$ -value is estimated by summing up all frequencies for  $r$  values that are more extreme than the observed  $r$  ( $= -0.3955$ ), i.e. all frequencies left of the vertical solid line and right of the vertical dashed line (at  $|r| = 0.3955$ ); the sum of these frequencies is divided by  $M$  yielding a relative frequency which is an estimate of  $p$ .

[PermutationTestCorB.R](#) [CorTestPerm.R](#)

## H.7 $\chi^2$ test

The  $\chi^2$  test is another goodness-of-fit-test that is applied when observed frequencies and especially count data are available. For the null hypothesis 'sample is from a probability distribution  $\mathcal{P}_1$ ' the test statistic is defined by

$$\chi^2 = \sum_{k=1}^n \frac{(f_{k,\text{obs}}/n - p_{k,\text{theor}})^2}{p_{k,\text{theor}}} \quad (\text{H.18})$$

where  $f_{k,\text{obs}}/n$  is the observed relative frequency for the outcome  $k$  and  $p_{k,\text{theor}}$  is the corresponding probability of the theoretical (hypothesized) distribution. The test statistic is the sum of the squared deviations between observed and theoretical values weighted by one divided by the theoretical probabilities, i.e. deviations at low theoretical probabilities ('rare events') contribute more. We will discuss an example with genetic background (Zar, 2010, p. 466, Example 22.1). The color of 100 flowers have been recorded: 84 are yellow and 16 are green. Based on Mendel's laws (assuming the color is controlled by a single gene that comes both as a dominant and a recessive allele) one hypothesizes that yellow and green should occur in the ratio of 3 to 1. I.e. the theoretical probabilities are  $p_1 = p_{\text{yellow}} = 0.75$  and  $p_2 = p_{\text{green}} = 0.25$  which for a sample size  $n = 100$  would suggest theoretical frequencies of 75 and 25. These numbers have to be compared to the observed frequencies. The test statistic  $\chi^2$  can be calculated by

$$\chi^2 = \frac{(84 - 75)^2}{75} + \frac{(16 - 25)^2}{25} = 4.32 \quad (\text{H.19})$$

$$= n \frac{(0.84 - 0.75)^2}{0.75} + \frac{(0.16 - 0.25)^2}{0.25} = 4.32 \quad (\text{H.20})$$

R code: fobs = c(84,16); fthe = c(75,25); pthe = fthe/sum(fthe); chisq.test(fobs,p=pthe)

### Interpretation of output of chisq.test()

1.  $\chi^2 = 4.32$  is the observed value of the test statistic
2.  $\text{df} = 1$  is the number of degrees of freedom ( $n = 2$  observations; 1 constraint: sum of relative frequencies = 1)
3.  $p\text{-value} = 0.03767$  = observed level of significance
4. The null hypothesis  $H_0$  'data follow partitioning of 3:1' is rejected on the (chosen) level of significance  $\alpha = 0.05$  because  $p = 0.038 < \alpha$  or, equivalently, the observed  $\chi^2_{\text{obs}} = 4.32$  is larger than the critical value  $\chi^2_{\alpha,\nu} = \chi^2_{0.05,1} = 3.84$  (Fig. H.12).
5. How to calculate  $\chi^2_{\alpha,\nu} = \chi^2_{0.05,1}$  in R? `qchisq(1-alpha,dv=nu)`

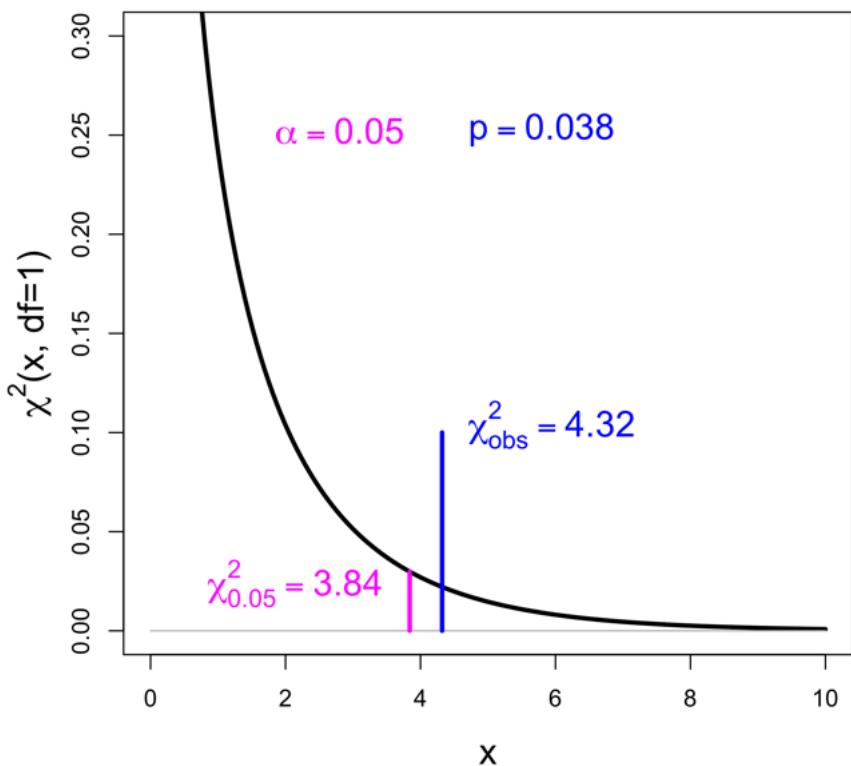


Figure H.12: Zar (2010, p. 466, Example 22.1): The null hypothesis  $H_0$  'data follow partitioning of 3:1' is rejected on the (chosen) level of significance  $\alpha = 0.05$  because  $p = 0.038 < \alpha$  or, equivalently, the observed  $\chi^2_{\text{obs}} = 4.32$  is larger than the critical value  $\chi^2_{\alpha,v} = \chi^2_{0.05,1} = 3.84$ . [ChiSquaredTestZar22d1.R](#)

## H.8 Pedestrian way of KS-test (Edelweiss) (\*)

The pedestrian way consists of three major steps: (I) estimate CDF from data, (II) construct hypothesized CDF of the continuous uniform distribution, (III) calculate the maximum absolute difference between data CDF and uniform CDF.

1. CDF estimated from data,  $CDF_{est}(x) = 0$  for  $x < x_1$ ; at each data point the  $CDF_{est}(x)$  increases by  $1/L$  and reaches 1 at  $x = x_L$ ;  $CDF_{est}(x) = 1$  for  $x > x_L$  (Fig. 12.29)
2. PDF of the continuous uniform distribution in the range from 1810 to 2270 m:  $PDF_{unif}(x) = (x - 1810)/(2270 - 1810) = (x - 1810)/460$  for  $1810 \leq x \leq 2270$  and 0 otherwise (Fig. H.13)
3. CDF corresponding to the PDF of the continuous uniform distribution in the range from 1810 to 2270 m: the CDF( $x$ ) is given by the integration from  $-\infty$  to  $x$  over the PDF  $\Rightarrow CDF_{unif}(x)$  is 0 for  $x < 1810$ , it increases linearly from 0 to 1 between 1810 and 2270, i.e.  $CDF_{unif}(x) = (x - 1810)/460$ , and it is 1 for  $x > 2270$  (Fig. 12.30).
4. Both CDFs in the same graph: Fig. H.14.
5. The test statistic  $D$  of the Kolmogorov-Smirnov test is defined as the maximum difference between the CDF estimated from the data ( $CDF_{est}(x_i)$ ) and the CDF (at the same heights) based on the null hypothesis ( $CDF_{unif}(x_i)$ ). More precisely, two sets of differences will be calculated, namely:

$$D_i = |CDF_{est}(x_i) - CDF_{unif}(x_i)| \quad \text{for } i = 1, 2, \dots, L \quad (\text{H.21})$$

$$D'_i = |CDF_{est}(x_{i-1}) - CDF_{unif}(x_i)| \quad \text{for } i = 2, 3, \dots, L. \quad (\text{H.22})$$

$$D'_1 = |0 - CDF_{unif}(x_1)| \quad (\text{based on } CDF_{est}(x_0) = 0) \quad (\text{H.23})$$

6. Here:

$$D_i = \{0.016, 0.019, 0.037, 0.006, 0.024, 0.090, 0.123, 0.144, 0.208, \color{red}{0.260535}, \\ 0.1021392, 0.149, 0.054\}$$

$$D'_i = \{0.093, 0.058, 0.040, 0.071, 0.053, 0.013, 0.046, 0.067, 0.131, 0.184, \\ 0.115, 0.072, 0.023\}$$

and thus  $D = 0.260535$  (Fig. H.14).

7. Calculate the observed level of evidence (p-value):  $p$  is usually calculated as an integral over the PDF of the test statistic (here: from the observed  $D = 0.260535$  to 1 which is the maximum possible value). For the KS-test a routine is available to calculate the CDF of the test statistic  $D$ , i.e. the integration is already done, and thus we have just to calculate how much of the CDF is above observed CDF value  $f(D, 13) = 0.7123 \Rightarrow p = 1 - 0.7123 = 0.2877$ .
8. The PDF corresponding to the KS-distribution (which is a CDF!) can be easily calculated by numerical differentiation (Fig. 12.31).

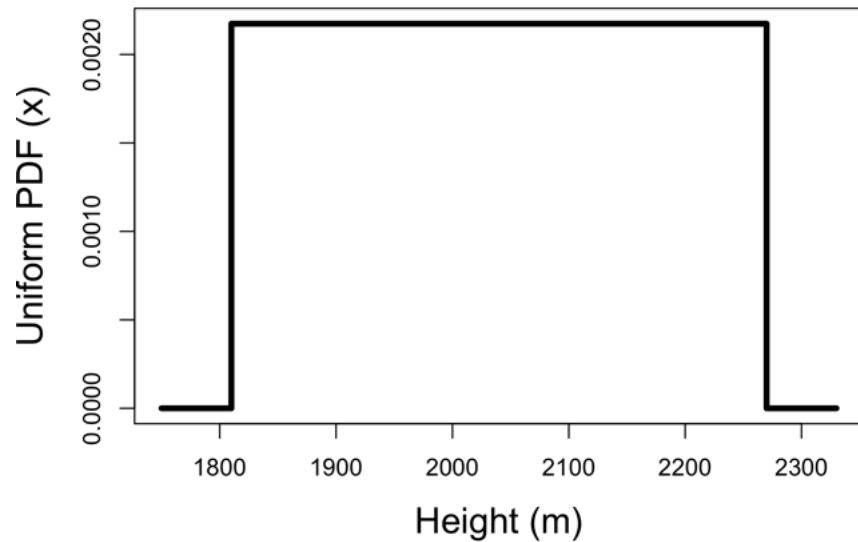


Figure H.13: PDF of the continuous uniform distribution in the range from 1810 to 2270 m.

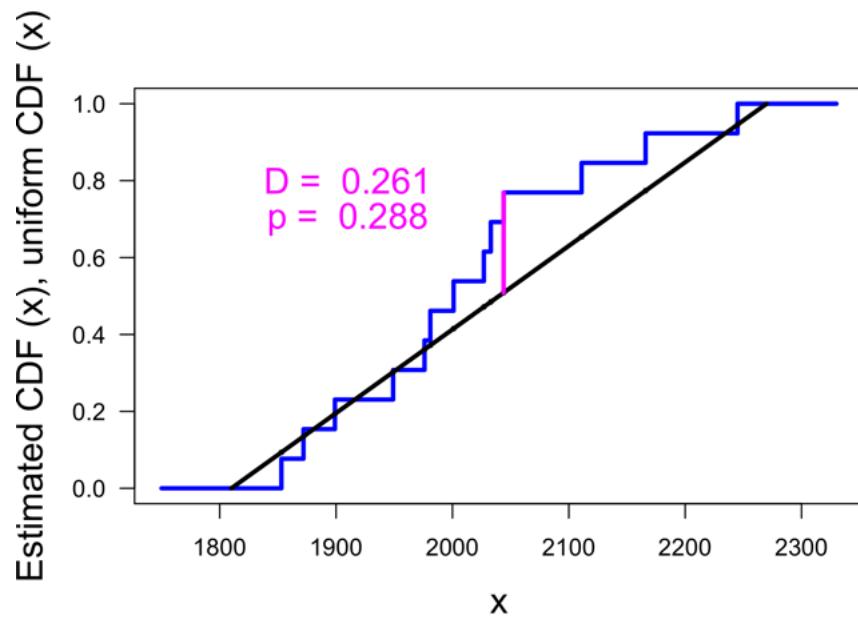


Figure H.14: CDFs & test statistic  $D$ : (1) estimated from data (plotted as staircase; black) & (2) for continuous uniform PDF between 1810 and 2270 m. The test statistic  $D$  is the largest difference between the CDFs at observations  $x_i$ . Here:  $D = 0.2605$  (indicated by the vertical red line).

$i$	$x_i$	$f_i$	$fc_i$	$CDF_{\text{est}}(x_i)$	$CDF_{\text{unif}}(x_i)$	$D_i$	$D'_i$
1	1853	1	1	0.077	0.093	0.0166	0.0935
2	1872	1	2	0.154	0.135	0.0191	0.0579
3	1899	1	3	0.231	0.193	0.0373	0.0396
4	1949	1	4	0.308	0.302	0.0055	0.0714
5	1976	1	5	0.385	0.361	0.0237	0.0532
6	1981	1	6	0.462	0.372	0.0898	0.0129
7	2001	1	7	0.538	0.415	0.1232	0.0463
8	2027	1	8	0.615	0.472	0.1436	0.0667
9	2033	1	9	0.692	0.485	0.2075	0.1306
10	2044	1	10	0.769	0.509	0.2605	0.1836
11	2111	1	11	0.846	0.654	0.1918	0.1149
12	2166	1	12	0.923	0.774	0.1492	0.0722
13	2245	1	13	1.000	0.946	0.0543	0.0226

Table H.1: Pedestrian way for KS-test of edelweiss data:  $i = 1, 2, \dots, 13$  observation numbers,  $x_i$  observed heights of edelweiss sightings,  $f_i$  frequency of observation,  $fc_i$  cumulative frequencies,  $CDF_{\text{est}}(x_i)$  CDF estimated from data, at  $x_i$ ,  $CDF_{\text{unif}}(x_i)$  CDF of continuous uniform distribution between 1810 and 2270 m, at  $x_i$ ,  $D_i$  first difference,  $D'_i$  second difference.

## H.9 Fair/unbiased coin? Bayesian approach

Tossing a coin = sampling from binomial PD with  $p$  = probability of success (here: head) in single trial and  $n$  = number of trials

Null hypothesis (unbiased coin)  $H_0: p = 1/2$  versus alternative hypothesis  $H_1: p \neq 1/2$ .

Prior probability for  $H_0: \rho_0 = 1/2 \Rightarrow$  prior probability for  $H_1: \rho_1 = (1 - \rho_0) = 1/2$ .

The posterior for  $H_0$  is given by (Robert, 2007, p. 231)

$$\text{Post}(p_0 = 1/2|x, n, \rho_0) = \frac{\overbrace{f_n(x|p_0)}^{\text{likelihood}} \overbrace{\rho_0}^{\text{prior}}}{\underbrace{\int f_n(x|p) \pi(p) dp}_{\text{normalization}}} = \frac{f_n(x|p_0) \rho_0}{f_n(x|p_0) \rho_0 + (1 - \rho_0) \underbrace{\int_{p \neq p_0} f_n(x|p) g_1(p) dp}_{\text{marginal likelihood}}} \quad (\text{H.24})$$

Likelihood  $f_n(x|p)$  = binomial distribution<sup>4</sup>:

$$f_n(x|p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (\text{H.25})$$

Marginal likelihood<sup>5</sup> for  $g_1 = 1$  (flat prior) reads

$$\int_0^1 \binom{n}{x} p^x (1 - p)^{n-x} dp = \frac{1}{n+1} \quad (\text{H.26})$$

The posterior of  $H_0$  is calculated as follows:

$$\begin{aligned} \text{Post}(p_0 = 1/2|x, n, \rho_0) &= \frac{\binom{n}{x} p_0^x (1 - p_0)^{n-x} \rho_0}{\binom{n}{x} p_0^x (1 - p_0)^{n-x} \rho_0 + \frac{1}{n+1} (1 - \rho_0)} \\ &= \frac{\binom{n}{x} 0.5^x (1 - 0.5)^{n-x} \rho_0}{\binom{n}{x} 0.5^x (1 - 0.5)^{n-x} \rho_0 + \frac{1}{n+1} (1 - \rho_0)} \\ &= \frac{\binom{n}{x} 0.5^n \rho_0}{\binom{n}{x} 0.5^n \rho_0 + \frac{1}{n+1} (1 - \rho_0)} = \frac{\binom{n}{x}}{\binom{n}{x} + \frac{1}{n+1} \frac{(1 - \rho_0)}{\rho_0} 2^n} \\ &= \frac{1}{1 + \frac{1}{n+1} \frac{(1 - \rho_0)}{\rho_0} 2^n \frac{x!(n-x)!}{n!}} \\ &= \frac{1}{1 + \frac{(1 - \rho_0)}{\rho_0} \frac{x!(n-x)!}{(n+1)!} 2^n} \end{aligned} \quad (\text{H.27})$$

(this is identical to Robert, 2007, p. 231) where  $x$  is the number of successes (here: head). The Bayes factor  $B_{01}$  is related to the posterior probability for  $H_0$  by (Robert, 2007, p. 231)

$$\text{Post}(p = 1/2|x, n, \rho_0) = \frac{1}{1 + \frac{1 - \rho_0}{\rho_0} \frac{1}{B_{01}}} \quad (\text{H.29})$$

---

<sup>4</sup>  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

<sup>5</sup> Strictly speaking one should exclude  $p = p_0$  from the integral, however, a single non-singular point does not contribute enough to change the value of the integral (measure theory).

and thus (by comparing the right-hand sides of the two equations above)

$$B_{01}(p = 1/2; n, x) = \frac{(n+1)!}{x! (n-x)! 2^n} \quad (\text{H.30})$$

(please note that  $B_{01}$  is independent of  $\rho_0$ ).

**Example:** What would one expect when  $n = 5$  and  $x = 3$ ? If  $p = 1/2$  one would expect a high probability for  $n/2$  successes (heads). Unfortunately,  $n$  is an odd number and thus one cannot expect a single maximum of probability at  $x = n/2 = 2.5$ . However, posterior values for  $x = 2$  or  $x = 3$  should be high. Thus observing  $x = 3$  should provide evidence for  $H_0$ . The number of trials is relatively small and thus we should not expect a high value for the Bayes factor.

Calculation:

$$\text{Post}(p = 1/2|x = 3, n = 5, \rho_0 = 1/2) = \frac{1}{1 + \frac{1 - \rho_0}{\rho_0} \frac{x! (n-x)!}{(n+1)!} 2^n} \quad (\text{H.31})$$

$$\begin{aligned} &= \frac{1}{1 + \frac{1 - 1/2}{1/2} \frac{\overbrace{3!}^{=6} \overbrace{(5-3)!}^{=2} \overbrace{2^5}^{=32}}{\underbrace{(5+1)!}_{=720}}} \\ &= \frac{1}{1 + \frac{2 \cdot 6 \cdot 32}{720}} = \frac{1}{1 + \frac{8}{15}} = \frac{1}{\frac{15}{15} + \frac{8}{15}} \\ &= \frac{15}{23} \approx 0.65 \end{aligned} \quad (\text{H.32})$$

$$B_{01}(p = 1/2; n = 5, x = 3) = \frac{(n+1)!}{x! (n-x)! 2^n} = \frac{(5+1)!}{3! (5-3)! 2^5} = \frac{720}{6 \cdot 2 \cdot 32} = \frac{15}{8} = 1.875 \quad (\text{H.33})$$

i.e. the Bayes factor is indicating some (not substantial) evidence for  $H_0$ ; stronger support for  $H_0$  is only possible with more trials (larger sample size  $n$ ).

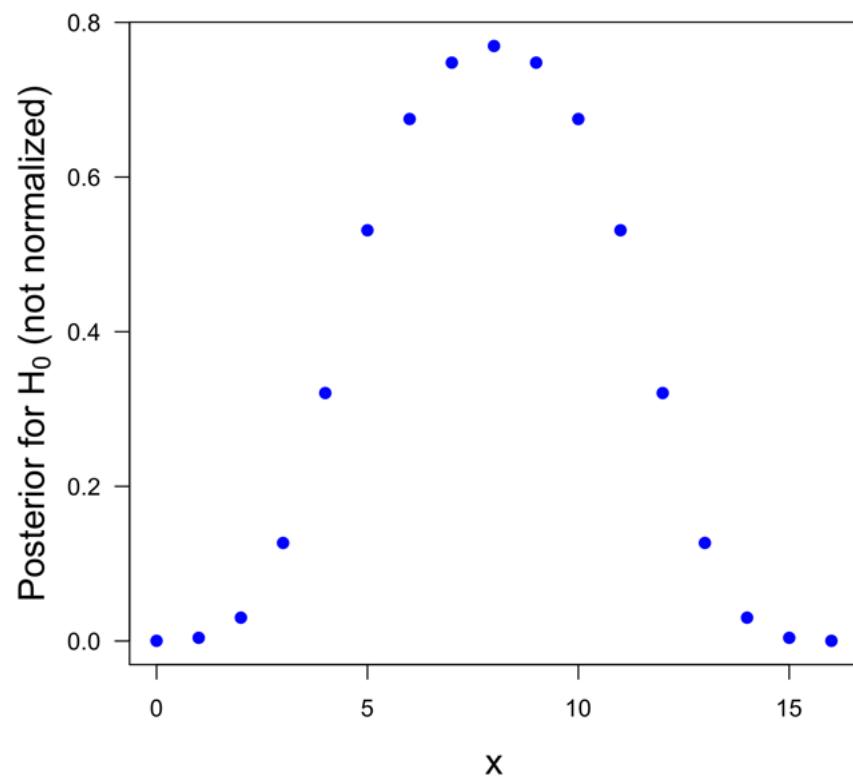


Figure H.15: Unbiased coin? The posterior for  $H_0$  (not normalized!) for  $n = 16$  trials.  
[SolutionRobert07FairCoin.R](#)

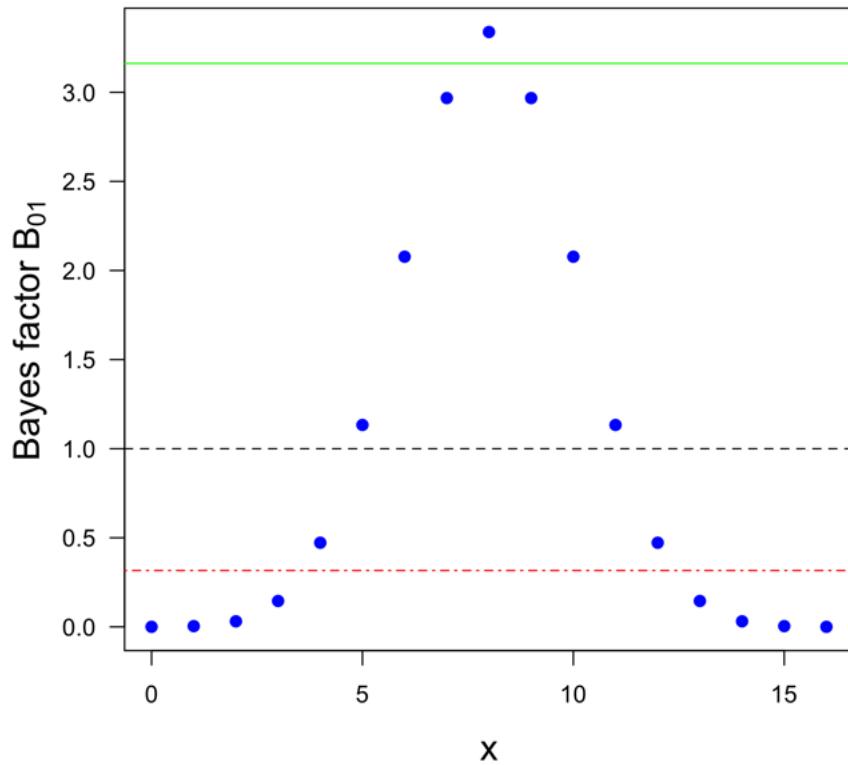


Figure H.16: Unbiased coin? The Bayes factor  $B_{01}$  for  $n = 16$  trials as a function of observed number of heads,  $x$ , (blue dots).  $B_{01} < 1/\sqrt{10} \approx 0.316$  provides substantial evidence against  $H_0$  (red line),  $B_{01} = 1$  is the border between evidence for or against  $H_0$  (black line), and  $B_{01} > \sqrt{10} \approx 3.16$  [SolutionRobert07FairCoin.R](#) (line 17: change sflag to 2)

### H.9.1 Slightly different derivation (\*)

The Bayesian approach starts from Bayes Theorem which reads in the context of hypothesis testing:

$$\underbrace{P(\text{hypothesis} | \text{data})}_{\text{posterior}} = \frac{\underbrace{P(\text{data} | \text{hypothesis})}_{\text{likelihood}} \cdot \underbrace{P(\text{hypothesis})}_{\text{prior}}}{\underbrace{P(\text{data})}_{\text{normalization constant}}} \quad (\text{H.34})$$

The goal is to calculate posteriors for  $H_0$  and  $H_1$  and to compare them (which is the larger one?) **Posterior ratio:**

$$\frac{P(H_0 | \text{data})}{P(H_1 | \text{data})} = \underbrace{\frac{P(\text{data} | H_0)}{P(\text{data} | H_1)}}_{\text{Bayes factor}} \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior ratio}}$$

We often have no bias towards  $H_0$  or  $H_1$  (no background knowledge). Or we want to be demonstratively unbiased. In both cases we would set the prior ratio equal to 1 (fifty-fifty chance for both hypotheses).

**Recommended song:** Frank Zappa: Fifty-fifty ('· · the odds are fifty-fifty that I have something to say · ·')

If prior ratio = 1  $\Rightarrow$  posterior ratio = Bayes factor

Likelihood: How likely is it to observe  $x$  (for example,  $x = 13$ ) when  $H_0$  is true?

$$P(x | H_0) = \underbrace{\binom{n}{x} 0.5^x (1 - 0.5)^{n-x}}_{\text{binomial}} = \binom{n}{x} 0.5^n$$

Likelihood: How likely is it to observe  $x$  (for example,  $x = 13$ ) when  $H_1$  is true?

$$P(x | H_1) = \underbrace{\binom{n}{x} p^x (1 - p)^{n-x}}_{\text{binomial}} \quad p \neq 0$$

How to get rid of the unknown  $p$  in the likelihood for  $H_1$ ?

$$P(x | H_1) = \binom{n}{x} p^x (1 - p)^{n-x} \quad p \neq 0$$

Integration! Integrating out the nuisance parameter  $p$ :

$$\int_0^1 \binom{n}{x} p^x (1 - p)^{n-x} g(p) dp$$

where  $g(p)$  is a prior distribution for the (hyper-)parameter  $p$ .

Remark: we should exclude the point  $p = 1/2$  from the integral because it belongs to  $H_0$  and not to  $H_1$ ; however, this single point does not change the value of the integral.

Choice of prior  $g(p)$ :  $g_1(p) = 1$  ('flat prior').

A bit of calculus & algebra

For the flat prior we have to calculate:

$$\int_0^1 \binom{n}{x} p^x (1 - p)^{n-x} dp = \frac{1}{n+1}$$

Thus the posterior ratio = the Bayes factor reads:

$$B_{01}(p = 1/2; n, x) = \binom{n}{x} 0.5^n (n+1) = \frac{(n+1)!}{x!(n-x)!2^n}$$

(compare, for example, Robert, 2007)

Small Bayes factors  $B_{01}$  would speak against  $H_0$ .

Large Bayes factors  $B_{01}$  would provide evidence for  $H_0$ .

## H.10 Sample from Poisson or from zero inflated Poisson distribution?

Here we will investigate two count data sets which might be samples from Poisson distributions. Both data sets contain a large number of zeros which would indicate a low mean rate  $\lambda$  of the Poisson distribution. An alternative hypothesis states that the sample stems from a zero inflated Poisson (ZIP) distribution. We will apply the Bayesian approach to hypothesis testing and use different priors.

Bayarri et al. (2008) list the following count data:

$x$	0	1	2	3	number of events
$y$	81	9	7	1	frequency of events

Table H.2: Example 1

$x$	0	1	2	3	4	number of events
$y$	38	26	8	2	1	frequency of events

Table H.3: Example 2

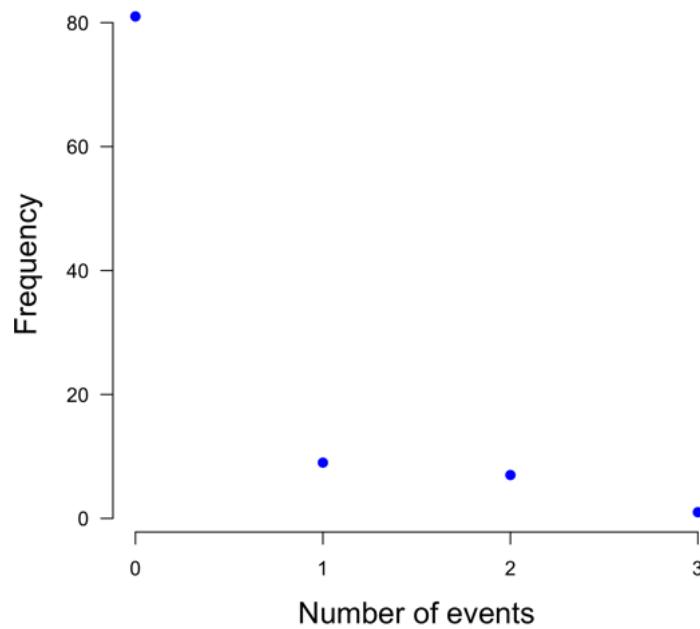


Figure H.17: Plot of count data for Example 1 (Table H.2): the data might stem from a Poisson distribution (with small mean rate  $\lambda < 1$ ), however, the number of zeros seems to be quite high given the rather slow decline of frequencies for number of events larger than zero. [BayesianHypBayarri1Ex.R](#)

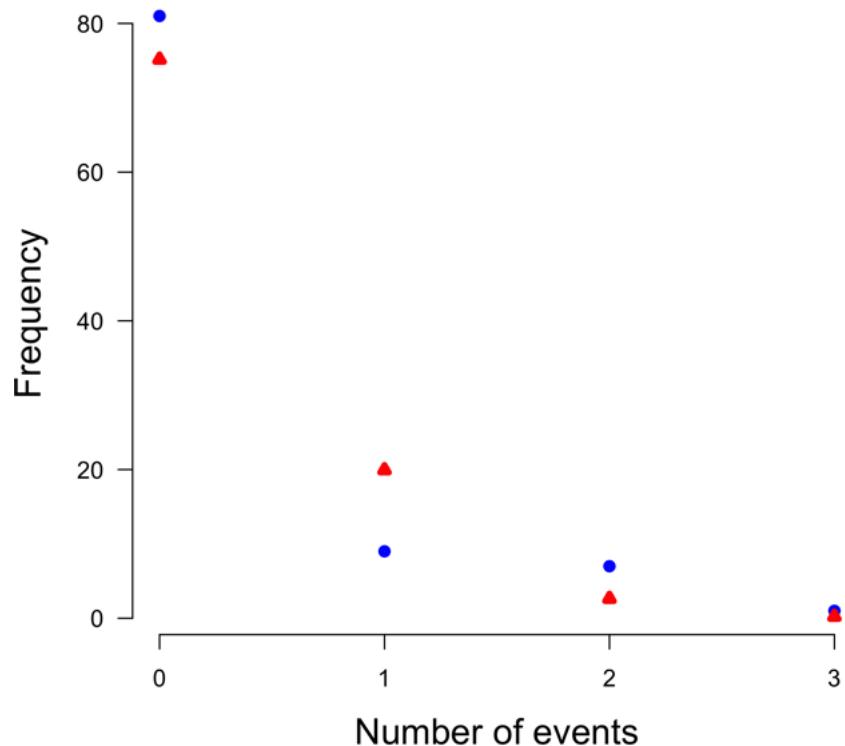


Figure H.18: Plot of count data (blue dots) for Example 1 (Table H.2) and predicted frequencies based on an estimated mean rate  $\hat{\lambda} = 0.27$  for a Poisson distribution. The fitted Poisson distribution underestimates the number of zeros and overestimates the number of ones, i.e. the estimated decrease of the frequencies for 0 to 1 is underestimated. This could be taken as an indication for a zero inflated Poisson distribution.

[BayesianHypBayarri1ExFit.R](#)

### H.10.1 Zero inflated Poisson (ZIP) distributions

Zero inflated Poisson (ZIP) distributions consist of a mixture of a Poisson distribution with mean rate  $\lambda$  and a ‘point’ distribution with probability  $p$  for a single point, namely  $x = 0$  (Figs. H.19 – H.20):

$$f_{\text{ZIP}}(x; \lambda, p) = p I(x = 0) + (1 - p) f_{\text{Poisson}}(x; \lambda), \quad x = 0, 1, 2, \dots \quad (\text{H.35})$$

where

$$f_{\text{Poisson}}(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots \quad (\text{H.36})$$

is the Poisson distribution,  $I()$  is the indicator or characteristic function<sup>6</sup>

$$I(x = 0) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \quad (\text{H.37})$$

and  $0 \leq p \leq 1$  is the zero inflation parameter.

The probability at  $x = 0$  is given by

$$f_{\text{ZIP}}(x = 0; \lambda, p) = p + (1 - p) e^{-\lambda}. \quad (\text{H.38})$$

For the example plotted in Fig. H.19 one obtains  $f_{\text{ZIP}}(x = 0; \lambda = 2.5, p = 0.4) \approx 0.45$  and  $f_{\text{Poisson}}(x = 0; \lambda = 2.5) \approx 0.08$ . The mean of ZIP is given by  $\mu_{\text{ZIP}} = (1 - p) \lambda$  and the variance of ZIP is not equal to its mean  $\mu$  or to the rate parameter  $\lambda$ :  $\sigma_{\text{ZIP}}^2 \neq \mu_{\text{ZIP}}$  and  $\sigma_{\text{ZIP}}^2 \neq \lambda$ .

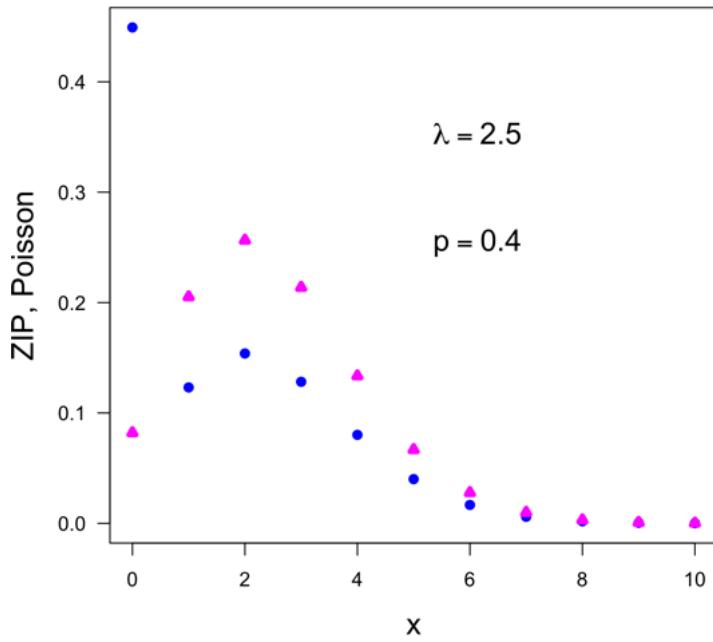


Figure H.19: Probabilities following a zero inflated Poisson (ZIP) distribution (blue dots,  $p = 0.4$ ) or a Poisson distribution (red triangles) both for  $\lambda = 2.5$ . The deviation between these distributions at  $x = 0$  is clearly visible. [BayesianHypPoissonZIP2d5.R](#)

<sup>6</sup>Instead of  $I(x = 0)$  one could also use here the Kronecker symbol  $\delta_{x0}$

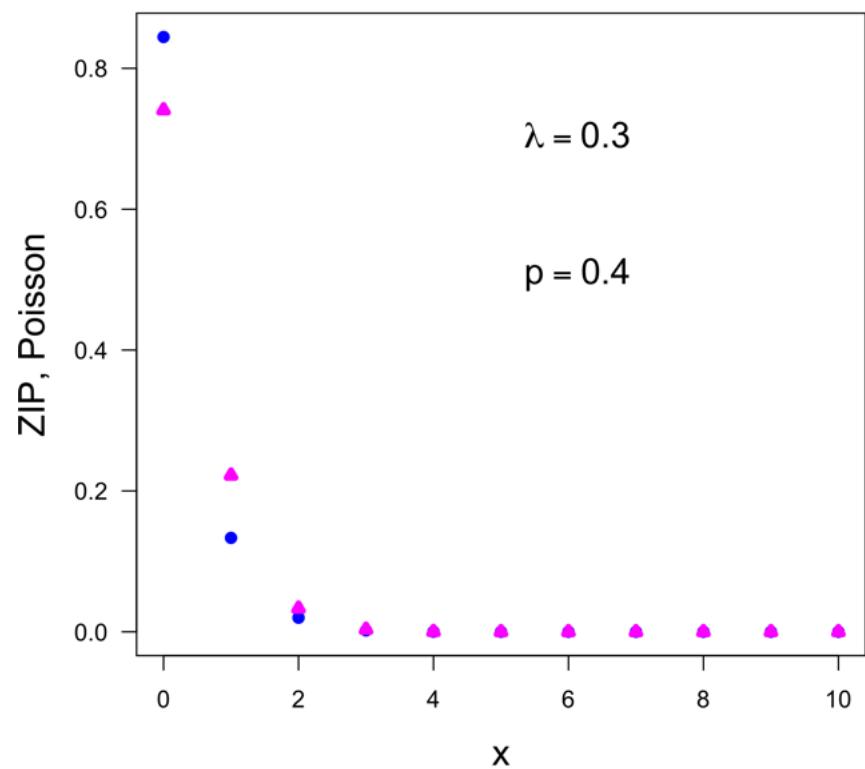


Figure H.20: Probabilities following a zero inflated Poisson (ZIP) distribution (blue dots,  $p = 0.4$ ) or a Poisson distribution (red triangles) both for  $\lambda = 0.3$ . The deviation between these distributions at  $x = 0$  are small.  
[BayesianHypPoissonZIP0d3.R](#)

## H.10.2 $H_0, H_1$ , priors, marginal likelihoods, Bayes factor (\*)

We will now apply the Bayesian approach to hypothesis testing which requires the choice of priors. In the Bayesian approach one formulates a null,  $H_0$ , and an alternative hypothesis,  $H_1$ . These two hypotheses often depend on unknown parameters for which we can and actually have to formulate priors. Based on the hypothesis, the priors for its parameters, and the data (observations) one calculates the marginal likelihood that does not depend on the parameters anymore ('the nuisance parameters have been integrated out'). The ratio of the marginal likelihoods for the two hypotheses is called the Bayes factor. Large (small) values of the Bayes factor speak for or against one of the hypotheses; values close to one do not allow a clear decision.

1. Formulation of  $H_0$  &  $H_1$ :

The null hypothesis,  $H_0$ , states that the sample stems from a Poisson distribution with unknown mean rate  $\lambda$ .

The alternative (or working) hypothesis,  $H_1$ , states that the sample stems from a zero inflated Poisson distribution with unknown mean rate  $\lambda$  and zero-inflation parameter  $p$ .

2. Calculate likelihoods for  $H_0$  &  $H_1$  (compare Appendix for details):

$$f_0(\mathbf{x}; \lambda) = \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \quad (\text{H.39})$$

Likelihood for the zero inflated Poisson (ZIP) distribution:

$$f_1(\mathbf{x}; \lambda, p) = \frac{[p + (1-p)e^{-\lambda}]^k (1-p)^{n-k} e^{-(n-k)\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \quad (\text{H.40})$$

where  $k = \sum_{i=1}^n I(x_i = 0)$  and  $0 \leq p \leq 1$  is the zero-inflation parameter.

3. Choose priors:

$H_0$  contains a single unknown ('nuisance') parameter, namely the mean rate  $\lambda > 0$  of the Poisson distribution. Here we choose the exponential distribution

$$\pi_0(\lambda) = e^{-\lambda} \quad (\text{H.41})$$

as prior. The exponential distribution is a proper prior, i.e. it is normalizable; actually it is already normalized to 1:  $\int_0^\infty e^{-\lambda} d\lambda = 1$ . Alternatives to this choice of prior  $\pi_0(\lambda)$  will be discussed below.

$H_1$  contains two unknown ('nuisance') parameter, namely the rate  $\lambda > 0$  of the zero depleted Poisson distribution and the zero-inflation parameter  $p$ . For  $\lambda$  we choose the same prior as for  $H_0$ , namely the exponential distribution. For  $p$  we can express our ignorance by a flat prior, which is actually normalizable because  $p$  can vary only over a finite range (between 0 and 1). I.e.

$$\pi_1(\lambda, p) = \pi_{1a}(\lambda) \cdot \pi_{1b}(p) = e^{-\lambda} \cdot 1 = e^{-\lambda}. \quad (\text{H.42})$$

4. Calculate marginal likelihoods ('integrate out nuisance parameters'):

Marginal likelihoods are defined as integrals over the product of likelihood and prior(s):

$$m_0(\mathbf{x}) = \int_0^\infty f_0(\mathbf{x}; \lambda) \pi_0(\lambda) d\lambda = \int_0^\infty \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} e^{-\lambda} d\lambda = \frac{s!}{(n+1)^{s+1} \prod_{i=1}^n x_i!} \quad (\text{H.43})$$

$$m_1(\mathbf{x}) = \int_0^\infty \int_0^1 f_1(\mathbf{x}; \lambda, p) \pi_1(\lambda, p) d\lambda dp = \frac{k! s!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} (n+1-j)^{-(s+1)} \quad (\text{H.44})$$

5. Calculate the Bayes factor ( $m_0/m_1$  or  $m_1/m_0$  whatever is more convenient):

$$B_{10}(x) = \frac{m_1}{m_0} = \frac{k!}{(n+1)!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \left(1 - \frac{j}{n+1}\right)^{-(s+1)} \quad (\text{H.45})$$

6. Scales of evidence (Jeffreys, 1961):

$B_{10} > 10$	strong evidence for $H_1$ (against $H_0$ )
$\sqrt{10} \approx 3.16 < B_{10} < 10$	substantial evidence for $H_1$ (against $H_0$ )
$1 < B_{10} < \sqrt{10} \approx 3.16$	slight evidence for $H_1$ (against $H_0$ )
$1/\sqrt{10} \approx 0.316 < B_{10} < 1$	slight evidence for $H_0$ (against $H_1$ )
$0.1 < B_{10} < 1/\sqrt{10} \approx 0.316$	substantial evidence for $H_0$ (against $H_1$ )
$B_{10} < 0.1$	strong evidence for $H_0$ (against $H_1$ ).

Applying the Bayesian test to the data of Example 1 (Table H.2) yields the Bayes factor  $B_{10}(x) = 214.8$ . According to the scales of evidence given by Jaynes (1961) such a high value would definitely speak for  $H_1$  and against  $H_0$ . This confirms our guess based on looking at the data.

### H.10.3 Choice of priors (discussion)

Bayarri et al. (2008) have chosen the Jeffreys' prior  $1/\sqrt{\lambda}$  instead of the exponential prior. They did this on purpose to show that even an inappropriate (not normalizable) prior<sup>7</sup> can work (give reasonable results) when applied to both hypotheses. The Bayes factor for the Jeffreys' prior reads

$$B_{10}^{\text{Jeffreys}}(x) = \frac{m_1}{m_0} = \frac{k!}{(n+1)!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \left(1 - \frac{j}{n}\right)^{-(s+1/2)} \quad (\text{H.46})$$

It yields  $B_{10}^{\text{Jeffreys}} = 223.1$  for the data of Example 1 (Table H.2) and thus leads to the same conclusion as based on the exponential prior.

Flat priors usually do not work when the (nuisance) parameter range is infinite.

Alternatives:

$$\pi_0(\lambda) = a e^{-a\lambda}, \quad a > 0 \quad (\text{H.47})$$

---

<sup>7</sup>The integral  $\int_0^\infty d\lambda / \lambda$  does not exist because of the singularity at  $\lambda = 0$ .

### H.10.4 Formulation as an estimation problem

Can the data be used to estimate  $\lambda$  and  $p$ ? A small (negligible) estimate for  $p$  would then speak against  $H_1$ . An ad-hoc method would be least squares where one (1) calculates the differences between observed relative frequencies  $fr_i = y_i / \sum y$  and probabilities  $p_i(x_i; \lambda, p)$  predicted from the zero inflated Poisson distribution, (2) sums up the squared differences, and (3) finds the minimum of the sums of squares:

$$SS(p, \lambda; \mathbf{y}(\mathbf{x})) = \sum_i (fr_i - p_i^{\text{ZIP}}(p, \lambda))^2 \rightarrow \text{minimum (by varying } p \text{ and } \lambda\text{)} \quad (\text{H.48})$$

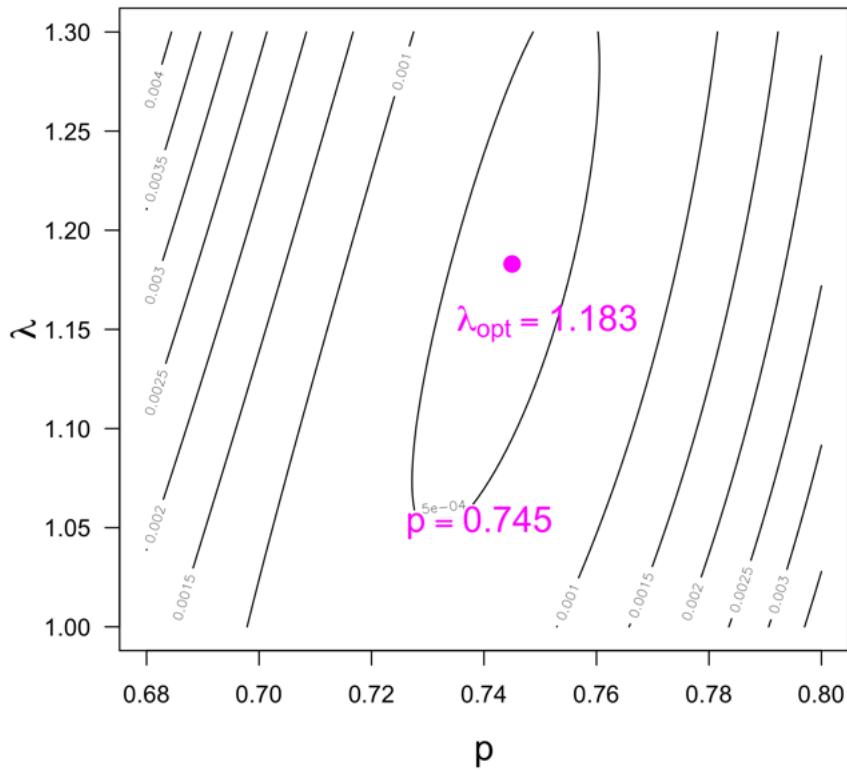


Figure H.21: Least-squares for observed frequencies  $y = \{81, 9, 7, 1\}$  for  $x = \{0, 1, 2, 3\}$ . The graph shows iso-contours of the sum of squares (squared differences between observed relative frequencies  $fr_i = y_i / \sum y$  and predicted probabilities of the zero inflated Poisson distribution) as a function of zero inflation parameter  $p$  and rate parameter  $\lambda$ . The optimal (in the least-squares sense) parameter values are  $\hat{p} = p_{\text{opt}} = 0.745$  and  $\hat{\lambda} = \lambda_{\text{opt}} = 1.183$ . [BayesianHypBayarriEst.R](#)

**Summary:** The large value of the estimated inflation parameter,  $\hat{p} = p_{\text{opt}} = 0.745$ , speaks against  $H_0$  and for  $H_1$ . The ZIP distribution with the estimated parameters fits the data very well (Fig. H.22). Thus for Example 1 estimation (using least squares<sup>8</sup>) is a valid alternative to testing, yielding at the same time optimal parameters that can be used for a fit to the data.

<sup>8</sup>The least squares method for simple linear regression can be derived from the basic rules of probability together under the conditions of independence of data, additive normal noise, and flat priors (Section J.3). These conditions are not given here and thus we apply least square in an **ad-hoc** manner. The good estimation results indicates that least square is a very robust method that often works although it can not be – or at has not been – justified by theory.

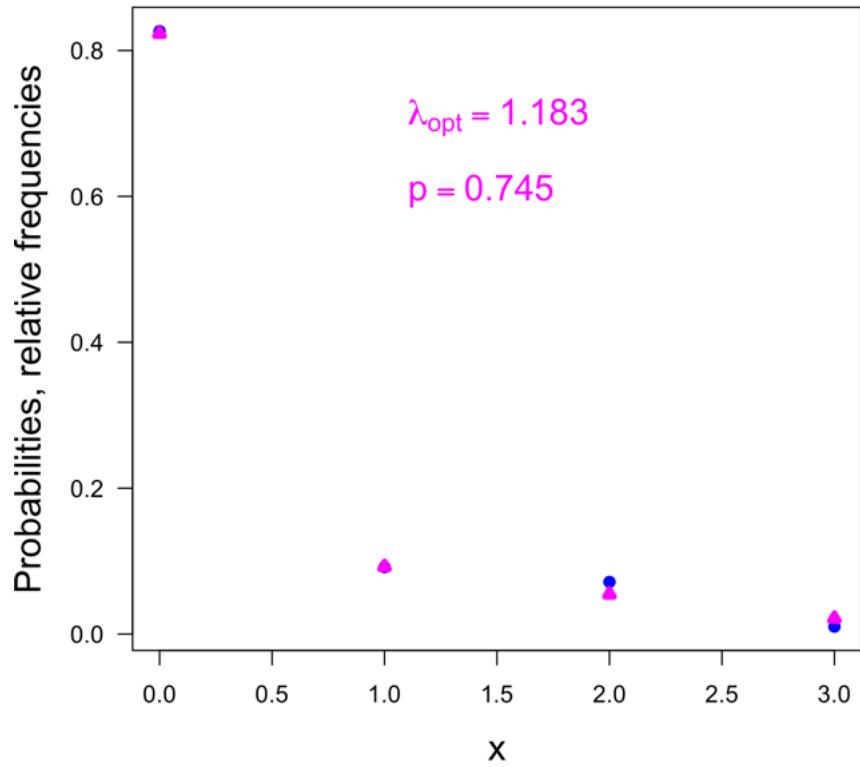


Figure H.22: Least-squares for observed frequencies  $y = \{81, 9, 7, 1\}$  for  $x = \{0, 1, 2, 3\}$ . The graph shows observed relative frequencies (blue dots) and predicted probabilities from the zero inflated Poisson distribution using the estimates  $\hat{p} = p_{\text{opt}} = 0.745$  and  $\hat{\lambda} = \lambda_{\text{opt}} = 1.183$ . [BayesianHypBayarriEstFit.R](#)

### H.10.5 Likelihoods of Poisson and zero inflated Poisson distributions (\*)

For calculation of the likelihoods one assumes independence of the data and thus one can apply the simplified product rule of probabilities.

Likelihoods of Poisson distribution:

$$f_0(\mathbf{x}; \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \quad (\text{H.49})$$

where  $s = \sum_{i=1}^n x_i$  (sum of data).

Likelihoods of zero inflated Poisson distribution:

$$f_0(\mathbf{x}; \lambda) = \prod_{i=1}^n \left[ p I(x_i = 0) + (1-p) \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right] \quad (\text{H.50})$$

$$\underset{\text{sort } x_i}{=} \underbrace{\left( \prod_{i=1}^k \left[ p + (1-p) e^{-\lambda} \right] \right)}_{x_i=0} \underbrace{\left( \prod_{i=k+1}^n (1-p) \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)}_{x_i>0} \quad (\text{H.51})$$

$$= \frac{1}{\prod_{i=1}^n x_i!} \left[ p + (1-p) e^{-\lambda} \right]^k (1-p)^{n-k} \lambda^s e^{-(n-j)\lambda} \quad (\text{H.52})$$

where 'sort  $x_i$ ' stands for sorting the data into the two groups (a) all  $x_i = 0$  and (b) all  $x_i > 0$ ; please note that  $\prod_{i=1}^n x_i! = \prod_{i=n-k}^n x_i!$  because (after sorting)  $x_i = 0$  and  $x_i! = 1$  for  $i = 1, 2, \dots, k$ .

### H.10.6 Marginal likelihoods for exponential prior (\*)

The marginal likelihood for the Poisson distribution under the exponential prior  $\exp(-\lambda)$  reads

$$m_0(\mathbf{x}) = \int_0^\infty \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} e^{-\lambda} d\lambda = \frac{\Gamma(s+1)}{(n+1)^{s+1} \prod_{i=1}^n x_i!} = \frac{s!}{(n+1)^{s+1} \prod_{i=1}^n x_i!} \quad (\text{H.53})$$

where

$$\int_0^\infty x^m e^{-ax} dx = \frac{\Gamma(m+1)}{a^{m+1}} \quad a > 0, \quad m > -1 \quad (\text{H.54})$$

has been used for  $m = s$  and  $a = n+1$ .

The marginal likelihood for the zero inflated Poisson distribution under the exponential prior for  $\lambda$  and the uniform prior for  $p$  can be calculated as follows

$$m_1(\mathbf{x}) = \frac{1}{\prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{k!}{j!(k-j)!} \int_0^\infty \int_0^1 p^j (1-p)^{n-j} e^{-(n-j)\lambda} \lambda^s e^{-\lambda} dp d\lambda \quad (\text{H.55})$$

$$= \frac{1}{\prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{k!}{j!(k-j)!} \frac{j!(n-j)!}{(n+1)!} \int_0^\infty e^{-(n+1-j)\lambda} \lambda^s d\lambda \quad (\text{H.56})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \int_0^\infty e^{-(n+1-j)\lambda} \lambda^s d\lambda \quad (\text{H.57})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \frac{\Gamma(s+1)}{(n+1-j)^{s+1}} \quad (\text{H.58})$$

$$= \frac{k! s!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} (n+1-j)^{-(s+1)} \quad (\text{H.59})$$

where the Euler integral of the 1. kind (also called the beta function)

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1} (1-p)^{\beta-1} dp = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (\text{H.60})$$

has been used for integer values  $\alpha = j+1, \beta = n-j+1$  (using  $\Gamma(m+1) = m!$  for  $m$  integer  $\geq -1$ )

$$\int_0^1 p^j (1-p)^{n-j} dp = \frac{\Gamma(j+1) \Gamma(n-j+1)}{\Gamma(n+2)} = \frac{j!(n-j)!}{(n+1)!} = \frac{1}{(n+1) \binom{n}{j}}. \quad (\text{H.61})$$

### H.10.7 Bayes factor for exponential prior (\*)

The Bayes factor is given by  $B_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$  or by  $B_{01}(\mathbf{x}) = m_0(\mathbf{x})/m_1(\mathbf{x})$  (whatever is more convenient). Here the Bayes factor for the exponential prior is defined by

$$B_{10}^{\text{exp}}(\mathbf{x}) = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \quad (\text{H.62})$$

$$= \frac{k! s!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} (n+1-j)^{-(s+1)} \frac{(n+1)^{s+1} \prod_{i=1}^n x_i!}{s!} \quad (\text{H.63})$$

$$= \frac{k!}{(n+1)!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \left(1 - \frac{j}{n+1}\right)^{-(s+1)} \quad (\text{H.64})$$

### H.10.8 Marginal likelihoods for Jeffreys' prior (\*)

The marginal likelihood for the Poisson distribution under Jeffreys' prior  $1/\sqrt{\lambda} = \lambda^{-1/2}$  reads

$$m_0(\mathbf{x}) = \int_0^\infty \frac{e^{-n\lambda} \lambda^s}{\prod_{i=1}^n x_i!} \lambda^{-1/2} d\lambda = \frac{\Gamma(s+1/2)}{n^{s+1/2} \prod_{i=1}^n x_i!} \quad (\text{H.65})$$

(compare Bayarri et al., 2008, p. 110) where

$$\int_0^\infty x^m e^{-ax} dx = \frac{\Gamma(m+1)}{a^{m+1}} \quad a > 0, \quad m > -1 \quad (\text{H.66})$$

has been used for  $m = s - 1/2$  and  $a = n$ .

The marginal likelihood for the zero inflated Poisson distribution under Jeffreys' prior for  $\lambda$  and the uniform prior for  $p$  can be calculated as follows

$$m_1(\mathbf{x}) = \frac{1}{\prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{k!}{j!(k-j)!} \int_0^\infty \int_0^1 p^j (1-p)^{n-j} e^{-(n-j)\lambda} \lambda^s \lambda^{-1/2} dp d\lambda \quad (\text{H.67})$$

$$= \frac{1}{\prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{k!}{j!(k-j)!} \frac{j!(n-j)!}{(n+1)!} \int_0^\infty e^{-(n-j)\lambda} \lambda^{s-1/2} d\lambda \quad (\text{H.68})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \int_0^\infty e^{-(n-j)\lambda} \lambda^{s-1/2} d\lambda \quad (\text{H.69})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \frac{\Gamma(s+1/2)}{(n-j)^{s+1/2}} \quad (\text{H.70})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \Gamma(s+1/2) (n-j)^{-(s+1/2)} \quad (\text{H.71})$$

(compare Bayarri et al., 2008, p. 110) using the Euler integral of the 1. kind

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1} (1-p)^{\beta-1} dp = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (\text{H.72})$$

for integer values  $\alpha = j+1, \beta = n-j+1$  (using  $\Gamma(m+1) = m!$  for  $m$  integer  $\geq -1$ )

$$\int_0^1 p^j (1-p)^{n-j} dp = \frac{\Gamma(j+1) \Gamma(n-j+1)}{\Gamma(n+2)} = \frac{j!(n-j)!}{(n+1)!} = \frac{1}{(n+1) \binom{n}{j}}. \quad (\text{H.73})$$

and the integral over  $d\lambda$  (Eq. H.66) for  $m = s - 1/2$  and  $a = n$ .

### H.10.9 Bayes factor for Jeffreys' prior (\*)

The Bayes factor<sup>9</sup> for Jeffreys' prior is defined by

$$B_{10}^{\text{Jeffreys}}(\mathbf{x}) = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \quad (\text{H.74})$$

$$= \frac{k!}{(n+1)! \prod_{i=1}^n x_i!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \Gamma(s+1/2) (n-j)^{-(s+1/2)} \frac{n^{s+1/2} \prod_{i=1}^n x_i!}{\Gamma(s+1/2)} \quad (\text{H.75})$$

$$= \frac{k!}{(n+1)!} \sum_{j=0}^k \frac{(n-j)!}{(k-j)!} \left(1 - \frac{j}{n}\right)^{-(s+1/2)} \quad (\text{H.76})$$

(compare Bayarri et al., 2008, p. 110, Eq. 2.9).

