

H. P. Wolf, P. Naeve, V. Tiemann: Statistik — aktiv mit R Überschriften und Graphiken zum Buch

Abbildungsverzeichnis	9
Vorwort	15
1 Datenanalyse? Daten? Statistik?	17
1.1 Was für Daten gibt es?	18
1.2 Wo kommen Daten her?	19
2 Univariate, exploratorische Analyse	21
2.1 Häufigkeitstabellen und deren Darstellung	22
2.2 Auswertung der Urliste: Lage und Variabilität	33
2.3 Die empirische Verteilungsfunktion	44
2.4 Besondere Strukturen einer Verteilung	48
2.5 Konzentrationsmessung – LORENZ und GINI	54
2.6 Fallstudie – das 6 aus 49 Lotto	58
3 Bivariate, exploratorische Analyse	69
3.1 Korrelation von Merkmalen	70
3.2 Der Vergleich zweier Merkmale	77
4 Auf zur Modellierung	79
4.1 Konzepte am Beispiel der Binomialverteilung	80
4.2 Verschiedene diskrete Verteilungen	93
4.3 Stetige Modellwelt	100
5 Casino-Statistik	113
5.1 Würfelfragen	114
5.2 Wahrscheinlichkeit — was ist das?	116
5.3 Rechnen mit Wahrscheinlichkeiten	117
5.4 Axiome der Wahrscheinlichkeitsrechnung	118
5.5 Zusammengesetzte Ereignisse	119
5.6 Kombinatorik für das Gleichmöglichkeitsmodell	122

5.7	Wahrscheinlichkeiten und Bedingungen	123
5.8	Abhängigkeit und Unabhängigkeit	124
5.9	Totale Wahrscheinlichkeit	125
5.10	Lernen aus Zusatzinformationen	126
5.11	Zusammengesetzte Zufallsexperimente	127
6	Parameterschätzungen	131
6.1	Datengrundlage	133
6.2	Zur Identifikation des Modelltyps	134
6.3	Stichproben- und Schätzfunktionen	139
6.4	Zur Konstruktion von Schätzfunktionen	146
6.5	Check des gefundenen Modells	155
7	Konfidenzintervalle	165
7.1	Konfidenzintervall für den Median	166
7.2	Was kostet der Wunsch?	167
7.3	Konstruktionsprinzip für Konfidenzintervalle	168
7.4	Konfidenzintervall für einen Anteil p	171
7.5	Fragen an Konfidenzintervalle	173
7.6	Konfidenzintervalle für die Normalverteilung	177
7.7	Anwendung: Raucherrisiken	178
7.8	Caveat – Mahnung	180
8	Statistik und BAYES	181
8.1	Ein Problem in klassischer Sicht	182
8.2	BAYES und der Euro	186
8.3	Prior – Sample – Posterior	190
8.4	Beta-Verteilung	192
8.5	Es hilft auch im Weltall	194
9	Testen	195
9.1	Kochen und Testen	196
9.2	Der Aufbau eines Tests	203
9.3	Der χ^2 -Test: Ein vielseitiger Geselle	207
9.4	Eine kleine Testgalerie	209
10	Regressionsanalyse	211
10.1	Eine Reise für den Überblick	212
10.2	Das lineare Regressionsmodell	217
10.3	Modell-Schätzung und -Check	218
10.4	Modell-Interpretation	226
10.5	Ausblick	233

11 R-Einführung	243
11.1 Hintergrund, Installation und erste Schritte mit R	244
11.2 Daten einlesen und Statistiken berechnen	246
11.3 Graphiken erstellen	247
11.4 R als Rechenmaschine	249
11.5 Bequemes Arbeiten mit diesem Buch in R	250
11.6 Statistische R-Idioms	252
11.7 Weitere Infos	253
11.8 Codechunks	254
11.9 Funktionen	255
11.10 Daten	256

Abbildungsverzeichnis

1.1	Erhebungsbogen zur Volkszählung	20
2.1	Stabdiagramm zu Merkmal „alter“	23
2.2	Tortendiagramm	24
2.3	Balkendiagramm	25
2.4	Bevölkerungspyramiden von 1910 und 1950	26
2.5	Bevölkerungspyramiden von 2001 und 2050	27
2.6	Größe und Mathenoten	28
2.7	Anzahl Bücher und Anzahl CDs	29
2.8	Alter	30
2.9	Histogramme von gewicht mit verschiedenen Klassendefinitionen	31
2.10	Histogramm: links falsch, rechts richtig	32
2.11	Dot-Plot: Anzahl Bücher	34
2.12	Ausreißerwirkung	35
2.13	Getrimmtes Mittel	36
2.14	Dot-Plot mit Quartilen	37
2.15	Bücheranzahlen	38
2.16	Stichprobe / alle Werte	39
2.17	Boxplots wiederholter Ziehungen	40
2.18	Einkommensgruppen in Deutschland	41
2.19	Range der zentralen $p*100\%$ der Werte gegen p	42
2.20	Variationskoeffizienten gegen Anteil Datenpunkte	43
2.21	Dot-Plot: Rangwertreihe	45
2.22	Empirische Verteilungsfunktion	46
2.23	\hat{F} bei vier verschiedenen Klassenbildungen	47
2.24	Histogramm, Dichtespur	49
2.25	Dichtespur Gewicht	50
2.26	Dichtespuren verschiedener Fensterbreiten	51
2.27	Box-Cox-Plot für Stichprobe und Gesamtdatensatz Bücher	52
2.28	Dichtespuren und Boxplots der transformierten Daten	53
2.29	Verteilung der Möbel-Klicks	55
2.30	Lorenz-Kurve zu Möbelklicks	56
2.31	Drei Lorenzkurven zu unterschiedlichen Situationen	57
2.32	Eine Werbung, deren Herkunft nicht mehr rekonstruierbar war	59

2.33	Häufigkeiten 1 bis 49 – 50 Jahre Lotto	60
2.34	Häufigkeiten der i -ten gezogenen Kugel, $i = 1, \dots, 6$	61
2.35	Häufigkeiten 1 bis 49 – 250 Jahre simuliert	62
2.36	Warten auf die 13 bzw. die 25	63
2.37	Zusätzliche Wartezeiten auf die 13 bzw. 25	64
2.38	Verteilung der Anzahl Treffer bei 2 516 Versuchen	65
2.39	Relative Häufigkeiten	66
2.40	Verteilung der Anzahl Treffer bei 13 000 Versuchen	67
3.1	Kundenverteilung	71
3.2	Zeilenprofile: P_1 , P_2 , P_3 sowie Summenzeile	72
3.3	Profilabweichungen	73
3.4	Größe / Umsatz	74
3.5	Größe / Umsatz mit Schwerpunkt	75
3.6	Einige vergleichende Plots	76
4.1	Simulation	81
4.2	Einfangen von p und Steuerung	82
4.3	Wahrscheinlichkeitsbaum	83
4.4	Pascalsches Dreieck	84
4.5	Wahrscheinlichkeitsfunktionen von Binomialverteilungen	85
4.6	Binomialverteilung	86
4.7	Binomialverteilung	87
4.9	Binomialverteilung mit Simulation	88
4.10	Erwartungswerte zweier 2-Punkte-Verteilungen	89
4.11	$E(X)$ bei <code>binom()</code>	90
4.12	Varianz	91
4.13	Einfangen von p und Steuerung	92
4.14	<code>dhyper()</code>	94
4.15	Drucker und Straftaten	95
4.16	Verteilung der χ^2 -Statistik	96
4.17	Binomial- und hypergeometrische Verteilung	97
4.18	Binomial- und Poisson-Verteilung	98
4.19	Verschiedene Poisson-Verteilungen	99
4.23	Geschätzte Dichte der Gleichverteilung	101
4.24	Summation von Gleichverteilungen	102
4.25	Drei Normalverteilungen	103
4.26	QQ-Plots zum Verteilungsvergleich	104
4.27	$f(x)$ geometrischer Verteilungen	105
4.28	Von der geometrischen zur Exponentialverteilung	106
4.29	Dichten von Exponentialverteilungen	107
4.30	Unfälle und <code>exp()</code>	108

4.31 Dichten verschiedener Gamma-Verteilungen	109
4.32 $F(x), \text{pois}$	110
4.33 $F_W(w)$	111
4.34 QQ-Plot	112
5.1 10 000 Simulationen	115
5.2 Venndiagramme	120
5.3 Mindestens eine Sechs	121
5.4 Drei Würfe	128
5.5 $P(\text{„ein Pasch“})$	129
6.1 Häufigkeiten der Unfälle pro Woche	132
6.2 Unfallzwischenzeiten	135
6.3 pois -Erkennungsplots Datensatz bzw. Simulation	136
6.4 exp -Erkennungsplots Datensatz bzw. Simulation	137
6.5 qqnorm -Plot	138
6.6 $\hat{\theta}_i$ -Realisationen	140
6.7 Schätzfunktionen	141
6.8 Dichten von t - und χ^2 -Verteilungen	142
6.9 \hat{F} und qqnorm -Plot von Mitteln aus $\text{exp}(1)$	143
6.10 \hat{f}_{Median} und $\hat{f}_{\bar{X}}$	144
6.11 Verteilung der Mediane	145
6.12 $\log(\text{Dateigrößen})$	147
6.13 $\log(\text{Dateigrößen})$	148
6.14 Likelihood-Funktion, unterschiedliche Ausschnitte	149
6.15 Entwicklung bis 10 bzw. 20	150
6.16 Entwicklung bis 100 bzw. 1 000	151
6.17 Entwicklung von \bar{X}	152
6.18 \bar{X} aus $\text{pois}(5)$	153
6.19 $\hat{f}_j(x)$	154
6.20 $\log(\text{Dateigrößen})$	156
6.21 Beispiele für an Beta-Verteilungen angepasste Normalverteilungen	157
6.22 Angepasste Dichten	158
6.23 QQ-Plots: Daten aus 2002 und 2004 gegen exp	159
6.24 Nach Reihenfolge	160
6.25 z ab September 2004	161
6.26 $f_{\hat{\lambda}}$ geschätzt	162
6.27 Normalverteilung	163
6.28 \log -Likelihood	164
7.1 α_u und α_o	169
7.2 Schwankungs- und Konfidenzintervall	170

7.3	Konfidenzintervalle	172
7.4	Einfluß von n , $1 - \alpha$	174
7.5	KI: n gegen $1 - \alpha$	175
7.6	KI: IL gegen n	176
7.7	KIs und Sterberate	179
8.1	Verlauf von $P(2 : 1)$	183
8.2	Likelihood	184
8.3	$\max(L(p; y), L(y/n; y))$	185
8.4	Einpunktverteilung	187
8.5	Beta-Verteilungen für kleine Parameterwerte	188
8.6	Beta-Verteilungen für große Parameterwerte	189
8.7	Schnapschuss vom Experiment	191
8.8	Dichten von Beta-Verteilungen	193
9.1	Kartoffeldaten	197
9.2	Einfache Box-and-Whisker bzw. geordnete Studentennittel	198
9.3	Kartoffel-Boxplots	199
9.4	Einfacher Binomialtest	200
9.5	Einfacher Konfidenztest	201
9.6	Empirische Verteilungsfunktionen	202
9.7	Kritische Bereiche des Binomialtests	204
9.8	Wunschverlauf der Güte	205
9.9	Power des Binomialtests	206
9.10	χ^2 -Verteilung für $\nu = 10$	208
10.1	Litern / km	213
10.2	Symbole des Geradenmodells	214
10.3	Verbrauch gegen Kilometer mit fbe-Gerade	215
10.4	Residualplot	216
10.5	KQ-Flächenquadrate	219
10.6	Verbrauchsdaten mit KQ-Gerade und Residualplot	220
10.7	Verbrauchsdaten mit KQ-Gerade und Residualplot	221
10.8	Verschiedene Residualplots	222
10.9	Nichtlinearitäten	223
10.10	Experiment zum Auswirkungsstudium	224
10.11	qqnorm-Plot der geschätzten Störungen	225
10.12	Zähler- und Nenner von R^2 , rechts: Punkte auf Geraden	227
10.13	ein Beispiel für $R^2 \approx 90\%$	228
10.14	R^2 mit einem moderaten bzw. einem schweren Ausreißer	229
10.15	Geraden zu extremen KIs	230
10.16	Prognosebereiche	231
10.17	distance gegen speed	232

10.18Inhaltsstoffe einiger Milchprodukte	234
10.19Residuen des Milchprodukte	235
10.20Anpassung einer Geraden und einer Parabel an cars	236
10.21Gehirn- / Körpergewichte	237
10.22Gewichte logarithmiert	238
10.23Luchsbestand von 1821 bis 1934	239
10.24Tankdaten mit Glättungskurve (links) und mit einem ange- passten Polynom 3. Grades (rechts)	240
10.25Beispiel für unterschiedliche Glättungen	241
10.26Teuerung gegen Anteil	242
11.1 Plot der CO ₂	245
11.2 Verschiedene type-Setzungen	248
11.3 Von open.wnt () erzeugte Oberfläche	251

Vorwort

1 Datenanalyse? Daten? Statistik?

1.1 Was für Daten gibt es?

1.2 Wo kommen Daten her?

VOLKSZÄHLUNG 1987

Personenbogen

Bitte Gemeinde angeben: _____

1 Geburtsangaben

a) Geburtsjahr _____

b) Geburtsmonat _____

2 Geschlecht

männlich _____

weiblich _____

3 Familienstand

ledig _____

verheiratet _____

verwitwet _____

geschieden _____

4 Rechtliche Zugehörigkeit zu einer Religionsgesellschaft

Römisch-katholische Kirche _____

Evangelische Kirche _____

Evangelische Freikirche _____

Jüdische Religionsgesellschaft _____

Islamische Religionsgemeinschaft _____

andere Religionsgesellschaften _____

keiner Religionsgesellschaft rechtlich zugehörig _____

5 Welche Staatsangehörigkeit haben Sie ?

deutsch _____

griechisch _____

italienisch _____

übrige EG-Staaten _____

Jugoslawisch _____

türkisch _____

sonstige/keine _____

6 Wird von Ihnen noch eine weitere Wohnung (Unterkunft/Zimmer) in der Bundesrepublik Deutschland einschließlich Berlin (West) bewohnt ?

nein _____

ja _____

Falls ja:

a) Für Verheiratete, die nicht dauernd getrennt leben: Ist die hiesige Wohnung die vorwiegend benutzte Wohnung der Familie ?

nein _____

ja _____

b) Für alle übrigen Personen: Ist die hiesige Wohnung die vorwiegend benutzte Wohnung ?

nein _____

ja _____

c) Außerdem für Erwerbstätige, Schüler/Studenten: Gehen Sie vorwiegend von der hiesigen Wohnung aus zur Arbeit oder Schule/Hochschule ?

nein _____

ja _____

7 Sind Sie erwerbstätig ?

Vollzeit (über 36 Std. in der Woche) _____

Teilzeit (bis zu 36 Std. in der Woche) _____

arbeitslos, arbeitsuchend _____

nicht erwerbstätig _____

Mehrere Antworten möglich

1) Auch Landwirte, mithelfende Familienangehörige, Auszubildende, Soldaten, Zivildienstleistende

Hausfrau, Hausmann _____

2) Maßgebend ist die normalerweise in der Woche geleistete Arbeitszeit

Schüler(in), Student(in) _____

8 Leben Sie überwiegend von

Erwerbs-, Berufstätigkeit _____

Arbeitslosengeld, -hilfe _____

Rente, Pension _____

eigenem Vermögen, Vermietung, Verpachtung, Allenteil _____

Zuwendungen, Unterhalt durch Eltern, Ehegatten usw. _____

sonstigen Unterstützungen (z. B. Sozialhilfe, BAföG) _____

2

1

Rechtsgrundlage :
Siehe Haushaltsmantelbogen oder Erläuterungsblatt, die Bestandteile der Erhebungsvordrucke sind,
Stichtag: 25. Mai 1987

Falls Sie einen Abschluss an einer allgemeinbildenden bzw. berufsbildenden Schule/Hochschule haben

9 Welchen höchsten allgemeinen Schulabschluss haben Sie ?

Volksschule, Hauptschule _____

Realschule/gleichwertiger Abschluss (z. B. Mittlere Reife) _____

Hochschulreife (Abitur), Fachhochschulreife _____

10 a) Welchen höchsten Abschluss an einer berufsbildenden Schule oder Hochschule haben Sie ?

