Grundlagen der Biometrie

Skriptenheft des IMB
zur Ausbildung im Fach „Medizinische Biometrie“
im Rahmen des Medizin- bzw. Public Health-Studiums
an der Medizinischen Fakultät Carl Gustav Carus
der TU Dresden

Sollten Ihnen Fehler oder Verbesserungsmöglichkeiten auffallen, dann teilen Sie uns das bitte mit. Danke!

imblehre@mail.zih.tu-dresden.de

Impressum:
Autoren: Prof. Dr. I. Röder, Dipl.-Math. M. Kuhn, Dr. I. Glauche, Ph. Schulze, T. Lange

In Teilen unter Verwendung des IMISE-Skriptes (Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig) bereitgestellt durch Prof. Dr. M. Löffler

Institut für Medizinische Informatik und Biometrie (IMB)
Technische Universität Dresden
Haus 105, Blasewitzer Straße 86
01307 Dresden
Tel.:(0351) 458 6060 Fax: (0351) 458 7222
`http://tu-dresden.de/med/imb`

Version 2.1.6 (10. März 2016)
© Alle Rechte vorbehalten. Nachdruck nur mit ausdrücklicher Genehmigung und Quellenangabe gestattet.

Inhaltsverzeichnis

Inhaltsverzeichnis **II**

Abbildungsverzeichnis **III**

Tabellenverzeichnis **IV**

1 Wahrscheinlichkeitsrechnung **1**
1.1 Zufallsexperiment . . . 1
1.2 Wahrscheinlichkeit . . . 3
1.3 Wahrscheinlichkeitsverteilungen . . . 9
1.4 Zusammenfassung . . . 18

2 Deskriptive Statistik **19**
2.1 Grundbegriffe der deskriptiven Statistik . . . 19
2.2 Kenngrößen/Maßzahlen . . . 25
2.3 Zusammenfassung . . . 30

3 Schätzen **31**
3.1 Punktschätzer . . . 31
3.2 Intervallschätzungen . . . 34
3.3 Zusammenfassung . . . 38

4 Statistisches Testen - Signifikanztests **39**
4.1 Einführung . . . 39
4.2 Statistische Hypothesen . . . 40
4.3 Testentscheidung . . . 41
4.4 Beurteilung des Testergebnisses . . . 42
4.5 Interpretation des Testergebnisses . . . 44
4.6 Versuchsplanung und Fallzahlberechnung . . . 45
4.7 Zusammenhang von Hypothesentesten und Konfidenzintervallschätzungen . . . 46
4.8 Multiples Testen . . . 47
4.9 Ablaufschema eines statistischen Signifikanztests . . . 48
4.10 Zusammenfassung . . . 49

5 Spezielle Testverfahren **50**
5.1 Überblick . . . 50
5.2 Lagetests . . . 52
5.3 Tests bezüglich relativer Häufigkeiten . . . 67
5.4 Anpassungstests . . . 77
5.5 Systematisierung der vorgestellten Tests . . . 80

6 Korrelation und Regression **81**
6.1 Korrelation . . . 81
6.2 Lineare Regression . . . 88

7 Ereigniszeitanalyse **98**
7.1 Einführung . . . 98
7.2 Deskription von Ereigniszeiten . . . 101
7.3 Vergleich der Überlebenszeiten mehrerer Gruppen . . . 107

8 Appendix **115**
8.1 Packages . . . 115
8.2 Kapitel 1 . . . 115

8.3 Kapitel 2 122
8.4 Kapitel 3 125
8.5 Kapitel 4 126
8.6 Kapitel 5 132
8.7 Kapitel 6 135
8.8 Kapitel 7 147

Abbildungsverzeichnis

1.1 Venn-Diagramme . . . 3
1.2 Auftreten von Krankheit K in Altersgruppen A1,A2,A3 . . . 7
1.3 Wahrscheinlichkeitsbaum . . . 9
1.4 Wahrscheinlichkeitsfunktion vs. Verteilungsfunktion . . . 11
1.5 Binomiale -Wahrscheinlichkeits- und Verteilungsfunktion . . . 14
1.6 Zusammenhang von Verteilungs- und Dichtefunktion . . . 15
1.7 Variation des Erwartungswertes . . . 16
1.8 Intervalgrenzen . . . 17
1.9 Logarithmische Normalverteilung . . . 18

2.1 Balken- und Kreisdiagramm . . . 23
2.2 absolute Häufigkeiten . . . 24
2.3 typische Verteilungstypen . . . 25
2.4 Fehlerbalken . . . 29
2.5 Boxplot . . . 29
2.6 Darstellung Boxplot und Histogramm . . . 30

3.1 Verteilung der Einzelwerte . . . 32
3.2 Verteilung der Mittelwerte . . . 33
3.3 Beispiel für Intervallschätzung . . . 35

4.1 Ablehnungsbereich . . . 42
4.2 ein- und zweisseitige Fragestellung . . . 42
4.3 Ablehnungsbereich a=0 . . . 43
4.4 signifikantes Testergebnis . . . 44
4.5 nicht signifikantes Testergebnis . . . 44
4.6 Verteilung der Prüfgröße . . . 46

5.1 Vergleich von unterschiedlichen Normalverteilungen . . . 52
5.2 Vergleich von t-Verteilung mit verschiedenen Freiheitsgraden . . . 53
5.3 Voraussetzungen . . . 62
5.4 QQ-Plot . . . 79

6.1 Verteilungformen . . . 81
6.2 Scatterplot . . . 82
6.3 Punktwolke . . . 83
6.4 positive Korrelation . . . 84
6.5 negative Korrelation . . . 84
6.6 Verteilung in Abhangigkeit von r . . . 84
6.7 Veranschaulichung von SP(X,Y) . . . 85
6.8 Exponentiale Verteilung . . . 86
6.9 Ausreißer . . . 87
6.10 Inhomogenitat . . . 87
6.11 Größenvergleich bei Vätern und Söhnen . . . 88
6.12 Regressionsgerade . . . 90
6.13 Regressionskoeffizient . . . 91
6.14 Residuenanalyse . . . 92
6.15 Muster in Residuen . . . 93
6.16 Trichterform der Residuen . . . 93
6.17 Residuenplot . . . 93
6.18 Normalverteilung der . . . 94
6.19 Konfidenzintervalle zur Vorhersage . . . 96

7.1 Studienzeit und Patientenzeit 99
7.2 Theorie Ueberlebenszeitkurve 102
7.3 Intervalle 103
7.4 Ueberlebenszeitkurve 105
7.5 Ueberlebenszeitkurve nach Tüchler 106
7.6 Mediane Überlebenszeit 107
7.7 Überlebenszeiten der beiden Gruppen 109
7.8 Rezidiv des Morbus Hodgkin 112
7.9 Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache 112
7.10 Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache oder Sekundärneoplasie 113
7.11 Tod jeglicher Ursache 113

Tabellenverzeichnis

1.1 Mehrfachwurf eines sechsseitigen Würfels . . . 5
1.2 Merkmale und Ausprägungen . . . 10
1.3 Wurfwahrscheinlichkeit . . . 11
1.4 Auftreten von k Komplikationen . . . 13

2.1 Patienten mit Mundhöhlenkarzinom . . . 20
2.2 Skalenklassifikation . . . 21
2.3 falsche Datenerfassung . . . 22
2.4 richtige Datenerfassung . . . 22
2.5 Register der Tumoren des Kiefer- und Gesichtsbereiches . . . 23
2.6 häufig verwendete Quantile . . . 28

3.1 HDL-Choleserolwerte . . . 34
3.2 (E: Exposition, $\overline{E}$: keine Exposition, D: krank, $\overline{D}$: gesund) . . . 37

4.1 Plasma - Magnesium - Spiegel . . . 39
4.2 Fehlentscheidungen . . . 41
4.3 Stichprobendaten . . . 42
4.4 theoretische Wahrscheinlichkeiten . . . 47

5.1 Arten von t-Tests . . . 52
5.2 Gruppenvergleich . . . 57
5.3 Gruppenvergleich vor und nach der Therapie . . . 57
5.4 Testvergleich . . . 58
5.5 systolischer Blutdruck bei 20 Patienten . . . 59
5.6 Blutdruckbehandlung bei weiblichen Infarktpatienten . . . 61
5.7 Cholesterinwerte der Kontrollgruppe . . . 63
5.8 Cholesterinwerte der Infarktgruppe . . . 63
5.9 Patientenzufriedenheit . . . 65
5.10 ANOVA Auswertung . . . 66
5.11 Übersicht zu Lagetests. Tests mit * werden im Skript nicht behandelt. . . . 66
5.12 Binomialtest für eine Ereignis-Wahrscheinlichkeit 1 . . . 67
5.13 Binomialtest für eine Ereignis-Wahrscheinlichkeit 2 . . . 67
5.14 Augenzahl beim Würfeln . . . 69
5.15 Kreuztabelle . . . 70
5.16 allgem. Vierfeldertafel . . . 70
5.17 beobachtete und erwartete Häufigkeit beim Vierfelder-Test . . . 71
5.19 Vierfeldertafel . . . 72
5.18 χ^2-Unabhängigkeitstest . . . 72
5.20 Verteilung der Leukopenie . . . 74
5.21 Therapieergebnisse . . . 75
5.22 Vierfeldertafel beim McNemar-Test SP: Stichprobe A: ein beliebiges dichotomes Merkmal 76
5.23 Vierfeldertafel der Leukozyten . . . 77
5.24 Verteilung der Blutgruppen . . . 78

6.1 systolischer Blutdruck . . . 82
6.2 Korrelation . . . 86

7.1 Überlebenszeiten der 6-Mercaptopurin Gruppe . . . 104
7.2 geschätzten Überlebenswahrscheinlichkeiten . . . 105
7.3 Schätzung der Überlebenszeiten Gruppe 2 . . . 109
7.4 Rechnung zum LogRank Test . . . 110

1 Wahrscheinlichkeitsrechnung

Jede Wiederholung eines realen Experimentes führt im Rahmen bestimmter Schwankungen zu einem anderen Ergebnis, selbst wenn man sich bemüht, die Versuchs- und Messbedingungen konstant zu halten. Diese unterschiedlichen Ergebnisse entstehen durch geringfügige, sich der Kontrolle entziehender Veränderungen der Versuchsbedingungen und durch Messfehler. Bei Experimenten an biologischen Versuchsobjekten kommt hinzu, dass diese nicht alle identisch sind, sondern individuelle Unterschiede aufweisen, die in ihrer Vielfalt nicht alle erfassbar und zu berücksichtigen sind. Das Versuchsergebnis ist daher für das einzelne Experiment nicht exakt vorhersagbar; die Vorhersage ist mit Unsicherheit behaftet. Man spricht bei solchen Experimenten von Zufallsexperimenten.
Die empirische Beobachtung und die theoretische Betrachtung von Erscheinungen mit Zufallscharakter haben jedoch gezeigt, dass bei einer großen Zahl gleichartig wiederholter Experimente summarische, quantitative Aussagen über die Versuchsergebnisse und den Grad der Sicherheit dieser Aussagen möglich sind. Die Beschreibung dieser Aussagen erfolgt mit dem Begriffssystem der Wahrscheinlichkeitstheorie.

1.1 Zufallsexperiment

Ein *Zufallsexperiment* bezeichnet einen mehrfach, prinzipiell in identischer Anordnung, durchführbaren Vorgang mit zufälligem Ausgang. Charakteristische Eigenschaften eines Zufallsexperimentes sind:

- Durchführung nach bestimmter (klar definierter) Vorschrift
- prinzipielle Wiederholbarkeit
- es sind mehrere (vorher bekannte) Ausgänge einer Realisierung möglich
- der Ausgang jeder einzelnen Realisierung ist ungewiss

Beispiel: 1.1.1

Münzwurf:

- mehrfaches Werfen einer Münze nach definierter Regel
- es gibt genau zwei mögliche, bekannte, aber nicht vorhersagbare Versuchsausgänge: *Wappen* oder *Zahl*

Hämatokritmessung bei Blutspendern:

- Beobachtung von mehreren („gleichartigen“) Spendern
- Hämatokrit immer positiv (Prozentwert) und im Einzelfall nicht vorhersagbar (Normalwert im Bereich von 37-50%)

1.1.1 Grundgesamtheit (Menge der Elementarereignisse)

Die Menge aller möglichen Ausgänge eines Zufallsexperiments bildet die *Grundgesamtheit* (GGH). Die Grundgesamtheit wird auch als Stichprobenraum bezeichnet und häufig mit Ω benannt. Die Grundgesamtheit umfasst die Menge aller *Elementarereignisse*. Elemantarereignisse sind die kleinsten, nicht zusammengesetzten Elemente der Grundgesamtheit. Sie sind daher einelementige Teilmengen von Ω.

Beispiel: 1.1.2

Spiel mit zwei sechsseitigen Würfeln:

Die Grundgesamtheit Ω besteht aus $6^2 = 36$ Elementarereignissen:

$$\begin{aligned}\Omega = \{ \quad & \{(1,1)\},\{(1,2)\},\{(1,3)\},\{(1,4)\},\{(1,5)\},\{(1,6)\},\\ & \{(2,1)\},\{(2,2)\},\{(2,3)\},\{(2,4)\},\{(2,5)\},\{(2,6)\},\\ & \{(3,1)\},\{(3,2)\},\{(3,3)\},\{(3,4)\},\{(3,5)\},\{(3,6)\},\\ & \{(4,1)\},\{(4,2)\},\{(4,3)\},\{(4,4)\},\{(4,5)\},\{(4,6)\},\\ & \{(5,1)\},\{(5,2)\},\{(5,3)\},\{(5,4)\},\{(5,5)\},\{(5,6)\},\\ & \{(6,1)\},\{(6,2)\},\{(6,3)\},\{(6,4)\},\{(6,5)\},\{(6,6)\};\ \}\end{aligned}$$

1.1.2 Das Ereignis

Teilmengen von Ω nennt man *Ereignisse*, die üblicherweise mit lateinischen Großbuchstaben bezeichnet werden (z.B. A, B etc.). Ein bestimmtes Ereignis kann durch ein oder mehrere Versuchsausgänge realisiert werden.

Beispiel: 1.1.3

Ereignis: Augensumme = 6 beim Werfen zweier sechsseitiger Würfel

- die Teilmenge A_6 beinhaltet alle Elementarereignisse, für die die Augensumme 6 ist
  $A_g = \{\ \{(1,5)\},\ \{(5,1)\},\ \{(2,4)\},\ \{(4,2)\},\ \{(3,3)\}\}$

Einige Ereignisse haben eine besondere Bedeutung und tragen daher spezielle Namen:

- **Elementarereignis:** 1-elementige Teilmengen von Ω. Elementarereignisse enthalten nur einen möglichen Versuchsausgang.
- **sicheres Ereignis:** Das Ereignis, welches die gesamte Grundgesamtheit umfasst. Das sichere Ereignis tritt daher bei jeder Realisierung mit Sicherheit ein.
- **unmögliches Ereignis:** die leere Menge ($\emptyset$). Dieses Ereignis tritt bei einer Realisierung des Zufallsexperimentes mit Sicherheit nicht auf.

1.1.3 Verknüpfungen von Ereignissen

Bei vielen Fragestellungen interessieren oft nicht nur einzelne Ereignisse sondern Kombinationen von Ereignissen. Solche Kombinationen können sowohl einschließenden als auch ausschließenden Charakter haben.

Beispiel: 1.1.4

Myokardinfarkt - Befunderhebung:

- keine krankheitsspezifischen Symptome **aber** EKG-Veränderungen
- Troponin positiv **und** EKG-Veränderungen
- charakteristische Schmerzsymptomatik **und** Troponin positiv **aber keine** EKG-Veränderungen

spezielle Patientengruppen:

- Diabetiker **und** Raucher **und** Hypercholesterinämie
- Myokardinfarkt, **aber kein** Diabetiker
- Mammakarzinom, **aber kein** Lymphknotenbefall

Die Beschreibung von Ereignissen in Mengendarstellung erlaubt die Anwendung der klassischen Mengenlogik. Mengenbeziehungen lassen sich wie folgt darstellen:

- **Schnittmenge** ($A \cap B$): sowohl A als auch B treten ein
- **Vereinigungsmenge** ($A \cup B$): entweder A oder B oder beide treten ein
- **Differenzmenge** ($A \setminus B$): A aber nicht B tritt ein
- **Negation** ($\overline{A}$): A tritt nicht ein [1]
- **Disjunkte Ereignisse** ($A \cap B = \emptyset$): A und B haben keine Elemente gemeinsam

Venn-Diagramme (benannt nach dem britischen Mathematiker *John Venn* (1834-1923)) haben sich als übliche Methode zur grafischen Darstellung solcher Mengenbeziehungen etabliert (vgl. Abb. 1.1).

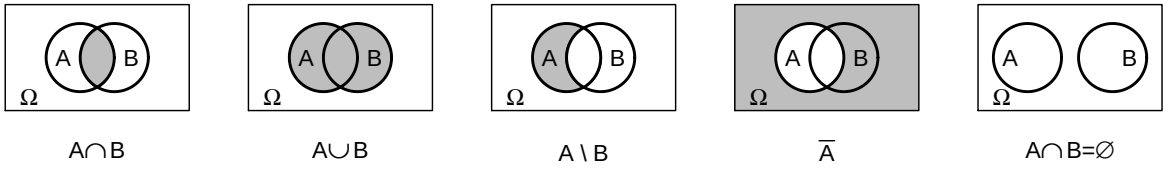


Abbildung 1.1: Venn-Diagramme

Beispiel: 1.1.5

Blutdruckmessung:
Bei der Blutdruckmessung wird sowohl der systolische als auch der diastolische Blutdruck (BD) gemessen. Beide Messwerte sind positiv und können nur Werte bis zu einem gewissen oberen Grenzwert annehmen. Dabei ist der systolische Blutdruck immer größer als der diastolische Blutdruck. Innerhalb dieser Beschränkungen können systolischer und diastolischer Blutdruck alle reellen Werte annehmen. Die Ereignisse SH und DH bezeichnen das Auftreten einer systolischen Hypertonie (systolischer BD über $140 mmHg$) bzw. diastolischen Hypertonie (diastolischer BD über $90 mmHg$). Folgende Ereigniskombinationen sind denkbar:

$SH \cup DH$	=	systolische **oder** diastolische Hypertonie (dies schließt das gleichzeitige Auftreten einer syst. und diast. Hypertonie ein, da sich beide Ereignisse nicht gegenseitig ausschließen)
$SH \cap DH$	=	systolische **und** diastolische Hypertonie
$\overline{SH}$	=	**keine** systolische Hypertonie

1.2 Wahrscheinlichkeit

1.2.1 Definition der Wahrscheinlichkeit

Zur Untersuchungen von Glücksspielen wurde von *Pierre-Simon Laplace* (1749-1827) ein heute als „klassisch“ bezeichneter Wahrscheinlichkeitsbegriff eingeführt. Er geht davon aus, dass alle Elementarereignisse gleichwahrscheinlich sind. Die Wahrscheinlichkeit eines Ereignisses A errechnet sich demnach aus dem

[1] $\overline{A}$ - sprich „A Komplement“

Quotienten der für A günstigen Ereignisse zur Anzahl aller möglichen Ereignisse, d.h.[2]

$$P(A) = \frac{\text{Anzahl der Elemente von } A}{\text{Anzahl der Elemente von } \Omega} = \frac{|A_g|}{|\Omega|} \tag{1.1}$$

Beispiel: 1.2.1

Skatblatt (32 Karten der Farben: K=Kreuz, P=Pik, H=Herz, S=Karo):

Berechnung der Wahrscheinlichkeit ein Ass aus einem vollständigen Skatblatt zu ziehen:

gegeben:

$$\begin{aligned}\Omega = \{ \quad & \{S_7\},\{S_8\},\{S_9\},\{S_{10}\},\{S_{Bube}\},\{S_{Dame}\},\{S_{König}\},\{S_{Ass}\},\\ & \{H_7\},\{H_8\},\{H_9\},\{H_{10}\},\{H_{Bube}\},\{H_{Dame}\},\{H_{König}\},\{H_{Ass}\},\\ & \{P_7\},\{P_8\},\{P_9\},\{P_{10}\},\{P_{Bube}\},\{P_{Dame}\},\{P_{König}\},\{P_{Ass}\},\\ & \{K_7\},\{K_8\},\{K_9\},\{K_{10}\},\{K_{Bube}\},\{K_{Dame}\},\{K_{König}\},\{K_{Ass}\}\ \}\end{aligned}$$

$$A_g = \text{alle Karten, die ein Ass sind} = \{\ \{S_{Ass}\},\{H_{Ass}\},\{P_{Ass}\},\{K_{Ass}\}\ \}$$

gesucht: $P(A_g)$

Lösung:

$$P(A_g) = \frac{|A_g|}{|\Omega|} = \frac{|\{\{S_{Ass}\},\{H_{Ass}\},\{P_{Ass}\},\{K_{Ass}\}\}|}{|\{\{S_7\},\{S_8\},\cdots,\{K_{Ass}\}\}|} = \frac{4}{32} = \frac{1}{8} = 12,5\%$$

Aufbauend auf der Laplaceschen Definition wurden folgende, allgemeinere Axiome für Wahrscheinlichkeiten aufgestellt. Diese Axiome setzen nicht voraus, dass alle Elementarereignisse mit gleicher Wahrscheinlichkeit eintreten. Die hier angegebenen Axiome sind eine *vereinfachte* Darstellung der von *Andrei N. Kolmogorow* (1903-1987) formulierten Definition:

1. **Nichtnegativität:** Für jedes Ereignis A gilt: $0 \leq P(A) \leq 1$
2. **Normierung:** Für das sichere Ereignis $S = \{\Omega\}$ gilt: $P(S) = 1$
3. **Additivität:** Für disjunkte Ereignisse A und B gilt: $P(A \cup B) = P(A) + P(B)$.