## H.11 Bayesian $t$ -test: details (\*)

### H.11.1 Likelihoods for $H_0$ and $H_1$

The likelihood for a random sample  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  from a normal distribution has been derived by Jeffreys (1961, p. 108–109). The deviation is repeated here, however, using a different notation.

The likelihood for a single data point  $x_i$  from a normal distribution with mean  $\mu$  and variance  $\sigma^2$  is given by

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (\text{H.77})$$

Assume that the data  $x_i$  are independent of each other, the likelihood for  $\mathbf{x}$  can be calculated simply as the product of the likelihoods for all  $x_i$  (according to the simplified product rule of probabilities)

$$f(\mathbf{x}; \mu, \sigma^2) = \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}} \quad (\text{H.78})$$

---

<sup>9</sup>The Bayes factor is given by  $B_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$  or by  $B_{01}(\mathbf{x}) = m_0(\mathbf{x})/m_1(\mathbf{x})$  (whatever is more convenient).

The sum in the exponent can be expressed in terms of the sample mean  $\bar{x}$  and the sample variance  $s^2$ . In this way one reduces the large number of data ( $n$ ) to the small size of information relevant in the current context (compare the concept of [sufficient statistics](#)).

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n [(x_i - \bar{x})(\bar{x} - \mu)]^2 \quad (\text{H.79})$$

$$= \underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{=(n-1)s^2} + 2(\bar{x} - \mu) \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} + \underbrace{\sum_{i=1}^n (\bar{x} - \mu)^2}_{=n(\bar{x} - \mu)^2} \quad (\text{H.80})$$

$$= (n-1)s^2 + n(\bar{x} - \mu)^2 \quad (\text{H.81})$$

$$= s'^2 + n(\bar{x} - \mu)^2 \quad (\text{H.82})$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{sample mean} \quad (\text{H.83})$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{sample variance} \quad (\text{H.84})$$

$$s'^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{modified sample variance} \quad (\text{H.85})$$

Thus one finally obtains the [likelihood](#)

$$f(\mathbf{x}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2} [(\bar{x} - \mu)^2 + s'^2]} \quad (\text{H.86})$$

The likelihood for  $\mu = 0$  (null hypothesis  $H_0$ ) reads

$$f_0(\mathbf{x}; \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2} [\bar{x}^2 + s'^2]} \quad (\text{H.87})$$

and for  $\mu$  arbitrary (alternative hypothesis  $H_1$ )

$$f_1(\mathbf{x}; \mu, \sigma^2) \equiv f(\mathbf{x}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2} [(\bar{x} - \mu)^2 + s'^2]}. \quad (\text{H.88})$$

## H.11.2 Marginal likelihood for $H_0$ using Jeffreys' prior

The marginal likelihood  $m_0$  for  $H_0$  is defined by

$$m_0 = \int_0^\infty f_0(\mathbf{x}; \sigma^2) \pi_0(\sigma) d\sigma \quad (\text{H.89})$$

where  $\pi_0(\sigma)$  is a prior for the unknown standard deviation,  $\sigma$ . Jeffreys (1961, p. 271, Eq. 18) uses the [prior  \$1/\sigma\$](#)  which is now-a-days called [Jeffreys' prior](#)<sup>10</sup>. The corresponding marginal likelihood  $m_0^J$  can be calculated as follows ('integrating out the nuisance parameter  $\sigma$ '):

$$m_0^J = \underbrace{\int_0^\infty (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2} [\bar{x}^2 + s'^2]} d\sigma}_{\text{likelihood}} \underbrace{\frac{1}{\sigma}}_{\text{prior}} \quad (\text{H.90})$$

<sup>10</sup>Please note that Jeffreys' prior is not a PDF (it cannot be normalized to 1 because of the singularity at  $\sigma = 0$ ) and it thus called an '[improper prior](#)'. Using improper prior may lead to unreasonable results if not handled with care. Jeffreys uses the same prior for  $H_1$  which works out fine.

$$= (2\pi)^{-n/2} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{n}{2\sigma^2} [\bar{x}^2 + s'^2]\right) d\sigma \quad \text{compare Jeffreys (1961, p. 271, Eq. 18)} \quad (\text{H.91})$$

$$= -(2\pi)^{-n/2} \int_\infty^0 z^{(n+1)/2} \exp\left(-\frac{n}{2} [\bar{x}^2 + s'^2]\right) \frac{z^{-3/2}}{2} dz \quad (\text{H.92})$$

$$= (2\pi)^{-n/2} \frac{1}{2} \int_0^\infty z^{(n-2)/2} \exp\left(-\frac{n}{2} [\bar{x}^2 + s'^2]\right) dz \quad (\text{H.93})$$

$$= (2\pi)^{-n/2} \frac{1}{2} \int_0^\infty z^m \exp(-az) dz \quad (\text{H.94})$$

$$= (2\pi)^{-n/2} \frac{1}{2} \frac{\Gamma(n/2)}{\left(\frac{n}{2} [\bar{x}^2 + s'^2]\right)^{n/2}} \quad (\text{H.95})$$

$$= (2\pi)^{-n/2} 2^{n/2-1} \frac{\Gamma(n/2)}{\left(n [\bar{x}^2 + s'^2]\right)^{n/2}} \quad \text{compare Jeffreys (1961, p. 273, Eq. 26)} \quad (\text{H.96})$$

Remark: in Jeffreys (1961, p. 273, Eq. 26) the constant  $(2\pi)^{-n/2}$  is omitted and the  $\Gamma$ -term is replaced by  $(n/2 - 1)!$  which is valid only when  $n$  is even.

Substitution:  $z = 1/\sigma^2 \Rightarrow dz/d\sigma = -2/\sigma^3$ ,  $d\sigma = \sigma^3 dz/2 = z^{-3/2} dz/2$ ; integration limits:  $\sigma \rightarrow 0 \Rightarrow z \rightarrow \infty$ ,  $\sigma \rightarrow \infty \Rightarrow z \rightarrow 0$

$$\int_0^\infty x^m e^{-ax} dx = \frac{\Gamma(m+1)}{a^{m+1}} \quad a > 0, \quad m > -1 \quad (\text{H.97})$$

with  $a = (n/2) [\bar{x}^2 + s'^2]$  and  $m = (n-2)/2 = n/2 - 1$ ,  $m+1 = n/2$ .

### H.11.3 Marginal likelihood for $H_1$ using Jeffreys' & Cauchy prior

The marginal likelihood  $m_1$  for  $H_1$  is defined by

$$m_1 = \int_{-\infty}^{\infty} d\mu \int_0^{\infty} d\sigma f_1(\mathbf{x}; \mu, \sigma^2) \pi_1(\mu, \sigma) \quad (\text{H.98})$$

where  $\pi_1(\mu, \sigma)$  is a prior for the unknown mean,  $\mu$ , and the unknown standard deviation,  $\sigma$ . For  $\sigma$  Jeffreys (1961, p. 271, Eq. 18) uses the [prior  \$1/\sigma\$](#)  which is now-a-days called [Jeffreys' prior](#). Jeffreys (1961, p. 268-269) performs a reparametrization of the likelihood for  $H_1$  by introducing the [effect size](#)<sup>11</sup>  $\delta = \mu/\sigma$ . He uses the [Cauchy prior](#)

$$\pi_{1a}(\delta) = \frac{1}{(1 + \delta^2) \pi} \quad (\text{H.99})$$

for the effect size<sup>12</sup>. The Cauchy prior is a PDF with median zero, undefined mean, and diverging variance ('fat tails'). The marginal likelihood can be calculated as follows

$$m_1^J = \int_{-\infty}^{\infty} \int_0^{\infty} f_1(\mathbf{x}; \delta, \sigma^2) \pi_1(\delta, \sigma) d\sigma d\delta \quad (\text{H.100})$$

$$= \int_{-\infty}^{\infty} \int_0^{\infty} \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}[(\bar{x}/\sigma - \delta)^2 + s'^2/\sigma^2]\right)}_{\text{likelihood}} \underbrace{\frac{1}{\sigma(1 + \delta^2)\pi}}_{\text{prior}} d\sigma d\delta \quad (\text{H.101})$$

$$= (2\pi)^{-n/2} \frac{1}{\pi} \int_{-\infty}^{\infty} \int_0^{\infty} \sigma^{-n-1} \exp\left(-\frac{n}{2}[(\bar{x}/\sigma - \delta)^2 + s'^2/\sigma^2]\right) \frac{1}{(1 + \delta^2)} d\sigma d\delta \quad (\text{H.102})$$

(compare Eq. 9 with prior Eq. (11) for  $\gamma = 1$  in Ly et al., 2016). Substitution (reparametrization):  $\delta = \mu/\sigma \Rightarrow d\delta/d\mu = 1/\sigma$ ,  $d\mu = \sigma d\delta$ , integration limits,  $\pm\infty$ , for  $\mu$  and  $\delta$  are identical. Please note that the integration given above starts immediately over  $d\delta$ , i.e. the factor  $\sigma$  from the substitution has been omitted.

"Jeffreys knew that this integral [Eq. (H.102)] is hard to compute and went to great lengths to compute an approximation that makes his Bayesian t-test usable in practice. Fortunately, we can now take advantage of computer software that can numerically solve the aforementioned integral and we therefore omit Jeffreys's approximation from further discussion. By a decomposition of a Cauchy distribution we obtain a Bayes factor of the following form:

$$B_{10;\gamma}(n, t) = \frac{\gamma \int_0^{\infty} (1 + ng)^{-1/2} \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-(\nu+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-\gamma^2/(2g)} dg}{\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}} \quad (\text{H.103})$$

" (Ly et al., 2016). The decomposition of the Cauchy prior is given, for example, in Liang et al. (2008). For  $\gamma = 1$ , Eq. H.103 reduces<sup>13</sup> to the inverse of Eq. 12.18.

Rouder et al., 2009 write: "To our knowledge, Equation 1<sup>14</sup> is novel. The derivation is straightforward and tedious and not particularly informative. Gönen et al. (2005) provided the analogous equation for the unit-information Bayes factor. Liang et al. (2008) provided the corresponding JZS Bayes factors for testing slopes in a regression model."

<sup>11</sup>The term 'effect size' is probably never used by Jeffreys (1961).

<sup>12</sup>For the justification of this choice compare Jeffreys (1961, p. 269-270).

<sup>13</sup>Eq. H.103 has been derived for the Cauchy prior with scale  $\gamma$ :  $\pi_{1a,\gamma}(\delta) = \frac{1}{\pi\gamma(1+(\delta/\gamma)^2)\pi}$

<sup>14</sup>Eq. 12.18 or  $B_{01;\gamma}(n, t)$  = the inverse of the right-hand-side of Eq. H.103 for  $\gamma = 1$

### H.11.4 $\sigma^2$ known, no prior (Rouder et al., 2009)

Let's first assume that the variance  $\sigma^2$  is known. In this case no prior for  $\sigma^2$  is required and we can directly compare the likelihoods for  $H_0$  ( $\mu = 0$ ) and  $H_1$  ( $\mu \neq 0$ ):

$$B_{10}^{\sigma^2 \text{ known}} = \frac{(2\pi\sigma^2)^{-n/2} e^{-\frac{n\mu^2 - 2n\bar{x}\mu + \sum_{j=1}^n x_j^2}{2\sigma^2}}}{(2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{j=1}^n x_j^2}{2\sigma^2}}} \quad (\text{H.104})$$

$$= e^{-\frac{n\mu^2 - 2n\bar{x}\mu}{2\sigma^2}} \quad (\text{H.105})$$

$$= \exp \left\{ -\frac{n}{2} \left( \frac{\mu}{\sigma} \right)^2 + n \frac{\bar{x}\mu}{\sigma\sigma} \right\} \quad (\text{H.106})$$

where the last form suggests measuring  $\mu$  and  $\bar{x}$  in units of the (known) standard deviation  $\sigma$ .

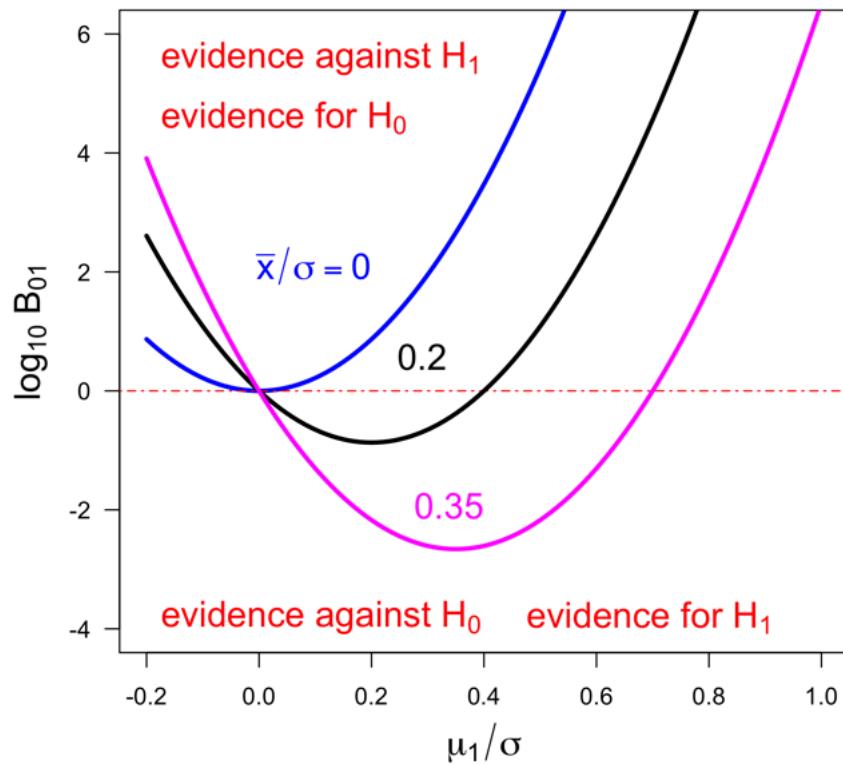


Figure H.23: Logarithm of the Bayes factor  $B_{01} = 1/B_{10}^{\sigma^2 \text{ known}}$  for known variance  $\sigma^2$  as a function of the hypothesized ( $H_1$ ) mean  $\mu_1$  (scaled by  $\sigma$ ) for scaled sample mean  $\bar{x}/\sigma = 0$  (blue line),  $= 0.2$  (black line), and  $= 0.35$  (magenta line) and sample size  $n = 100$ . The graph is a (slightly modified) version of Fig. 2A in Rouder et al. (2009). [Rouder09Fig2A.R](#)

### Discussion of results (Fig. H.23, H.24)

Values  $B_{01} > 1$  ( $\log B_{01} > 0$ ) would speak against  $H_1 (\mu \neq 0)$  and for  $H_0 (\mu = 0)$  whereas values  $B_{01} < 1$  ( $\log B_{01} < 0$ ) would speak against  $H_0 (\mu = 0)$  and for  $H_1 (\mu \neq 0)$ .

1. Shown are logarithms of Bayes factors  $B_{01}$  as function of the hypothesized mean value  $\mu_1$  for three different values of scaled sample mean values.
2. All three curves meet at  $(\mu_1 = 0, \log B_{01} = 0)$  because at  $\mu_1 = 0$  the alternative and null hypothesis are identical ( $H_1 \equiv H_0$ ) and thus  $B_{01} = 1$  and  $\log B_{01} = 0$ .
3. When the sample mean  $\bar{x}$  is zero, any alternative hypothesis with  $\mu_1 \neq 0$  would have lower likelihood than the null hypothesis and thus the Bayes factor would increase with increasing  $|\mu_1|$  (blue solid line).
4. When the scaled sample mean  $\bar{x}/\sigma$  is 0.2, the alternative hypothesis  $H_1$  is maximal supported ( $B_{01} < 1, \log B_{01} < 0$ ) when  $\mu_1/\sigma = \bar{x}/\sigma$ :  $B_{01}(\mu_1/\sigma = 0.2 = \bar{x}/\sigma) = 0.135$  or  $B_{10}(\mu_1/\sigma = 0.2 = \bar{x}/\sigma) = 7.4$ . For  $\mu_1/\sigma < 0$  and  $\mu_1/\sigma > 2\bar{x}/\sigma = 0.4$  there is evidence against  $H_1$ .
5. Why is  $B_{01} > 1$  ( $\log B_{01} > 0$ ) for  $\mu_1/\sigma > 2\bar{x}/\sigma$ ? Because in this range, the magnitude of the distance  $(\bar{x} - \mu_1)/\sigma$  is larger than the magnitude of the distance  $(\bar{x} - 0)/\sigma$ :

$$\frac{|\bar{x} - \mu_1|}{\sigma} > \frac{|\bar{x}|}{\sigma} \quad (\text{H.107})$$

The same applies for  $\mu_1/\sigma < 0$ .

6. When the scaled sample mean  $\bar{x}/\sigma$  is 0.35, the alternative hypothesis  $H_1$  is maximal supported ( $B_{01} < 1, \log B_{01} < 0$ ) when  $\mu_1/\sigma = \bar{x}/\sigma$ :  $B_{01}(\mu_1/\sigma = 0.35 = \bar{x}/\sigma) = 0.022$  or  $B_{10}(\mu_1/\sigma = 0.35 = \bar{x}/\sigma) = 457.1$ . For  $\mu_1/\sigma < 0$  and  $\mu_1/\sigma > 2\bar{x}/\sigma = 0.7$  there is evidence against  $H_1$ .

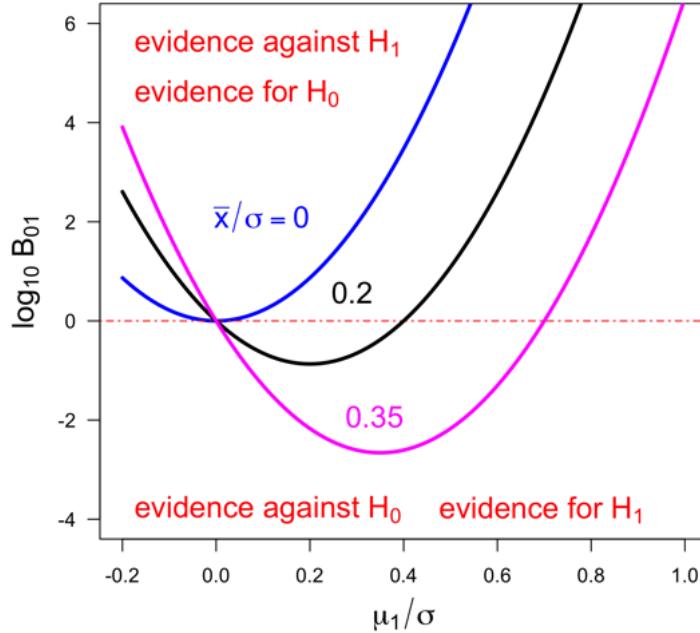


Figure H.24: Small version of Fig. H.23

**What can be learned from this exercise?** In the Bayesian approach to hypothesis testing one specifies priors for the unknown (nuisance) parameters and gets rid of these nuisance parameters by integration over the product of likelihood and prior ('integrating out nuisance parameters'). This can be seen as the calculation of a weighted (weighted by the prior) mean over many (infinitely many in case of continuously varying parameters) different hypotheses. If the prior for  $H_1$  is very broad, the Bayes factor  $B_{01}$  will be very large (speaking against  $H_1$ ) because  $H_1$  contains many point hypotheses outside the range  $0 \leq \mu_1 \leq 2\bar{x}$  with large prior weights. This is one reason why flat priors will not work over infinite regions. Thus one has to choose the priors carefully in order to not always reject  $H_1$ . Or to quote Rouder et al. (2009): "This simple example illustrates that as the alternative is placed farther from the observed data, the Bayes factor increasingly favors the null<sup>15</sup>. Moreover, when the alternative is unrealistically far from the data, the Bayes factor provides unbounded support for the null hypothesis over this alternative. This insight that unrealistic alternatives yield support for the null will be utilized in specifying appropriate alternatives."

### H.11.5 Normal prior for $\mu, \sigma^2$ known (Rouder et al., 2009)

As a first trial, Rouder et al. (2009) consider for the hypothesis  $H_1$  normal distributions (with mean zero and variances  $\sigma_\mu^2$  measured in units of the known variance  $\sigma^2$ ) as prior for  $\mu$ . The Bayes factor  $B_{01}$  thus reads

$$B_{01}^{\text{normal prior}} = \frac{\overbrace{(2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{j=1}^n x_j^2}{2\sigma^2}}}^{\text{likelihood for } H_0}}{\int_{-\infty}^{+\infty} d\mu \underbrace{(2\pi\sigma^2)^{-n/2} e^{-\frac{n\mu^2 - 2n\bar{x}\mu + \sum_{j=1}^n x_j^2}{2\sigma^2}}}_{\text{likelihood for } H_1} \underbrace{\frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{\mu^2}{2\sigma_\mu^2}}}_{\text{prior for } \mu}} \quad (\text{H.108})$$

$$= \frac{\sqrt{2\pi\sigma_\mu^2}}{\int_{-\infty}^{+\infty} d\mu e^{-\frac{n\mu^2 - 2n\bar{x}\mu}{2\sigma^2}} e^{-\frac{\mu^2}{2\sigma_\mu^2}}} \quad (\text{H.109})$$

$$= \frac{\sqrt{2\pi\sigma_\mu^2}}{\int_{-\infty}^{+\infty} d\mu e^{-\frac{n\sigma_\mu^2\mu^2 - 2n\bar{x}\sigma_\mu^2\mu + \sigma^2\mu^2}{2\sigma^2\sigma_\mu^2}}} \quad (\text{H.110})$$

$$= \frac{\sqrt{2\pi\sigma_\mu^2}}{\int_{-\infty}^{+\infty} d\mu e^{-\frac{(a\mu - c)^2 - c^2}{2\sigma^2\sigma_\mu^2}}} \quad (\text{quadratic complement}) \quad (\text{H.111})$$

$$= \frac{\sqrt{2\pi\sigma_\mu^2}}{e^{\frac{c^2}{2\sigma^2\sigma_\mu^2}} \int_{-\infty}^{+\infty} d\mu e^{-\frac{(a\mu - c)^2}{2\sigma^2\sigma_\mu^2}}} \quad (\text{H.112})$$

<sup>15</sup>'The null' is statistical slang for the null hypothesis  $H_0$ .

$$= \frac{\frac{c^2}{\sqrt{2\pi\sigma_\mu^2} e^{-\frac{c^2}{2\sigma^2\sigma_\mu^2}}}{z^2}}{\frac{1}{\sqrt{n\sigma_\mu^2 + \sigma^2}} \int_{-\infty}^{+\infty} dz e^{-\frac{c^2}{2\sigma^2\sigma_\mu^2}}} \quad (\text{substitution } z = a\mu - c) \quad (\text{H.113})$$

$$= \frac{\frac{\sqrt{2\pi\sigma_\mu^2} e^{-\frac{c^2}{2\sigma^2\sigma_\mu^2}}}{z^2}}{\frac{\sqrt{2\pi\sigma_\mu^2} e^{-\frac{c^2}{2\sigma^2\sigma_\mu^2}}}{\sqrt{n\sigma_\mu^2 + \sigma^2} \underbrace{\int_{-\infty}^{+\infty} dz \frac{1}{\sqrt{2\pi\sigma^2\sigma_\mu^2}} e^{-\frac{z^2}{2\sigma^2\sigma_\mu^2}}}_{} = 1}} \quad (\text{H.114})$$

$$= \frac{\frac{c^2}{\sqrt{n\sigma_\mu^2 + \sigma^2} e^{-\frac{c^2}{2\sigma^2\sigma_\mu^2}}}}{\sigma} \quad (\text{H.115})$$

$$= \frac{\frac{n^2\bar{x}^2\sigma_\mu^4}{\sqrt{n\sigma_\mu^2 + \sigma^2} e^{-\frac{2\sigma^2\sigma_\mu^2(n\sigma_\mu^2 + \sigma^2)}{\sigma}}}}{\sigma} \quad (\text{H.116})$$

$$= \frac{\frac{n^2\bar{x}^2\sigma_\mu^2}{\sqrt{n\sigma_\mu^2 + \sigma^2} e^{-\frac{2\sigma^2\sigma_\mu^2(n\sigma_\mu^2 + \sigma^2)}{\sigma}}}}{\sigma} \quad (\text{H.117})$$

$$= \frac{\sqrt{n\sigma_\mu^2 + \sigma^2}}{\sigma} e^{-\frac{n^2(\bar{x}/\sigma)^2(\sigma_\mu/\sigma)^2}{2n(\sigma_\mu/\sigma)^2 + 2}} \quad (\text{H.118})$$

Find quadratic complement:

$$(n\sigma_\mu^2 + \sigma^2)\mu^2 - 2n\bar{x}\sigma_\mu^2\mu = a^2\mu^2 - b\mu + c^2 - c^2 = (a\mu - c)^2 - c^2 \quad (\text{H.119})$$

where

$$a^2 = n\sigma_\mu^2 + \sigma^2 \quad (\text{H.120})$$

$$b = 2n\bar{x}\sigma_\mu^2 = 2ac \quad (\text{H.121})$$

$$c^2 = \frac{b^2}{4a^2} = \frac{n^2\bar{x}^2\sigma_\mu^4}{n\sigma_\mu^2 + \sigma^2} \quad (\text{H.122})$$

Substitution:  $z = a\mu - c \Rightarrow dz/d\mu = a$ ,  $d\mu = \frac{dz}{a} = \frac{dz}{\sqrt{n\sigma_\mu^2 + \sigma^2}}$ , integration limits do not change because  $z \rightarrow \pm\infty$  when  $\mu \rightarrow \pm\infty$ .

The solution given by Rouder et al. (2009, footnote 3, p. 236; slight change of notation:  $n \equiv N$ ) is identical to the expression derived above:

$$B_{01} = \left( \phi \frac{\sigma_\mu^2}{\sigma^2} \right)^{1/2} \exp \left( -\frac{n^2\bar{x}^2}{2\phi\sigma^2} \right) \quad (\text{H.123})$$

$$= \left( n \frac{\sigma_\mu^2}{\sigma^2} + 1 \right)^{1/2} \exp \left( - \frac{n^2 \bar{x}^2 \sigma_\mu^2}{2\sigma^2 (n\sigma_\mu^2 + \sigma^2)} \right) \quad (\text{H.124})$$

$$= \frac{1}{\sigma} \left( n\sigma_\mu^2 + \sigma^2 \right)^{1/2} \exp \left( - \frac{n^2 (\bar{x}/\sigma)^2 (\sigma_\mu/\sigma)^2}{2n(\sigma_\mu/\sigma)^2 + 2} \right) \quad (\text{H.125})$$

where

$$\phi = n + \frac{\sigma^2}{\sigma_\mu^2} = \frac{n\sigma_\mu^2 + \sigma^2}{\sigma_\mu^2} \quad (\text{H.126})$$

### H.11.6 Unit-information prior (Rouder et al., 2009)

Effect size is defined by  $\delta = \mu/\sigma$ . The normal PDF with mean  $\mu_\delta = 0$  (no bias with respect to sign) and variance  $\sigma_\delta^2 = 1$  (emphasis on small effect sizes; large effects, namely  $|\delta| \gg 1$ , are so obvious that they do not need application of a hypothesis test) is chosen as prior for  $\delta$ .

$$B_{01}^{\text{normal prior}} = \frac{(2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{j=1}^n x_j^2}{2\sigma^2}}}{\int_{-\infty}^{+\infty} d\mu (2\pi\sigma^2)^{-n/2} e^{-\frac{n\mu^2 - 2n\bar{x}\mu + \sum_{j=1}^n x_j^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{-\frac{\delta^2}{2\sigma_\delta^2}}} \quad (\text{H.127})$$

$$= \frac{1}{\sigma \int_{-\infty}^{+\infty} d\delta e^{-\frac{n\delta^2 - 2n(\bar{x}/\sigma)\delta}{2}} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{-\frac{\delta^2}{2\sigma_\delta^2}}} \quad (\text{H.128})$$

$$\underbrace{\sigma_\delta^2}_{=1} = \frac{1}{\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} d\delta e^{-\frac{n\delta^2 - 2n(\bar{x}/\sigma)\delta}{2}} e^{-\frac{\delta^2}{2}}} \quad (\text{H.129})$$

$$= \frac{1}{\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} d\delta e^{-\frac{(n+1)\delta^2 - 2n(\bar{x}/\sigma)\delta}{2}}} \quad (\text{H.130})$$

$$= \frac{1}{\frac{\sigma}{\sqrt{2\pi}} e^{\frac{c^2}{2}} \int_{-\infty}^{+\infty} d\delta e^{-\frac{(a\delta - c)^2}{2}}} \quad (\text{H.131})$$

$$= \frac{1}{\frac{\sigma}{a} e^{\frac{c^2}{2}} \underbrace{\int_{-\infty}^{+\infty} dz \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}_{=1}} \quad (\text{H.132})$$

$$= \frac{1}{\frac{\sigma}{n+1} e^{\frac{n^2(\bar{x}/\sigma)^2}{2(n+1)}}} \quad (\text{H.133})$$

$$= \frac{1}{\frac{\sigma}{n+1} \exp \left\{ \frac{n^2(\bar{x}/\sigma)^2}{2(n+1)} \right\}} \quad (\text{H.134})$$

Substitution 1:  $d\delta/d\mu = 1/\sigma \Rightarrow d\mu = \sigma d\delta$ ; integration limits  $\pm\infty$  do not change.

Quadratic complement:

$$(n+1)\delta^2 - 2n(\bar{x}/\sigma)\delta = a^2\delta^2 - b\delta + c^2 - c^2 = (a\delta - c)^2 - c^2 \quad (\text{H.135})$$

where

$$a^2 = n+1 \quad (\text{H.136})$$

$$b = 2n(\bar{x}/\sigma) = 2ac \quad (\text{H.137})$$

$$c^2 = \frac{b^2}{4a^2} = \frac{n^2(\bar{x}/\sigma)^2}{n+1} \quad (\text{H.138})$$

Substitution 2:  $z = a\delta - c \Rightarrow dz/d\delta = a, d\delta = dz/a$

## H.12 Wilcoxon-Mann-Whitney test

The 2-sample t-test is valid for samples from normal distributions with equal variances. Often these prerequisites are not fulfilled and non-parametric tests can be an alternative. Wilcoxon (1945) proposed a test that is based on ranks of data.<sup>16</sup> Mann & Whitney (1947) modified the test statistic of Wilcoxon and extended the test to cases where the sizes of the two samples can be different from each other. Today, tests with various different test statistics run under identical or similar names which may lead to different p-values when applying routines from different statistical packages. Here, we will use the test statistic U originally proposed by Mann & Whitney (1947).

Various non-parametric tests are based on the ranks of data. For the Wilcoxon-Mann-Whitney test one considers two samples:  $\mathbf{x} = x_1, x_2, \dots, x_n$  with  $n$  data points and  $\mathbf{y} = y_1, y_2, \dots, y_m$  with  $m$  data points. The two samples are combined to a single entity  $\mathbf{xy} = x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$  of length  $n + m$ . Ordering of the entries in  $\mathbf{xy}$  according to size (from small to large) leads to a certain sequence of values as, for example,  $y_3, y_2, x_5, y_7, x_2, \dots, y_6$ . The position in this sequence is called the rank. When several data possess identical values<sup>17</sup> the ranks of these data are first summed up and than divided by the number of identical values (see example below). Wilcoxon (1945) proposed the test statistic  $T$  which is given by the sum of the ranks of sample  $\mathbf{y}$ . Mann & Whitney (1947) proposed instead the test statistic  $U$  which is given by

$$U = m \cdot n + \frac{m(m+1)}{2} - T \quad (\text{H.139})$$

which can vary between 0 and  $m \cdot n$ . The calculation of the exact CDF( $U$ ) under the null hypothesis is tedious (compare Section H.14) when sample sizes are small ( $n$  or  $m$  or both  $< 9$ ); for larger sample sizes, a normal approximation (Subsection H.14.1) yield good results. In the routine **MannWhitney()** (written in R) the normal approximation is applied for large sample sizes and probabilities estimated by Monte Carlo simulations are used for small sample sizes.

### Example 1 (samples from normal PDFs, small sample sizes)

Given two samples

$$\mathbf{x} = 0.022, -0.904, 0.413, 0.187, 0.231 \quad (\text{H.140})$$

$$\mathbf{y} = 0.536, 1.784, 1.399, 0.295 \quad (\text{H.141})$$

of length  $n = 5$  and  $m = 4$ , respectively. Combining these two samples yields

$$\mathbf{xy} = 0.022, -0.904, 0.413, 0.187, 0.231, 0.536, 1.784, 1.399, 0.295 \quad (\text{H.142})$$

with ranks

$$\mathbf{r}_{\mathbf{xy}} = 2, 1, 6, 3, 4, 7, 9, 8, 5 \quad (\text{H.143})$$

i.e.  $x_2 = -0.904$  (smallest value in  $\mathbf{xy}$ ) has rank 1 and  $y_3 = 1.784$  (largest value in  $\mathbf{xy}$ ) has rank 9. The sum of ranks  $T$  for  $\mathbf{y}$  reads

$$T = 7 + 9 + 8 + 5 = 29 \quad (\text{H.144})$$

and thus the test statistic  $U$  is

$$U = m \cdot n + \frac{m(m+1)}{2} - T = 4 \cdot 5 + \frac{4(4+1)}{2} - 29 = 1 \quad (\text{H.145})$$

The probability to observe such a low  $U$  value or lower  $U$  values ('left tail') under the null hypothesis  $H_0$  'both samples are from the same population' is  $p = 0.016$  and thus we reject  $H_0$  on the significance level  $\alpha = 0.05$ . This result is consistent with common sense: by looking at the data we recognize that there is little overlap between the two samples ( $\mathbf{y}$  is probably from a population with larger mean value).

<sup>16</sup>Gelman et al. (2020, p. 97) discuss the Wilcoxon rank test from a Bayesian point of view suggesting to interpret the introduction of ranks as a non-linear transformation of data.

<sup>17</sup>When sampling from continuously varying population the probability for such cases (called 'ties') is almost zero, however, a finite number of recorded digits or sampling from discrete populations can easily lead to ties.

**Example 2 (count data, small sample sizes, with ties):**  
given two samples

$$x = 4, 1, 7, 8, 9 \quad (\text{H.146})$$

$$y = 2, 3, 1, 5 \quad (\text{H.147})$$

of length  $n = 5$  and  $m = 4$ , respectively. Combining these two samples yields

$$xy = 4, 1, 7, 8, 9, 2, 3, 1, 5 \quad (\text{H.148})$$

with ranks

$$r_{xy} = 5, 1.5, 7, 8, 9, 3, 4, 1.5, 6 \quad (\text{H.149})$$

where the ranks of  $x_2 = 1$  and  $y_3 = 1$  both have rank 1.5 (mean of ranks 1 and 2; 'tie'). The sum of ranks  $T$  for  $y$  is

$$T = 3 + 4 + 1.5 + 6 = 14.5 \quad (\text{H.150})$$

and thus the test statistic  $U$  is

$$U = m \cdot n + \frac{m(m+1)}{2} - T = 4 \cdot 5 + \frac{4(4+1)}{2} - 14.5 = 15.5 \quad (\text{H.151})$$

The probability to observe such a  $U$  value or lower  $U$  values ('left tail') under the null hypothesis  $H_0$  'both samples are from the same population' is  $p = 0.905$  or to observe such a  $U$  value or larger  $U$  values ('right tail') is  $p = 0.095$  and thus one can not reject  $H_0$  on the significance level  $\alpha = 0.05$  (more data are required!).

**Further reading:** Birnbaum (1956), Wilcoxon et al. (1970)

**Exercise 79 Tapeworms**

Ott & Longnecker (2001, p. 272-274) consider two samples of count data

$$y_d = \{18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7\} \quad (\text{H.152})$$

$$y_u = \{40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39\} \quad (\text{H.153})$$

where  $y_d$  are the number of tapeworms in stomachs of sheep treated with a drug (supposed to be against worms) and  $y_u$  are the number of tapeworms in untreated sheep. The null hypothesis  $H_0$  is 'drug-treated and untreated sheep have the same amount of worms in their stomachs' (or in other words: the drug has no effect). The alternative hypothesis  $H_a$  is 'drug-treated and untreated sheep do not have the same amount of worms in their stomachs' (or in other words: the drug has an effect; whether the effect is positive or negative is another question).

The requirements for the t-test are not fulfilled: (1) the histograms of both samples (Fig. H.25) are far away from what to expect for samples from normal distributions and (2) count data (integers!) cannot (by definition) be normally distributed. Apply the Wilcoxon-Mann-Whitney test and compare the results with those from the t-test.

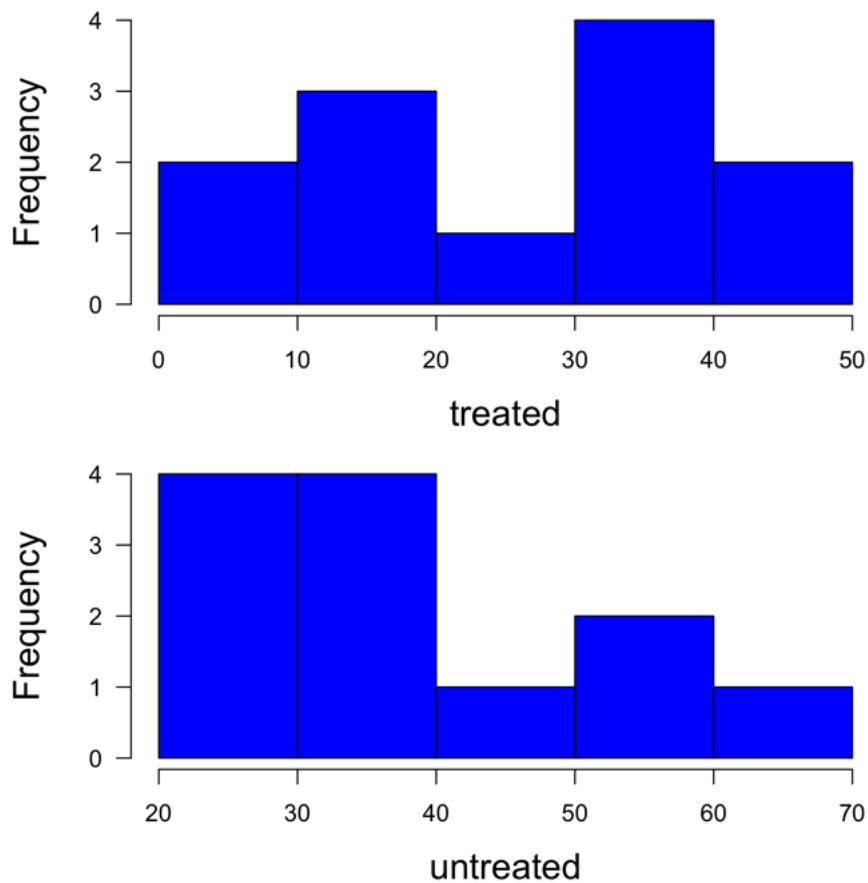


Figure H.25: Histograms of the tapeworm data. [Solution-2-sampleEqualMeans.R](#)

**Exercise 80 Heights of males & females**

Zar (2010, p. 164) discusses a simple example. The heights (in cm) of male and female students read

$$x_1 = \{193, 188, 185, 183, 180, 175, 170\} \quad \text{males} \quad (\text{H.154})$$

$$x_2 = \{178, 173, 168, 165, 163\} \quad \text{females.} \quad (\text{H.155})$$

The null hypothesis  $H_0$  shall be 'male and female students are the same height'. The alternative hypothesis  $H_a$  is 'male and female students are not the same height'. Apply the Wilcoxon-Mann-Whitney test.

## H.13 Wilcoxon paired-sample test

When data come in pairs (for example, when measuring two quantities on the same individual) they are often related to each other and thus not truly independent. In these cases, one calculates the pairwise differences which can be seen as independent from each other. The Wilcoxon paired-sample test is applied to these differences.

### H.13.1 Wilcoxon (1945) example

Given the stand of wheat (units not specified) under two treatments

$$A = (209, 200, 177, 169, 159, 169, 187, 198) \quad (\text{H.156})$$

$$B = (151, 168, 147, 164, 166, 163, 176, 188) \quad (\text{H.157})$$

one obtains the differences

$$d = x - y = c(58, 32, 30, 5, -7, 6, 11, 10). \quad (\text{H.158})$$

The null hypothesis  $H_0$  is 'no difference between treatments'. By looking at the differences we can already guess that  $H_0$  should be rejected because there are much more positive (7) than negative differences (1) whereas a true  $H_0$  would favor a similar number of positive and negative differences. However, the test statistics  $T_+$  and  $T_-$  are not just based on the number of positive or negative differences but take into account the ranks of the absolute differences. Why the ranks of absolute differences and not just the ranks of differences? Because all negative differences would have smaller ranks than positive differences and there would be already a large difference when  $H_0$  is true. The ranks of the magnitude of the difference ( $|d|$ ) read

$$r = (8, 7, 6, 1, 3, 2, 5, 4). \quad (\text{H.159})$$

i.e.  $r_5 = 3$  for  $d_5 = -7$  although it is the smallest (= most negative) difference. By multiplication with the signs of the differences one obtains the signed ranks

$$\mathbf{sr} = (8, 7, 6, 1, -3, 2, 5, 4) \quad (\text{H.160})$$

The test statistics  $T_+$  and  $T_-$  are defined by summing over all positive or negative signed ranks, respectively, and taking the absolute value of these sums. For the example discussed here, one obtains

$$T_+ = 8 + 7 + 6 + 1 + 2 + 5 + 4 = 33 \quad (\text{H.161})$$

$$T_- = |-3| = 3 \quad (\text{H.162})$$

If  $H_0$  is true one expects similar values for  $T_+$  and  $T_-$ . This is obviously not the case here. The critical value for the two-sided (i.e. violation of  $H_0$  because mean difference of population is larger or small than 0) Wilcoxon paired-sample test, a level of significance chosen as  $\alpha = 0.05$ , and  $n = 8$  differences (which is equal to the number of degrees of freedom  $v$  because we do not have to estimate a population parameter) is given by

$$T_{\alpha(2),v} = T_{0.05(2),8} = 3 \quad (\text{H.163})$$

(Zar, 2010, Table B.12 in appendix). The null hypothesis  $H_0$  is rejected when  $T_+$  or  $T_-$  are smaller than or equal to  $T_{\alpha(2),v}$  which is the case here because

$$T_- = 3 \leq 3 = T_{0.05(2),8} \quad (\text{H.164})$$

The p-value for the two-sided Wilcoxon paired-sample test is calculated by

$$p = \sum_{i=0}^{\min(T_-, T_+)} p_i + 1 - \sum_{j=0}^{\max(T_-, T_+)} p_j \quad (\text{H.165})$$

where the  $p_i$  are the probabilities to obtain  $T_+ = i$  or  $T_- = i$ , respectively. For the data of Wilcoxon (1945) one obtains  $p = 0.0385$  ('exact' p-value based on Monte Carlo simulations) and  $p = 0.0357$  based on the normal approximation (i.e. even for a small data set with  $n = 8$  pairs, the normal approximation yields a value that is close (about 10% difference) to the exact value). [NHST-pairedWilcoxon.R](#) [WilcoxonPaired.R](#)

### H.13.2 Wilcoxon paired-sample test: Zar (2010, Example 9.4)

Given the measurements of hind leg ( $x$ ) and foreleg lengths ( $y$ )

$$x = (142, 140, 144, 144, 142, 146, 149, 150, 142, 148) \quad \text{hind leg lengths (cm)} \quad (\text{H.166})$$

$$y = (138, 136, 147, 139, 143, 141, 143, 145, 136, 146) \quad \text{foreleg lengths (cm)} \quad (\text{H.167})$$

one obtains the differences (in cm)

$$d = x - y = c(4, 4, -3, 5, -1, 5, 6, 5, 6, 2). \quad (\text{H.168})$$

The null hypothesis  $H_0$  is 'no difference between length of hind leg and foreleg'. By looking at the differences one can already guess that  $H_0$  should be rejected because there are much more positive (8) than negative differences (2) whereas a true  $H_0$  would favor a similar number of positive and negative differences. However, the test statistics  $T_+$  and  $T_-$  are not just based on the number of positive or negative differences but take into account the ranks of the absolute differences. Why the ranks of absolute differences and not just the ranks of differences? Because all negative differences would have smaller ranks than positive differences and there would be already a large difference when  $H_0$  is true. The ranks of the magnitude of the difference ( $|d|$ ) read

$$r = (4.5, 4.5, 3, 7, 1, 7, 9.5, 7, 9.5, 2), \quad (\text{H.169})$$

i.e.  $r_3 = 3$  for  $d_3 = -3$  although it is the smallest (= most negative) difference and the first two differences possess the same non-integer rank values (4.5) because their absolute differences are equal to each other ('tie'). By multiplication with the signs of the differences one obtains the [signed ranks](#)

$$\mathbf{sr} = (4.5, 4.5, -3, 7, -1, 7, 9.5, 7, 9.5, 2). \quad (\text{H.170})$$

The test statistics  $T_+$  and  $T_-$  are defined by summing over all positive or negative signed ranks, respectively, and taking the absolute value of these sums. For the example discussed here, one obtains

$$T_+ = 4.5 + 4.5 + 7 + 7 + 9.5 + 7 + 9.5 + 2 = 51 \quad (\text{H.171})$$

$$T_- = |-3 - 1| = 4 \quad (\text{H.172})$$

If  $H_0$  is true one expects similar values for  $T_+$  and  $T_-$ . Here this is obviously not the case. The critical value for the two-sided Wilcoxon paired-sample test,  $\alpha = 0.05$ , and  $n = 10 = \nu^{18}$  is given by

$$T_{\alpha(2), \nu} = T_{0.05(2), 10} = 8 \quad (\text{H.173})$$

(Zar, 2010, p. 185 & Table B.12 in appendix). The null hypothesis  $H_0$  is rejected when  $T_+$  or  $T_-$  are smaller than or equal to  $T_{\alpha(2), \nu}$  which is the case here because

$$T_- = 4 \leq 8 = T_{0.05(2), 10} \quad (\text{H.174})$$

The p-value for the two-sided Wilcoxon paired-sample test is calculated by

$$p = \sum_{i=0}^{\min(T_-, T_+)} p_i + 1 - \sum_{j=0}^{\max(T_-, T_+)} p_j \quad (\text{H.175})$$

where the  $p_i$  are the probabilities to obtain  $T_+ = i$  or  $T_- = i$ , respectively. For the data of Zar (2010, Example 9.4) one obtains  $p = 0.0141$  ('exact' p-value based on Monte Carlo simulations) and  $p = 0.0166$  based on the normal approximation (i.e. even for a small data set with  $n = 10$  pairs, the normal approximation yields a value that is close (about 10% difference) to the exact value).

---

<sup>18</sup> $\nu = n$  because no constraint (calculation of parameter) is necessary.

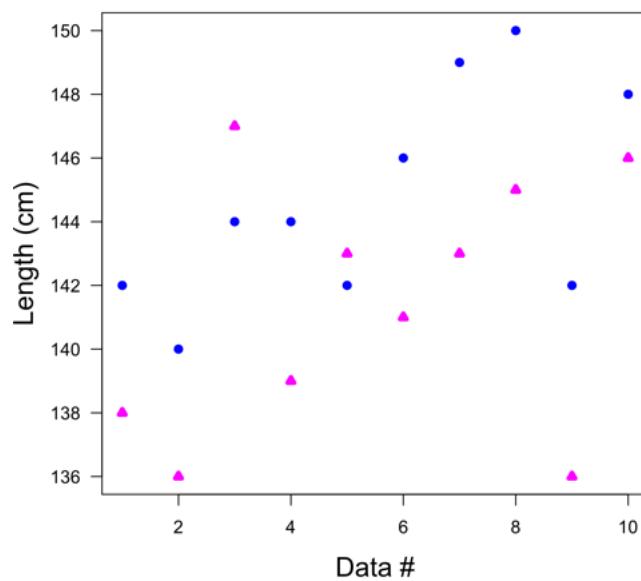


Figure H.26: Zar (2010, Example 9.4): lengths (in cm) of hind leg and foreleg of deers.  
[NHST-pairedWilcoxon9d4Zar.R](#) (line 12: set sflag to 1)

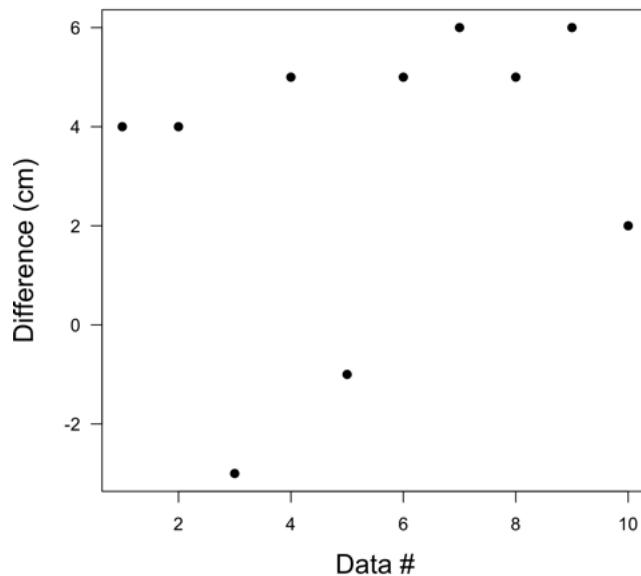


Figure H.27: Zar (2010, Example 9.4): difference in lengths of hind leg and foreleg of deers.  
[NHST-pairedWilcoxon9d4Zar.R](#) (line 12: set sflag to 2)

## H.14 Wilcoxon-Mann-Whitney test: calculation of probabilities (\*)

The cumulative distribution function  $CDF(U)$  for the test statistic  $U$  for large sample sizes  $n$  and  $m$  approaches the CDF for a normal distribution and thus a good approximation of the observed significance level  $p$  ( $p$ -value) is easy to calculate if  $n > 8$  and  $m > 8$  (Subsection H.14.1). However, for small sample sizes ( $n$  or  $m \leq 8$ ) the method for calculating  $CDF(U)$  proposed by Mann & Whitney (1947) based on recurrence relationships is tedious (Subsection H.14.2). An efficient (and mathematical interesting) alternative is based on the probability generating function for the Mann-Whitney statistic  $U$  (Di Buccianico, 1999, van de Wiel, 2000; Subsection H.14.3). Finally,  $CDF(U)$  can be estimated by Monte Carlo simulations (Subsection H.14.4).

### H.14.1 Normal approximation (\*)

When both  $n$  and  $m$  are  $> 8$ , the cumulative distribution function for the test statistic  $U$  ( $CDF_U$ ) approaches the CDF for the normal distribution with  $\mu = n \cdot m / 2$  and variance  $\sigma^2 = n \cdot m (n + m + 1) / 12$ . The standardized statistic

$$z = \frac{U - \mu}{\sigma} \quad (\text{H.176})$$

follows the CDF of the standard normal distribution,  $\Phi$ . Thus the probability  $p$  of the left tail is calculated by

$$p = \int_{-\infty}^z \Phi(x) dx \quad (\text{H.177})$$

or in R by  $p = \text{pnorm}(z)$ . For the right tail, the probability is given by

$$p = \int_z^{+\infty} \Phi(x) dx \quad (\text{H.178})$$

or in R by  $p = 1 - \text{pnorm}(z)$ .

### H.14.2 Recurrence relations (\*)

Test statistic  $U$  is related to the test statistic  $T$  (= sum of the ranks of  $y$ 's in the ordered sequence of  $x$ 's and  $y$ 's) proposed by Wilcoxon (1945):

$$U = mn + \frac{m(m+1)}{2} - T \quad (\text{H.179})$$

where  $n$  is the number of data in sample  $x$  and  $m$  is the number of data in sample  $y$ . The task is to derive the probability distribution for  $U$  when the null hypothesis is true. Since it is only the relation between  $x_i$  and  $y_j$  that matters one can replace each  $x_i$  by 0 and each  $y_j$  by 1.  $U$  is number of times a 1 proceeds a 0 (for example,  $U = 0$  for 0, 0, 1, 1 and  $U = 3$  for 1, 0, 1, 0). Mann & Whitney (1947) derived recursion relations for the

(1) cumulative probabilities  $\tilde{p}_{nm}(U)$  (= probability to observe test statistic  $\leq U$ )

$$\tilde{p}_{nm}(U) = \tilde{p}_{n-1m}(U-m) + \tilde{p}_{nm-1}(U) \quad (\text{H.180})$$

where  $\tilde{p}_{ij}(U) = 0$  if  $U < 0$  and  $\tilde{p}_{i0}(U), \tilde{p}_{0j}(U)$  are 0 or 1 according to  $U \neq 0$  or  $U = 0$  (examples:  $\tilde{p}_{30}(1) = 0$ ;  $\tilde{p}_{30}(0) = 1$ ) and for

(2) probabilities  $p_{nm}(U)$  for single  $U$  values

$$p_{nm}(U) = \frac{n}{n+m} p_{n-1m}(U-m) + \frac{m}{n+m} p_{nm-1}(U) \quad (\text{H.181})$$

where

$$p_{ij}(U) = 0 \quad \text{if } U < 0 \quad (\text{H.182})$$

$$p_{i0}(U) = 0 \quad \text{if } U \neq 0 \quad (\text{H.183})$$

$$p_{i0}(U) = 1 \quad \text{if } U = 0 \quad (\text{H.184})$$

$$p_{0j}(U) = 0 \quad \text{if } U \neq 0 \quad (\text{H.185})$$

$$p_{0j}(U) = 1 \quad \text{if } U = 0 \quad (\text{H.186})$$

Examples:  $p_{30}(1) = 0$ ;  $p_{30}(0) = 1$ . Other start values for the recurrence relation can be derived easily by writing down all possible sequences of  $n$  zeros and  $m$  ones (all with equal probability because of the Principle of Indifference), calculating  $U$ , and finally the probability for a certain  $U$  value.

Example A (direct calculation):  $n = 3, m = 1 \Rightarrow 4$  sequences with equal probability:

$$1. (0, 0, 0, 1) \Rightarrow U = 0$$

$$2. (0, 0, 1, 0) \Rightarrow U = 1$$

$$3. (0, 1, 0, 0) \Rightarrow U = 2$$

$$4. (1, 0, 0, 0) \Rightarrow U = 3$$

$$\Rightarrow p_{31}(0) = 1/4, p_{31}(1) = 1/4, p_{31}(2) = 1/4, p_{31}(3) = 1/4$$

Example A (applying the recurrence relation Eq. H.181):  $n = 3, m = 1 \Rightarrow$

$$p_{31}(0) = \frac{3}{4} \underbrace{p_{21}(-1)}_{= 0 \text{ (Eq. H.182)}} + \frac{1}{4} \underbrace{p_{30}(0)}_{= 1 \text{ (Eq. H.184)}} = 1/4 \quad (\text{H.187})$$

$$p_{31}(1) = \frac{3}{4} \underbrace{p_{21}(0)}_{= 1/3} + \frac{1}{4} \underbrace{p_{30}(1)}_{= 1 \text{ (Eq. H.183)}} = 1/4 \quad (\text{H.188})$$

$n = 2, m = 1 \Rightarrow 3$  sequences with equal probability:

$$1. (0, 0, 1) \Rightarrow U = 0$$

$$2. (0, 1, 0) \Rightarrow U = 1$$

$$3. (1, 0, 0) \Rightarrow U = 2$$

$$\Rightarrow p_{21}(0) = 1/3$$

$$p_{31}(2) = \frac{3}{4} \underbrace{p_{21}(1)}_{= 1/3} + \frac{1}{4} \underbrace{p_{30}(2)}_{= 0} = 1/4 \quad (\text{H.189})$$

$$p_{31}(3) = 1 - p_{31}(0) - p_{31}(1) - p_{31}(2) = 1/4 \quad (\text{H.190})$$

Accordingly, the cumulative probabilities  $\tilde{p}_{nm}(U)$  for example A read:

$$\tilde{p}_{31}(0) = p_{31}(0) = 1/4 \quad (\text{H.191})$$

$$\tilde{p}_{31}(1) = p_{31}(0) + p_{31}(1) = 1/4 + 1/4 = 1/2 \quad (\text{H.192})$$

$$\tilde{p}_{31}(2) = p_{31}(0) + p_{31}(1) + p_{31}(2) = 1/4 + 1/4 + 1/4 = 3/4 \quad (\text{H.193})$$

$$\tilde{p}_{31}(3) = 1 \quad (\text{H.194})$$

The first three values<sup>19</sup> are identical to the values listed in Table 1 of Mann & Whitney (1947).