Berufsfachschule (ohne Berufsschule) _____

Fachschule _____

Fachhochschule (Ing.-Schule, höhere Fachschule) _____

Hochschule (einschließlich Lehrerausbildung) _____

b) Welche Hauptfachrichtung hat dieser Abschluss ? _____

11 Falls Sie eine praktische Berufsausbildung (z. B. Lehre) abgeschlossen haben:

a) Auf welchen Lehrberuf bezog sich diese Ausbildung ? _____

b) Wie lange dauerte diese Ausbildung ? (Jahre): _____

12 Bitte Name und Anschrift Ihrer Arbeitsstätte oder Schule/Hochschule angeben.

Name _____

Strasse/Kassennummer: _____

PLZ _____

Gemeinde: _____

13 Welches Verkehrsmittel benutzen Sie hauptsächlich (längste Strecke) auf dem Hinweg zur Arbeit oder Schule/Hochschule ?

kein Verkehrsmittel (zu Fuß) _____

Fahrrad _____

Pkw _____

U-Bahn, S-Bahn, Straßenbahn _____

Eisenbahn _____

Bus, sonst. öffentl. Verkehrsmittel _____

sonstiges (Motorrad, Moped, Mofa) _____

14 Wieviel Zeit benötigen Sie normalerweise für den Hinweg zur Arbeit oder Schule/Hochschule ?

entfällt, da auf gleichem Grundstück _____

unter 15 Minuten _____

15 bis unter 30 Minuten _____

30 bis unter 45 Minuten _____

45 bis unter 60 Minuten _____

60 Minuten und mehr _____

15 Sind Sie zur Zeit tätig als

Facharbeiter(in) _____

sonstige(r) Arbeiter(in) _____

Angestellter(r) _____

Auszubildende(r) _____

Beamter/Beamtin, Richter(in), Soldat, Zivildienst. _____

mit bezahlten Beschäftigten _____

Selbständige(r) ohne bezahlte Beschäftigte _____

mithelfende(r) Familienangehörige(r) _____

16 Zu welchem Wirtschaftszweig (Branche, Behörde) gehört der Betrieb (Firma, Dienststelle), in dem Sie tätig sind ? _____

17 Welche Tätigkeit, welchen Beruf üben Sie aus ? _____

18 Falls Sie eine Nebenberufstätigkeit ausüben, handelt es sich um eine

landwirtschaftliche _____

nichtlandwirtschaftliche _____

NUR VOM STATISTISCHEN LANDESAMT AUSZUFÜLLEN

Geburtsjahr (1 a) 2, 3 u. 4 Stelle	Hauptfachrichtung (10 b)	Erlernter Beruf (11 a)	Dauer (11 b)
Arbeitsstätte, Schule/Hochschule (Pendler) Land (12), Gemeinde (12)	Strasse (12)	Hausnummer (12)	
Wirtschaftszweig (16)	Ausgewählte Tätigkeit (17)		