1.2.2 Relative Häufigkeit und Wahrscheinlichkeit

Vor Durchführung eines einzelnen Versuchs ist eine individuelle Vorhersage über das Auftreten eines bestimmten Ereignisses A nicht möglich. Beobachtet man jedoch das Eintreten von A bei einer großen Anzahl von Versuchen unter gleichen Versuchsbedingungen, so stellt man unter Umständen bestimmte, auf die Gesamtheit der Versuchswiederholungen bezogene Gesetzmäßigkeiten fest. Betrachtet man das Verhältnis aus der **absoluten Häufigkeit** $h_a(A)$ für das Auftreten von A und der Gesamtanzahl der Versuche n, so zeigt sich, dass der Quotient $h_r(A) = h_a(A)/n$ für große n bei mehrfacher Wiederholung des Versuches annähernd gleiche Werte annimmt. Der Quotient $h_r(A)$ wird als **relative Häufigkeit** des Ereignisses A bezeichnet und zeigt eine deutliche Stabilität, wenn die Anzahl der in einer Serie durchgeführten Versuche hinreichend groß ist. Die relative Häufigkeit h_r schwankt um den Wert der theoretischen Wahrscheinlichkeit des Ereignisses A und nähert sich ihm um so mehr, je größer die Anzahl der Versuchswiederholungen ist. Die mathematische Begründung dafür gibt das „**Gesetz der großen Zahlen**“ von *Jakob I. Bernoulli* (1655-1705). Diese Konvergenzeigenschaft rechtfertigt die experimentelle Bestimmung von Wahrscheinlichkeiten durch relative Häufigkeiten in genügend großen Stichproben.

[2] $|\ldots|$ bezeichnet die Mächtigkeit einer Menge, d.h. die Anzahl ihrer Elemente.

Beispiel: 1.2.2

Wurf mit einem idealen sechsseitigen Würfel:

Betrachtung des Falls, dass eine Eins gewürfelt wird − also das Ereignis $\{1\}$

Anzah der Würfe n	Anzahl der gewürfelten Einsen $h_a(\{1\})$	rel. Häufigkeit von von $\{1\}$ $h_r(\{1\})$
50	5	0.1
100	13	0.13
500	88	0.176
1000	159	0.159
5000	822	0.1644

Tabelle 1.1: Mehrfachwurf eines sechsseitigen Würfels

- Wahrscheinlichkeit des Ereignisses $\{1\}$: $\mid \Omega \mid = 6$ und $\mid A_g \mid = 1$
  $\Rightarrow \mathrm{P}(A_g) = \frac{1}{6} = 0,1\overline{6}$

1.2.3 Abhängigkeit und bedingte Wahrscheinlichkeit

Um die Wahrscheinlichkeit für das Auftreten bestimmter Ereigniskonstellationen zu berechnen, ist es wichtig zu wissen, in welcher Beziehung die Ereignisse zueinander stehen, d.h. ob sie voneinander abhängig oder unabhängig sind.
Zwei Ereignisse heißen voneinander **stochastisch abhängig**, wenn das Auftreten des einen Ereignisses die Wahrscheinlichkeit für das Auftreten des anderen Ereignisses beeinflusst. Anderenfalls heißen die Ereignisse **stochastisch unabhängig**. Beispielsweise sind systolischer und diastolischer Blutdruck eines Patienten voneinander abhängig, während der systolische Blutdruck bei zwei unverwandten, völlig getrennt lebenden Menschen voneinander unabhängig ist.
Zur Beschreibung des Sachverhaltes stochastisch abhängiger/unabhängiger Ereignisse wird der Begriff **bedingte Wahrscheinlichkeit** eingeführt. Es handelt sich dabei um die Wahrscheinlichkeit eines Ereignisses A unter der Bedingung, dass ein anderes Ereignis B eingetreten ist, abgekürzt durch $P(A|B)$. Mathematisch ist die bedingte Wahrscheinlichkeit definiert als

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \tag{1.2}$$

Zwei zufällige Ereignisse A und B sind stochastisch unabhängig, wenn die bedingte Wahrscheinlichkeit $P(A|B)$ von der Bedingung B nicht abhängt, d.h. wenn gilt: $P(A|B) = P(A)$.

Beispiel: 1.2.3

Erhöhtes Risiko für ein Burnout-Syndrom bei Lehrern $P(A|B) \geq P(A)$:

- Ereignis A = Burnout-Syndrom
- Bedingung B = Lehrer

Erhöhtes Thromboserisiko unter gleichzeitiger Behandlung mit Antikonzeptiva $P(A|B) \geq P(A)$:

- Ereignis A = Auftreten einer Thrombose
- Bedingung B = Behandlung mit Antikonzeptiva

1.2.4 Rechenregeln

Aufbauend auf den Axiomen in Abschnitt 1.2.1 lassen sich einige wichtige Rechenregeln für Wahrscheinlichkeiten ableiten.

Addition von Wahrscheinlichkeiten. Für die Vereinigung zweier Ereignisse A und B gilt der **Additionssatz**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.3}$$

Der Additionssatz vereinfacht sich für disjunkte Ereignisse (d.h. $A \cap B = \emptyset$):

$$P(A \cup B) = P(A) + P(B) \tag{1.4}$$

Beispiel: 1.2.4

Seien A, B und C Krankheitssymptome:

gegeben:
Die Auftretenswahrscheinlichkeiten betragen:
$P(A) = 0,5$; $P(B) = 0,3$; $P(C) = 0,2$; $P(A \cap B) = 0,25$; $P(B \cap C) = 0,05$; $P(A \cap C) = 0$

gesucht:

(a) Wahrscheinlichkeit, dass Symptom A **oder** B auftritt

(b) Wahrscheinlichkeit, dass Symptom A **oder** C auftritt

(c) Wahrscheinlichkeit, dass Symptome A **und** B gemeinsam auftreten **oder** C allein auftritt

Lösung:

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,5 + 0,3 - 0,25 = 0,55$
(b) $P(A \cup C) = P(A) + P(C) - P(A \cap C) = 0,5 + 0,2 + 0 = 0,7$
(c) $P((A \cap B) \cup (C \setminus B)) = P(A \cap B) + P(C \setminus B) = P(A \cap B) + P(C) - P(B \cap C) = 0,25 + 0,2 - 0,05 = 0,4$

Wahrscheinlichkeit für komplementäre Ereignisse. Aus der Wahrscheinlichkeit $P(A)$ für das Ereignis A ergibt sich die Wahrscheinlichkeit für das komplementäre Ereignis $\overline{A}$ als

$$P(\overline{A}) = 1 - P(A). \tag{1.5}$$

Multiplikation von Wahrscheinlichkeiten. Auf der Basis der Definition der bedingten Wahrscheinlichkeit erhält man durch Umstellung der Gleichung den **Multiplikationssatz**, der angibt mit welcher Wahrscheinlichkeit die Ereignisse A und B gemeinsam auftreten:

$$P(A \cap B) = P(A|B) \cdot P(B) \tag{1.6}$$

Im Fall der stochastischen Unabhängigkeit von A und B vereinfacht sich der Multiplikationsatz.

$$P(A \cap B) = P(A) \cdot P(B) \tag{1.7}$$

Satz von der totalen Wahrscheinlichkeit. Seien $A_1, A_2, \ldots, A_n$ n zufällige Ereignisse, die einander paarweise ausschließen (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$ mit $i, j = 1, 2, \ldots, n$) und deren Vereinigung das sichere Ereignis ergibt ($A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$). Weiter sei $P(A_i) > 0$ für alle $i = 1, 2, \ldots, n$. Dann berechnet sich die Wahrscheinlichkeit für ein beliebiges Ereignis B über die folgende Beziehung:

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i) \tag{1.8}$$

Beispiel: 1.2.5

gegeben: Wahrscheinlichkeiten für das Auftreten von Krankheit K in Altersgruppen A_1, A_2, A_3 (z.B. ≤ 18 Jahre, $19 - 59$ Jahre, ≥ 60 Jahre):

$$P(K|A_1) = 0,05;\ P(K|A_2) = 0,2;\ P(K|A_3) = 0,5$$

außerdem seien die Altersanteile in der Bevölkerung bekannt:

$$P(A_1) = 0,1\ ;\ P(A_2) = 0,7\ ;\ P(A_3) = 0,2$$

gesucht: Wahrscheinlichkeit für das Auftreten der Krankheit $P(K)$ unabhängig vom Alter; d.h. die Wahrscheinlichkeit, dass eine zufällig aus der Gesamtbevölkerung ausgewählte Person auch erkrankt ist.

Lösung:

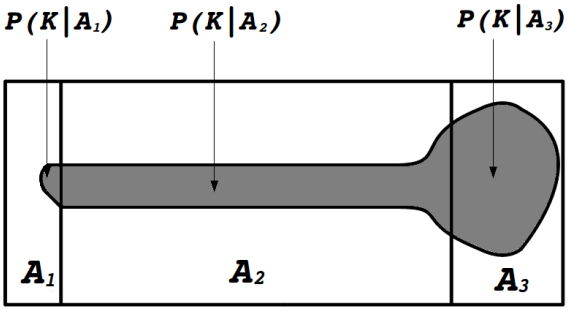


Abbildung 1.2: Auftreten von Krankheit K in Altersgruppen A1,A2,A3

$$\begin{aligned} P(K) &= P(K|A_1) \cdot P(A_1) + P(K|A_2) \cdot P(A_2) + P(K|A_3) \cdot P(A_3) \\ &= 0,05 \cdot 0,1 + 0,2 \cdot 0,7 + 0,5 \cdot 0,2 \\ &= 0,005 + 0,14 + 0,1 \\ &= 0,245 \end{aligned}$$

Satz von Bayes (Bayessche Formel). Unter den gleichen Voraussetzungen wie beim Satz der totalen Wahrscheinlichkeit, gilt für die **bedingte Wahrscheinlichkeit** des Ereignisses A unter der Bedingung B:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \cdot P(A_j)} \tag{1.9}$$

Diese Beziehung wird als *Bayessche Formel* bezeichnet. Für den speziellen Fall zweier Ereignisse A und B vereinfacht sich dieser Ausdruck zu:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})} \tag{1.10}$$

Die Bayessche Formel geht zurück auf den englischen Geistlichen *Thomas Bayes* (1702-1761) und erlaubt die Berechnung der bedingten Wahrscheinlichkeit $P(A|B)$ aus den komplementären bedingten Wahr-

scheinlichkeiten $P(B|A)$ und $P(B|\overline{A})$. Die Formel von Bayes findet u.a. Anwendung bei der Berechnung der Vorhersagewerte eines diagnostischen Tests aus den Güteparametern **Sensitivität** und **Spezifität**.

Beispiel: 1.2.6

Diagnostischer Test

Zu einem diagnostischen Test ist bekannt, dass Sensitivität bei 85% und Spezifität bei 75% liegen.
Problem:
Es könnte sein, dass Sie sich angesteckt haben. Der diagnostischen Test zeigt einen negativen bzw. positiven Befund. Sie erfahren, dass im Mittel 10 von 1000 Menschen erkrankt sind (d.h. Prävalenz = 1%).

Naheliegende Fragen:
(a) Wie wahrscheinlich ist es bei **positivem** Testergebnis **infiziert** zu sein?
(b) Wie wahrscheinlich ist es bei **negativem** Testergebnis **nicht infiziert** zu sein?
(c) Wie wahrscheinlich ist es bei **positivem** Testergebnis **nicht infiziert** zu sein?
(d) Wie wahrscheinlich ist es bei **negativem** Testergebnis **infiziert** zu sein?

Bezeichungen:
$T+$:= Test positiv $T-$:= Test negativ
$I+$:= infiziert $I-$:= nicht infiziert

gegeben:
Sensitivität: $P(T+|I+) = 0,85$; Spezifität: $P(T-|I-) = 0,75$;
Prävalenz: $P(I+) = \dfrac{10}{1000} = 0,01$

gesucht:
(a) **positiver prädiktiver Wert (PPW)**: $P(I+|T+)$
Lösung:
zu (a):
Nach dem Satz von Bayes erhält man:

$$P(I+|T+) = \frac{P(T+|I+)\cdot P(I+)}{P(T+)} = \frac{P(T+|I+)\cdot P(I+)}{P(T+|I+)\cdot P(I+) + P(T+|I-)\cdot P(I-)}$$

Zwischenrechnung:

$$\begin{aligned} P(T+|I-) &= 1 - P(T-|I-) = 0,25 \\ P(I-) &= 1 - P(I+) = 0,99 \end{aligned}$$

$$\begin{aligned} P(I+|T+) &= \frac{0,85\cdot 0,01}{0,85\cdot 0,01 + 0,25\cdot 0,99} = \frac{0,0085}{0,256} \\ &\approx 0,033 = 3,3\% \end{aligned}$$

Bemerkung: Der **negative prädiktive Wert (NPW)**, d.h. $P(I-|T-)$, ist analog zu berechnen. NPW bzw. PPW geben an, mit welcher Wahrscheinlichkeit ein *negativ* getesteter Patient *nicht erkrankt* ist, bzw. mit welcher Wahrscheinlichkeit ein *positiv* getesteter Patient wirklich *erkrankt* ist.
Lösung für:
(b) $P(I-|T-) \approx 99,8\%$ (c) $P(I-|T+) \approx 96,7\%$ (d) $P(I+|T-) \approx 0,2\%$

1.2.5 Wahrscheinlichkeitsbaum

Mit Hilfe von Wahrscheinlichkeitsbäumen kann man hierarchisch strukturierte Ereignisfelder darstellen. Das folgende Beispiel veranschaulicht, wie man die Wahrscheinlichkeit einer bestimmten Ereigniskonstellation mit Hilfe eines Wahrscheinlichkeitsbaumes bestimmen kann.

Beispiel: 1.2.7

Verhältnis von Knaben- und Mädchengeburten für Familien mit 3 Kindern.
Beachte: Bei der Geburt der Kinder handelt es sich um unabhängige Ereignisse; d.h. bedingte Wahrscheinlichkeiten sind gleich den unbedingten Wahrscheinlichkeiten.

gegeben:
Wahrscheinlichkeit für Knaben- bzw. Mädchengeburten: $P(K) = 0,51$; $P(M) = 0,49$

gesucht:
Wahrscheinlichkeiten für alle mögl. Kombinationen, d.h.: $P(KKK)$, $P(KKM)$, $P(KMK)$, $P(KMM)$, $P(MKK)$, $P(MKM)$, $P(MMK)$, $P(MMM)$
(*Beachte:* wir schreiben kurz $P(KKK)$ anstatt $P(K \cap K \cap K)$, etc.)

Lösung:

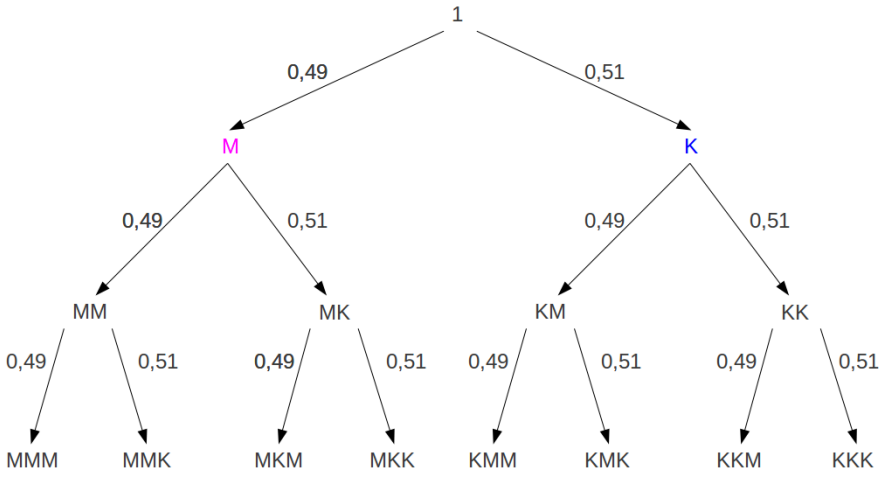


Abbildung 1.3: Wahrscheinlichkeitsbaum

Multiplikation der Wahrscheinlichkeit einer bestimmten Ereigniskonstellationen entlang des Pfades:

$$\begin{aligned} P(KKK) &= P(K) \cdot P(K|K) \cdot P(K|KK) = P(K) \cdot P(K) \cdot P(K) \\ &= 0,51 \cdot 0,51 \cdot 0,51 = 0,133 = 13,3\% \\ P(MKK) &= P(M) \cdot P(K|M) \cdot P(K|MK) = P(M) \cdot P(K) \cdot P(K) \\ &= 0,49 \cdot 0,51 \cdot 0,51 = 0,127 = 12,7\% \end{aligned}$$

Die weiteren Kombinationen können analog berechnet werden.

1.3 Wahrscheinlichkeitsverteilungen

1.3.1 Zufallsvariable

Eine *Zufallsvariable (ZV)* ordnet den Ergebnissen eines Zufallsexperiments eine (reelle) Zahl zu. D.h. eine Zufallsvariable ist eine theoretische Größe (eine Funktion), die bestimmte Zahlenwerte X_i (Ausprägun-

gen) in Abhängigkeit vom Zufall (d.h. gemäß bestimmter Verteilungen, s. u.) annimmt. Ein konkret beobachteter Versuchsausgang x_i wird als *Realisierung* der Zufallsvariable X bezeichnet.

Beispiel: 1.3.1

Merkmale und Ausprägungen:

Merkmal	ZV	Ausprägungen	Werte der ZV
Geschlecht	X	männlich; weiblich	z.B. X=1; X=2
Heilungsverlauf	Y	besser;gleich; schlechter	z.B. Y=1; Y=0; Y=-1
Gewicht	G	Gewicht in kg	Zahlenwert der Gewichtsmessungen

Tabelle 1.2: Merkmale und Ausprägungen

1.3.2 Wahrscheinlichkeitsverteilung diskreter Zufallsvariablen

Bei Zufallsexperimenten mit abzählbar vielen Versuchsausgänge $X_1, X_2, ..., X_n$ werden die entsprechenden Zufallsvariablen X als *diskrete Zufallsvariablen* bezeichnet. Für diesen Fall seien die Auftretenswahrscheinlichkeiten der Versuchsausgänge $X_1, X_2, ..., X_n$ gegeben durch: $p_1, p_2, ..., p_n$. Die Zuordnung der Wahrscheinlichkeiten zu den Versuchsausgängen nennt man *Wahrscheinlichkeitsfunktion* der diskreten Zufallsvariablen X; man schreibt:

$$P(X = x_j) = p_j \ , \ j = 1, 2, ..., n \tag{1.11}$$

wobei die Summe der Einzelwahrscheinlichkeiten aller möglichen Versuchausgänge auf 1 normiert ist, d.h.:

$$\sum_{j=1}^{n} p_j = 1 \tag{1.12}$$

Grafisch kann man die Wahrscheinlichkeitsfunktion in einem Stabdiagramm darstellen (siehe Beispiel 1.14).