Example B (applying the recurrence relation Eq. H.181):  $n = 3, m = 2$

$$p_{32}(0) = \frac{3}{5} \underbrace{p_{22}(-2)}_{= 0 \text{ (Eq. H.182)}} + \frac{2}{5} \underbrace{p_{31}(0)}_{= 1/4 \text{ (Eq. H.187)}} = 1/10 \quad (\text{H.195})$$

$$p_{32}(1) = \frac{3}{5} \underbrace{p_{22}(-1)}_{= 0 \text{ (Eq. H.182)}} + \frac{2}{5} \underbrace{p_{31}(1)}_{= 1/4 \text{ (Eq. H.188)}} = 1/10 \quad (\text{H.196})$$

$$p_{32}(2) = \frac{3}{5} \underbrace{p_{22}(0)}_{= 1/6 \text{ (Eq. H.198)}} + \frac{2}{5} \underbrace{p_{31}(2)}_{= 1/4 \text{ (Eq. H.189)}} = 1/5 \quad (\text{H.197})$$

---

<sup>19</sup>The forth value ( $\tilde{p}_{31}(3) = 1$ ) was left out in order to save space.

$$n = 2, m = 2 \Rightarrow \frac{(n+m)!}{n!m!} = \frac{4!}{2!2!} = 6 \text{ sequences with equal probability:}$$

1.  $(0, 0, 1, 1) \Rightarrow U = 0$
  2.  $(0, 1, 0, 1) \Rightarrow U = 1$
  3.  $(0, 1, 1, 0) \Rightarrow U = 2$
  4.  $(1, 0, 0, 1) \Rightarrow U = 2$
  5.  $(1, 0, 1, 0) \Rightarrow U = 3$
  6.  $(1, 1, 0, 0) \Rightarrow U = 4$
- $\Rightarrow$

$$p_{22}(0) = 1/6 \quad (\text{H.198})$$

$$p_{22}(1) = 1/6 \quad (\text{H.199})$$

$$\begin{aligned} p_{32}(3) &= \frac{3}{5} \underbrace{p_{22}(1)}_{=1/6 \text{ (Eq. H.199)}} + \frac{2}{5} \underbrace{p_{31}(3)}_{=1/4 \text{ (Eq. H.190)}} = 1/5 \\ &= 1/6 \end{aligned} \quad (\text{H.200})$$

Accordingly, the cumulative probabilities  $\tilde{p}_{nm}(U)$  for example B read:

$$\tilde{p}_{32}(0) = p_{32}(0) = 1/10 = 0.1 \quad (\text{H.201})$$

$$\tilde{p}_{32}(1) = p_{32}(0) + p_{32}(1) = 1/10 + 1/10 = 1/5 = 0.2 \quad (\text{H.202})$$

$$\tilde{p}_{32}(2) = p_{32}(0) + p_{32}(1) + p_{32}(2) = 1/10 + 1/10 + 1/5 = 2/5 = 0.4 \quad (\text{H.203})$$

$$\tilde{p}_{32}(3) = p_{32}(0) + p_{32}(1) + p_{32}(2) + p_{32}(3) = 1/10 + 1/10 + 1/5 + 1/5 = 3/5 = 0.6 \quad (\text{H.204})$$

$$\tilde{p}_{32}(4) = 1 \quad (\text{H.205})$$

### H.14.3 Generating function (\*)

Mann & Whitney (1947) calculated exact probabilities for small sizes for the test statistic  $U$  under the null hypothesis by applying recursion relations. This is a tedious way to obtain probabilities. An alternative is to use the probability generating function for the Mann-Whitney statistic  $U$  (Di Buccianico, 1999, van de Wiel, 2000)

$$\sum_{k=0}^{n+m} p_{nmk} x^k = \frac{n! m!}{(n+m)!} \frac{\prod_{k=m+1}^{m+n} (1-x^k)}{\prod_{k=1}^n (1-x^k)} = f(x; n, m) = f(x) \quad (\text{H.206})$$

where  $n$  and  $m$  are the sizes of samples  $x$  and  $y$ , respectively, and  $p_{nmk}$  are the probabilities for observing  $U = k$ . The probabilities  $p_{nmk}$  can be derived by comparing coefficients with the Taylor expansion of the generating function  $f(x)$  around 0:

$$\sum_{k=0}^{n+m} p_{nmk} x^k = f(0) + \left. \frac{df}{dx} \right|_{x=0} x + \frac{1}{2!} \left. \frac{d^2 f}{dx^2} \right|_{x=0} x^2 + \frac{1}{3!} \left. \frac{d^3 f}{dx^3} \right|_{x=0} x^3 + \dots \quad (\text{H.207})$$

Accordingly, the probability for  $U = 0$  is given by

$$p(n, m, U = 0) = f(0; n, m) = \frac{n! m!}{(n+m)!} = \frac{1}{\binom{n+m}{n}} \quad (\text{H.208})$$

and the probabilities for  $U = k > 0$  by the  $k$ th derivative at 0 divided by  $k!$

$$p(n, m, U = k) = \frac{1}{k!} \left. \frac{d^k f}{dx^k} \right|_{x=0}. \quad (\text{H.209})$$

It is quite remarkable that the generating function (Fig. H.28 for  $n = 3 = m$ ) in a small environment of  $x = 0$  contains information about all these probabilities.

As an example, the case of  $n = 3, m = 3$  is considered here. The generating function reads

$$f(x; n = 3, m = 3) = \underbrace{\frac{3! 3!}{6!}}_{= 0.05} \frac{(1 - x^4)(1 - x^5)(1 - x^6)}{(1 - x)(1 - x^2)(1 - x^3)} \quad (\text{H.210})$$

(Fig. H.28). Numerical calculation of the derivatives using the usual symmetric difference quotients (2 and 3 point, respectively, for the 1. and 2. derivative) gives reasonable results. However, this approach does not give meaningful results for higher derivatives. Thus it is necessary to evaluate analytical derivatives of the generation function.<sup>20</sup> In R one can formulate the generation function as an '**expression**'. Applying the differentiation operator '**D**' on the expression yields another (unevaluated) expression than can be differentiated further. Evaluation of the resulting expressions at  $x = 0$  yields the searched for probabilities (compare R code below and Fig. H.29 for the probabilities for single  $U$  values and for the cumulative probability distribution). Estimates from Monte Carlo simulations (Fig. H.30) yield similar results.

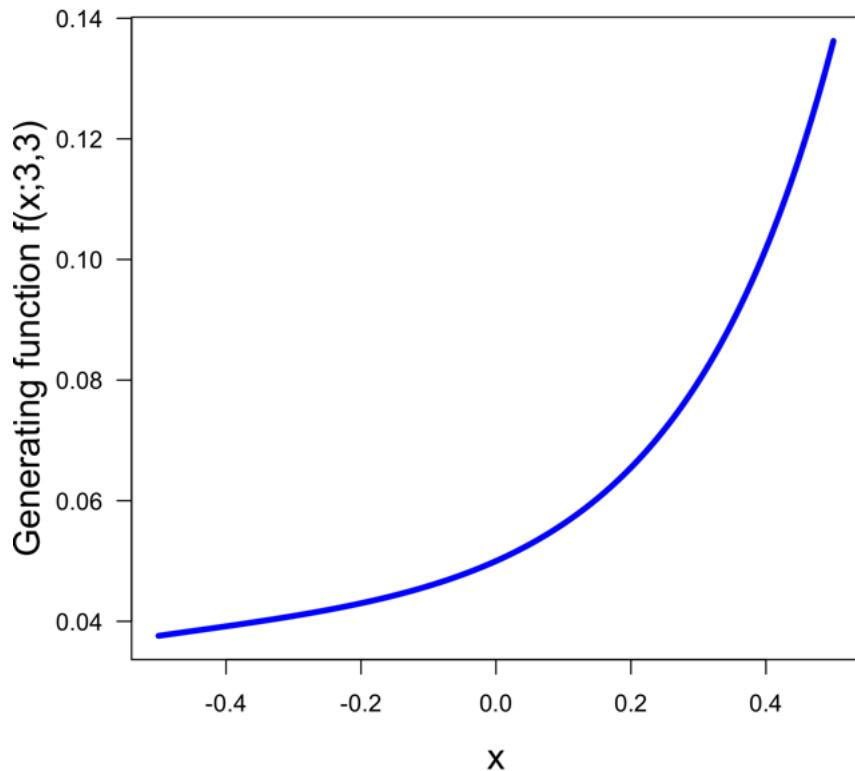


Figure H.28: The probability generating function for the Mann-Whitney statistic  $U$  (Eq. H.206) for  $n = 3$ ,  $m = 3$ . [MannWhitneyGenFct.R](#)

<sup>20</sup>Mathematica code for this purpose can be found in Di Buccianico (1999) or in van de Wiel (2000).

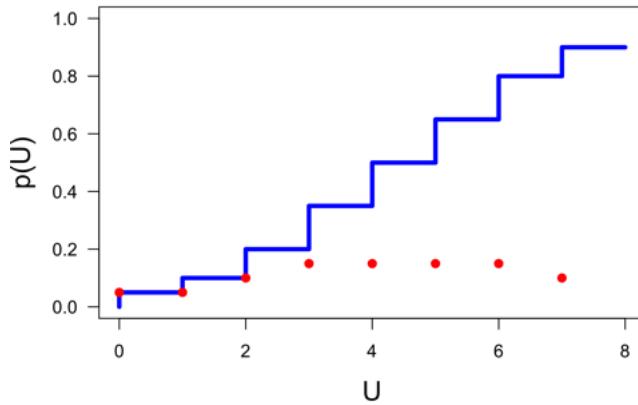


Figure H.29: The probabilities  $p(n = 3, m = 3, U)$  for the Mann-Whitney test statistic  $U$  calculated from the generating function (red dots; up to  $U = 7$  only because of increasing computational time for higher derivatives) and the corresponding CDF (blue line; identical to values listed in Table 1 of Mann & Whitney, 1947). [MannWhitneyGenFct-pU.R](#)

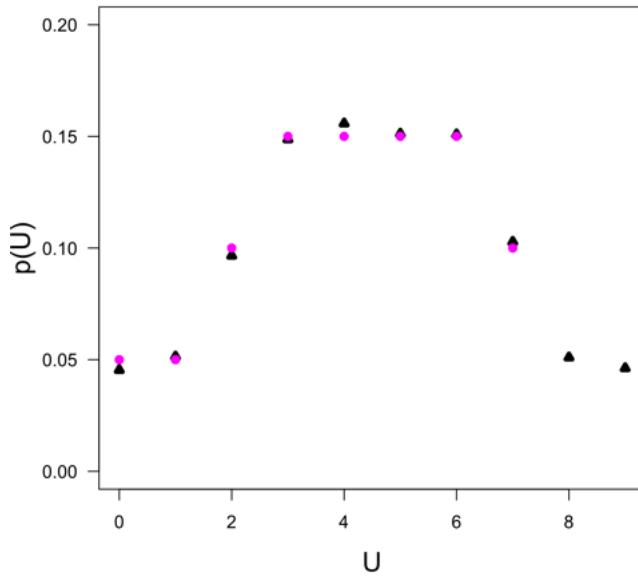


Figure H.30: The probabilities  $p(n = 3, m = 3, U)$  for the Mann-Whitney test statistic  $U$  calculated from the generating function (magenta dots; up to  $U = 7$  only because of increasing computational time for higher derivatives) and estimated by Monte Carlo simulations ( $M = 10^4$  runs, black triangles).

[MannWhitneyGenFct-3pU.R](#)

### H.14.4 Monte Carlo (\*)

Estimates of the probabilities  $p(U)$  by Monte Carlo simulations is straight forward. The ranks for single entries  $xy_a$  of the combined sample  $\mathbf{xy}$  can take on values between 1 and 15. The minimum of the Wilcoxon test statistic  $T$  is

$$j_{\min} = \sum_{j=1}^m j = \frac{m(m+1)}{2} \quad (\text{H.211})$$

and the maximum is

$$j_{\max} = \sum_{j=n+1}^{n+m} j = m \cdot n + \frac{m(m+1)}{2} \quad (\text{H.212})$$

(Exercise 81). Accordingly, the test statistic  $U$  can vary between 0 and  $m \cdot n$ . The Monte Carlo simulation proceeds as follows:

1. Generate many ( $M = 10^6$  or more) random samples of sizes  $n$  and  $m$  from the same population (null hypothesis!) as, for example, from the standard normal distribution.
2. Calculate the test statistic  $U$ .
3. Count the frequency for each  $U$  value.
4. Finally, calculate the relative frequencies (= divide frequencies by  $M$ ) and use these values as estimates for probabilities for single  $U$  values.
5. Estimate the CDF( $U$ ) by successively adding up the probability estimates.

The results of this procedure for  $n = 8$  and  $m = 7$  is shown in Figs. [H.31](#) and [H.32](#).

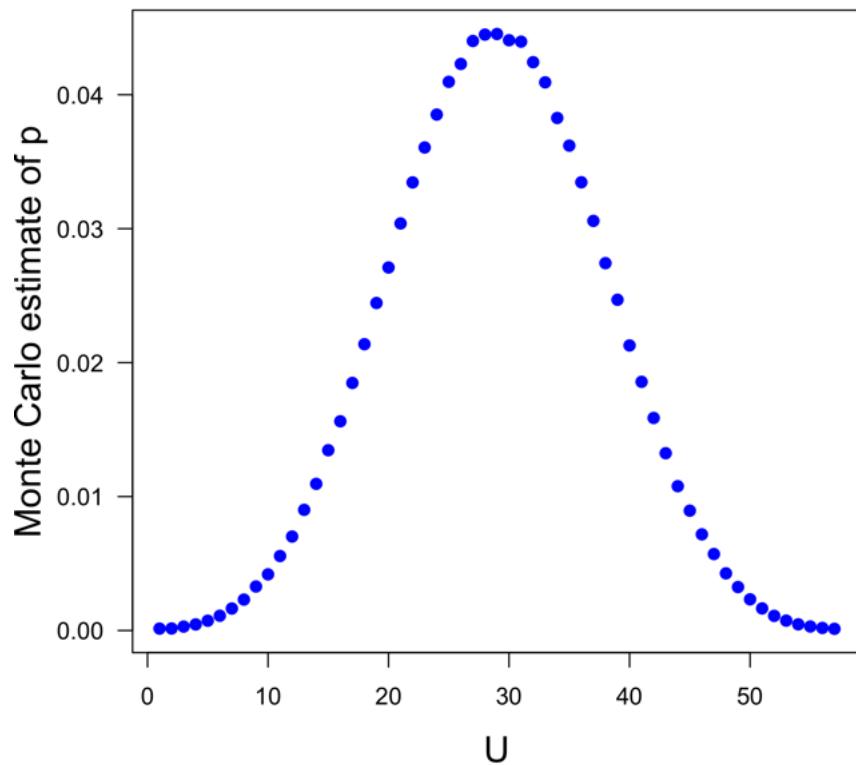


Figure H.31: Monte Carlo estimates ( $M = 10^6$  runs) of probabilities  $p(U)$  for  $n = 8, m = 7$ . The distribution is symmetric (and therefore Mann & Whitney, 1947, list only 'half' of the cumulative distributions in Table I). The imagined (continuous!) envelope of the (discrete!) probability distribution would almost look like a normal PDF. [MannWhitney2n8m7MC.R](#) [WMWpMC.R](#)

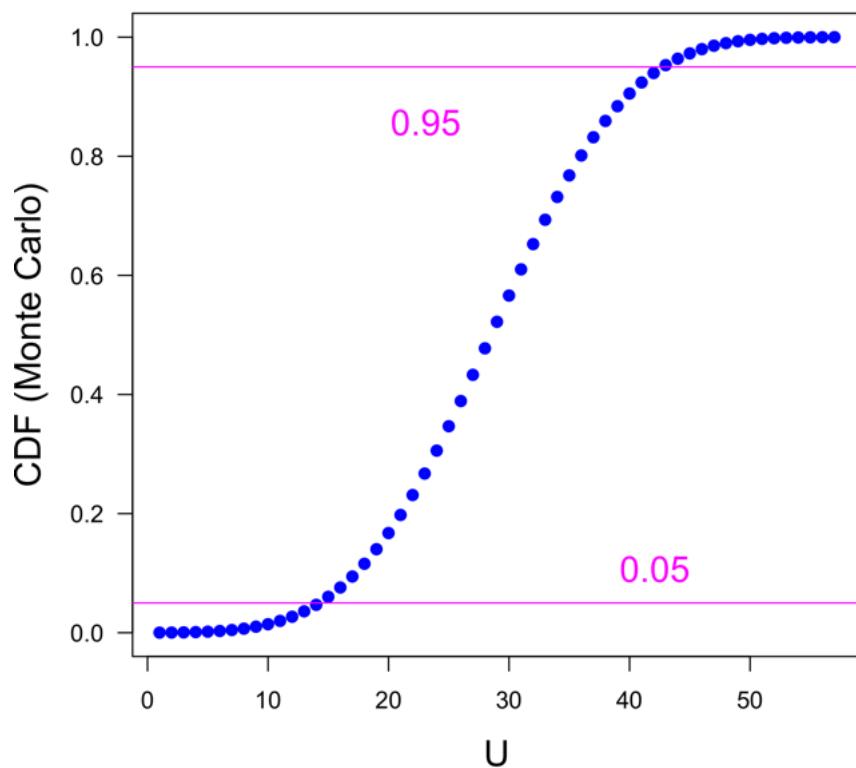


Figure H.32: Monte Carlo estimates ( $M = 10^6$  runs) of cumulative probabilities for  $n = 8, m = 7$ .  
[MannWhitney3n8m7MC.R](#) [WMWpMC.R](#)

**Exercise 81 Prove sum formula (\*)**

Prove (Eq. H.212), i.e.

$$\sum_{j=n+1}^{n+m} j = m \cdot n + \frac{m(m+1)}{2}.$$

### H.14.5 Critical values of the Wilcoxon paired-sample test (\*)

Critical values  $T_{\alpha(1),n}$  and  $T_{\alpha(2),n}$  for the Wilcoxon paired-sample test can be found for  $n$  up to 100 and various  $\alpha$  in Zar (2010, Table B.12; based on McCornack, 1965). An alternative to a look-up table is estimation of the critical values by Monte Carlo simulations (code listed below;  $M = 10^5$  runs). The estimates for  $T_{0.05(2),n}$  for  $n = 7, 8, \dots, 100$  are mostly identical to the values listed in Zar (2010, Table B.12); deviations are small (for example, for  $T_{0.05(2),99}$ : 1911 versus 1913).

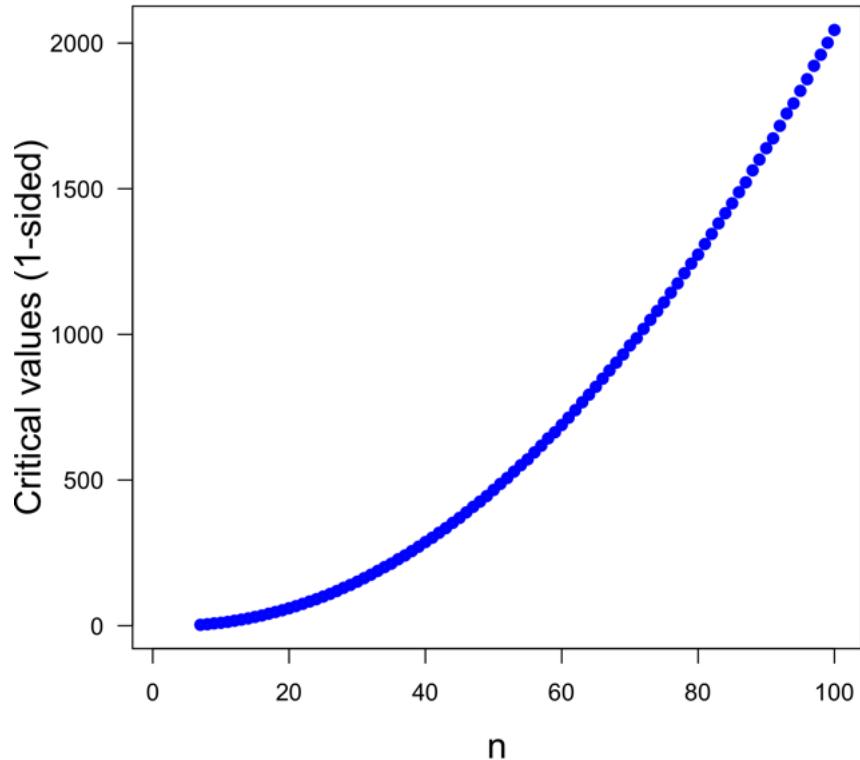


Figure H.33: Critical values  $T_{0.05(1),n}$  for the Wilcoxon paired-sample test estimated by Monte Carlo simulations with  $M = 10^5$  runs. [WilcoxonPairedCritValMC1.R](#)

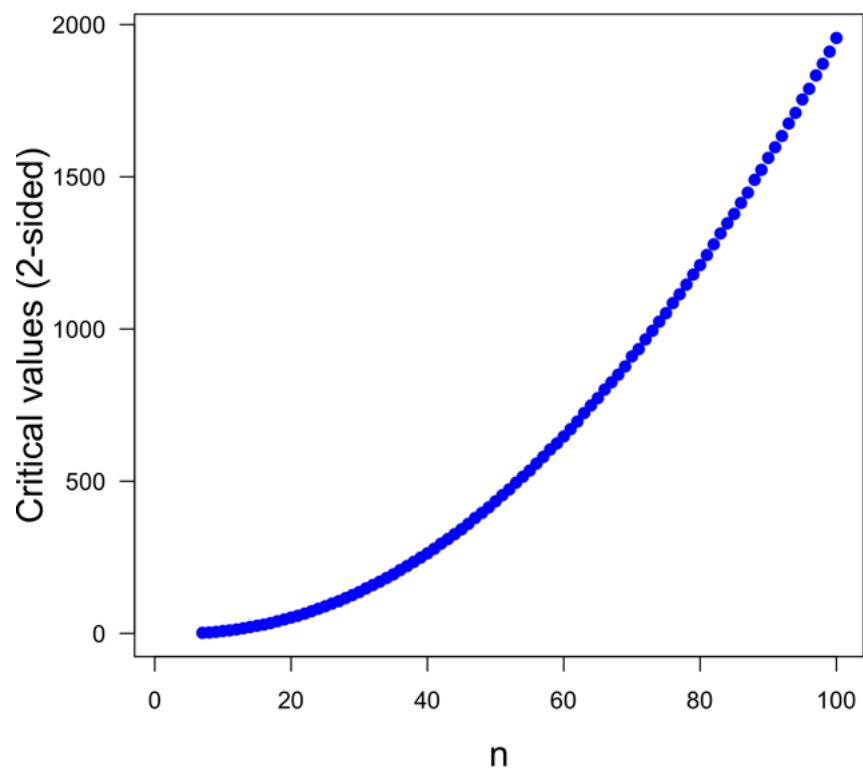


Figure H.34: Critical values  $T_{0.05(2),n}$  for the Wilcoxon paired-sample test estimated by Monte Carlo simulations with  $M = 10^5$  runs. [WilcoxonPairedCritValMC2.R](#)

## H.15 Zar (2010, Example 8.1): the *t*-test procedure

From the data (here: two samples  $X_1, X_2$ ) one calculates the **test statistic *t*** which is defined as the difference in sample mean values divided by an estimate of the standard error of the difference under the null hypothesis.

1. Data (units: min):

$$X_1 = \{8.8, 8.4, 7.9, 8.7, 9.1, 9.6\} \quad \text{for drug B} \quad (\text{H.213})$$

$$X_2 = \{9.9, 9.0, 11.1, 9.6, 8.7, 10.4, 9.5\} \quad \text{for drug G} \quad (\text{H.214})$$

2. Number of data per sample:  $n_1 = 6, n_2 = 7$ .
3. Degrees of freedom ( $\nu_j = n_j - 1, j = 1, 2$ ):  $\nu_1 = 5, \nu_2 = 6; \nu = \nu_1 + \nu_2 = 11$ .
4. Sample mean values (= estimates of the mean values  $\mu_j$  of the corresponding statistical populations):

$$\hat{\mu}_1 = \bar{X}_1 = 8.75 \text{ min} \quad (\text{H.215})$$

$$\hat{\mu}_2 = \bar{X}_2 = 9.74 \text{ min} \quad (\text{H.216})$$

where the little hat above  $\mu$  indicates the 'estimate of the true mean'. The sample mean  $\bar{X}$  is an 'optimal' (unbiased, consistent, efficient) estimator for the population mean.

5. Sum of squares for each sample:

$$SS_1 = \sum_{k=1}^{n_1} (X_{1,k} - \bar{X}_1)^2 = 1.6950 \text{ min}^2 \quad (\text{H.217})$$

$$SS_2 = \sum_{k=1}^{n_2} (X_{2,k} - \bar{X}_2)^2 = 4.0171 \text{ min}^2 \quad (\text{H.218})$$

The sum of squares are proportional to the sample variances.

6. Pooled variance  $s_p^2$  (will be used as best estimate for the variance  $\sigma^2$  of the statistical population in the null hypothesis):

$$s_p^2 = \frac{SS_1 + SS_2}{\nu_1 + \nu_2} = 0.5193 \text{ min}^2 \quad (\text{H.219})$$

Remark: for  $SS_2 = 0$  and  $\nu_2 = 0$ ,  $s_p^2$  reduces to the best estimate of the variance  $\sigma^2$  from a single sample:

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (X_{1,k} - \bar{X}_1)^2 \quad (\text{H.220})$$

7. The variance of the differences of the mean values is estimated by

$$s_{\bar{X}_1 - \bar{X}_2}^2 = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} = 0.16 \text{ min}^2 \quad (\text{H.221})$$

(the variance of the differences of the means is proportional to the variance  $\sigma^2$  of the statistical population and decreases with the number of data because of compensation of positive and negative deviations from the mean).

8. The estimate of the standard error of the difference reads

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = 0.4 \text{ min} \quad (\text{H.222})$$

9. The test statistic  $t$  is given by the differences of the sample means divided by the estimate of the standard error of the difference:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = -2.4765 \quad (\text{H.223})$$

Please note that the **test statistic  $t$  is a dimensionless quantity**.

10. Calculate the observed level of significance (*p*-value) by integration from  $|t| = 2.4765$  to  $+\infty$  over the *t*-distribution (which is the pdf of the test statistic  $t$ ) and double the results because one adds up probabilities from both tails ('two-tail' or 'two-sided' test):

$$p = 2 \int_{|t|}^{+\infty} f(t'; \nu) dt' \quad (\text{H.224})$$

where

$$f(t; \nu) = \frac{1}{\sqrt{\nu \pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} \quad (\text{H.225})$$

is the *t*-distribution (the probability density function for the test statistic  $t$ ; Fig. 6.11).

### Exercise 82 Modified two-sided *t*-test

Use

$$a = \frac{|\bar{x} - \mu|}{\frac{\hat{\sigma}}{\sqrt{n}}} \quad (\text{H.226})$$

instead of

$$t = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \quad (\text{H.227})$$

as test statistic. How is  $a$  distributed when  $H_0$  is true. Generate the equivalent to Fig. 12.5 for the test statistic  $a$  and the data set  $x$  given in Section 1.5.

## H.16 The Lilliefors distribution: estimate CDF by Monte Carlo simulation

'To date, tables for this distribution [Lilliefors distribution = PDF of the test statistic  $D$  of the Lilliefors test] have been computed only by Monte Carlo methods.' (Wikipedia, Lilliefors test, 15.9. 2016)

Tables of  $p$ -values for the Lilliefors test were given by Lilliefors (1967) and van Soest (1967) based on Monte Carlo simulations with  $M = 1000$  runs. Molin & Abdi (1998) gave improved estimates based on  $M = 10^5$  runs and fitted an analytical formula to the results for the  $p$ -values as a function of sample size  $n$  and test statistic  $D$  (here cited by Abdi & Molin, 2007):

$$b_0 = 0.37872256037043 \quad (\text{H.228})$$

$$b_1 = 1.30748185078790 \quad (\text{H.229})$$

$$b_2 = 0.08861783849346 \quad (\text{H.230})$$

$$A = \frac{-(b_1 + n) + \sqrt{(b_1 + n)^2 - 4 b_2 (b_0 - D^2)}}{2 b_2} \quad (\text{H.231})$$

$$\begin{aligned} p(n, D) = & -0.37782822932809 + 1.67819837908004 * A \\ & -3.02959249450445 * A^2 + 2.80015798142101 * A^3 \\ & -1.39874347510845 * A^4 + 0.40466213484419 * A^5 \\ & -0.06353440854207 * A^6 + 0.00287462087623 * A^7 \\ & + 0.00069650013110 * A^8 - 0.00011872227037 * A^9 \\ & + 0.00000575586834 * A^{10} \end{aligned} \quad (\text{H.232})$$

According to Abdi & Molin (2007), this formula give results that are correct for the first 2 digits. However, the formula can be applied only over a limited range of  $D$  values. At  $n = 30$ , for example,  $p(n, D)$  is  $> 1$  for  $D < 0.074$  and negative for  $D > 0.207$  (Fig. H.35).

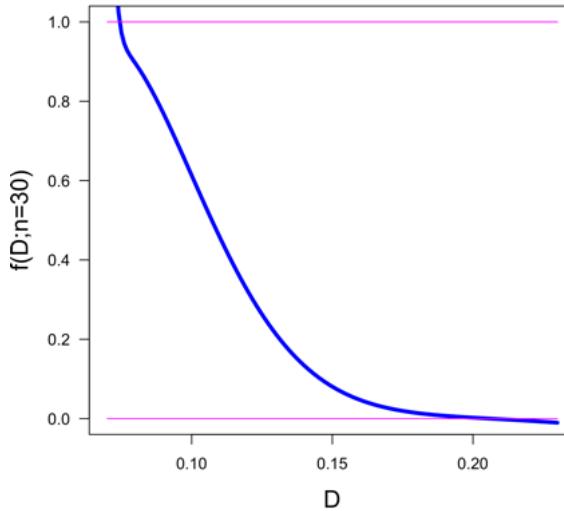


Figure H.35: Lilliefors test: estimated  $p$ -values for sample size  $n = 30$  as a function of test statistic  $D$  (based on analytical formula by Abdi & Molin, 2007). [LillieforsPolynomial.R](#)

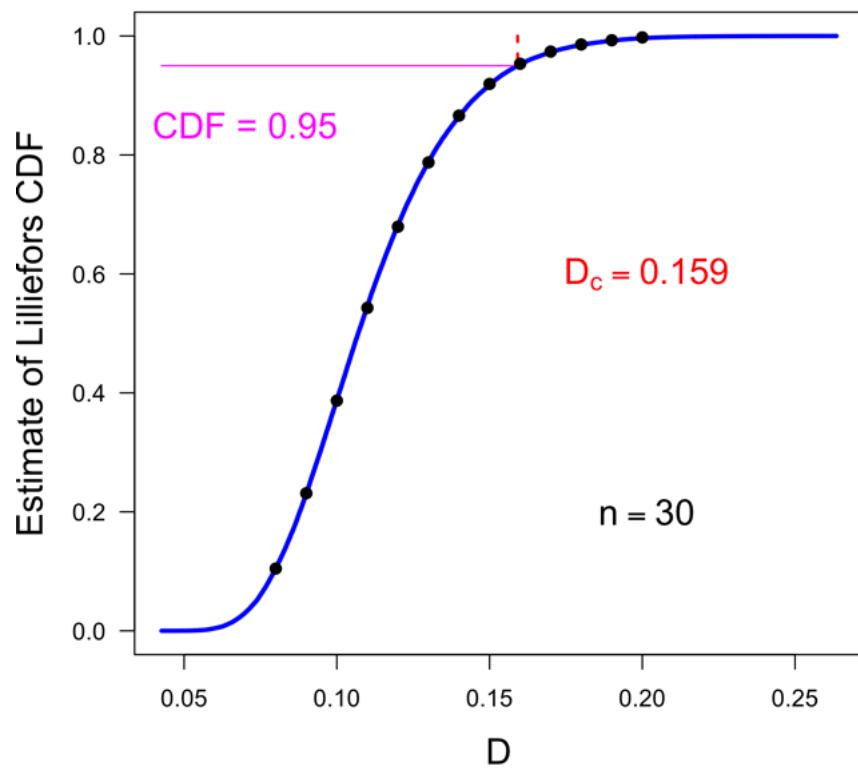


Figure H.36: Estimate p-values and critical value  $D_c$  ( $\alpha = 0.05$ ) for the Lilliefors test: the blue solid line shows the estimate of the CDF for the test statistic  $D$  based on  $M = 10^5$  Monte Carlo simulations, black dots are calculated from the analytical formula given by Abdi & Molin (2007; CDF values =  $1 - p(n, D)$ , Eq. H.232).  
[LillieforsCDF-MC.R](#) [LilliePapprox.R](#)

A few comments about the Monte Carlo code are in order. The Lilliefors test is based on the null hypothesis  $H_0$  = 'standardized random sample is from the standard normal distribution'. The test statistic  $D$  is equal to that of the KS-test, namely the maximal absolute difference between the CDF of the standard normal PDF and the CDF estimated from the sample. As usual in significance tests, the distribution function  $f(n, D)$  ( $n$  is the sample size) of the test statistic  $D$  is calculated just based on  $H_0$  (an alternative hypothesis can be formulated, however, does not play any role in the calculations). In order to estimate  $f(n, D)$  one can sample from any normal population because each sample will be standardized. For the sake of simplicity, one samples from the standard normal distribution. The standardized sample is called  $z$  (a common notation for standardized samples). In order to calculate the test statistic  $D$  for  $z$  one calls the KS-test which uses the same test statistic (this saves some programming and makes the code more compact; this option is probably not the fastest way to calculate  $D$ ). Repeating this procedure many times (here  $M = 10^5$  times) yields many  $D$  values in a certain range between 0 and 1 (very small and very large numbers are rare; for  $n = 30$  the range between  $D = 0.05$  and 0.25 is most populated). Now one sorts the many  $D$  values from small to large: large  $D$  speak against  $H_0$ . One uses the sorted  $D$  values to estimate the cumulative distribution function (CDF): a staircase with height of each step equal to  $1/M = 10^{-5}$  and thus in a plot not recognizable as single steps. For a given level of significance  $\alpha$  (here:  $\alpha = 0.05$ ) one can calculate from the CDF the critical value  $D_c$ , which is given as the location where the CDF is equal to  $(1 - \alpha) = 0.95$  (values  $D \geq D_c$  would lead to rejection of  $H_0$ ). The CDF can also be used to calculate the  $p$ -value for any  $D$ :  $p$  is given by 1 minus the CDF value for  $D$ , i.e. large for small  $D$  and small for large  $D$ , or in short notation  $p = 1 - \text{CDF}(D)$ . Finally, one compares the estimates with the analytical formula fitted by Molin & Abdi (1998) and Abdi & Molin (2007) to their Monte Carlo estimates (they also used  $M = 10^5$ ): the values agree very well over the most interesting range ( $0.08 < D < 0.2$  for  $n = 30$ ; problems with negative values or values  $> 1$  outside this range have been discussed above).

## H.17 Hypotheses testing: history

History of hypotheses testing largely based on Lehmann (1993):

1. "Throughout the 19th century, testing was carried out rather informally. It was roughly equivalent to calculating an (approximate)  $p$  value and rejecting the hypothesis if this value appeared to be sufficiently small. These early approximate methods required only a table of the normal distribution." (Lehmann, 1993)
2. 1875 Friedrich Robert Helmert:  $\chi^2$ -distribution
3. 1900 Karl Pearson introduced the chi-squared test<sup>21</sup>
4. 1908 Student: small sample test of the mean,  $t$ -distribution,  $t$ -test
5. Significance testing: Fisher (several papers before 1925) and Fisher (1925, Statistical Methods for Research Workers);  $\alpha = 0.05$ <sup>22</sup>
6. 1928 Neyman and Pearson asked for the origin of test statistics: 'Why these rather than some others?' (Lehmann, 1993); alternative hypothesis; false rejection (type I error, Error I), false acceptance (type II error, Error II);
7. Neyman & Pearson (1933): power of the test (the rejection probability as a function of the alternative).
8. Jeffreys (1939, 1948, 1961) Bayesian tests with prior, likelihood, posterior, Bayes' factor
9. 1951 Fisher in a letter to W.E. Hick (cited in Lehmann, 1993): "I am a little sorry that you have been worrying yourself at all with that unnecessarily portentous approach to tests of significance represented by the Neyman and Pearson critical regions, etc. In fact, I and my pupils throughout the world would never think of using them. . . ."
10. 1955 Fisher argues against type II errors.
11. 1994 "Berger, Brown and Wolpert (1994) approached the issue of choice of the conditioning statistic from the perspective of seeking a unification between conditional frequentist testing and Bayesian testing . . ." Berger (2003)
12. 2003 Berger: Could Fisher, Jeffreys and Neyman have agreed? "The program of developing conditional frequentist tests for the myriad of testing scenarios that are considered in practice today will involve collaboration of frequentists and objective Bayesians. This is because the most direct route to determination of a suitable conditional frequentist test, in a given scenario, is the Bayesian route, thus first requiring determination of a suitable objective Bayesian procedure for the scenario."
13. Jaynes (2003, Probability – The Logic of Science) Bayesian tests with prior, likelihood, posterior, Bayes' factor

"Physicists mostly have not heard of  $p$ -values; they tend to talk in terms of the number of  $\sigma$  (standard errors) from the null.

Those who have heard of them mostly interpret them incorrectly.

Those who understand  $p$ -values know their use is difficult . . ."

Berger (2011, talk)

---

<sup>21</sup>Compare Plackett (1983) for a discussion of Pearson (1900) and its consequences.

<sup>22</sup>First occurrence in literature; compare, however, Cowles and Davis, 1982.



## Appendix I

# Beyond hypothesis testing: Bayesian inference

### Observations, likelihood, priors, posterior for original vaquita data

The following input parameters are specified: area  $A = 2276.3 \text{ km}^2$ , effort (total length of transects) in 1997  $L97 = 514.3 \text{ km}$ , effort in 2008  $L08 = 872.3 \text{ km}$ , mean group size  $S = 1.86^1$ , i.e. observation of pairs seems to be quite common, truncation distance (maximum perpendicular distance over which observation efforts are made) in 1997  $W97 = 3 \text{ km}$  and in 2008  $W08 = 4 \text{ km}$ .

The 'core' data are the perpendicular distances (km) in 1997 and 2008, respectively:  
 $pd97 = \{0.58, 0.38, 0.3, 0.62, \dots, 2.21, 0.86, 1.53\}$ , sample size  $n97 = 88$   
 $pd08 = \{0.42, 1.11, 3.6, 0.53, \dots, 0.51, 1.11, 2.75\}$ , sample size  $n97 = 88$  (for the complete list see the R code below).

The likelihood for the original vaquita data is a function of 5 variables

$$L(D97, esw97, g_0, esw08, D.\delta | \text{data}) \quad (\text{I.1})$$

where  $D97$  (individuals  $\text{km}^{-2}$ ) is the abundance density of vaquita in 1997,  $esw97$  and  $esw08$  are the effective strip half-width in 1997 and 2008, respectively,  $g_0$  is the detection probability on the transect, and  $D.\delta$  is the change of abundance density of vaquita (negative values indicate decrease in abundance). The ranges, likelihoods, and priors for the relevant variables are explained below. Under the assumption of independence the likelihoods and priors for various variables can be multiplied.

---

<sup>1</sup>The group size is not taken into account in the R code provided by Gerrodette (2011) and it will be also ignored in the following discussion.

### Likelihood and prior for the detection probability on the transect

The detection probability on the transect,  $g_0$ , can vary between 0.1 and 1. It follows a beta distribution with parameters  $\alpha = 4$  and  $\beta = 3$  (Fig. I.1)

$$L_{g_0} = \mathcal{B}(g_0; \alpha = 4, \beta = 3) \quad (\text{I.2})$$

with maximum at  $g_0 = 0.6$ . The prior for  $g_0$  follows the identical distribution.

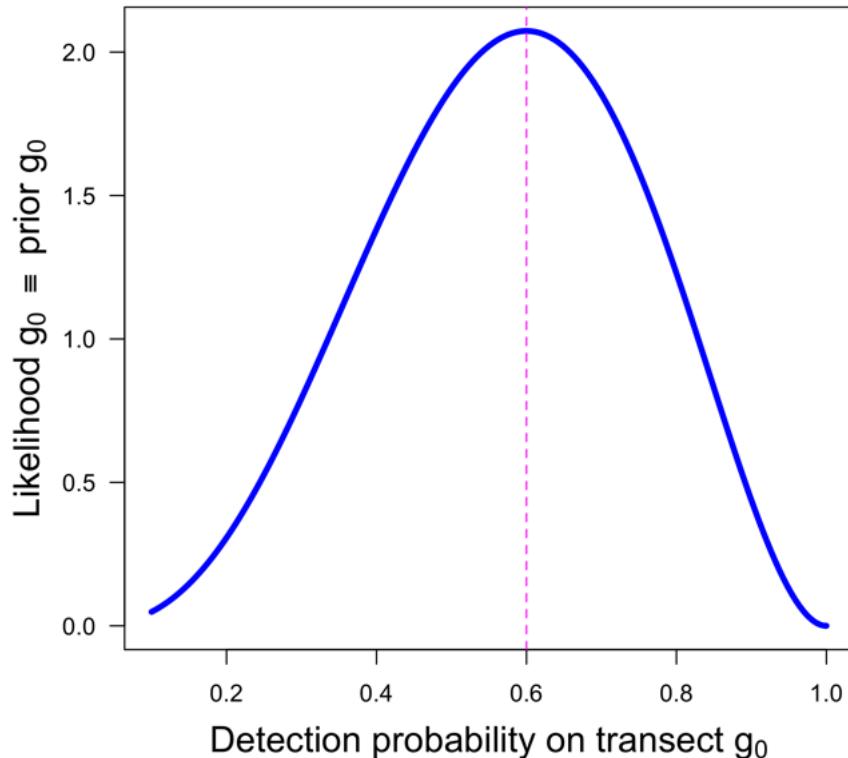


Figure I.1: The likelihood and prior for  $g_0$  are identical and given by the beta PDF with parameters  $\alpha = 4$  and  $\beta = 3$ : note that for this choice of parameters the PDF is slightly asymmetric around the maximum at  $g_0 = 0.6$  or 60% detection probability.

### The density in 1997, $D97$ , and its prior distribution

The abundance density of vaquitas in 1997,  $D97$ , can vary over a wide range from 0.01 to 0.45 individuals  $\text{km}^{-2}$ . This density range is used to estimate the number of vaquitas in the area,  $N_a$ , which is requested as parameter in the binomial likelihood distribution for the number of vaquita detections in 1997,  $n_{97}$  (see below). A uniform ('flat') prior is assigned to  $D97$  (Fig. I.2) because no knowledge is currently available that would suggest the choice of an informative prior.

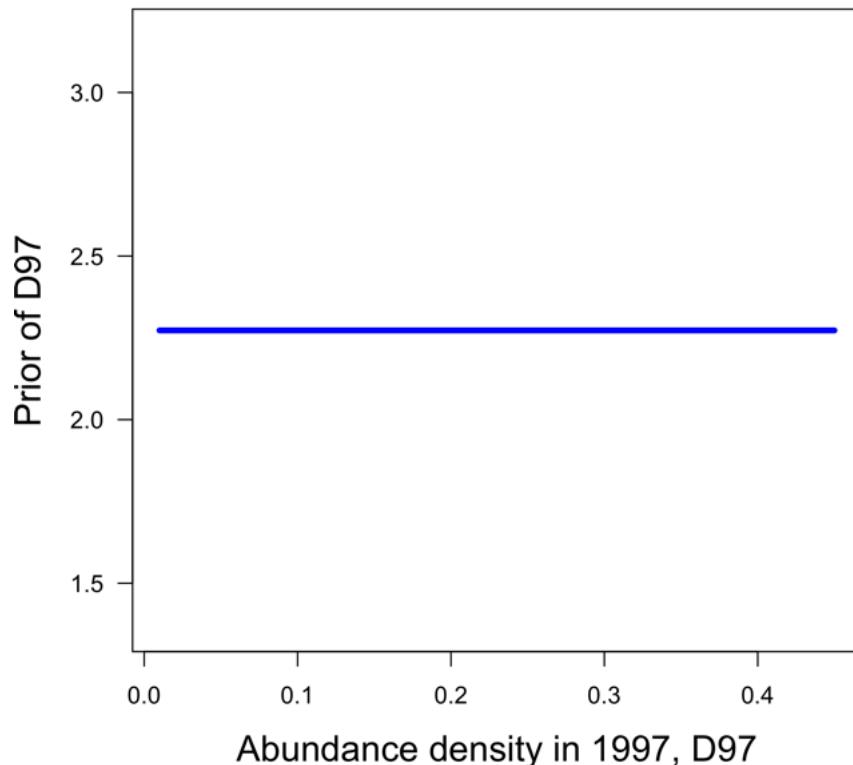


Figure I.2: A uniform ('flat') prior is assigned to  $D97$  because no knowledge is currently available that would suggest the choice of an informative prior.

### Range and prior for the effective strip half-width in 1997, $esw97$

The effective strip half-width in 1997,  $esw97$ , can vary between 1.1 and 2.0 km. It is requested as parameter in the likelihoods for  $n97$  and  $pd97$  (see below). A normal prior is assigned to  $esw97$  with mean  $\mu = 1.9$  and standard deviation  $\sigma = 2.0$  and thus the prior varies only slightly (almost a flat prior) between 0.1841 and 0.1995 (maximum at  $\mu = 1.9$ , Fig. I.3).

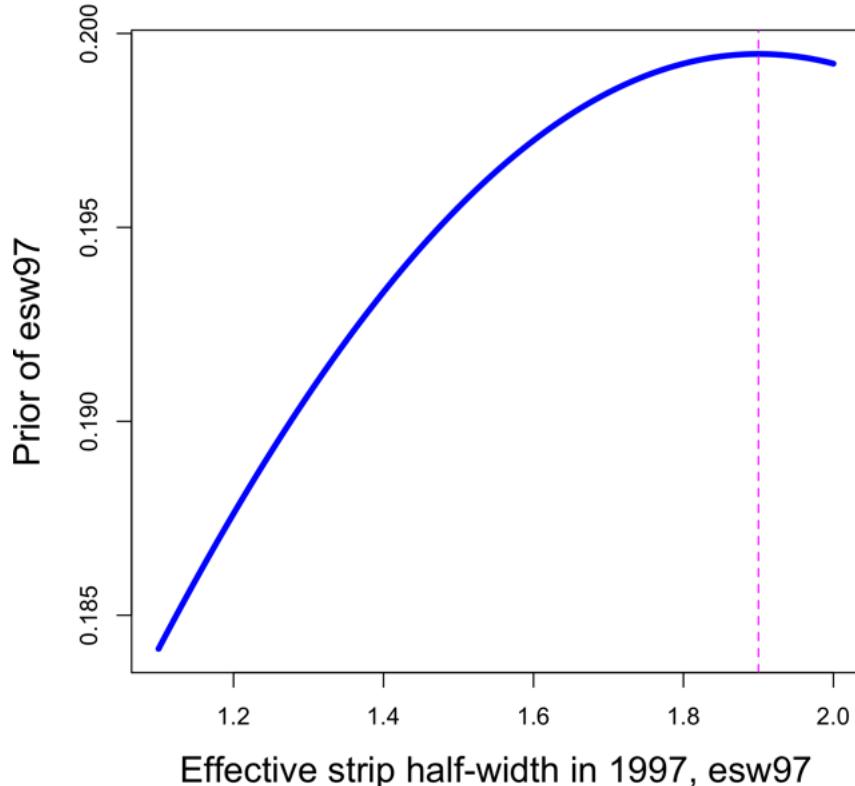


Figure I.3: A normal prior is assigned to  $esw97$  with mean  $\mu = 1.9$  and standard deviation  $\sigma = 2.0$  (blue solid line) and thus the prior varies only slightly (almost a flat prior) between 0.1841 and 0.1995 (maximum at  $\mu = 1.9$ , indicated by vertical green dashed line).

### Range and prior for the effective strip half-width in 2008, $esw08$

The effective strip half-width in 2008,  $esw08$ , can vary between 1.1 and 2.0 km. It is requested as parameter in the likelihoods for  $n08$  and  $pd08$  (see below). A normal prior is assigned to  $esw08$  with mean  $\mu = 1.9$  and standard deviation  $\sigma = 1.0$  and thus the prior varies between 0.2897 and 0.3989 (maximum at  $\mu = 1.9$ , Fig. I.3).

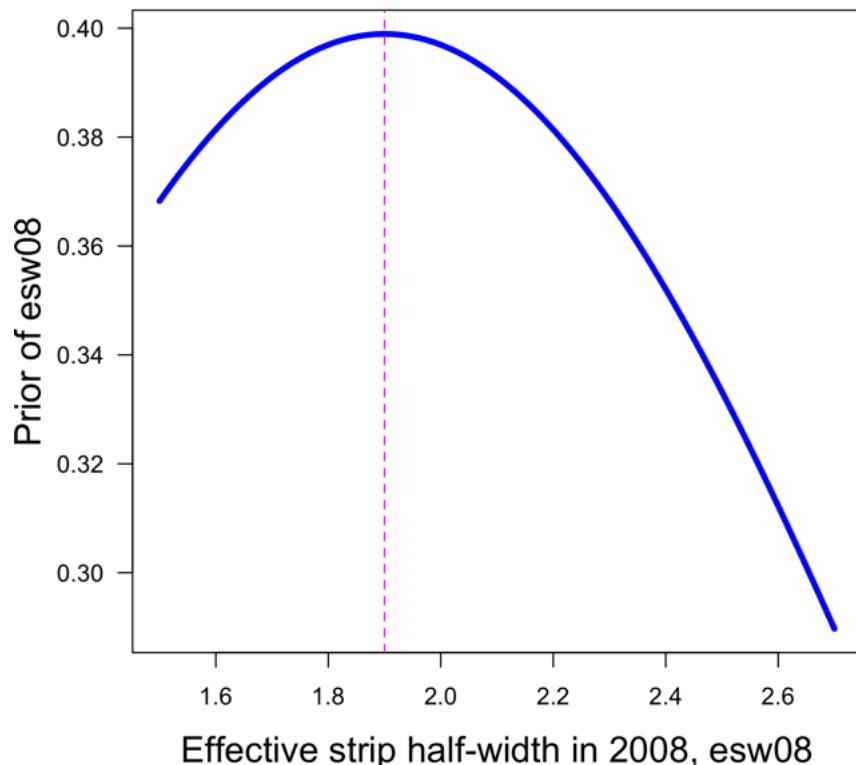


Figure I.4: A normal prior is assigned to  $esw08$  with mean  $\mu = 1.9$  and standard deviation  $\sigma = 1.0$  (blue solid line) and thus the prior varies between 0.2897 and 0.3989 (maximum at  $\mu = 1.9$ , indicated by vertical green dashed line).

**Detection probability,  $P_a97$** 

The 'detection probability' is usually estimated by the effective strip half-width,  $esw$ , divided by the truncation length, i.e.  $P_a = esw/W$ . The R code provided by Gerrodette (2011) encompasses another factor that is neither explained in Eguchi & Gerrodette (2009) nor in Gerrodette (2011):

$$P_a = \frac{esw}{W} \cdot 2 \cdot \left( \Phi \left( \sqrt{\frac{\pi}{2}} \cdot \frac{W}{esw} \right) - 0.5 \right) \quad (\text{I.3})$$

where  $\Phi()$  is the CDF of the standard normal PDF. Both factors in the argument of  $\Phi()$  are  $> 1$  (the truncation length,  $W$ , is always  $\geq$  the effective strip half-width;  $\sqrt{\frac{\pi}{2}} = 1.25$ ) and thus  $1 \geq \Phi \left( \sqrt{\frac{\pi}{2}} \cdot \frac{W}{esw} \right) > \Phi \left( \sqrt{\frac{\pi}{2}} \right) = 0.89$  and  $1 \geq 2 \left( \Phi \left( \sqrt{\frac{\pi}{2}} \cdot \frac{W}{esw} \right) - 0.5 \right) > 0.79$ , i.e. this factor can reduce the value of  $P_a$  by not more than 21%. For the actual truncation lengths and effective strip half-widths in 1997 and 2008 the reduction is not larger than 6.5%.

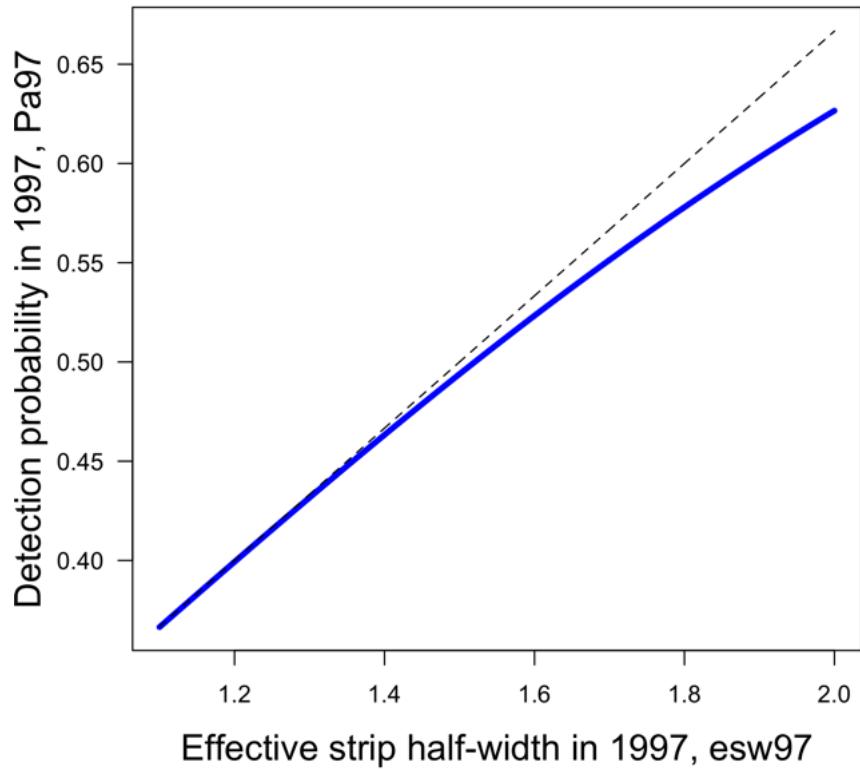


Figure I.5: The 'detection probability' in 1997,  $P_a97$  as a function of the effective strip half-width (blue solid line). The approximation by the ratio of effective strip half-width divided by the truncation length (dashed black line) deviates by not larger than 6.5% from detection probability used by Gerrodette (2011).

### The number of observations in 1997, $n_{97}$

Each sighting is modeled as a Bernoulli trial (success or failure) with probability of success in a single trial (detection probability),  $P_a$ . Under the assumption of independence of the observations, the probability for sighting  $n_{97}$  vaquitas is given by the binomial distribution

$$\mathcal{B}(n_{97}; N_a, P_a) = \binom{N_a}{n_{97}} P_a^{n_{97}} (1 - P_a)^{N_a - n_{97}} \quad (\text{I.4})$$

where  $N_a$  is the number of vaquitas in the area.

$$N_a(D97, g_0) = 2 [D97 \cdot W97 \cdot L97 \cdot g_0] \quad (\text{I.5})$$

is two times the rounded (the square brackets indicate rounding) product of the abundance density,  $D97$ , the truncation length,  $W97$ , the effort (total length of transects, km),  $L97$ , and the detection probability on the transect,  $g_0$  (Fig. I.6). Please note that the binomial distribution for  $n_{97}$  depends on 3 independent variables ( $D97, g_0, esw97$ ).

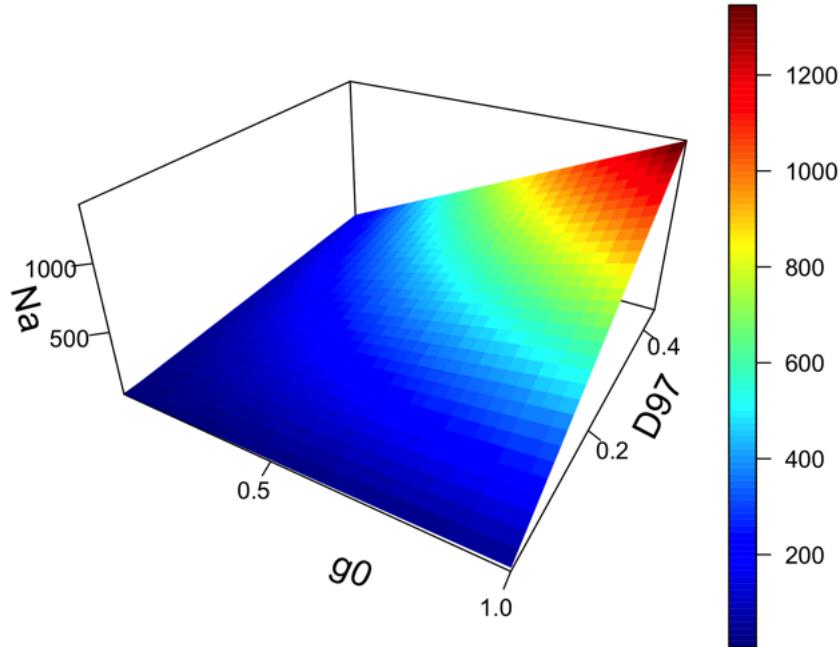


Figure I.6: Number of vaquitas in the area,  $N_a$ , depending on the abundance density,  $D97$ , and the detection probability on the transect,  $g_0$ .

### Likelihood for the perpendicular distances

The likelihoods for the perpendicular distances<sup>2</sup> in 1997,  $pd97 \geq 0$ , are half-normal PDFs (Subsection C.3.1) with  $\mu = 0$  ( $\mu = 0$  corresponds to the point on the transect from which the perpendicular distance is measured; please note that  $\mu$  is a parameter characterizing the half-normal PDF and it is *not* the true mean of this distribution) and  $\sigma_j = esw97_j \sqrt{2}/\sqrt{\pi}$ ,  $j = 1, 2, \dots, J$ . The true mean,  $\mu_{HN}$ , of the half-normal distribution is  $\mu_{HN} = \sigma\sqrt{2}/\sqrt{\pi}$ . Please note that Gerrodette (2011) left out a multiplicative factor of 2 in the calculation of the likelihood (in the R code `dnorm(pd97,0,esw97[j]*sqrt(2/pi))` instead of `2 dnorm(pd97,0,esw97[j]*sqrt(2/pi))`); however, this causes no problem because multiplicative factors do not shift the position of the maximum of the likelihood or the posterior and they cancel out when normalizing the posterior. The likelihoods for the perpendicular distances in 2008,  $pd08 \geq 0$ , are calculated in the same manner as for 1997, i.e. half-normal PDFs with  $\mu = 0$  and  $\sigma_j = esw08_j \sqrt{2}/\sqrt{\pi}$ ,  $j = 1, 2, \dots, J$ .

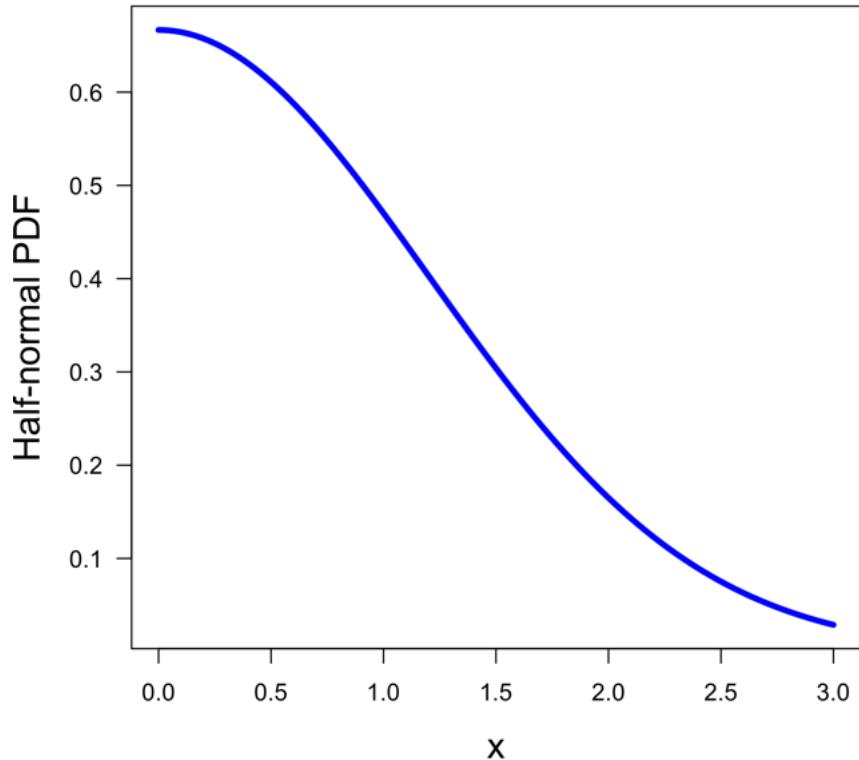


Figure I.7: The half-normal PDF for  $\mu = 0$  and  $\sigma = 1.1968$  (corresponding to  $esw97 = 1.5$ ). The true mean of the half-normal distribution is  $\mu_{HN} = \sigma\sqrt{2}/\sqrt{\pi} = 0.9549$ .