Abb. 1.1: Erhebungsbogen zur Volkszählung

2 Univariate, exploratorische Analyse

2.1 Häufigkeitstabellen und deren Darstellung

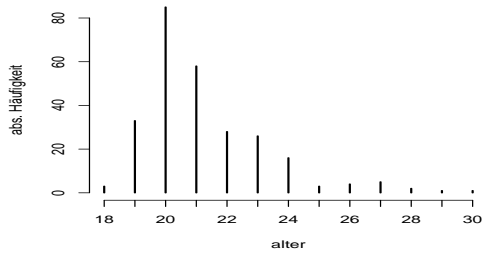


Abb. 2.1: Stabdiagramm zu Merkmal „alter“

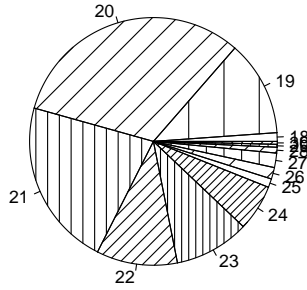


Abb. 2.2: Tortendiagramm

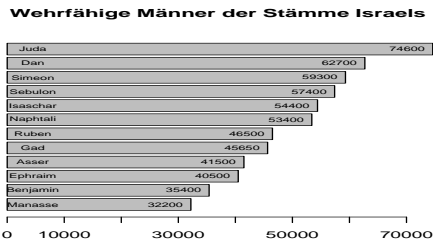


Abb. 2.3: Balkendiagramm

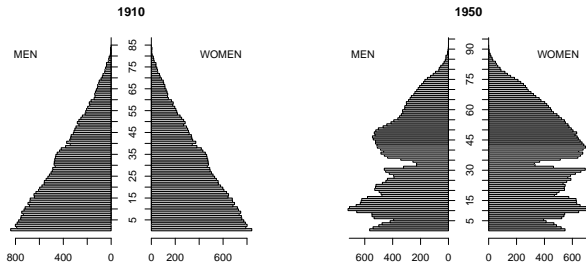


Abb. 2.4: Bevölkerungspyramiden von 1910 und 1950

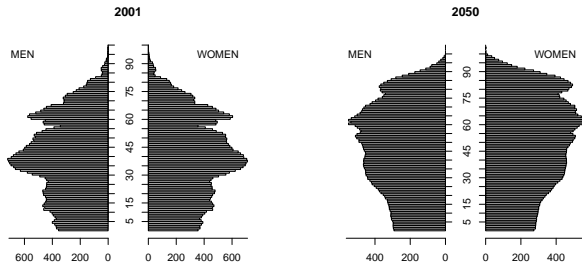


Abb. 2.5: Bevölkerungspyramiden von 2001 und 2050

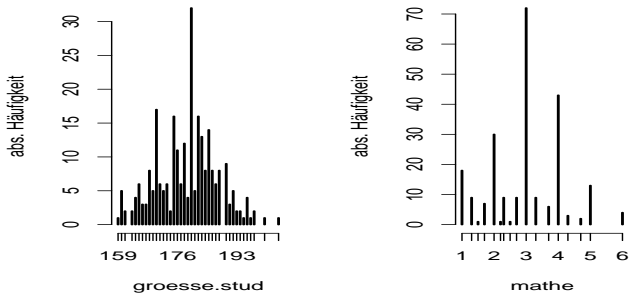


Abb. 2.6: Größe und Mathenoten

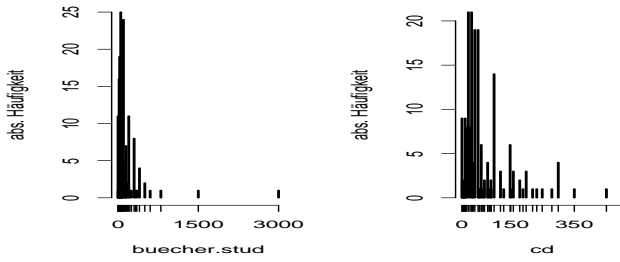


Abb. 2.7: Anzahl Bücher und Anzahl CDs

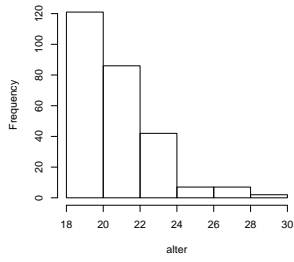


Abb. 2.8: Alter

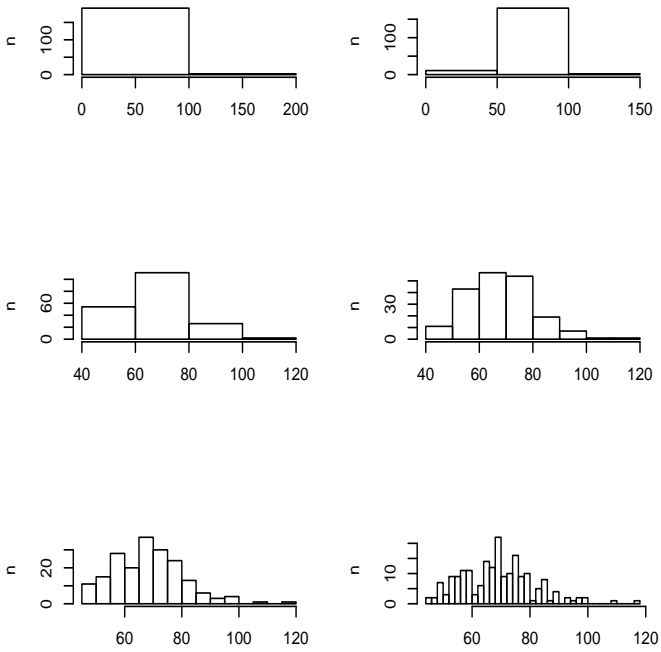


Abb. 2.9: Histogramme von `gewicht` mit verschiedenen Klassendefinitionen

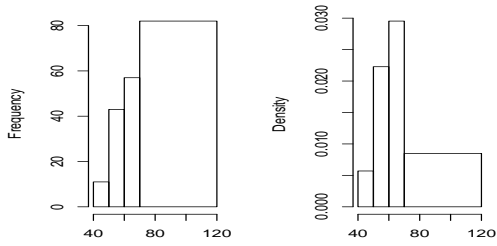


Abb. 2.10: Histogramm: links falsch, rechts richtig

2.2 Auswertung der Urliste: Lage und Variabilität

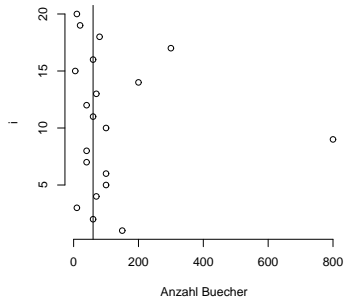


Abb. 2.11: Dot-Plot: Anzahl Bücher

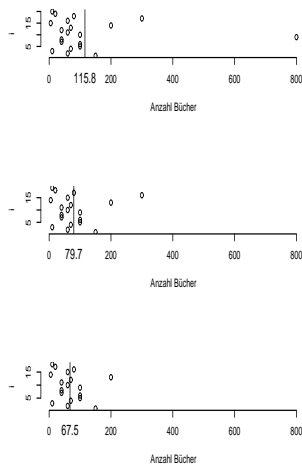


Abb. 2.12: Ausreißerwirkung

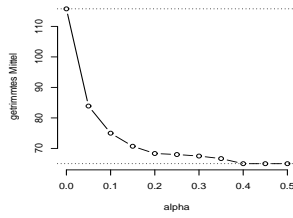


Abb. 2.13: Getrimmtes Mittel

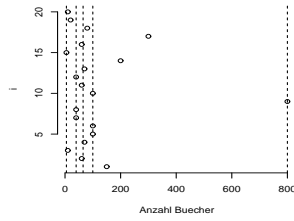


Abb. 2.14: Dot-Plot mit Quartilen

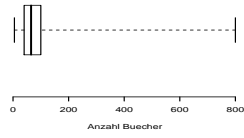


Abb. 2.15: Bücheranzahlen

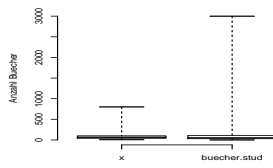


Abb. 2.16: Stichprobe / alle Werte

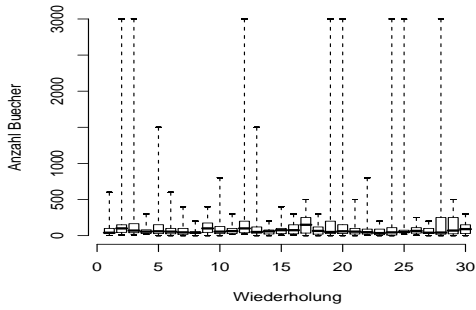


Abb. 2.17: Boxplots wiederholter Ziehungen

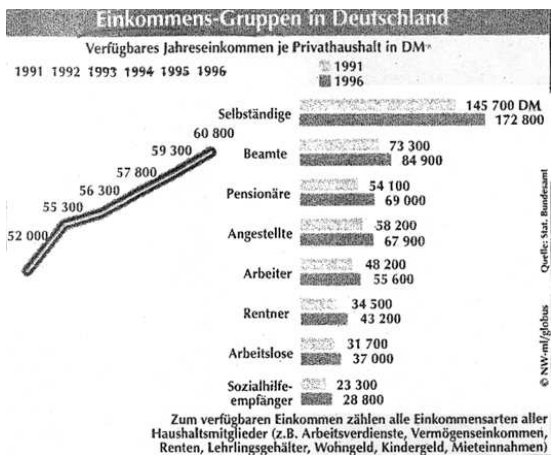


Abb. 2.18: Einkommensgruppen in Deutschland

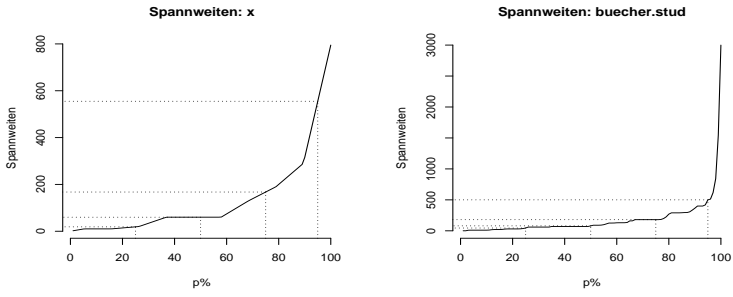


Abb. 2.19: Range der zentralen $p \cdot 100\%$ der Werte gegen p

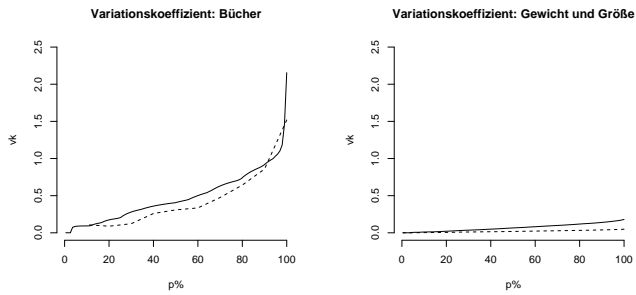


Abb. 2.20: Variationskoeffizienten gegen Anteil Datenpunkte

2.3 Die empirische Verteilungsfunktion

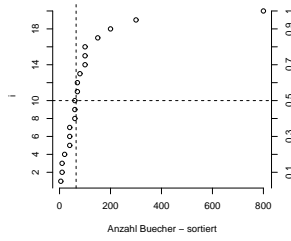


Abb. 2.21: Dot-Plot: Rangwertreihe

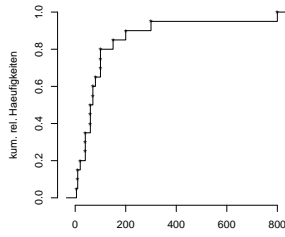


Abb. 2.22: Empirische Verteilungsfunktion

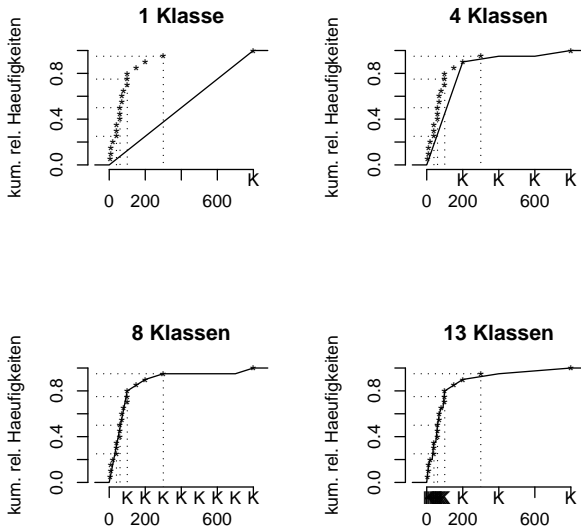


Abb. 2.23: \hat{F} bei vier verschiedenen Klassenbildungen

2.4 Besondere Strukturen einer Verteilung

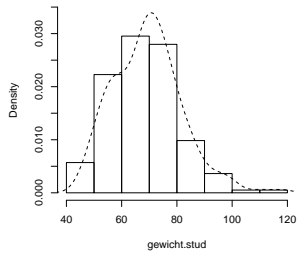


Abb. 2.24: Histogramm, Dichtespur

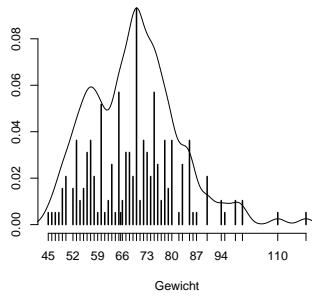


Abb. 2.25: Dichtespur Gewicht

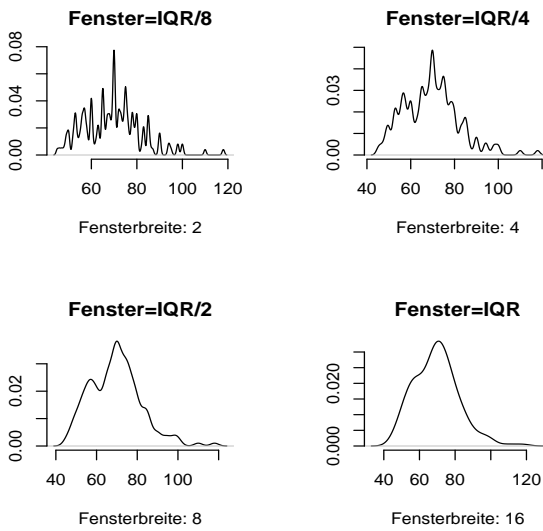


Abb. 2.26: Dichtespuren verschiedener Fensterbreiten

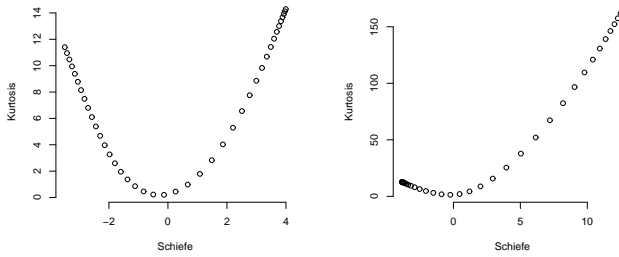


Abb. 2.27: Box-Cox-Plot für Stichprobe und Gesamtdatensatz Bücher

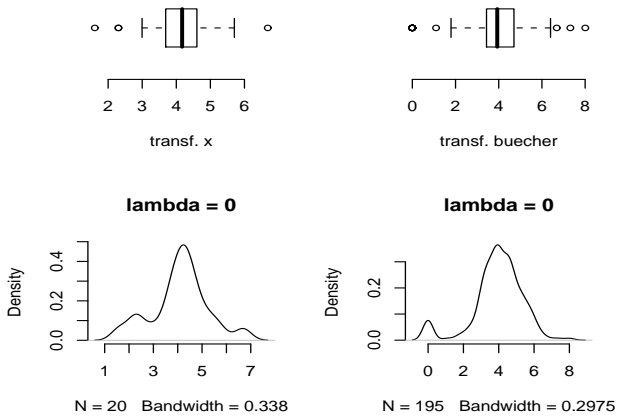


Abb. 2.28: Dichtespuren und Boxplots der transformierten Daten

2.5 Konzentrationsmessung – LORENZ und GINI

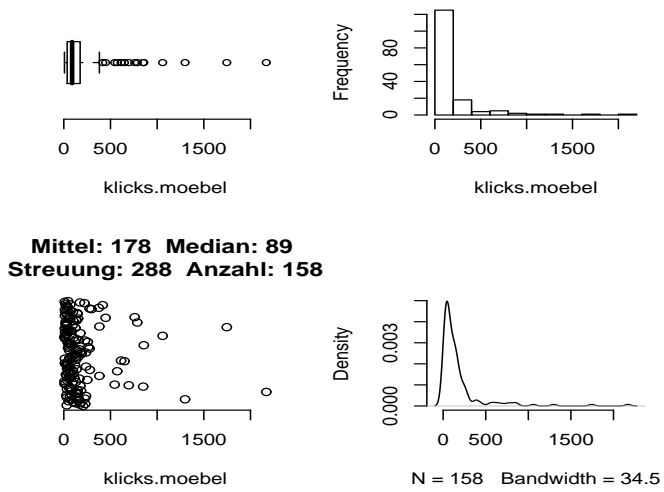


Abb. 2.29: Verteilung der Möbel-Klicks

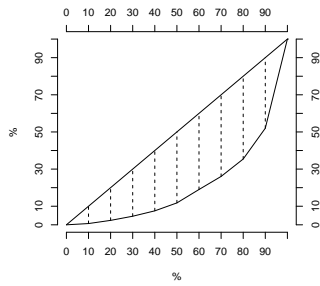


Abb. 2.30: Lorenz-Kurve zu Möbelclicks

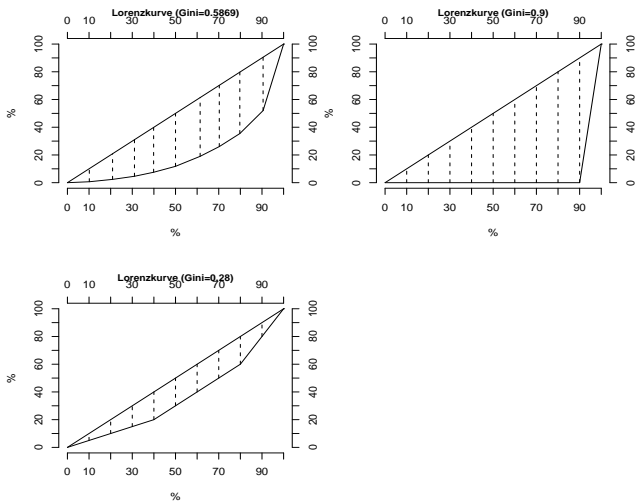


Abb. 2.31: Drei Lorenzkurven zu unterschiedlichen Situationen

2.6 Fallstudie – das 6 aus 49 Lotto

„Und nun die Lotto-Prognose für Samstag, den 31. Oktober 1998.“

Unser Angebot: Wir erhöhen die Gewinnchancen Ihrer Leser!

Zugegeben, die Vorstellung, daß wir Lottozahlen mit annähernder Genauigkeit prognostizieren können, ist ungewöhnlich aber wahr. Überzeugen Sie sich selbst und testen Sie unser Angebot Sie werden begeistert sein!



Wie genau ist unsere Lotto-Prognose?

Die Gewinnzahlen vom 31.10.98: 2 5 6 24 26 44

Wir können keinen Sechser vorausberechnen, auch keinen Fünfer. Allerdings ist unsere Prognose immer ziemlich nah dran. Und das ist ausschlaggebend. So spielen Ihre Leser nicht 6 aus 49, sondern etwa 6 aus 24, wenn sie einige Varianten in der Nähe unserer Kombination tippen. Deshalb erhöhen sich die Gewinnchancen für Ihre Leser wesentlich.

Wie funktioniert unsere Lotto-Prognose?

Unsere Prognose basiert nicht auf der Wahrscheinlichkeitstheorie, denn unser Expertenteam hat nachgewiesen, daß solche Prognosen nicht funktionieren können, weil die Lototrommel a) kein idealer Zufallsgenerator ist und b) 49 Kugeln wenig mit den Gesetzen der großen Zahlen zutun haben. Unsere Prognose basiert auf der Annahme, daß diese Trommel mit 49 Kugeln ein kompliziertes physikalisches System ist. Unsere Software simuliert dieses System, lernt aus jeder Ziehung, paßt sich an und berechnet Zahlenfelder, die bevorzugt für die nächste Ziehung in Betracht kommen. In dieser Software stecken zwei Promotionen, eine Habilitation und 11 Jahre wissenschaftliche Kleinarbeit.

Wir bieten Ihnen exklusiv für Ihr Medium in Ihrem Verbreitungsgebiet die Prognose an. Natürlich auch mit einer Lieferung gratis für die Ziehung in der kommenden Woche, damit Sie Ihren eigenen Test machen können. *