Ist die Wahrscheinlichkeitsfunktion der diskreten Zufallsvariable X bekannt, so lässt sich die Wahrscheinlichkeit dafür berechnen, dass die Zufallsvariable X einen Wert kleiner oder gleich x annimmt:

$$F(x) = P(X \leq x) = \begin{cases} 0 \text{ falls } x < \ x_1 \\ \sum_{i=1}^{j} \ p_i \text{ falls } \ x_j \leq x \leq \ x_{j+1} \\ 1 \text{ falls } x \geq \ x_n \end{cases} \tag{1.13}$$

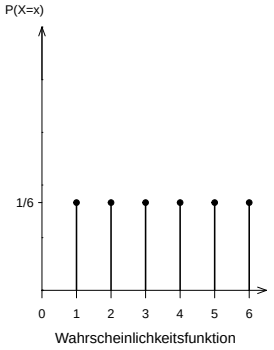
$F(x)$ wird als **Verteilungsfunktion** der diskreten Zufallsvariable X bezeichnet. Die grafische Darstellung der Verteilungsfunktion einer diskreten Zufallsvariablen ist eine Treppenkurve mit Sprüngen der Höhe p_j an den Stellen x_j (siehe Beispiel 1.14).

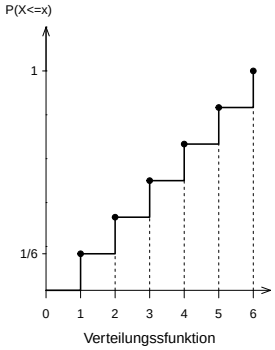
Beispiel: 1.3.2

Wahrscheinlichkeitsverteilung beim Werfen eines sechsseitigen Würfels:

x	$P(X = x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Tabelle 1.3: Wurfwahrscheinlichkeit





$$f(x) = P(X = x_i) = p_i \qquad F(x) = P(X \leq x)$$
$$\sum_{i=1}^{n} p_i = 1 \qquad = \sum_{k:\, x_k \leq x} p_k$$

Abbildung 1.4: Wahrscheinlichkeitsfunktion vs. Verteilungsfunktion

Erwartungswert und Varianz diskreter Verteilungen. Die Verteilung vieler Zufallsvariablen kann durch **Erwartungswert** $E(X)$ und **Varianz** $Var(X)$ charakterisiert werden.
Der Erwartungswert $E(X)$ ist das mit den Werten der Wahrscheinlichkeitsfunktion (Einzelwahrscheinlichkeiten) gewichtete Mittel der möglichen Realisierungen der Zufallsgröße:

$$E(X) = \sum_{j=1}^{n} x_j \cdot p_j \tag{1.14}$$

Der Erwartungswert ist ein häufig verwendetes Maß für die mittlere Lage der Wahrscheinlichkeitsverteilung.
Die Varianz $Var(X)$ ist das gewichtete Mittel der quadratischen Abweichungen vom Erwartungswert:

$$Var(X) = E([X - E(X)]^2) = \sum_{i=1}^{n} [\, X_i - E(X)]^2 \cdot p_i \tag{1.15}$$

Die Quadratwurzel aus der Varianz bezeichnet man als *Standardabweichung* SD[3]:

$$SD(X) = \sqrt{Var(X)} \tag{1.16}$$

[3] Abkürzung aus dem Englischen: *standard deviation*

Beispiel: 1.3.3

Erwartungswert (E), Varianz (Var) und Standardabweichung (SD) für vier Ereignisse mit unterschiedlichen Auftretenswahrscheinlichkeiten:

gegeben:

Ereignis X:	1	2	3	4
P(X=x):	0,2	0,3	0,4	0,1

gesucht:

$E(\ X_{1,2,3,4})$
$Var(\ X_{1,2,3,4})$
$SD(\ X_{1,2,3,4})$

Lösung:

$$\begin{aligned}
E(X_{1,2,3,4}) &= 0,2 \cdot 1 + 0,3 \cdot 2 + 0,4 \cdot 3 + 0,1 \cdot 4 \\
&= 0,2 + 0,6 + 1,2 + 0,4 \\
&= 2,4 \\
Var(X_{1,2,3,4}) &= 0,2 \cdot (1-2,4)^2 + 0,3 \cdot (2-2,4)^2 + 0,4 \cdot (3-2,4)^2 + 0,1 \cdot (4-2,4)^2 \\
&= 0,2 \cdot 1,96 + 0,3 \cdot 0,16 + 0,4 \cdot 0,36 + 0,1 \cdot 2,56 \\
&= 0,392 + 0,048 + 0,144 + 0,256 = 0,84 \\
SD(\ X_{1,2,3,4}) &= \sqrt{Var(\ X_{1,2,3,4})} = \sqrt{0,84} = 0,917
\end{aligned}$$

Binomialverteilung. Es gibt eine Vielzahl von Verteilungen diskreter Zufallsvariablen. Beispielhaft wird im Folgenden die Binomialverteilung erläutert, die es gestattet, binäre Zufallsvariablen zu betrachten.
Für eine Zufallsvariable X wird angenommen, dass sie nur zwei Werte annehmen kann, z.B. Erfolg ($X = 1$) und Nichterfolg ($X = 0$). Ist die Wahrscheinlichkeit für das Auftreten des Erfolges p, so tritt mit der Wahrscheinlichkeit $q = 1-p$ kein Erfolg auf. Eine derartige Zufallsvariable wird als Bernoulli-Variable bezeichnet; derartige Experimente als Bernoulli-Experimente mit Trefferwahrscheinlichkeit $p = P(X = 1)$. Man sagt, X ist $B(1, p)$ verteilt.
Wird das Bernoulli-Experiment n mal wiederholt, wobei die Einzelexperimente unabhängig voneinander durchgeführt werden, kann jedes mögliche Ergebnis des Gesamtexperimentes durch eine n-stellige Folge von Einsen (Erfolg) und Nullen (Nichterfolg) beschrieben werden. Es gibt 2^n verschiedene Versuchsausgänge. Man kann sie in Gruppen mit gleicher Erfolgszahl k (k Einsen, $n-k$ Nullen) zusammenfassen. Diese Gruppen haben je nach Erfolgszahl k verschiedenen Umfang.
Für $k = 0$ (kein einziger Erfolg) gibt es nur eine einzige Möglichkeit:

$$000\ldots0$$

Das gleiche gilt für $k = n$ (kein einziger Nichterfolg):

$$111\ldots1$$

Dagegen gibt es für $k = 1$ sowie für $k = n-1$ je n verschiedene Ausgänge:

$$\begin{array}{lll}
100\cdots0 & \text{bzw.} & 011\cdots1 \\
010\cdots0 & & 101\cdots1 \\
001\cdots0 & & 110\cdots1 \\
\vdots & & \vdots \\
000\cdots1 & & 111\cdots0
\end{array}$$

Im allgemeinen Fall liefert die Kombinatorik als Anzahl der Versuchsausgänge (k Erfolge bei n Wieder-

holungen) den Binomialkoeffizienten[4;5]

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-k+1)}{1 \cdot 2 \cdot 3 \cdot ... \cdot k}. \tag{1.17}$$

Per Definition gilt:

$$\binom{n}{0} = \binom{n}{n} = 1. \tag{1.18}$$

Alle Versuchsausgänge mit gleicher Erfolgszahl haben die gleiche Wahrscheinlichkeit. Wegen der Unabhängigkeit der Bernoulli-Experimente gilt der Multiplikationssatz. Damit ist die Wahrscheinlichkeit eines einzelnen Versuchsausgangs

$$p^k \cdot (1-p)^{n-k}. \tag{1.19}$$

Alle Versuchsausgänge sind zueinander disjunkt. Damit ist der Additionssatz (Formel 1.3) der Wahrscheinlichkeitsrechnung anwendbar. Fasst man alle Versuchsausgänge mit k Treffern zum Ereignis „k Erfolg bei n Wiederholungen“ zusammen, erhält man für dieses die Wahrscheinlichkeit

$$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}. \tag{1.20}$$

Es ist zweckmäßig, die Erfolgszahl als neue Zufallsgröße Y einzuführen. Dann ergibt sich für jedes $k = 0, 1, \ldots, n$

$$P(Y = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}. \tag{1.21}$$

Die diskrete Zufallsvariable Y ist binomialverteilt. Schreibweise: $Y \sim B(n, p)$.

Beispiel: 1.3.4

Auftreten von k Komplikationen bei Behandlung von $n = 3$ Patienten:

Binomialverteilung:
$n = 3$ Patienten, $p = P(\text{Komplikation}) = 0.1$, P(keine Komplikation) $= 1 - p = 0.9$

mögliche Anzahl k der Komplikationen	mögliche Realisierungen dieser Anzahl	$P(Y = k)$ Wahrscheinlichkeit für das Auftreten von genau k Komplikationen	Wahrscheinlichkeit $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$
0	000	$P(Y = 0)$	$1 \cdot p^0 \cdot (1-p)^3 = 0,729$
1	100 010 001	$P(Y = 1)$	$3 \cdot p^1 \cdot (1-p)^2 = 0,243$
2	110 101 011	$P(Y = 2)$	$3 \cdot p^2 \cdot (1-p)^1 = 0,027$
3	111	$P(Y = 3)$	$1 \cdot p^3 \cdot (1-p)^0 = 0,001$

Tabelle 1.4: Auftreten von k Komplikationen

Eine $B(n, p)$ verteilte Zufallsgröße Y hat:

- den Erwartungswert: $E(Y) = n \cdot p$
- die Varianz: $Var(Y) = n \cdot p \cdot (1-p)$

[4] $\binom{n}{k}$ - sprich „n über k“
[5] n! - sprich „n Fakultät“ z.B. $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$; per Definition gilt: $0! := 1$

- die Standardabweichung: $\sqrt{Var(Y)} = \sqrt{n \cdot p \cdot (1-p)}$

Beispiel: 1.3.5

Die Abbildung zeigt Wahrscheinlichkeitsfunktion und Verteilungsfunktion für die Anzahl von k Rauchern unter $n = 4$ zufällig ausgewählten Infarktpatienten unter der Annahme, dass 60 Prozent aller Infarktpatienten rauchen ($p = 0,6$).

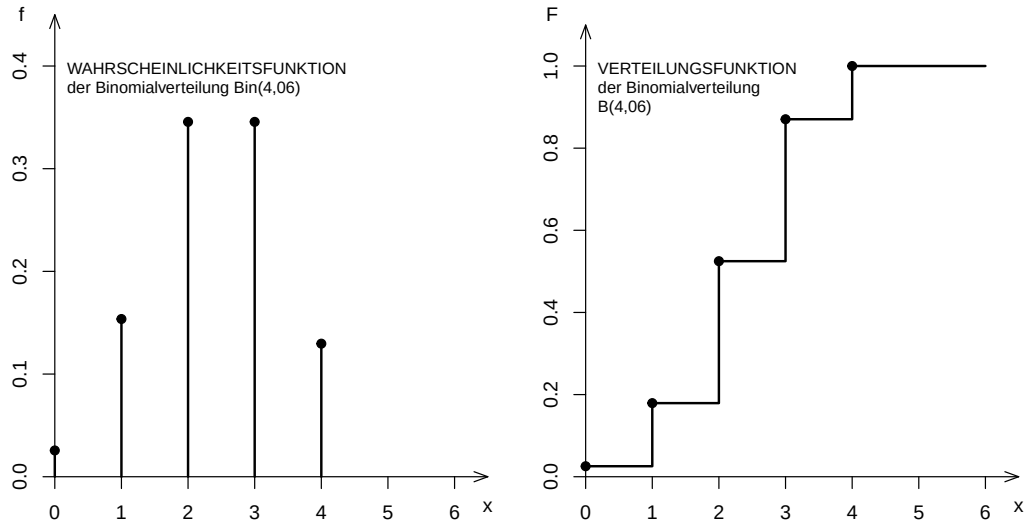


Abbildung 1.5: Binomiale -Wahrscheinlichkeits- und Verteilungsfunktion

Zwar sind die beiden maximalen Werte der Wahrscheinlichkeitsfunktion gleich, aber das Stabdiagramm ist trotzdem asymmetrisch, weil die Raucher überwiegen.

Erwartungswert	$E(X) = 4 \cdot 0,6 = 2,4$
Varianz	$Var(X) = 4 \cdot 0,6 \cdot 0,4 = 0,96$
Standardabweichung	$\sqrt{Var(X)} = \sqrt{0,96} \approx 0,98$

1.3.3 Wahrscheinlichkeitsverteilung stetiger Zufallsvariablen

Stetige Zufallsvariablen beschreiben Merkmale mit einem kontinuierlichen Wertebereich quantitativ, insbesondere Messgrößen. Sie können in einem bestimmten Intervall der reellen Zahlen jeden beliebigen Wert annehmen. Das heißt aber auch, dass stetige Zufallsvariablen unendlich viele Ausprägungen annehmen können.

Beispiel: 1.3.6

Betrachten Sie beispielsweise das Geburtsgewicht einer Kohorte von reifgeborenen männlichen Neugeborenen, so treten um 3200g sehr viele Meßpunkte relativ dicht gedrängt auf. Im Gegensatz dazu wird unterhalb von 2600g und oberhalb von 3800g die Dichte der Meßwerte stark abnehmen.

Da die Gesamtwahrscheinlichkeit in diesem Fall auf unendlich viele mögliche Werte (beliebige Messgenauigkeit vorausgesetzt) verteilt werden muss, ist die Wahrscheinlichkeit für das Auftreten eines ganz be-

stimmten Geburtsgewichtes gleich Null. Aus diesem Grund wird bei stetigen Zufallsvariablen die **Wahrscheinlichkeitsdichte** $f(x)$ benutzt. Die Wahrscheinlichkeitsdichte entspricht damit der Wahrscheinlichkeitsfunktion diskreter Zufallsvariablen, bei der die Gesamtwahrscheinlichkeit nur auf eine abzählbare Anzahl von Ereignissen verteilt wird. Für stetige Zufallsvariablen gilt, dass jeder einzelne Wert die Wahrscheinlichkeit Null besitzt.
Für die Betrachtung stetiger Zufallsvariablen wird dem unendlichen Intervall $x \leq X \leq x + dx$ die Wahrscheinlichkeit $f(x) \cdot dx$ zugeordnet, die angibt, mit welcher Wahrscheinlichkeit eine Realisierung von X in dieses Intervall fällt. Für endliche Intervalle $x_1 \leq X \leq x_2$ wird zur Berechnung der Wahrscheinlichkeit über die Wahrscheinlichkeitsdichte integriert, d.h. es gilt:

$$P(\, x_1 \leq X \leq \, x_2) = \int_{x_1}^{x_2} f(x) \cdot dx \tag{1.22}$$

In der grafischen Darstellung der Wahrscheinlichkeitsdichte $f(x)$ über x (vgl. Abb. 1.6) entspricht die Fläche zwischen Kurve und x-Achse im Intervall von x_1 bis x_2 der Wahrscheinlichkeit, mit der die Zufallsvariable X Werte innerhalb dieses Intervalls annimmt. In dieser Interpretation entspricht der Flächeninhalt unter der Gesamtkurve der Wahrscheinlichkeit des sicheren Ereignisses und hat daher den Wert 1.
Ordnet man jedem Wert x auf der Abszisse die Wahrscheinlichkeit zu, mit der die Zufallsvariable X Werte bis zu der Zahl x annimmt, erhält man wie bei der diskreten Zufallsvariable die Verteilungsfunktion $F(x)$. Für jedes x ist $F(x)$ somit die Fläche unter der Dichtefunktion $f(x)$ von $-\infty$ bis zur Stelle x:

$$F(X) = P(X \leq x_1) = \int_{-\infty}^{x_1} f(x) \cdot dx. \tag{1.23}$$

Die grafische Darstellung der Verteilungskurve ergibt eine stetige, monoton steigende Kurve (vgl. Abb. 1.6).

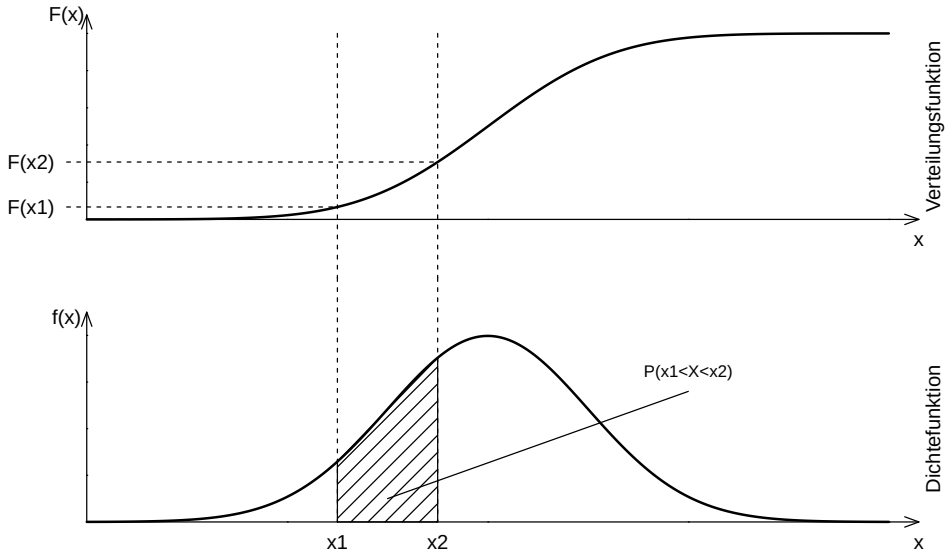


Abbildung 1.6: Zusammenhang zwischen Verteilungsfunktion und Dichtefunktion stetiger Zufallsvariablen

Die Wahrscheinlichkeit dafür, dass X in das Intervall von x_1 bis x_2 fällt, lässt sich aus der Verteilungsfunktion berechnen:

$$P(\, x_1 \leq X \leq \, x_2) = F(\, x_2) - F(\, x_1). \tag{1.24}$$

Erwartungswert und Varianz stetiger Zufallsvariable. Erwartungswert $E(X)$ und Varianz $Var(X)$ errechnen sich in Analogie zu den Aussagen für diskrete Zufallsvariablen:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx \tag{1.25}$$

$$Var(X) = \int_{-\infty}^{+\infty} [x - E(x)]^2 \cdot f(x)dx \tag{1.26}$$

Die Standardabweichung ergibt sich ebenso aus der Varianz:

$$SD(X) = \sqrt{Var(X)} \tag{1.27}$$

Normalverteilung. Die Normalverteilung ist eine der wichtigsten Verteilungen stetiger Zufallsvariablen. Sie lässt sich in vielen Fällen anwenden, um die Verteilung von Messwerten um einen Mittelwert zu beschreiben. Die Normalverteilung wurde erstmals im Zusammenhang mit der Analyse von astronomischen Messungen durch *Carl Friedrich Gauß* (1777-1855) im Jahre 1809 beschrieben.
Eine Normalverteilung tritt z.B. auf, wenn die Variabilität der Zufallsgrößen dadurch bedingt ist, dass sehr viele voneinander unabhängige Einflüsse additiv aufeinander einwirken. Das heißt, die Summe vieler unabhängiger, beliebig verteilter Zufallsvariablen folgt einer Normalverteilung. Diese Eigenschaft wird auch als *Zentraler Grenzwertsatz* bezeichnet.
Der Kurvenverlauf der Dichtefunktion $f(x)$ der Normalverteilung besitzt folgende Eigenschaften:

- glockenförmig
- symmetrisch um den Erwartungswert
- Wendepunkte im Abstand der Standardabweichung vom Erwartungswert
- nähert sich nach $-\infty$ und $+\infty$ asymptotisch der x-Achse

Die Dichtefunktion einer normalverteilten Zufallsgröße wird eindeutig durch zwei Parameter charakterisiert: Erwartungswert μ und Varianz σ^2. Die Dichtefunktion $f(x)$ ist definiert als

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.28}$$

Die Zufallsgröße X heißt dann $\mathcal{N}(\mu, \sigma^2)$-normalverteilt. Die zugehörige Verteilungsfunktion $F(x)$ lautet:

$$F(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \int_{-\infty}^{x'} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \, dx' \tag{1.29}$$

Die Parameter der Normalverteilungskurve haben anschauliche Bedeutungen. Der Erwartungswert μ legt die Lage des Maximums der zu $f(x)$ gehörenden Glockenkurve fest, die Standardabweichung σ deren Breite. Diese Eigenschaft wird in der Abbildung 1.7 veranschaulicht, in der sowohl Erwartungswert ($\mu = 5; \mu = 0$) als auch Standardabweichung ($\sigma = 1,5; \sigma = 1$) variiert werden.