<sup>2</sup>The perpendicular distances are the ‘core data’ in animal surveys.

### The number of observations in 2008, $n_{08}$

The likelihood for the observations in 2008,  $n_{08}$ , is calculated similarly to the one for 1997, however, taking into account changes in abundance density. Each sighting is modeled as a Bernoulli trial (success or failure) with probability of success in a single trial (detection probability),  $P_a$ . Under the assumption of independence of the observations, the probability for sighting  $n_{08}$  (groups of) vaquitas is given by the binomial distribution

$$\mathcal{B}(n_{08}; N_a, P_a) = \binom{N_a}{n_{08}} P_a^{n_{08}} (1 - P_a)^{N_a - n_{08}} \quad (\text{I.6})$$

where the number of vaquitas in the area

$$N_a(D97_i + D.\delta\_{delta}_j, g_0) = \begin{cases} 2 [(D97_i + D.\delta\_{delta}_j) \cdot W08 \cdot L08 \cdot g_0] & \text{if } D97_i + D.\delta\_{delta}_j > 0 \\ 0 & \text{otherwise} \end{cases} \quad (\text{I.7})$$

is two times the rounded (the square brackets indicate rounding) product of the abundance density in 2008,  $D97_i + D.\delta\_{delta}_j$  if  $>$  0 otherwise, the truncation length,  $W08$ , the effort (total length of transects, km),  $L08$ , and the detection probability on the transect,  $g_0$  (Fig. I.8). Please note that the binomial distribution for  $n_{08}$  depends on 4 independent variables ( $D97$ ,  $D.\delta\_{delta}$ ,  $g_0$ ,  $esw97$ ) and thus visualization is not done here.

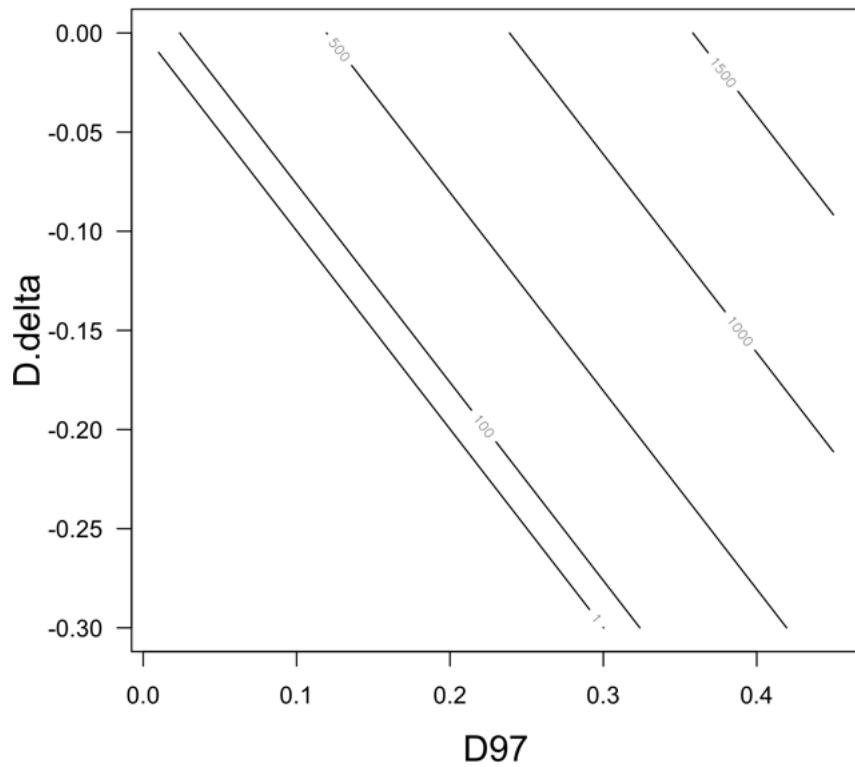


Figure I.8: Number of vaquitas in the area,  $N_a$ , depending on the abundance density in 1997,  $D97$  and the change in abundance density,  $D.\delta\_{delta}$ , at detection probability on the transect  $g_0 = 0.6$ .



# Appendix J

## Straight line fitting

### J.1 Simple linear regression: Bayesian approach

The Bayesian approach to simple linear regression is discussed in full detail in Zeller (1971, p. 58ff). Some of his results are given here and illustrated using artificial data.

Data<sup>1</sup> (Fig. J.1):  $(x_i, y_i), i = 1, 2, \dots, n$ .

Model:<sup>2</sup>

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad (\text{J.1})$$

where  $\beta_1$  is the intercept,  $\beta_2$  is the slope, and noise  $u_i \sim \mathcal{N}(0, \sigma^2)$ .

Likelihood:

$$p(\mathbf{y}|\mathbf{x}, \beta_1, \beta_2, \sigma) \propto \frac{1}{\sigma^n} \exp \left[ -\frac{1}{2\sigma^2} \sum (y_i - \beta_1 - \beta_2 x_i)^2 \right], \quad (\text{J.2})$$

Diffuse prior<sup>3</sup>

$$p(\beta_1, \beta_2, \sigma) \propto \frac{1}{\sigma} \quad -\infty < \beta_1, \beta_2 < \infty, \quad 0 < \sigma < \infty. \quad (\text{J.3})$$

Joint posterior for the unknown parameters:

$$\begin{aligned} p(\beta_1, \beta_2, \sigma | \mathbf{y}, \mathbf{x}) &\propto \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{1}{2\sigma^2} \sum (y_i - \beta_1 - \beta_2 x_i)^2 \right] \\ &\propto \frac{1}{\sigma^{n+1}} \times \exp \left\{ -\frac{1}{2\sigma^2} \left[ \nu s^2 + n (\beta_1 - \hat{\beta}_1)^2 \right. \right. \\ &\quad \left. \left. + (\beta_2 - \hat{\beta}_2)^2 \sum x_i^2 + 2 (\beta_1 - \hat{\beta}_1) (\beta_2 - \hat{\beta}_2) \sum x_i \right] \right\}. \end{aligned} \quad (\text{J.4})$$

where<sup>4</sup>  $\nu = n - 2$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad (\text{J.5})$$

$$s^2 = \nu^{-1} \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \quad (\text{J.6})$$

---

<sup>1</sup>The artificial data were generated similar to the procedure used in Section 14.1, however, with a much smaller sample size ( $n = 10$  instead of 170) in order to eventually recognize differences between the marginal posterior distributions (for intercept, slope, and noise level) from normal distributions.

<sup>2</sup>Note that in section the notation follows Zellner (1971), i.e. the intercept, for example, is denoted by  $\beta_1$  (in contrast to other sections in these lecture notes where  $\beta_0$ ).

<sup>3</sup>Assumption:  $\beta_1, \beta_2$ , and  $\log \sigma$  are uniformly and independently distributed.

<sup>4</sup>Degrees of freedom  $\nu$  = number of data - 2 constraints (= estimates for  $\beta_1$  and  $\beta_2$ ) =  $10 - 2 = 8$ .

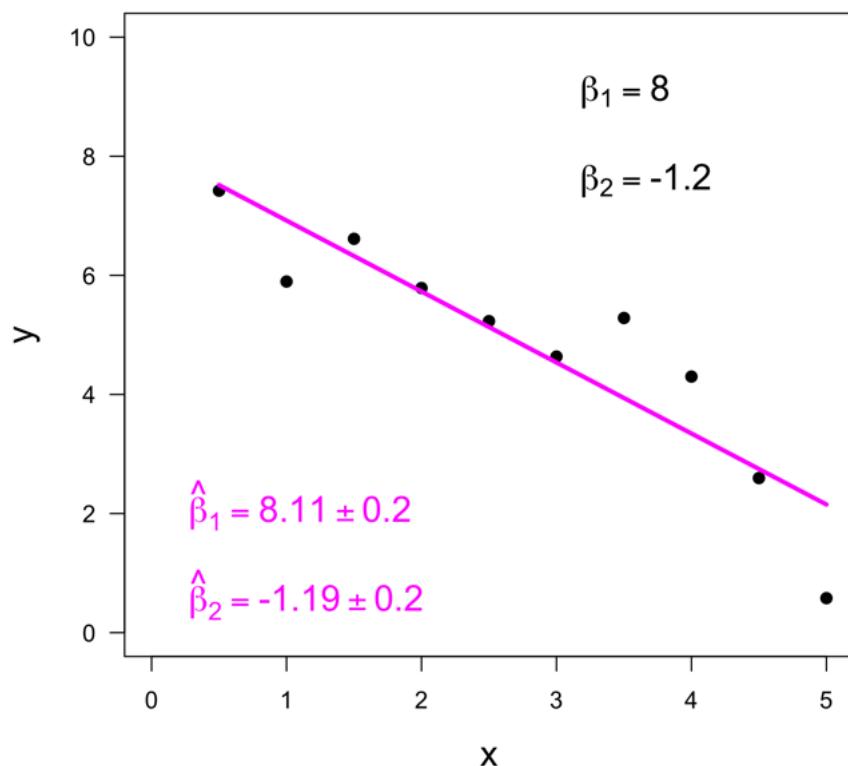


Figure J.1: Data (blue dots; sample size  $n = 10$ ), regression line (magenta), true ( $\beta_1, \beta_2$ ) and estimated ( $\hat{\beta}_1, \hat{\beta}_2$ ) intercept and slope. R code: [SLRBayesian.R](#) for sflag = 1

with  $\bar{y} = n^{-1} \sum y_i$  and  $\bar{x} = n^{-1} \sum x_i$ .

Note that the diffuse prior (Eq. J.3) has no impact on the optimal maximum likelihood estimates for the intercept, the slope, and the variance of noise (Eqs. J.5 – J.6).

The marginal<sup>5</sup> posterior distributions for the intercept and slope read (Figs. J.2 – J.3):

$$p(\beta_1 | \mathbf{y}, \mathbf{x}) \propto \left[ \nu + \frac{\sum (x_i - \bar{x})^2}{s^2 \sum x_i^2 / n} (\beta_1 - \hat{\beta}_1)^2 \right]^{-(\nu+1)/2}, \quad -\infty < \beta_1 < \infty, \quad (\text{J.7})$$

$$p(\beta_2 | \mathbf{y}, \mathbf{x}) \propto \left[ \nu + \frac{\sum (x_i - \bar{x})^2}{s^2} (\beta_2 - \hat{\beta}_2)^2 \right]^{-(\nu+1)/2}, \quad -\infty < \beta_2 < \infty. \quad (\text{J.8})$$

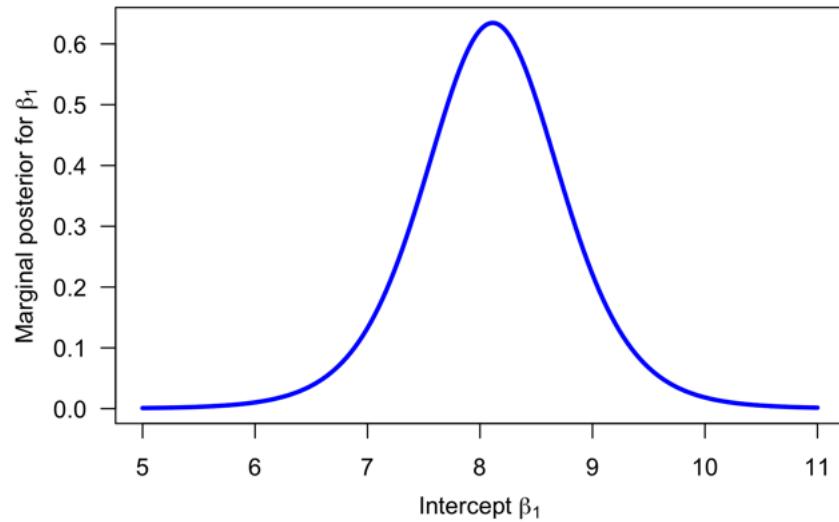


Figure J.2: Marginal posterior for the intercept  $\beta_1$ ,  $p(\beta_1 | \mathbf{y}, \mathbf{x})$ , (Zellner, 1971, Eq. 3.11) normalized by numerical integration. R code: [SLRBayesian.R](#) for sflag = 2

It can be shown further that the transformed quantities

$$t = \left[ \frac{\sum (x_i - \bar{x})^2}{s^2 \sum x_i^2 / n} \right]^{1/2} (\beta_1 - \hat{\beta}_1) \quad (\text{J.9})$$

and

$$t = \frac{\beta_2 - \hat{\beta}_2}{s / \left[ \sum (x_i - \bar{x})^2 \right]^{1/2}} \quad (\text{J.10})$$

both follow  $t$  distributions with  $\nu$  degrees of freedom. This allows us to make inferences about  $\beta_1$  and  $\beta_2$  using properties of the  $t_\nu$  distribution.

The marginal posterior of  $\sigma$  follows an inverted gamma distribution (Zellner, 1971, Eq. (A.37b), Fig. J.4):

$$p(\sigma | \nu, s) = \frac{2}{\Gamma(\nu/2)^{\nu/2}} \left( \frac{\nu s^2}{2} \right)^{\nu/2} \frac{1}{\sigma^{\nu+1}} e^{-\nu s^2 / (2\sigma^2)}, \quad 0 < \sigma < \infty, \quad (\text{J.11})$$

<sup>5</sup>Marginal = here, the dependencies on all except for a single parameter of interest have been integrated out.

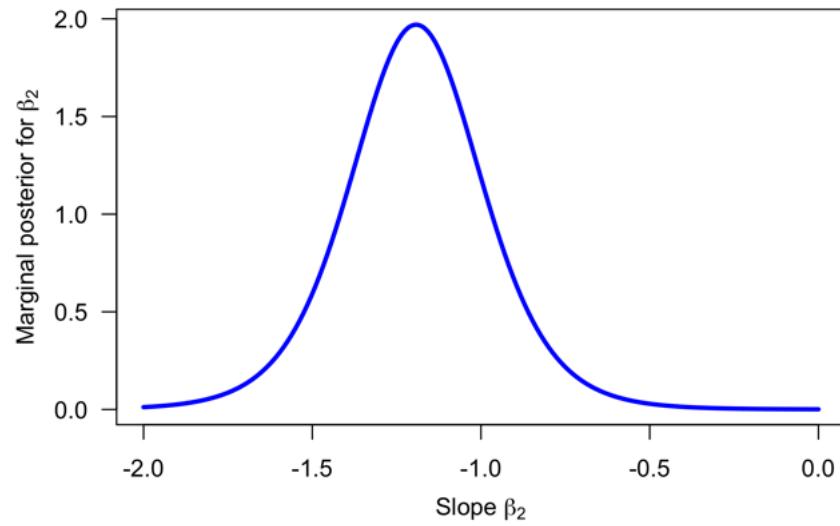


Figure J.3: Marginal posterior for the slope  $\beta_2$ ,  $p(\beta_2|y, x)$ , (Zellner, 1971, Eq. 3.12) normalized by numerical integration. R code: [SLRBayesian.R](#) for sflag = 3

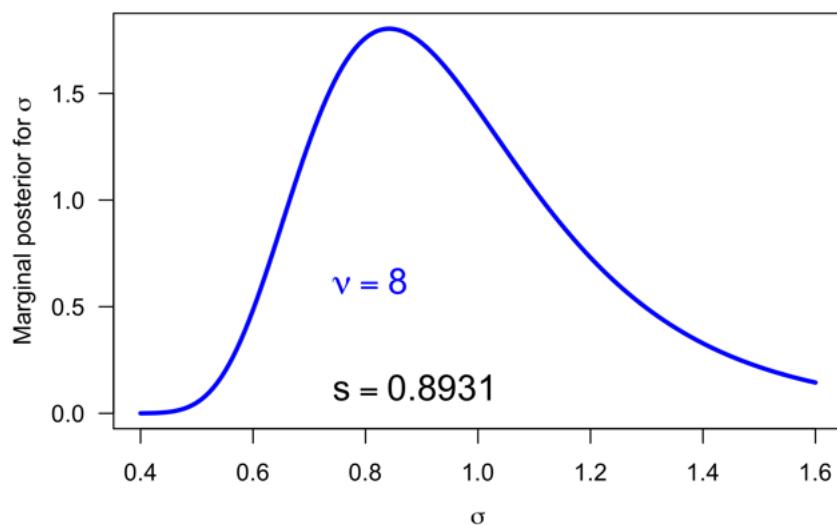


Figure J.4: Marginal posterior for  $\sigma$ ,  $p(\sigma|\nu,s)$ , (Zellner, 1971, Eq. A.37b) for  $\nu = 8$  and  $s = 0.8931$ . R code: [SLRBayesian.R](#) for `sflag = 4`

**Exercise 83 Normal approximations SLR marginal posterior PDFs**

- (1) Calculate means and standard deviations for the marginal posterior distributions of intercept (Eq. J.7), slope (Eq. J.8), and  $\sigma$  (Eq. J.11).
- (2) Add the normal approximations to the plots given above (Figs. J.2, J.3, J.4) and discuss obvious differences between distributions and normal approximations.

## J.2 Proof of the Gauss-Markov theorem (\*)

The following proof is based on Wikipedia (Gauss-Markov theorem, assessed 4 November 2020). We consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (\text{J.12})$$

with  $\boldsymbol{\epsilon} \sim (\mu = 0, \sigma^2)$ . The [ordinary least squares \(OLS\) estimator](#) for this model (Eq. J.12) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (\text{J.13})$$

where  $\mathbf{X}'$  is the transpose of  $\mathbf{X}$ ; it minimizes the sum of squares of residuals

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left( y_i - \sum_{j=1}^K \hat{\beta}_j X_{ij} \right)^2. \quad (\text{J.14})$$

Let

$$\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y} \quad (\text{J.15})$$

be a linear estimator of  $\boldsymbol{\beta}$  with

$$\mathbf{C} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\text{OLS}} + \mathbf{D} \quad (\text{J.16})$$

where  $\mathbf{D}$  is a  $K \times n$  ( $K$  = number of slopes + 1 for the intercept;  $n$  = sample size) non-zero matrix. As we are restricting to [unbiased](#) estimators, i.e.  $E[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ , minimum mean squared error implies minimum variance.<sup>6</sup>

The goal is therefore to show that such an estimator has a variance no smaller than that of  $\hat{\boldsymbol{\beta}}$ . First we will calculate the expectation of  $\tilde{\boldsymbol{\beta}}$  and use the constraint of unbiasedness to obtain a constraint on  $\mathbf{D}$ , namely  $\mathbf{D}\mathbf{X} = \mathbf{0}$ :

$$\begin{aligned} E[\tilde{\boldsymbol{\beta}}] &= E[\mathbf{C}\mathbf{y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &\stackrel{\text{Eq. (J.17)}}{=} E\left[\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{=\text{const.}} \mathbf{X}\boldsymbol{\beta} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}}_{=\text{const.}} \boldsymbol{\epsilon}\right] \\ &\stackrel{\text{Eqs. (J.18–J.19)}}{=} \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}\right)\mathbf{X}\boldsymbol{\beta} + \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}\right)\underbrace{E[\boldsymbol{\epsilon}]}_{=0} \\ &= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{=\mathbf{I}_K} \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{I}_K + \mathbf{D}\mathbf{X})\boldsymbol{\beta} \end{aligned}$$

where we have used ( $X_1, X_2$  random variables;  $c$  is a constant)

$$E[X_1 + X_2] = E[X_1] + E[X_2] \quad (\text{J.17})$$

$$E[cX_1] = cE[X_1] \quad (\text{J.18})$$

$$E[c] = c \quad (\text{J.19})$$

and  $\mathbf{I}_K$  is  $K \times K$  identity matrix.  $\tilde{\boldsymbol{\beta}}$  is unbiased, i.e.  $E[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ , if and only if

$$\mathbf{D}\mathbf{X} = \mathbf{0}. \quad (\text{J.20})$$

<sup>6</sup>Please note that minimum variance does not necessarily imply 'small' variance and thus unbiased estimators might sometimes do a better job with respect to variance (compare, for example, the discussion of ridge regression in Section 17.6.4).

Now, we will calculate the variance of  $\tilde{\beta}$ :

$$\begin{aligned}
 \text{Var}(\tilde{\beta}) &= \text{Var}(C\mathbf{y}) \\
 &= C \text{Var}(\mathbf{y}) C' \quad (\text{valid for square matrices } C) \\
 &= \sigma^2 C C' \\
 &= \sigma^2 \left( (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' + \mathbf{D} \right) \left( \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + \mathbf{D}' \right) \\
 &= \sigma^2 \left( (\mathbf{X}' \mathbf{X})^{-1} \underbrace{\mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1}}_{=I} + (\mathbf{X}' \mathbf{X})^{-1} \underbrace{\mathbf{X}' \mathbf{D}'}_{= (\mathbf{D}' \mathbf{X})' = 0, \text{ Eq. (J.20)}} + \underbrace{\mathbf{D}' \mathbf{X}}_{= 0, \text{ Eq. (J.20)}} (\mathbf{X}' \mathbf{X})^{-1} + \mathbf{D} \mathbf{D}' \right) \\
 &= \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} + \sigma^2 \mathbf{D} \mathbf{D}' \\
 &= \text{Var}(\hat{\beta}) + \underbrace{\sigma^2 \mathbf{D} \mathbf{D}'}_{> 0}
 \end{aligned}$$

Since  $\mathbf{D} \mathbf{D}'$  is a positive semidefinite matrix,  $\text{Var}(\tilde{\beta})$  exceeds  $\text{Var}(\hat{\beta})$  by a positive semidefinite matrix. **q.e.d.**

### J.3 From Bayes' Theorem to least-squares (\*)

This section is meant for mathematically inclined readers and could be skipped by others. However, it is worth following the derivation of least-squares from Bayes' Theorem: [for me it caused an aha or light bulb moment!](#)

The starting point is the product rule of probability theory

$$P(A \cap B | I) = P(A | B \cap I) P(B | I) = P(B | A \cap I) P(A | I) \quad (\text{J.21})$$

which by slight rearrangement leads to [Bayes' Theorem](#)

$$P(A | B \cap I) = \frac{P(B | A \cap I) P(A | I)}{P(B | I)}$$

for propositions  $A$  and  $B$ . In the context of parameter estimation one considers the propositions  $A$  = 'model & model parameters' ('parameters' for short) and  $B$  = 'data', and thus Bayes' Theorem reads<sup>7</sup>

$$P(\text{parameters} | \text{data}) = \frac{P(\text{data} | \text{parameters}) \cdot P(\text{parameters} | I)}{P(\text{data} | I)} \quad (\text{J.22})$$

The goal is to calculate the [posterior](#) which is the PDF for the model parameter(s) in the light of data  $(y_k(x_k), k = 1, 2, \dots, n)$ :

$$P(\text{parameter} | \text{data}) = P(\beta | y_k(x_k), k = 1, 2, \dots, n) = \text{posterior} \quad (\text{J.23})$$

In the case of one model parameter  $\beta$  only (here:  $\beta$  = slope of straight line through origin), one hopes finding a unimodal<sup>8</sup> PDF and take the [location of the maximum as the optimal value of  \$\beta\$](#)  and the [standard deviation of the posterior PDF as the uncertainty of the model parameter](#).

There are three terms on the right-hand-side of Eq. J.22. The likelihood distribution

$$P(\text{data} | \text{parameters}) = P(y_k(x_k), k = 1, 2, \dots, n | \beta) \quad (\text{J.24})$$

a measure for the likelihood to observe the data  $y_k(x_k), k = 1, 2, \dots, n$  from a statistical population characterized by the model parameter  $\beta$ . By a switch of perspective one obtains from the likelihood distribution the [likelihood function](#) (or likelihood for short)

$$f(\text{parameters} | \text{data}) = f(\beta | y_k(x_k), k = 1, 2, \dots, n) \quad \text{likelihood} \quad (\text{J.25})$$

is a measures of the likelihood for the model parameter  $\beta$  given the data  $y_k(x_k), k = 1, 2, \dots, n$ ; please note that in general the likelihood function  $f()$  is not normalized to 1, i.e. it is not a PDF; of course normalization is possible, however, usually not required. In simple linear regression one makes the (implicit) assumption that the [observations  \$y\_k\$  result from an exact model \(here: straight line through origin\) plus normal noise with zero mean and variance  \$\sigma\_k^2\$](#) . Thus the likelihood to observe  $y_1$  is given by the [normal distribution](#)

$$\mathcal{N}(y_1; \mu_1 = \beta x_1, \sigma_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(\beta x_1 - y_1)^2}{2\sigma_1^2}} = P(y_1(x_1) | \beta) \quad (\text{J.26})$$

By a switch of perspective one obtains the likelihood (function) of  $\beta$  given  $y_1(x_1)$

$$f(\beta; y_1(x_1)) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(\beta x_1 - y_1)^2}{2\sigma_1^2}} \quad (\text{J.27})$$

<sup>7</sup>Please note that the notation is again simplified by leaving out ' $\cap I$ '.

<sup>8</sup>unimodal = one maximum

The likelihood of  $\beta$  given  $y_2(x_2)$  reads accordingly

$$f(\beta; y_2(x_2)) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(\beta x_2 - y_2)^2}{2\sigma_2^2}} \quad (\text{J.28})$$

What's the likelihood for  $\beta$  given observations  $y_1(x_1)$  and  $y_2(x_2)$ ? For observations that are **independent** of each other one can apply the simplified product rule of probabilities (Eq. 4.9)

$$\begin{aligned} f(\beta | y_1(x_1), y_2(x_2)) &= \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(\beta x_1 - y_1)^2}{2\sigma_1^2}} \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(\beta x_2 - y_2)^2}{2\sigma_2^2}} \\ &= e^{-\sum_{k=1}^2 \frac{(\beta x_k - y_k)^2}{2\sigma_k^2}} \prod_{k=1}^2 \left( \frac{1}{\sigma_k \sqrt{2\pi}} \right) \end{aligned} \quad (\text{J.29})$$

The generalization to  $n$  independent data is obvious and reads

$$f(\beta | y_k(x_k), k = 1, 2, \dots, n) = e^{-\sum_{k=1}^n \frac{(\beta x_k - y_k)^2}{2\sigma_k^2}} \prod_{k=1}^n \left( \frac{1}{\sigma_k \sqrt{2\pi}} \right) \quad (\text{J.30})$$

The second term in the numerator of the right-hand-side of Eq. J.22 is the **prior**

$$P(\text{parameters} | I) = P(\beta | I) = \text{prior}. \quad (\text{J.31})$$

It is a PDF<sup>9</sup> for various values of the model parameter(s) before ('prior to' in the logical sense) taking into account the data, however, based on the background knowledge  $I$ . Or, in other words, it is the mathematical expression for the background information or often ignorance with respect to the model parameters.

If one does not know anything about the prior distribution of the model parameters one can use a constant as prior. This is called the '**flat prior**'. Here nothing is known about the slope  $\beta$  (not even if it is positive or negative!) and thus the flat prior is chosen.

The term in the denominator of the right-hand-side of Eq. J.22

$$P(\text{data} | I) = P(y_k(x_k), k = 1, 2, \dots, n | I)$$

is the PDF for observing the data just based on the background knowledge, however, without specifying any model (straight line or exponential function or cos-function or  $\dots$ ). It seems impossible to assign a PDF for this term. However, this is not a drawback at all because one can use this term as a **normalization factor**. In general, the product of the prior and the likelihood function does not yield a PDF because the resulting expression is not normalized to 1. In order to obtain a posterior PDF one uses  $P(\text{data} | I)$  as **normalization factor**.

Thus finally one obtain the posterior PDF, which here (for flat prior) is essentially (except for normalization) the likelihood function

$$f(\beta | y_k(x_k), k = 1, 2, \dots, n) = c e^{-\sum_{k=1}^n \frac{(\beta x_k - y_k)^2}{2\sigma_k^2}} \prod_{k=1}^n \left( \frac{1}{\sigma_k \sqrt{2\pi}} \right) \quad (\text{J.32})$$

where  $c$  is the normalization constant.

---

<sup>9</sup>Some priors are actually not PDFs because they cannot be normalized. These priors are called 'improper priors'. An important example is the 'flat prior' where  $P(\beta | I)$  is a constant over in infinite range.

Now one searches for the maximum of the posterior  $P(\beta | y_k(x_k), k = 1, 2, \dots, n) = P(\beta | \text{data})$ : its location will be used as optimal estimate  $\hat{\beta}$  for the slope  $\beta$ . When, i.e. at which value of  $\beta$ , is the posterior maximal? The normalization constant  $c$  has no impact on the location of the maximum. The same applies for the factor

$$\prod_{k=1}^n \left( \frac{1}{\sigma_k \sqrt{2\pi}} \right) \quad (\text{J.33})$$

which (for given  $\sigma_k$  values) is a constant and independent of  $\beta$ . One is left with

$$e^{- \sum_{k=1}^n \frac{(\beta x_k - y_k)^2}{2 \sigma_k^2}} \quad (\text{J.34})$$

The term

$$\sum_{k=1}^n \frac{(\beta x_k - y_k)^2}{2 \sigma_k^2} \quad (\text{J.35})$$

in the exponent of the exponential function is obviously non-negative because of the squares (and in real applications positive). The posterior becomes maximal when this term becomes minimal and thus one can reformulate the optimality condition as

$$\sum_{k=1}^n \frac{(\beta x_k - y_k)^2}{2 \sigma_k^2} \Rightarrow \text{minimum} \quad (\text{J.36})$$

In other words, (generalized) least-squares has been derived from Bayes' Theorem!!!

**Discussion:**

- Let us review the assumptions that were made to obtain generalized least squares: (1) Additive normally distributed noise ( $\Rightarrow$  the likelihood function for a single data  $y_k$  is a normal distribution), (2) independence of data ( $\Rightarrow$  the likelihood for the complete data set can be calculated as the product of the likelihoods for each single data point), (3) a flat prior is used, i.e. the posterior is essentially (except for the normalization) equal to the likelihood.
- Special case of equal noise level: when the noise levels, as measured by the standard deviations  $\sigma_k$  or the variances  $\sigma_k^2$ , are equal to each other, Eq. J.36 simplifies to

$$\sum_{k=1}^n (\beta x_k - y_k)^2 \Rightarrow \text{minimum} \quad (\text{J.37})$$

which can be expressed as

$$\sum_{k=1}^n (\text{model prediction} - \text{observed response})^2 \Rightarrow \text{minimum} \quad (\text{J.38})$$

which is valid for other models<sup>10</sup>.

- The residuals  $r_k$  are defined as the differences between observations  $y_k$  and model predictions  $\beta x_k$ :

$$r_k = y_k - \beta x_k. \quad (\text{J.39})$$

- The residuals can be used to check two prerequisites of simple linear regression: (1) normality of the additive noise (plot histogram of residuals or estimate density of residuals, apply Shapiro-Wilk test to residuals) and (2) homogeneous noise level (look for pattern in scatterplot of residuals).

---

<sup>10</sup>In the derivation of Eq. J.37 no specific features of the 'straight line through the origin' model have been used: this model was only used to give an example and to avoid more complicated notation.

- Usually no detailed information is available about the noise level before application of linear regression. The variance of the noise can be estimated from the residuals as

$$\hat{\sigma}^2 = \frac{1}{\nu} \sum k = 1^n r_k^2 \quad (\text{J.40})$$

with  $\nu = n - 1$  for straight line through origin model (1 model parameter) or  $\nu = n - 2$  for straight line model (2 model parameter, namely intercept and slope).

- If the noise is non-normal or the noise level varies with  $x$ , one can either transform (for example, taking the logarithm if all  $y_k$  are positive) or prescribe a function  $\sigma(x)$  describing the variation of the noise level and use this function in the application of general least squares using Eq. J.36.

## J.4 Analytic solution of the least squares straight line problem (\*)

The sum of least squares (fitting a straight line, simple linear regression) is given by

$$\text{SLS}(\underbrace{\alpha, \beta}_{\text{unknown}}; \underbrace{x_k, y_k}_{\text{given, data}}) = \sum_{k=1}^n \left( \underbrace{\beta_0 + \beta x_k}_{\text{model}} - \underbrace{y_k}_{\text{data}} \right)^2 \quad (\text{J.41})$$

The location of the minimum of  $\text{SLS}(\beta_0, \beta)$  can be calculated as follows. The first derivatives of SLS with respect to  $\beta_0$  and  $\beta$  are set to zero (necessary condition for minimum)  $\Rightarrow$

$$\hat{\beta}_{0,\text{opt}} = \frac{v p - w q}{u v - w^2} \quad (\text{J.42})$$

$$\hat{\beta}_{\text{opt}} = \frac{u q - w p}{u v - w^2} \quad (\text{J.43})$$

where

$$u = 2 \sum_{k=1}^n x_k^2, \quad v = 2 n, \quad w = 2 \sum_{k=1}^n x_k \quad (\text{J.44})$$

$$p = 2 \sum_{k=1}^n x_k y_k, \quad q = 2 \sum_{k=1}^n y_k \quad (\text{J.45})$$

### Derive analytic solution

Necessary condition for minimum: first derivatives of SLS (Eq. J.41) with respect to  $\beta_0$  and  $\beta$  must vanish

$$\frac{\partial \text{SLS}}{\partial \beta_0} = \sum_{k=1}^n 2 (\beta_0 + \beta x_k - y_k) \quad (\text{J.46})$$

$$= 2 n \beta_0 + 2 \beta \sum_{k=1}^n x_k - 2 \sum_{k=1}^n y_k \quad (\text{J.47})$$

$$= v \beta_0 + w \beta - q = 0 \quad (\text{J.48})$$

and

$$\frac{\partial \text{SLS}}{\partial \beta} = 2 \sum_{k=1}^n x_k (\beta_0 + \beta x_k - y_k) \quad (\text{J.49})$$

$$= 2 \beta_0 \sum_{k=1}^n x_k + 2 \beta \sum_{k=1}^n x_k^2 - 2 \sum_{k=1}^n x_k y_k \quad (\text{J.50})$$

$$= w \beta_0 + u \beta - p = 0 \quad (\text{J.51})$$

where

$$p = 2 \sum_{k=1}^n x_k y_k, \quad q = 2 \sum_{k=1}^n y_k, \quad u = 2 \sum_{k=1}^n x_k^2, \quad v = 2 n, \quad w = 2 \sum_{k=1}^n x_k \quad (\text{J.52})$$

Eliminate  $\beta$  terms: multiply Eq. J.48 by  $u$  and Eq. J.51 by  $-w$ , then add up  $\Rightarrow$

$$\beta_0 (u v - w^2) = q u - p w \quad (\text{J.53})$$

and thus

$$\beta_0 = \frac{q u - p w}{u v - w^2} \quad (\text{J.54})$$

and

$$\beta = \frac{p - w \beta_0}{u} \quad (\text{J.55})$$

## J.5 Fit polynomial to data

Polynomials

$$y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots = \sum_{k=0}^K \beta_k x^k \quad (\text{J.56})$$

are also linear models in the sense that they are linear in their coefficients  $\beta_k$ . If we again assume that the predictor  $x$  is non-stochastic (or small uncertainties can be neglected in the current context), the noise in  $y$  is additive and stems from a normal population with mean  $\mu = 0$  and unknown variance  $\sigma^2$  that does not vary with  $x$  (homoscedasticity or 'no pattern in noise'), we can again apply ordinary least-squares by applying the R routine `lm()`. An example using artificial data is shown in Fig. J.5.

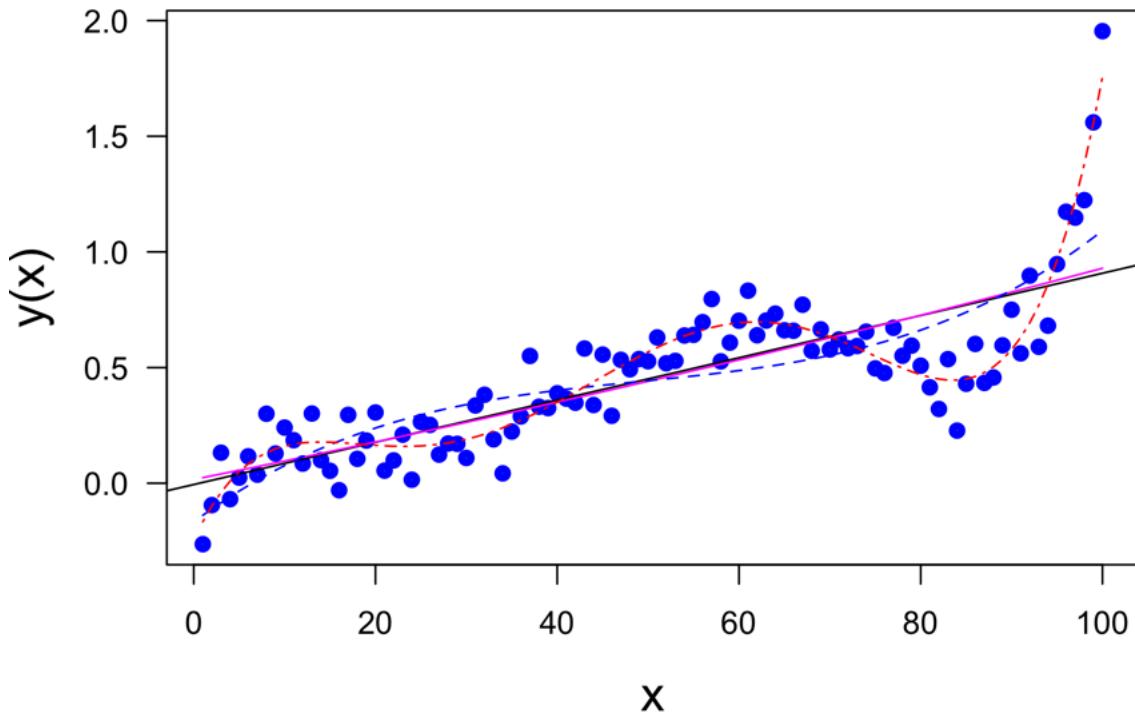


Figure J.5: Ordinary least-squares fit to an artificial data set. 4 different polynomials have been fitted to the data: (1) up to linear in  $x$  (straight line, black), (2) up to second order in  $x$  (magenta curve), (3) up to third order in  $x$  (blue dashed line), (4) up to fifth order in  $x$  (red dash-dotted line). The fit by the fifth order polynomial looks best. A quantitative measure for comparing the 4 models can be obtained from the AIC values: -24.91, -23.12, -31.61, -161.32. The AIC values for the models differ by less than 2, and the quadratic model is not better than the linear model. The cubic model is clearly better than the first two models and the fifth order model is clearly the best. [PolynomialFit.R](#)

# Appendix K

## Errors in variables (Appendix)

### K.1 Maximum likelihood estimation (MLE) aka Deming regression

Although maximum likelihood estimation (MLE) might look like the method of choice, this approach is facing surprisingly a number of difficulties even under the assumption of normal distributions (compare, for example, Zellner, 1971, and Casella & Berger, 2002, and references therein for further details). Additional information is required to calculate estimates from data using ML. The most popular ML estimator is based on the assumption that the ratio of the error variances,  $\lambda = \sigma_x^2 / \sigma_y^2$ , is known:

$$\hat{\beta}_{\text{MLE}} = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}^2} \quad (\text{K.1})$$

$$\hat{\beta}_{0,\text{MLE}} = \bar{y} - \hat{\beta}_{\text{MLE}} \bar{x} \quad (\text{K.2})$$

where the sum of squares ( $S_{xx}$ ,  $S_{yy}$ ,  $S_{xy}$ ) are defined as usual (Eqs. 15.8–15.12). This estimator is also known as ‘Deming regression’ (Deming, 1943). If one applies Eq. K.1 to the Jitjareonchai et al. (2006) data with the true value of  $\lambda = 4/9$  one obtains  $\hat{\beta}_{\text{MLE}} = 2.79$  and  $\hat{\beta}_{0,\text{MLE}} = 17.5$ . MLE seems to work quite good if good guesses for  $\lambda$  are available. However, I’m not aware of any sensible method for estimating  $\lambda$  from the data.<sup>1</sup>

#### Exercise 84 Deming and ML estimator are identical

Deming (1943, p. 184) gave the following estimator of the slope

$$\hat{\beta}_{\text{Deming}} = \frac{(S_{yy} - \eta S_{xx}) + \sqrt{(S_{yy} - \eta S_{xx})^2 + 4\eta S_{xy}^2}}{2 S_{xy}^2}$$

where  $\eta = \sigma_\epsilon^2 / \sigma_\delta^2 = 1/\lambda$  and the other symbols like in this section. Deming further remarked: “This is equivalent to a result obtained by Kummell in 1876, Karl Pearson in 1901, and Gini in 1921.”

Show that this estimator is identical to the maximum likelihood estimator derived by Casella & Berger (2002, Eq. 12.2.16).

---

<sup>1</sup>Legendre & Legendre (2012, p. 552) write: “Contrary to the sample variance, the error variance on  $x$  and  $y$  cannot be estimated from the data.”

## K.2 Markov Chain Monte Carlo (MCMC)

*Markov Chain Monte Carlo (MCMC) is a method to derive estimates in case of high dimensional problems. In errors in variables problems (Chapter 15) the number of unknowns increases with sample size. In the current section the errors in variables problem will be treated by MCMC using the Gibbs sampler approach. This section is based on Jitjareonchai et al. (2006) who used a priori values for sum of squares for  $x$  and  $y$  to construct informative priors. The method is applied to the data set  $(x, y)$  (Eqs. 15.6 – 15.7) given by Jitjareonchai et al. (2006)*

**Further reading (MCMC):** Gilks, Richardson, Spiegelhalter (1995), Robert & Casella (2009)

The errors in variables model and assumptions are described in Section 15.2. The likelihood function for a single  $(x_i, y_i)$  is given by

$$\mathcal{L}(x_i, y_i | \beta, \beta_0, \sigma_x^2, \sigma_y^2, \xi_i) \propto \frac{1}{\sigma_x} \exp \left[ -\frac{(x_i - \xi_i)^2}{2\sigma_x^2} \right] \times \frac{1}{\sigma_y} \exp \left[ -\frac{(y_i - \beta\xi_i - \beta_0)^2}{2\sigma_y^2} \right] \quad (\text{K.3})$$

$$\propto \frac{1}{\sigma_x \sigma_y} \exp \left[ -\frac{(x_i - \xi_i)^2}{2\sigma_x^2} - \frac{(y_i - \beta\xi_i - \beta_0)^2}{2\sigma_y^2} \right] \quad (\text{K.4})$$

and thus the likelihood function for  $(x, y)$  reads

$$\mathcal{L}(x, y | \beta, \beta_0, \sigma_x^2, \sigma_y^2, \xi_i) \propto \prod_{i=1}^n \frac{1}{\sigma_x \sigma_y} \exp \left[ -\frac{(x_i - \xi_i)^2}{2\sigma_x^2} - \frac{(y_i - \beta\xi_i - \beta_0)^2}{2\sigma_y^2} \right] \quad (\text{K.5})$$

$$\propto \frac{1}{\sigma_x^n \sigma_y^n} \exp \left\{ -\sum_{i=1}^n \left[ \frac{(x_i - \xi_i)^2}{2\sigma_x^2} + \frac{(y_i - \beta\xi_i - \beta_0)^2}{2\sigma_y^2} \right] \right\} \quad (\text{K.6})$$

The number of unknowns is large, namely  $n \xi_i$ s,  $\beta$ ,  $\beta_0$ ,  $\sigma_x$ ,  $\sigma_y$ , i.e.  $n + 4$  unknowns, and thus always larger than the number of observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . In the context of the ML analysis of the structural form of EVM, Zellner (1971, p. 128) writes: "Thus we cannot obtain estimates of all parameters unless prior information is available to reduce the number of unknown parameters." and (p. 132) "Further, it should be appreciated that since prior information is needed to identify unknown parameters no matter how large the sample size is this prior information will exert an important influence on posterior inferences."

In order to derive the posterior for the unknowns, one needs priors for  $\beta$ ,  $\beta_0$ ,  $\sigma_x^2$ ,  $\sigma_y^2$ . For  $\beta$ ,  $\beta_0$  flat priors (over large ranges) can be applied; the corresponding constants,  $\text{prior}(\beta) = \text{const}$  and  $\text{prior}(\beta_0) = \text{const}_0$ , do not have to be specified.<sup>2</sup> It is well known that the sum of squares,  $P_x = \sum_{i=1}^n (x_i - \bar{x})^2$ , divided by  $\sigma_x^2$  follow the  $\chi^2$  distribution with  $v = n$  degrees of freedom (Section C.3.3, Exercise 70), i.e. the probability density for  $z = P_x / \sigma_x^2$  is  $\chi^2(z; v = n)$ :

$$\chi^2(z; v) = \frac{z^{(v/2-1)} e^{-z/2}}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} \quad \text{for } z > 0 \quad (\text{K.7})$$

or

$$\chi^2\left(P_x / \sigma_x^2; v\right) = \frac{\left(P_x / \sigma_x^2\right)^{(v/2-1)} e^{-\left(P_x / \sigma_x^2\right)/2}}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} \quad (\text{K.8})$$

$$\propto \left(\sigma_x^2\right)^{-(v/2-1)} \exp\left(-\frac{P_x}{2\sigma_x^2}\right) = p(\sigma_x^2; v, P_x) \quad (\text{K.9})$$

where in the last transformation we have dropped multiplicative factors that are independent of  $\sigma_x^2$ .<sup>3</sup> For given

<sup>2</sup>These constants would 'drop out' in the final normalization of the posterior distribution. However, we will see that in the MCMC method an explicit normalization is not required.

<sup>3</sup>This is – after correcting a typo – identical to (A-9) given by Jitjareonchai et al. (2006):

$$Df(v_x) \propto v_x^{-\left(\frac{k}{2}-1\right)} \exp\left(-\frac{P_x}{2v_x}\right) \quad (A-9)$$

$$k \equiv v, v_x \equiv \sigma_x^2.$$

$P_x$ , Eq. K.9 can be used as an informative prior for  $\sigma_x^2$ . Analogously, given  $P_y$ , an informative prior for  $\sigma_y^2$  is

$$p(\sigma_y^2; \nu, P_y) \propto (\sigma_y^2)^{-(\nu/2-1)} \exp\left(-\frac{P_y}{2\sigma_y^2}\right) \quad (\text{K.10})$$

Given the likelihood function (Eq. K.6) and the priors (Eqs. K.9 – K.10), the posterior reads (except for normalization)

$$\mathcal{P}(\beta, \beta_0, \sigma_x^2, \sigma_y^2, \xi_i | \mathbf{x}, \mathbf{y}) \propto (\sigma_x^2)^{-(\nu/2-1)} \exp\left(-\frac{P_x}{2\sigma_x^2}\right) (\sigma_y^2)^{-(\nu/2-1)} \exp\left(-\frac{P_y}{2\sigma_y^2}\right) \quad (\text{K.11})$$

$$\times \frac{1}{\sigma_x^n \sigma_y^n} \exp\left\{-\sum_{i=1}^n \left[\frac{(x_i - \xi_i)^2}{2\sigma_x^2} + \frac{(y_i - \beta\xi_i - \beta_0)^2}{2\sigma_y^2}\right]\right\} \quad (\text{K.12})$$

The Markov Chain Monte Carlo (MCMC) using the Gibbs sampler is an iterative process that proceeds as follows:

1. At  $t = 0$  ( $t$  is the iteration index) select starting parameter values for  $\beta_t, \beta_{0,t}, \sigma_{x,t}^2, \sigma_{y,t}^2$ .
2. Take a random sample for each element of  $\xi$  from<sup>4</sup>

$$\begin{aligned} f(\xi_{i,t+1} | \mathbf{x}, \mathbf{y}, \sigma_{x,t}^2, \sigma_{y,t}^2, \beta_{0,t}, \beta_t) \\ \propto \mathcal{N}\left\{\left(\frac{1}{\sigma_{x,t}^2} + \frac{\beta_t^2}{\sigma_{y,t}^2}\right)^{-1} \left(\frac{x_i}{\sigma_{x,t}^2} + \frac{\beta_t}{\sigma_{y,t}^2}(y_i - \beta_{0,t})\right), \left(\frac{1}{\sigma_{x,t}^2} + \frac{\beta_t^2}{\sigma_{y,t}^2}\right)^{-1}\right\} \end{aligned} \quad (\text{K.13})$$

A few comments are in order:

- (1) Eq. (K.13) is derived from the posterior (Eq. K.12) by discarding terms that do not contain  $\xi_i$  (compare Jitjareonchai et al. (2006) for more details) and by adding indices for the iteration step, namely  $t + 1$  on the left hand side and  $t$  on the right hand side of Eq. (K.13).
- (2) All parameter values on the right hand side of Eq. (K.13) are from iteration step  $t$  and thus known at step  $t + 1$ .
- (3) The term

$$\left(\frac{1}{\sigma_{x,t}^2} + \frac{\beta_t^2}{\sigma_{y,t}^2}\right)^{-1} = \frac{\sigma_{x,t}^2 \sigma_{y,t}^2}{\sigma_{y,t}^2 + \beta_t^2 \sigma_{x,t}^2} \quad (\text{K.14})$$

in the argument of the normal distribution in Eq. (K.13) is the variance (not the standard deviation).

- (4) Normalization of Eq. (K.13) is automatically done by random sampling in R using **rnorm()** with the mean and the standard deviation from Eq. (K.13).

3. Random sampling of  $\beta$  from

$$f(\beta_{t+1} | \mathbf{x}, \mathbf{y}, \sigma_{y,t}^2, \beta_{0,t}, \xi_{i,t+1}) \propto \mathcal{N}\left\{\frac{\sum_{i=1}^n \xi_{i,t+1} (y_i - \beta_{0,t})}{\sum_{i=1}^n \xi_{i,t+1}^2}, \frac{\sigma_{y,t}^2}{\sum_{i=1}^n \xi_{i,t+1}^2}\right\} \quad (\text{K.15})$$

Remarks:

- (1) Eq. (K.15) is derived from the posterior (Eq. K.12) by discarding terms that do not contain  $\beta$  and by adding indices for the iteration step, namely  $t + 1$  on the left hand side and  $t$  or, namely for the already updated values  $\xi_i, t + 1$  on the right hand side of Eq. (K.15).
- (2) All parameter values on the right hand side of Eq. (K.15) are known from iteration step  $t$  or are already updated at step  $t + 1$ .

<sup>4</sup>Eq. (A-3) in Jitjareonchai et al. (2006), however, after correcting a typo:  $\beta_t$  instead of  $\beta_i^2$ .

4. Random sampling of  $\beta_0$  from

$$f(\beta_{0,t+1} | \mathbf{x}, \mathbf{y}, \beta_{t+1}, \xi_{i,t+1}) \propto \mathcal{N} \left\{ \sum_{i=1}^n \left( \frac{y_i - \beta_{t+1} \xi_{i,t+1}}{n} \right), \frac{\sigma_y^2}{n} \right\} \quad (\text{K.16})$$

Remarks:

- (1) Note that all iteration indices on the right hand side are  $t + 1$  because  $\beta_{t+1}$  and all  $\xi_{i,t+1}$  have been updated already.
- (2) The second argument of the normal distribution,  $\sigma_y^2/n$ , is a variance (not a standard deviation).

5. Take a random sample from  $\chi_{2n}^2$ , yielding  $\chi_{2n,t+1}^2$ . Calculate the sum of squares

$$S_{x,t+1} = \sum_{i=1}^n (x_i - \xi_{i,t+1})^2 \quad (\text{K.17})$$

and set  $\sigma_{x,t+1}^2$  to  $(P_x + S_{x,t+1})/\chi_{2n,t+1}^2$ .

Remark:

The inclusion of the (assumed to be known a priori)  $P_x$  operates like an anchor that keeps  $\sigma_{x,t+1}^2$  near a certain position. However, the position is not strictly fixed because it varies from iteration to iteration with  $S_x$ . An additional variation results from the randomness of  $\chi_{2n,t+1}^2$ .

6. Take a random sample,  $\chi_{2n,t+1}^2$ , from  $\chi_{2n}^2$ . Calculate the sum of squares

$$S_{y,t+1} = \sum_{i=1}^n (y_i - \beta_{t+1} \xi_{i,t+1} - \beta_{0,t+1})^2 \quad (\text{K.18})$$

and set  $\sigma_{y,t+1}^2$  to  $(P_y + S_{y,t+1})/\chi_{2n,t+1}^2$ .

7. Set  $t$  to  $t + 1$  and repeat steps 2 through 7.

### K.2.1 Start values and prior parameters

The start values for  $\beta$ ,  $\beta_0$ ,  $\sigma_x^2$ ,  $\sigma_y^2$  were purposely chosen significantly different from their true values (Table K.1). The values of the prior parameters  $P_x$ ,  $P_y$ ,  $k$  were taken from Jitjareonchai et al. (2006; here listed in Table K.1; compare also Fig. K.1).

Parameter	$\beta$	$\beta_0$	$\sigma_x^2$	$\sigma_y^2$	$P_x$	$P_y$	$k$
Start value	10	-35	2.68	14.38	14.7	44.3	4
True value	3	10	4	9			

Table K.1: Start values and prior parameters ( $P_x$ ,  $P_y$ ,  $k$ )

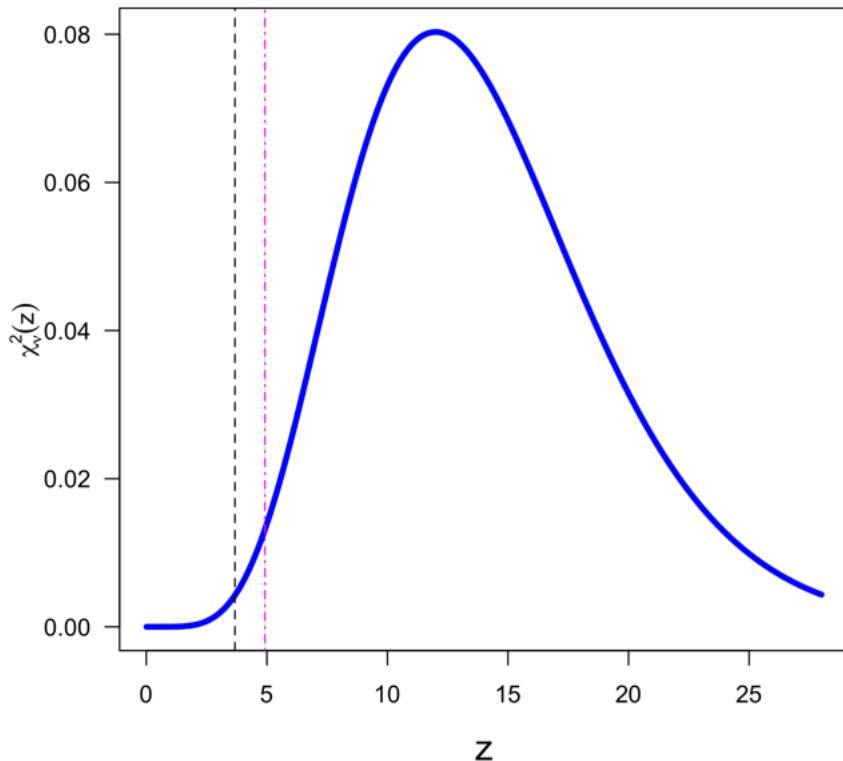


Figure K.1: The chi-squared distribution for  $\nu = 10 + 4 = 14$  degrees of freedom ( $\chi_{14}^2$ , blue solid line) and the (small) anchors  $P_x/\sigma_x^2$  (vertical black broken line) and  $P_y/\sigma_y^2$  (vertical magenta dash-dotted line).

R code: [Jitjareonchai06MCMC.R](#) (set: `sflag = 11`)

## K.2.2 Results

MCMC using the Gibbs sampler with  $M = 10^6$  cycles (iterations) was performed. Our results can not be compared one-to-one with those of Jitjareonchai et al. (2006) because we had to make our own choice for the start values of  $\sigma_x^2$  and  $\sigma_y^2$  and for the seed of the random number generators. However, the plots of our time series and density estimates look usually quite similar to the plots given in Jitjareonchai et al. (2006).

The first 2000 cycles of the slope and the intercept values are displayed in Fig. K.2 (compare also Fig. 2 in Jitjareonchai et al., 2006). A ‘burn-in’ phase with large deviations from, however, a clear tendency towards a long term mean value is clearly recognizable. The burn-in phase is caused by the large deviation of the start values of slope and intercept from their true values; it lasts for about 400 cycles (indicated by vertical magenta broken lines in Fig. K.2).

The MCMC estimates of slope and intercept and their uncertainties can be calculated as the mean  $\pm$  the standard deviation of the time series beyond the burn phase. Although it might be sufficient to cut out the first 400 cycles before calculating the mean and the variance of the time series, we follow Jitjareonchai et al. (2006) in dropping the first 1.5% of the time series (15000 cycles given  $M = 10^6$ ). The slope and the intercept are estimated by calculating the mean over the remaining 98.5% of the time series yielding  $\hat{\beta}_{\text{MCMC}} = 2.64 \pm 0.29$  and  $\hat{\beta}_0, \text{MCMC} = 20.3 \pm 6.1$  (uncertainties are  $\pm 1$  standard deviation). Density estimates for slope and intercepts are displayed in Figs. K.4 and K.5 (compare also Fig. 3 in Jitjareonchai et al., 2006). Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels are shown in Fig. K.6.