Abb. 2.32: Eine Werbung, deren Herkunft nicht mehr rekonstruierbar war

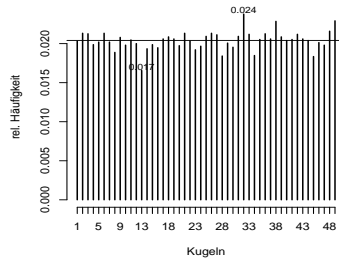


Abb. 2.33: Häufigkeiten 1 bis 49
– 50 Jahre Lotto

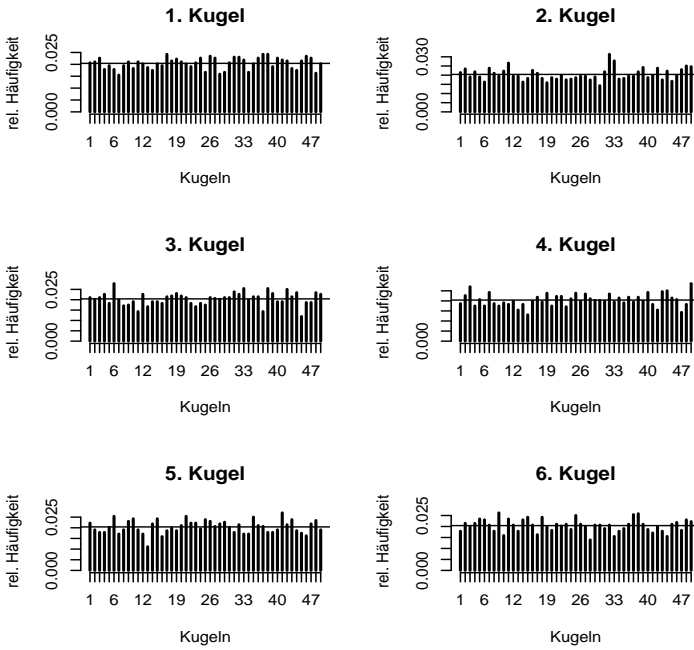


Abb. 2.34: Häufigkeiten der i -ten gezogenen Kugel, $i = 1, \dots, 6$

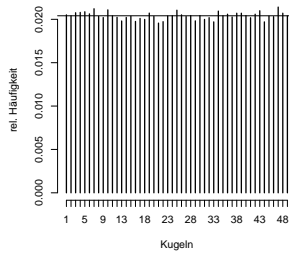


Abb. 2.35: Häufigkeiten 1 bis 49
– 250 Jahre simuliert

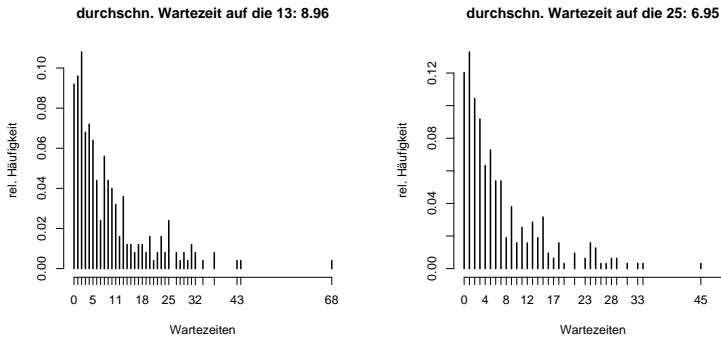


Abb. 2.36: Warten auf die 13 bzw. die 25

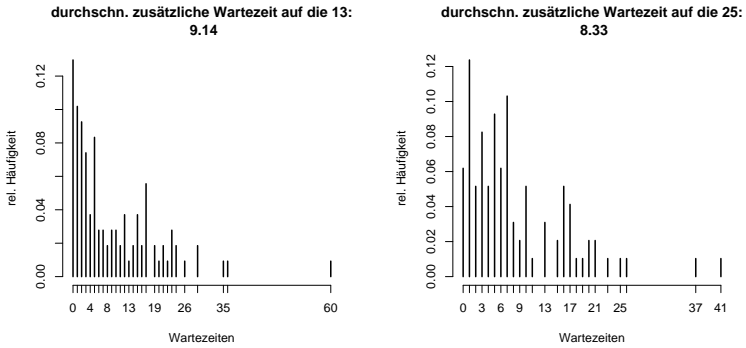


Abb. 2.37: Zusätzliche Wartezeiten auf die 13 bzw. 25

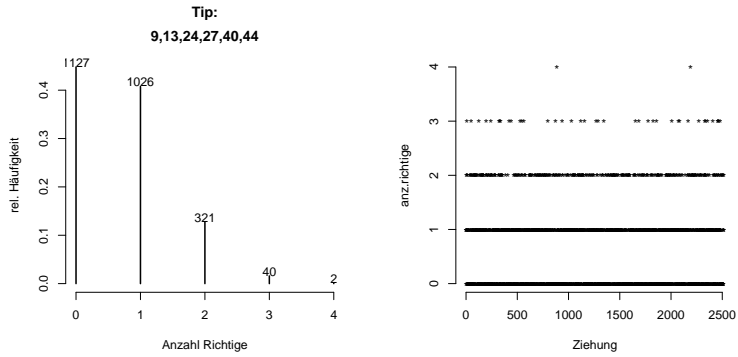


Abb. 2.38: Verteilung der Anzahl Treffer bei 2516 Versuchen

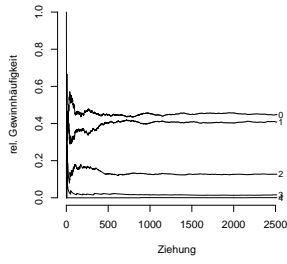


Abb. 2.39: Relative Häufigkeiten

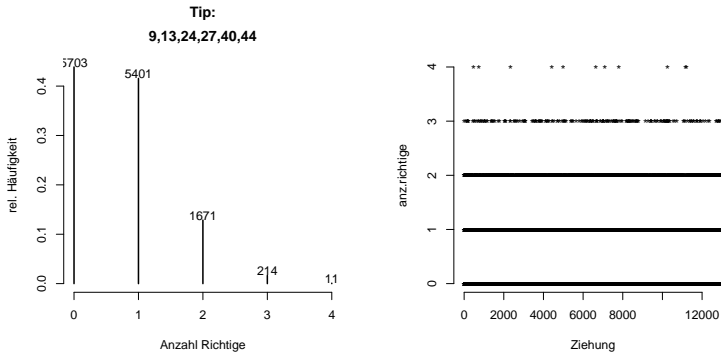


Abb. 2.40: Verteilung der Anzahl Treffer bei 13 000 Versuchen

3 Bivariate, exploratorische Analyse

3.1 Korrelation von Merkmalen

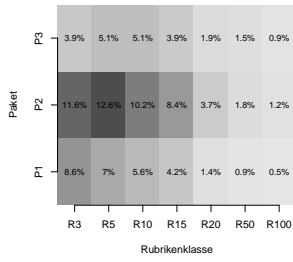


Abb. 3.1: Kundenverteilung

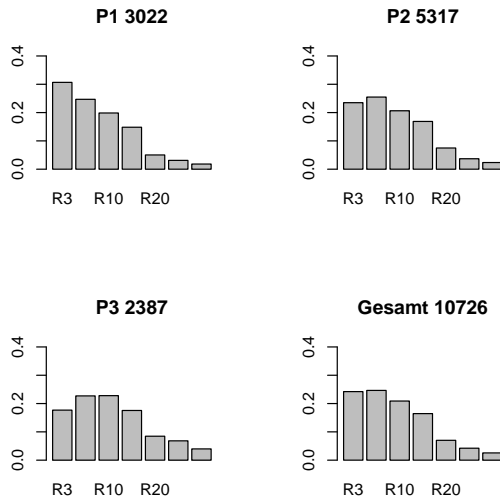


Abb. 3.2: Zeilenprofile: P1 , P2 , P3 sowie Summenzeile

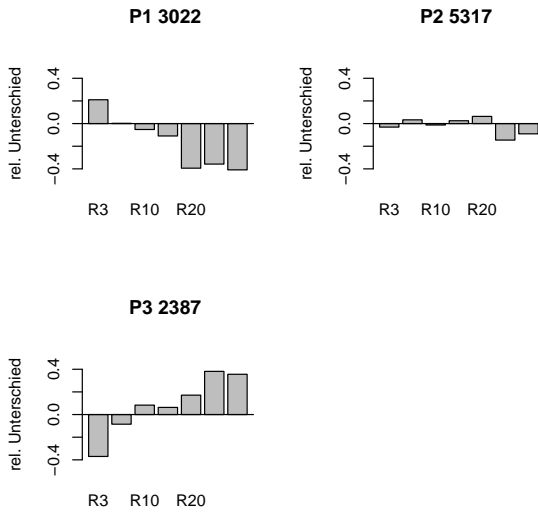


Abb. 3.3: Profilabweichungen

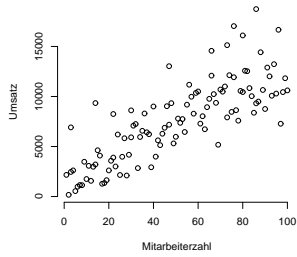


Abb. 3.4: Größe / Umsatz

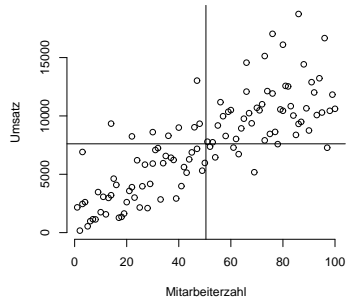


Abb. 3.5: Größe / Umsatz mit Schwerpunkt

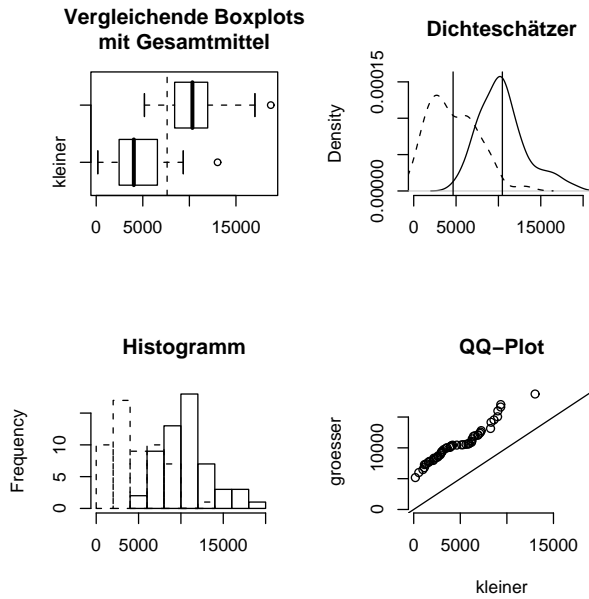


Abb. 3.6: Einige vergleichende Plots

3.2 Der Vergleich zweier Merkmale

4 Auf zur Modellierung

4.1 Konzepte am Beispiel der Binomialverteilung

Erfolgsprozentsatz in GG=50,
n.stpr=50, WD=1000, Zufall=13

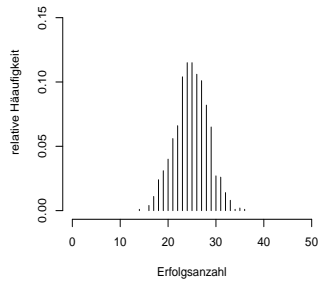


Abb. 4.1: Simulation

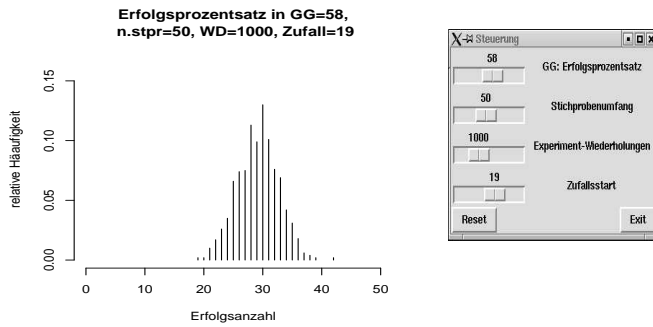


Abb. 4.2: Einfangen von p und Steuerung

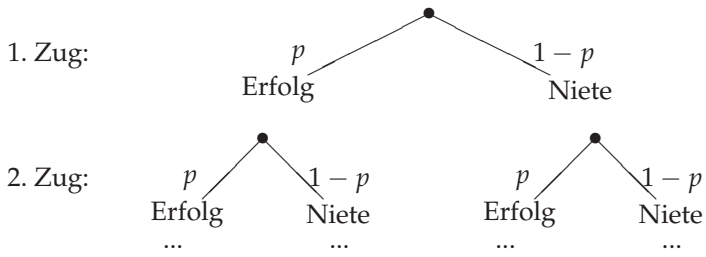


Abb. 4.3: Wahrscheinlichkeitsbaum

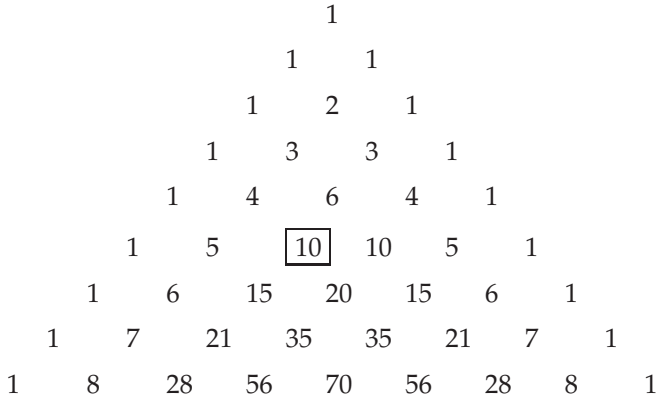


Abb. 4.4: Pascalsches Dreieck

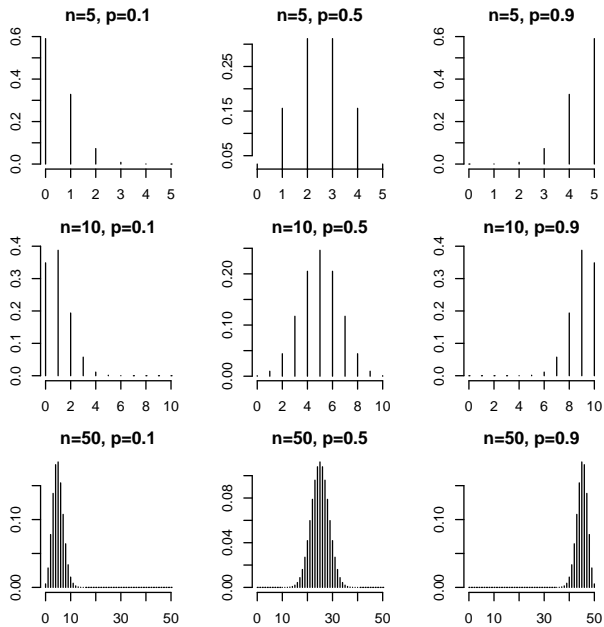


Abb. 4.5: Wahrscheinlichkeitsfunktionen von Binomialverteilungen

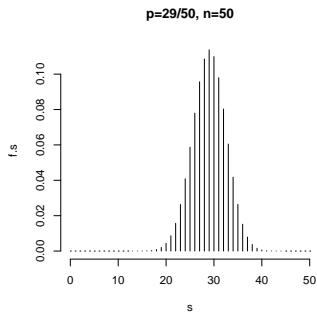


Abb. 4.6: Binomialverteilung

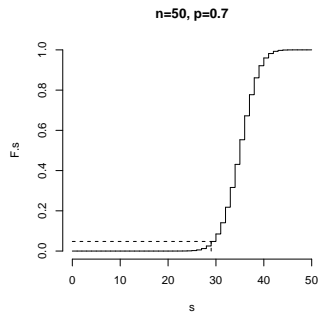


Abb. 4.7: Binomialverteilung

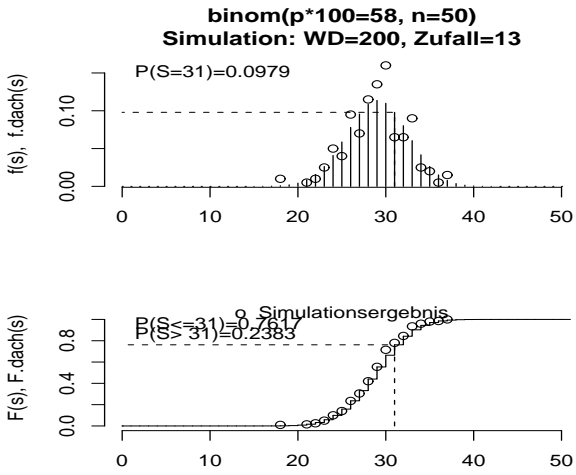


Abb. 4.9: Binomialverteilung mit Simulation

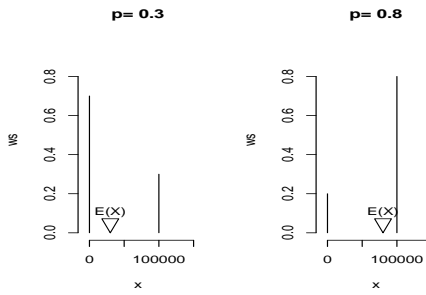


Abb. 4.10: Erwartungswerte zweier 2-Punkte-Verteilungen

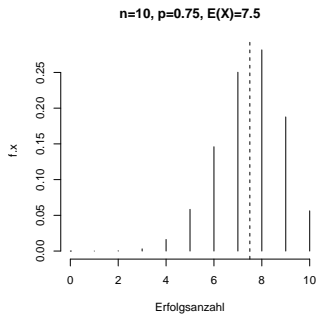


Abb. 4.11: $E(X)$ bei `binom()`

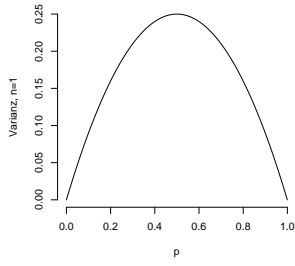


Abb. 4.12: Varianz

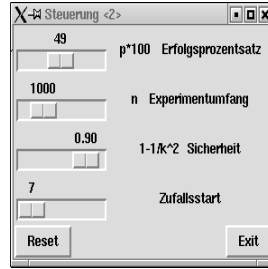
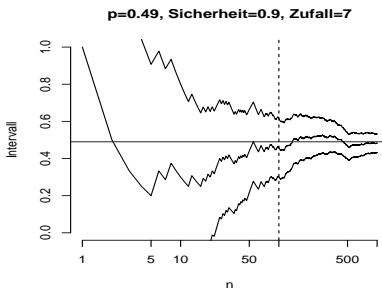


Abb. 4.13: Einfangen von p und Steuerung

4.2 Verschiedene diskrete Verteilungen

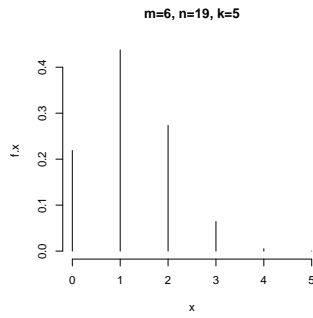


Abb. 4.14: `dhyper()`

Beziehung	Raub	Mord	Σ	Preis		Note	
				pro Seite	GUT	SCHLECHTER	Σ
verwandt	627	254	881	< 3 Cent	4	5	9
bekannt	5010	99	5109	> 3 Cent	3	7	10
Σ	5637	353	5990	Σ	7	12	19