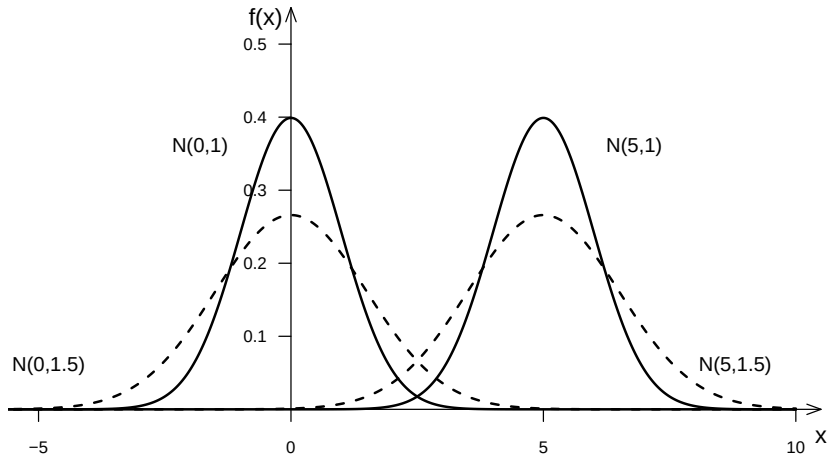


Abbildung 1.7: Variation des Erwartungswertes und der Standardabweichung

Die Intervalle, deren Intervallgrenzen symmetrisch zu μ liegen und um ganzzahlige Vielfache von σ von μ entfernt sind ($1\sigma-, 2\sigma-$, und $3\sigma - Intervall$), sind von besonderer Bedeutung (vgl. Abb. 1.8). Es lassen sich damit Standardbereiche bestimmen, die angeben, wieviel Prozent der Werte in bestimmten Intervallen zu erwarten sind (siehe Abb. 1.8):

68,3% der Messwerte liegen zwischen $\mu - \sigma$ und $\mu + \sigma$

95,5% der Messwerte liegen zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$

99,7% der Messwerte liegen zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$

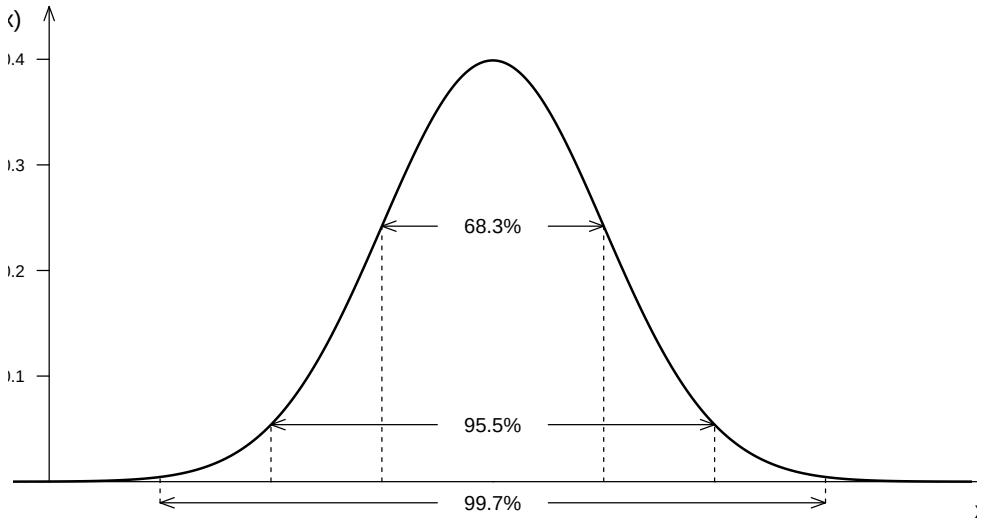


Abbildung 1.8: Intervalgrenzen

Beispiel: 1.3.7

Labordiagnostik:

- Für normalverteilte Messgrößen werden oftmals $\pm 2\sigma-$Intervalle als Normbereiche bezeichnet.
- Diese Intervalle beziehen sich dabei auf die Verteilung einer geeignet definierten, gesunden Bevölkerungsstichprobe.
- Deutet man die Flächeninhalte unterhalb der Dichtefunktion als Wahrscheinlichkeit folgt daraus, dass die Wahrscheinlichkeit für das Auftreten von Messwerten im 2σ-Bereich 95,4% beträgt. Folglich liegt bei 4,6% der Gesunden der Wert außerhalb des Normbereiches.

Standardnormalverteilung. Subtrahiert man von einer $\mathcal{N}(\mu, \sigma^2)$-verteilten Zufallsgröße X den Erwartungswert μ (Verschiebung) und teilt das Ergebnis durch die Wurzel aus der Varianz σ^2 (Normierung), so resultiert eine neue Zufallsgröße Z, die sogenannte Standardisierung von X.

$$Z = \frac{X - \mu}{\sigma} \tag{1.30}$$

Diese ist dadurch charakterisiert, dass sie den Erwartungswert $E(Z) = 0$ und die Varianz $Var(Z) = 1$ hat.

Ihre Wahrscheinlichkeitsdichte ist

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{1.31}$$

und ihre Verteilungsfunktion

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dy. \tag{1.32}$$

Die Verteilung von Z ist die Standardnormalverteilung; man schreibt $Z \sim N(0, 1)$.

Abgeleitete Verteilungen. Normalverteilungen kommen in der Praxis häufig, aber bei Weitem nicht immer vor. Andere Zufallsgrößen sind beispielweise log-normalverteilt, d.h. logarithmiert man die Messwerte, so entsteht eine Normalverteilung (vgl. Beispiel 1.21). Dies gilt z.B. für Zellpopulationszahlen wie Leukozytenwerte.

Beispiel: 1.3.8

Logarithmische Normalverteilung:
Bei 5000 Patienten mit fortgeschrittenem Morbus Hodgkin wurde die Leukozytenzahl im Blut gemessen. Die folgende Abbildung 1.9 zeigt ein Histogramm der Leukozyten. In der linken Abbildung sind die Leukozytenwerte etwas rechtsschief (Normalverteilungskurve kommt nur unvollständig zur Anpassung).

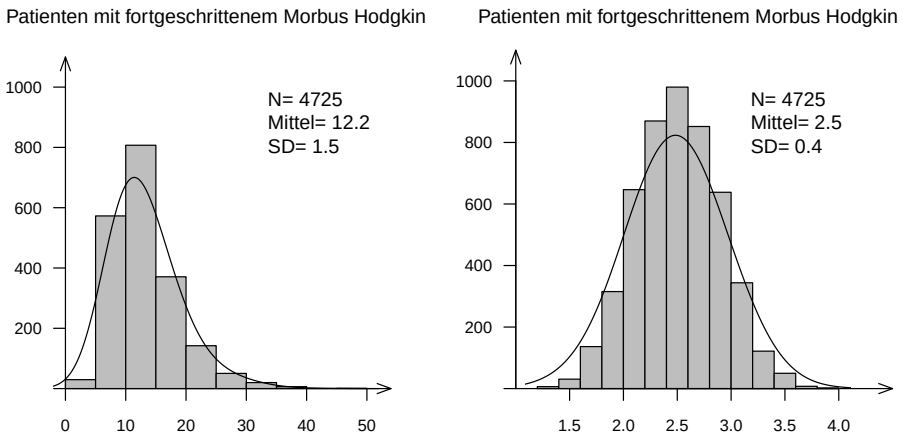


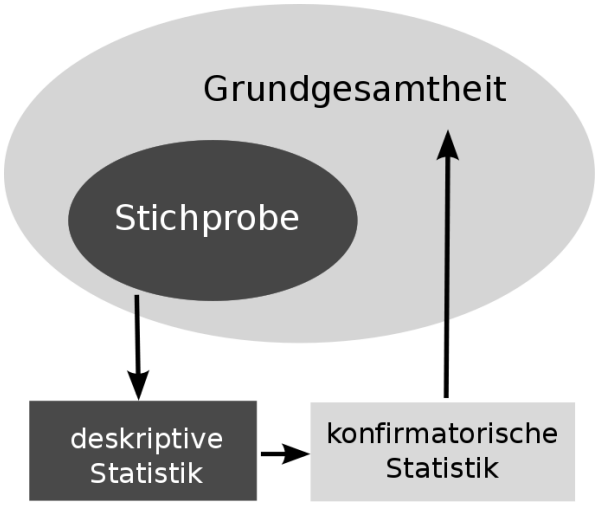
Abbildung 1.9: Logarithmische Normalverteilung

- eine Transformation der Messwerte erleichtert es, die zu charakterisierenden Verteilungsformen darzustellen
- das Logarithmieren der Leukozytenwerte bewirkt die Transformation zu einer annährend symmetrischen Verteilung

1.4 Zusammenfassung

- Zufallsexperimente sind nur im Rahmen einer statistischen Unsicherheit vorhersagbar
- Ereignisse sind Ergebnisse von Zufallsexperimenten; sie entsprechen damit den individuellen Realisierungen innerhalb einer empirischen Stichprobe
- Wahrscheinlichkeitsverteilungen sind geeignet, empirischen Verreilungen in den wesentlichen Eigenschaften (Lage, Form, Breite) zu approximieren
- die symmetrische glockenförmige Normalverteilung ist eine der wichtigsten Verteilungen stetiger Zufallsvariablen

2 Deskriptive Statistik



Deskriptive Verfahren werden angewendet, um *konkrete Stichproben* aus einer definierten *Grundgesamtheit* zu beschreiben. Dazu werden Tabellen, Diagramme sowie charakteristische Kenngrößen und Maßzahlen verwendet, die im Folgenden vorgestellt werden.

2.1 Grundbegriffe der deskriptiven Statistik

Die deskriptive Statistik beschäftigt sich mit der Aufbereitung, Systematisierung und Beschreibung empirisch gewonnener Daten. Folgende Grundbegriffe sind dabei von fundamentaler Bedeutung:

- **Beobachtungseinheit:** Die Beobachtungseinheit beschreibt die kleinste Einheit, an der bestimmte Beobachtungen gemacht bzw. Daten erhoben werden.
- **Grundgesamtheit:** Die Menge aller möglichen Beobachtungseinheiten nennt man Grundgesamtheit (GGH). Kenngrößen, welche die Grundgesamtheit beschreiben (wie beispielsweise der Mittelwert μ), heißen Parameter und werden häufig mit *griechischen Buchstaben* bezeichnet.

Beispiel: 2.1.1

Beobachtungseinheit	**Grundgesamtheit**
an einem Mundhöhlenkarzinom erkrankter Patient	alle an einem Mundhöhlenkarzinom erkrankten Patienten
Neugeborenes	alle Neugeborenen einer Klinik
Erythrozyt eines Menschen	alle Erythrozyten eines Menschen

- **Stichprobe:** Eine Teilmenge, welche die Grundgesamtheit repräsentiert (z.B. durch Zufallsauswahl aus Grundgesamtheit), nennt man Stichprobe. Sie ist die Menge der tatsächlich untersuchten Beobachtungseinheiten. Kenngrößen, welche die Stichprobe beschreiben, heißen statistische Maßzahlen

und werden meistens mit *lateinischen Buchstaben* bezeichnet, z.B. $\overline{x}$ - arithmetischer Mittelwert der Stichprobe.

- **Merkmal:** Merkmale sind Eigenschaften, die die Beobachtungseinheiten charakterisieren. Wenn die Merkmale im Rahmen statistischer Analysen betrachtet werden, spricht man auch von *Variablen*.
- **Merkmalsausprägung:** Merkmalsausprägungen beschreiben Werte, die von den Merkmalen bzw. Variablen angenommen werden können.

Beispiel: 2.1.2

Beobachtungseinheit	**Merkmal**	**Merkmalsausprägungen**
Patient mit Mundhöhlenkarzinom	Lokalisation	Zunge, Gaumen
Neugeborenes	Gewicht	3670g, 4400g
Erythrozyt	Durchmesser	2 μm, 10 μm

Tabelle 2.1: Patienten mit Mundhöhlenkarzinom

- **Zielgröße:** Die Zielgröße definiert das Merkmal, welches in einer bestimmten Studie untersucht werden soll.
- **Einflussgröße:** Einflussgrößen sind zusätzliche Merkmale, die die Zielgröße beeinflussen können. Dabei unterscheidet man weiter zwischen:
  - **Erklärende Einflussgrößen:** Stehen im Zentrum der Studie und sollen potentielle Veränderungen der Zielgröße erklären.
  - **Begleitmerkmale (Covariable):** Ebenfalls erfasste Merkmale, welche die Zielgröße beeinflussen (können), die aber nicht im Zentrum der Studie stehen.
- **Störgrößen:** Als Störgrößen werden unbekannte oder nicht erfasste Einflussgrößen bezeichnet. Dabei wird weiterhin unterschieden zwischen:
  - **Nicht verzerrende Störgrößen:** Dazu gehört die zufallsbedingte Variabilität der Ergebnisse (z.B. Messfehler), die nicht zu vermeiden ist. Durch geeignete Studienplanung kann der Einfluss nicht verzerrender Störgrößen minimiert werden.
  - **Verzerrende Störgrößen (Confounder):** Als Confounder werden im Allgemeinen unbekannte Hintergrundmerkmale bezeichnet, die die Zielgröße beeinflussen und dadurch zu einer Verzerrung der Ergebnisse führen können.

Beispiel: 2.1.3

Untersuchung des Einflusses der Dosis eines Medikamentes (z.B. Carboanhydrasehemmer) auf den Augeninnendruck bei Patienten mit Winkelblockglaukom anhand der neudiagnostizierten Fälle an einem speziellen Krankenhaus im Zeitraum des nächsten Jahres.

- **Beobachtungseinheit:** Augen
- **Merkmale/Variablen:**
  - **Zielgröße:** Augeninnendruck (Ausprägung: Druck in mmHg)
  - **Erklärende Einflussgröße:** Dosis des Medikamentes
  - **Covariablen:** z.B. Patient, Alter
  - **Störgrößen:** z.B. unbekannte Begleiterkrankungen
- **Grundgesamtheit:** Patienten mit Winkelblockglaukom
- **Stichprobe:** neudiagnostizierten Fälle an einem speziellen Krankenhaus im Zeitraum des nächsten Jahres

2.1.1 Merkmalsskalen

Zur Beschreibung von Merkmalsausprägungen stehen verschiedene Skalen zur Verfügung.

- **Nominalskala:** Merkmalsausprägungen unterscheiden sich nur begrifflich (z.B. Haarfarbe, Geschlecht).
- **Ordinalskala (Rangskala):** Merkmalsausprägungen lassen sich in einer Rangfolge anordnen (z.B. Zensuren, klinische Scores).

  → Nominal und ordinal skalierte Merkmale werden auch beide als **qualitative** oder **kategoriale** Merkmale bezeichnet.
- **Metrische Skala:** Die Merkmalsausprägungen werden durch ein Messverfahren ermittelt (d.h. durch Vergleich mit einer Messskala, entweder *stetig* oder *diskret*). Man unterscheidet weiterhin intervallskalierte Daten, für welche die Angabe einer Differenz, aber nicht die eines Verhältnisses sinnvoll ist (z.B. Temperatur in °C) und verhältnisskalierte Daten, für welche auch Verhältnisse angegeben werden können (z.B. Gewicht).

Beispiel: 2.1.4

Klassifikation	**Merkmal**	**Merkmals-ausprägung**	**Zuordnen von Zahlen**	**Skala**
quantitativ stetig	Gewicht $[kg]$	80, 56, 90	messen	metrisch
quantitativ diskret	Zahl der kariösen Zähne	0,1,2	zählen	metrisch
qualitativ ordinal	Schulnoten	1,2,3,4,5,6	vergeben von Rängen	ordinal
qualitativ nominal	Tumorlokalisation	Niere, Leber	kodieren	nominal

Tabelle 2.2: Skalenklassifikation

Metrisch skalierte, quantitative Merkmale besitzen den höchsten Informationsgehalt. Merkmalsarten können teilweise ineinander umgewandelt werden. So können z.B. metrisch skalierte Merkmale durch Klassenbildung in qualitativ ordinale Merkmale überführt werden.

2.1.2 Datenerfassung, Datenkodierung

Eine gut strukturierte Erfassung von Daten zur weiteren Beschreibung und Auswertung sind ein wesentlicher Schritt zur Deduktion wissenschaftlicher Resultate. Dabei hat es sich in den meisten Fällen als sinnvoll erwiesen, möglichst alle relevanten Daten in *einer* Datei abzulegen. Dabei sollte jeder Beobachtungseinheit eine Zeile zugewiesen werden. Die zugehörigen Merkmale (Variablen) werden dann spaltenweise erfasst.
Die Kodierungen von Daten (Benennung von Merkmalsausprägungen) sollte in Abhängigkeit der konkreten Fragestellung festgelegt werden. Dabei haben sich einige Regeln als sehr nützlich erwiesen:

- möglichst *Rohdaten erfassen*; Zusammenfassung von Daten (z.B. Farbklassen: blau / gelb) können bei Bedarf später realisiert werden
- zur statistischen Auswertung sind *numerische Kodierungen* oft einfacher zu handhaben (z.B. statt alt / jung, besser 1 / 2)
- *Variablenspalten nicht mehrfach belegen*; kann ein Merkmal verschiedene Ausprägungen annehmen ist es u.U. sinnvoll jede Merkmalsausprägung einzeln (d.h. mit Hilfe von „Dummy"-Variablen) als *vorhanden* bzw. *nicht vorhanden* zu kodieren (siehe Beispiel 2.5)

Beispiel: 2.1.5

Datenerfassung/-kodierung:

falsch

PatientenID	Symptom
1	1
2	1,2
3	2
4	1,2
5	1

Tabelle 2.3: falsche Datenerfassung

richtig

PatientenID	Symptom 1	Symptom 2
1	1	0
2	1	1
3	0	1
4	1	1
5	1	0

Tabelle 2.4: richtige Datenerfassung

Die Behandlung und Erfassung *fehlender Werte* erfordert besondere Sorgfalt. Fehlende Werte treten auf, wenn bestimmte Merkmale bei einzelnen Beobachtungseinheiten nicht erfasst werden konnten bzw. „verlorengegangen" sind. Fehlende Werte sind zu kennzeichnen (z.B. „N.A."), um diese bei Auswertungen entsprechend beachten zu können. Die Kodierung von fehlenden Werten mit einer Null (0) ist **ungeeignet** und eine häufige Fehlerquelle.
Treten fehlenden Werte zufällig und in kleiner Zahl auf, sind sie im Allgemeinen unkritisch für eine Analyse der Daten. Treten Sie hingegen in großer Zahl und/oder systematisch auf, dann können sie zu erheblichen Verzerrungen der Resultate führen.

2.1.3 Häufigkeit

Eine Häufigkeitsverteilung beschreibt die Auftretenshäufigkeit einzelner Ausprägungen eines Merkmals innerhalb einer Stichprobe. Dabei wird zwischen absoluter und relativer Häufigkeit unterschieden.

Die **absolute Häufigkeit** gibt an, wie oft ein diskretes Merkmal A innerhalb einer Stichprobe beobachtet wurde. Unter der Annahme, dass das Merkmal k verschiedene Werte $A_1, \cdots, A_k$ annehmen kann, bezeichnet h_a die absolute Auftretenshäufigkeit des Wertes A_i bei n-maliger Beobachtung (Stichprobengröße n, wobei $\sum_{i=1}^{k} n_i = n$).
Im Gegensatz dazu bezeichnet die **relative Häufigkeit** h_r den Anteil der Auftretenshäufigkeit A_i an allen Beobachtungen: $h_r = \frac{h_a}{n}$ mit $(0 \leq h_r \leq 1)$.

Zur grafischen Darstellung von absoluten und relativen Häufigkeiten eignen sich beispielsweise Balken- und Kreisdiagramme.

- **Balkendiagramme** werden im rechtwinkligen Koordinatensystem dargestellt. Dabei wird der Merkmalsausprägung auf der Abszisse eine absolute oder relative Häufigkeiten auf der Ordinate zugeordnet.
- **Kreisdiagramme** eignen sich zur Darstellung relativer Häufigkeiten. Dabei wird die Winkelsumme proportional zu den berechneten relativen Häufigkeiten aufgeteilt (Flächen $\sim$ relative Häufigkeit).