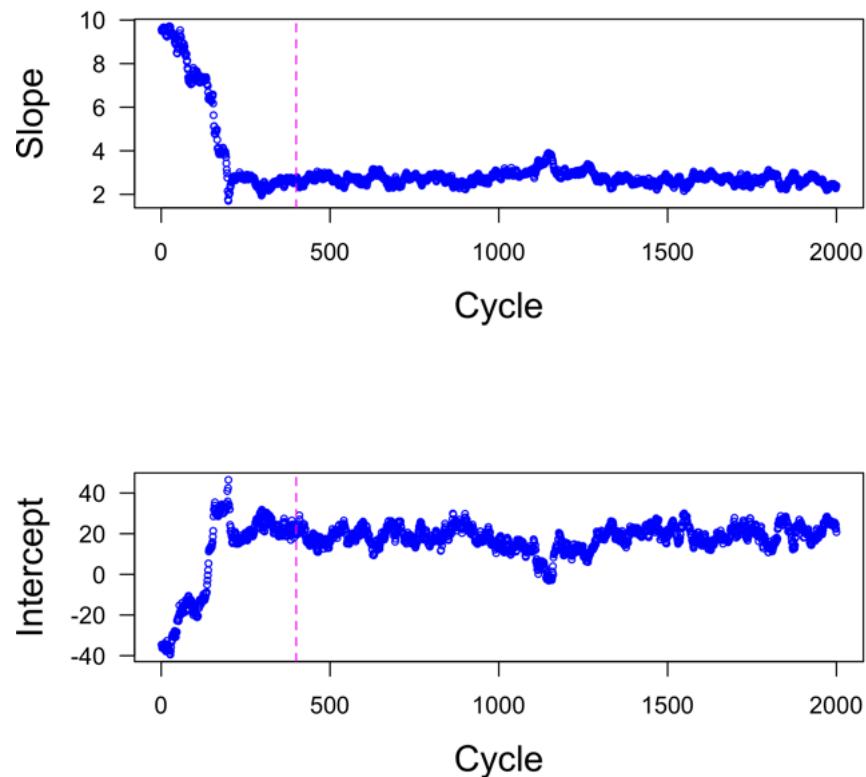
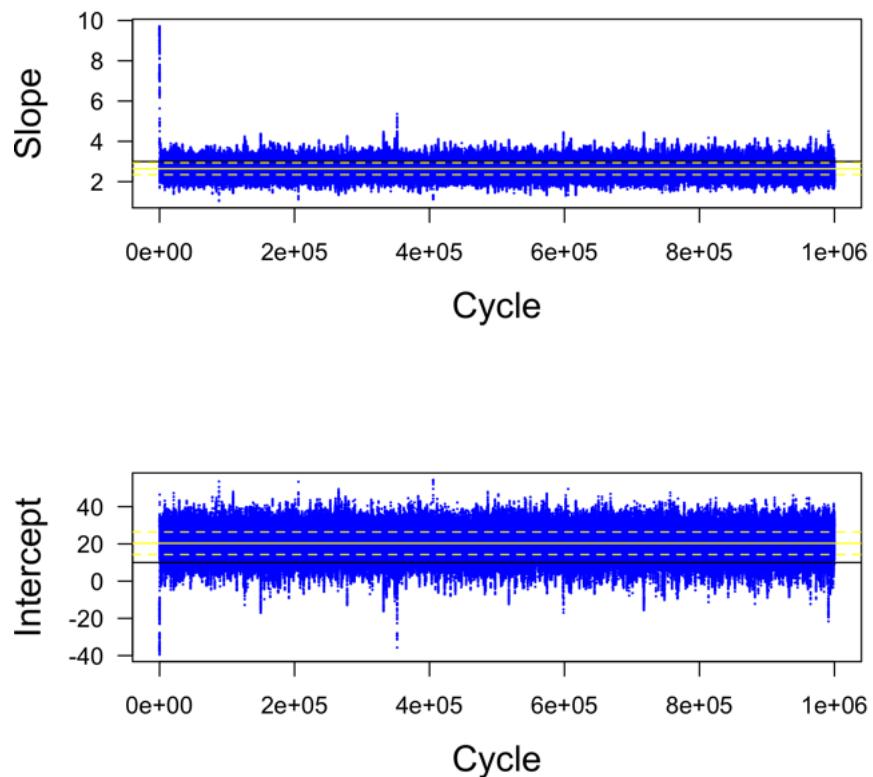


Figure K.2: **The first 2000 cycles of slope and intercept.** A 'burn-in' phase with large deviations from the true values, however, a clear tendency towards a long term mean value is clearly recognizable. The burn-in phase is caused by the large deviation of the start values of slope and intercept from their true values; it lasts for less than 400 cycles (indicated by vertical magenta broken lines).

R code: [Jitjareonchai06MCMC.R](#) (set: `sflag = 2`)



**Figure K.3: The complete time series of slope and intercept.** The values of slope and intercept keep on fluctuating with a constant level after the burn-in phase. The horizontal yellow lines indicate the estimates of slope and intercept (mean values of 98.5% of the time series; solid lines) and the mean  $\pm$  one standard deviation (broken lines); the horizontal black lines indicate the true values.

R code: [Jitjareonchai06MCMC.R](#) (set: `sflag = 3`)

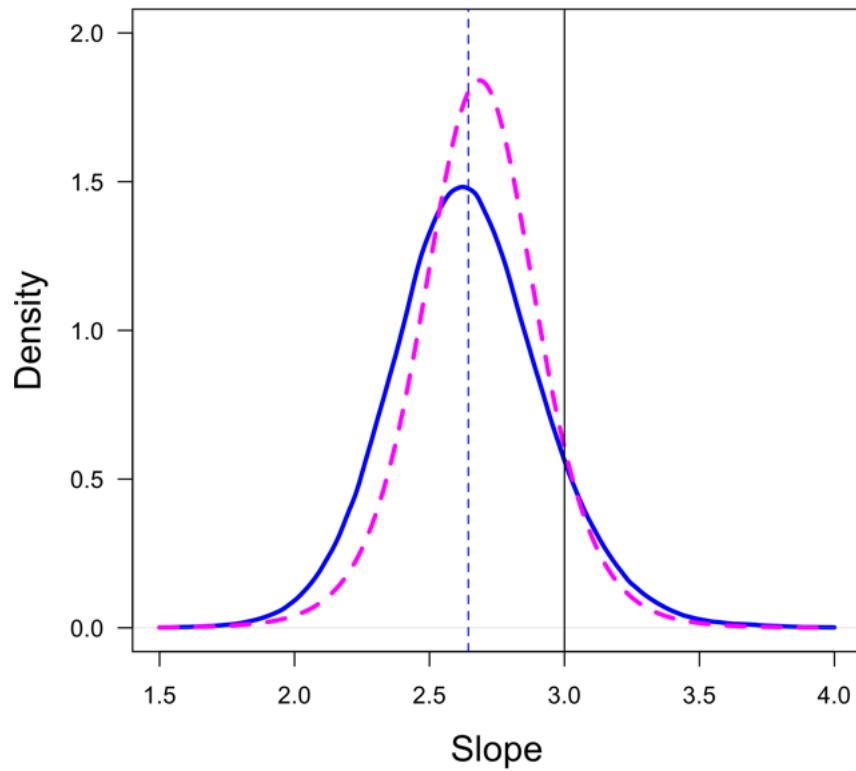


Figure K.4: **Density estimate for the slope,  $\beta$ , based on 98.5% of the MCMC cycles (blue solid line) and a normal approximation (magenta broken line). The true value of the slope is indicated by the vertical black solid line and its estimate (mean value) by the vertical blue broken line.**

R code: [Jitjareonchai06MCMC.R](#) (set: `sflag = 4`)

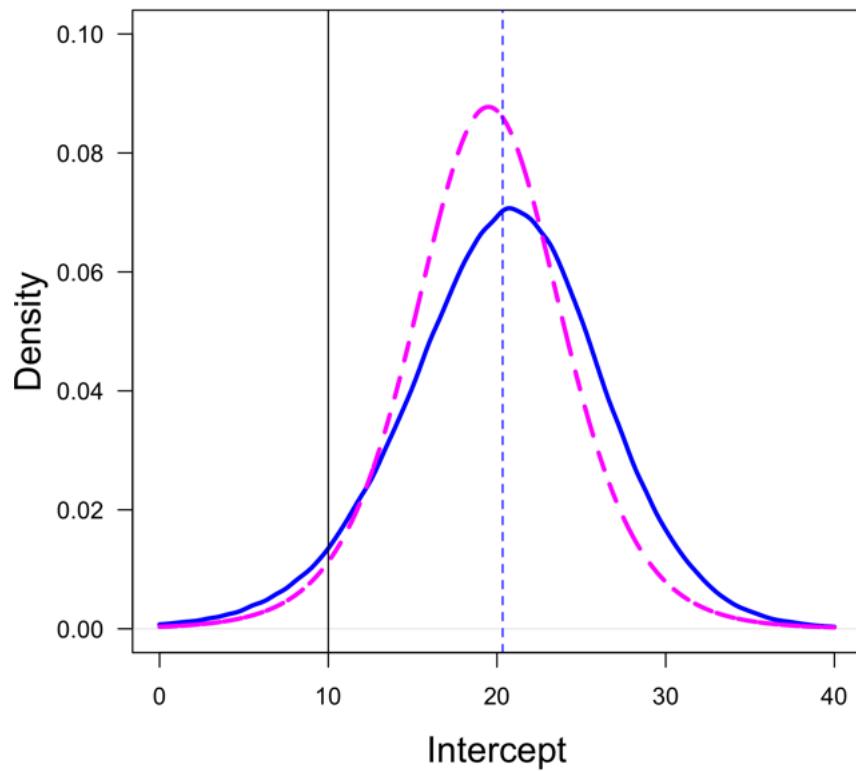


Figure K.5: **Density estimate for the intercept,  $\beta_0$ , based on 98.5% of the MCMC cycles (blue solid line)** and a normal approximation (magenta broken line). The true value of the intercept is indicated by the vertical black solid line and its estimate (mean value) by the vertical blue broken line.

R code: [EIVMCMC2504.R](#) (set: `sflag = 5`)

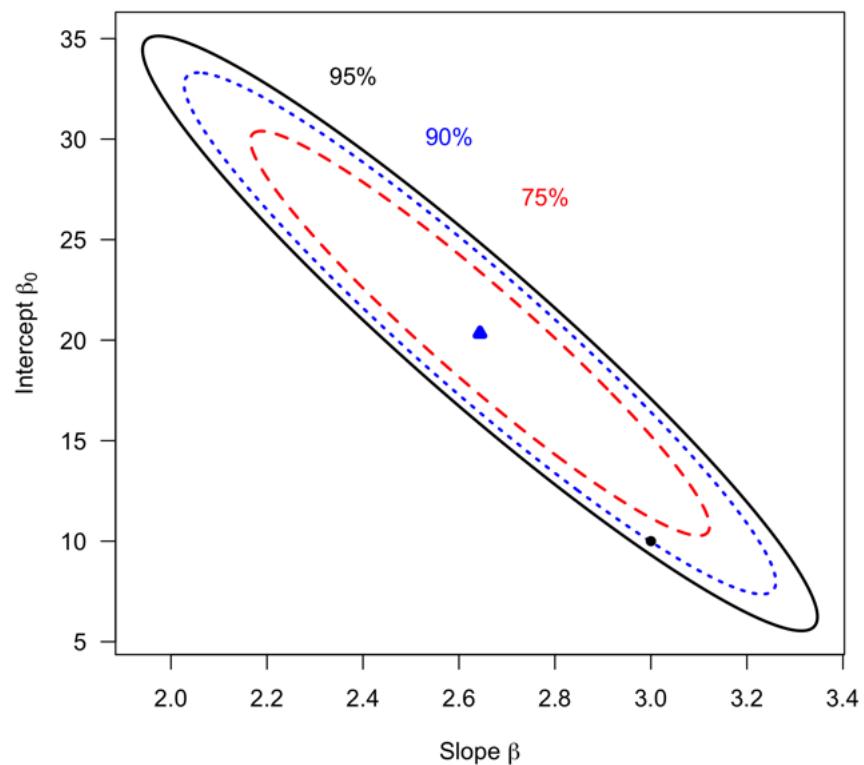


Figure K.6: **Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels.** The estimated values  $(\hat{\beta}_{MCMC}, \hat{\beta}_0_{MCMC})$  are indicated by the blue triangle in the center of the confidence ellipses. The true values  $(\beta, \beta_0) = (3, 10)$  are indicated by the black dot.

R code: [EIVMCMC2504.R](#) (set: `sflag = 6`)

The MCMC method also allows estimating the  $\xi_i$  and  $\gamma_i$  values (again averages over 98.5% of the time series):

$$\hat{\xi}_i = \{18.8719, 27.8795, 25.7495, 30.0524, 13.9042, 14.068, 17.2546, 13.5963, 26.5295, 12.5999\} \quad (\text{K.19})$$

$$\hat{\gamma}_i = \{70.2537, 93.896, 88.302, 99.5902, 57.2147, 57.647, 66.0097, 56.4086, 90.3415, 53.7964\} \quad (\text{K.20})$$

Estimated densities of  $\xi_i$ ,  $i = 1, 2, \dots, 6$  are displayed in Fig. K.7 (compare also Fig. 4 in Jitjareonchai et al., 2006). Jitjareonchai et al. (2006, Table 2) gave the following estimates:

$$\xi_i^{(J)} = \{18.8592, 28.0712, 25.8652, 30.2092, 13.7835, 13.9688, 17.212, 13.4836, 26.5858, 12.4958\} \quad (\text{K.21})$$

$$\gamma_i^{(J)} = \{70.3378, 93.4900, 87.9729, 98.9411, 57.5709, 58.0115, 66.1884, 56.7921, 89.8429, 54.2846\} \quad (\text{K.22})$$

The true values<sup>5</sup> are

$$\xi^* = \{18.3075, 27.4873, 25.3590, 29.8017, 14.2793, 16.4007, 18.2391, 15.3589, 26.7848, 12.6755\} \quad (\text{K.23})$$

$$\gamma^* = \{64.9225, 92.4619, 86.0770, 99.4051, 52.8379, 59.2021, 64.7173, 56.0767, 90.3544, 48.0265\} \quad (\text{K.24})$$

and thus one can calculate the mean squared errors (MSE)

$$\text{MSE}_{\hat{\xi}, \xi^*} = \frac{1}{n} \sum_{i=1}^{n=10} (\hat{\xi}_i - \xi_i^*)^2 \quad (\text{K.25})$$

and analogue for  $\text{MSE}_{\hat{\gamma}, \gamma^*}$  and for the raw data,  $\text{MSE}_{x, \xi^*}$ ,  $\text{MSE}_{y, \gamma^*}$  (Table K.2). The MSE values for the estimates are smaller than those for the raw data, i.e. the Gibbs sampler estimates are closer to the true values.

$\text{MSE}_{x, \xi^*}$	$\text{MSE}_{\hat{\xi}^{(J)}, \xi^*}$	$\text{MSE}_{\hat{\xi}, \xi^*}$	$\text{MSE}_{y, \gamma^*}$	$\text{MSE}_{\hat{\gamma}^{(J)}, \gamma^*}$	$\text{MSE}_{\hat{\gamma}, \gamma^*}$
1.2498	1.1871	1.0417	12.2986	10.0112	9.2110

Table K.2: Mean squared errors (MSE)

<sup>5</sup> $\xi^*$  and  $\gamma^*$  in Table 2 of Jitjareonchai et al., 2006.

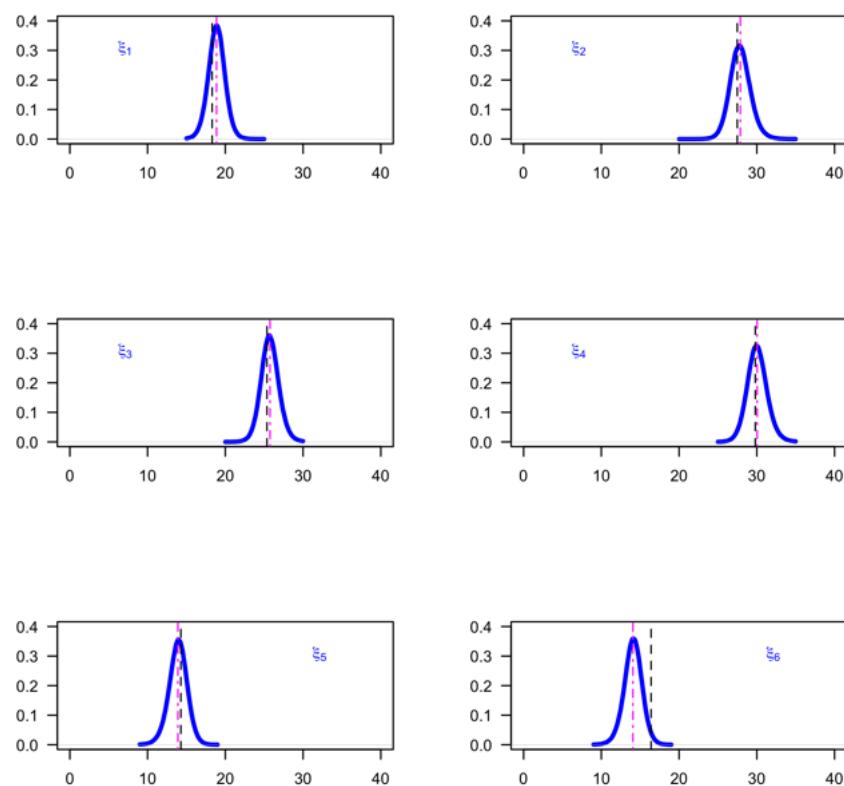


Figure K.7: Density estimates for  $\xi_i$ ,  $i = 1, 2, \dots, 6$  (compare also Fig. 4 Jitjareonchai et al., 2006).  
R code: [EIVMCMC2504.R](#) (set: `sflag = 7`)

Finally, we plot the density estimates for the variances  $\sigma_x^2$  and  $\sigma_y^2$  Figs. K.8 – K.9.

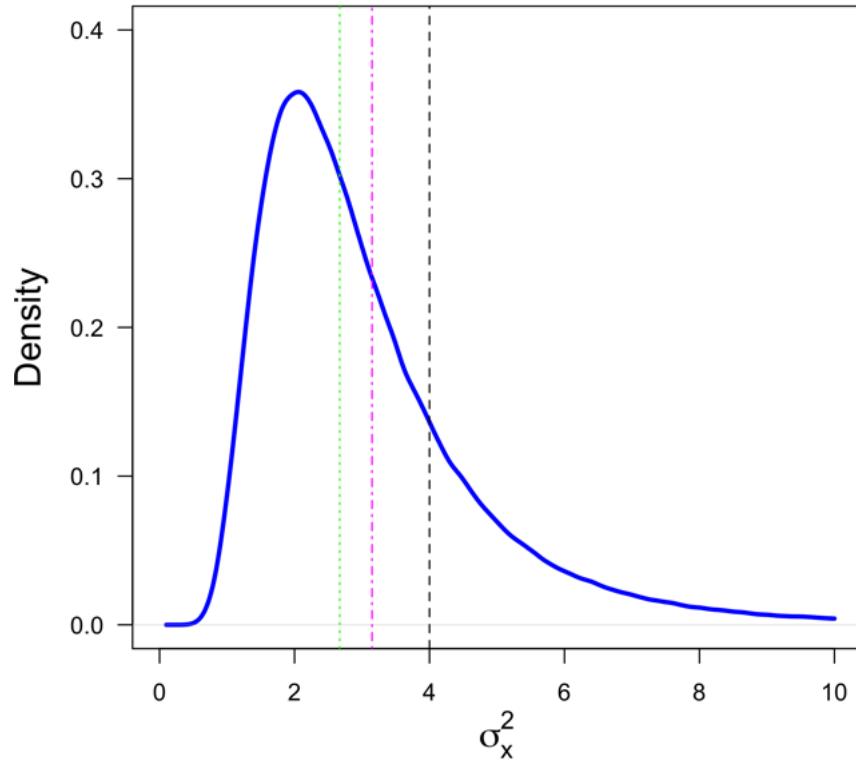


Figure K.8: **Density estimate for  $\sigma_x^2$ .** The vertical lines indicate true value (black dashed), the mean (magenta dash-dotted), and the median (green dotted).

R code: [EIVMCMC2504.R](#) (set: `sflag = 8`)

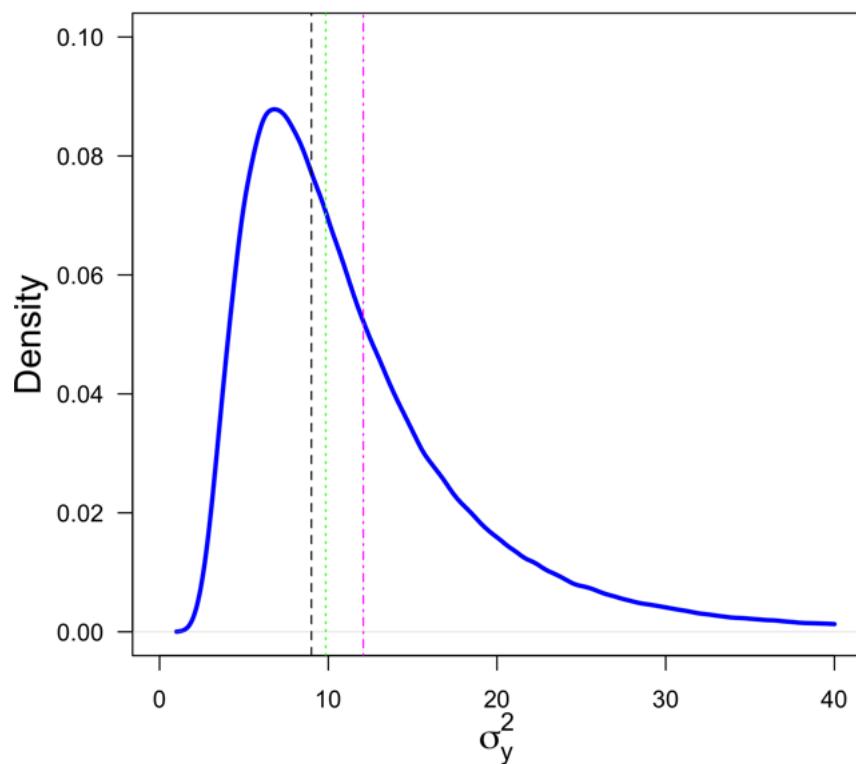


Figure K.9: **Density estimate for  $\sigma_y^2$ .** The vertical lines indicate true value (black dashed), the mean (magenta dash-dotted), and the median (green dotted).

R code: [EIVMCMC2504.R](#) (set: `sflag = 9`)

### K.3 Artificial data sets: regressions & MCMC

In order to illustrate the various methods for dealing with errors in variables (EIV) problems we will use artificial data sets. This has the advantage that one knows the exact model values and that one can study the behavior of the applied methods dependent on the model parameters. The sample sizes of the data set used here are a bit larger than that of the sample from Jitjareonchai et al. (2006) and they differ between each other in the variances of  $x$  and  $y$ .

To generate an artificial data set one has to specify the following parameters: slope  $\beta$ , intercept  $\beta_0$ , standard deviations  $\sigma_x$  and  $\sigma_y$ , sample size  $n$ , mean  $\mu_\xi$  and standard deviation  $\sigma_\xi$  to generate a random sample of  $\xi$  values.<sup>6</sup> The chosen parameter values are listed in Table .

Parameter	Data set 1	Data set 2	Data set 3	Data set 4
$n$	24	24	48	48
$\sigma_\xi$	1	2	1	4
Remarks	more spread in $\xi$	more data	more data & spread	

Table K.3: Chosen parameter values for generation of artificial data sets. Parameters for all data sets:  $\beta = 3$ ,  $\beta_0 = 2.5$ ,  $\sigma_x = 1$ ,  $\mu_\xi = 3$ ,  $\sigma_\gamma = 1.5$ . Remarks: in comparison to data set 1.

#### K.3.1 Start values and prior parameters for MCMC

The start values for  $\beta$ ,  $\beta_0$ ,  $\sigma_x^2$ ,  $\sigma_y^2$  for Monte Carlo Markov Chain (MCMC) were purposely chosen significantly different from their true values (Table K.4).

Parameter	$\beta$	$\beta_0$	$\sigma_x^2$	$\sigma_y^2$	$P_x$	$P_y$	$k$
Start value	2.7	1.25	0.01	0.0225	0.94	2.115	n
True value	3.0	2.50	1.00	2.2500			

Table K.4: Start values and prior parameters ( $P_x$ ,  $P_y$ ,  $k$ )

<sup>6</sup>Note that many other ways to generate  $\xi$  values might be reasonable, i.e. depending on the context the  $\xi$  could be set, for example, over a range with equidistant distances or sampled from a uniform distribution.

### K.3.2 Results

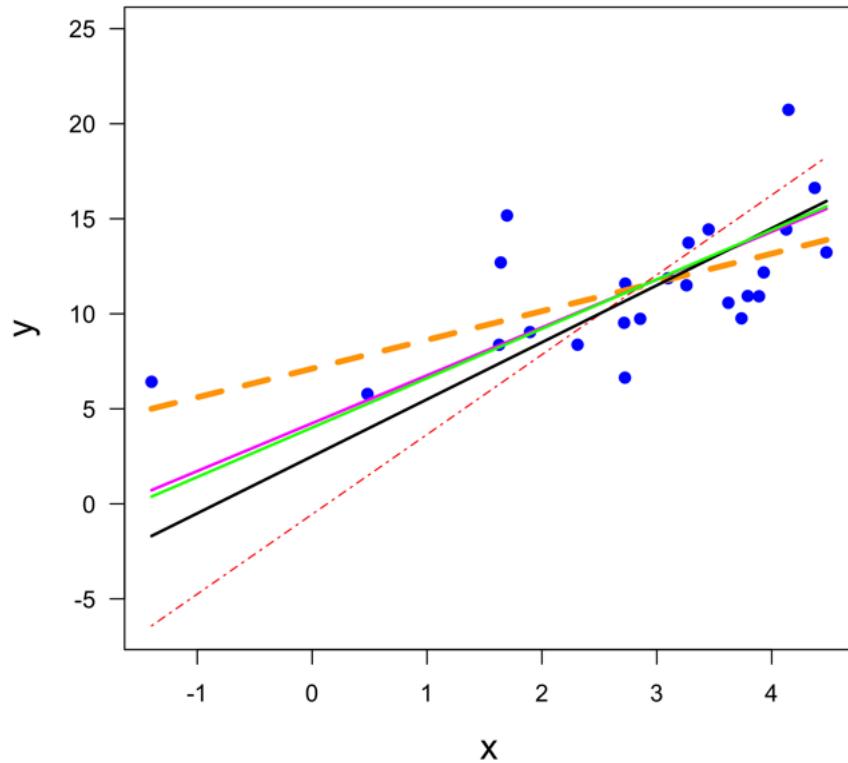


Figure K.10: **Data (blue dots) and estimates of straight lines for data set 1:** true line (black solid line,  $\beta = 3$ ,  $\beta_0 = 2.5$ ), regression of  $y$  on  $x$  (orange dashed line), regression via  $x$  on  $y$  (red dash-dotted line), geometric line (magenta solid line), MCMC (green solid line). Estimates of slopes and intercepts for all methods applied and all for data sets are listed in Tables K.5 – K.6. R code: [MCMCartificial2504.R](#) (set: `DataSet = 1` and `sflag = 1`)

The estimates of intercepts are often largely different from the true values especially when most data lie at some distance from the  $y$  axis because then small errors in slope estimates usually imply large deviations in intercept estimates.

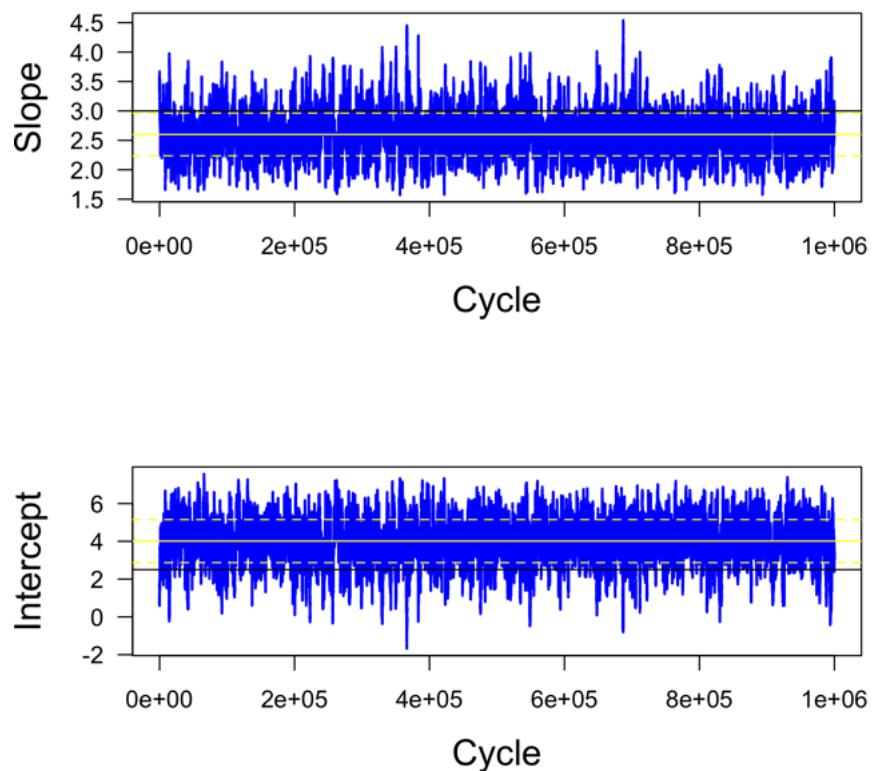


Figure K.11: **The complete time series of slope and intercept for data set 1.** A burn-in phase is hardly visible because the start values lie within the range of fluctuations. The horizontal yellow lines indicate the estimates of slope and intercept (solid lines) and the mean  $\pm$  one standard deviation (broken lines); the horizontal black lines indicate the true values.

R code: [MCMCartificial2504.R](#) (set: `DataSet = 1` and `sflag = 2`)

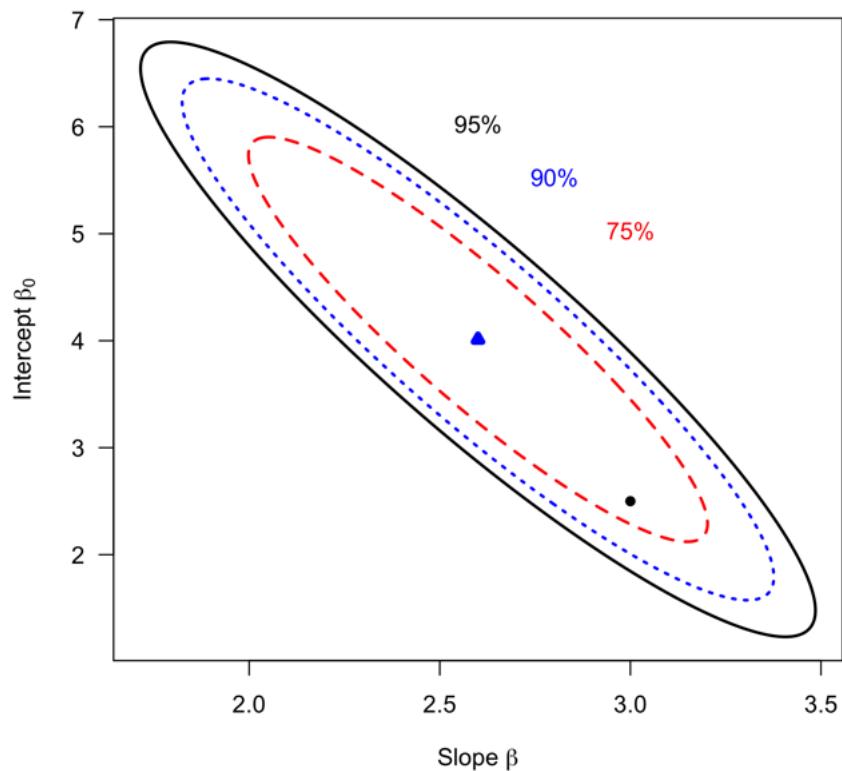


Figure K.12: **Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels for data set 1.** The estimated values  $(\hat{\beta}_{MCMC}, \hat{\beta}_0_{MCMC})$  are indicated by the blue triangle in the center of the confidence ellipses. The true values  $(\beta, \beta_0) = (3, 2.5)$  are indicated by the black dot.

R code: [MCMCartificial2504.R](#) (set: `DataSet = 1` and `sflag = 3`)

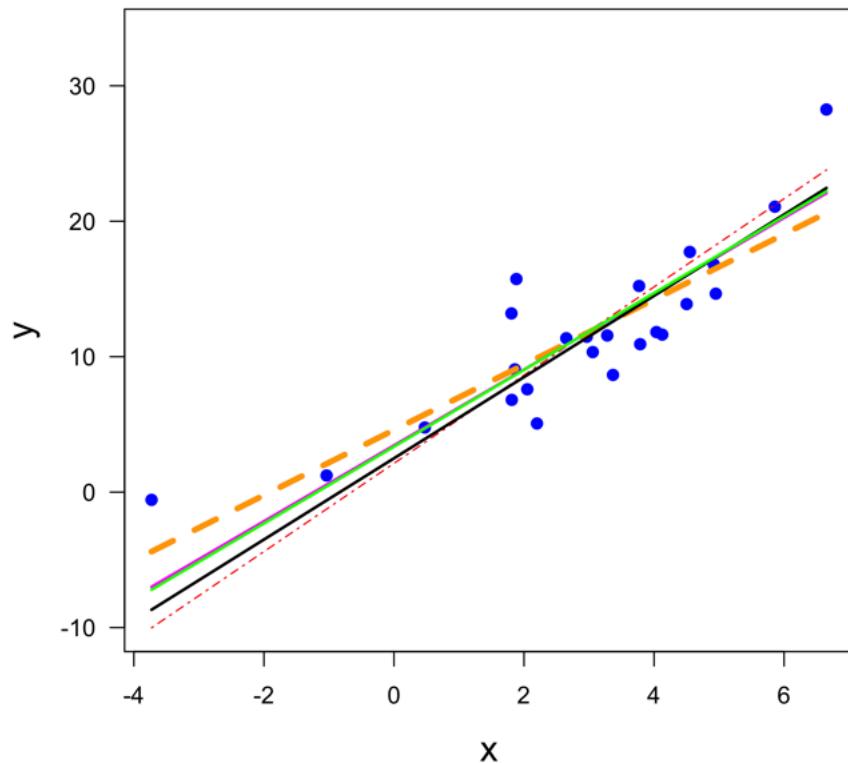


Figure K.13: **Data (blue dots) and estimates of straight lines for data set 2:** true line (black solid line), regression of  $y$  on  $x$  (orange dashed line), regression via  $x$  on  $y$  (red dash-dotted line), geometric line (magenta solid line), MCMC (green solid line). Estimates of slopes and intercepts for all methods applied and all for data sets are listed in Tables K.5 – K.6. The lines based on regression of  $y$  on  $x$  and on MCMC are closest to the true line. R code: [MCMCartificial2504.R](#) (set: `DataSet = 2` and `sflag = 1`)

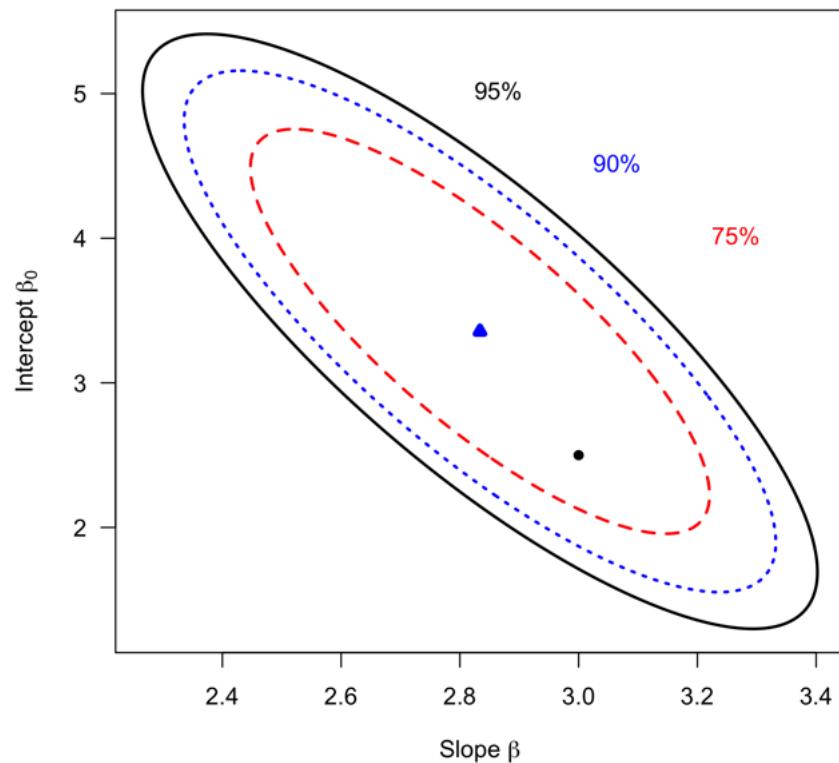


Figure K.14: **Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels for data set 2.** The estimated values  $(\hat{\beta}_{MCMC}, \hat{\beta}_0_{MCMC})$  are indicated by the blue triangle in the center of the confidence ellipses. The true values  $(\beta, \beta_0) = (3, 2.5)$  are indicated by the black dot.

R code: [MCMCartificial2504.R](#) (set: `DataSet = 2` and `sflag = 3`)

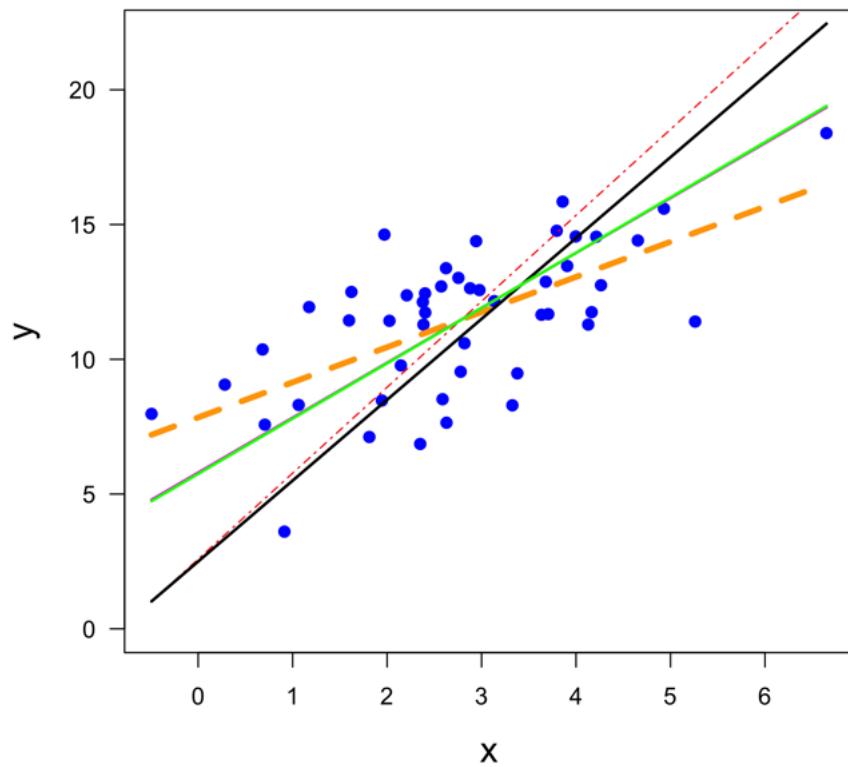


Figure K.15: **Data (blue dots) and estimates of straight lines for data set 3:** true line (black solid line,  $\beta = 3$ ,  $\beta_0 = 2.5$ ), regression of  $y$  on  $x$  (orange dashed line), regression via  $x$  on  $y$  (red dash-dotted line), geometric line (magenta solid line), MCMC (green solid line). Estimates of slopes and intercepts for all methods applied and all for data sets are listed in Tables K.5 – K.6. R code: [MCMCartificial2504.R](#) (set: `DataSet = 3` and `sflag = 1`)

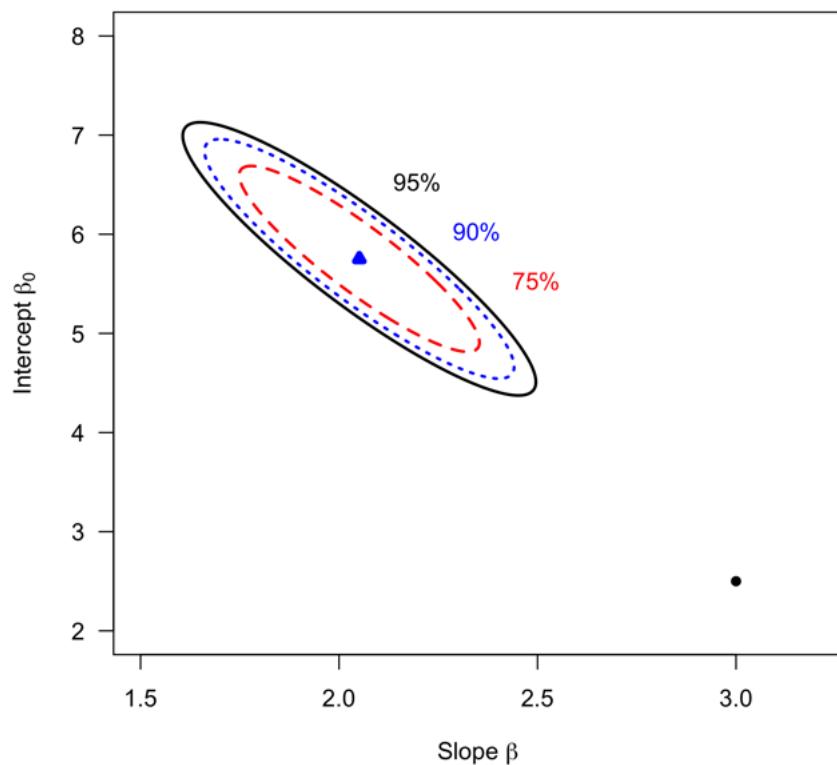


Figure K.16: **Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels for data set 3.** The estimated values ( $\hat{\beta}_{\text{MCMC}}, \hat{\beta}_0, \text{MCMC}$ ) are indicated by the blue triangle in the center of the confidence ellipses. The true values  $(\beta, \beta_0) = (3, 2.5)$  are indicated by the black dot.

R code: [MCMCartificial2504.R](#) (set: `DataSet = 3` and `sflag = 3`)

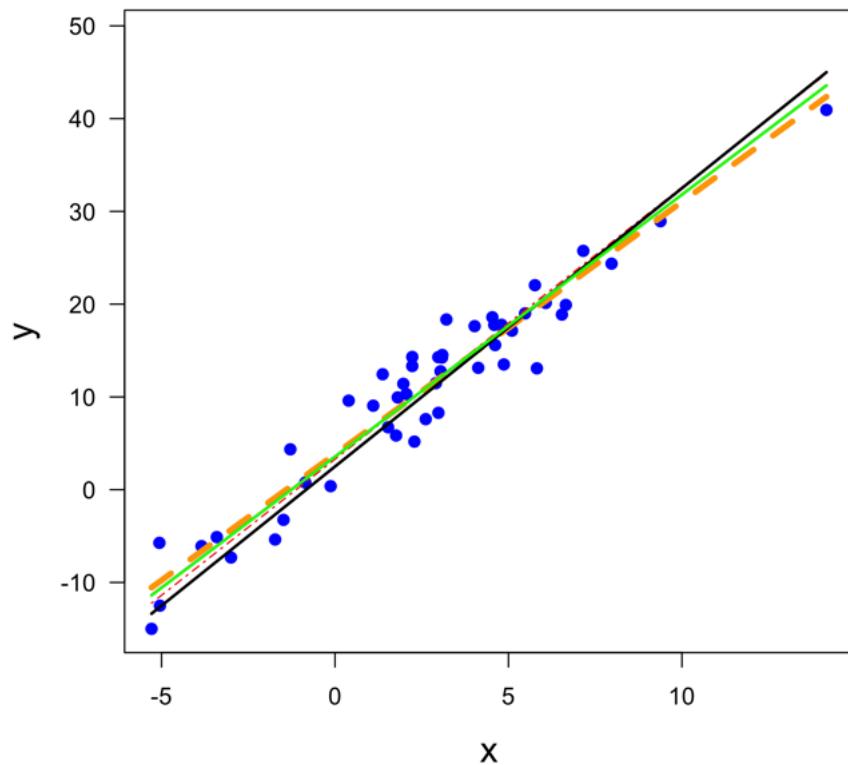
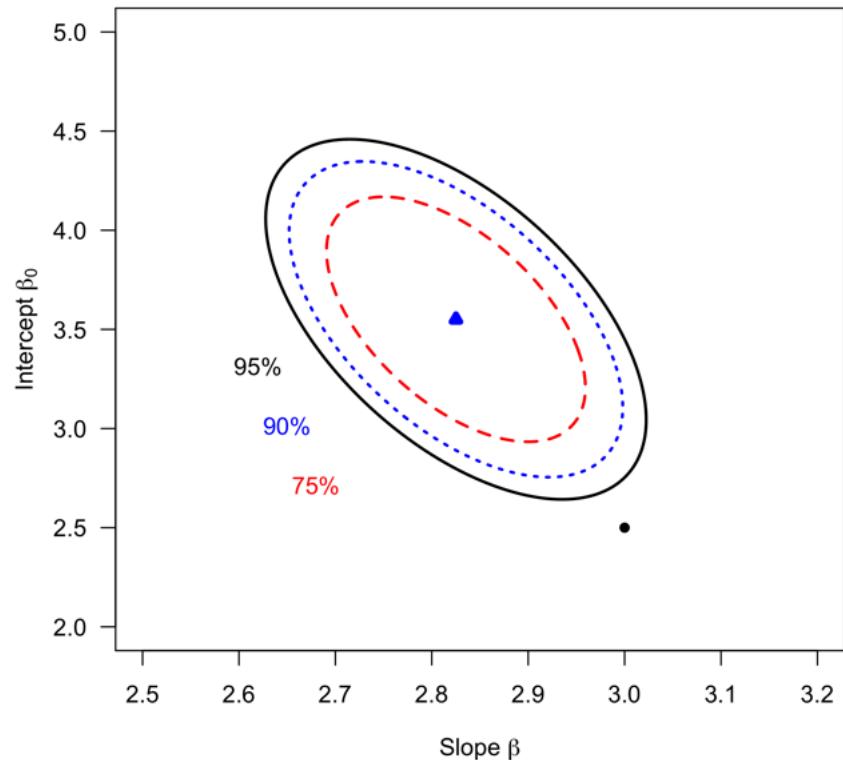


Figure K.17: **Data (blue dots) and estimates of straight lines for data set 4:** true line (black solid line), regression of  $y$  on  $x$  (orange dashed line), regression via  $x$  on  $y$  (red dash-dotted line), geometric line (magenta solid line), MCMC (green solid line). Estimates of slopes and intercepts for all methods applied and all for data sets are listed in Tables K.5 – K.6. R code: [MCMCartificial2504.R](#) (set: `DataSet = 4` and `sflag = 1`)



**Figure K.18: Joint confidence regions for slope,  $\beta$ , and intercept,  $\beta_0$ , at 95%, 90% and 75% probability levels for data set 4.** The estimated values ( $\hat{\beta}_{\text{MCMC}}, \hat{\beta}_{0,\text{MCMC}}$ ) are indicated by the blue triangle in the center of the confidence ellipses. The true values  $(\beta, \beta_0) = (3, 2.5)$  are indicated by the black dot. The uncertainty estimates of slope and intercept for data set 4 seem to be too small and thus the black point indicating the true values  $(\beta = 3, \beta_0 = 2.5)$  lies outside the 95% confidence ellipse despite the best (together with data set 2) MCMC slope estimate.

R code: [MCMCartificial2504.R](#) (set: DataSet = 4 and sflag = 3)

	Data set 1	Data set 2	Data set 3	Data set 4	Remarks
True slope	3.00	3.00	3.00	3.00	
Regression $y$ on $x$	$1.51 \pm 0.43$	$2.41 \pm 0.30$	$1.30 \pm 0.23$	$2.72 \pm 0.11$	<b>lm0</b>
Regression $y$ on $x$	$1.51 \pm 0.37$	$2.41 \pm 0.37$	$1.30 \pm 0.22$	$2.72 \pm 0.10$	based on Isobe et al. (1990)
Regression via $x$ on $y$	$4.20 \pm 1.19$	$3.26 \pm 0.42$	$3.19 \pm 0.58$	$2.93 \pm 0.13$	<b>lm0</b>
Regression via $x$ on $y$	$4.20 \pm 1.37$	$3.26 \pm 0.55$	$3.19 \pm 0.56$	$2.93 \pm 0.11$	based on Isobe et al. (1990)
Bisection line	$2.31 \pm 0.47$	$2.78 \pm 0.42$	$1.92 \pm 0.21$	$2.82 \pm 0.10$	based on Isobe et al. (1990)
Geometric line	$2.52 \pm 0.57$	$2.80 \pm 0.42$	$2.04 \pm 0.23$	$2.82 \pm 0.10$	based on Isobe et al. (1990)
MCMC	$2.60 \pm 0.36$	$2.83 \pm 0.23$	$2.05 \pm 0.18$	$2.83 \pm 0.08$	

Table K.5: Slopes &amp; slope estimates for four different data sets and various estimation methods

	Data set 1	Data set 2	Data set 3	Data set 4	Remarks
True intercept	2.50	2.50	2.50	2.50	
Regression $y$ on $x$	$7.12 \pm 1.35$	$4.59 \pm 1.10$	$7.84 \pm 0.72$	$3.83 \pm 0.52$	<b>lm0</b>
Regression $y$ on $x$	$7.12 \pm 1.23$	$4.59 \pm 1.15$	$7.84 \pm 0.72$	$3.83 \pm 0.55$	based on Isobe et al. (1990)
Regression via $x$ on $y$	$-0.55 \pm 3.52$	$2.12 \pm 1.70$	$2.59 \pm 2.16$	$3.28 \pm 0.64$	<b>lm0</b>
Regression via $x$ on $y$	$-0.55 \pm 4.28$	$2.12 \pm 1.91$	$2.59 \pm 1.67$	$3.28 \pm 0.53$	based on Isobe et al. (1990)
Bisection line	$4.85 \pm 1.44$	$3.52 \pm 1.40$	$6.11 \pm 0.75$	$3.56 \pm 0.53$	based on Isobe et al. (1990)
Geometric line	$4.24 \pm 1.75$	$3.45 \pm 1.43$	$5.79 \pm 0.81$	$3.56 \pm 0.53$	based on Isobe et al. (1990)
MCMC	$4.01 \pm 1.14$	$3.36 \pm 0.84$	$5.75 \pm 0.57$	$3.55 \pm 0.37$	

Table K.6: Intercepts &amp; intercept estimates for four different data sets and various estimation methods

### K.3.3 Discussion of MCMC results and summary

The estimated slopes and intercepts for data set 1 are significantly different from the true value; the coefficient of determination  $R^2$  for simple linear regression of  $y$  on  $x$  is only 0.36. The main reason seems to be that the spread of the true values in  $\xi$ -direction ( $\sigma_\xi = 1$ ) is comparable to the uncertainty in  $x$  ( $\sigma_x = 1$ ). In data set 2 the spread of the  $\xi$  values has been doubled ( $\sigma_\xi = 2$ ) while all other parameters for generating artificial data have been kept. The estimated slopes (Table K.5) and lines (Fig. K.13) are much closer to the true slope and true intercept, respectively. In data set 3 we use the same parameters as in data set 1 except that we doubled the sample size: this does not lead to better estimates (Fig. K.15) supporting the insightful discussion of EIV problems by Zellner (1971). Doubling the sample size and increasing the spread in  $\xi$  ( $\sigma_\xi = 4$ ) in data set 4 leads to better estimates (Fig. K.17). The uncertainty estimates of slope and intercept for data set 4 seem to be too small and thus the black point indicating the true values ( $\beta = 3, \beta_0 = 2.5$ ) lies outside the 95% confidence ellipse (Fig. K.18) despite the best (together with data set 2) MCMC slope estimate.

**Summary** Of the methods applied here, the geometric line and the bisection line can work quite good. Estimation using MCMC yields slope estimates often close to the geometric line. A more detailed investigation of the impact of different initial values for MCMC should be performed. The goodness of slope estimates depends essentially on the spread of the  $\xi_i$  values (measured by  $\sigma_\xi$ ) compared to the uncertainty in  $x$  (measured by  $\sigma_x$ ): for  $\sigma_\xi \gg \sigma_x$  one can expect good slope estimates (this seems to be the case for the Redfield data, Section 15.1) whereas for  $\sigma_\xi \leq \sigma_x$  slope estimates can be far off from true values.

## K.4 Orthogonal linear regression

Another regression method that takes  $x$  and  $y$  data into account in a symmetric way is the orthogonal regression. The basic idea is to draw from each data point  $(x_i, y_i)$  a line that is orthogonal to the estimated line, then square the orthogonal distance between the line and the data point, add up all this squared distances, and finally minimize the sum of these squares (Fig. K.19). Although this method is easy to grasp and looks promising because of the symmetric treatment of  $x$  and  $y$ , it is known since a long time that orthogonal regression is not scale invariant. This is illustrated by a simple example where the distance,  $y$ , is measured first in meters and later scaled to centimeters: the estimated slopes using original and scaled data, -0.3903 m/s and -0.8379 m/s, respectively, differ by more than a factor 2. Thus in general orthogonal regression should not be applied; an except might be data with identical units of  $x$  and  $y$ , especially when both variables are dimensionless.

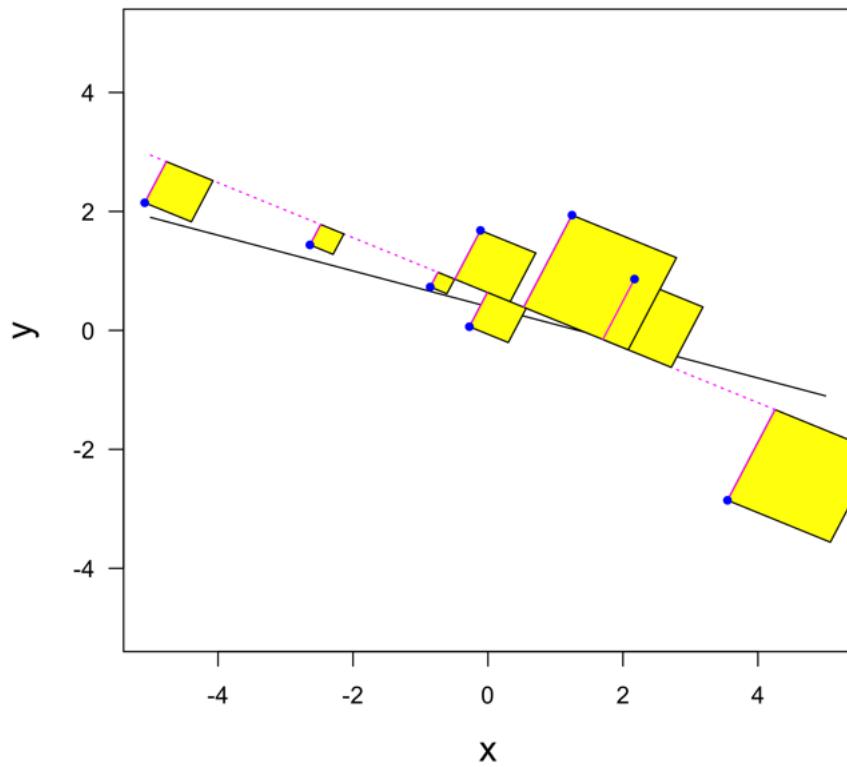


Figure K.19: The true line (black solid line) and the line estimated by orthogonal regression (magenta dashed line) which is derived by minimizing the squares of the orthogonal distances between the line and the data (blue dots), i.e. by minimizing the sum of the yellow areas.

R code: [OrthogonalRegression2504.R](#)

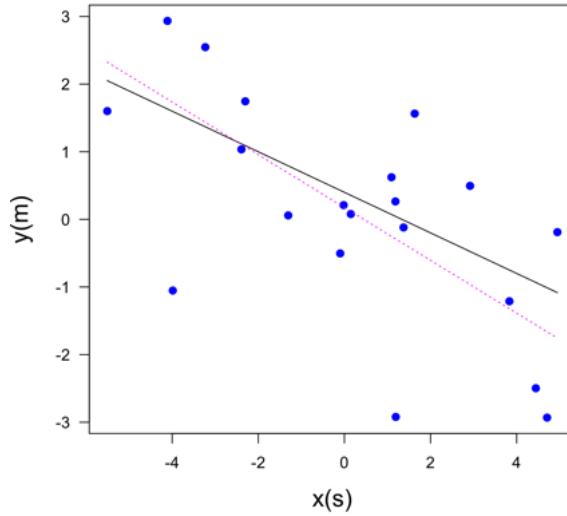


Figure K.20: Orthogonal regression of the original artificial data where  $y$  is measured in meters and  $x$  is in seconds. The estimated slope by orthogonal regression is  $\hat{\beta} = -0.39$  m/s (a velocity in minus  $y$  direction).  
**R code:** [OrthogonalRegressionScaling2504.R](#)

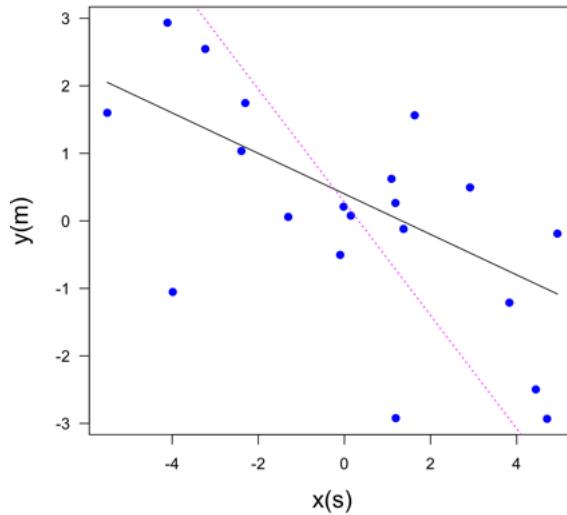


Figure K.21: Orthogonal regression of the scaled artificial data where  $y$  is measured in centimeters and  $x$ , as before, is in seconds. The estimated slope by orthogonal regression is  $\hat{\beta} = -84$  cm/s or  $-0.84$  m/s. This estimate differs by a factor of about 2 from the estimate based on the original data.  
**R code:** [OrthogonalRegressionScaling2504.R](#)

**Exercise 85 Pearson (1901): orthogonal linear regression**

Karl Pearson (1901) gave a numerical example to illustrate different slope estimation approaches: orthogonal linear regression, regression of  $y$  on  $x$ , and regression via  $x$  on  $y$ . His data read

$$x = \{0, 0.9, 1.8, 2.6, 3.3, 4.4, 5.2, 6.1, 6.5, 7.4\} \quad (\text{K.26})$$

$$x = \{5.9, 5.4, 4.4, 4.6, 3.5, 3.7, 2.8, 2.8, 2.4, 1.5\} \quad (\text{K.27})$$

Calculate the slope estimates using his data and equations. Do you obtain the same slope estimates as with the expressions from Isobe et al. (1990)? Do you obtain the same standard deviations ( $\sigma_x, \sigma_y$ ) as Pearson?