Abb. 4.15: Drucker und Straftaten

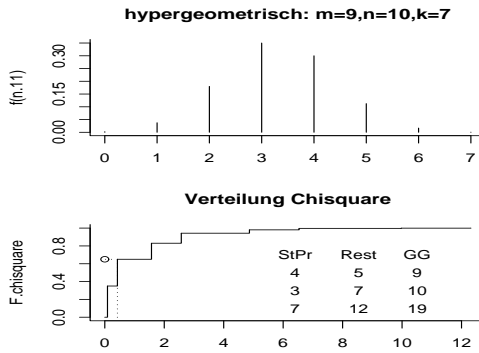


Abb. 4.16: Verteilung der χ^2 -Statistik

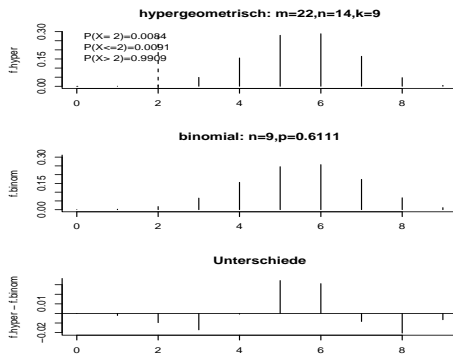


Abb. 4.17: Binomial- und hypergeometrische Verteilung

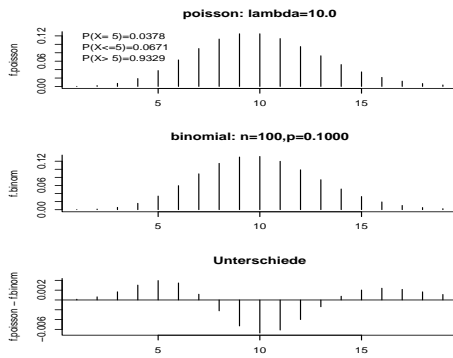


Abb. 4.18: Binomial- und Poisson-Verteilung

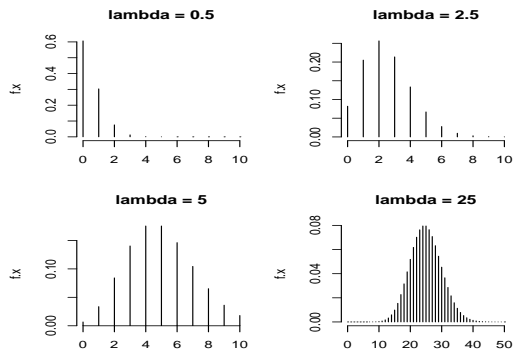


Abb. 4.19: Verschiedene Poisson-Verteilungen



4.3 Stetige Modellwelt

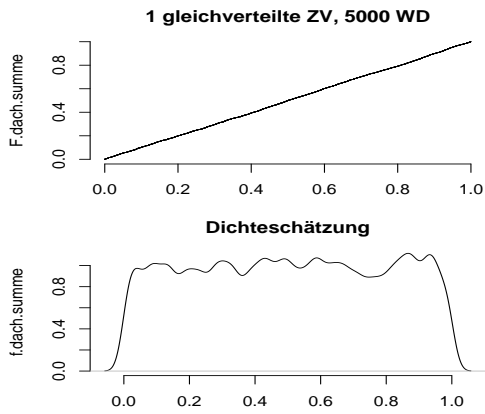


Abb. 4.23: Geschätzte Dichte der Gleichverteilung

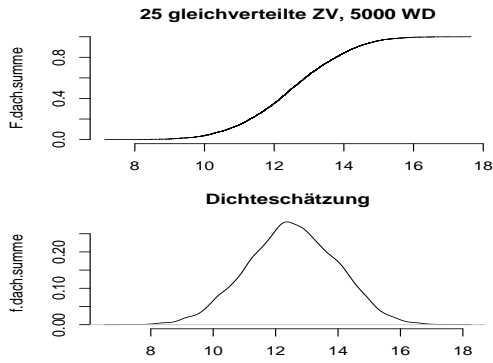


Abb. 4.24: Summation von Gleichverteilungen

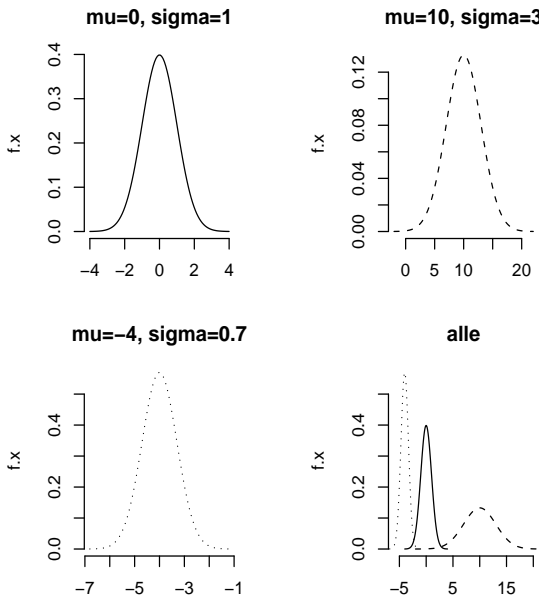


Abb. 4.25: Drei Normalverteilungen

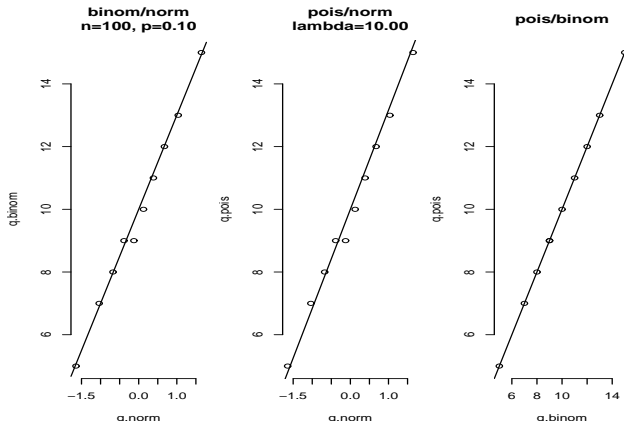


Abb. 4.26: QQ-Plots zum Verteilungsvergleich

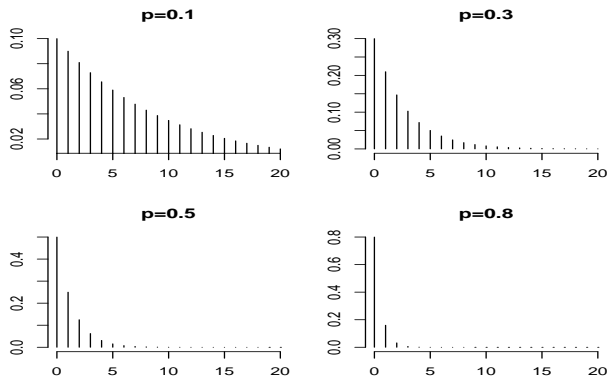


Abb. 4.27: $f(x)$ geometrischer Verteilungen

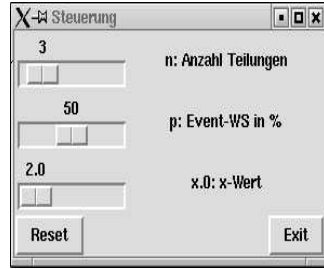
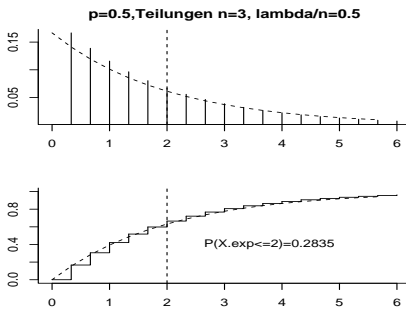