Beispiel: 2.1.6

Register der Tumoren des Kiefer- und Gesichtsbereiches der DDR:
Erfassung von Tumorerkrankungen der Lippe und Mundhöhle des Zeitraumes 1966 - 1980 (2481 Patienten mit Primärerkrankung)

		absolute Häufigkeit	relative Häufigkeit	Gültige Prozente	Kumulierte Prozente
Gültig	oberfl. $\leq$ 5mm	677	0,273	27,4	27,4
	tief $>$ 5mm	1613	0,65	65,2	92,6
	penetrierend	183	0,074	7,4	100,0
	Gesamt	2473	0,997	100,0	
Fehlend	System	8	0,003		
Gesamt		2481	1		

Tabelle 2.5: Register der Tumoren des Kiefer- und Gesichtsbereiches

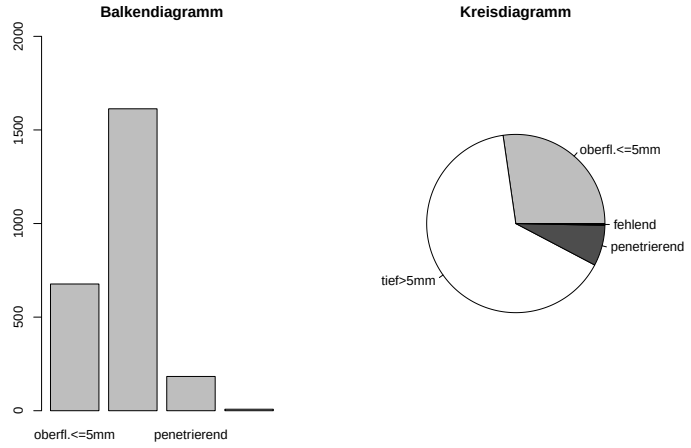


Abbildung 2.1: Balken- und Kreisdiagramm

Klassenbildung. Bei stetigen Merkmalen und bei diskrete Merkmalen mit sehr vielen Ausprägungen ist es häufig sinnvoll, zur grafischen Darstellung eine Klassenbildung anzuwenden. Die Klassen sollten dabei so definiert werden, das sie den Merkmalsraum angemessen repräsentieren. Sehr viele, schmale Klassen werden schnell unübersichtlich und der zugrunde liegende Verteilungstyp ist unter Umständen nur schwer erkennbar. Andererseits birgt die Beschränkung auf eine zu geringe Anzahl sehr breiter Klassen die Gefahr des Informationsverlustes. Zur Wahl der Klassenbreite lassen sich ein paar einfache Faustregeln aufstellen:

- Die Klassenanzahl k sollte etwa der Wurzel des Stichprobenumfangs n entsprechen: $k \approx \sqrt{n}$.
- Weniger als drei Klassen sind nicht sinnvoll.
- Die Klassen sollten möglichst die gleiche Breite haben (Ausnahme: bei „Ausreißern“ am Rand der Häufigkeitsverteilung können größere Klassenbreiten sinnvoll sein).

Histogramm. Histogramme sind eine typische Form der graphischen Darstellung für die Häufigkeitsverteilung metrisch skalierter Merkmale. Dabei werden über den eingeteilten, und auf der Abszisse abgetragenen Klassen angrenzende Rechtecke errichtet, deren Flächeninhalt der (relativen) Klassenhäufigkeit entspricht.

Beispiel: 2.1.7

absolute Häufigkeit verschiedener Altersklassen innerhalb einer Studienpopulation:

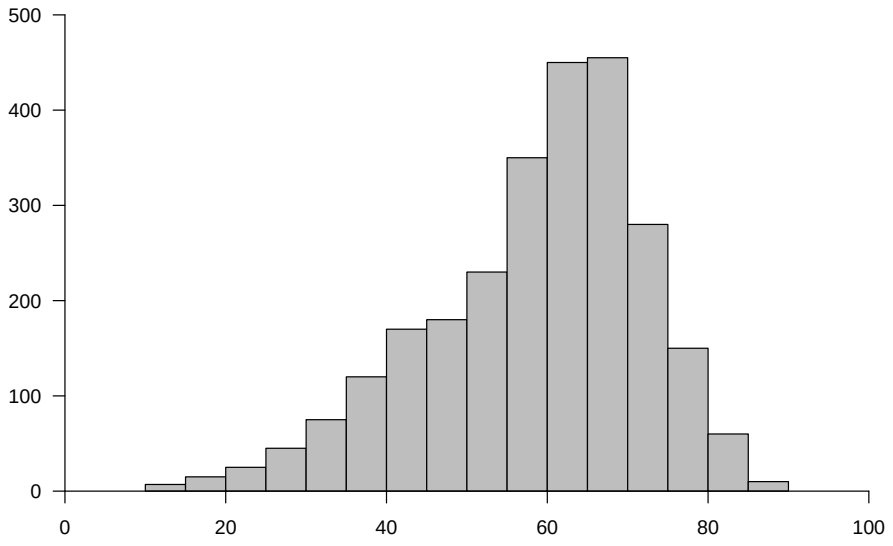


Abbildung 2.2: absolute Häufigkeiten

Histogramme sind geeignet, einen visuellen Eindruck über die Häufigkeitsverteilung einer Merkmalsausprägung zu erhalten. Diese Darstellung gestattet es, Rückschlüsse auf die zugrunde liegende Verteilung des beobachteten Merkmales innerhalb der Stichprobe zu ziehen. Beispielsweise lassen sich Aussagen zur Symmetrie und zur Schiefe einer Verteilung treffen. Typische Verteilungstypen sind in der Abbildung 2.3 dargestellt.

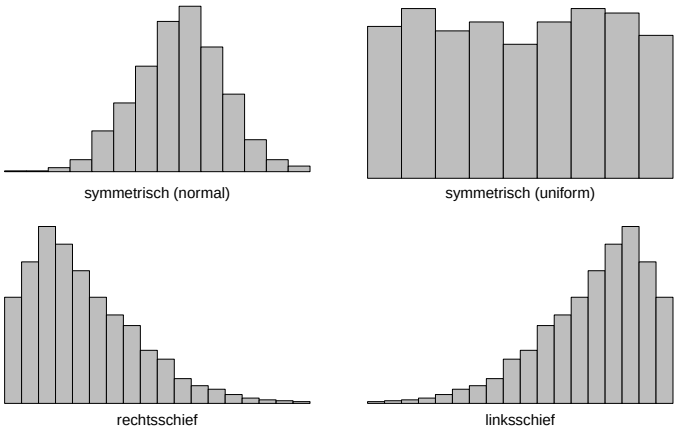


Abbildung 2.3: typische Verteilungstypen

2.2 Kenngrößen/Maßzahlen

Maßzahlen dienen der Charakterisierung von erfassten Daten hinsichtlich bestimmter Eigenschaften. Sie sind geeignet die Verteilungen bzw. deren Charakteristika in kompakter Form zu beschreiben. Bei derartigen Maßzahlen (oder Kenngrößen) unterscheidet man im Wesentlichen drei Arten:

- **Lagemaße:** Charakterisieren in welchen Bereich sich die Stichprobenwerte konzentrieren (im Allgemeinen, Bereich hoher Auftretenshäufigkeit)
- **Streungsmaße (Dispersionsmaße):** Charakterisieren die Variabilität (die Streubreite) der beobachteten Werte
- **Formmaße:** Charakterisieren die Verteilungsform

2.2.1 Lagemaße:

Lagemaße (auch Lokalisationsmaße genannt) dienen der Charakterisierung der erfassten Daten bezüglich der Lage des Zentrums der Verteilung.

Das Arithmetische Mittel $\overline{x}$ ist das bekannteste und am weitesten verbreitete Lagemaß. Häufig wird das Arithmetische Mittel auch als Mittelwert oder Durchschnittswert bezeichnet. Das Arithmetische Mittel wird berechnet als Quotient aus der Summe aller beobachteten Werte und der Anzahl der Werte n:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

Beispiel: 2.2.1

Arithmetisches Mittel:

gegeben: $x = \{1, 1, 2, 3, 4, 4, 4, 6, 6, 8\}$, $n = 10$

gesucht: $\overline{x}$

Lösung: $\overline{x} = \frac{1}{10}(1 + 1 + 2 + 3 + 4 + 4 + 4 + 6 + 6 + 8) = 4$

Der Modalwert. x_{mod} (auch Modus genannt) bezeichnet die Merkmalsausprägung mit der größten Häufigkeit. Bei Daten, welche in Klassen eingeteilt sind, gibt man die *modale Klasse* (Klassen mit der höchsten Besetzungsszahl) an. Es ist zu beachten, das der Modalwert nur bei *eingipfligen* Verteilungen eindeutig bestimmt ist, da er ein *lokales* Dichte-/Häufigkeits- Maximum repräsentiert. Man bezeichnet solche Verteilungen als unimodal. *Zwei-* oder *mehrgipflige* Verteilungen bezeichnet man als bi- bzw. multimodal. Für derartige Verteilungen ist der Modalwert unter Umständen nicht eindeutig definierbar.

Beispiel: 2.2.2

Modalwert:

$x = \{1, 1, 2, 3, 4, 4, 4, 6, 6, 8\} \rightarrow x_{\text{mod}} = 4$

$y = \{1, 2, 2, 2, 2, 3, 3, 4, 5, 6, 6, 7, 7, 7, 7, 8, 9\} \rightarrow y_{\text{mod}} = \{2, 7\}$

Der Median. $\tilde{x} = \tilde{x}_{0,5}$ halbiert eine der Größe nach geordnete Beobachtungsreihe $x_1, ..., x_n$, d.h. 50% der Beobachtungen haben einen Wert größer (oder gleich) dem Median und 50% der Beobachtungen einen Wert kleiner (oder gleich) dem Median. Daher wird der Median über eine Rangfolge der Stichprobenwerte: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ ermittelt.

$$\tilde{x} = \tilde{x}_{0,5} = \begin{cases} x_{\frac{n+1}{2}} & \text{für } n \text{ ungerade} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{für } n \text{ gerade} \end{cases} \tag{2.2}$$

Der Median wird häufig auch als Zentralwert oder 50%-Quantil bezeichnet.

Beispiel: 2.2.3

Median:
Bei Patienten mit Tumoren des Kiefer- und Gesichtsbereiches wird die mittlere Zeit vom Auftreten der ersten Symptome bis zum Arztbesuch in Wochen gemessen.

- Zeit bis Arztbesuch von 10 Patienten:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
12	2	24	2	70	4	1	13	20	24

- ordnen:

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$	$x^{(9)}$	$x^{(10)}$
1	2	2	4	12	13	20	24	24	70

- n gerade: $\tilde{x}_{0.50} = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) = \frac{1}{2}(x_{\frac{10}{2}} + x_{\frac{10}{2}+1}) = \frac{1}{2}(x_5 + x_6)$
- Median (ist Mittelwert aus 5. und 6. Element der geordneten Reihe):

  $\tilde{x}_{0.50} = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(12 + 13) = 12,5$

  (Das arithmetische Mittel ist: $\bar{x} = 17,2$)

Quantile. Quantile sind Werte, die eine Beobachtungsreihe $x_1,\ x_2,\ ,\ldots,\ x_n$ in nicht überlappende Abschnitte teilen. Die Bestimmung der Quantile setzt mindestens das Vorhandensein einer Rangliste $x_1 \leq\ x_2 \leq \cdots \leq\ x_n$ voraus. Dabei ist das $\tilde{x}_\alpha$-Quantil $(0 < \alpha < 1)$ so definiert, dass $\alpha \cdot 100\%$ aller Beobachtungswerte die Bedingung $\leq\ x_\alpha$ erfüllen, daher[1]

$$\tilde{x}_\alpha = \begin{cases} x_i \text{ mit } i = \lceil (n \cdot \alpha) \rceil & \mid\ n \cdot \alpha\ \notin \mathbb{N} \\ \frac{1}{2}(\ x_i +\ x_{i+1}) & \mid\ n \cdot \alpha\ \in \mathbb{N} \end{cases} \tag{2.3}$$

Beispiel: 2.2.4

Quantile:

Register der Tumoren des Kiefer- und Gesichtsbereiches

gesucht: $\tilde{x}_{0.25}$ und $\tilde{x}_{0.5}$ der Zeiten vom Auftreten des 1. Symptoms bis zum Arztbesuch mit:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
12	2	24	2	70	4	1	13	20	24

Lösung: Ordnen der Zahlenwerte.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	2	2	4	12	13	20	24	24	70

Fortsetzung des Beispiels

für $\tilde{x}_{0.25}$:

$\rightarrow\ n \cdot \alpha = 10 \cdot 0,25 = 2,5 (\notin \mathbb{N})$

$\rightarrow\ \tilde{x}_\alpha\ = x_i$ mit $i\ = \lceil 10 \cdot 0,25 \rceil\ = \lceil 2,5 \rceil\ = 3$
(*vgl. Fußnote der Formel* 2.3)

$\rightarrow\ \tilde{x}_{0.25}\ =\ x_3 = 2$

für $\tilde{x}_{0.5}$:

$\rightarrow\ n \cdot \alpha = 0,5 \cdot 10 = 5 = i$

$\rightarrow\ \tilde{x}_\alpha = \frac{1}{2}(\ x_i +\ x_{i+1}) = \frac{1}{2}(x_5 + x_6)$

$\rightarrow\ \tilde{x}_{0.5} = \frac{1}{2}(12 + 13) = 12,5$

Quantile lassen sich ebenso für Verteilungen $f(x)$ von Zufallsvariablen definieren. In diesem Fall liegen links vom α-Quantil $100 \cdot \alpha\%$ der Zufallswerte:

$$\alpha = \int_{-\infty}^{\tilde{x}_\alpha} f(x)dx \tag{2.4}$$

Einige Quantile werden besonders häufig verwendet und tragen spezielle Bezeichnungen (siehe Tabelle 2.6).

2.2.2 Streuungsmaße

Streuungsmaße geben an, wie stark die beobachteten Daten (um einen zentralen Wert) streuen. Sie sind ein Maß für die empirisch beobachtete Heterogenität in einer betreffenden Stichprobe.

Empirische Varianz und Standardabweichung. Als Pendant zum arithmetischen Mittel ist die empirische Varianz s^2 (bzw. die empirische Standardabweichung SD) das am meisten genutzte Streuungsmaß.

[1] $\lceil ... \rceil$ - bezeichnet die obere Gaußklammer - ordnet einer reellen Zahl die nächst größere ganze Zahl zu, z.B.: $\lceil 1,3 \rceil = 2$.

Perzentile	$\tilde{x}_{0.01}, \tilde{x}_{0.02}, \ldots, \tilde{x}_{0.99}$	(synonym: $P_1, P_2, \ldots, P_{99}$)
Dezile	$\tilde{x}_{0.10}, \tilde{x}_{0.20}, \ldots, \tilde{x}_{0.90}$	(synonym: $D_1, D_2, \ldots, D_9$)
Quartile	$\tilde{x}_{0.25}, \tilde{x}_{0.50}, \tilde{x}_{0.75}$	(synonym: Q_1, Q_2, Q_3)
Median	$\tilde{x}_{0.50}$	(synonym: Z)

Tabelle 2.6: häufig verwendete Quantile

Die empirische Varianz s^2 beschreibt die mittlere quadratische Abweichung der Daten von ihrem Mittelwert:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.5}$$

Die Zahl

$$SD = \sqrt{s^2} \tag{2.6}$$

heißt empirische Standardabweichung (SD - standard deviation) des Merkmals A.

Die Spannweite (Range). R ist das am einfachsten zu berechnende Streuungsmaß. Sie beschreibt die Differenz zwischen dem größten und kleinsten Werte innerhalb einer Stichprobe.

$$R = x_{max} - x_{min} = x_n - x_1 \tag{2.7}$$

Der Interquantilabstand. I_{50} ist ein weiteres Streuungsmaß. Er entspricht dem Bereich einer Verteilung, der die mittleren 50%, der Beobachtungen enthält, und wird errechnet aus der Differenz des 3. und 1. Quartils:

$$I_{50} = \tilde{x}_{0,75} - \tilde{x}_{0,25} \tag{2.8}$$

Der Variationskoeffizient. V ist geeignet, Varianzen sinnvoll zu normieren. Häufig treten bei größeren Datenwerten auch (natürlicher Weise) größere Variationen auf. Der Variationskoeffizient berücksichtigt diesen Effekt, indem die Standardabweichung auf das arithmetische Mittel bezogen wird:

$$V = \frac{SD}{\overline{x}} \tag{2.9}$$

2.2.3 Grafische Veranschaulichung von Verteilungscharakteristiken

Zur deskriptiven Beschreibung und für den Vergleich verschiedener Stichproben ist es häufig angezeigt, wichtige Maßzahlen einer empirischen Stichprobe kompakt und übersichtlich darzustellen. Dazu haben sich verschiedene Abbildungstypen etabliert.

Fehlerbalken. Die Darstellung von Fehlerbalken ist geeignet, die Variabilität der Einzelwerte zu dokumentieren. Dabei entspricht die Länge der Fehlerbalken um den Mittelwert $\overline{x}$ der Standardabweichung SD des Einzelwertes.

Beispiel: 2.2.5

Fehlerbalken: Darstellung von arithmetischem Mittel ($\overline{x}$) und Standardabweichung (SD) des Alters getrennt für Patienten mit Lippen- und Mundhöhlenkarzinom.

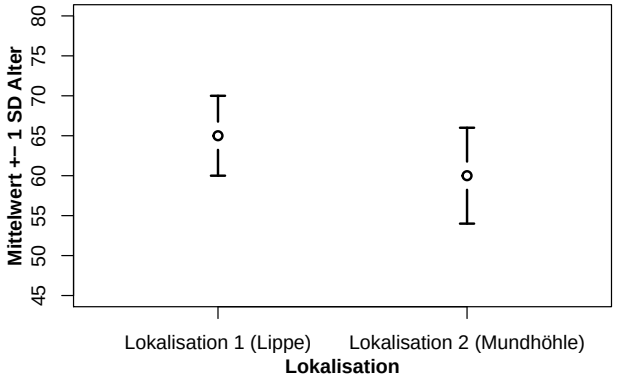


Abbildung 2.4: Fehlerbalken

Boxplots. Boxplots bieten die Möglichkeit, mehrere wichtige Kenngrößen einer Verteilung gleichzeitig darzustellen. In der allgemeinen Form werden sowohl Median, 1. und 3. Quartil sowie Minimum und Maximum der Stichprobe angegeben. Als Ausreißer werden Messwerte bezeichnet, die mehr als das 1,5-fache des Interquartilsabstandes I_{50} vom 1. bzw. 3. Quartil abweichen. Die Definition der Ausreißer ist allerdings nicht eindeutig und wird von verschiedenen Statistikprogrammen unterschiedlich behandelt.

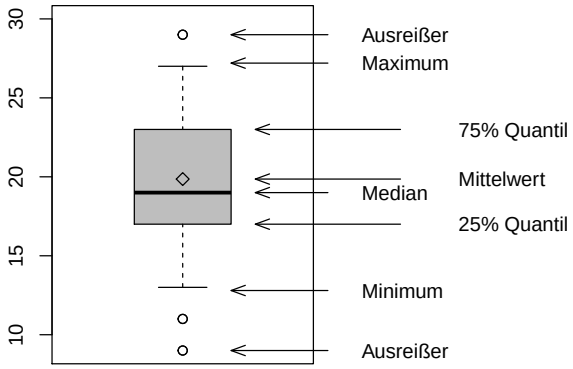


Abbildung 2.5: Boxplot

Die parallele Anordnung von Boxplots für verschiedene Stichproben vermittelt häufig einen ersten Eindruck über Ähnlichkeiten und Unterschiede bestimmter Lagemaße. Dabei ist allerdings zu berücksichtigen, dass bestimmte Verteilungstypen, wie beispielsweise bimodale Verteilungen, von Boxplots nur unzureichend charakterisiert werden können (vgl. Abb.: 2.6). An dieser Stelle sind Histogrammdarstellungen unersetzlich (vgl. Abschnitt 2.1.3).