## K.5 Splitting methods

Splitting methods were proposed in the 1940ies (Wald, 1940; Bartlett, 1949) and 1950ies (Gibson & Jowett, 1957), i.e. well before the invention of personal computers and the era of big data. These ad-hoc methods are easy to grasp, however, to the best of my knowledge, they come without uncertainty estimates. Nowadays, they are more of historical interest. The numerical estimates given below are based on the Jitjareonchai et al. (2006) data Eqs. (15.6-15.7).

### K.5.1 Split up sorted data into two groups (Wald, 1940)

Wald (1940) suggested the following simple estimator for the slope:

1. Sort the data according to size of  $x$ -data  $\Rightarrow x_s$
2. split the sorted  $x$  data into two groups of equal size: small and large  $x$  data  $\Rightarrow x_1$  and  $x_2$  and the corresponding  $y$  data  $y_1$  and  $y_2$ ;
3. calculate the mean values of  $x_1, x_2, y_1, y_2 \Rightarrow \bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{y}_2$
4. estimate the slope by

$$\hat{\beta}_{\text{Wald}} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} = 2.80 \quad (\text{K.28})$$

5. and the intercept by

$$\hat{\beta}_{0,\text{Wald}} = \bar{y}_1 - \hat{\beta}_{\text{Wald}} \cdot \bar{x}_1 = 17.2 \quad (\text{K.29})$$

### K.5.2 Split up sorted data into three groups (Bartlett, 1949)

Bartlett (1949) suggested a modification of the Wald approach.

1. Sort the data according to size of  $x$ -data  $\Rightarrow x_s$
2. split the sorted  $x$  data into three groups of (approximately) equal size: small, medium, and large  $x$  data  $\Rightarrow x_1, x_2, x_3$  and the corresponding  $y$  data  $y_1, y_2, y_3$ ;
3. calculate the mean values of  $x_1, x_3, y_1, y_3 \Rightarrow \bar{x}_1, \bar{x}_3, \bar{y}_1, \bar{y}_3$
4. estimate the slope by

$$\hat{\beta}_{\text{Bartlett}} = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} = 2.68 \quad (\text{K.30})$$

5. and the intercept by

$$\hat{\beta}_{0,\text{Bartlett}} = \bar{y}_1 - \hat{\beta}_{\text{Bartlett}} \cdot \bar{x}_1 = 19.5 \quad (\text{K.31})$$

The Barlett estimates are a bit further away from the true values than the Wald estimates. A reason could be the smaller group size when splitting into more groups, especially at small sample sizes.

### K.5.3 Split up sorted data into three groups (Gibson & Jowett, 1957)

Gibson & Jowett (1957) suggested to split up the sorted  $x$  data into three groups in the group size ratio of (about) 1:2:1 and calculate the slope again from the mean values of the two outer groups (same as Bartlett, however, with smaller group sizes). For  $n = 10$  one could use the two smallest  $x$  values for  $x_1$  and the two largest values as  $x_3$  yielding

$$\hat{\beta}_{\text{Gibson}} = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} = 2.41 \quad (\text{K.32})$$

and

$$\hat{\beta}_{0,\text{Gibson}} = \bar{y}_1 - \hat{\beta}_{\text{Gibson}} \cdot \bar{x}_1 = 23.3 \quad (\text{K.33})$$

The Gibson & Jowett estimates are even worse than those of the Bartlett estimates. This might be caused by the extremely small group sizes.

### K.5.4 Application of the splitting methods to an artificial data set

In this subsection we apply the splitting methods discussed above to the first artificial data set of sample size  $n = 24$  used in Section K.3. Wald (1940) proposed to split the data set into two equally large groups with small versus large  $x$  values, calculate the centroids of these two groups, and finally put a line through these two centroids. This yields for data set 1  $\hat{\beta}_{\text{Wald}} = 1.85$  (Fig. K.22) which is close to the estimate from regression of  $y$  on  $x$  ( $1.51 \pm 0.43$ ). Slope estimates for all data sets are listed in Table K.5). Bartlett (1949) proposed splitting into three groups of equal sample size and using the centroids of the two outer groups to fix the line. This yields  $\hat{\beta}_{\text{Bartlett}} = 1.56$  (Fig. K.23). Gibson & Jowett (1957a,b) suggested a split into three groups of size ratio 1:2:1 and again taking the two outer groups to fix the line. For the data set 1 this yields  $\hat{\beta}_{\text{GJ}} = 1.61$  (Fig. K.24). Note that no uncertainty estimates are given. A comparison of the three splitting-based estimators with each other and with other estimation methods at larger sample sizes is left to the reader.

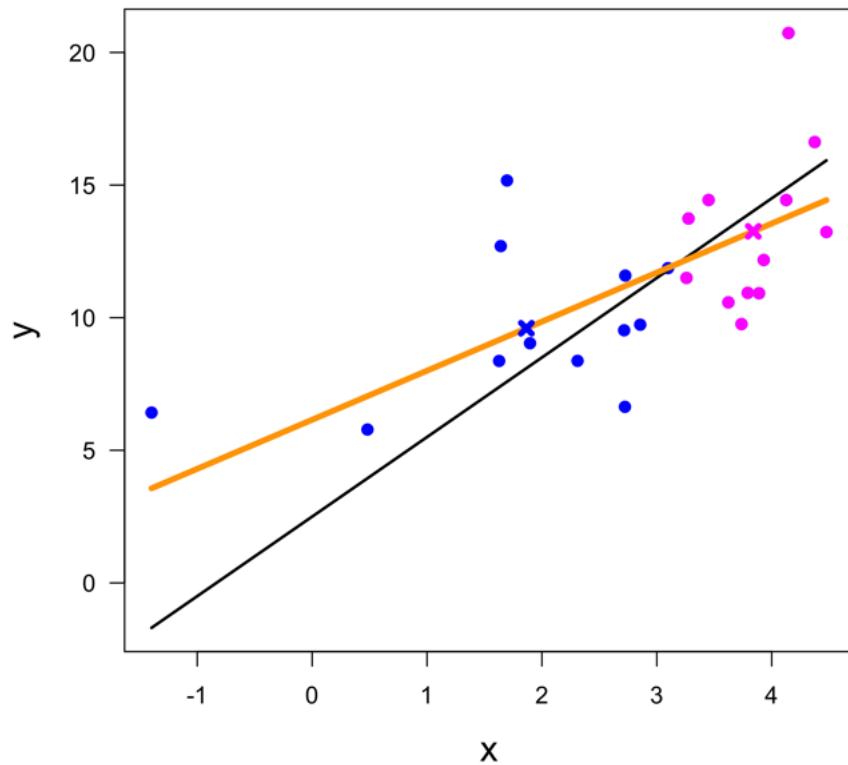


Figure K.22: Estimating the slope following Wald (1940) by splitting the data set into two groups (low  $x_i$  = blue dots; high  $x_i$  = magenta dots) and drawing a straight line through the centroids of the two data groups (blue and magenta crosses, respectively). The estimate  $\hat{\beta}_{\text{Wald}} = 1.85$  is smaller than the true value  $\beta = 3.00$ .

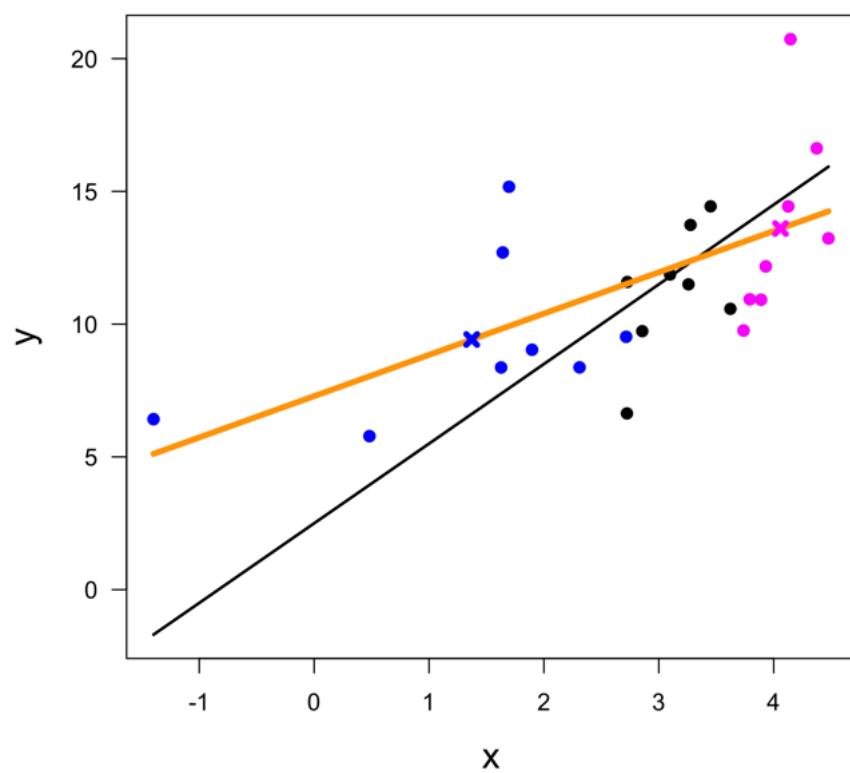


Figure K.23: Estimating the slope following the splitting method proposed by Bartlett (1949).

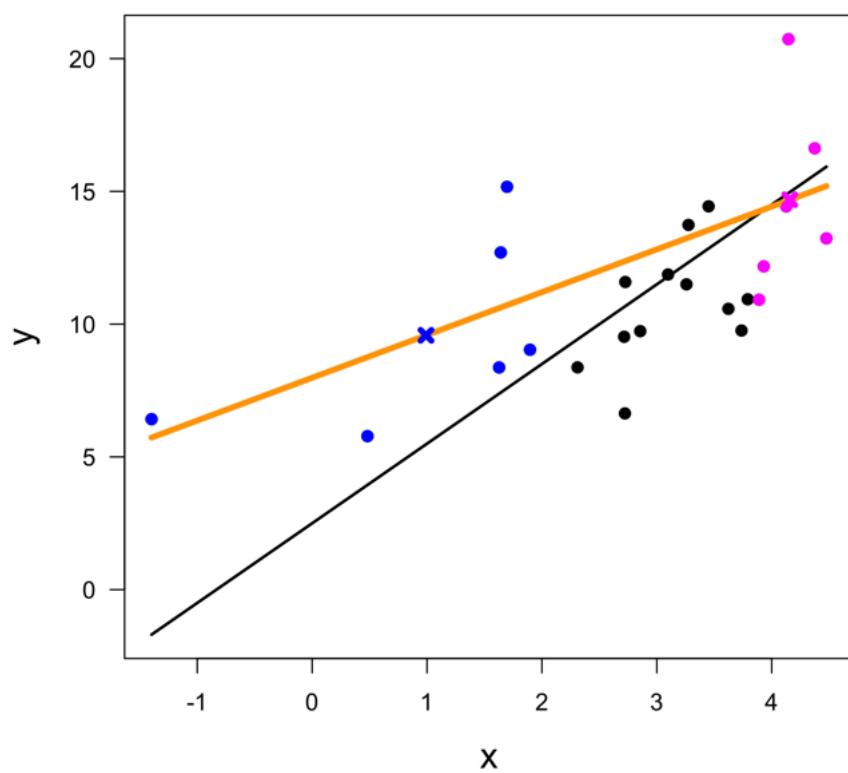


Figure K.24: Estimating the slope following the splitting method proposed by Gibson & Jowett (1957a,b).

# Appendix L

## Collinearity

### L.1 From slopes for scaled data to intercept and slopes for the original data (\*)

*Multilinear regression (MLR) has been applied to both the original data and the unit length scaled data. The question arises how these two sets of model parameters are related to each other. For applications it is most important to derive the intercept and the slopes for the original data from the slopes of the scaled data. This rather technical question is addressed in the current section.*

The intercept,  $\hat{\beta}_{0,o}$ , and the slopes,  $\hat{\beta}_{j,o}$ , for the original data  $X_j$  will be calculated from the slope estimates  $\hat{\beta}_{j,s}$  for unit length scaled predictors  $x_j$ . On the right-hand-side of Eq. 17.24 there are terms of the form

$$\hat{\beta}_{j,s} x_j \quad (\text{L.1})$$

that one would like to convert to

$$\hat{\beta}_{j,o} X_j. \quad (\text{L.2})$$

The scaled predictors  $x_j$  and the original predictors  $X_j$  are related by

$$x_j = \frac{X_j - \bar{X}_j}{S_{X_j}} \quad (\text{L.3})$$

(Eqs. 17.15 and 17.16) and thus

$$\hat{\beta}_{j,s} x_j = \hat{\beta}_{j,s} \frac{X_j - \bar{X}_j}{S_{X_j}} = \hat{\beta}_{j,o} X_j + c_j \quad (\text{L.4})$$

with

$$\hat{\beta}_{j,o} = \frac{\hat{\beta}_{j,s}}{S_{X_j}} \quad (\text{L.5})$$

$$c_j = \frac{-\hat{\beta}_{j,s} \bar{X}_j}{S_{X_j}} \quad (\text{L.6})$$

On the left-hand-side of Eq. 17.24 one would like to convert the scaled response  $y$  to the original response  $Y$ :<sup>1</sup>

$$y = \frac{Y - \bar{Y}}{S_Y} \quad (\text{L.7})$$

---

<sup>1</sup>The scaling factor  $S_Y$  could be defined analogously to the one for  $X_j$ , i.e.  $S_Y = \sqrt{\sum(Y - \bar{Y})^2}$ . Actually scaling of  $Y$  is not required for ridge regression and thus  $S_Y = 1$  works fine.

and thus

$$Y = S_Y y + \bar{Y} \quad (\text{L.8})$$

Finally one obtains the following equation for the predicted response values  $Y_p$

$$Y_p = S_Y y + \bar{Y} = \bar{Y} + S_Y \sum_k (\hat{b}_{j,o} X_j + c_j) = \hat{\beta}_{0,o} + \sum_j \hat{\beta}_{j,o} X_j \quad (\text{L.9})$$

with

$$\hat{\beta}_{0,o} = \bar{Y} + S_Y \sum_k c_j = \bar{Y} + \sum_j \frac{-S_Y \hat{\beta}_{j,s} \bar{X}_j}{S_{X_j}} \quad (\text{L.10})$$

$$\hat{\beta}_{j,o} = S_Y \hat{b}_{j,o} = \frac{S_Y \hat{\beta}_{j,s}}{S_{X_j}} \quad (\text{L.11})$$

The values of the intercept  $\hat{\beta}_{0,o}$  and the slopes  $\hat{\beta}_{j,o}$  are identical to the values derived by MLR of the original data (Table 17.8). This ‘un-scaling’ of the model parameters looks like an unnecessary detour, however, it will be essential for the ridge regression discussed in the next section.

## L.2 Acetylene data: Montgomery & Peck (1982) choice of predictors

*Ridge regression of the acetylene data has been discussed by various authors (including Marquardt & Snee, 1975; Snee, 1977; Smith & Campbell, 1980; Montgomery & Peck, 1982; Charnes et al., 1986). Unfortunately, most of them use a particular choice of predictors (they first scale the three observed predictors, then generate an extended set of predictors by forming quadratic terms from the three scaled predictors, and finally scale the quadratic terms) which makes ‘un-scaling’ of the slopes a bit tedious. Several numerical values in some of the articles differ from values obtained by using R codes; the reasons for this deviations are not known (rounding errors, typos?). In the main text we followed a somewhat different choice of predictors in that we first form quadratic terms from the original data and then scale all predictors. In the current section we follow Montgomery & Peck (1982) in order to allow comparison with literature values.*

Montgomery & Peck (1982) apply unit length scaling to quadratic terms formed from unit length scaled predictors  $x_1$ ,  $x_2$  and  $x_3$ , i.e. for example

$$x_{4,\text{MP82}} = \text{scaled}[x_1 \cdot x_2] = \text{scaled}[\text{scaled}(X_1) \cdot \text{scaled}(X_2)]. \quad (\text{L.12})$$

In general,

$$x_{4,\text{MP82}} = \text{scaled}[\text{scaled}(X_1) \cdot \text{scaled}(X_2)] \neq \text{scaled}[X_1 \cdot X_2] = x_4 \quad (\text{L.13})$$

and thus the unit length scaled predictors  $x_{j,\text{MP82}}$ ,  $j = 4, 5, \dots, 9$  look different from those in Table 17.3. Calculating the estimated slopes for the original predictors as, for example,  $X_1 \cdot X_2$  is tedious for the Montgomery & Peck scaling (see below), however, yields the same results as our scaling applied in Section 17.2.3.

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$
0.28022	-0.22544	-0.23106	-0.33783	-0.02098	0.30974	0.07828	-0.04091	-0.03436
0.28022	-0.15704	-0.23106	-0.25390	-0.02098	0.23679	0.07828	-0.13259	-0.03436
0.28022	-0.06584	-0.23514	-0.14199	-0.02592	0.14076	0.07828	-0.20382	-0.02719
0.28022	0.04817	-0.22289	-0.00210	-0.01111	0.01975	0.07828	-0.21088	-0.04833
0.28022	0.20777	-0.21881	0.19375	-0.00617	-0.14055	0.07828	-0.06774	-0.05512
0.28022	0.48138	-0.23106	0.52948	-0.02098	-0.44410	0.07828	0.59301	-0.03436
0.04003	-0.32577	-0.00255	-0.00410	0.25902	0.07314	-0.29746	0.15287	-0.23562
-0.04003	-0.22544	-0.01887	-0.02168	0.26184	0.08894	-0.29746	-0.04091	-0.23430
-0.04003	-0.06584	-0.06784	-0.04966	0.27030	0.08992	-0.29746	-0.20382	-0.21829
-0.04003	0.04817	-0.11680	-0.06964	0.27876	0.04334	-0.29746	-0.21088	-0.18421
-0.04003	0.20777	-0.05152	-0.09762	0.26748	0.01990	-0.29746	-0.06774	-0.22564
-0.04003	0.48138	0.00561	-0.14558	0.25761	0.08177	-0.29746	0.59301	-0.23552
-0.36029	-0.32577	0.35653	0.45274	-0.29604	-0.46680	0.32877	0.15287	0.24361
-0.36029	-0.22544	0.47078	0.29447	-0.47378	-0.42060	0.32877	-0.04091	0.59999
-0.36029	-0.06584	0.42182	0.04267	-0.39761	-0.05888	0.32877	-0.20382	0.43520
-0.36029	0.20777	0.37285	-0.38899	-0.32143	0.42688	0.32877	-0.06774	0.28850

Table L.1: Matrix  $\mathbf{X}_{MP82}$  of the unit length scaled predictors where  $x_4 = \text{scaling}[\text{scaling}(X_1) \cdot \text{scaling}(X_2)] = \text{scaling}[x_1 \cdot x_2]$  etcetera (compare Table 8.3 in Montgomery & Peck, 1982). Note that in the appendix we leave out the index MP82 for Montgomery & Peck scaled predictors. Montgomery & Peck (1982) use the notation ' $x_1 x_2$ ' etc. which could be confusing because the product of two unit length scaled predictors is not unit length scaled; comparison of numerical values, however, shows that the authors refer to the unit length scaled values of the quadratic terms.

1.00000	0.22363	-0.95820	-0.13242	0.44282	0.20554	-0.27075	0.03096	-0.57679
0.22363	1.00000	-0.24023	0.03869	0.19226	-0.02307	-0.14771	0.49755	-0.22391
-0.95820	-0.24023	1.00000	0.19499	-0.66053	-0.27412	0.50096	-0.01751	0.76515
-0.13242	0.03869	0.19499	1.00000	-0.26485	-0.97448	0.24631	0.39790	0.27477
0.44282	0.19226	-0.66053	-0.26485	1.00000	0.32352	-0.97224	0.12583	-0.97217
0.20554	-0.02307	-0.27412	-0.97448	0.32352	1.00000	-0.27927	-0.37460	-0.35853
-0.27075	-0.14771	0.50096	0.24631	-0.97224	-0.27927	1.00000	-0.12359	0.89366
0.03096	0.49755	-0.01751	0.39790	0.12583	-0.37460	-0.12359	1.00000	-0.15798
-0.57679	-0.22391	0.76515	0.27477	-0.97217	-0.35853	0.89366	-0.15798	1.00000

Table L.2: The matrix  $\mathbf{X}'\mathbf{X}$  is of correlation form, i.e. symmetric matrix with ones on the diagonal and off-diagonal elements between -1 and +1 (based on  $\mathbf{X}_{MP82}$ ).

### L.3 MLR of acetylene data

Slope	Estimate	Uncertainty
$\hat{\beta}_1$	0.3365	$\pm 0.3789$
$\hat{\beta}_2$	<b>0.2335</b>	<b><math>\pm 0.0258</math></b>
$\hat{\beta}_3$	-0.6759	$\pm 0.5102$
$\hat{\beta}_4$	<b>-0.4800</b>	<b><math>\pm 0.1090</math></b>
$\hat{\beta}_5$	<b>-2.034</b>	<b><math>\pm 1.585</math></b>
$\hat{\beta}_6$	-0.2657	$\pm 0.1167$
$\hat{\beta}_7$	-0.8345	$\pm 0.8212$
$\hat{\beta}_8$	<b>-0.0904</b>	<b><math>\pm 0.0348</math></b>
$\hat{\beta}_9$	-1.001	$\pm 0.6653$

Table L.3: Estimates of slopes  $\beta_k$  and their uncertainties for the unit length scaled acetylene data (based on  $X_{MP82}$ ). The uncertainties of  $\hat{\beta}_1$ ,  $\hat{\beta}_3$ ,  $\hat{\beta}_5$ , and  $\hat{\beta}_7$  are large (same order as their magnitudes) and thus  $\beta_1$ ,  $\beta_3$ ,  $\beta_5$ , and  $\beta_7$  are not significantly different from zero. Only  $\beta_2$ ,  $\beta_4$ , and  $\beta_8$  are significantly different from zero ( $t$ -tests with  $\alpha = 0.05$ ).

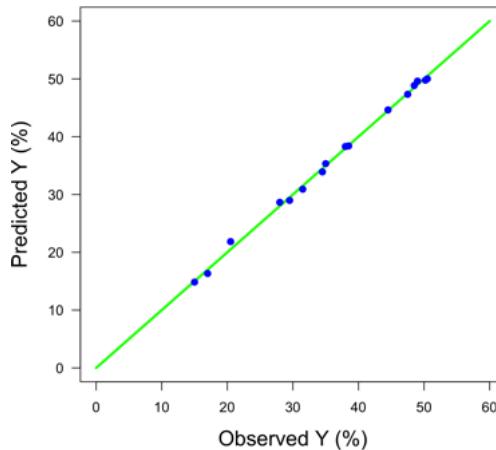


Figure L.1: Predicted response  $Y_{predicted}$  (based on MLR) over  $Y_{observed}$  (blue dots) and the 1-to-1 relationship (green line): the predicted values are very close to the observed values; coefficient of determination  $r^2 = 0.9977$  (based on  $X_{MP82}$ ). [AcetyleneULSpred.R](#)

### L.3.1 Extrapolation

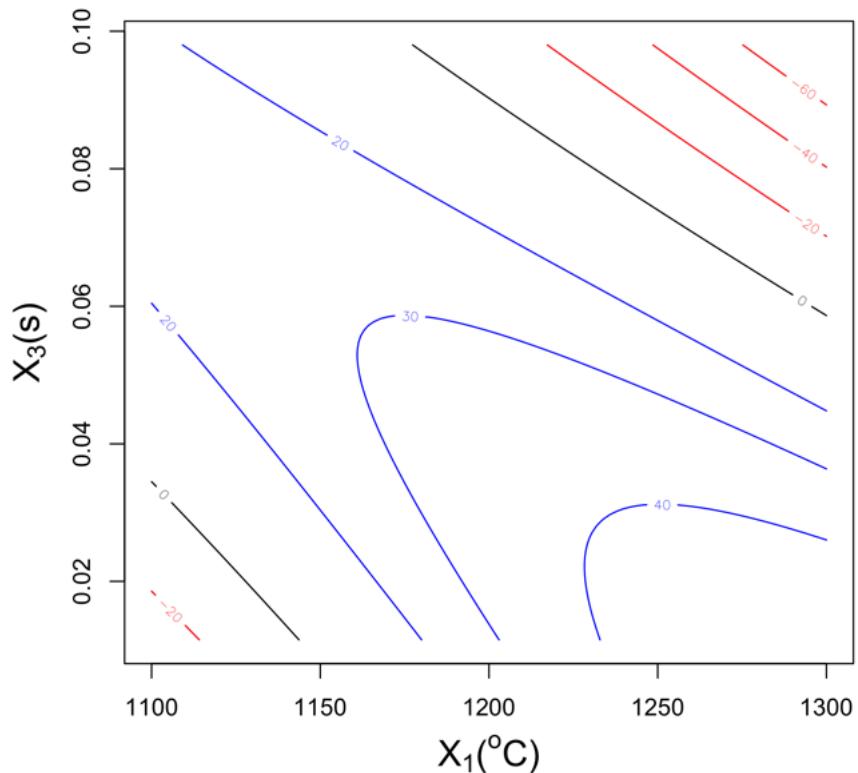


Figure L.2: Iso-contour plot for the predicted response  $Y$  (conversion of n-heptane to acetylene) for  $X_2 = 12.4 \text{ mol mol}^{-1} \text{ H}_2$  to n-heptane ratio (based on  $X_{\text{MP82}}$ ).  $X_1$  and  $X_3$  values vary between their observed minimal and maximal values. Negative response values (regions with red contour lines) make no sense because 'negative conversion' (from acetylene to n-heptane) is not possible. [AcetyleneULS-extra.R](#)

## L.4 MLR and ridge regression for Montgomery & Peck (1982) choice of predictors (\*)

**Step 1** Choose the predictor variables.

In Model 1 we will consider in addition to  $X_1$ ,  $X_2$ , and  $X_3$  the following quadratic combinations as predictors:

$$X_1 X_2, \quad X_1 X_3, \quad X_2 X_3, \quad X_1^2, \quad X_2^2, \quad X_3^2 \quad (\text{L.14})$$

where quadratic terms formed by multiplying different predictor, for example,  $X_i X_j$  with  $i \neq j$ , are called 'interaction' terms.

**Step 2** The scaling of the quadratic terms applied by Montgomery & Peck (1982) is a bit peculiar, because they do not scale the product  $X_1 X_2$  directly. Instead they first scale  $X_1$  and  $X_2$  yielding  $x_1$  and  $x_2$ , then multiply the two scaled predictors yielding  $U_4 = x_1 x_2$  (which is not scaled), and finally scale  $U_4$  yielding  $x_4$ . In short:

$$x_4 = \text{scaling}(\text{scaling}(X_1) \text{ scaling}(X_2)) \quad (\text{L.15})$$

The result of this procedure is different from

$$x_4^{\text{DWG}} = \text{scaling}(X_1 X_2) \quad (\text{L.16})$$

applied by Wolf-Gladrow (Chapter 17). Unit length scaling of the predictors  $X_1$  is defined by

$$x_1 = (X_1 - \bar{X}_1) / S_{X_1} \quad (\text{L.17})$$

where  $\bar{X}_1$  is the sample mean and  $S_{X_1}$  is the square root of the sum of squares

$$S_{X_1} = \sqrt{\sum_i (X_{1,i} - \bar{X}_1)^2}. \quad (\text{L.18})$$

The other predictors ( $X_2$ ,  $X_3$ ,  $U_4 = x_1 x_2, \dots$ ) and the response variable  $Y$  are unit length scaled as well. The scaled predictors are lumped together in the scaled predictor matrix  $\mathbf{X}$  (Table L.1). As a consequence of unit length scaling, the matrix  $\mathbf{X}'\mathbf{X}$  (Table L.2) looks like a correlation matrix, i.e. it is symmetric, has ones on the diagonal, and the off-diagonal elements are in the range  $-1$  to  $+1$ .

**Step 3** Set up the model:

The scaled response variable  $y$  is related to the scaled predictors  $x_j$  by

$$y = \beta_{1,s} x_1 + \beta_{2,s} x_2 + \dots + \beta_{9,s} x_9 + \epsilon \quad (\text{L.19})$$

where  $\beta_{j,s}$  are the model parameters (in the current context called 'slopes') and  $\epsilon$  is additive normal noise with zero mean and (unknown) variance  $\sigma^2$ . This can be written in more compact matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \quad (\text{L.20})$$

where the columns of the matrix  $\mathbf{X}$  consists of the scaled predictors  $x_1, \dots, x_9$ .

**Step 4** MLR = least-squares solution:

Multiplication of Eq. L.20 from left by the transpose of  $\mathbf{X}$ ,  $\mathbf{X}'$ , yields

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}_s + \mathbf{X}'\epsilon \quad (\text{L.21})$$

It can be shown that

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (\text{L.22})$$

yields a solution in the least-squares sense.  $\mathbf{X}'\mathbf{X}$  has the form of a correlation matrix, i.e. ones on the diagonal and off-diagonal elements in the range  $-1$  to  $+1$ . It is difficult to invert when the magnitude of off-diagonal elements becomes large, i.e.  $|XX_{i,j}|$  close to 1 for  $i \neq j$ . This is the case when some of the predictors are strongly

correlated or anti-correlated (collinearity problem). Even if the numerical evaluation yield values for  $\hat{\beta}_s$  that fit the observed data quite well, the solution is often not satisfactory at all because the length of the vector  $\hat{\beta}_s$  is large, leading to large terms in Eq. L.20 that cancel each other in order to yield a good fit. Addition of a few new data will often change  $\hat{\beta}_c$  dramatically and thus the interpretation of single  $\hat{\beta}_{j,s}$  values as changes due to variations in the corresponding predictor becomes meaningless (attribution of response to certain causes not possible; attribution problem).

**Step 5** Convert solution values  $\hat{\beta}_{j,s}$  derived for scaled predictors and response to coefficients applicable to original data, i.e. the goal is to find coefficients  $\hat{\beta}_{j,o}$  that can be used for predicting the original (not scaled) response  $Y$  from the original predictors  $X_1, X_2, X_3, X_4 = X_1 X_2$  etc.:

$$Y_{\text{pred}} = \hat{\beta}_{0,o} + \hat{\beta}_{1,o} X_1 + \hat{\beta}_{2,o} X_2 + \hat{\beta}_{3,o} X_3 + \hat{\beta}_{4,o} X_1 X_2 + \hat{\beta}_{5,o} X_1 X_3 + \hat{\beta}_{6,o} X_2 X_3 + \hat{\beta}_{7,o} X_1^2 + \hat{\beta}_{8,o} X_2^2 + \hat{\beta}_{9,o} X_3^2. \quad (\text{L.23})$$

where  $\hat{\beta}_{0,o}$  is the intercept.

On the right-hand-side of Eq. L.20 we have terms of the form

$$\hat{\beta}_{j,s} x_j \quad (\text{L.24})$$

which we would like to convert to

$$\hat{\beta}_{j,o} X_j. \quad (\text{L.25})$$

The scaled predictor  $x_1$  and the original predictor  $X_1$  are related by

$$x_1 = \frac{X_1 - \bar{X}_1}{S_{X_1}} \quad (\text{L.26})$$

and thus

$$\hat{\beta}_{1,s} x_1 = \hat{\beta}_{1,s} \frac{X_1 - \bar{X}_1}{S_{X_1}} = \hat{b}_1 X_1 + c_1 \quad (\text{L.27})$$

with

$$\hat{b}_1 = \frac{\hat{\beta}_{1,s}}{S_{X_1}} \quad (\text{L.28})$$

$$c_1 = \frac{-\hat{\beta}_{1,s} \bar{X}_1}{S_{X_1}} \quad (\text{L.29})$$

The same 'unscaling' applies for the second and third predictors.

The conversion for the other terms is a bit tedious:

$$\hat{\beta}_{4,s} x_4 = \hat{\beta}_{4,s} \text{ scaling}(x_1 x_2) = \hat{\beta}_{4,s} \frac{x_1 x_2 - \bar{x}_1 \bar{x}_2}{S_{x_1 x_2}} \quad (\text{L.30})$$

$$= \hat{\beta}_4 \frac{(X_1 - \bar{X}_1)(X_2 - \bar{X}_2) - (\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)}{S_{X_1} S_{X_2} S_{x_1 x_2}} \quad (\text{L.31})$$

$$= \hat{\beta}_{4,s} \frac{X_1 X_2 - X_1 \bar{X}_2 - X_2 \bar{X}_1 + 2 \bar{X}_1 \bar{X}_2 - \bar{X}_1 \bar{X}_2}{S_{X_1} S_{X_2} S_{x_1 x_2}} \quad (\text{L.32})$$

I.e.  $\hat{\beta}_{4,s} x_4$  not only determines

$$\hat{b}_4 = \frac{\hat{\beta}_{4,s}}{S_{X_1} S_{X_2} S_{x_1 x_2}} \quad (\text{L.33})$$

but also impacts  $\hat{b}_1$  and  $\hat{b}_2$ .

The same 'un-scaling' applies for  $\hat{\beta}_{5,s} x_5$  and  $\hat{\beta}_{6,s} x_6$  yielding

$$\hat{\beta}_{5,s} x_5 = \hat{\beta}_{5,s} \frac{X_1 X_3 - X_1 \bar{X}_3 - X_3 \bar{X}_1 + 2 \bar{X}_1 \bar{X}_3 - \bar{X}_1 \bar{X}_3}{S_{X_1} S_{X_3} S_{x_1 x_3}} \quad (\text{L.34})$$

$$\hat{\beta}_6 x_6 = \hat{\beta}_{6,s} \frac{X_2 X_3 - X_2 \bar{X}_3 - X_3 \bar{X}_2 + 2 \bar{X}_2 \bar{X}_3 - \bar{X}_2 \bar{X}_3}{S_{X_2} S_{X_3} S_{x_2 x_3}} \quad (\text{L.35})$$

and

$$\hat{b}_5 = \frac{\hat{\beta}_{5,s}}{S_{\mathbf{X}_1} S_{\mathbf{X}_3} S_{\mathbf{x}_1 \mathbf{x}_3}} \quad (\text{L.36})$$

$$\hat{b}_6 = \frac{\hat{\beta}_{6,s}}{S_{\mathbf{X}_2} S_{\mathbf{X}_3} S_{\mathbf{x}_2 \mathbf{x}_3}} \quad (\text{L.37})$$

plus terms impacting  $\hat{b}_1, \hat{b}_2$ , and  $\hat{b}_3$ .

The 'un-scaling' for the  $\hat{\beta}_{7,s} \mathbf{x}_7$  is given by

$$\hat{\beta}_{7,s} \mathbf{x}_7 = \hat{\beta}_{7,s} \text{scaling}(\mathbf{x}_1^2) = \hat{\beta}_{7,s} \frac{\mathbf{x}_1^2 - \bar{\mathbf{x}}_1^2}{S_{\mathbf{x}_1^2}} \quad (\text{L.38})$$

$$= \hat{\beta}_{7,s} \frac{(\mathbf{X}_1 - \bar{\mathbf{X}}_1)^2 - (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_1)^2}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} \quad (\text{L.39})$$

$$= \hat{\beta}_{7,s} \frac{\mathbf{X}_1^2 - 2\bar{\mathbf{X}}_1 \mathbf{X}_1 + 2\bar{\mathbf{X}}_1^2 - \bar{\mathbf{X}}_1^2}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} \quad (\text{L.40})$$

and thus

$$\hat{b}_7 = \frac{\hat{\beta}_{7,s}}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} \quad (\text{L.41})$$

plus a term impacting  $\hat{b}_1$ . The same 'un-scaling' applies for  $\hat{\beta}_{8,s} \mathbf{x}_8$  and  $\hat{\beta}_{9,s} \mathbf{x}_9$  yielding

$$\hat{\beta}_{8,s} \mathbf{x}_8 = \hat{\beta}_{8,s} \frac{\mathbf{X}_2^2 - 2\bar{\mathbf{X}}_2 \mathbf{X}_2 + 2\bar{\mathbf{X}}_2^2 - \bar{\mathbf{X}}_2^2}{S_{\mathbf{X}_2}^2 S_{\mathbf{x}_2^2}} \quad (\text{L.42})$$

$$\hat{\beta}_{9,s} \mathbf{x}_9 = \hat{\beta}_{9,s} \frac{\mathbf{X}_3^2 - 2\bar{\mathbf{X}}_3 \mathbf{X}_3 + 2\bar{\mathbf{X}}_3^2 - \bar{\mathbf{X}}_3^2}{S_{\mathbf{X}_3}^2 S_{\mathbf{x}_3^2}} \quad (\text{L.43})$$

and

$$\hat{b}_8 = \frac{\hat{\beta}_{8,s}}{S_{\mathbf{X}_2}^2 S_{\mathbf{x}_2^2}} \quad (\text{L.44})$$

$$\hat{b}_9 = \frac{\hat{\beta}_{9,s}}{S_{\mathbf{X}_3}^2 S_{\mathbf{x}_3^2}} \quad (\text{L.45})$$

plus terms impacting  $\hat{b}_2$  and  $\hat{b}_3$ .

On the right-hand-side of Eq. L.20 we like to convert the scaled response  $\mathbf{y}$  to the original response  $\mathbf{Y}$ :

$$\mathbf{y} = \frac{\mathbf{Y} - \bar{\mathbf{Y}}}{S_{\mathbf{Y}}} \quad (\text{L.46})$$

$\Rightarrow$

$$\mathbf{Y} = S_{\mathbf{Y}} \mathbf{y} + \bar{\mathbf{Y}} \quad (\text{L.47})$$

Finally one obtains the following equation for the predicted response values

$$Y_p = S_{\mathbf{Y}} \mathbf{y} + \bar{\mathbf{Y}} = \bar{\mathbf{Y}} + S_{\mathbf{Y}} \sum_k \left( b \hat{B}_j X_j + c B_j \right) = \hat{\beta}_{0,o} + \sum_j \hat{\beta}_{j,o} X_j \quad (\text{L.48})$$

with

$$\hat{b}\hat{B}_1 = \frac{\hat{\beta}_{1,s}}{S_{\mathbf{X}_1}} - \hat{\beta}_{4,s} \frac{\bar{X}_2}{S_{\mathbf{X}_1} S_{\mathbf{X}_2} S_{\mathbf{x}_1 \mathbf{x}_2}} - \hat{\beta}_{5,s} \frac{\bar{X}_3}{S_{\mathbf{X}_1} S_{\mathbf{X}_3} S_{\mathbf{x}_1 \mathbf{x}_3}} - \hat{\beta}_{7,s} \frac{2 \bar{X}_1}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} \quad (\text{L.49})$$

$$\hat{b}\hat{B}_2 = \frac{\hat{\beta}_{2,s}}{S_{\mathbf{X}_2}} - \hat{\beta}_{4,s} \frac{\bar{X}_1}{S_{\mathbf{X}_1} S_{\mathbf{X}_2} S_{\mathbf{x}_1 \mathbf{x}_2}} - \hat{\beta}_{6,s} \frac{\bar{X}_3}{S_{\mathbf{X}_2} S_{\mathbf{X}_3} S_{\mathbf{x}_2 \mathbf{x}_3}} - \hat{\beta}_{8,s} \frac{2 \bar{X}_2}{S_{\mathbf{X}_2}^2 S_{\mathbf{x}_2^2}} \quad (\text{L.50})$$

$$\hat{b}\hat{B}_3 = \frac{\hat{\beta}_3}{S_{\mathbf{X}_3}} - \hat{\beta}_{5,s} \frac{\bar{X}_1}{S_{\mathbf{X}_1} S_{\mathbf{X}_3} S_{\mathbf{x}_1 \mathbf{x}_3}} - \hat{\beta}_{6,s} \frac{\bar{X}_2}{S_{\mathbf{X}_2} S_{\mathbf{X}_3} S_{\mathbf{x}_2 \mathbf{x}_3}} - \hat{\beta}_{9,s} \frac{2 \bar{X}_3}{S_{\mathbf{X}_3}^2 S_{\mathbf{x}_3^2}} \quad (\text{L.51})$$

$$\hat{b}\hat{B}_4 = \hat{b}_4 = \frac{\hat{\beta}_{4,s}}{S_{\mathbf{X}_1} S_{\mathbf{X}_2} S_{\mathbf{x}_1 \mathbf{x}_2}} \quad (\text{L.52})$$

$$\hat{b}\hat{B}_5 = \hat{b}_5 = \frac{\hat{\beta}_{5,s}}{S_{\mathbf{X}_1} S_{\mathbf{X}_3} S_{\mathbf{x}_1 \mathbf{x}_3}} \quad (\text{L.53})$$

$$\hat{b}\hat{B}_6 = \hat{b}_6 = \frac{\hat{\beta}_{6,s}}{S_{\mathbf{X}_2} S_{\mathbf{X}_3} S_{\mathbf{x}_2 \mathbf{x}_3}} \quad (\text{L.54})$$

$$\hat{b}\hat{B}_7 = \hat{b}_7 = \frac{\hat{\beta}_{7,s}}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} \quad (\text{L.55})$$

$$\hat{b}\hat{B}_8 = \hat{b}_8 = \frac{\hat{\beta}_{8,s}}{S_{\mathbf{X}_2}^2 S_{\mathbf{x}_2^2}} \quad (\text{L.56})$$

$$\hat{b}\hat{B}_9 = \hat{b}_9 = \frac{\hat{\beta}_{9,s}}{S_{\mathbf{X}_3}^2 S_{\mathbf{x}_3^2}} \quad (\text{L.57})$$

$$\hat{\beta}_{0,o} = \bar{Y} - S_Y \left( \frac{\hat{\beta}_{1,s} \bar{X}_1}{S_{\mathbf{X}_1}} + \frac{\hat{\beta}_{2,s} \bar{X}_2}{S_{\mathbf{X}_2}} + \frac{\hat{\beta}_{3,s} \bar{X}_3}{S_{\mathbf{X}_3}} - \hat{\beta}_{4,s} \frac{2 \bar{X}_1 \bar{X}_2 - \bar{X}_1 \bar{X}_2}{S_{\mathbf{X}_1} S_{\mathbf{X}_2} S_{\mathbf{x}_1 \mathbf{x}_2}} - \hat{\beta}_{5,s} \frac{2 \bar{X}_1 \bar{X}_3 - \bar{X}_1 \bar{X}_3}{S_{\mathbf{X}_1} S_{\mathbf{X}_3} S_{\mathbf{x}_1 \mathbf{x}_3}} \right. \quad (\text{L.58})$$

$$\left. - \hat{\beta}_{6,s} \frac{2 \bar{X}_2 \bar{X}_3 - \bar{X}_2 \bar{X}_3}{S_{\mathbf{X}_2} S_{\mathbf{X}_3} S_{\mathbf{x}_2 \mathbf{x}_3}} - \hat{\beta}_{7,s} \frac{2 \bar{X}_1^2 - \bar{X}_1^2}{S_{\mathbf{X}_1}^2 S_{\mathbf{x}_1^2}} - \hat{\beta}_{8,s} \frac{2 \bar{X}_2^2 - \bar{X}_2^2}{S_{\mathbf{X}_2}^2 S_{\mathbf{x}_2^2}} - \hat{\beta}_{9,s} \frac{2 \bar{X}_3^2 - \bar{X}_3^2}{S_{\mathbf{X}_3}^2 S_{\mathbf{x}_3^2}} \right) \quad (\text{L.59})$$

$$\hat{\beta}_{j,o} = S_Y \hat{b}\hat{B}_j \quad (\text{L.60})$$



# Appendix M

## Priors (appendix)

### M.1 History of priors, especially 'non-informative'

- 1814 Laplace (1814) usually assigns the constant ('flat') prior if nothing is known a-priori
- 1961 Jeffreys (1939, 1948, 1961) derives what is now called Jeffreys priors: the constant ('flat') prior for location parameters that can vary within a finite interval (say between 0 and 1, as for example for probabilities) or over the whole range from  $-\infty$  to  $\infty$  and the  $1/\theta$  prior for scale parameters that can vary between 0 and  $\infty$ . Both the prior for the scale parameter and also the prior for the location parameter applied to the infinite range are improper, i.e. they can not be normalized to 1 (the integrals yield  $\infty$ ), which might lead to improper posteriors. Despite their impropriety they often turn out to work fine.
- 1961 Raiffa & Schlaifer (1961) extensively applied 'conjugate priors'. Compare also Gelman (2020) for various applications.
- 1979 Bernardo (1979, 2005) introduces so-called 'reference priors' based on the Kullback-Leibler divergence or relative entropy. For further development compare Berger et al. (2009).
- 1985 The remark by Berger (1985) sounds like he was given up (at this moment):<sup>1</sup>  
"Perhaps Laplace was right, in a practical sense, to simply pretend that unknown parameters had constant priors." Berger (1985, p. 187)
- To appreciate the impact of a non-flat prior it is useful to calculate and analyze the posterior based on the flat prior. The same applies for scale parameters: compare the posteriors based on Jeffreys prior and reference prior when these priors are different from each other.
- 1996 Kass & Wasserman (1996): The selection of prior distributions by formal rules
- 2003 Jaynes (2003) uses transformation group methods to assign non-informative priors. He also gives nice examples for informative priors.

---

<sup>1</sup>However, see his later work especially together with J. Bernardo.

## M.2 Numerical estimation of reference priors

A method for numerical estimation of reference priors can be found in Bernardo (2005) or Berger et al. (2009): it will be given below with slightly different notation. The method is applied to two examples: (1) estimation of the rate of an exponential population and (2) estimation the mode of an asymmetric triangular population. The first example yields numerical values consistent with the Jeffreys prior for a scale parameter, whereas the second example results in a reference prior that is different from Jeffreys flat prior for a location parameter over a finite interval.

The reference prior can be estimated as follows (based on Bernardo, 2005):

1. Starting values:

Choose a moderate value  $K$  for the artificial sample sizes<sup>2</sup>.

Choose an arbitrary positive function  $h(\theta)$ , say  $h(\theta) = 1$ , i.e. the flat prior.

Choose the number  $M$  of samples to be simulated ( $M$  is the number of Monte Carlo runs).

2. The reference prior will be estimated for a number of (discrete)  $\theta$  values,  $\theta_d$ , that cover the interesting range of  $\theta$ . These values can be chosen with finer resolution in regions where the prior shows larger variations. For any given  $\theta_d$  value, **repeat** for  $j = 1, \dots, M$ :

Simulate a random sample  $\{x_1, \dots, x_K\}$  of size  $K$  from the distribution  $p(x|\theta_d)$ . For independent data  $x_i$ ,  $i = 1, \dots, K$  the joint likelihood  $L(x|\theta_d)$  is just the product  $\prod_{i=1}^K p(x_i|\theta_d)$

$$L(x|\theta_d) = \prod_{i=1}^K p(x_i|\theta_d) \quad (\text{M.1})$$

Compute numerically the integrals

$$c_j = \int_{\Theta} L(x|\theta) h(\theta) d\theta \quad (\text{M.2})$$

which, for the flat prior  $h(\theta) = 1$ , simplifies to

$$c_j = \int_{\Theta} L(x|\theta) d\theta \quad (\text{M.3})$$

where  $\Theta$  stands for the whole definition range of  $\theta$  (also called 'parameter space').

Evaluate

$$r_j(\theta_d) = \log [L(x|\theta)/c_j]. \quad (\text{M.4})$$

3. Compute the estimate of the reference prior as the exponent of the mean of all  $M r_j$  values

$$\pi(\theta_d) = \exp \{ \text{mean} [r_j(\theta_d)] \} \quad (\text{M.5})$$

and **store** the pair  $\{\theta_d, \pi(\theta_d)\}$ .

4. **Repeat** routines (2) and (3) for all  $\theta_d$  values for which the pair  $\{\theta_d, \pi(\theta_d)\}$  is required.

---

<sup>2</sup>In the examples presented in Bernardo (2005)  $K$  is in the range 500 to 2500. The value depends on the problem under consideration and the desired uncertainty.

### M.2.1 Estimation of the rate of an exponential population

The goal is to estimate the rate parameter  $\theta$  of an exponential population from a sample  $\mathbf{x} = \{x_1, \dots, x_n\}$ . The exponential PDF reads (Section C.3.10)

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0 \quad (\text{M.6})$$

where  $\theta > 0$  is the rate. Thus  $\theta$  is a scale parameter: can vary between 0 and  $\infty$ . In the Bayesian approach to parameter estimation we have to assign a prior. If we don't know nothing about  $\theta$  before looking at the data (a-priori), we assign a non-informative prior. Following Jeffreys (1961) the non-informative prior is  $1/\theta$ . This Jeffreys prior is improper because it cannot be normalized to 1: the integral  $\int_0^\infty 1/\theta d\theta$  yields infinity. Despite the impropriety of the Jeffreys prior we will obtain a proper posterior.

The likelihood  $L(\mathbf{x}|\theta)$  for the independent data  $x_i, i = 1, \dots, n$  reads

$$L(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i} = \theta^n e^{-\theta n \bar{x}} \quad (\text{M.7})$$

where the sample mean  $\bar{x}$  is the sufficient statistic. The likelihood function  $LF(\theta|\mathbf{x})$  reads

$$LF(\theta|\mathbf{x}) = \theta^n e^{-\theta n \bar{x}}. \quad (\text{M.8})$$

Using Jeffreys prior we obtain the posterior

$$p(\theta|\mathbf{x}) \propto LF(\theta|\mathbf{x}) \times \frac{1}{\theta} = \theta^{n-1} e^{-\theta n \bar{x}} \quad (\text{M.9})$$

which can be normalized giving

$$p(\theta|\mathbf{x}) = \theta^{n-1} e^{-\theta n \bar{x}} \frac{(n \bar{x})^n}{(n-1)!} = \frac{n^n}{(n-1)!} \bar{x}^n \theta^{n-1} e^{-\theta n \bar{x}} \quad (\text{M.10})$$

where we have used the integral for integer values of  $k \geq 0$

$$\int_0^\infty x^k e^{-ax} dx = \frac{k!}{a^{k+1}} \quad a > 0 \quad (\text{M.11})$$

with  $k = n - 1$  and  $a = \bar{x} n$ .

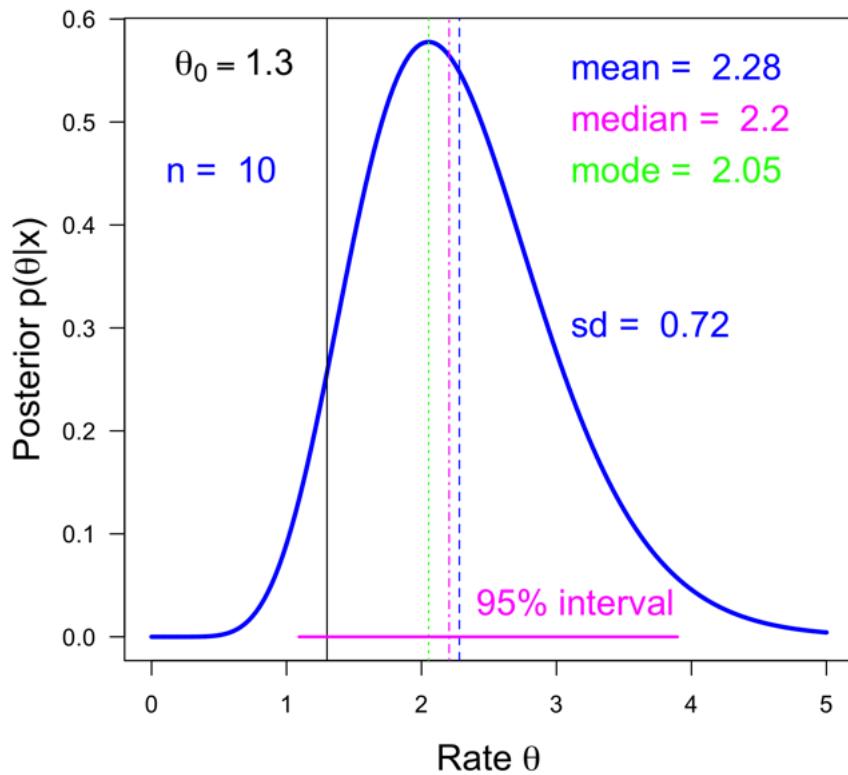


Figure M.1: Posterior  $p(\theta|x)$  (blue solid line; Eq. M.10) for a random sample of size  $n = 10$  from an exponential population with rate constant  $\theta_0 = 1.3$ . The mean, median, and mode of the posterior are all somewhat larger than the true value of the rate constant. However, for the small sample size the standard deviation  $sd = 0.72$  is quite large and the true value  $\theta_0 = 1.3$  lies within the 95% interval (magenta solid line). [ExpPostn10.R](#)

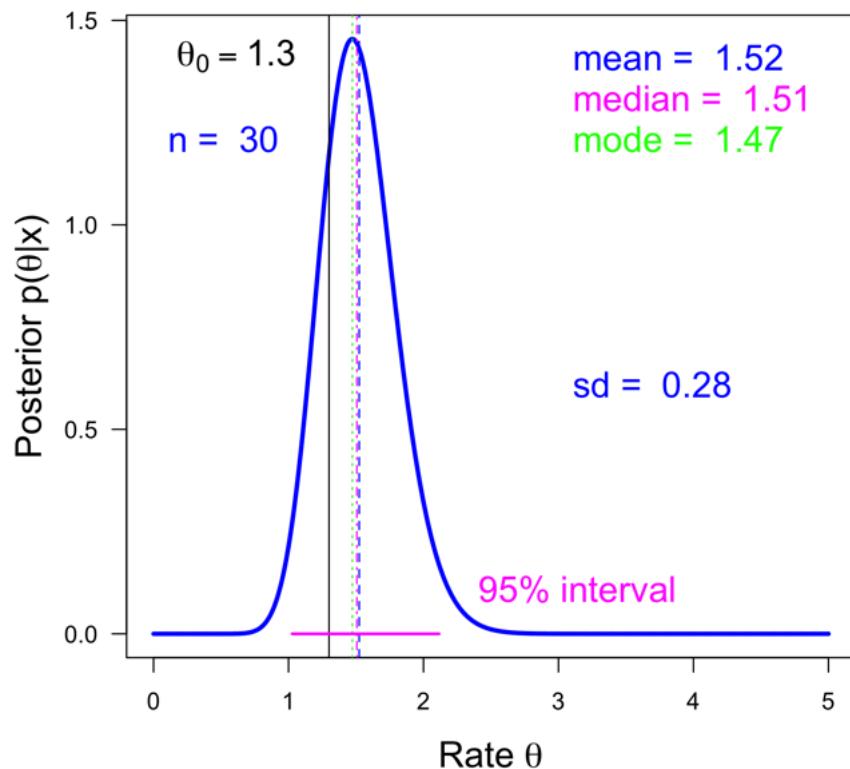


Figure M.2: Posterior  $p(\theta|x)$  (blue solid line; Eq. M.10) for a random sample of size  $n = 30$  from an exponential population with rate constant  $\theta_0 = 1.3$ . The mean, median, and mode of the posterior are all somewhat larger than the true value of the rate constant. However, for sample size  $n = 30$  the standard deviation  $sd = 0.28$  is already much smaller than for  $n = 10$  and the differences between the mean, median, or mode on the one hand and the true value  $\theta_0 = 1.3$  is less than one standard deviation ('within  $\pm 1\sigma$ '). [ExpPostn30.R](#)

## M.2.2 Numerical estimation of reference prior for exponential population

We will now apply the recipe for the estimation of a reference prior to the exponential population (Eq. M.6). The example with exponential distribution and flat trial prior  $h(\theta) = 1$  is relative easy to handle because a simple sufficient statistic, namely the sample mean, exists and the calculation of the factors  $c_j$  can be done analytically. The **joint likelihood** is given by (Eq. M.7; replace  $n$  by  $K$  in current context):

$$L(\mathbf{x}|\theta) = \prod_{i=1}^K f(x_i; \theta) = \theta^K e^{-\theta K \bar{x}}. \quad (\text{M.12})$$

The  $c_j$  are readily calculated:

$$c_j = \int_0^\infty L(\mathbf{x}|\theta) d\theta = \int_0^\infty \theta^K e^{-\theta K \bar{x}} d\theta = \frac{K!}{(K \bar{x})^{K+1}} \quad (\text{M.13})$$

where we have used again integral (Eq. M.11). The following steps, namely calculating  $r_j$ , averaging, and finally the prior estimates, is straightforward. Finally the estimated prior values are rescaled such that the prior for  $\theta = 1$  is 1 which allows comparison with the Jeffreys prior  $1/\theta$  (Fig. M.3). From this figure one can conclude: the reference prior is identical to the Jeffreys prior. Although this result is not surprising<sup>3</sup>, it shows that the numerical methods seems to work very well.

In the numerical implementation for large values of  $K$  one runs into overflow problems. Thus a relative small  $K$  of 25 has been chosen by Bernardo (2005). For  $K = 90$  we already detected overflow problems for some  $\theta_d$  values. Fig. M.3 shows numerical results for  $K = 75$  (red dots) compared to the analytical prior  $\propto 1/\theta$  (blue line). Better agreement of the numerical results with the analytical expression of the prior could be reached in principle by numerical calculations with more digits (see R library **Rmpfr** (R Multiple Precision Floating-Point Reliable)) because this would allow an increase of  $K$  values.

---

<sup>3</sup>The result of the next example will be more surprising!