Abb. 4.28: Von der geometrischen zur Exponentialverteilung

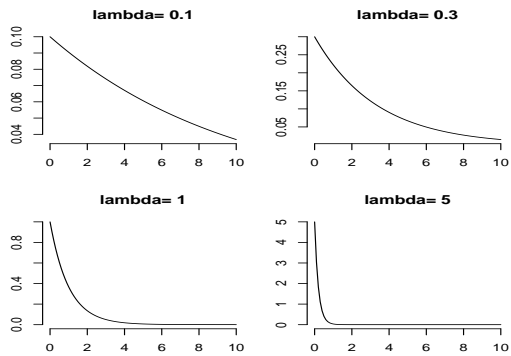


Abb. 4.29: Dichten von Exponentialverteilungen

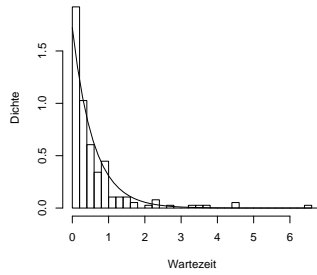


Abb. 4.30: Unfälle und $\exp()$

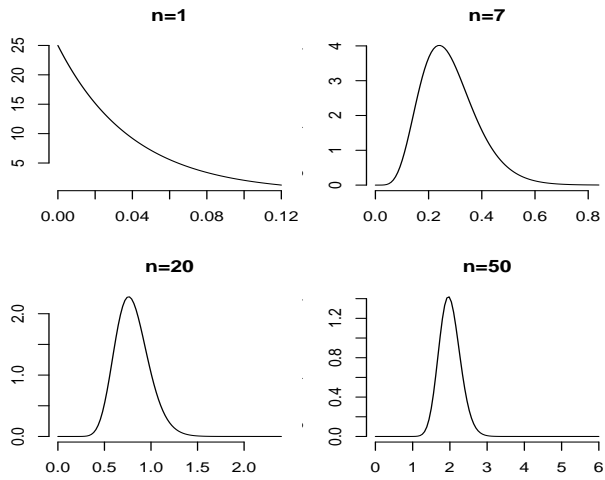


Abb. 4.31: Dichten verschiedener Gamma-Verteilungen

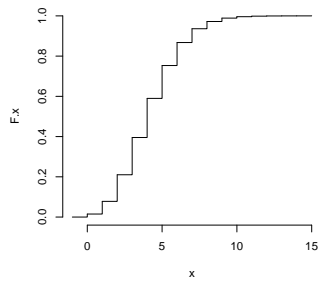


Abb. 4.32: $F(x)$, pois

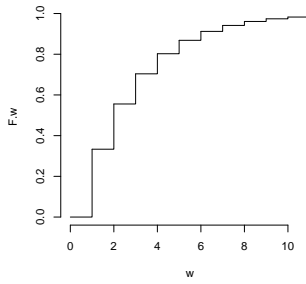


Abb. 4.33: $F_W(w)$

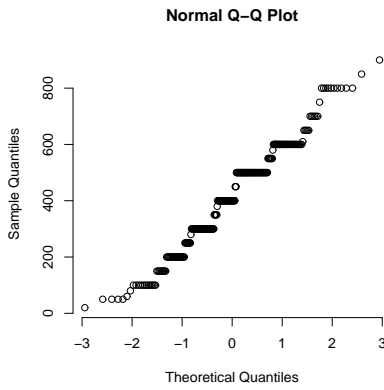


Abb. 4.34: QQ-Plot

5 Casino-Statistik

5.1 Würfelfragen

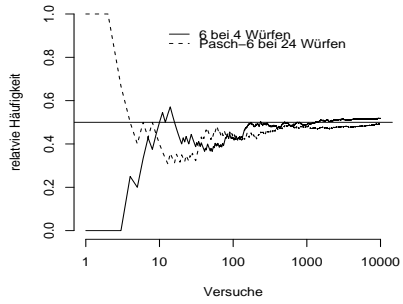


Abb. 5.1: 10 000 Simulationen



5.2 Wahrscheinlichkeit — was ist das?

5.3 Rechnen mit Wahrscheinlichkeiten



5.4 Axiome der Wahrscheinlichkeitsrechnung

5.5 Zusammengesetzte Ereignisse

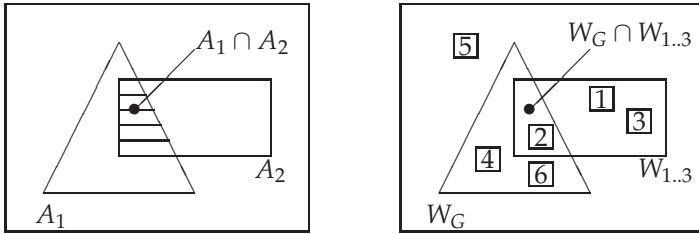


Abb. 5.2: Venndiagramme

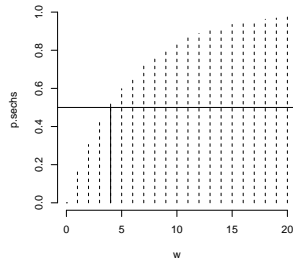


Abb. 5.3: Mindestens eine Sechs



5.6 Kombinatorik für das Gleichmöglichkeitsmodell

5.7 Wahrscheinlichkeiten und Bedingungen



5.8 Abhängigkeit und Unabhängigkeit

5.9 Totale Wahrscheinlichkeit



5.10 Lernen aus Zusatzinformationen



5.11 Zusammengesetzte Zufallsexperimente

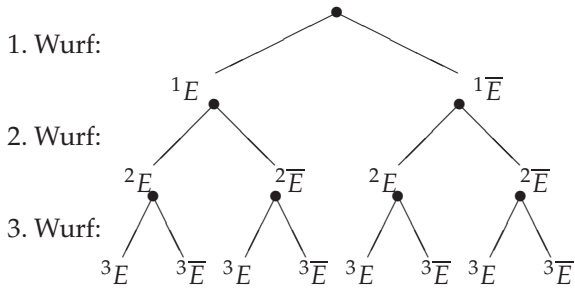


Abb. 5.4: Drei Würfe

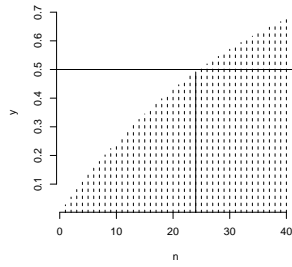


Abb. 5.5: $P(\text{„ein Pasch“})$

6 Parameterschätzungen

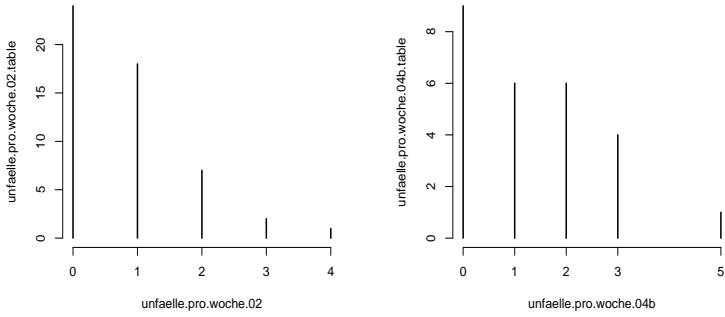


Abb. 6.1: Häufigkeiten der Unfälle pro Woche

6.1 Datengrundlage

6.2 Zur Identifikation des Modelltyps

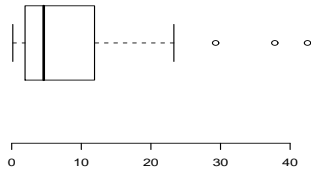


Abb. 6.2: Unfallzwischenzeiten

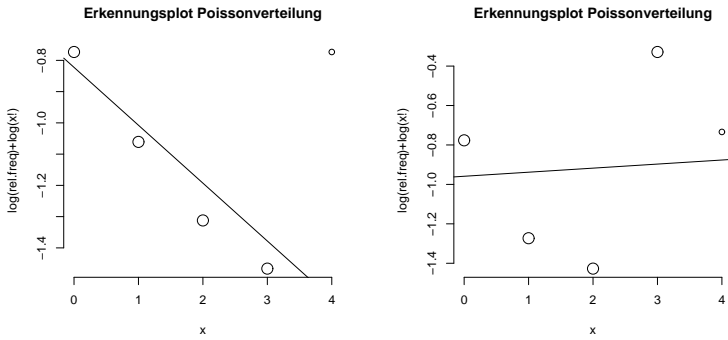


Abb. 6.3: pois-Erkennungsplots Datensatz bzw. Simulation

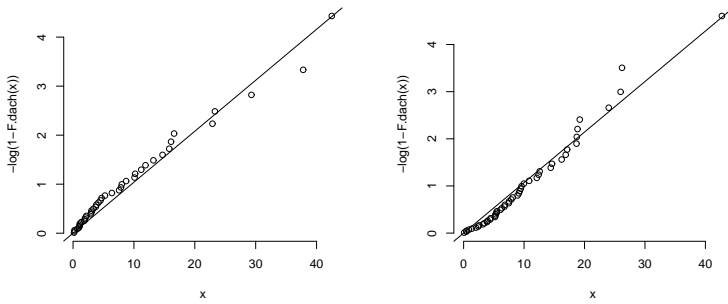


Abb. 6.4: exp-Erkennungsplots Datensatz bzw. Simulation

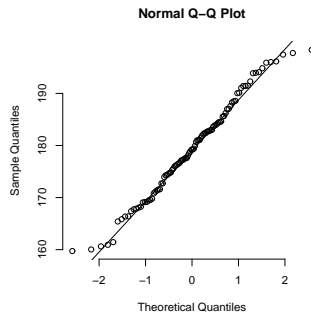


Abb. 6.5: qqnorm-Plot

6.3 Stichproben- und Schätzfunktionen

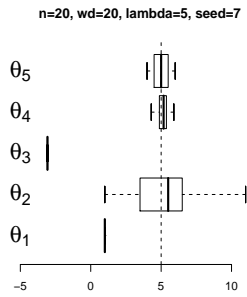


Abb. 6.6: $\hat{\theta}_i$ -Realisationen

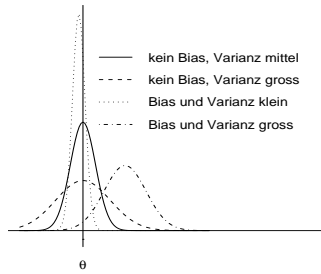


Abb. 6.7: Schätzfunktionen

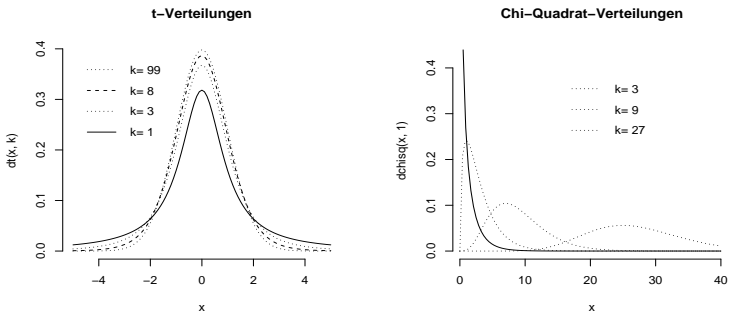


Abb. 6.8: Dichten von t- und χ^2 -Verteilungen

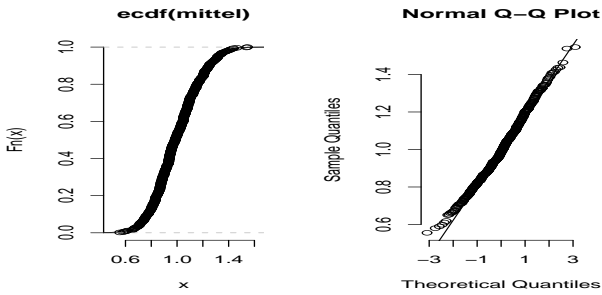


Abb. 6.9: \hat{F} und qqnorm-Plot von Mitteln aus $\exp(1)$

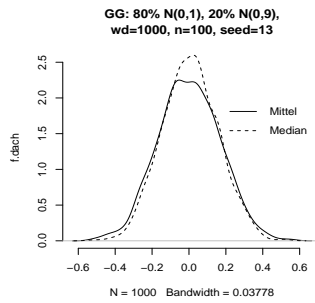


Abb. 6.10: \hat{f}_{Median} und $\hat{f}_{\bar{X}}$

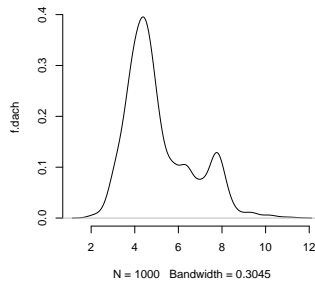


Abb. 6.11: Verteilung der Mediane



6.4 Zur Konstruktion von Schätzfunktionen

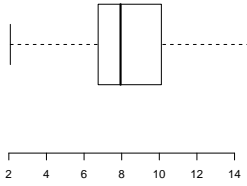


Abb. 6.12: $\log(\text{Dateigrößen})$

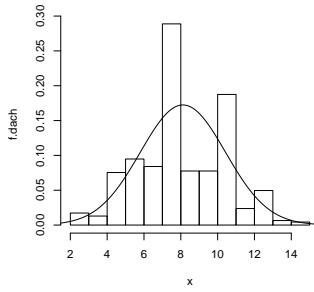


Abb. 6.13: $\log(\text{Dateigrößen})$

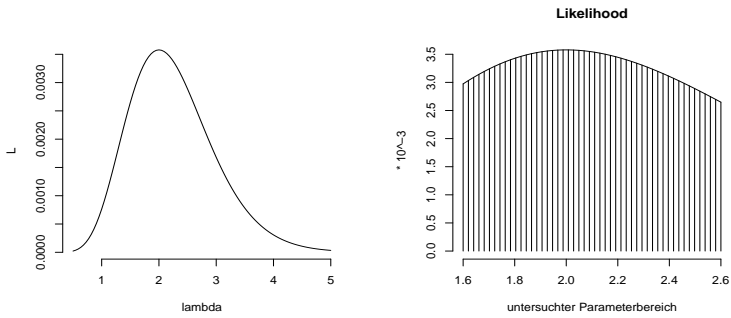


Abb. 6.14: Likelihood-Funktion, unterschiedliche Ausschnitte

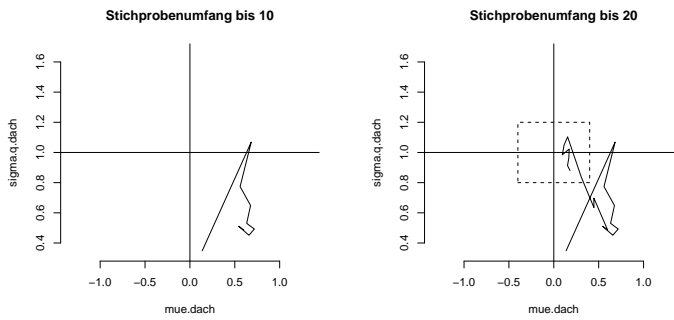


Abb. 6.15: Entwicklung bis 10 bzw. 20

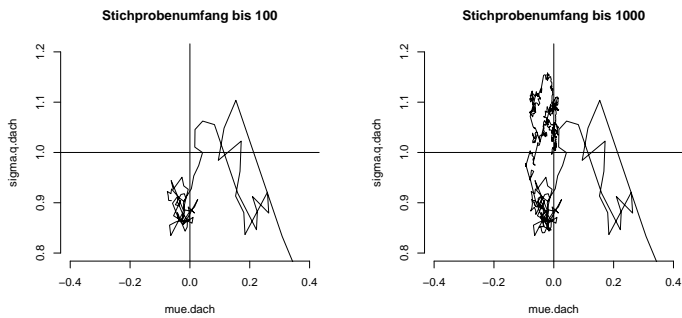


Abb. 6.16: Entwicklung bis 100 bzw. 1 000

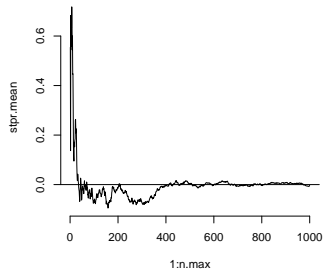


Abb. 6.17: Entwicklung von \bar{X}

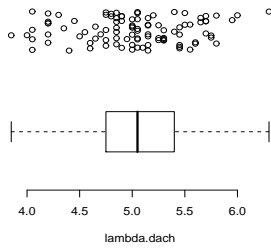


Abb. 6.18: \bar{X} aus `pois(5)`

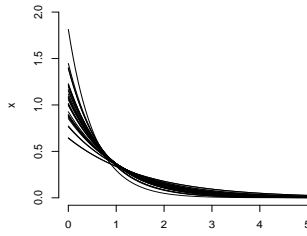


Abb. 6.19: $\hat{f}_j(x)$

6.5 Check des gefundenen Modells

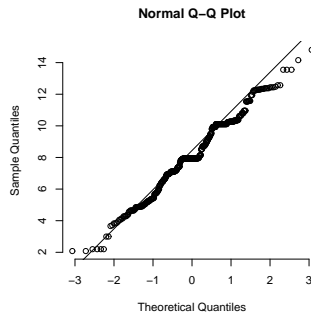


Abb. 6.20: $\log(\text{Dateigrößen})$

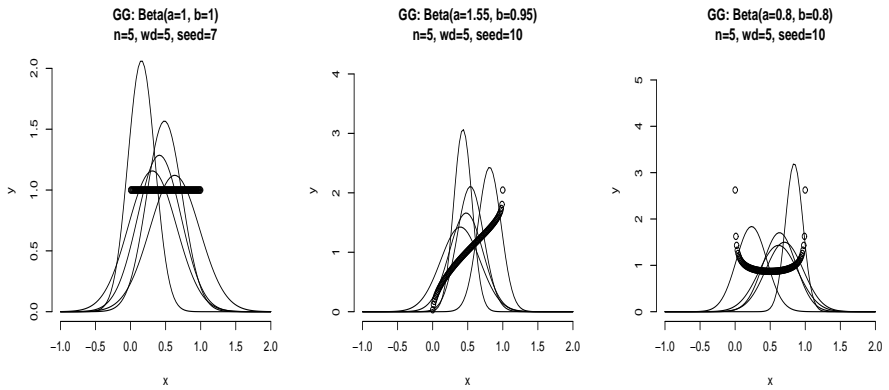


Abb. 6.21: Beispiele für an Beta-Verteilungen angepasste Normalverteilungen

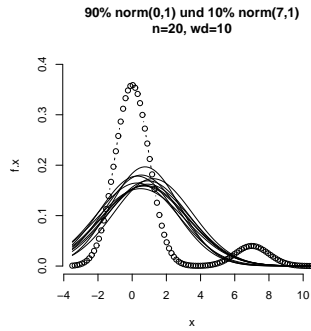


Abb. 6.22: Angepasste Dichten

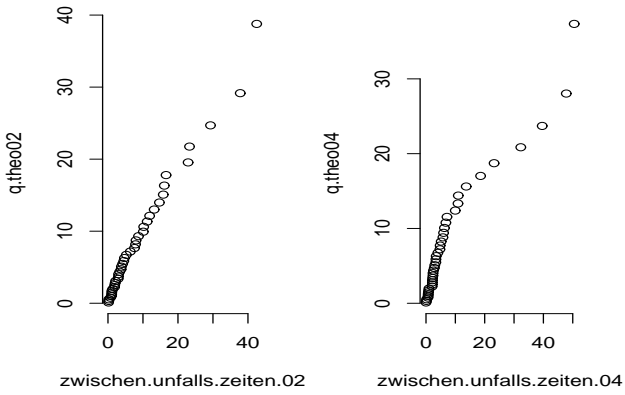


Abb. 6.23: QQ-Plots: Daten aus 2002 und 2004 gegen exp

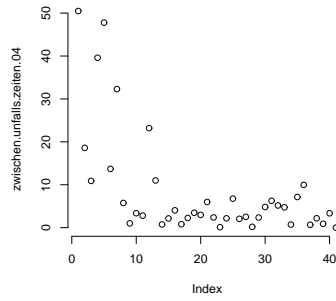


Abb. 6.24: Nach Reihenfolge

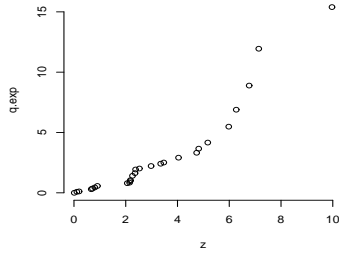