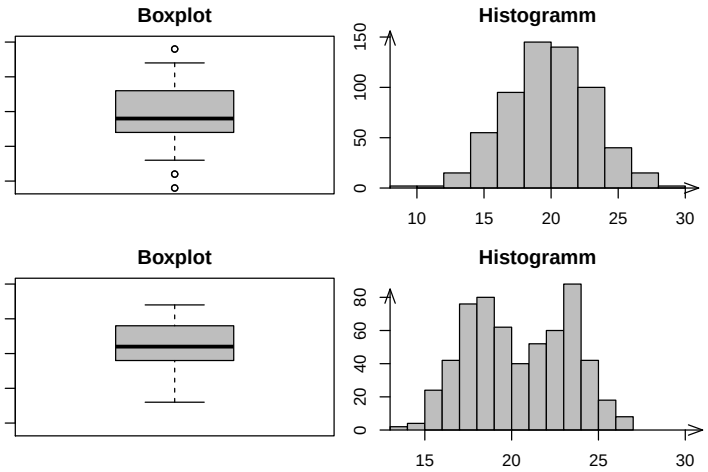


Abbildung 2.6: Darstellung Boxplot und Histogramm

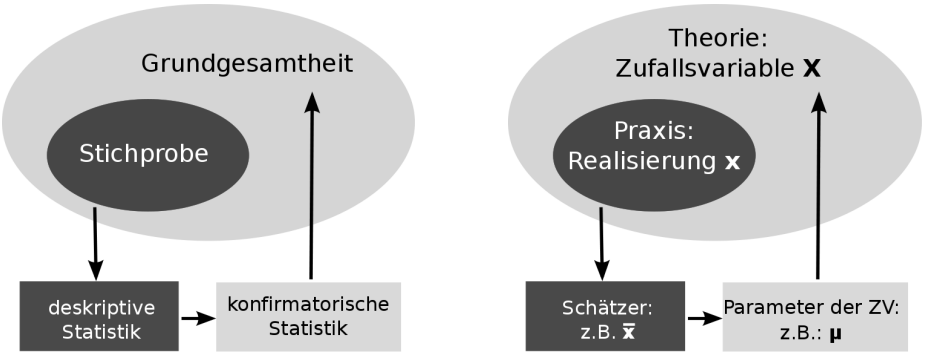
2.3 Zusammenfassung

- als Stichprobe wird eine repräsentative Teilmenge der Grundgesamtheit bezeichnet, an der Untersuchungen durchgeführt werden
- Merkmale beschreiben charakteristische Eigenschaften der Beobachtungseinheiten, Merkmale können auf verschiedenen Skalen klassifiziert werden (z.B. nominal, ordinal, metrisch)
- Lagemaße (wie Mittelwert und Median) und Streuungsmaße (wie Standardabweichung und Varianz) sind geeignet, die erhobenen Merkmale innerhalb einer Stichprobe in kompakter Form zu beschreiben

3 Schätzen

Die alleinige deskriptive Analyse von stichprobenbasierten Daten reicht zur Beantwortung wissenschaftlicher Fragestellungen in der Regel nicht aus. Ein Grund dafür ist, dass sie keine Angabe über die Genauigkeit bzw. die Vertrauenswürdigkeit der Daten in Bezug auf die zugrundeliegende Population der eigentlich interessierenden Untersuchungsobjekte macht. Da man aber in der Regel auf die Untersuchung von Stichproben angewiesen ist, benötigt man Methoden, die es zulassen, von der Stichprobe auf die zugrundeliegende Population zurückzuschließen.

Mit Hilfe der Methoden der schließenden Statistik werden aufgrund der Stichprobenresultate sowie geeigneter statistischer Modelle, Aussagen über die Grundgesamtheit gemacht (siehe Bild, links). D.h. man betrachtet auf Basis der Stichprobe konkrete Realisierungen einer Zufallsvariable(ZV) und benutzt diese im Rahmen der schließenden Statistik für Schlussfolgerungen bzgl. der interessierenden Grundgesamtheit, die durch eine entsprechende Zufallsvariable charakterisiert werden kann (siehe Bild, rechts).



3.1 Punktschätzer

Punktschätzungen liefern Werte für (unbekannte) Parameter. Um diese Werte zu erhalten, benötigt man eine Berechnungsvorschrift, welche als **Schätzfunktionen** bzw. **Schätzer** bezeichnet werden. Die Schätzfunktion legt fest, wie aus den Werten einer Stichprobe ein Schätzwert für einen theoretischen Parameterwert berechnet wird.
Beachte: Punktschätzer liefern keine Aussage über die Größe von (möglichen) Abweichungen des Schätzwertes vom gesuchten „wahren“ Parameter.
Zur Charakterisierung der Qualität von Schätzfunktionen werden eine Reihe von Eigenschaften benutzt. Wichtige Beispiele dieser Eigenschaften sind die **Konsistenz** und die **Erwartungstreue**. *Konsistenz* bedeutet, dass die Schätzfunktion um so genauere Ergebnisse liefert, je mehr Beobachtungen zur Verfügung stehen. *Erwartungstreue* bedeutet, dass ein Schätzer (der im Einzelfall aufgrund der Zufälligkeit der Stichprobe kaum dem „wahren“ Wert entsprechen wird) im Mittel den „wahren“ zu schätzenden Wert liefert. Eine systematische Abweichung des mittleren Schätzwertes vom „wahren“ Wert heißt **Bias** oder Verzerrung des Schätzers. Für einen erwartungstreuen Schätzer ist der Bias Null.

3.1.1 Beispiele für Punktschätzer

Das **arithmetische Mittel** ist ein konsistenter und erwartungstreuer Schätzer[1] für den Erwartungswert (μ) einer Zufallsvariable. Es wird zumeist mit $\overline{X}$ (als Schätzfunktion) bzw. $\overline{x}$ (als konkrete Realisierung)

[1] Schätzer werden oft durch ein $\widehat{\ }$ über dem Variablennamen als solche kenntlich gemacht.

bezeichnet und folgender Maßen berechnet:

$$\overline{X} = \widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.1}$$

Die **empirische Streuung** ist ein konsistenter und erwartungstreuer Schätzer für die Varianz (σ^2) einer Zufallsvariable. Sie wird mit S^2 (als Schätzfunktion) bzw. s^2 (als konkrete Realisierung) bezeichnet und folgender Maßen berechnet:

$$S^2 = \widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{3.2}$$

Die **relative Häufigkeit** ist ein konsistenter und erwartungstreuer Schätzer für die Erfolgswahrscheinlichkeit (p). Sie wird oft mit π bezeichnet und folgender Maßen berechnet:

$$\pi = \widehat{p} = \frac{\text{Anzahl der Erfolge}}{\text{Gesamtanzahl}} \tag{3.3}$$

3.1.2 Verteilungseigenschaften von Schätzern

Schätzer sind selbst Zufallsvariablen. D.h. sie unterliegen ebenfalls einer Verteilung und sind keine festen Werte; verschiedene Stichproben ergeben verschiedene Schätzwerte.

Beispiel: 3.1.1

Die Körpergröße in einer ausgewählten Population sei normalverteilt mit einem Erwartungswert von 175 cm und einer Streuung von 100 cm (d.h. $\mathcal{N}(175, 10^2)$). Angenommen der Erwartungswert sei unbekannt und soll anhand einer Stichprobe (SP) vom Umfang $n = 10$ geschätzt werden. Bei mehrmaliger Erhebung ($m = 10$) einer solchen SP wird diese aufgrund der zufälligen Auswahl der Stichprobenelemente unterschiedliche Werte annehmen. Die Verteilung der Einzelwerte entspricht der Verteilung der zugrundeliegenden Population.
Betrachtung der einzelnen Stichproben und der Verteilung der **Einzelwerte**:

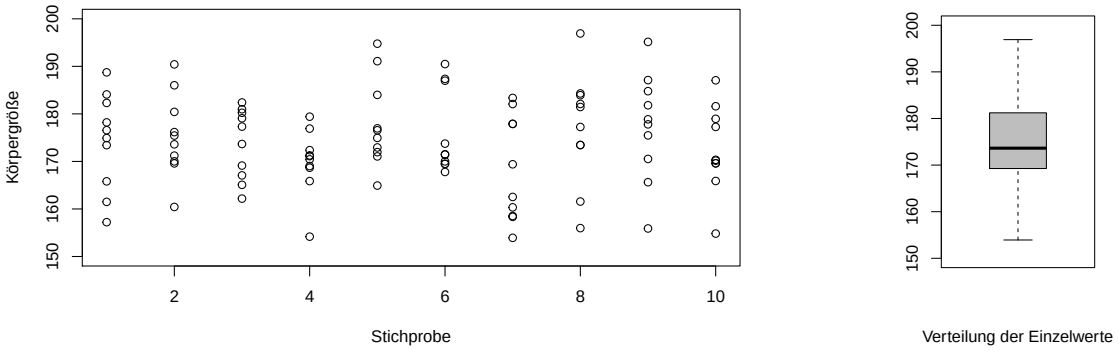


Abbildung 3.1: Verteilung der Einzelwerte

Fortsetzung des Beispiels

Betrachtung der einzelnen SP-mittelwerte und der Verteilung der **Mittelwerte**:

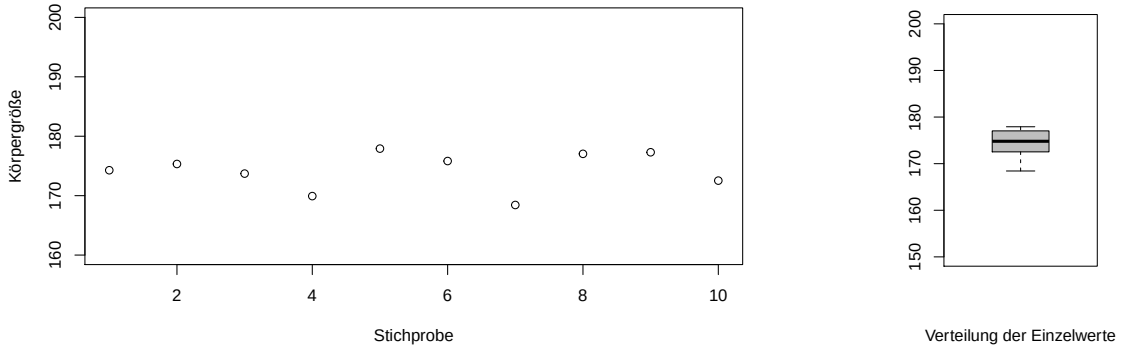


Abbildung 3.2: Verteilung der Mittelwerte

Wie im Beispiel 3.1 zu sehen, ist die Streuung der Stichprobenmittelwerte kleiner als die Streuung der Einzelwerte. Die Streuung der Mittelwerte kann bei Kenntnis der Streuung der zugrundeliegenden Population berechnet werden indem man letzteren Wert (d.h. σ^2) durch die Stichprobengröße n teilt. Entsprechend erhält man die Standardabweichung der Mittelwerte aus der Standardabweichung der Originaldaten indem man diese durch $\sqrt{n}$ dividiert. Man bezeichnet die Standardabweichung der Mittelwerte in der Literatur auch oft als **SEM (Standard Error of the Mean)**:

$$SEM = \frac{S}{\sqrt{n}} \tag{3.4}$$

Standardabweichung (Standard Deviation, SD) und Standardfehler (SEM) lassen sich unter Kenntnis von n in einander umrechnen, haben jedoch inhaltlich völlig andere Bedeutungen: Die Standardabweichung beschreibt die **Streuung von Daten** in einer Population. Sie ist ein Heterogenitätsmaß und repräsentiert *die mittlere Abweichung der Einzelwerte vom Mittelwert* und kann aus einer Stichprobe durch die Wurzel aus der empirischen Streuung geschätzt werden:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Im Gegensatz dazu liefert der Standardfehler des Mittelwertes ein Maß der **Präzision** einer Punktschätzung des Erwartungswertes. Der SEM ist damit ein *Maß für die Genauigkeit der Schätzung des Mittelwertes*. Wobei sich der Wert der Schätzung der Varianz einer Zufallsvariablen mit zunehmender Stichprobengröße nicht systematisch ändert, so nimmt die Genauigkeit dieser Schätzung mit zunehmender Stichprobengröße zu, d.h. *der SEM wird mit zunehmender Stichprobengröße kleiner*.

Beispiel: 3.1.2

Die Güte der Mittelwertschätzung hängt von der Fallzahl ab.
Von n = 10 (25,100,1000) Frauen wurden die HDL-Cholesterolwerte $\left[\frac{mg}{dl}\right]$ erhoben.

		n = 10	**n = 25**	**n = 100**	**n = 1000**
Mittelwert	$\overline{x}$	55.73	53.41	60.17	55.41
SD	s	18.40	15.13	18.04	15.37
SEM	$\frac{s}{\sqrt{n}}$	5.82	3.03	1.80	0.49

Tabelle 3.1: HDL-Cholesterolwerte

3.2 Intervallschätzungen

Wie oben dargelegt, sind Punktschätzer Zufallsvariablen und liefern per se keine Aussage darüber, wie sehr der Schätzwert vom „wahren“ Wert abweicht. Es stellt sich also die Frage, wie sehr man einem Punktschätzwert vertrauen kann. Es wäre wünschenswert, wenn anstatt eines einzelnen Wertes eine Umgebung angeben werden könnte, die eine Aussage über die „Genauigkeit“ der Schätzung zulässt. Eine solche Möglichkeit bieten sogenannte **Intervallschätzer**. Diese liefern im Ergebnis der Schätzung ein Intervall, das mit einem angebbaren Maß an Vertrauen (Konfidenz) den „wahren“ Parameter enthält.

3.2.1 Konfidenzintervall / Vertrauensbereich

Eine spezielle Intervallschätzung ist das **Konfidenzintervall (KI)**. Ein 95% - Konfidenzintervall erhält man durch Anwendung eines speziellen Schätzverfahrens, welches eine obere und eine untere Intervallgrenze liefert, so dass bei wiederholter Anwendung dieses Verfahrens 95% der resultierenden Intervalle den „wahren“ Parameterwert enthalten. Man kann ein KI auch so konstruieren, dass mehr oder weniger als 95% der Intervalle den „wahren“ Parameterwert enthalten. Allgemein wird dieser Anteil (bzw. die Wahrscheinlichkeit dafür den „wahren“ Wert zu enthalten) als **Konfidenzniveau** bezeichnet und gibt es in der Form $1 - \alpha$ an. Alpha ist hierbei die Irrtumswahrscheinlichkeit, d.h. die Wahrscheinlichkeit, dass das konstruierte Intervall den „wahren“ Parameterwert fälschlicher Weise nicht enthält. Im Beispiel 3.3 ist das Konfidenzniveau 0.95 bzw. 95%; die Irrtumwahrscheinlichkeit α beträgt damit 0.05 bzw. 5%.

Beachte:

- Das Konfidenzniveau ist frei wählbar
- Obwohl 95% einen gängigen Wert darstellt, sollte immer im Einzelfall geprüft werden, welches Konfindenzniveau (bzw. welche Irrtumswahrscheinlichkeit) adäquat ist
- Bei gegebenem α ist das KI umso kürzer, je größer der Stichprobenumfang ist
- Bei festem Stichprobenumfang ist das Intervall umso größer, je größer das Konfidenzniveau $(1 - \alpha)$ gewählt wird bzw. umso kürzer je kleiner das Konfidenzniveau (und damit je größer der tolerierte Fehler) gewählt wird

Korrekt interpretiert besagt ein KI, dass bei vielfacher Anwendung des Verfahrens auf unabhängige Stichproben zu erwarten ist, dass in $(1 - \alpha) \cdot 100\%$ der Fälle der „wahre“ (gesuchte) Parameterwert im jeweiligen Konfidenzintervall liegt. D.h. für ein konkret ermitteltes KI kann mit einer Wahrscheinlichkeit von $(1 - \alpha)$ **vermutet** werden, dass der „wahre“ Parameterwert im Intervall liegt. Ob der „wahre“ Parameterwert allerdings **im konkreten Fall** wirklich im KI liegt, bleibt unbekannt! Im Fall einer einzelnen vorliegenden Stichprobe, liegt der „wahre“ Parameterwert entweder im Intervall **oder** nicht!

Beispiel: 3.2.1

Der mittlere HDL-Cholesterolwert bei Frauen kann als normalverteilt mit einem Mittelwert von $55,1\frac{mg}{dl}$ und einer Standardabweichung von $11,7\frac{mg}{dl}$ angenommen werden. Bei Ziehung von 20 unabhängigen Stichproben mit jeweiligem Umfang n=25 haben sich die in Abbildung 3.3 dargestellten arithmetischen Mittelwerte jeder Stichprobe und die entsprechenden 95% - Konfidenzintervalle ergeben. Man sieht, dass von den 20 dargestellten Konfidenzintervallen 19 (d.h. 95%) den wahren Parameter von $55,1\frac{mg}{dl}$ (waagerechte Linie) überdecken (vgl. Abb.: 3.3).

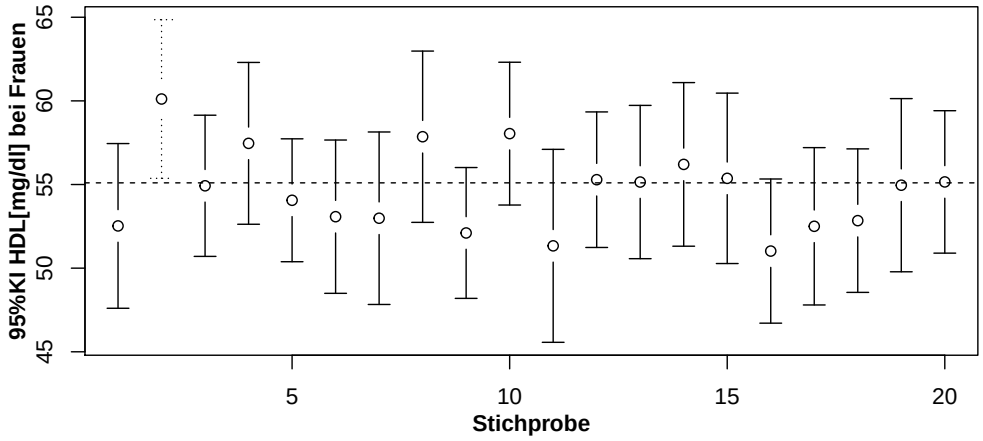


Abbildung 3.3: Beispiel für Intervallschätzung

3.2.2 Einschub:

Satz 3.1 (Zentraler Grenzwertsatz (ZGWS)) *Seien $X_1,\ X_2, \cdots,\ X_n$ unabhängige, identisch verteilte Zufallsvariablen mit dem Erwartungswert μ und der Varianz σ^2. Es gilt: für große n ist die Summe $X_1 + X_2 + \cdots + X_n$ asymptotisch normalverteilt mit dem Erwartungswert $n \cdot \mu$ und der Varianz $n \cdot \sigma^2$. Dies wiederum bedeutet, dass die normierte Größe Z_n asymtotisch standardnormal verteilt ist:*

$$Z_n = \frac{(\sum_{i=1}^{n} X_i) - n \cdot \mu}{\sqrt{n} \cdot \sigma} = (\overline{X} - \mu) \cdot \frac{\sqrt{n}}{\sigma} \sim \mathcal{N}(0,1) \tag{3.5}$$

Dieser zentrale Satz der Wahrscheinlichkeitstheorie bildet die Grundlage der vielfach verwendeten Annahme, dass aus einem (additiven) Zusammentreffen vieler unabhängiger Zufallseinflüsse im Resultat von einer Normalverteilung ausgegangen werden kann. Eine spezifische Folgerung dieses Satzes ist es auch, dass die Verteilung der Mittelwerte, die aus (unabhängigen) Stichproben des Umfangs n aus *derselben Grundgesamtheit* (mit Erwartungswert μ und Varianz σ^2) hervorgehen, asymptotisch (Faustregel: $n > 25$) einer Normalverteilung mit Erwartungswert μ und Varianz σ^2/n folgt. Dies gilt unabhängig von der konkreten Verteilung der Grundgesamtheit!

3.2.3 Konfidenzintervall für den Erwartungswert bei bekannter Varianz σ^2

Aus dem ZGWS folgt die Normalverteilung von Mittelwerten. Deshalb gilt für die Wahrscheinlichkeit P, den „wahren“ Wert μ in 95% der Fälle innerhalb des Konfidenzintervalls zu lokalisieren:

$$P(-1,96 \leq (\overline{X} - \mu) \cdot \frac{\sqrt{n}}{\sigma} \leq 1,96) = 0,95$$

bzw. durch Umstellung der Ungleichungen

$$P(\overline{X} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}) = 0,95$$

Der Wert von 1,96 entspricht dabei dem 0,975 Quantil der Standard-$\mathcal{N}$-Verteilung. Das resultierende 95%-Konfidenzintervall für den Mittelwert lautet damit:

$$\left[\overline{X} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Ein allgemeines Konfidenzintervall zum Konfidenzniveau $(1-\alpha)$ für den Mittelwert kann folgender Maßen angegeben werden:

$$\left[\overline{X} - \; z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + \; z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{3.6}$$

hierbei ist $z_{1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung.