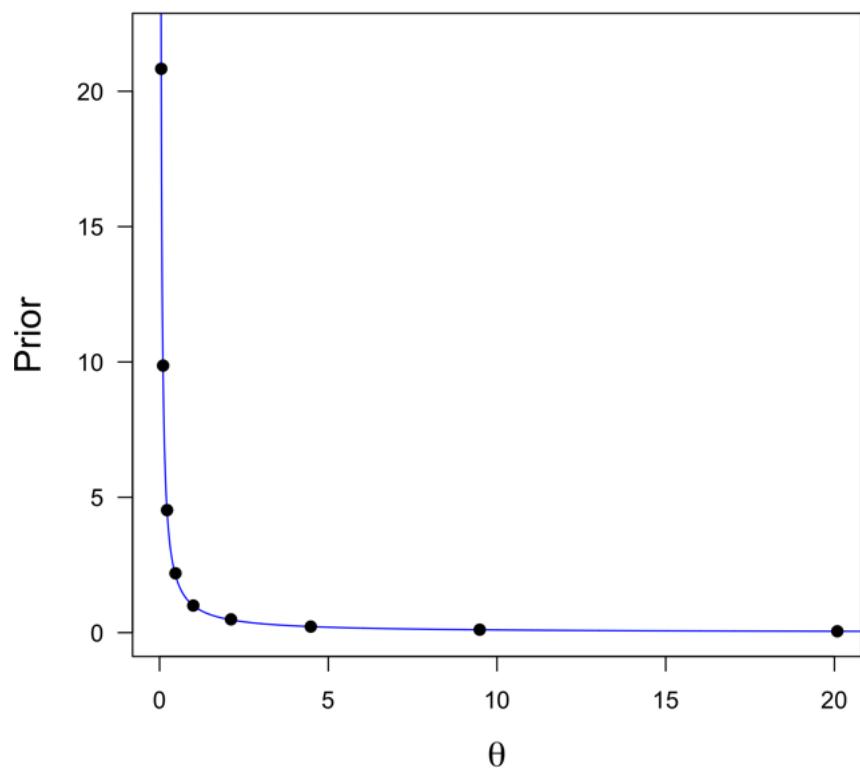


Figure M.3: Estimate of the reference prior for exponential populations: the numerically estimated values (red dots) fall on the blue line representing Jeffreys prior  $1/\theta$ . [RefPriorExpBernardo.R](#)

### M.2.3 Estimation of the mode of an asymmetric triangular distribution

*Estimation a reference prior for asymmetric triangular distributions is more difficult compared to the exponential distribution problem. Berger et al. (2009) write 'The nonsymmetric triangular distribution does not possess a useful reduced sufficient statistic. Also, although  $\log[p(x|\theta)]$  is differentiable for all  $\theta$  values, the formal Fisher information function is strictly negative, so Jeffreys prior does not exist. ... Analytical derivation of the reference prior does not seem to be feasible in this example, but there is an interesting heuristic argument which suggests that the  $Be(\theta|1/2, 1/2)$  prior is indeed the reference prior for the problem.'*

The asymmetric standard triangular distribution on  $(0, 1)$ ,

$$p(x|\theta) = \begin{cases} 2x/\theta, & \text{for } 0 < x \leq \theta, \\ 2(1-x)/(1-\theta), & \text{for } \theta < x \leq 1, \end{cases} \quad 0 < \theta < 1$$

has a unique mode at  $\theta$  (Fig. M.4).

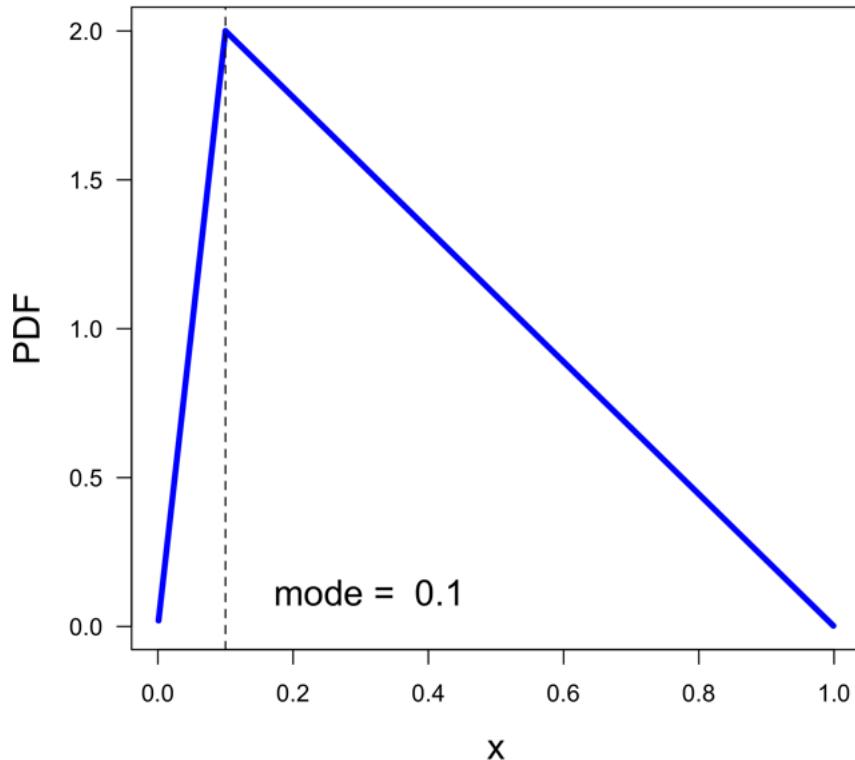


Figure M.4: The asymmetric standard triangular distribution with mode 0.1. [TriangAsymPDF.R](#)

Given a sample  $\mathbf{x} = \{x_1, \dots, x_M\}$  where all  $x_i, i = 1, \dots, M$  are independent of each other, we can write the joint likelihood as a product of the likelihoods for single data  $x_i$ :

$$L(\mathbf{x}|\theta) \propto \prod_{i=1}^M L(x_i|\theta). \quad (\text{M.14})$$

However, this expression cannot be simplified (as mentioned above: a sufficient statistic does not exist). Random numbers from the nonsymmetric triangular PDF can be generated using the R routine `rtriangle()` in the package `triangle` (Figs. M.5 – M.7); compare Exercise 87 for a more pedestrian way. For small sample sizes ( $M = 10^3$ ) the modes of estimated probability densities are shifted towards  $x = 0.5$  (Figs. M.5 and M.7). This already indicates that estimating the mode from small samples might be difficult and a non-flat prior with more weight towards extreme (small, large) values could be useful.

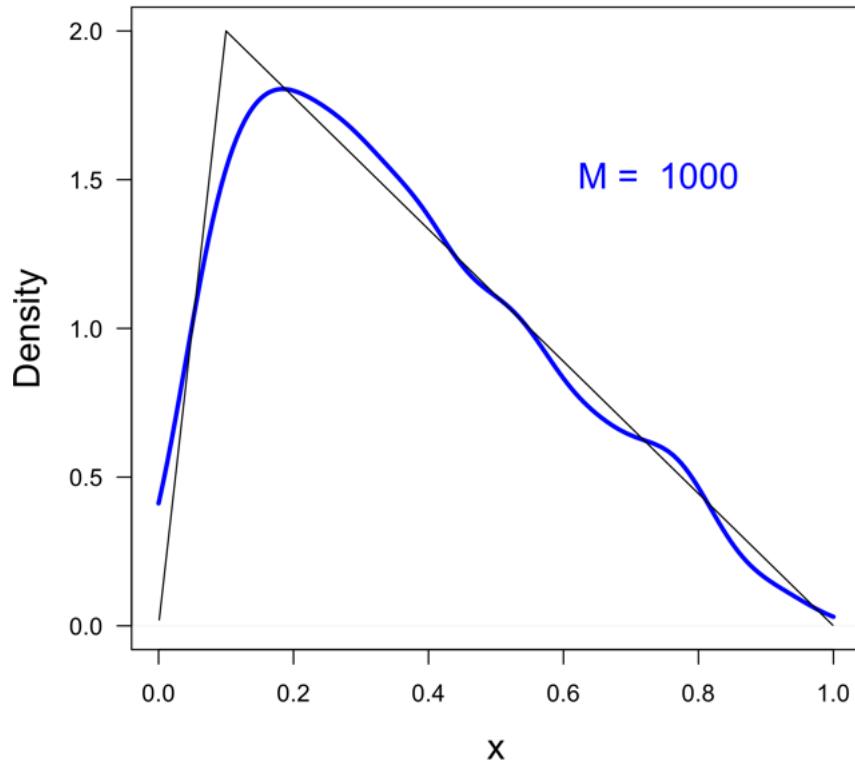


Figure M.5: Random numbers (estimated density based on  $M = 10^3$  numbers, blue thick line) from the asymmetric standard triangular distribution with mode 0.1 (black thin line). The mode of the density estimate is larger than the true mode. [TriangAsymRandom.R](#)

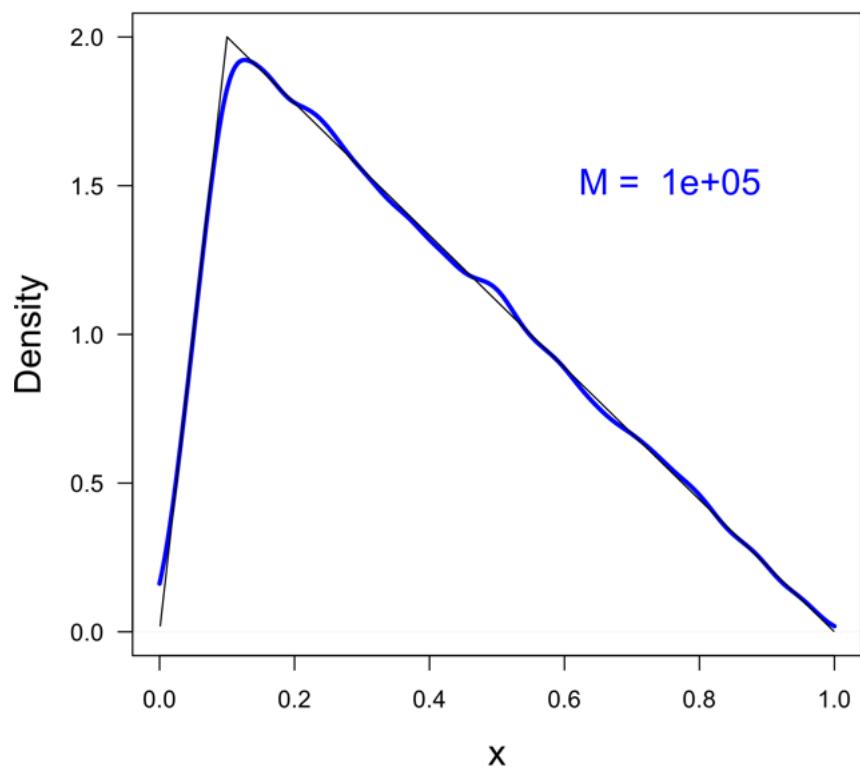


Figure M.6: Random numbers (estimated density based on  $M = 10^5$  numbers, blue thick line) from the asymmetric standard triangular distribution with mode 0.1 (black thin line). At larger sample size the mode of the density estimate is only slightly larger than the true mode. [TriangAsymRandom.R](#)

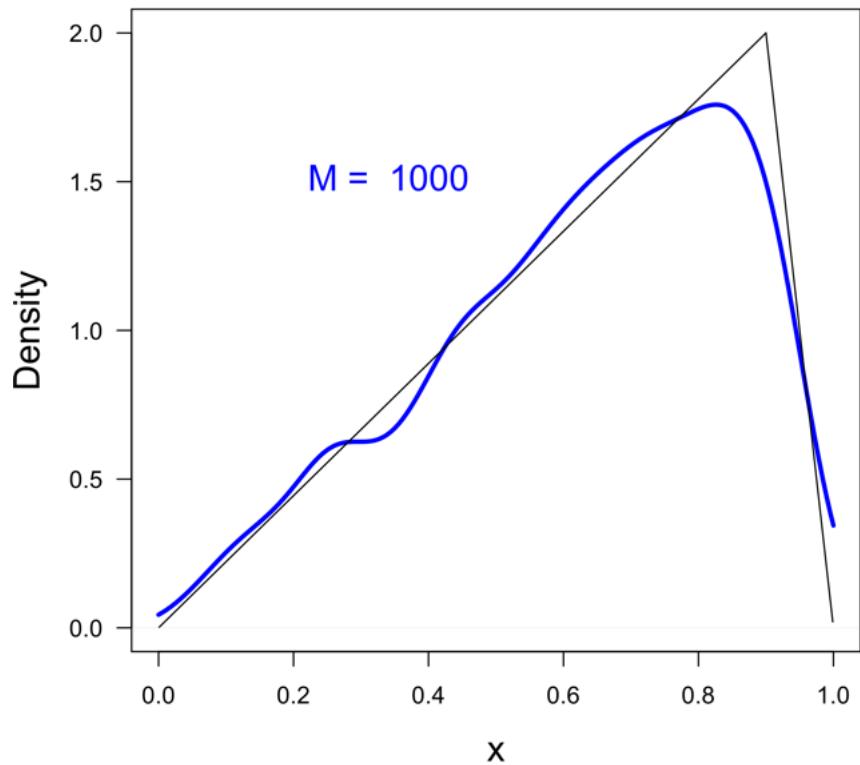


Figure M.7: Random numbers (estimated density based on  $M = 10^3$  numbers, blue thick line) from the asymmetric standard triangular distribution with mode 0.9 (black thin line). The mode of the density estimate is smaller than the true mode. [RNasymTriangle.R](#)

### M.2.4 Numerical estimation of a reference prior for asymmetric triangular PDF

We will now use the recipe for the numerical estimation of a reference prior given in the beginning of this section. We choose the following values: number of Monte Carlo runs  $M = 500$ , sample size  $K = 100$ , flat prior  $h(\theta) = 1$ , 13 discrete  $\theta$  values  $\{0.01, 0.05, 0.1, 0.17, 0.27, 0.37, 0.5, 0.63, 0.73, 0.83, 0.9, 0.95, 0.99\}$  covering the whole range with finer resolutions at the lower and upper boundary. Random values are generated by using the R routine `rtriangle()` from package `triangle`. Likelihoods for single data values are calculated using the routine `dtriangle()` from the same package. The joint likelihood for a sample of size  $K$  is given by the product of the likelihoods for single data. The main numerical effort goes into the integrals for  $c_j$  (Eq. M.3). The integrands were coded in the form of functions in order to apply the routine `integrate()`. However, for large sample sizes  $K$  this routine stops with the error message 'extremely bad integrand behaviour'. Thus we calculated the integrand for a fine equidistant grid of argument values and calculated the integral by summing up the integrand values times the grid resolution. The calculation of the  $r_j$  (Eq. M.4) and the prior estimates  $\pi(\theta_d)$  (Eq. M.5) is straightforward. Finally, the prior values are rescaled to  $\pi(1/2) = \text{Beta}(0.5; \alpha = 1/2, \beta = 1/2) = 2/\pi$ .<sup>4</sup> The results are shown in Fig. M.8 together with the Beta-prior mentioned in the above quotation by Berger et al. (2009): **the numerical estimates indeed support the conjecture that  $\text{Beta}(\theta; \alpha = 1/2, \beta = 1/2)$  is the reference prior for asymmetric standard triangular PDFs.**

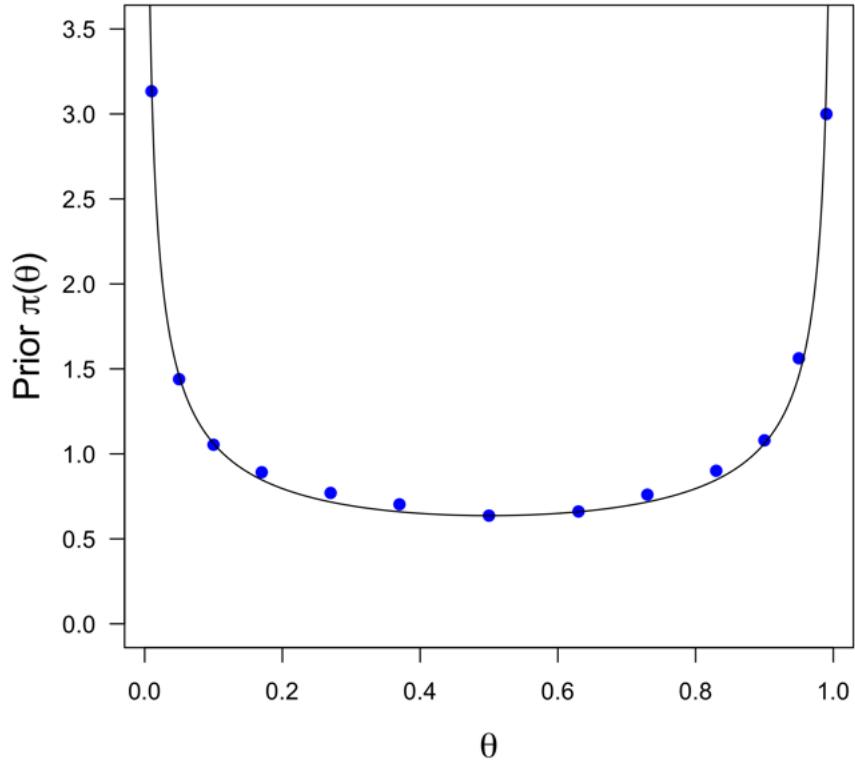


Figure M.8: Numerical estimates of the reference prior for the asymmetric standard triangular PDF (blue dots) and the Beta distribution  $\text{Beta}(\theta; \alpha = 1/2, \beta = 1/2)$ . [RefPriorAsyTri.R](#)

<sup>4</sup>Note that  $\pi$  is used here as a function name (for the prior) as well as for the constant  $\pi = 3.14\dots$  which is hopefully not too confusing.

### M.2.5 Analyzing the posterior and estimating the mode

Using the Beta-prior  $\text{Beta}(\theta; \alpha = 1/2, \beta = 1/2)$  and the likelihood (Eq. M.14) we obtain the posterior

$$L(x|\theta) \propto \prod_{i=1}^M L(x_i|\theta). \quad (\text{M.15})$$

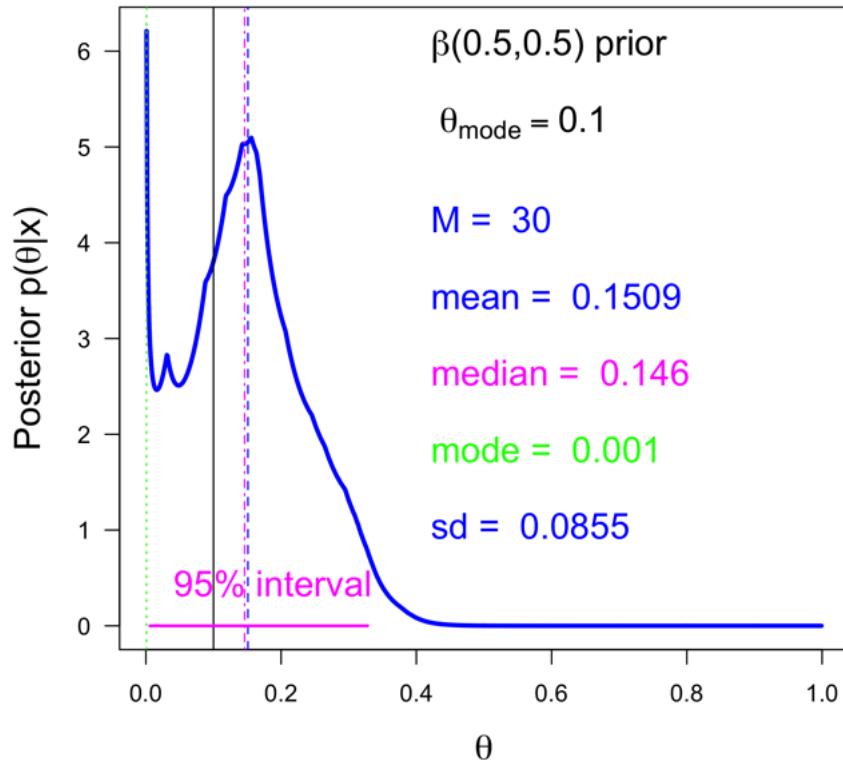


Figure M.9: The posterior based on the Beta prior  $\text{Beta}(\theta; \alpha = 0.5, \beta = 0.5)$  for an artificial data set of size  $M = 30$  from a triangular population the mode  $\theta_{\text{mode}} = 0.1$ . [TriangPostBetaPrior.R](#)

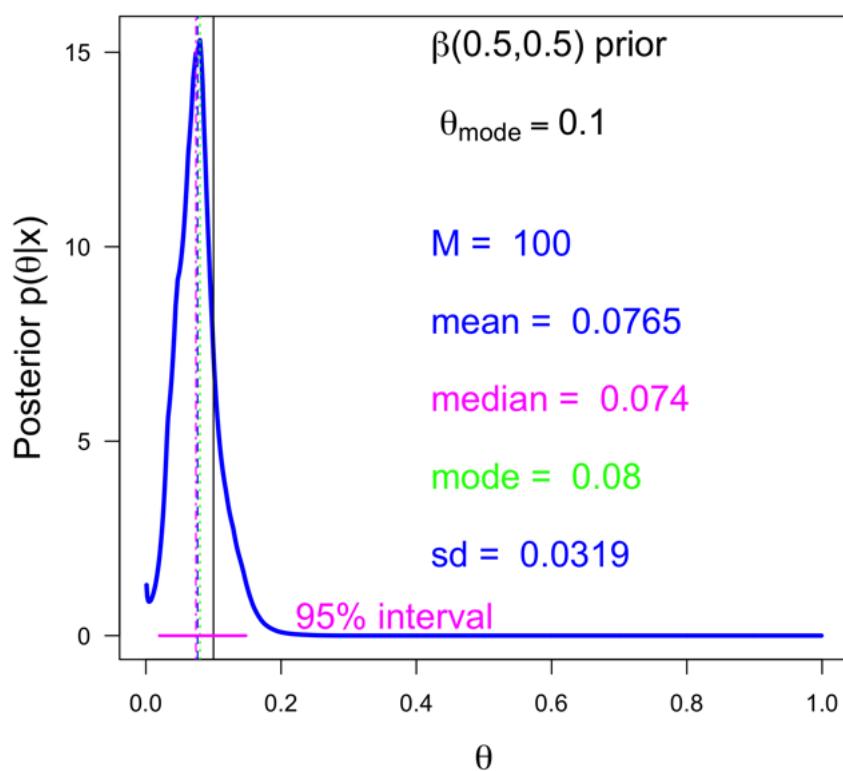


Figure M.10: The posterior based on the Beta prior  $\text{Beta}(\theta; \alpha = 0.5, \beta = 0.5)$  for an artificial data set of size  $M = 100$  from a triangular population the mode  $\theta_{\text{mode}} = 0.1$ . [TriangPostBetaPrior.R](#)

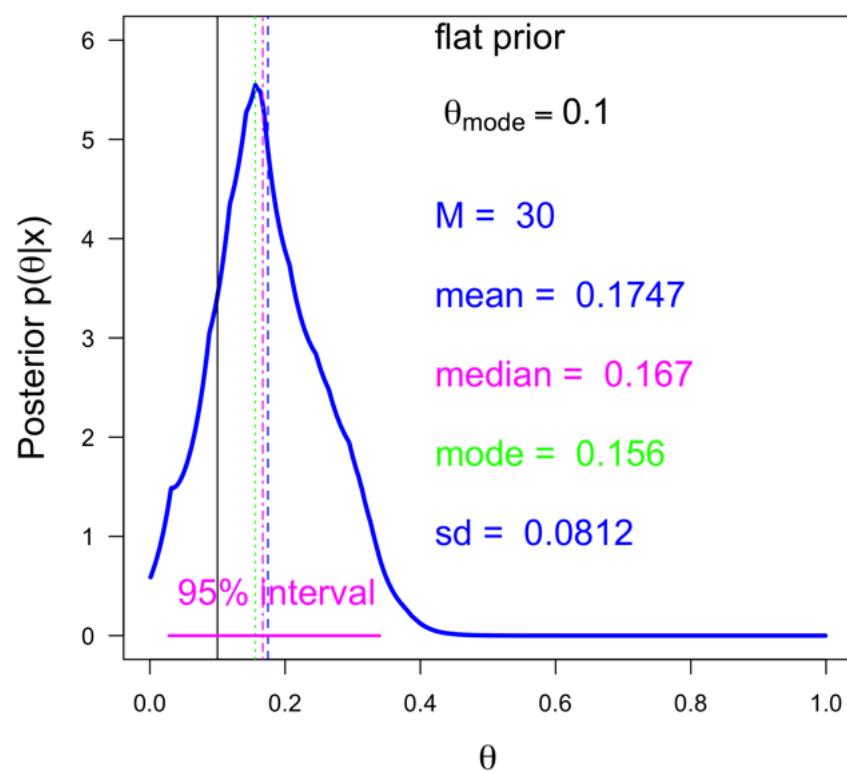


Figure M.11: The posterior based on the flat prior for an artificial data set of size  $M = 30$  from a triangular population the mode  $\theta_{\text{mode}} = 0.1$ . [TriangPostBetaPrior.R](#)

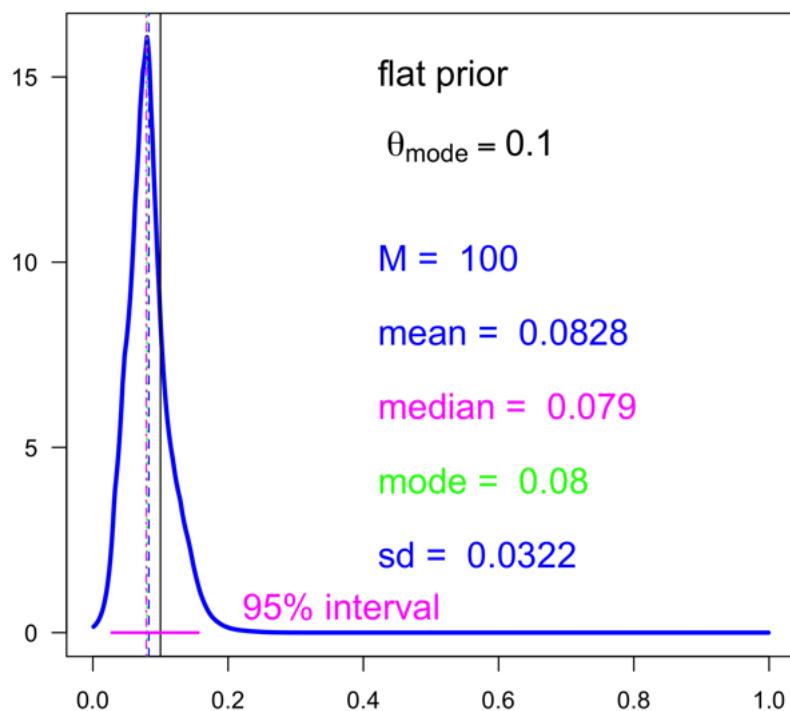


Figure M.12: The posterior based on the flat for an artificial data set of size  $M = 100$  from a triangular population the mode  $\theta_{\text{mode}} = 0.1$ . [TriangPostBetaPrior.R](#)

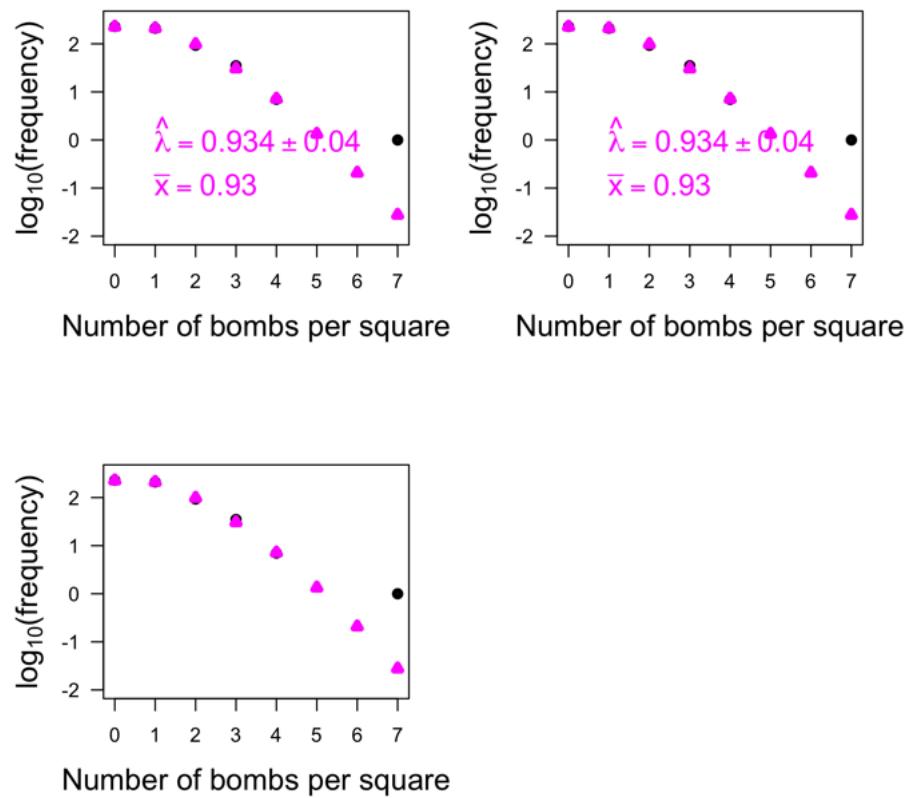


Figure M.13: Posteriors based on different priors ( $\beta$  prior  $\text{Beta}(\theta; 1/2, 1/2)$  in upper panels, flat prior in lower panels) for artificial data with different sample size ( $M = 30$  in left panels,  $M = 100$  in right panels) from the triangular population with  $\theta_{\text{mode}} = 0.1$ . The mean of the population (0.151 based on  $\beta$  prior, 0.175 based on flat prior) overestimates the true value for small sample sizes (left panels). [TriangPostBetaPrior.R](#)

**Discussion** (based on single artificial samples):

- For small sample sizes ( $M = 30$ , Fig. M.13 left panels), the mean of the population (0.151 based on  $\beta$  prior, 0.175 based on flat prior) overestimates the true value ( $\theta_{\text{mode}} = 0.1$ ). The  $\beta$  prior does a better job than flat prior in that the mean of the prior is closer to the true mode. The estimated posterior based on the beta prior is bimodal and the mode (0.001, Fig. M.9 and Fig. M.13 upper left panel) largely underestimates the true value. The estimated posterior based on the flat prior (Fig. M.11 and Fig. M.13 lower left panel) is unimodal showing that the peak at small  $\theta$  in the posterior based on the beta prior is largely driven by the prior.
- For large sample sizes ( $M = 100$ , Fig. M.13 right panels), the mean of the population (0.076 based on  $\beta$  prior, 0.083 based on flat prior) underestimates the true value ( $\theta_{\text{mode}} = 0.1$ ). As expected, the standard deviation of the mean is smaller for the larger sample size. A bit surprisingly, the flat prior does a better job than the  $\beta$  prior in that the mean of the posterior is closer to the true mode. The estimated posterior based on the beta prior is unimodal (Fig. M.10 and Fig. M.13 upper right panel) and the mode gives an estimate (0.08) which underestimates the true value. The estimated posterior based on the flat prior (Fig. M.12 and Fig. M.13 lower right panel) is unimodal and mode gives an estimate (0.08) which underestimates the true value.

The discussion is not exhaustive because the statements are based on single artificial samples. A more detailed investigation (Monte Carlo simulation) will be left to the reader: Is the underestimation at  $\theta_{\text{mode}} = 0.1$  typical? How do the estimates vary with sample size? How they vary with  $\theta_{\text{mode}}$ ?

**Exercise 86 Estimate rate constant of exponential population: flat prior**

*In Section M.2.1 we estimated the rate constant of an exponential population using the Jeffreys prior  $1/\theta$ . Calculate the posterior for the flat prior and compare the resulting mean, median, and mode for our sample  $x$  of size  $n = 10$  with those based on Jeffreys prior.*

**Exercise 87 Random numbers from an asymmetric triangular distribution**

*In Section B.1.1 we have shown how to generate random numbers from a symmetric triangular or 'tent' distribution. Generalize the method to asymmetric triangular distributions.*

# Appendix N

## R: tips & tricks

### N.1 Communication with the operating system

One can communicate with the operation system of a computer by applying special **R** commands:

**getwd()** for 'get/print working directory'

**setwd('/Myroot/myworkingdirectory')** for changing/setting the working directory to a working directory with path '/Myroot' and name 'myworkingdirectory'

**mydir = getwd()**

**list.files(path=**mydir**)** print names of all files in current working directory

**list.files(path=**mydir**,pattern='.**R**)** print names of all files with extension '.R' in the current working directory

On Macs one can use, in addition, many unix commands by calling the routine **system()**. Here are a few examples:

**system('pwd')** for 'print working directory'

**system('ls \*.R')** listing all files with extension '.R'

Another difference between Macs and PCs refers to the opening of graphic windows (especially for 'large' plots, e.g. plots with many panels). On Macs this can be done by calling

**quartz(title='I love my Mac',10,5)** = opening a graphic window with title 'I love my Mac', a width of 10 inches (yes, inches!), and a hight of 5 inches.

On a PC one can open a graphic window by the command

**windows(title='I love my PC')**

### N.2 Assign and print

Assigning a value to a variable:

```
x <- 5+3  
x = 5+3 # This works as well!
```

Results are not printed ('echoed'). One can enforce an echo by putting the command in round brackets

```
(x = 5+3) # Give me an echo!
```

### N.3 Generate arrays when length is not known ('dynamic array size')

In certain applications one likes to store numbers (or other items) in an array, however, without knowing in advance the length of the array. A simple way to deal with such cases in R is to first generate an empty array by the command '`x = c()`' where `c()` stands for concatenate and '`x`' is just a name. A number  $r$  can be added (actually appended) to this array by the command '`x = c(x,r)`'. Here is a simple (rather silly) example:

### N.4 Missing functions/routines: `erf()`, `erf.inv()`, ...

Although the error function and its inverse are often used in the context of statistics, R does not provide these as built-in functions. However, one can use

```
erf <- function(x) 2 * pnorm(x * sqrt(2)) - 1
erf.inv <- function(x) qnorm((x + 1)/2)/sqrt(2)
```

R routine for non-standardized  $t$  distribution:

```
mydnst = function(x,location,scale,df) {
  # density of non-standardized t PDF; DWG 6/2021
  tstat = (x-location)/scale; return(dt(tstat,df)/scale)
  # factor 1/scale in density stems from d tstat/dx = 1/scale
}
myrnst = function(M,location,scale,df) {
  # random numbers from non-standardized t PDF; DWG 6/2021
  tstat = rt(M,df); return(tstat*scale+location)
}
```

### N.5 Find indices of minimum values

Given an array of numbers  $A$ . One wants to know the minimum of  $A$  and thus calls `min(A)`. However, one also wants to know the index of the minimum value. This can be obtained by calling `which.min(A)`. Everything is analogue for the maximum.

The call `which.min(A)` works as well when  $A$  is a matrix. However, in case of matrices one obtains one index only (as if the  $n \times m$  matrix is considered as a one-dimensional array of length  $n \cdot m$ ). In order to obtain both indices (row and column) one can call `which(A == min(A), arr.ind = TRUE)`.

### N.6 Numerical integration in 1D: `integrate()`

Numerical integration in 1D can be performed using the R routine `integrate()`. A simple function (for which an analytic solution is known) was used on purpose in order to compare results from numerical and analytical integration ( $a = 2.1$ ):

$$\int_0^{1.2} \cos(ax) dx = \frac{1}{a} \sin(ax) \Big|_0^{1.2} = \frac{1}{2.1} \sin(2.1 \cdot 1.2) \approx 0.2773 \quad (\text{N.1})$$

R code: [NumericalIntegration1D.R](#)

## N.7 95% intervals: `quantile()`

Suppose you have a large sample  $x_i, i = 1, 2, \dots, M$  from a statistical population and you want to find an interval, that covers 95% of the sample. This can be done by applying the R routine `quantile()`. In the example below, we take a random sample  $x$  of  $M = 1000$  numbers from the standard normal distribution and plot a histogram of the data (Fig. N.1). The analytical value for the 95% interval of the standard normal distribution is  $[-1.9600, 1.9600]$ . Applying `quantile()` to the random sample yields  $[-2.0675, 1.904]$ .

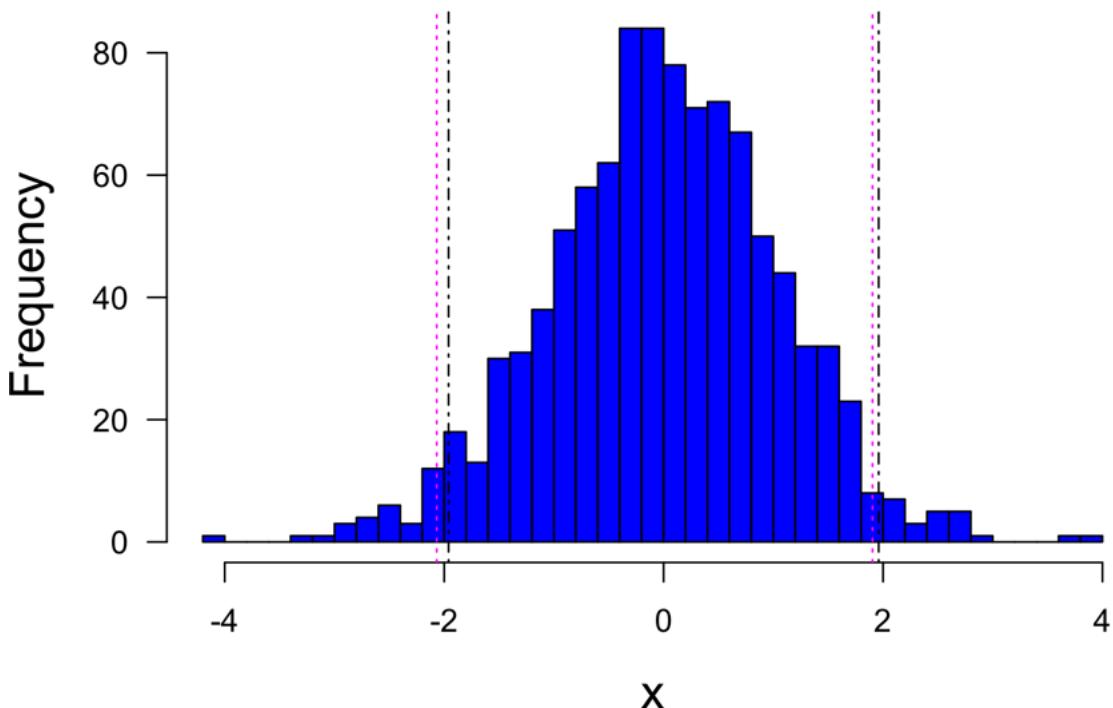


Figure N.1: Histogram of  $M = 1000$  random numbers from the standard normal distribution. The analytical value for the 95% interval of the standard normal distribution is  $[-1.9600, 1.9600]$  (black dash-dotted lines). Applying `quantile()` to the random sample yields  $[-2.0675, 1.904]$  (magenta dashed lines).

R code: [QuantileRTricks.R](#)

## N.8 Plots

### N.8.1 How to change position of axis labels

The automatic positioning of axis labels does not give always satisfying results, especially when superscripts are included in the label (Fig. N.2). The positioning can be improved by plotting the axis label separately using the R routine `title()` and specifying the distance from the axis by assigning a value for 'line' (Fig. N.2 and R code below).

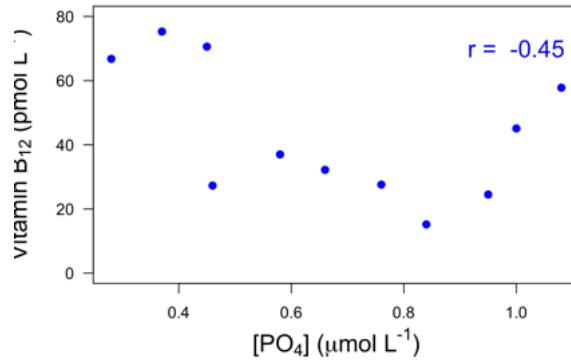


Figure N.2: Vitamin B<sub>12</sub> over PO<sub>4</sub> (data from Sañudo-Wilhelmy et al., 2006): note that the superscript <sup>-1</sup> at the end of the y-axis label is cut off. R code (flag=1): [YLabelRTricks.R](#)

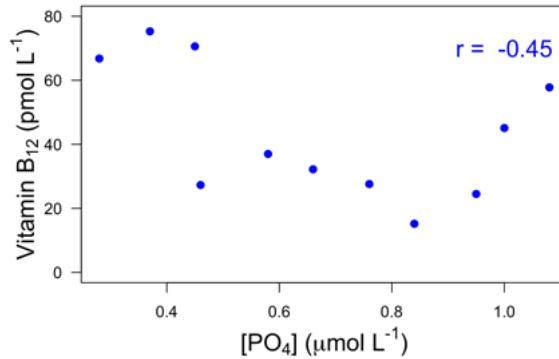


Figure N.3: Vitamin B<sub>12</sub> over PO<sub>4</sub> (data from Sañudo-Wilhelmy et al., 2006) R code (flag=2): [YLabelRTricks.R](#)

## N.8.2 How get rid of axes, axes labels, and frame

```
plot(..., col.axis='white',xaxt='n',yaxt='n',xlab='',ylab='',bty='n',...)  
compare, for example, Fig. 8.1
```

## N.8.3 Mathematical annotation

Mathematical annotation in R can be added using `text()` or, for axes, `title(yaxis = ...)`. Values (numbers) can be converted to character strings by using `as.character`. The basic R routine for generating, for example, Greek letters with sub- or super-scripts is `expression()`. The application of `expression()` is often tedious or awful and time-consuming. The package `latex2exp` allows to use the well-known L<sup>A</sup>T<sub>E</sub>X commands (with small modifications, especially by doubling backslashes) to formulate desired terms and to 'translate' them into 'R-expressions'. I found no easy way for combining Greek text with values using `expression()` or `TeX()` from `latex2exp`. However, this combination is possible using `bquote()`. Some information on various mathematical expressions within `expression()` or `bquote()` is available under `?plotmath`; example:  $\pm$  is expressed as `x %+-% y`. Examples of all three routines – `expression()`, `TeX()`, `bquote()` – are shown in Fig. N.4.

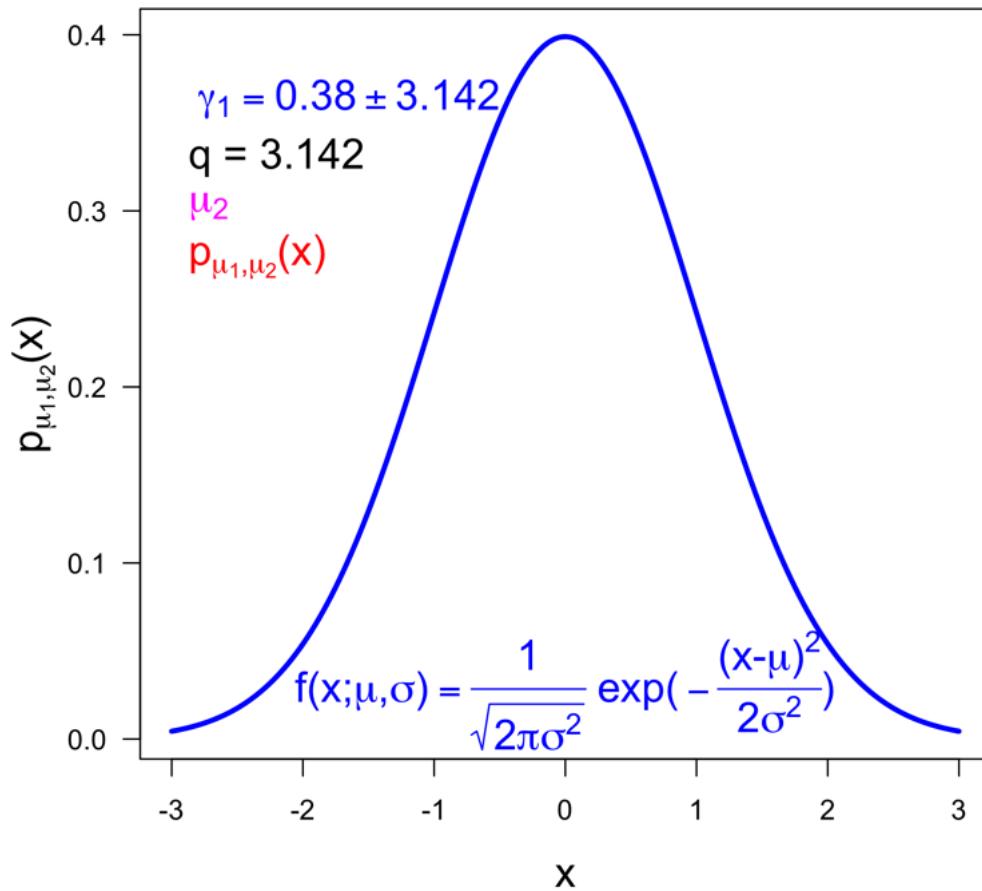


Figure N.4: Mathematical annotation in R [RtipsLaTeX2exp.R](#)

## N.9 How to obtain specific output from summary?

Although the **R** routine **summary()** provides a lot of information, it is sometimes not obvious how to obtain specific output that one likes to use in further calculations. Here is an example where one would like to obtain the *p-value* or the *F value* of an ANOVA. The amount of brackets required to do so is impressive!

**R code: ANOVA Zar (2010, Example 10.1) & how to create Fig. 12.19**

```
print('file: R_Zar10Ex10d1_160726.R')
print('3 = ANOVA Zar (2010, Example 10.1) (7/2016)')
mydata=read.table('Zar10Ex10d1data.txt',header=T)
# ANOVA:
summary(aov(weight ~ group,data=mydata))
#          Df Sum Sq Mean Sq F value Pr(>F)
# group      3  338.9  112.98  12.04 0.000283 ***
# Residuals  15  140.8   9.38
q = summary(aov(weight ~ group,data=mydata))
p = q[[1]][, 5][1] # yes, this is obvious!
Fvalue = q[[1]][, 4][1]
print(c(p,' p'))
print(c(Fvalue,' Fvalue'))
# [1] '0.000283012228461253' ' p'
# [1] '12.0404038515471' ' Fvalue'
```

## N.10 How to get rid of NAs? x[!is.na(x)]

The generic function **is.na()** indicates which elements are missing (NA = not available). Here is an example:

x = c(5,NA,8)

Application of **is.na()**, i.e. **is.na(x)**, yields

FALSE TRUE FALSE

However, our goal is to get the negation, i.e. TRUE if the element is a number and FALSE if it is not available.

Negation can be obtained by putting an exclamation mark (!) in front of **is.na(x)**. **!is.na(x)** yields:

TRUE FALSE TRUE

This result can now be used to get rid of the NAs:

xc = **x[!is.na(x)]**

which yields:

5 8

## N.11 Data formats

### N.11.1 Factors

'Tell **R** that a variable is nominal by making it a factor. The factor stores the nominal values as a vector of integers in the range  $[1 \dots k]$  (where  $k$  is the number of unique values in the nominal variable), and an internal vector of character strings (the original values) mapped to these integers.'

### N.11.2 Lists

'An ordered collection of objects (components). A list allows you to gather a variety of (possibly unrelated) objects under one name.' (Quick **R**, Data Types) <https://www.statmethods.net/input/datatypes.html>

### N.11.3 Data frames

Data frames are more general than matrices, in that different columns can have different modes (numeric, character, factor, etc.).

Some R routines require input of data in a format that is called **data frame()**. From three data vectors of different types (numbers, characters, logical variables)

```
A = c(2.1, 3.5, -1.2)
```

```
B = c('q','p','a')
```

$C = c(\text{TRUE}, \text{FALSE}, \text{FALSE})$  a data frame is generated by the command

`dfr = data.frame(A,B,C)` It can be viewed by calling various routines:

`View(dfr)`

`head(dfr)`

`tail(dfr)`

### N.11.4 Application of data frames: ANOVA & Tukey HSD post-hoc test

Given samples  $X_k$  of equal size from 5 groups  $k = 1, 2, \dots, 5$ , we ask whether all true mean values  $\mu_k$  are equal (= null hypothesis  $H_0$ ). In order to guess what the test will yield, one calculates the sample mean  $\pm$  one standard deviation (an estimate of the true values):

$$\bar{X}_1 \pm s_{d1} = 32.1 \pm 3.2$$

$$\bar{X}_2 \pm s_{d2} = 40.2 \pm 2.5$$

$$\bar{X}_3 \pm s_{d3} = 44.1 \pm 3.1$$

$$\bar{X}_4 \pm s_{d4} = 41.1 \pm 3.7$$

$\bar{X}_5 \pm s_{d5} = 58.3 \pm 3.0$  The sample standard deviations are quite similar (from 2.5 to 3.7). The sample mean  $\bar{X}_1$  is much smaller (several standard errors<sup>1</sup>) than all other sample means and thus the true mean  $\mu_1$  is most probably different from all other true mean values. The sample mean  $\bar{X}_5$  is much larger (several standard errors) than all other sample means and thus the true mean  $\mu_5$  is most probably different from all other true mean values. The sample means  $\bar{X}_2$ ,  $\bar{X}_3$ , and  $\bar{X}_4$  are quite similar and thus they are most probably not different from each other.

Application of ANOVA leads to a tiny  $p$ -value ( $p = 3.95 \cdot 10^{-12}$ ) and thus  $H_0$  is rejected based on the chosen significance level  $\alpha = 0.05$ . The Tukey HSD post-hoc test leads to the results that one was guessing based on the sample mean  $\pm$  standard error.

## N.12 Read from Excel file

In the ANOVA sections 12.3.2 and 12.3 we pasted the data into the R code or loaded an ascii (.txt) file. Here we will read the data directly from an Excel file [Zar10Ex10d1x210223.xlsx](#) This can be done with the routine `read_xlsx` from the package `readxl`: R code: [ReadFromExcel.R](#)

## N.13 Restrict samples by applying simple conditions

When analyzing multivariate data sets one often likes to exclude samples where single predictors violate certain conditions. In the following example we like to exclude all samples where the first predictor ( $X_1$ ) is negative and all samples where the fourth predictor ( $X_4$ ) is larger than 4.0. This can be done as follows:

1. `q = (X1 >= 0) & (X4 <= 4)`

`q` is an array of TRUE and FALSE values. Please note that the two constraints are connected by a single ampersand (&); this is in contrast to connections between two single logical values where two ampersand are applied as, for example, in

---

<sup>1</sup>For sample size  $n = 6$ , the standard error of the mean is smaller than the standard deviation by a factor of  $\sqrt{6} \approx 2.4$

```
(X1[1] >= 0) && (X4[1] <= 4)
```

2. By applying q as 'index' to all predictors (X1, X2, X3, X4) and to the response (Y), one obtains predictor and response values that are restricted to the samples with q[i] = TRUE:

```
X1r = X1[q]
```

etc.

An example using the data from Zar (2010, Example 20.1; used already in Section 16.2.4) can be found here: **R** code: [RestrictSamples.R](#)

## N.14 If ... else ...

If A (inclusive) or B are not available, then set C also to 'not available (NA), else set C to A divided by B:

```
if((is.na(A)) | (is.na(B))) {C = NA
} else {C = A/B}
```

## N.15 Obtain source code of library routines

The source code of **R** routines can be obtained as follows:

**libraryname:::routinename**

for example:

**LearnBayes:::laplace**

# Appendix O

## Notation & Abbreviations

Table O.1: Notation: Latin letters

Symbol	Meaning
$p$	observed level of significance ('p-value')
$SS$	sum of squares

Table O.2: Notation: Greek letters

Symbol	Meaning
$\alpha$	(chosen) level of significance
$\beta$	slope (linear regression)
$\Gamma()$	gamma function
$\epsilon$	noise
$\lambda$	mean number of events (Poisson distribution)
$\log(x)$ or $\ln(x)$	natural logarithm of $x$
$\mu$	true mean value
$\sigma$	true standard deviation
$\sigma^2$	true variance

Table O.3: Notation: calligraphic letters

Symbol	Meaning
$\mathcal{B}()$	binomial distribution
$\mathcal{F}()$	F-distribution
$\mathcal{H}()$	hypergeometric distribution
$\mathcal{N}(x; \mu, \sigma)$	normal distribution
$\mathcal{O}(\epsilon)$	order of epsilon
$\mathcal{P}(k; \lambda)$	Poisson distribution

Table O.4: Notation: miscellaneous

Symbol	Meaning
$\bar{A}$	not $A$
,	read as 'and': $A, B = A \text{ and } B$
	read as 'given': $A B = A \text{ given } B$
$\cap$	and (for proposals); intersection (in set theory)
$\cup$	inclusive or (for proposals); union (in set theory)

Table O.5: Abbreviations

Abbreviation	Meaning
CLT	Central Limit Theorem
KS	Kolmogorov-Smirnov
MLE	Maximum Likelihood Estimate
MLR	Multiple Linear Regression
MSE	Mean Squared Error
PD	Probability Distribution (discrete)
PDF	Probability Density Function (continuous)
PLS	Partial Least Squares
q.e.d.	quod erat demonstrandum (Latin) = what was to be demonstrated

## O.1 Glossary

**Bernoulli process** = a discrete stochastic process with only two possible outcomes (0 or 1, failure or success) in a single trial (other stochastic processes are: Cauchy process, Poisson process, Wiener process)

**BLUE** = best linear unbiased estimator

**CDF** = cumulative distribution function

**c.i. (or CI)** = credible interval (or credibility set, Casella & Berger, 2002) (Bayesian) or confidence interval (frequentist); please note that credible intervals and confidence interval are not identical

**count data** = non-negative integer values (0, 1, 2, ...) from counting individuals, events, etc.

**critical value** of the test statistic = value of the test statistic determining the border of rejection region; the critical value depends on the degrees of freedom,  $\nu$ , and on the chosen level of significance,  $\alpha$ .

**degrees of freedom** = number of data minus number of constraints (example: if the mean value of  $n$  data points is known (= 1 constraint) there remain  $\nu = n - 1$  degrees of freedom).

**density (slang)** = probability density function (PDF) or probability distribution (PD)

**deviance** = - two times the natural logarithm of the maximum likelihood ( $-2 \log \text{Lik}$ ); the deviance is a measure for the goodness-of-fit (Section )

**distribution function (R slang)** = cumulative distribution function (CDF); examples (**pnorm()**, (**pt()**

**effect size** = mean divided by standard deviation (true:  $\mu/\sigma$ ; observed:  $\bar{x}/s$ )

**empirical Bayes** = procedures (estimation, hypothesis testing) in which prior distributions are estimated from data

**estimator**  $\Rightarrow$  point estimator

**false negative (slang)** = Type II error = failure to reject a false null hypothesis

**false positive (slang)** = Type I error = incorrect rejection of the null hypothesis

**fat tail(s) (slang)** = more probability in the tail(s) compared to another PDF

**heteroscedasticity** = noise level is varying (from Ancient Greek *skedasis* = 'dispersion'); contrast: homogeneous

**homoscedasticity** = homogeneous noise level (from Ancient Greek *skedasis* = 'dispersion') contrast: heteroscedasticity

**i.i.d. or iid** = independent and identically distributed, i.e. (1) each random variable has the same probability distribution as the others and (2) all are mutually independent

**interaction** = (statistical slang) in multiple linear regression, terms consisting of products of predictors (quadratic or higher order) are called 'interactions'

**inverse probability** = an old notation (named by Augustus de Morgan; used between 1838 and 1945) for 'Bayesian analysis'

**level of significance**  $\alpha = \alpha \cdot 100\%$  = chosen size of rejection region;  $\alpha = 0.05 \Rightarrow$  rejection region contains 0.05 (5%) of the whole probability (1; 100%)

**location parameter** (see under **scale parameter**)

**mode** = location of a (local) maximum of a PDF; if the PDF has only a single maximum it is called unimodal.

**nested model**: model B is nested in model A when setting specific parameters in model B to zero leads to model A

**nuisance parameters** = parameters that are present in a model but are not of direct inferential interest

**observed level of significance (p-value)**: the probability for the observed test statistic or less favorable values under the assumption that the null hypothesis is true

**odds**: the odds for probability  $p$  is  $o = p/(1-p) \Rightarrow p = o/(1+o)$ .

**one-sided test** = alternative name for one-tailed test.

**one-tailed test** = test whether estimated value is larger (lower) than hypothesized value; examples of null hypotheses: mean value  $\mu_1$  is lower than  $\mu_2$  (one-tailed t-test), variance  $\sigma_1^2$  is larger than  $\sigma_2^2$  (variance ratio test); compare also: two-tailed test.

**outlier** = an observational point that lies far away from the other data; an outlier is often a falsely recorded data point, however, sometimes it is the most interesting observation

**overdispersion** = presence of greater variability (variance, statistical dispersion) in a data set than expected (expectation could, for example, be based on a Poisson distribution fitting the sample mean)

**parameter** = an unknown quantity in a statistical model, usually denoted with Greek letters (examples: mean  $\mu$ , standard deviation  $\sigma$ , intercept  $\beta_0$ ); estimates of parameters are called 'statistics' and are usually denoted with a little hat atop the Greek letters ( $\hat{\mu}$ ,  $\hat{\sigma}$ ,  $\hat{\beta}_0$ ) or by Latin letters ( $m$ ,  $s$ ,  $b_1$ ) or other symbols ( $\bar{x}$  for sample mean).

**p-value** (slang) = observed level of significance: "the probability of observing data as extreme or more extreme than that actually observed assuming the null hypothesis is true" (McShane et al., 2019)

**point estimator** 'A point estimator is any function  $W(X_1, \dots, X_n)$  of a sample; that is, any statistic is a point estimator.' Casella & Berger (2003, p. 311); the vagueness of the definition reflects the fact that various choices are possible (compare, for example, the arithmetic mean, the median, winsorized mean etc.).

**point null** = statistical slang for a point (or sharp) null hypothesis

**power of a test** (Neyman-Pearson) = probability of correctly rejecting the null hypothesis, given the alternative hypothesis is true =  $1 - \text{Type I error probability} = 1 - \beta$

**precision** = inverse of the variance

**regression to the mean** = 'taller mothers tend to have daughters who are shorter than them, and shorter mothers tend to have taller daughters' (the same phenomenon holds for fathers and sons) or, more general, the phenomenon 'occurs because part of the reason for the initial extreme case was chance, and this is unlikely to repeat to the same extent' (Spiegelhalter, 2019)

**residuals** = difference between observed and fitted values

**residuals, standardized** = residuals divided by square root of mean square error

**residuals, studentized** = residuals divided its standard error

**sample size** = number of data points in the sample

**sample space**: 'The set,  $\mathcal{S}$ , of all possible outcomes of a particular experiment is called the *sample space* for the experiment.' (Casella & Berger, 2002) The sample space can be discrete (for example, tossing a coin can yield head (H) or tail (T) and thus  $\mathcal{S} = \{H, T\}$ ) or continuous (for example, temperatures  $T$  between 0°C and 100°C; here we consider the temperature as a real number although we are not able to measure it with arbitrary precision).<sup>1</sup>

**scale parameter** = a parameter describing the dispersion (spread) of a PDF; example: the variance  $\sigma^2$  of the normal PDF  $\mathcal{N}(x; \mu, \sigma^2)$  is a scale parameter; remark: 0 is a natural lower bound for scale parameters. Contrast: **location parameters** as, for example, the mean  $\mu$  of the normal PDF can vary all over the place (no natural bounds).