Abb. 6.25: z ab September 2004

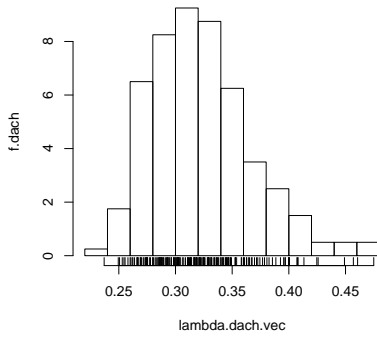


Abb. 6.26: $f_{\hat{\lambda}}$ geschätzt

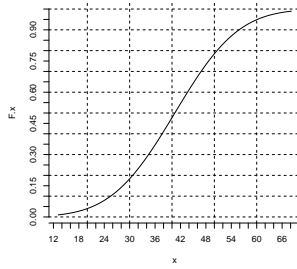


Abb. 6.27: Normalverteilung

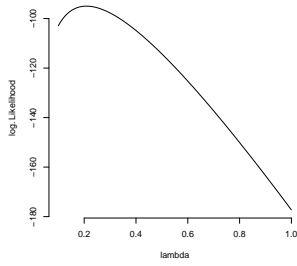


Abb. 6.28: log-Likelihood

7 Konfidenzintervalle



7.1 Konfidenzintervall für den Median

7.2 Was kostet der Wunsch?



7.3 Konstruktionsprinzip für Konfidenzintervalle

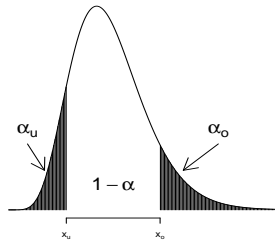


Abb. 7.1: α_u und α_o

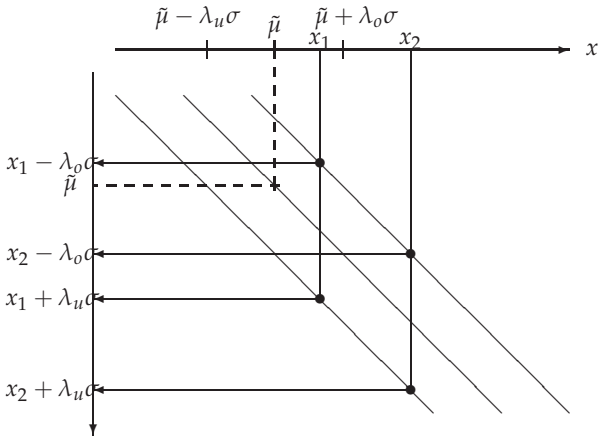


Abb. 7.2: Schwankungs- und Konfidenzintervall

7.4 Konfidenzintervall für einen Anteil p

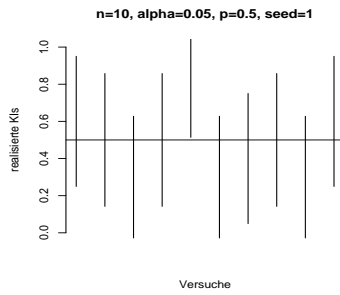


Abb. 7.3: Konfidenzintervalle



7.5 Fragen an Konfidenzintervalle

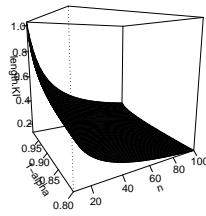


Abb. 7.4: Einfluß von $n, 1 - \alpha$

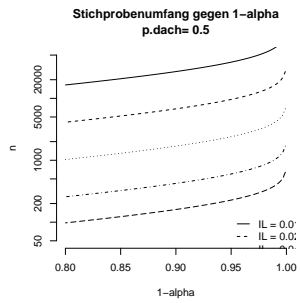


Abb. 7.5: KI: n gegen $1 - \alpha$

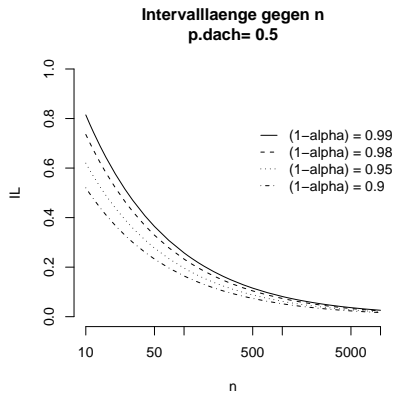


Abb. 7.6: KI: IL gegen n

7.6 Konfidenzintervalle für die Normalverteilung



7.7 Anwendung: Raucherrisiken

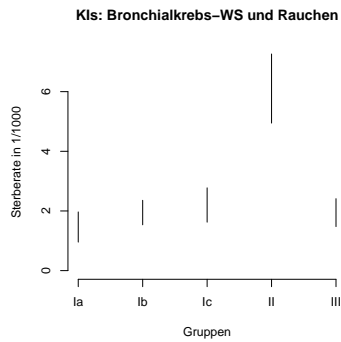


Abb. 7.7: KIs und Sterberate



7.8 Caveat – Mahnung

8 Statistik und BAYES



8.1 Ein Problem in klassischer Sicht

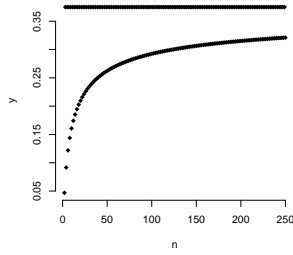


Abb. 8.1: Verlauf von $P(2 : 1)$

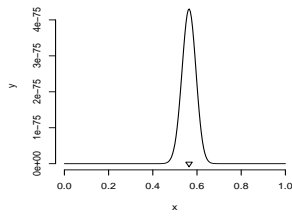


Abb. 8.2: Likelihood

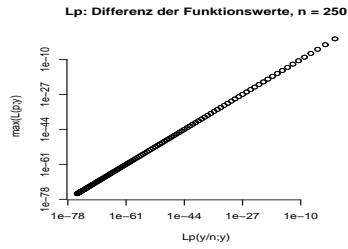


Abb. 8.3: $\max(L(p; y), L(y/n; y))$



8.2 BAYES und der Euro

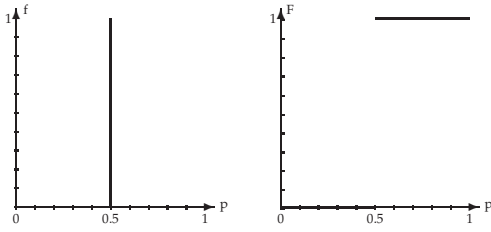


Abb. 8.4: Einpunktverteilung

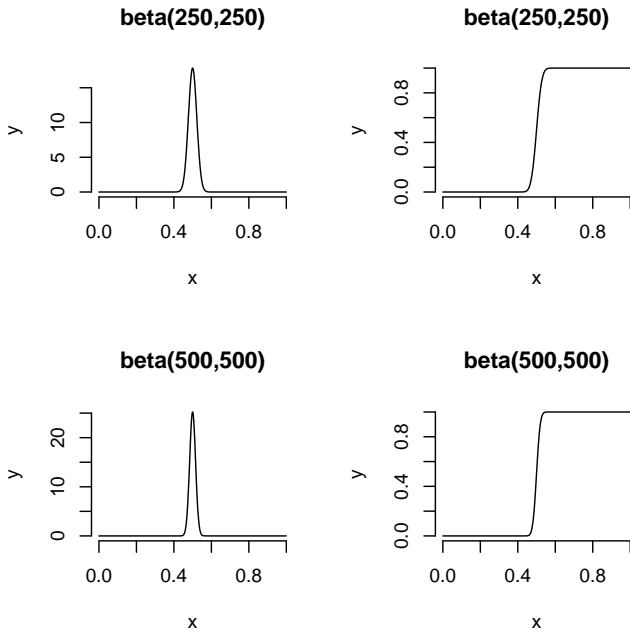


Abb. 8.5: Beta-Verteilungen für kleine Parameterwerte

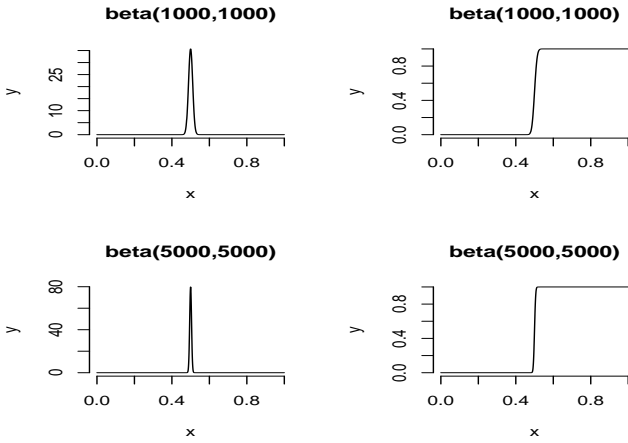


Abb. 8.6: Beta-Verteilungen für große Parameterwerte



8.3 Prior – Sample – Posterior

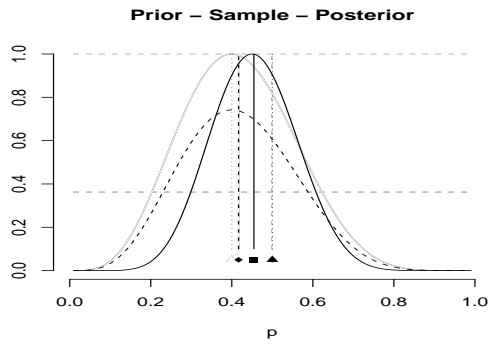


Abb. 8.7: Schnapschuss vom Experiment



8.4 Beta-Verteilung

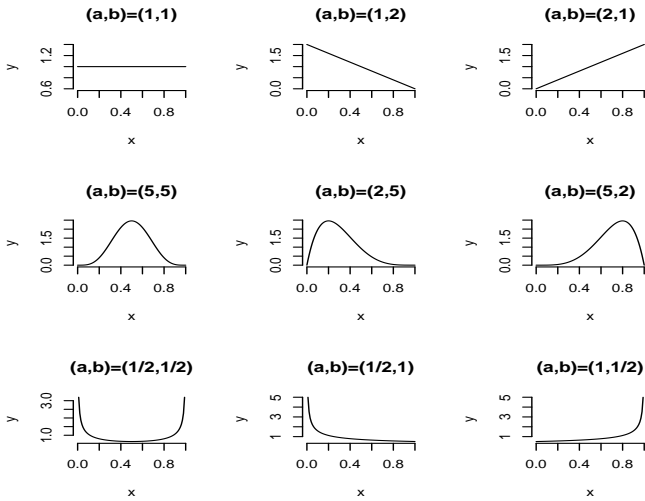


Abb. 8.8: Dichten von Beta-Verteilungen



8.5 Es hilft auch im Weltall

9 Testen



9.1 Kochen und Testen

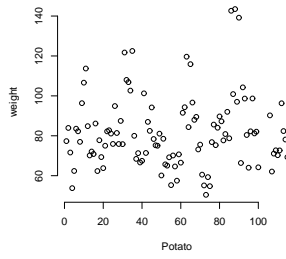


Abb. 9.1: Kartoffelfeldaten

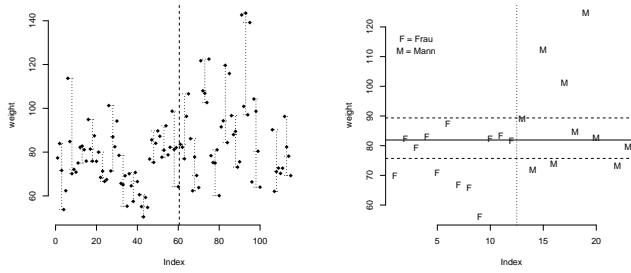


Abb. 9.2: Einfache Box-and-Whisker bzw. geordnete Studentennittel

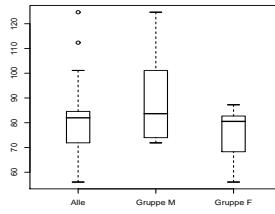


Abb. 9.3: Kartoffel-Boxplots

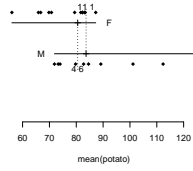


Abb. 9.4: Einfacher Binomialtest

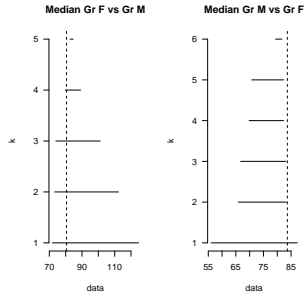


Abb. 9.5: Einfacher Konfidenztest

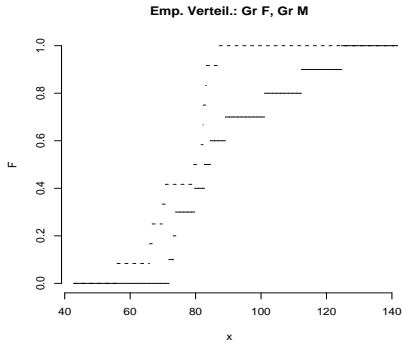


Abb. 9.6: Empirische Verteilungsfunktionen

9.2 Der Aufbau eines Tests

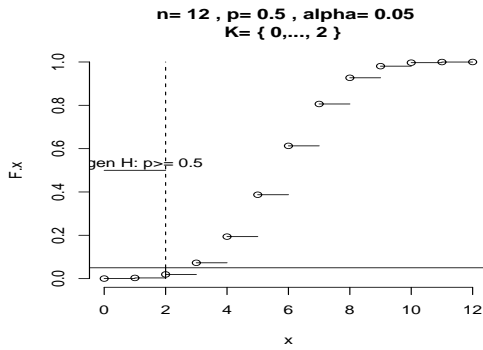


Abb. 9.7: Kritische Bereiche des Binomialtests

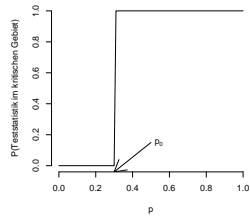


Abb. 9.8: Wunschverlauf der Güte

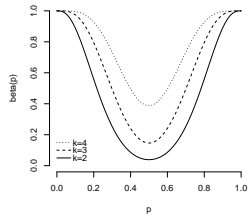


Abb. 9.9: Power des Binomialtests

9.3 Der χ^2 -Test: Ein vielseitiger Geselle

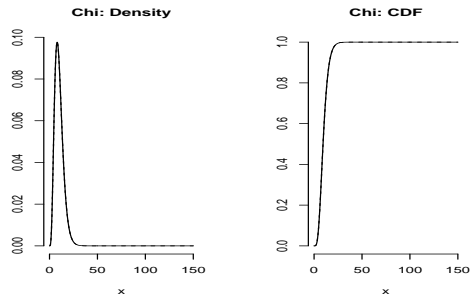


Abb. 9.10: χ^2 -Verteilung für $\nu = 10$

9.4 Eine kleine Testgalerie

10 Regressionsanalyse



10.1 Eine Reise für den Überblick

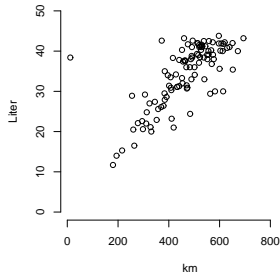


Abb. 10.1: Litern / km

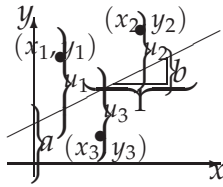


Abb. 10.2: Symbole des Geradenmodells

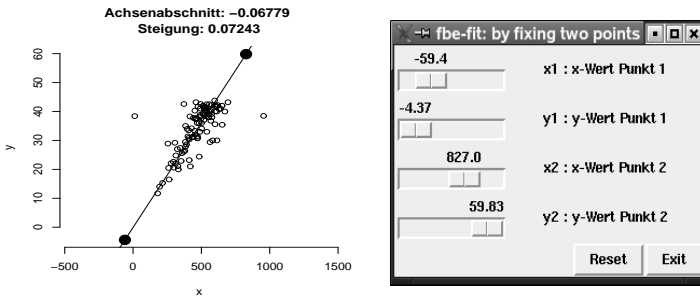


Abb. 10.3: Verbrauch gegen Kilometer mit fbe-Gerade

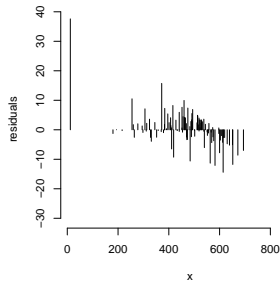


Abb. 10.4: Residualplot

10.2 Das lineare Regressionsmodell



10.3 Modell-Schätzung und -Check

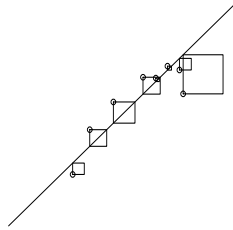


Abb. 10.5: KQ-Flächenquadrate

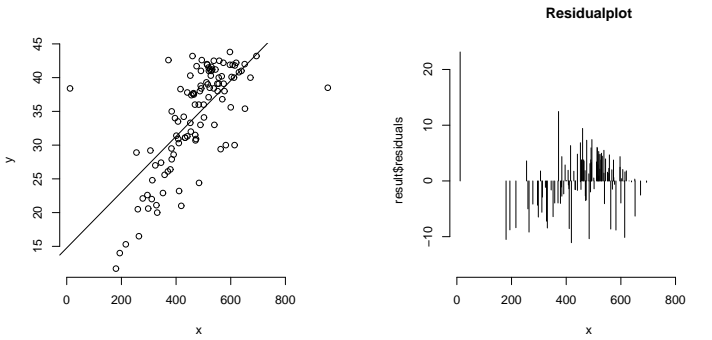


Abb. 10.6: Verbrauchsdaten mit KQ-Gerade und Residualplot

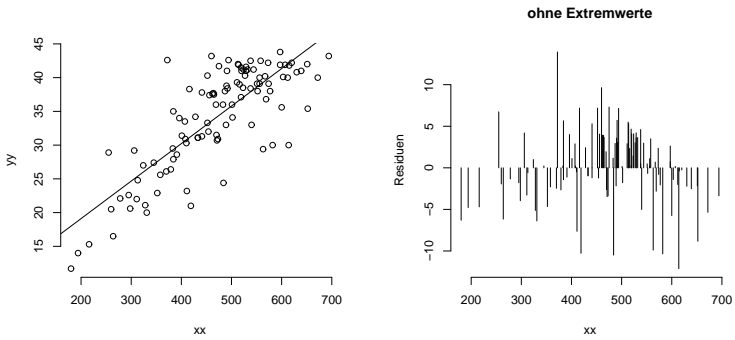
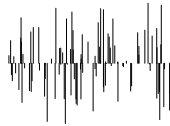