Beispiel: 3.2.2

In einem Labor ist durch Langzeiterfahrung bekannt, dass ein bestimmter Enzym-Messwert als normalverteilt mit einer **theoretischen Standardabweichung** von 1,5 i.E. angenommen werden kann. Aus einer konkreten Stichprobe von 25 Enzymbestimmungen ergab sich ein arithmetisches Mittel von $\overline{x} = 22,3$ i.E. Das zugehörige 95%- Konfidenzintervall für den Erwartungswert des Enzyms wird wie folgt berechnet:

Da ein 95% KI gesucht ist, muss zunächst das 0,975 Quantil der Normalverteilung bestimmt werden[a]:

$$z_{1-\frac{\alpha}{2}} = \; z_{0,975} = 1,96$$

Damit kann mit der Formel 3.6 das 95%-KI berechnet werden:

$$22,3 - 1,96\frac{1,5}{\sqrt{25}} \leq \mu \leq 22,3 + 1,96\frac{1,5}{\sqrt{25}}$$

$$21,71 \leq \mu \leq 22,89$$

Interpretation: Der unbekannte mittlere Wert des Enzyms in der Grundgesamtheit (d.h. der Erwartungswert) kann mit einer Wahrscheinlichkeit von 95% im Intervall $[21,71;\; 22,89]$ vermutet werden.

[a] Quantile sind in Tabellen nachlesbar bzw. durch entsprechende Statistiksoftware berechenbar

3.2.4 Konfidenzintervall für den Erwartungswert bei unbekannter Varianz σ

Die Aussagen im Abschnitt 3.2.3 setzen die Kenntnis der Standardabweichung σ voraus. Zumeist ist diese aber nicht bekannt und muss z.B. durch die empirische Standardabweichung S geschätzt werden. In diesem Fall ist die standardisierte Schätzung des Erwartungswertes $(\overline{X} - \mu) \cdot \frac{\sqrt{n}}{S}$ **nicht** *standardnormalverteilt* sondern *t-verteilt*. Dies hat zur Folge, dass das Quantil der Standardnormalverteilung ($z_{1-\frac{\alpha}{2}}$) bei der Berechnung des KI durch das **Quantil der *t*-Verteilung** ($t_{n-1;1-\frac{\alpha}{2}}$) ersetzt werden muss, d.h. das $(1-\alpha)$-Konfidenzintervall für den Mittelwert bei unbekannter Streuung ergibt sich aus:

$$\left[\overline{X} - \; t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \; ; \; \overline{X} + \; t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right] \tag{3.7}$$

Beachte: Die genaue Form einer **t - Verteilung** hängt vom Umfang der Stichprobe n ab. Diese Abhängigkeit wird durch den sogenannten **Freiheitsgrad** f charakterisiert. Freiheitsgrade werden im vorliegenden Fall aus dem Stichprobenumfang n mittels $f = n-1$ berechnet. D.h., je nach Stichprobengröße muss das Quantil der jeweils gültigen t-Verteilung mit $f = n-1$ Freiheitsgraden verwendet werden.

Beispiel: 3.2.3

Anhand einer Stichprobe von 25 Frauen wurde ein mittlerer HDL-Cholesterolwert von $\overline{x} = 54{,}9\ \frac{mg}{dl}$ und eine **empirische Standardabweichung** von $S = 14,8\frac{mg}{dl}$ ermittelt. Das 95%- Konfidenzintervall für das Populationsmittel des HDL-Cholesterolwertes wird wie folgt berechnet:
Da die Standardabweichung aus der Stichprobe geschätzt wurde, muss eine t-Verteilung genutzt werden. Da ein 95% KI gesucht wird, bestimmt man zunächst das 0,975-Quantil der t-Verteilung mit $25 - 1 = 24$ Freiheitsgraden. Es lautet:

$$t_{24;\,0{,}975} = 2,06.$$

Damit kann mit der Formel 3.7 das 95% KI berechnet werden:

$$54,9 - 2,06\frac{14,8}{\sqrt{25}} \leq \mu \leq 54,9 + 2,06\frac{14,8}{\sqrt{25}}$$

$$48,83 \leq \mu \leq 61,05$$

Interpretation: Der unbekannte mittlere HDL Cholesterolwert in der zugrundeliegenden Population (gemessen in $\left[\frac{mg}{dl}\right]$) kann mit einer Wahrscheinlichkeit von 95% im Intervall $[48,83;\ 61,05]$ vermutet werden.

3.2.5 Konfidenzintervall für eine Wahrscheinlichkeit

Als weiteres Beispiel für die Berechnung eines Konfidenzintervalls für einen konkreten Parameter sei an dieser Stelle das Konfidenzintervall für eine unbekannte (Erfolgs-)Wahrscheinlichkeit vorgestellt. Hierbei beschränken wir uns auf die Version des Konfidenzintervalls, welches auf einer Normalverteilungsapproximation einer binomialverteilen Zufallsgröße beruht und nutzen die relative Häufigkeit als Schätzer für die unbekannte Wahrscheinlichkeit. Da das hier vorgestellte Verfahren nur approximativ gilt, setzten wir voraus, dass $n \cdot \widehat{p} > 5$ und $n \cdot (1 - \widehat{p}) > 5$ gilt. D.h. wir gehen davon aus, dass der Stichprobenumfang n nicht zu klein und die relative Häufigkeit $\widehat{p}$ nicht zu extrem (d.h. nicht zu nahe an 0 oder 1) ist. In diesem Fall kann man ein Konfidenzintervall für die unbekannte Wahrscheinlichkeit p folgender Maßen angeben:

$$\left[\widehat{p} - \frac{1}{2n} - \ z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\ ;\ \widehat{p} + \frac{1}{2n} + \ z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] \tag{3.8}$$

3.2.6 Konfidenzintervalle für wichtige Risikomaße

Ohne näher auf Anwendung und Interpretation der Risikomaße selbst einzugehen (siehe hierzu Literatur, z.B. Sachs & Hedderich: *Angewandte Statistik*), sollen an dieser Stelle kurz die KI für das **Relative Risiko (RR)** sowie das **Odds Ratio (OR)** angegeben werden.
Hierzu gehen wir von einer Vierfeldertafel (Tab.: 3.2) aus, in der Krankheitsstatus und Risikofaktor (Exposition) folgender Maßen gelistet sind:

	D	$\overline{D}$	Gesamt
E	a	b	a+b
$\overline{E}$	c	d	c+d
Gesamt	a+c	b+d	n=a+b+c+d

Tabelle 3.2: (E: Exposition, $\overline{E}$: keine Exposition, D: krank, $\overline{D}$: gesund)

Relatives Risiko

Das Relative (Erkrankungs-)Risiko bzgl. der Exposition berechnet sich aus:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)}. \tag{3.9}$$

Das zugehörige $(1-\alpha)\cdot 100\%$-KI lautet:

$$\left[e^{ln(RR) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}} \quad ; \quad e^{ln(RR) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}} \right] \tag{3.10}$$

Odds Ratio

Das Odds Ratio zu Tabelle 3.2 berechnet sich aus:

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \cdot d}{b \cdot c}. \tag{3.11}$$

Das zugehörige $(1-\alpha)\cdot 100\%$-KI lautet:

$$\left[e^{ln(OR) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} \quad ; \quad e^{ln(OR) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} \right] \tag{3.12}$$

3.3 Zusammenfassung

- Schätzer (d.h. auch Grenzen von Konfidenzintervallen) sind selbst Zufallsgrößen
- Bei Angabe von Punktschätzern (möglichst) immer eine Intervallschätzung angeben, um Aussage zum Grad der Abweichung der Schätzung vom zu schätzenden Wert zu erhalten
- Standardabweichung (SD) und Standardfehler (SE) messen unterschiedliche Dinge: SD ist Maß der Datenvariabilität; SEM (SE des Mittelwertes) ist Maß der Genauigkeit der Mittelwertschätzung
- Ein Konfidenzintervall liefert einen Bereich in dem man den „wahren“ zu schätzenden Parmeterwert mit vorgegebener Wahrscheinlichkeit (Konfidenzniveau) vermuten kann

4 Statistisches Testen - Signifikanztests

Ein statistischer Test überprüft Aussagen über eine oder mehrere Grundgesamtheiten anhand gegebener Stichproben aus diesen Grundgesamtheiten.Solche Aussagen, oft auch Hypothesen genannt, beziehen sich auf Parameter, die eine interessierende Eigenschaft der Grundgesamtheit(en) beschreiben.Zum Beispiel kann ein Test eine Entscheidung darüber fällen, ob der Median zweier Grundgesamtheiten statistisch signifikant verschieden ist oder nicht.
Damit gehört der statistische Test wie auch das Schätzen 3 zur schließenden Statistik.Im Unterschied zum Schätzen werden beim Testen keine Parameter geschätzt, sondern das Resultat eines Tests ist eine statistisch begründete Entscheidung für oder gegen eine Hypothese zu bestimmten Parametern der beteiligten Grundgesamtheiten.
Aussagen, wie:

> „Der Unterschied zwischen Behandlungs- und Kontrollgruppe bzgl. ihrer Cholesterinwerte ist auf dem 1%-Signifikanzniveau gesichert (p = 0,0023).“
>
> „Es besteht ein statistisch signifikanter Zusammenhang zwischen Alter und systolischem Blutdruck (p = 0,016).“
>
> „Aufgrund der Datenlage kann kein Unterschied zwischen den Überlebenszeiten der beiden zu vergleichenden Therapiegruppen nachgewiesen werden (p = 0,49675).“

sind Interpretationen von Ergebnissen statistischer Tests. Um diese Aussagen richtig zu verstehen und zu bewerten, ist es notwendig, die Grundidee statistischer Tests zu kennen.

4.1 Einführung

Statistische Tests überprüfen Aussagen über eine oder mehrere Grundgesamtheiten. In diesem Kapitel beschreiben wir **Signifikanztests**, die zum Ziel haben, einen Unterschied zwischen den Grundgesamtheiten als statistisch signifikant nachzuweisen.
Daneben gibt es auch andere statistische Tests wie zum Beispiel Äquivalenztests, bei denen gerade die Gleichwertigkeit (etwa zweier Medikamente) gezeigt werden soll. Vieles, was in diesem Kapitel gesagt wird, ist für alle statistische Tests gültig. Allerdings ist insbesondere die Hypothesenformulierung bei Äquivalenztests anders als bei Signifikanztests, die wir hier beschreiben.
Das Konzept des Signifikanztests soll anhand des Beispiels 4.1 verdeutlicht werden.

Beispiel: 4.1.1

In einer Studie wurde bei insulinpflichtigen Diabetikern und einer Kontrollgruppe der Plasma - Magnesium - Spiegel $[\frac{\text{mmol}}{\text{l}}]$ gemessen. Die Ergebnisse sind in nachfolgender Tabelle zusammengefasst:

	$\bar{x}$	s	n
Kontrollgruppe	0,81	0,057	140
Diabetiker	0,719	0,068	227

Tabelle 4.1: Plasma - Magnesium - Spiegel

Frage: Unterscheidet sich der mittlere Plasma-Magnesium-Spiegel der insulinpflichtigen Diabetiker von dem der Kontrollen?

Zur Beantwortung derartiger Fragen gehen statistische Tests wie folgt vor:

- In einer sogenannten Nullhypothese trifft man die Aussage, dass kein Effekt vorliegt. D.h. in Beispiel 4.1, das der Magnesium-Spiegel bei Diabetikern im Grunde gleich dem der Kontrollgruppe ist und die beobachteten Unterschiede nur zufällige Schwankungen sind.

- Unter dieser Voraussetzung berechnet man die Wahrscheinlichkeit, mindestens diesen Unterschied (d.h. in Beispiel 4.1 die Differenz von $0{,}719\frac{\text{mmol}}{\text{l}}$ und $0{,}81\frac{\text{mmol}}{\text{l}}$) zu beobachten.
- Ist diese Wahrscheinlichkeit (sehr) groß, so spricht nichts dagegen, an der Nullhypothese, d.h. einem „rein zufälligen“ Unterschied festzuhalten. Ist sie dagegen (sehr) klein, d.h. ist es sehr unwahrscheinlich, einen so großen Unterschied im Magnesium-Spiegel unter der gegebenen Nullhypothese (d.h. kein Unterschied) zu beobachten, schlussfolgert man, dass die Gleichheitsvoraussetzung nicht beibehalten werden sollte. Das heißt, die beiden Populationen unterscheiden sich.

4.2 Statistische Hypothesen

Zu Beginn des statistischen Tests muss man klar die Aussage über die Grundgesamtheit(en) formulieren, die getestet werden soll(en). Solche Aussagen werden auch *Hypothesen* genannt. Typischerweise geht es dabei um einen Parameter der beteiligten Grundgesamtheit(en) wie zum Beispiel den Mittelwert, den Median, die Varianz oder eine Erfolgswahrscheinlichkeit. Man stellt zwei komplementäre Hypothesen auf, von denen eine wahr sein muss:

- Die **Nullhypothese** H_0 beschreibt den Fall, dass *kein systematischer Effekt vorliegt*, sondern von einem Zufallseffekt ausgegangen werden kann. Von dieser Hypothese wird zunächst ausgegangen.
- Die **Alternativhypothese** H_A ist zur Nullhypothese komplementär und drückt aus, dass *ein Effekt vorliegt*. Häufig ist die Alternativhypothese die innovative und interessante Aussage, die statistisch untermauert werden soll.

Beispiel: 4.2.1

Für oben genanntes Beispiel lauten die Hypothesen folgendermaßen:

- **Nullhypothese**

  $$H_0 \ : \ \mu_1 = \mu_2 \ \text{ bzw. } \ \mu_1 - \mu_2 = 0$$

  Der mittlere Magnesium-Spiegel μ_1 der insulinabhängigen Diabetespatienten ist gleich dem mittleren Magnesium-Spiegel μ_2 der Kontrollen bzw. die beobachteten Abweichungen sind zufällig (Zufallseffekt).

- **Alternativhypothese**

  $$H_A \ : \ \mu_1 \neq \mu_2 \ \text{ bzw. } \ \mu_1 - \mu_2 \neq 0$$

  Der mittlere Magnesium-Spiegel μ_1 der insulinabhängigen Diabetespatienten ist ungleich dem mittleren Magnesium-Spiegel μ_2 der Kontrollen. Die beobachteten Mittelwerte unterscheiden sich signifikant (systematischer Effekt).

Weil die Nullhypothese meist die Populationsparameter, zu denen der Test eine Aussage macht, genau festlegt, kann man anhand der tatsächlichen Beobachtungen in den Stichproben beurteilen, ob diese zur Nullhypothese passen. Im obigen Beispiel impliziert etwa die Nullhypothese $H_0 : \mu_1 - \mu_2 = 0$, dass die Differenz der Stichprobenmittelwerte $\bar{x}_1 - \bar{x}_2$ im Mittel 0 sein muss. Die Alternativhypothese legt die Populationsparameter nicht bzgl. konkreter Werte, sondern nur bzgl. eines potentiellen (wie auch immer gearteten) Unterschiedes fest. Im Beispiel 4.2 werden völlig verschiedene Situationen, sagen wir $\mu_1 - \mu_2 = 5,3$ oder $\mu_1 - \mu_2 = -11$, durch die Alternativhypothese abgedeckt.
Lassen sich die tatsächlichen Beobachtungen der Stichprobe durch die Bedingungen der Nullhypothese nur unzulänglich erklären, werden die ursprünglichen Annahmen aus der Nullhypothese als unhaltbar verworfen.
Bei einem statistischen Test wird also nichts „bewiesen“; Verwerfen der Nullhypothese wird aufgrund einer Wahrscheinlichkeitsaussage lediglich statistisch untermauert. In diesem Fall wird indirekt auf die Gültigkeit der Alternativhypothese geschlossen. Diese Vorgehensweise wird auch als **statistisches Falsifizierungsprinzip** bezeichnet.

4.2.1 Zweiseitige und einseitige Fragestellung

Die oben genannten Hypothesen entsprechen einer zweiseitigen Fragestellung, d.h. die Nullhypothese wird abgelehnt, wenn $\mu_1 < \mu_2$ (In Beispiel 4.2 ist der mittlere Magnesium-Spiegel der Diabetiker kleiner als der der Kontrollen) **oder** $\mu_1 > \mu_2$ ist.
Einseitige Fragestellungen sind angemessen, wenn aufgrund von Vorinformationen Abweichungen nur in eine Richtung zu erwarten sind. Die Nullhypothese bleibt dann unverändert, aber die Alternativhypothese H_A ist **entweder** $\mu_1 < \mu_2$ **oder** $\mu_1 > \mu_2$.

4.3 Testentscheidung

Aufgrund einer vorliegenden Stichprobe muss der Test zwischen der Nullhypothese und der Alternativhypothese entscheiden. Der **Ablehnungsbereich (kritischer Bereich, rejection area)** besteht aus all den möglichen Realisierungen der Stichprobe, die zu einer Ablehnung der Nullhypothese führen. Umgekehrt liegen Realisierungen einer Stichprobe im **Akzeptanzbereich (acceptance area)**, wenn die Nullhypothese nicht abgelehnt werden kann. Wie die Abgrenzung dieser beiden Bereiche realisiert wird, ist zunächst nicht vorgegeben, und es sind eine Vielzahl von Möglichkeiten denkbar. In der Praxis definiert man die beiden Bereiche häufig so, dass bestimmte Fehlentscheidungen des Tests nicht zu häufig auftreten. Diese Herangehensweise soll im Folgenden näher erläutert werden.

4.3.1 Fehlentscheidungen

Da Stichproben zufällig zustande kommen, besteht die Gefahr einer Fehlentscheidung, wenn von der Stichprobe auf die Grundgesamtheit geschlossen wird. Mögliche Fehler werden in Tabelle 4.2 veranschaulicht.

	H_0 **falsch**	H_0 **richtig**
H_0 **ablehnen**	richtig signifikant $1-\beta$	falsch signifikant Fehler 1. Art: α
H_0 **beibehalten**	falsch nicht signifikant Fehler 2. Art: β	richtig nicht signifikant $1-\alpha$

Tabelle 4.2: Fehlentscheidungen

Der Fehler 1. Art α wird auch als **Signifikanzniveau** (Irrtumswahrscheinlichkeit) bezeichnet und gibt die Wahrscheinlichkeit an, die Nullhypothese abzulehnen, obwohl sie richtig ist. Gilt statt der Nullhypothese die Alternativhypothese, d.h. es liegt ein Effekt in einer gewissen Größe vor, dann ist der **Fehler zweiter Art** β die Wahrscheinlichkeit, dass der Test die falsche Nullhypothese beibehält. Positiv ausgedrückt bedeutet dies, dass bei diesem spezifischen Effekt der Test mit einer Wahrscheinlichkeit $1-\beta$ die Nullhypothese richtigerweise ablehnt. Diese Wahrscheinlichkeit $1-\beta$ nennt man auch **Güte** oder auch **Power** des Tests bei gegebenem spezifischem Effekt. Die Power des Tests spielt bei der Fallzahlplanung vor der Durchführung einer Untersuchung eine entscheidende Rolle, siehe dazu auch Abschnitt 4.6.

4.3.2 Testgröße (Teststatistik, Prüfgröße)

Auf dem Weg hin zu einer Testentscheidung berechnen viele Standardtests zunächst aus der Stichprobe eine geeignete Testgröße (Prüfgröße), die ein Maß für den beobachteten Effekt ist. Zum Beispiel ist für die Frage nach einem Unterschied in der mittleren Lage zweier Population eine normierte Differenz der Mittelwerte von zwei Stichproben aus den beiden Populationen eine sinnvolle Testgröße (in Beispiel 4.3 wird dies anhand des 2-Stichproben t-Testes realisiert): je größer der mittlere Unterschied in den Stichproben betragsmäßig ausfällt, desto größer wird auch der Betrag dieser Testgröße sein. Dies und einige weitere klassische Beispiele für statistische Tests und deren Prüfgrößen werden im nächsten Kapitel vorgestellt. Typischerweise ist die Verteilung der Testgröße unter der Nullhypothese (d.h.: „kein Effekt“) bekannt, und man wählt den Ablehnungsbereich so, dass die Wahrscheinlichkeit für einen Fehler 1. Art (Ablehnen einer richtigen Nullhypothese) nicht größer ist als ein vorgegebenes Signifikanzniveau α, zum Beispiel $\alpha = 0,01$ (vgl. Abb.: 4.1).