**s.d.** = standard deviation

**standard error** = (a) standard deviation of a statistic and (b) the standard error of the mean is often just called the standard error

**s.e.m or SEM** = standard error of the mean = standard deviation / square root of sample size

**size of a test** (statistical slang) = probability for Type I error, usually denoted by  $\alpha$

<sup>1</sup>Casella & Berger (2002) make the distinction between countable (finite number of elements in the sample space or infinite many elements that can be mapped to the natural numbers) and uncountable (for example, all real numbers between 0 and 1) sample spaces. This is, however, what is meant by discrete versus continuous, a notation more familiar to non-mathematicians who have never heard of the difference between  $\aleph_0$  and  $\aleph_1$ .

**standardization** of a sample  $x$  = (sample - sample mean)/standard deviation of sample:  $z = (x - \bar{x})/s$  (also known as z-transformation, z-scores)

**statistical model** = 'formal representation of the relationships between variables, which we can use for the desired explanation or prediction' (Spiegelhalter, 2019)

**tail(s)** (of a probability density function) = region(s) far away from the location of the maximum (see also: fat tails)

**testing, Bayesian** = approach by Jeffreys, Jaynes and others using priors and likelihoods to calculate posteriors for null and working hypotheses

**testing, hypothesis** = approach by Neyman, Pearson and others using significance and power

**testing, significance** = approach by Fisher and others using significance levels (chosen level of significance  $\alpha$ , observed level of significance  $p = p\text{-value}$ )

**test statistic** or statistic for short: the definition 'A test statistic ... a function of the sample' (Casella & Berger, 2002, p. 374) is rather general, showing that its choice can be a bit arbitrary; one may expect, however, that the test statistic contains useful information about the quantity that we would like to know or estimate (compare the construction and interpretation of the test statistic  $t$  as an example).

**two-sided test** = alternative name for two-tailed test.

**two-tailed test** = test whether estimated value is different from hypothesized value and it does not matter whether it is smaller or larger; examples of null hypotheses: equal mean values (two-sample two-tailed t-test), equal variances (variance ratio test); compare also: one-tailed test.

**type I error** = probability for rejecting a true null hypothesis, usually denoted by  $\alpha$ , also called 'size of a test'

**type II error** = probability for rejecting a false null hypothesis, usually denoted by  $\beta$

**unbiased point estimator** = an estimator with expectation ('mean value') equal to the true value ('gives correct result in the mean'); examples: sample mean for  $\mu$  and sample variance (division by  $n - 1$ ) for  $\sigma^2$

**underdispersion** = presence of less variability (variance, statistical dispersion) in a data set than expected (expectation could, for example, be based on a Poisson distribution fitting the sample mean)

**variance covariate** = an explanatory variable that is used in the variance of the residuals

**Wilkinson notation** = a short symbolic notation for specifying models of linear regression (Section 16.3)

**z-scores** of a sample  $x$  = (sample - sample mean)/standard deviation of sample:  $z = (x - \bar{x})/s$  (also known as standardization, z-transformation)

**z-transformation** of a sample  $x$  = (sample - sample mean)/standard deviation of sample:  $z = (x - \bar{x})/s$  (also known as standardization, z-scores)



## Appendix P

# Postface: long version of preface

*In the preface I will address a few topics like my motivation for writing the script, readership, scope of the text, R codes, exercises, different schools (Bayesian, frequentist). If you are interested in a certain topic you can skip this and most other sections and jump directly to what you choose from the table of contents, the index, or from a search of a keyword.*

### My motivation for writing the script

In 2004, Christine Klaas approached me with questions about Bayesian data analysis. She believed I could be of immediate help given my background in physics and mathematics. At the time I knew nothing about the Bayesian approach and little about statistics which, as for probably the vast majority of scientists (a guesstimate, without applying statistics), was not on my top ten list of most exciting topics. I, nevertheless, had a look at several text books on statistics used in various fields of natural sciences (including ecology, climate sciences), but was often disappointed by the explanations of the principles or the lack thereof (correction factors popping up here and there, leaving me in a state of confusion).

Luckily, I came across the fantastic book by Ed Jaynes (*Probability - The Logic of Science*, 2003) which had just been published; it took me about two months on my sofa to digest it. Soon after, I was asked to give a MSc course on 'statistics for biologists'. I have been teaching this course since 2007 usually twice a year: for MSc students in 'marine biology' at the University of Bremen and, in slightly modified form, for PhD students from various disciplines including biology, geology, and physics in the framework of graduate schools (POLMAR at the AWI or GLOMAR at the University of Bremen).

In the beginning, my course was excessively theoretical and too difficult to grasp for most students given their background knowledge in mathematics. Over the years, I introduced more and more examples and solutions to problems by writing relatively short (10 to 20 lines) R codes. I had to learn and use R at the request of the students. I thank all students for their patience and for asking stimulating questions while I tried to explain the basic ideas behind the various statistical procedures. Without their feedback my script would have been quite different! The script has been written mainly thinking of Christine, myself and my students. However, I hope that it will be helpful to others as well.

**Exercise:** Falling in love with statistics? Who's singing?

"I've nothing much to offer  
There's nothing much to take  
**I'm an absolute beginner**  
And I'm absolutely sane  
As long as we're together  
The rest can go to hell  
**I absolutely love you**  
But we're absolute beginners  
With eyes completely open  
But nervous all the same"

### Readership

The readership that I had in mind are students and scientists with a background in biology, geology, or other natural sciences who struggle with the analysis of observations from the field or from experiments. In large parts, the script is written for beginners, i.e. I will start with discussing mean, variance, standard deviation, etcetera.

**Solution of the exercise:** David Bowie, Absolute Beginners

<https://www.youtube.com/watch?v=UgOCjm7QOog>

### Quotations:

'On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul  
...'

'One sees, from this Essay, that **the theory of probabilities is basically just common sense reduced to calculus ...'**

Pierre-Simon Laplace, Théorie Analytique des Probabilités, 1814

'Equations are your friends, not your enemies.'

### Mathematics

It is not possible to avoid mathematics in data analysis and thus the text will contain quite a number of equations. However, I have limited the number of equations in the main text by leaving out many derivations of formulas or proofs of theorems (some can be found, however, in appendices). Equations used in the main text are discussed (interpreted) having in mind students with limited inclination to and knowledge of mathematics; one goal of my course is to teach how to read equations as a step towards how to write equations. High-school knowledge of mathematics (differentiation = calculate the slope of a curve; integration = calculate the area between a curve and the x-axis) should be enough to understand most of the main text. I will not discuss measure theory.

### Plots ('seeing is believing')

The script contains a lot of plots for displaying data, results, or functions. These plots are meant to help you developing 'a feeling' about what to expect from applying statistical procedures or what's behind a particular formula.

### Common sense

Common sense has already been mentioned by Laplace (quotation given above) and is stressed again and again by Jaynes (2003). You should use common sense to guess the results of parameter estimation or hypothesis testing. By looking at the data (plots!) you can guess, for example, (1) for linear regression whether the slope is positive or negative or not significantly different from zero or (2) whether two mean values are equal or not. If the result of the statistical procedure is different from your expectation you either made a mistake or you found a result which is most interesting. The results of statistical procedures can sometimes sound or look counter-intuitive, however, at second thought, they should be always consistent with common sense.

### Scope of the text

Statistics is a wide field that can not be covered in a single text. The number of commonly applied methods is huge (the number of different hypothesis test is most probably larger than 100) and new methods are developed in response to new challenges, for example, the large data sets produced in molecular biology. The script is restricted to two main topics as mentioned in the title: parameter estimation (including estimation of uncertainties) and hypothesis testing. Many important topics like time series analysis or ordination are left out. I have tried to explain the basic principles behind various statistical procedures and hope that this will help grasping other methods not discussed in the script.

### Main and appendix

I have split up various chapters in the parts recommended to everyone (main body of the text) and the parts necessary for a deeper understand and/or requiring more background knowledge in mathematics.

### Different schools: Bayesians versus frequentist

A difficulty for beginners as well as experienced scientists is the existence of different schools of thought on

probabilities.<sup>1</sup> The so-called Bayesians and frequentists do not agree on what is meant by probability (the degree of plausibility or the infinite limit of relative frequencies) and they have developed different methods for hypothesis testing. In the 20th century, various members of these different schools were ‘strong personalities’ and fighting each other. It required an alternation of generations to take a less emotional view, to recognize what is functioning and what not, and even to find syntheses; this is still an ongoing process. Many Bayesian methods are conceptually much more sound, however, are often computationally intensive and thus became available only with the fast development of computational capacities in the second half of the 20th century. In the script I will discuss methods based on Bayesian as well as non-Bayesian approaches. My feeling is that although significance testing does not work in general, it is doing fine for many simple hypotheses and for these cases one will make almost always identical decisions when taking into account the different scales (Fisher’s scale for p-values versus Jeffreys’ scale for Bayes’ factors). The Bayesian approach has the advantage of providing conceptual connections between different methods: from the basic rules of probability, one immediately obtains Bayes’ Theorem that is then used, by appropriate interpretation of the involved terms, for parameter estimation leading directly to least squares and generalized least squares.

### Monte Carlo simulations

Many interesting questions can not be answered by analytical mathematical methods (‘no simple formula exists or has been found so far’). Even if analytical solutions exist, the mathematical proofs are sometimes too difficult to follow by non-mathematicians. Monte Carlo simulations can often give ‘quick and dirty’ answers: ‘quick’, because a few lines of code and a bit of computer time can give amazing insights, and ‘dirty’, because it is not a proper mathematical proof. A good example is the famous question whether to divide by the sample size  $n$  or by  $n - 1$  in order to calculate (actually: estimate) the variance (Section 10.3). Monte Carlo simulations will be used in various contexts throughout the script.

### R codes

R is a very powerful computing language. Students asked for it because it is available for free and is running on various computers including PCs and MACs. Routines for many statistical procedures are available or can be added by loading additional libraries. Learning R was a bit painful because the explanations of routines and the examples given by R are often cryptic (to say the least). Fortunately, more information is available on the Web (for example, the [Quick R](#) site). I recommend to use RStudio which includes an editor, a command window (‘console’), and a window for plots, help etcetera. RStudio is available for free from the Web. I provide R codes for all calculations and plots. The results of some routines (for example, `lm()` for simple linear regression) are discussed in detail. The R codes are provided ‘as is’ without warranty of any kind. But they are mine, so you can’t sell them.

### Exercises

Exercises are spread throughout the text. Most exercises are relatively easy to solve often by writing or just modifying an already given R code. They can help learning how to apply the various procedures.

### How to read the script

If you are an absolute beginner in probability and statistics, you have to learn a number of concepts ‘at the same time’. This includes probabilities, basic rules of probabilities, statistical populations, random sampling, estimation, probability distributions, probability density functions, and a few others. The concepts are discussed in the first chapters (until and including Chapter 6) of the script which you probably have to read twice in order to obtain a full understanding. Afterwards or if you are already a bit more experienced you can jump directly into other chapters.

---

<sup>1</sup>... and on statistical inference. David Spiegelhalter (2019, p. 305) writes: “I must make now an admission on behalf of the statistical community. The formal basis for learning from data is a bit of a mess. Although there have been numerous attempts to produce a single unifying theory of statistical inference, none has been fully accepted. It is no wonder mathematicians tend to dislike teaching statistics.”

## Literature

My favorite books on data analysis/statistics:

1. David Spiegelhalter, *The Art of Statistics – Learning from Data*, 2019.  
*Excellent text for beginners (almost no equations!) and senior scientists (giving precise definitions of terms and explaining also what terms do not mean, and much more ...). Reading this book is a must for anybody interested in statistics!*
2. Edward T. Jaynes, *Probability Theory – The Logic of Science*, 2003.  
*The 'New Testament'<sup>2</sup> of the Bayesian approach to data analysis.*
3. Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, New York, John Wiley, 1971.
4. D.S. Sivia, and J. Skilling: *Data analysis: a Bayesian tutorial*, 2006.  
*A nice (and shorter than Jaynes, 2003) introduction to the Bayesian approach.*
5. Casella, G. & R.L. Berger, *Statistical Inference*, Duxbury Pacific Grove, CA, 2002.
6. Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, third edition, Chapman & Hall, London, 2020.
7. J.H. Zar, *Biostatistical Analysis*, 2010.  
*Discussing in detail (data!) the most commonly applied null hypothesis significance tests and regression models.*

Further reading (books):

1. Efron, B. & R. Tibshirani, *An introduction to the bootstrap*, CRC Press, 1994.  
*bootstrap methods were developed by Efron (1979); so Bradley Efron could be called 'Mr. Bootstrap'.*
2. Efron, B., *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press, 2010.  
*More advanced; no software included*
3. Efron, B. & T. Hastie, *Computer age statistical inference*, student edition: algorithms, evidence, and data science, Cambridge University Press, 2021.  
*Including empirical Bayes, bootstrap; more advanced; no accompanying software; available on the Web*
4. Mudelsee, M., *Climate time series analysis: classical statistical and bootstrap methods*, Second Edition, Springer, 2014.  
*time series analysis (not covered in my script); accompanying algorithms and software; written by physicist*
5. Wasserman, L., *All of statistics: a concise course in statistical inference*, Springer, 2004.  
*wide range of methods; more advanced; no accompanying software*
6. Zuur, A.F., E.N. Ieno, & G.M. Smith, *Analysing Ecological Data*, Springer, 2007.  
*available also as E-book; R codes available on the Web*
7. Zuur, A.F., E.N. Ieno, N.J. Walker, A.A. Saveliev, & G.M. Smith, *Mixed effects models and extensions in ecology with R*, Springer, New York.  
*R codes available on the Web*

Further reading (articles):

1. McNutt, M., 2014. Raising the bar. *Science*, 345(6192), pp.9-9.

*"... it is not realistic to expect that a technical reviewer, chosen for her or his expertise in the topical subject matter or experimental protocol, will also be an expert in data analysis. For that reason, with much help from the American Statistical Association, **Science** has established, effective 1 July 2014, a Statistical Board of Reviewing Editors (SBoRE), consisting of experts in various aspects of statistics and data analysis, to provide better oversight of the interpretation of observational data."*

---

<sup>2</sup>The 'Old Testament' is: Harold Jeffrey, *Theory of Probability*, 1961.

- 
2. Raftery, A.E., Carriquiry, A.L., Daniels, M.J., Gatsonis, C., Goodman, S.N., Herring, A.H. and Reid, N.M., 2023. PNAS establishes a Statistical Review Committee. *Proceedings of the National Academy of Sciences*, 120(47), p.e2317870120. (*following the lead of Marcia McNutt and Science*)  
Most other journals (including **Nature**) have not yet implemented similar procedures.

**Notation** Many authors denote random variables by capital letters and the corresponding values by lower-case letters. I do not follow this convention because it can be confusing for beginners and it is usually clear from the context whether one is referring to the random variable or its value.

Vectors are denoted by bold face letters and in the notation I do not discern between column or row vectors (which are transpose to each other) because it is usually clear from the context which form is meant or required.

#### Music, movies

Statistics is not the most favored and usually not the most entertaining course for master students in biology. Thus, from time to time, I play some music ('Absolute beginners') or mention movies ('Whatever works') that are related to the topics of my course (as, for example, the different eye colors of David Bowie to the product rule of probabilities) or just to attract again the attention of the students or because I like it.

Little Wing from The Jimi Hendrix Experience, Axis: Bold as Love (1967)

<https://www.youtube.com/watch?v=iaE4s3m8UOQ>

# Index

- Absolute Beginners (David Bowie), 678  
aha moment, 595  
AIC, 316, 349  
AIC, Akaike information criterion, 318  
AICc, 316  
Akaike information criterion, AIC, 316, 318  
Amaldi, Eduardo, 15  
angle bisection, 295  
Anscombe's quartet, 61  
Axis, 681
- Bayes factor functions, BFFs, 203  
Bayes' Theorem, 74, 79, 272  
Bayesians, 69  
Bernoulli process, 401  
Bernoulli, Jacob, 69  
Bernoulli, Theorem of , 409  
best unbiased estimator, 498  
beta function, 125, 541, 542  
BFFs, Bayes factor functions, 203  
biased estimation, 348  
biasing parameter, 343  
bimodal, 103  
binomial distribution, negative, 447  
bisection, 295  
bisection line, 301  
bisection, angle, 295  
bivariate normal PDF, 130  
BLUE, best linear unbiased estimator, 273  
Bonzo Dog Doo-Dah Band, 119  
bootstrapping, 186  
Bowie, David, 73, 678  
box plot, 46  
Breusch-Pagan test, 276
- cancer, kidney, 162  
Cauchy distribution, 458  
Cauchy prior, 211  
CDF, 432, 437  
CDF, cumulative distribution function, 100  
CDF, normal, 559  
Central Limit Theorem, 118, 141, 468  
Central Limit Theorem (CLT), 133  
central tendency, 53  
Chadwick, James, 15  
Chandrasekhar, Subrahmanyan, 15
- CLT, Central Limit Theorem, 118, 133  
coefficient of determination, 63  
common sense, 72, 84, 86, 133, 678  
common sense, Laplace, 678  
confidence bands, 278  
confidence interval, 278  
constrained linear inversion method, 342  
correlation, 63  
correlation coefficient, 515  
correlation test, 519  
correlation, Kendall, 64  
correlation, Pearson, 63  
correlation, Spearman, 64  
covariance, 58  
Cox, R.T., 69  
credibility interval, CRI, 24  
credibility set, 28, 673  
credibility sets, 200  
cross entropy, 422  
cumulative distribution function (CDF), 100
- de Moivre, Abraham, 133  
Deming regression, 265, 284  
deviance, 388  
directed divergence, 422  
dispersion, 53  
dispersion parameter, 449  
distribution, geometric, 449  
distribution-free correlation test, 519
- effect size, 211  
entropy, cross, 422  
entropy, Maximum Entropy Principle, 82  
entropy, relative, 422  
entropy, Shannon, 82  
estimator, best unbiased, 498  
Euler integral, 125  
Euler integral, 1. kind, 541, 542  
EVM, 298  
expectation, 101, 131, 483  
expectation, numeric, 201
- fatal horse kicks, 375  
Fehlerfortpflanzungsgesetz, 141  
Fermi, Enrico, 15  
Fisher, 207, 503, 512, 575

- Frank Zappa, 430  
frequentist, 69
- gamma function, incomplete, 464  
gamma PDF, standard, 473  
Gauss, C.F., 272  
Gauss-Markov theorem, 273, 348  
generating function, 559, 561  
generating function, probability, 559, 561  
geometric distribution, 449  
geometric line, 295, 301  
geometric mean, 54  
Gibrat's law, 468  
Gibrat, R., 468  
GLM, 449
- harmonic mean, 54  
Hendrix, Jimi, 681  
heteroscedastic, 273  
heteroscedasticity, 273  
heteroskedasticity test, 276  
homoscedastic, 273  
homoscedasticity, 273  
homoskedasticity test, 276  
horse, 375
- incomplete gamma function, 464  
independence, 274  
infector, 379  
infectee, 379  
inverse prediction, 289
- Jaynes, E.T., 69  
Jeffreys, 575  
Jeffreys prior, 210  
Jeffreys' prior, 211, 543  
Jeffreys, Harold, 458  
Jeffreys, reparametrization, 210  
Jeffreys, scales, 211
- kidney cancer, 162  
King, Ben E., 422  
Kolmogorov, Andrei, 69  
Kullback-Leibler information, 318
- Lagrange function, 85  
Lagrange multiplier, 85  
Laplace, P.-S., 678  
lattice Boltzmann model (LBM), 422  
least-squares, 39, 272  
Legendre, A.-M., 272  
Lennon, John, 422  
likelihood  $\mathcal{L}$ , 151  
likelihood function  $L$ , 151  
Likelihood Ratio Test (LRT), 151  
line, bisection, 301
- line, geometric, 295, 301  
linear regularization, 342  
Lineweaver-Burk transformation, 367  
link function, 399  
Little Wing, 681  
logit, 394  
lognormal PDF, 250, 252, 466  
LRT, Likelihood Ratio Test, 151
- MADN, 57  
Mann-Whitney, 559, 561  
marginal PDF, 131  
marginalization, 79, 410, 428  
Markov Chain Monte Carlo, MCMC, 602  
MaxEnt, 82  
MaxEnt distributions, 417  
Maximum Entropy Principle, 69, 82, 85  
Maximum Likelihood Estimation (MLE), 151  
MCMC, 602  
mean squared error (MSE), 175  
mean, geometric, 54  
mean, harmonic, 54  
mean, Winsorized, 54  
median, 54, 103  
Michaelis-Menten, 361  
MLE, Maximum Likelihood Estimation, 151  
mode, 103  
Moivre's equation, 162  
Monod, 361  
Monte Carlo, 437, 559, 562, 564  
Monte Carlo simulation, 124, 133, 143, 159, 161, 165, 177, 457, 508, 518, 559, 679  
MSE (mean squared error), 175  
multimodal, 103  
multivariate PDF, 130  
music: Bonzo Dog Doo-Dah Band, 119  
music: David Bowie, 678  
music: Marteria, Supernova, 19  
music: Queen, A kind of magic, 192  
music: Rolling Stones, 401  
music: Rolling Stones, Memory Motel, 297  
music: Rolling Stones, No expectation, 101  
music: Stand by me, 422  
music: Steamhammer, 353
- negation of a proposition, 70  
negative binomial distribution, 447  
neutrino, 14  
noise, 118  
non-parametric correlation test, 519  
normal PDF, bivariate, 130  
nuisance parameter, 211
- odds, 394  
OLR, 299

orthogonal regression, 627  
 overdispersion, 109, 392, 450, 674  
 p-value, physicists, 575  
 Pólya, 133  
 Pauli, Wolfgang, 15  
 PDF, lognormal, 466  
 PDF, marginal, 131  
 PDF, multivariate, 130  
 point estimator, 674  
 point null, 674  
 Poisson distribution, 93, 109, 175, 375, 450  
 Poisson distribution, zero inflated, 442  
 Poisson distribution, zero truncated, 444  
 Poisson regression, 388  
 precision, 118  
 prediction, inverse, 289  
 Principle of Indifference, 69, 77, 82, 84, 560  
 Principle of Insufficient Reason, 84  
 Principle of Maximum Relative Entropy, 422  
 Prinzip vom unzureichenden Grunde, 84  
 prior, Cauchy, 211  
 prior, Jeffreys, 210, 211  
 probability generating function, 559, 561  
 probability, basic rules, 70  
 probability, conditional, 70  
 probability, generalized sum rule, 73  
 probability, product rule, 72  
 probability, sum rule, 73  
 probit function, 116  
 product rule, 72  
 product rule, simplified, 72  
 propagation of uncertainties, 133  
 quantile function, 100, 116  
 quantile(), 665  
 R, random numbers, 93  
 random numbers, 118, 431  
 random numbers, tent distribution, 431  
 ranks, signed, 557  
 Redfield ratios, 287, 296  
 reduced major-axis, 295  
 regression, orthogonal, 627  
 relative entropy, 422  
 relative frequencies, 15  
 residual deviance, 388  
 residuals, 273, 674  
 rf(), 94  
 ridge regression, 342  
 ridge trace, 343  
 rnorm(), 94  
 RStudio, 679  
 rt(), 94  
 runif(), 93

sample space, 98  
 scales of evidence, 211  
 Scheffé bands, 278  
 set.seed(), 93  
 Shannon entropy, 82, 85  
 signed ranks, 557  
 small schools, 162  
 standard error, 575  
 standard error of estimate, 278  
 standard error of the difference, 570  
 standard error of the regression, 278  
 standard gamma PDF, 473  
 standardized uniform distribution, 120  
 sufficient statistics, 192, 543  
 symmetry, 84  
 tent distribution, 431  
 test, Breusch-Pagan, 276  
 test, correlation, 515, 519  
 test, heteroskedasticity, 276  
 test, homoskedasticity, 276  
 Theorem of Bernoulli, 409  
 Tikhonov regularization, 342  
 time series, 73  
 trade-off, 318, 343  
 transformation law, 432  
 unbiased, 491  
 unbiased estimator, 348  
 underdispersion, 109  
 unimodal, 103  
 variance, 165, 318  
 variance, estimation of, 165  
 von Mises, Richard, 69  
 We are normal, 119  
 Wilkinson notation, 322  
 Winsorized mean, 54  
 Zappa, Frank, 430, 530  
 Zero Inflated Poisson distribution, 442  
 zero inflation parameter, 534  
 zero truncated Poisson distribution, 444  
 ZIP, Zero Inflated Poisson, 442

# Bibliography

- [1] Abdi, Hervé and Molin, Paul. Lilliefors/Van Soest's test of normality. *Encyclopedia of measurement and statistics*, pages 540–544, 2007.
- [2] Adcock, Robert James. A problem in least squares. *The Analyst*, 5(2):53–54, 1878.
- [3] Adrain, Robert. Research concerning the probabilities of the errors which happen in making observations. *The Analyst (or Mathematical Museum)*, 1(4):93–109, 1808.
- [4] Aitkin, M.A., B. Francis, and J. Hinde. *Statistical modelling in GLIM 4*, volume 32. Oxford University Press, USA, 2005.
- [5] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [6] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012.
- [7] Barker, F., Y.C. Soh, and R.J. Evans. Properties of the geometric mean functional relationship. *Biometrics*, pages 279–281, 1988.
- [8] Barlow, J., T. Gerrodette, and G. Silber. First estimates of vaquita abundance. *Marine Mammal Science*, 13(1):44–55, 1997.
- [9] Barlow, R.J. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. Wiley, 1989 [reprinted 1999].
- [10] Barlow, R.J. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. Wiley, 1999.
- [11] Bartlett, Michael S. Fitting a straight line when both variables are subject to error. *Biometrics*, 5(3):207–212, 1949.
- [12] Bartlett, M.S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937.
- [13] Bates, Douglas and Mächler, Martin and Bolker, Ben and Walker, Steve. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.
- [14] Bates, Douglas and Mächler, Martin and Bolker, Ben and Walker, Steve. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [15] Belenky, Gregory and Wesensten, Nancy J and Thorne, David R and Thomas, Maria L and Sing, Helen C and Redmond, Daniel P and Russo, Michael B and Balkin, Thomas J. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12, 2003.
- [16] Bellerby, R.G.J., A. Olsen, T. Johannessen, and P. Croot. A high precision spectrophotometric method for on-line shipboard seawater pH measurements: the automated marine pH sensor (AMpS). *Talanta*, 56:61–69, 2002.

- [17] Bendat, J.S. and A.G. Piersol. *Random Data: Analysis and Measurement Procedures*. Wiley-Interscience, New York, 1972.
- [18] Benjamin, D.J., J.O. Berger, M. Johannesson, B.A. Nosek, E.-J. Wagenmakers, R. Berk, K.A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C.D. Chambers, M. Clyde, T.D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A.P. Field, M. Forster, E.I. George, R. Gonzalez, S. Goodman, E. Green, D.P. Green, A. Greenwald, J.D. Hadfield, L.V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D.J. Hruschka, K. Imai, G. Imbens, J.P. A. Ioannidis, M. Jeon, J.H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S.E. Maxwell, M. McCarthy, D. Moore, S.L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T.H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F.D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D.J. Watts, C. Winship, R.L. Wolpert, Y. Xie, C. Young, J. Zinman and V.E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6, 2018.
- [19] Bennett, Keith D. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132(1):155–170, 1996.
- [20] Berger, James O. *Statistical decision theory and Bayesian analysis*. Springer, New York, 1985.
- [21] Berger, James O and Bernardo, José M and Sun, Dongchu. The formal definition of reference priors. *The Annals of Statistics*, pages 905–938, 2009.
- [22] Berger, J.O. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.
- [23] Berger, J.O. An Overview of Objective Bayesian Analysis [presentation]. available from homepage of jim Berger: <http://www2.stat.duke.edu/~berger/>, 2011.
- [24] Berger, J.O., L. Brown, and R. Wolpert. A unified conditional frequentist and bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.*, 22:1787–1807, 1994.
- [25] Bernardo, J.M. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.
- [26] Bernardo, José M. Reference analysis. *Handbook of statistics*, 25:17–90, 2005.
- [27] Bernoulli, Jakob. *Ars Conjectandi*. Thurnisiorum, Basel, 1713 [reprinted in: Die Werke von Jakob Bernoulli, Vol. 3, p. 107-286, Birkhaeuser, Basel, 1975].
- [28] Best, DJ and Roberts, DE. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379, 1975.
- [29] Birnbaum, Z.W. On a use of the Mann-Whitney statistic. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.
- [30] Bliss, Chester I. The method of probits. *Science*, 79(2037):38–39, 1934.
- [31] Borcard, D., F. Gillet, and P. Legendre. *Numerical ecology with R*. Springer, 2011.
- [32] Bowden, David C and Graybill, Franklin A. Confidence bands of uniform and proportional width for linear models. *Journal of the American Statistical Association*, 61(313):182–198, 1966.
- [33] Bretthorst, G.L. An introduction to model selection using probability theory as logic. In Heidbreder, G.R., editor, *Maximum Entropy and Bayesian Methods - Proceedings of the 13th International Workshop, Santa Barbara, California, August 1-5, 1993*, pages 1–42. Kluwer Academic Publishers, Dodrecht, Holland, 1996.
- [34] Burnham, Kenneth P and Anderson, David R and Huyvaert, Kathryn P. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65(1):23–35, 2011.

- [35] Burnham, K.P. and D.R. Anderson. *Model Selection and Multimodel Inference – A Practical Information-Theoretic Approach*. Second Edition, Springer-Verlag, 2002.
- [36] Campbell, Janet W. The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, 100(C7):13237–13254, 1995.
- [37] Carvalho, Luis. An improved evaluation of Kolmogorov's distribution. *J. Stat. Softw.*, 2015a.
- [38] Carvalho, Luis. Package 'kolmim'. 2015b.
- [39] Casella, George and Roger L. Berger. *Statistical inference*. Duxbury Pacific Grove, CA, 2002 [reprinted 2015].
- [40] Castelvecchi, Davide and Gibney, Elizabeth. Particle's surprise mass threatens to upend the standard model. *Nature*, 7 April 2022.
- [41] CDF Collaboration and Aaltonen, T and Amerio, S and Amidei, D and Anastassov, A and Annovi, A and Antos, J and Apollinari, G and Appel, JA and Arisawa, T and others. High-precision measurement of the W boson mass with the CDF II detector. *Science*, 376(6589):170–176, 2022.
- [42] Cheng, Chi-Lun and Van Ness, John W. *Statistical regression with measurement error*. Arnold, 1999.
- [43] Clarke, R.D. An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72(3):481–481, 1946.
- [44] Conover, W.J., M.E. Johnson, and M.M. Johnson. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4):351–361, 1981.
- [45] Conway, Richard W and Maxwell, William L. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136, 1962.
- [46] Cowan Jr, C.L., F. Reines, F.B. Harrison, H.W. Kruse and A.D. McGuire. Detection of the free neutrino: a confirmation. *Science*, 124(3212):103–104, 1956.
- [47] Cowles, M. and C. Davis. On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5):553–558, 1982.
- [48] Cowpertwait, P.S.P. and A.V. Metcalfe. *Introductory time series with R*. Springer Science & Business Media, 2009.
- [49] Cox, David Roxbee. *Principles of Statistical Inference*. Cambridge University Press, 2006.
- [50] Cox, D.R. Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, pages 357–372, 1958.
- [51] Cox, R.T. Probability, frequency and reasonable expectation. *Am. J. Phys.*, 14:1–14, 1946.
- [52] Cox, R.T. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.
- [53] Cox, R.T. *Algebra of Probable Inference*. The Johns Hopkins University Press, 1961.
- [54] Cox, R.T. On inference and inquiry, an essay in inductive logic. In Levine, R.D. and M. Tribus, editor, *The Maximum Entropy Formalism*, pages 119–167. MIT Press, Cambridge, 1979.
- [55] Daniels, C.J., A.J. Poulton, W.M. Balch, E. Marañón, T. Adey, B.C. Bowler, P. Cermeño, A. Charalambopoulou, D.W. Crawford, D. Drapeau, and others. A global compilation of coccolithophore calcification rates. *Earth System Science Data*, 10(4):1859–1876, 2018.
- [56] David, Herbert A. First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, 49(2):121–133, 1995.
- [57] de Moivre, A. (1733). Approximatio ad Summam Terminorum Binomii  $(a + b)^n$  in Seriem expansi; Photographic reproduction in Archibald, R.C. *Isis*, 8:671–683, 1926.

- [58] De Moivre, Abraham. *Miscellanea analytica*. *Tonson and Watts, London*, 1730, 1967.
- [59] Deming, William Edwards. *Statistical adjustment of data*. Wiley, 1943.
- [60] Dettinger, M.D. and M. Ghil. Seasonal and interannual variations of atmospheric CO<sub>2</sub> and climate. *Tellus B*, 50(1):1–24, 1998.
- [61] Di Bucchianico, A. Combinatorics, computer algebra and the Wilcoxon-Mann-Whitney test. *Journal of Statistical Planning and Inference*, 79(2):349–364, 1999.
- [62] Dickson, A.G., J.D. Afghan, and G.C. Anderson. Reference materials for oceanic CO<sub>2</sub> analysis: a method for the certification of total alkalinity. *Marine Chemistry*, 80(2):185–197, 2003.
- [63] Dobson, Annette J. and Barnett, Adrian. *An introduction to generalized linear models*. CRC press, 2008.
- [64] Draper, N.R. Straight line regression when both variables are subject to error. *Kansas State University Conference on Applied Statistics in Agriculture*, pages 1–18, 1992.
- [65] Draper, N.R. and H Smith. *Applied Regression Analysis, Third Edition*. John Wiley and Sons, New York, 1998.
- [66] Draper, N.R and H. Smith. *Applied regression analysis*. John Wiley & Sons, 2014.
- [67] Efron, B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.
- [68] Efron, B. and A. Gous. Scales of evidence for model selection: Fisher versus Jeffreys. *IMS Lecture Notes-Monograph Series, Hayward, CA*, pages 208–256, 2001.
- [69] Efron, Bradley. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010 [paperback edition 2013].
- [70] Efron, Bradley. Bayes' theorem in the 21st century. *Science*, 340(6137):1177–1178, 2013.
- [71] Efron, Bradley and Hastie, Trevor. *Computer age statistical inference – algorithms, evidence, and data science*, volume 6. Cambridge University Press, 2021 [first published 2016].
- [72] Efron, Bradley and Tibshirani, Robert J. *An introduction to the bootstrap*. CRC Press, 1994.
- [73] Eguchi, Tomoharu and Gerrodette, Tim. A Bayesian approach to line-transect analysis for estimating abundance. *Ecological Modelling*, 220(13-14):1620–1630, 2009.
- [74] Fahrmeir, L., T. Kneib, S. Lang, and B. Marx. *Regression – Models, Methods and Applications*. Springer, 2013.
- [75] Firth, D. Generalized linear models. *Statistical theory and modelling*, pages 55–82, 1991.
- [76] Fisher, Ronald A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.
- [77] Fisher, Ronald A. On the interpretation of  $\chi^2$  from contingency tables, and the calculation of  $P$ . *Journal of the Royal Statistical Society*, pages 87–94, 1922b.
- [78] Fisher, Ronald Aylmer. *Statistical Methods for Research Workers, [first edition]*. Oxford University Press, 1925.
- [79] Fuller, Wayne A. *Measurement error models*. John Wiley & Sons, 1987.
- [80] Gafarian, A. V. Confidence bands in straight line regression. *Journal of the American Statistical Association*, 59(305):182–213, 1964.
- [81] Garcia, H.E., R.A. Locarnini, T.P. Boyer, J.I. Antonov, M.M. Zweng, O.K. Baranova, and D.R. Johnson. World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate), S. Levitus, Ed. NOAA Atlas NESDIS 71, U.S. Government Printing Office, Washington, D.C., 398 pp, 2010.

- [82] Gauss, K.F. *Theoria Motus Corporum Celestium*. Perthes, Hamburg, 1809 [English translation: Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections, Dover, New York, 1963].
- [83] Geider, Richard and La Roche, Julie. Redfield revisited: variability of C: N: P in marine microalgae and its biochemical basis. *European Journal of Phycology*, 37(1):1–17, 2002.
- [84] Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*, 3rd ed. Chapman & Hall, London, 2020.
- [85] Gerrodette, Tim. Inference without significance: measuring support for hypotheses rather than rejecting them. *Marine Ecology*, 32(3):404–418, 2011.
- [86] Gerrodette, Tim and Taylor, Barbara L and Swift, René and Rankin, Shannon and Jaramillo-Legorreta, Armando M and Rojas-Bracho, Lorenzo. A combined visual and acoustic estimate of 2008 abundance, and change in abundance since 1997, for the vaquita, *Phocoena sinus*. *Marine Mammal Science*, 27(2):E79–E100, 2011.
- [87] Gibson, Wendy M. and Jowett, Geoffrey H. Three-Group Regression Analysis: Part II. Multiple Regression Analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 6(3):189–197, 1957.
- [88] Gibson, W.M. and G.H. Jowett. Three-Group Regression Analysis: Part I. Simple Regression Analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 6(2):114–122, 1957.
- [89] Gilks, W.R., S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.
- [90] Gillard, Jonathan. An overview of linear structural models in errors in variables regression. *REVSTAT–Statistical Journal*, 8(1):57–80, 2010.
- [91] Gillard, Jonathan. Method of moments estimation in linear regression with errors in both variables. *Communications in Statistics-Theory and Methods*, 43(15):3208–3222, 2014.
- [92] Gillard, Jonathan William. *Errors in variables regression: What is the appropriate model?* Cardiff University (United Kingdom), 2007.
- [93] Gini, Corrado. . *Metron*, 1(3):63, 1921.
- [94] Good, IJ. What are degrees of freedom? *The American Statistician*, 27(5):227–228, 1973.
- [95] Good, Irving J. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963.
- [96] Greenacre, M. and R. Primicerio. Multivariate data analysis for ecologists. *Foundation BBVA*, Madrid, 2013.
- [97] Greenwood, Major and Yule, G. Udny. The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine*, 8(Sect Epidemiol State Med):113, 1915.
- [98] Guisan, A., T.C. Edwards Jr, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3):89–100, 2002.
- [99] Gurland, John and Tripathi, Ram C. A simple approximation for unbiased estimation of the standard deviation. *The American Statistician*, 25(4):30–32, 1971.
- [100] Harremoës, Peter. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Transactions on Information Theory*, 47(5):2039–2041, 2001.
- [101] Haussner, Rudolf. *Kommentare zu 'Wahrscheinlichkeitsrechnung (Ars Conjectandi)' von Jakob Bernoulli*. Wilhelm Engelmann, Leipzig, 1899.
- [102] Helmert, F.R. Über die Bestimmung des wahrscheinlichen Fehlers aus einer endlichen Anzahl wahrer Beobachtungsfehler. *Zeitschrift für Mathematische Physik*, 20:300–303, 1875.

- [103] Helmert, F.R. Über die Formeln für den Durchschnittsfehler. *Astronomische Nachrichten*, 85(22-23):354–366, 1875.
- [104] Helmert, F.R. Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit in Zusammenhang stehende Fragen. *Zeitschrift für Mathematische Physik*, 21:192–218, 1876a.
- [105] Helmert, F.R. Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers direkter Beobachtungen gleicher Genauigkeit. *Astronomische Nachrichten*, 88:113–132, 1876b.
- [106] Hilbe, J.M. *Negative binomial regression*. Cambridge University Press, 2011.
- [107] Hildreth, Clifford. Bayesian statisticians and remote clients. *Econometrica: Journal of the Econometric Society*, pages 422–438, 1963.
- [108] Holtzman, Wayne H. The unbiased estimate of the population variance and standard deviation. *The American journal of psychology*, 63(4):615–617, 1950.
- [109] Howell, David C. Generating Data with a Specified Correlation. [http://www.uvm.edu/~dhowell/StatPages/More\\_Stuff/CorrGen.html](http://www.uvm.edu/~dhowell/StatPages/More_Stuff/CorrGen.html), 2013.
- [110] Huang, Huei-Chung and Niu, Yi and Qin, Li-Xuan. Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software: Supplementary Issue: Sequencing Platform Modeling and Analysis. *Cancer informatics*, 14:CIN-S21631, 2015.
- [111] Huber, Peter J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- [112] Huber, Peter J. The 1972 Wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, pages 1041–1067, 1972.
- [113] Huber, Peter J. and E.M. Ronchetti. *Robust statistics*. Springer, 2011.
- [114] Hurvich, Clifford M and Tsai, Chih-Ling. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.
- [115] Husmann, Eva and Klaas, Christine. Testing the use of the silica deposition fluorescent probe PDMPO to estimate in situ growth rates of diatoms. *Limnology and Oceanography: Methods*, 20(9):568–580, 2022.
- [116] Isobe, Takashi and Feigelson, Eric D and Akritas, Michael G and Babu, Gutti Jogesh. Linear regression in astronomy. *The astrophysical journal*, 364:104–113, 1990.
- [117] Jaramillo-Legorreta, Armando and Rojas-Bracho, Lorenzo and Brownell Jr, Robert L and Read, Andrew J and Reeves, Randall R and Ralls, Katherine and Taylor, Barbara L. Saving the vaquita: immediate action, not more data. *Conservation Biology*, pages 1653–1655, 2007.
- [118] Jaynes, E.T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957a.
- [119] Jaynes, E.T. Information theory and statistical mechanics. ii. *Physical review*, 108(2):171, 1957b.
- [120] Jaynes, E.T. Where do we stand on maximum entropy? (1978). In Rosenkrantz, R.D., editor, *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, pages 210–314. Kluwer, Dordrecht, 1989.
- [121] Jaynes, E.T. *Probability Theory - The Logic of Science*. Cambridge University Press, 727 pp., 2003.
- [122] Jeffreys, Harold. *Theory of Probability*. Clarendon Press, Oxford, 1939.
- [123] Jeffreys, Harold. *Theory of Probability*. second edition, Clarendon Press, Oxford, 1948.
- [124] Jeffreys, Harold. *Theory of Probability*. third edition, Oxford Clarendon Press, 1961 [reprinted 2003].
- [125] Jenkins, G.M. and D.G. Watts. *Spectral analysis*, pp. 541. Holden Day, 1968.

- [126] Jitjareonchai, J.J., P.M. Reilly, T.A. Duever, and D.B. Chambers. Parameter estimation in the error-in-variables models using the Gibbs sampler. *The Canadian Journal of Chemical Engineering*, 84(1):125–138, 2006.
- [127] Johnson, Kenneth A and Goody, Roger S. The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry*, 50(39):8264–8269, 2011.
- [128] Johnson, V.E., S. Pramanik, and R. Shudde. Bayes factor functions for reporting outcomes of hypothesis tests. *Proceedings of the National Academy of Sciences*, 120(8):e2217331120, 2023.
- [129] Kahneman, Daniel. *Thinking, fast and slow*. Macmillan, 2011.
- [130] Kass, Robert E and Wasserman, Larry. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.
- [131] Keeling, Ralph F. Measuring correlations between atmospheric oxygen and carbon dioxide mole fractions: A preliminary study in urban air. *Journal Of Atmospheric Chemistry*, 7(2):153–176, 1988.
- [132] Kermack, K.A. and J.B.S. Haldane. Organic correlation and allometry. *Biometrika*, 37(1/2):30–41, 1950.
- [133] Keynes, J.M. *A Treatise on Probability*. Macmillan, London, 1921 [reprinted by Harper & Row, New York, 1962].
- [134] Kolmogoroff, A.N. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin [reprinted 1973, English translation: Foundations of the theory of probability, 1950], 1933.
- [135] Kullback, S. and R.A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [136] Kullback, Solomon. Information and statistics. *J. Wiley, New York*, 1959 [reprinted by Dover 1968].
- [137] Kummell, C.H. Reduction of observation equations which contain more than one observed quantity. *The Analyst*, pages 97–105, 1879.
- [138] Kunugi, T., T. Tamura, T and T. Naito. New acetylene process uses hydrogen dilution. *Chemical Engineering Progress*, 57(11):43–49, 1961.
- [139] Lambert, Diane. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.
- [140] Laplace, Pierre Simon. *Essai philosophique sur les probabilités*. Courcier Imprimeur, Paris, 1814.
- [141] Leamer, Edward E. Multicollinearity: a Bayesian interpretation. *The review of economics and statistics*, pages 371–380, 1973.
- [142] Leamer, Edward E. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated, 1978.
- [143] Leemis, Lawrence M. Relationships among common univariate distributions. *The American Statistician*, 40(2):143–146, 1986.
- [144] Legendre, Adrien-Marie. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.
- [145] Legendre, P. and Legendre, L.F.J. *Numerical ecology*. Amsterdam Elsevier Science, 1983.
- [146] Legendre, P. and Legendre, L.F.J. *Numerical ecology, Third English Edition*, pp. 990. Elsevier, 2012.
- [147] Lehmann, E.L. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Amer. Statist. Assoc.*, 88:1242–1249, 1993.
- [148] Lehmann, Erich L. and Casella, George. *Theory of point estimation*. Springer Science & Business Media, 2006.

- [149] Levene, Howard. Robust tests for equality of variances. In Olkin, I., editor, *Contributions to probability and statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press, 517 pp, 1960.
- [150] Lewis, Michael. *The Undoing Project*. Penguin Books, 2017.
- [151] Liang, Feng and Paulo, Rui and Molina, German and Clyde, Merlise A and Berger, Jim O. Mixtures of  $\text{g}$  priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.
- [152] Lilliefors, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.
- [153] Limpert, Eckhard, Stahel, Werner A and Abbt, Markus. Log-normal distributions across the sciences: keys and clues. *BioScience*, 51(5):341–352, 2001.
- [154] Lineweaver, H. and D. Burk. The determination of enzyme dissociation constants. *Journal of the American chemical society*, 56(3):658–666, 1934.
- [155] Lisman, JHC and Van Zuylen, MCA. Note on the generation of most probable frequency distributions. *Statistica Neerlandica*, 26(1):19–23, 1972.
- [156] Lord, Dominique and Geedipally, Srinivas Reddy and Guikema, Seth D. Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis: An International Journal*, 30(8):1268–1276, 2010.
- [157] Loredo, Thomas J. From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In *Maximum entropy and Bayesian methods*, pages 81–142. Springer, 1990.
- [158] Loredo, Thomas J and Lamb, Donald Q. Bayesian analysis of neutrinos observed from supernova SN 1987A. *Physical Review D*, 65(6):063002, 2002.
- [159] Lüroth, J. Vergleichung von zwei Werthen des wahrscheinlichen Fehlers. *Astronomische Nachrichten*, 87(14):209–220, 1876.
- [160] MacArthur, Robert H. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, 43(3):293, 1957.
- [161] Mann, Henry B. and Whitney, Donald R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- [162] Maronna, R., D. Martin, and V. Yohai. *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.
- [163] Marquardt, Donald W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.
- [164] Marquardt, D.W. and R.D. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.
- [165] Martin, T.G., B.A. Wintle, J.R. Rhodes, P.M. Kuhnert, S.A. Field, S.J. Low-Choy, A.J. Tyre, and H.P. Possingham. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8(11):1235–1246, 2005.
- [166] McClelland, H.L.O., I. Halevy, D.A. Wolf-Gladrow, D. Evans, and A.S. Bradley. Statistical uncertainty in paleoclimate proxy reconstructions. *Geophysical Research Letters*, 48(15):e2021GL092773, 2021.
- [167] McCornack, Robert L. Extended tables of the Wilcoxon matched pair signed rank statistic. *Journal of the American Statistical Association*, 60(311):864–871, 1965.
- [168] McCullagh, P. and A.J. Nelder. *Generalized linear models, Second Edition*. London England Chapman and Hall, 1989.
- [169] McNutt, Marcia. Raising the bar. *Science*, 345(6192):9–9, 2014.

- [170] McShane, B.B., D. Gal, A. Gelman, C. Robert, and J.L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.
- [171] Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- [172] Metropolis, Nicholas. The beginning of the Monte Carlo method. *Los Alamos Science*, 15(584):125–130, 1987.
- [173] Michaelis, Leonor and Menten, Maud L. Die Kinetik der Invertinwirkung. *Biochem. Z.*, 49:333–369, 1913.
- [174] Miller, RG. *Simultaneous statistical inference*, Vol. 196. Springer, 1981.
- [175] Molin, Paul and Abdi, Hervé. New table and numerical approximations for Kolmogorov-Smirnov/Lilliefors/Van Soest normality test. Technical report, Technical report, University of Bourgogne, 1998.
- [176] Monod, J. *Recherches sur la Croissance des Cultures Bactériennes*. Hermann, Paris, 1942.
- [177] Monod, J. The growth of bacterial cultures. *Annual Review of Microbiology*, 3:371, 1949.
- [178] Montgomery, Douglas C. and Peck, Elizabeth A. *Introduction to linear regression analysis*. John Wiley & Sons, 1982.
- [179] Montgomery, Douglas C and Peck, Elizabeth A and Vining, G Geoffrey. *Introduction to linear regression analysis*. John Wiley & Sons, 2015.
- [180] Mudelsee, M. *Climate time series analysis: classical statistical and bootstrap methods, Second Edition*. Springer, 2014.
- [181] Mudelsee, Manfred. Unbiased proxy calibration. *Mathematical Geosciences*, pages 1–28, 2023.
- [182] Narasimhan, T.N. The dichotomous history of diffusion. *Physics today*, 62(7):48–53, 2009.
- [183] Nelder, J.A. and R.W.M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.
- [184] Neyman, J. and E.S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference, part i. and part ii. *Biometrika*, 20A:175–240, 263–294, 1928.
- [185] Neyman, J. and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. London Ser. A*, 231:239–337, 1933. introduction of the power concept.
- [186] Osborne, Jason W and Waters, Elaine. Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, 8(2), 2002.
- [187] Park, S.Y. and A.K. Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, 2009.
- [188] Pearl, J. and D. Mackenzie. *The Book of Why*. Penguin, 2019.
- [189] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 5(50):157–175, 1900.
- [190] Pearson, Karl. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- [191] Pinheiro, J. and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2000.
- [192] Plackett, R.L. Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59–72, 1983.

- [193] Pólya, G. Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Math. Zeit.*, 8:171–181, 1920 [reprinted in Pólya, 1984, Vol. IV].
- [194] Pynchon, Thomas. *Gravity's rainbow*. Viking Press, 1973.
- [195] Pynchon, Thomas. *Gravity's rainbow*. Pan Books, London, 1975.
- [196] Raftery, A.E., A.L. Carriquiry, M.J. Daniels, C. Gatsonis, S.N. Goodman, A.H. Herring, and N.M. Reid. Pnas establishes a statistical review committee. *Proceedings of the National Academy of Sciences*, 120(47):e2317870120, 2023.
- [197] Raiffa, Howard and Schlaifer, Robert. *Applied Statistical Decision Theory*. Division of Research, Harvard Business School, Boston, MA, 1961.
- [198] Redfield, Alfred C. The biological control of chemical factors in the environment. *American scientist*, 46(3):230A–221, 1958.
- [199] Redfield, Alfred C. The influence of organisms on the composition of seawater. *The sea*, 2:26–77, 1963.
- [200] Redfield, Alfred Clarence. On the proportions of organic derivatives in a sea water and their relation to the composition of plankton. *James Johnstone memorial volume*, pages 177–192, 1934.
- [201] Riggs, D.S., J.A. Guarnieri, and S. Addelman. Fitting straight lines when both variables are subject to error. *Life sciences*, 22(13-15):1305–1360, 1978.
- [202] Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.
- [203] Robert, C. and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 1998.
- [204] Robert, C. and G. Casella. *Introducing Monte Carlo Methods with R*. Springer Science & Business Media, 2009.
- [205] Robert, C. and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- [206] Robinson, M.D., D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
- [207] Rouder, Jeffrey N and Morey, Richard D and Speckman, Paul L and Province, Jordan M. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5):356–374, 2012.
- [208] Rouder, J.N., P.L. Speckman, D. Sun, R.D. Morey, and G. Iverson. Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009.
- [209] Rutherford, E., H. Geiger, and H. Bateman. The probability variations in the distribution of  $\alpha$  particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20(118):698–707, 1910.
- [210] Sañudo-Wilhelmy, S. A., C. J. Gobler, M. Okbamichael, and G. T. Taylor. Regulation of phytoplankton dynamics by vitamin B<sub>12</sub>. *Geophys. Res. Lett.*, 33, L04604:doi:10.1029/2005GL025046, 2006.
- [211] Sañudo-Wilhelmy, SA and Gobler, CJ and Okbamichael, M and Taylor, GT. Regulation of phytoplankton dynamics by vitamin B<sub>12</sub>. *Geophysical research letters*, 33(4), 2006.
- [212] Scheffé, Henry. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110, 1953, [corrections 1959].
- [213] Scheffé, Henry. *The Analysis of Variance*. Wiley, New York, 1959.
- [214] Shannon, C.E. A Mathematical Theory of Communication. *Bell Systems Tech. J.*, 27:379–423, 623–645, 1948.
- [215] Shore, J.E. and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory, IEEE Transactions on*, 26(1):26–37, 1980.

- [216] Sivia, Devinderjit and Skilling, John. *Data analysis: a Bayesian tutorial*. Oxford University Press, USA, 2006.
- [217] Spiegelhalter, David. *The Art of Statistics – Learning from Data*. Penguin Books, 2019.
- [218] Spiegelhalter, David. *The Art of Uncertainty: How to Navigate Chance, Ignorance, Risk and Luck*. Penguin, 2024.
- [219] Sterner, Robert W. and Elser, James J. *Ecological stoichiometry*. Princeton University Press, 2017.
- [220] Stigler, Stephen M. *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press, 1999.
- [221] Stigler, Stephen M. *Statistics on the table: The history of statistical concepts and methods*. third printing 2002, Harvard University Press, 2002.
- [222] Student [William Sealy Gosset]. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.
- [223] Taylor, B.N. and C.E. Kuyatt. NIST Technical Note 1297. *Guidelines for evaluating and expressing the uncertainty of NIST measurement results*, pages i–24, 1994.
- [224] Teissier, Georges. La relation d'allometrie sa signification statistique et biologique. *Biometrics*, 4(1):14–53, 1948.
- [225] Tikhonov, Andrey Nikolayevich. On the stability of inverse problems [in Russian]. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.
- [226] Trafimow, D. Editorial. *Basic and Applied Social Psychology*, 36(1):1–2, 2014.
- [227] Trafimow, D. and M. Marks. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015.
- [228] Tukey, J.W. *Exploratory data analysis (preliminary edition)*. 1970.
- [229] Tukey, J.W. *Exploratory data analysis*. Addison-Wesley, 1977.
- [230] Tyrrell, Toby. The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature*, 400(6744):525–531, 1999.
- [231] van de Wiel, Mark Adrianus. *Exact distributions of distribution-free test statistics*. Ph. D. Thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2000.
- [232] van Soest, J. Some experimental results concerning tests of normality. *Statistica Neerlandica*, 21(1):91–97, 1967.
- [233] Vance, Erik. Requiem for the Vaquita. *Scientific American*, 317(2):30–39, 2017.
- [234] von Mises, R. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1):52–99, 1919.
- [235] von Mises, Richard. *Wahrscheinlichkeit, Statistik und Wahrheit*. Springer-Verlag, 1928.
- [236] von Storch, H. and F.W. Zwiers. *Statistical analysis in climate research*. Cambridge University Press, 484 pp., 2001 [reprinted 2003].
- [237] Vrieze, Scott I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228, 2012.
- [238] Wainer, Howard. The most dangerous equation. *American Scientist*, 95(3):249, 2007.
- [239] Wainer, Howard and Zwerling, Harris L. Evidence that smaller schools do not improve student achievement. *Phi Delta Kappan*, 88(4):300–303, 2006.

- [240] Wald, Abraham. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.
- [241] Walker, Helen M. Degrees of freedom. *Journal of Educational Psychology*, 31(4):253–269, 1940.
- [242] Wang, J., W.W. Tsang, and G. Marsaglia. Evaluating Kolmogorov’s distribution. *Journal of Statistical Software*, 11(3):1–11, 2004.
- [243] Wasserman, Larry. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.
- [244] Wasserstein, R.L. (ed.). ASA statement on statistical significance and *p*-values. *The American Statistician*, 70(2):131–133, 2016.
- [245] Weber, P., K. Binder, S. Krauss. Why can only 24% solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness. *Frontiers in psychology*, 9, 2018.
- [246] Wehrl, A. General properties of entropy. *Reviews of Modern Physics*, 50(2):221, 1978.
- [247] Weissgerber, T.L., N.M. Milic, S.J. Winham, and V.D. Garovic. Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS biology*, 13(4):e1002128, 2015.
- [248] Wickham, H. and L. Stryjewski. 40 years of boxplots. *Am. Statistician*, page 2011, 2011.
- [249] Wilcoxon, Frank. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.
- [250] Wilcoxon, Frank and Katti, S.K. and Wilcox, Roberta A. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1970.
- [251] Wilkinson, G.N. and C.E. Rogers. Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22(3):392–399, 1973.
- [252] Williams, Matt N, Grajales, Carlos Alberto Gómez and Kurkiewicz, Dason. Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, 18(1):11, 2013.
- [253] Wolf-Gladrow, D., I. Borrione, G. Shirodkar, M.U. Gauns, M. Vadlamani, and C. Klaas. In a sea of crumbling icebergs. *Journal of Glaciology*, 71:e76, 2025.
- [254] Wolf-Gladrow, D.A. *Lattice-gas cellular automata and lattice Boltzmann models: an introduction*. Springer Science & Business Media, 2000.
- [255] Woosley, Stan and Janka, Thomas. The physics of core-collapse supernovae. *Nature Physics*, 1(3):147–154, 2005.
- [256] Yang, Yuhong. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.
- [257] York, D. Least-squares fitting of a straight line. *Canadian Journal of Physics*, 44(5):1079–1086, 1966.
- [258] Yu, C.H. Degrees of freedom. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 363–365. Springer, Berlin, Heidelberg, 2011.
- [259] Zar, J.H. *Biostatistical Analysis*. Prentice Hall, New Jersey, 5th edition, 2010.
- [260] Zeebe, R.E. and D. Wolf-Gladrow. *CO<sub>2</sub> in Seawater: Equilibrium, Kinetics, Isotopes: Equilibrium, Kinetics, Isotopes*. Elsevier, 2001.
- [261] Zellner, Arnold. *An introduction to Bayesian analysis in econometrics*. Wiley, New York [reprinted 1996 as ‘Wiley Classics’ edition], 1971.
- [262] Zuur, A.F., E.N. Ieno, and G.M. Smith. *Analysing Ecological Data*. Springer, New York, 2007.
- [263] Zuur, A.F., E.N. Ieno, N.J. Walker, A.A. Saveliev, and G.M. Smith. *Mixed effects models and extensions in ecology with R*. Springer, New York, 2009.