Abb. 10.7: Verbrauchsdaten mit KQ-Gerade und Residualplot



strukturloser Residualplot



Ausreißer

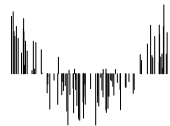


var(u) mit x steigend



x

x \leftrightarrow y nicht linear



x

Abb. 10.8: Verschiedene Residualplots

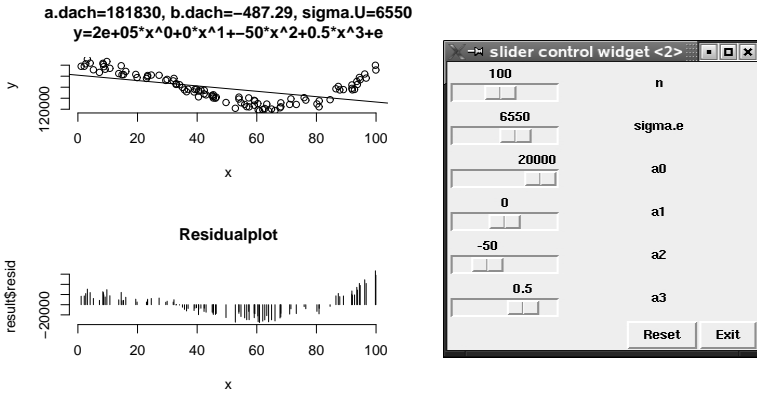


Abb. 10.9: Nichtlinearitäten

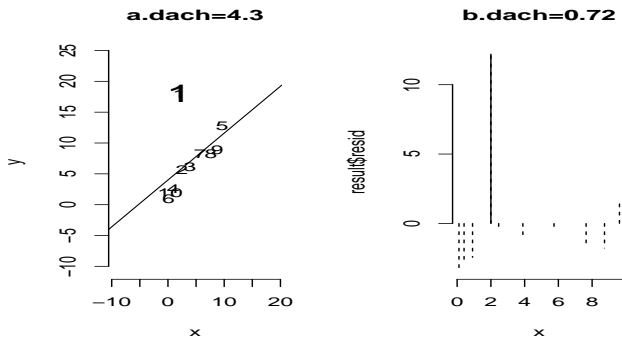


Abb. 10.10: Experiment zum Auswirkungsstudium

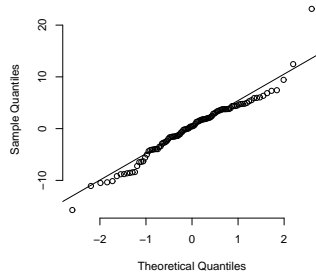


Abb. 10.11: qqnorm-Plot der geschätzten Störungen



10.4 Modell-Interpretation

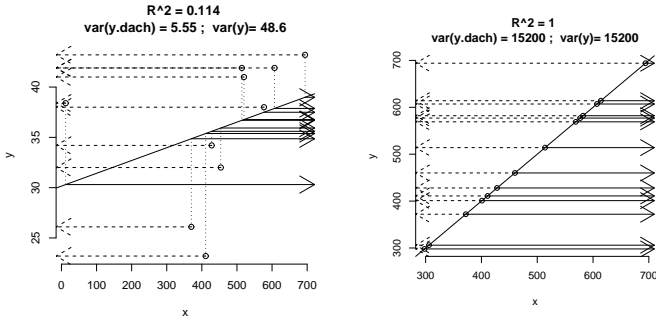


Abb. 10.12: Zähler- und Nenner von R^2 , rechts: Punkte auf Geraden

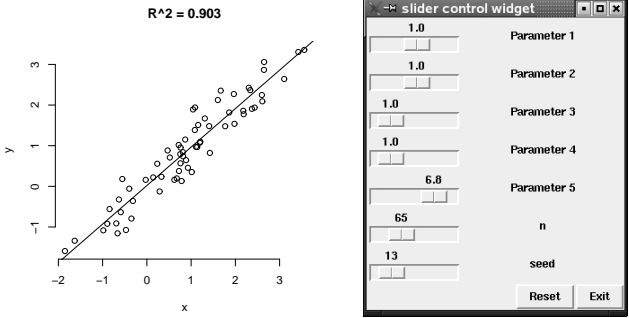


Abb. 10.13: ein Beispiel für $R^2 \approx 90\%$

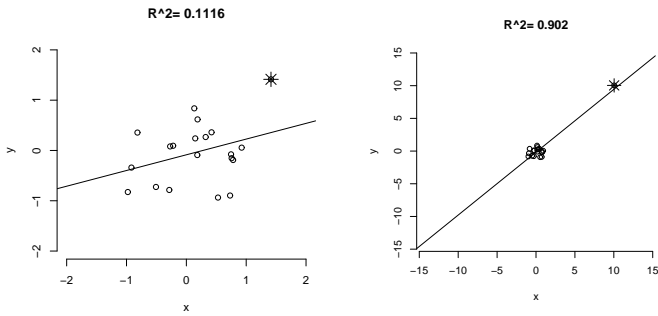


Abb. 10.14: R^2 mit einem moderaten bzw. einem schweren Ausreißer

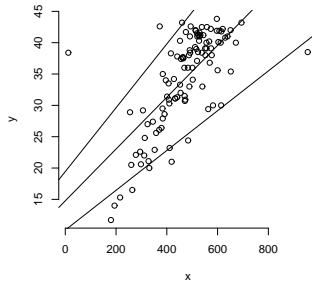


Abb. 10.15: Geraden zu extremen KIs

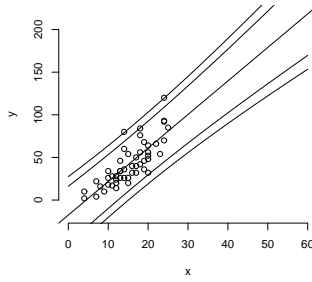


Abb. 10.16: Prognosebereiche

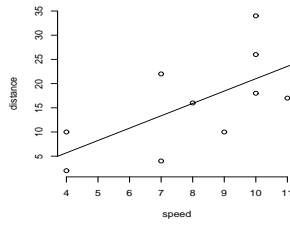


Abb. 10.17: distance gegen speed

10.5 Ausblick

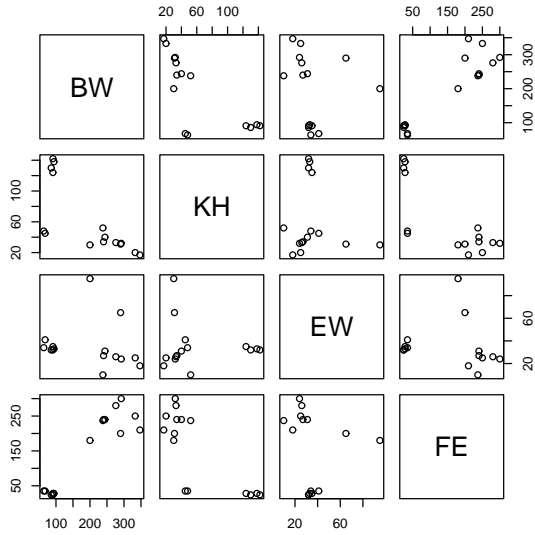


Abb. 10.18: Inhaltsstoffe einiger Milchprodukte

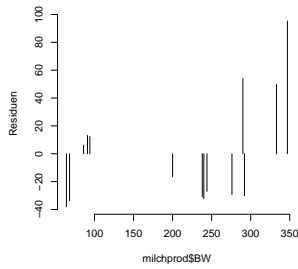


Abb. 10.19: Residuen des Milchprodukte

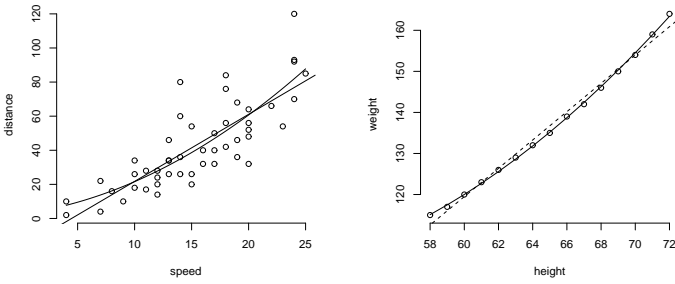


Abb. 10.20: Anpassung einer Geraden und einer Parabel an cars

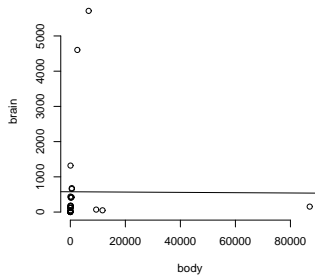


Abb. 10.21: Gehirn- / Körpergewichte

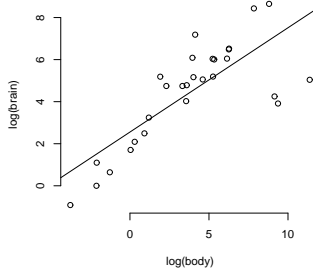


Abb. 10.22: Gewichte logarithmiert

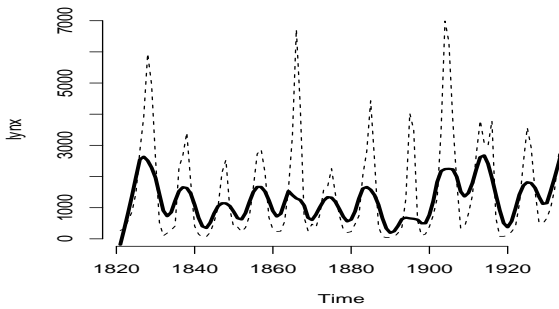


Abb. 10.23: Luchsbestand von 1821 bis 1934

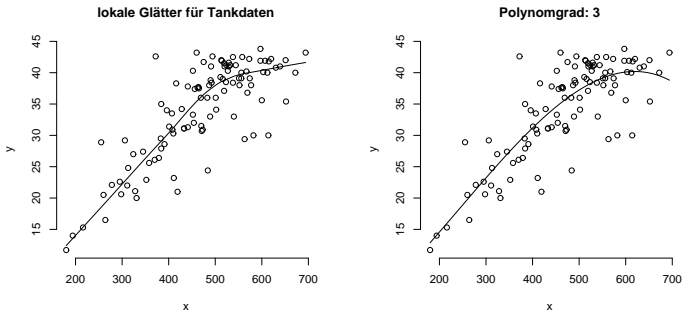


Abb. 10.24: Tankdaten mit Glättungskurve (links) und mit einem angepassten Polynom 3. Grades (rechts)

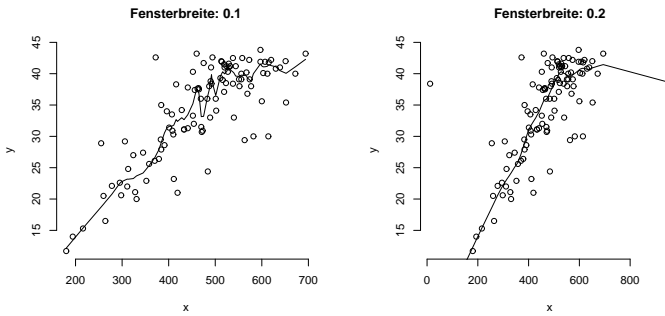


Abb. 10.25: Beispiel für unterschiedliche Glättungen

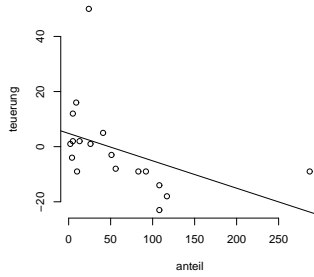


Abb. 10.26: Teuerung gegen Anteil

11 R-Einführung



11.1 Hintergrund, Installation und erste Schritte mit **R**

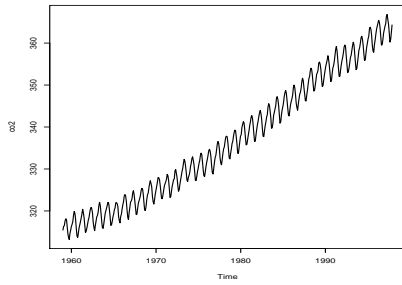


Abb. 11.1: Plot der CO₂



11.2 Daten einlesen und Statistiken berechnen

11.3 Graphiken erstellen

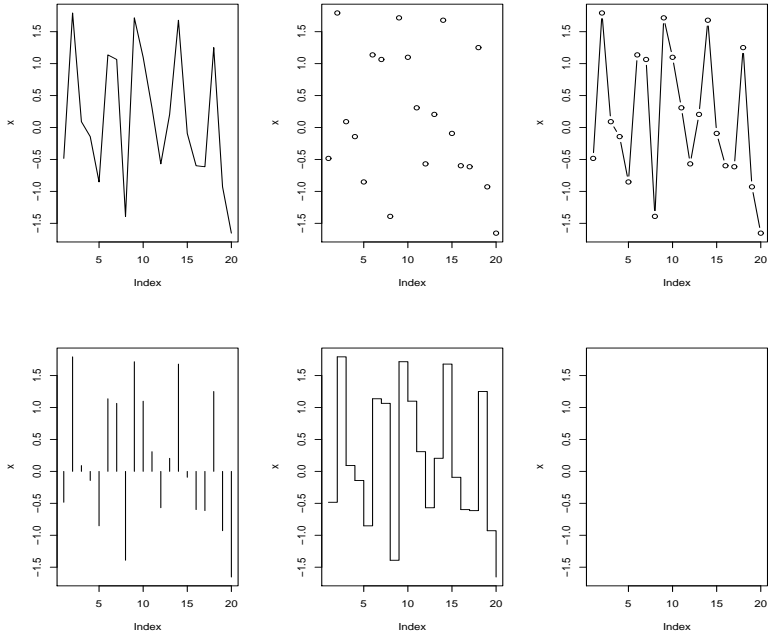


Abb. 11.2: Verschiedene type-Setzungen



11.4 R als Rechenmaschine



11.5 Bequemes Arbeiten mit diesem Buch in R



Abb. 11.3: Von `open.wnt()` erzeugte Oberfläche



11.6 Statistische R-Idioms

11.7 Weitere Infos



11.8 Codechunks

11.9 Funktionen



11.10 Daten