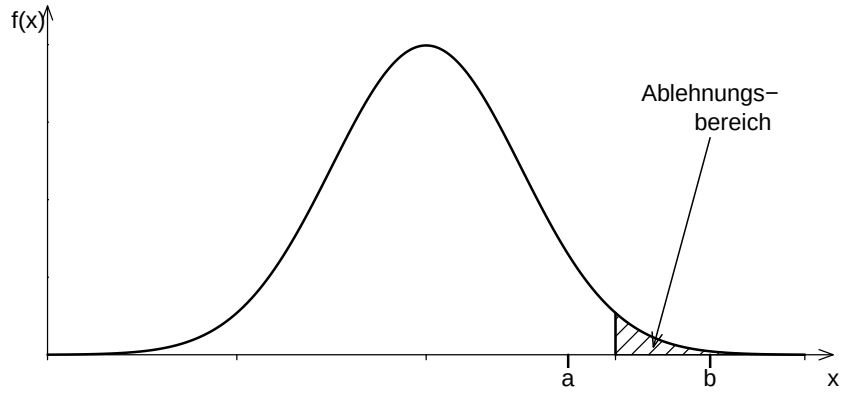


Abbildung 4.1: Ablehnungsbereich

Der Ablehnungsbereich wird unterschiedlich ausfallen, je nachdem, ob einseitig oder zweiseitig getestet wird, Die Abbildung 4.2 zeigt bei einer standard-normalverteilten Testgröße für $\alpha = 0,05$ den Annahme- und Ablehnungsbereich jeweils bei ein- und zweiseitigen Fragestellungen.

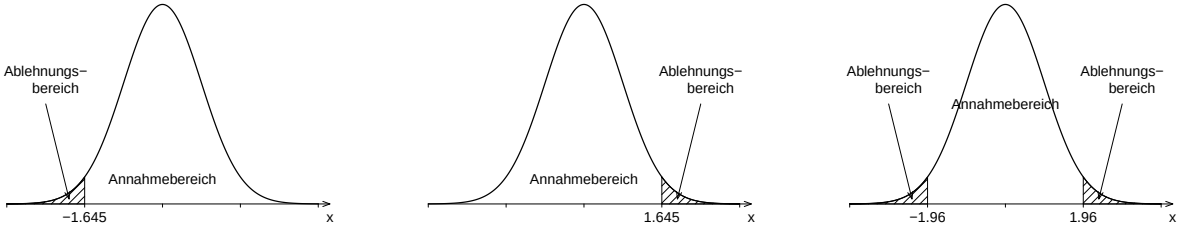


Abbildung 4.2: ein- und zweisseitige Fragestellung

Beispiel: 4.3.1

In diesem Beispiel kommt ein 2-Stichproben t-Test zur Anwendung. In Kapitel 5 wird auf diesen Test genauer eingegangen werden.

Daten der Stichproben:

	$\bar{x}$	s	n
Kontrolle	0,81	0,057	140
Diabetiker	0,719	0,068	227

Tabelle 4.3: Stichprobendaten

Testgröße für den 2-SP t-Test zum Mittelwertvergleich:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{0,81 - 0,719}{\sqrt{\frac{0,057^2}{140} + \frac{0,068^2}{227}}} = 13,79$$

4.4 Beurteilung des Testergebnisses

Wie im Abschnitt 4.3 schon angedeutet trifft ein statistischer Test eine Entscheidung zwischen der Null- und der Alternativhypothese. Statt der binären Testentscheidung kann man aus dem Test auch eine Wahrscheinlichkeitsaussage (den sogenannte p-Wert) ableiten.

4.4.1 Test als Entscheidungsprozedur

Liegt die Testgröße, die man anhand konkreter Stichprobendaten errechnet hat, im Ablehnungsbereich, kann man die Nullhypothese auf dem vorgegebenem Signifikanzniveau ablehnen. Ansonsten ist der Test nicht signifikant, und die Nullhypothese kann nicht verworfen werden.

Beispiel: 4.4.1

Der berechnete Wert der Prüfgröße im Beispiel 4.3 ($t = 13,79$) liegt bei zweiseitiger Fragestellung und einem vorgegebenen Signifikanzniveau von $\alpha = 0,05$ rechts vom Quantil der Standardnormalverteilung $z_{0,975} = 1,96$, d.h. im Ablehnungsbereich.

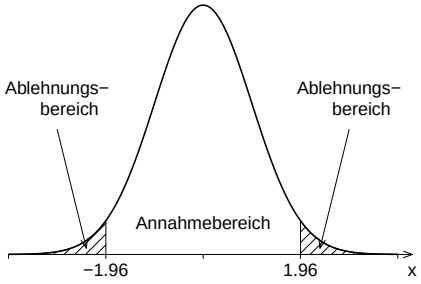


Abbildung 4.3: Ablehnungsbereich a=0.05

Daraus schlussfolgert man, dass die Nullhypothese auf dem gegebenen Signifikanzniveau von $\alpha = 5\%$ abgelehnt wird. Der Vergleich lautet:

$$\text{berechnetete Prüfgröße} = 13{,}79 > \text{Quantil der Prüfgröße} = 1{,}96$$

Bemerkung: Hier sei nochmals darauf hingewiesen, dass die Quantile für Prüfgrößen in Tabellen nachgeschlagen werden bzw. durch Statistiksoftware berechnet werden können.

4.4.2 Test als p-Wert-Prozedur

Ein alternatives Konzept zur Testentscheidung basiert auf dem p-Wert. Der p-Wert gibt die Wahrscheinlichkeit dafür an, bei gültiger Nullhypothese H_0 die beobachtete Testgröße oder ein unter H_0 noch größeren Betrag der Testgröße (d.h. in der Regel jeder noch weiter von Null entferntere Wert) zu erhalten. In diesem Sinn spricht ein großer p-Wert eher für die Nullhypothese; ein kleiner p-Wert eher dagegen. Das im Vorhinein gewählte Signifikanzniveau α markiert die Grenze, ab wann der Test signifikant ausfällt: der Test ist genau dann signifikant, wenn der p-Wert kleiner ist als α.
In den Abbildungen 4.4 und 4.5 wird der Zusammenhang zwischen Testgröße ('T' in den Abbildungen 4.4 und 4.5), p-Wert und Testentscheidung bei Irrtumswahrscheinlichkeit α veranschaulicht. Der p-Wert entspricht der Fläche unter der Verteilungskurve der Testgröße bei gültiger H_0, die jenseits der errechneten Testgröße liegt (schraffierte Flächen). Bei zweiseitiger Fragestellung sind das die Flächen an beiden Rändern. Ganz analog entspricht das Signifikanzniveau α der Fläche unter derselben Kurve, die jenseits eines bestimmten Quantils ('K' in den Abbildungen 4.4 und 4.5) der Verteilung der Testgröße unter H_0 liegt. Bei zweiseitiger Fragestellung sind das die Flächen an beiden Rändern jenseits des $(1 - \frac{\alpha}{2})$-Quantils. Bei einseitiger Fragestellung ist es die Fläche an einem Rand jenseits des $(1 - \alpha)$-Quantils. Da p-Werte aus der konkreten Stichprobe berechnet werden, spricht man auch vom *empirischen* oder *nominellen Signifikanzniveau.*[1]
Statistische Tests werden in der Regel mit einer Statistik-Software berechnet, wobei das Ergebnis des Tests mit Angabe des p-Wertes erfolgt.

[1] Eine weitere Sprechweise für den p-Wert ist *Überschreitungswahrscheinlichkeit*

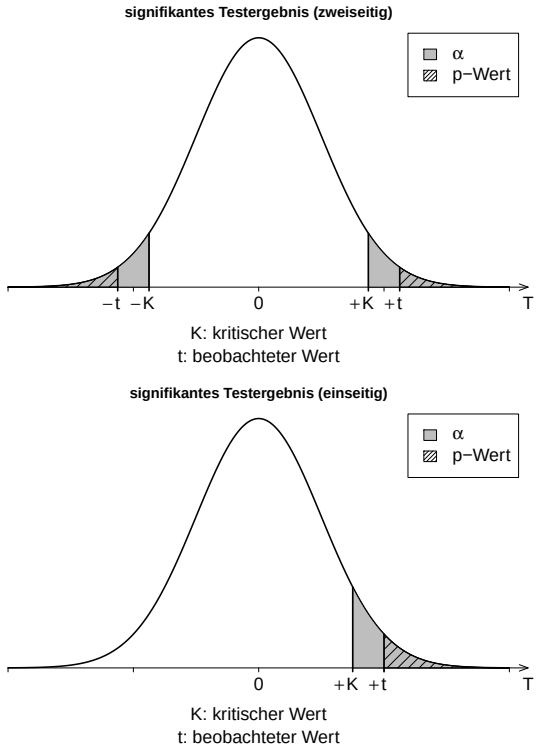


Abbildung 4.4: Bei signifikantem Testergebnis sind die schraffierten Flächen (der p-Wert) kleiner als das Signifikanzniveau α.

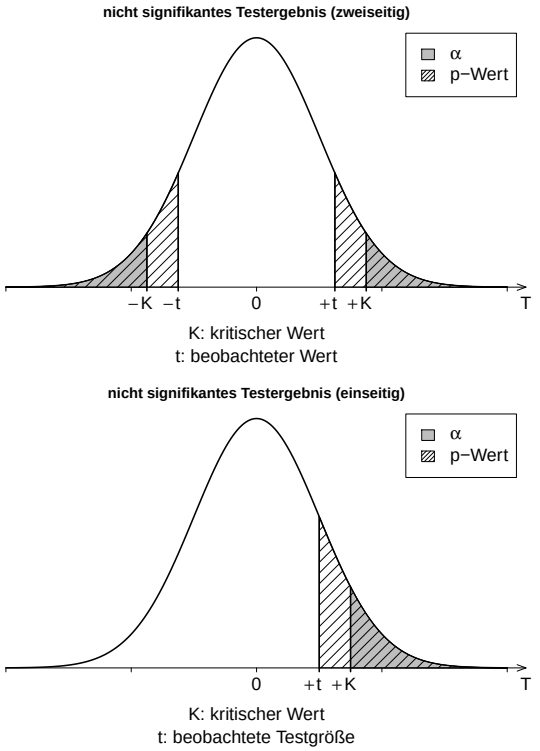


Abbildung 4.5: Bei nicht signifikantem Testergebnis sind die schraffierten Flächen (der p-Wert) größer als das Signifikanzniveau α.

Beispiel: 4.4.2

Für unser Beispiel wurde ein sehr kleiner p-Wert berechnet. Die Angabe „$p = 0,0000$“ des Computerprogramms bedeutet, dass sehr viele Nullen nach dem Komma folgen, die nicht mehr angezeigt werden. Sehr kleine p - Werte, wie im Beispiel berechnet, gibt man daher i.A. mit $p < 0,00005$ an. Das heißt, die Wahrscheinlichkeit eine Differenz von 0,81 - 0,719 = 0,091 (oder eine noch größere Differenz) unter der Annahme, dass sich die Mittelwerte nur zufällig unterscheiden, zu beobachten, ist so selten, dass man nicht davon ausgehen kann, dass die Diabetiker und Kontrollgruppe der gleichen Grundgesamtheit entstammen. Der berechnete p-Wert ist kleiner als das vorgegebene Signifikanzniveau von $\alpha = 0,05\%$. Die Nullhypothese wird abgelehnt.

Die Angabe von p-Werten ist bei der Beurteilung von Testergebnissen deutlich informativer als der Vergleich der berechneten Werte mit den Quantilen der Prüfgröße. Es ist also zu empfehlen, bei der Interpretation von Testergebnissen p-Werte mit anzugeben. Zur Angabe des p-Werts gehört allerdings immer auch die Hypothese, der verwendete Test und die Stichprobengröße, da sie alle den p-Wert beeinflussen und nur mit all diesen Information der p-Wert interpretierbar wird.

4.5 Interpretation des Testergebnisses

Nach der Testentscheidung muss das Ergebnis für die konkrete Fragestellung interpretiert werden.

Interpretationen. Es gibt verschiedene Arten, das Testergebnis in Worte zu fassen. Mögliche Interpretationen für das Beispiel 4.5 lauten:

- Die Wahrscheinlichkeit, derartige Mittelwertunterschiede zu beobachten, unter der Annahme, dass die Abweichungen zufällig sind, ist kleiner als $0,00005$, d.h. die Wahrscheinlichkeit eines Zufallseffekts ist extrem klein. Die beiden Mittelwerte unterscheiden sich signifikant.

- Die Nullhypothese wird auf dem vorgegebenen 5% - Niveau abgelehnt. Insulinpflichtigen Diabetiker haben im Mittel einen niedrigeren Magnesium - Spiegel als die Kontrollen ($p < 0,00005$).
- Der mittlere Plasma-Magnesium-Spiegel bei insulinpflichtigen Diabetikern unterscheidet sich signifikant vom mittleren Plasma-Magnesium-Spiegel der Kontrollen mit $p < 0,00005$ (t-Test, $n_1 = 140$, $n_2 = 227$).
- Der mittlere Plasma-Magnesium-Spiegel von $0,719 \frac{\text{mmol}}{\text{l}}$ in der Gruppe der insulinpflichtigen Diabetiker weicht signifikant vom mittleren Wert der Kontrollgruppe von 0,81 $\frac{\text{mmol}}{\text{l}}$ ($p < 0,00005$ t-Test, $n_1 = 140$, $n_2 = 227$) ab. Das heißt, Diabetiker und Kontrollgruppe entstammen nicht der gleichen Grundgesamtheit.

Fehlinterpretationen Es gibt auch einige Möglichkeiten, das Testergebnis falsch zu deuten, und wir wollen an dieser Stelle auf mögliche Fallstricke hinweisen.
Das *Beibehalten* der Nullhypothese bedeutet **nicht**, dass die Nullhypothese in jedem Fall wahr ist. Es ist vielmehr so zu verstehen, dass man einen Zufallsbefund nicht ausschließen kann bzw. dass die Fallzahl nicht ausreichend war, um den beobachteten Effekt als signifikant auszuweisen. Ein eventuell vorhandener Unterschied war nicht nachweisbar.
Das *Verwerfen* der Nullhypothese bedeutet **nicht**, dass die Nullhypothese H_0 zwangsläufig falsch ist. Die Wahrscheinlichkeit eines solchen falsch positiven Testergebnisses (d.h. der Test verwirft eine gültige Nullhypothese) wird zwar durch das Signifikanzniveau α beschränkt, aber die Möglichkeit ist doch existent. Außerdem ist zu beachten, dass der statistisch signifikante Effekt gar nicht klinisch relevant sein könnte. Bei hohen Fallzahlen können bereits kleinste Effekte statistisch signifikant werden. Kurz gesagt, statistische Signifikanz bedeutet **nicht** automatisch eine praktische Relevanz des positiven Testergebnis für den Anwender. Eine Verbindung zwischen Signifikanz und Relevanz wird erst durch eine Fallzahlberechung vor Durchführung des Tests bzw. durch eine detaillierte Betrachtung von Effektgröße, Fallzahl und verwendeter Testgröße möglich.

4.6 Versuchsplanung und Fallzahlberechnung

Wir haben gesehen, dass der kritische Wert, der den Ablehnungsbereich begrenzt, durch das Signifikanzniveau und die Wahrscheinlichkeitsverteilung unter der Nullhypothese festgelegt wird. Dieses Testkriterium wird auch dann angewendet, wenn in Wirklichkeit nicht die Nullhypothese H_0 gilt, sondern ein gewisser Effekt vorhanden ist und damit die Alternativhypothese H_A gilt.
Da die Testgröße ein Maß für den Effekt ist, ist dann die Verteilung der Testgröße relativ zur Verteilung unter H_0 verschoben. Bezeichnet H_A^{speziell} die Situation eines Effekts einer bestimmten Größe (im Fall eines Mittelwertvergleichs zweier Stichproben zum Beispiel etwa H_A^{speziell}: $\mu_1 - \mu_2 = 5,75$), dann sind die möglichen Fehlentscheidungen des Tests in folgender Grafik durch die schraffierten Flächen hervorgehoben (siehe Abb.: 4.6).

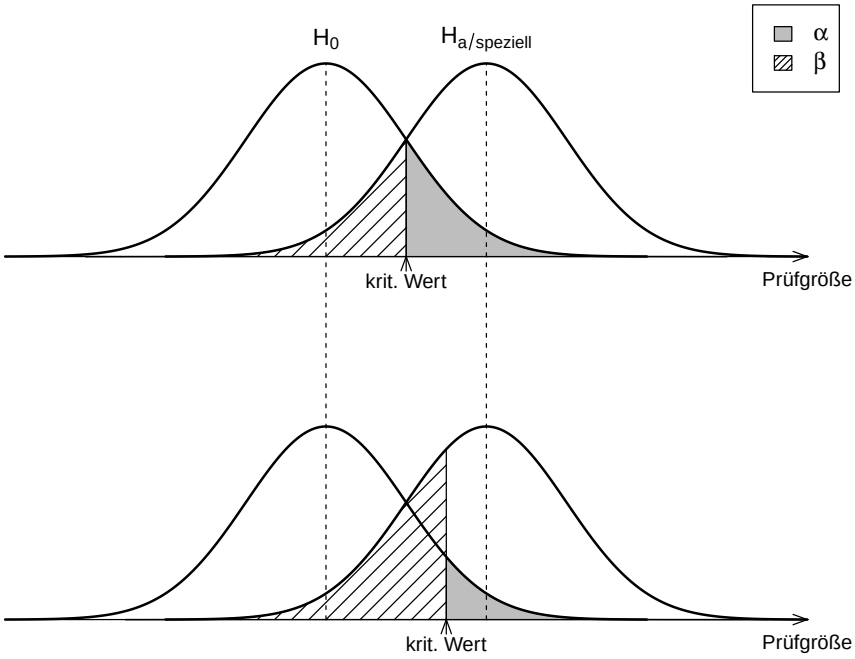


Abbildung 4.6: Verteilung der Prüfgröße

Wie man sieht, hängen die Fehler α und β gegenläufig voneinander ab. Wird der kritischen Wert nach rechts verschoben, wird α zwar kleiner, aber β wird größer. Die Kontrolle des Fehlers α ist einfach, da man die Verteilung der Testgröße unter der Nullhypothese kennt und den Ablehnungsbereich gerade gemäß α gewählt hat.
Die Wahrscheinlichkeit, einen vorhandenen Unterschied einer bestimmten Größe auch aufzudecken, ist die **Power** $1 - \beta$ eines Tests. Sie hängt nicht nur von der Effektgröße und dem gewählten Signifikanzniveau α ab, sondern wird auch von der untersuchten Fallzahl (Stichprobengröße) beeinflusst. Je größer die Stichprobe, desto steiler verlaufen die Kurven und desto kleiner wird β bei festem α.
Diese Informationen nutzt man bei der Planung von Studien. Das heißt, man kann im Vorfeld von Untersuchungen unter Kenntnis der genannten Größen den benötigten minimalen Stichprobenumfang ausrechnen, um eine vorgegebene Power zu erreichen. Ganz wesentlich dabei ist, welcher Unterschied Δ aufgedeckt werden soll, d.h. welche spezielle Alternativhypothese man zeigen will. Die konkrete Frage lautet, welcher Unterschied ist auch relevant für den Anwender und nicht, welcher Unterschied ist möglich. Im klinischen Kontext bedeutet dies: Zu kleine Studien decken selbst große vorhandene Unterschiede nicht auf, so dass therapeutische Effekte möglicherweise nicht erkannt werden. (Zu) Große Studien, können auch sehr kleine Unterschiede aufdenken, welche ggf. medizinisch irrelevant sein können.

4.7 Zusammenhang von Hypothesentesten und Konfidenzintervallschätzungen

Es besteht ein enger Zusammenhang zwischen Schätzen und Testen, den wir an unserem Beispiel verdeutlichen wollen:

Beispiel: 4.7.1

95% - Konfidenzintervall des Plasma - Magnesium - Spiegels

$$\bar{x}\, t_{f;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + \; t_{f;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

a) bei insulinpflichtigen Patienten

$$0,719 - 1,96 \cdot \frac{0,068}{\sqrt{227}} \leq \mu \leq 0,719 + 1,96 \cdot \frac{0,068}{\sqrt{227}}$$

$$0,71 \leq \mu \leq 0,727$$

b) Kontrollgruppe

$$0,81 - 1,96 \cdot \frac{0,057}{\sqrt{140}} \leq \mu \leq 0,81 + 1,96 \cdot \frac{0,057}{\sqrt{140}}$$

$$0,801 \leq \mu \leq 0,819$$

Die beiden Konfidenzintervalle überlappen sich nicht, so dass mit einer Sicherheit von 95% davon ausgegangen werden kann, dass eine signifikante Differenz zwischen den Gruppen existiert. Zusätzlich zu der Aussage über die Differenz der Gruppen liefern die Konfidenzintervalle Schätzungen für die wahren Mittelwerte in den Grundgesamtheiten der beiden Gruppen.
Beachte: Im Allgemeinen kann nicht von einer Umkehrung dieser Aussage ausgegangen werden, d.h. es ist möglich, dass sich die Konfidenzintervalle überlappen und gleichzeitig eine signifkante Differenz vorliegt.

4.8 Multiples Testen

Durch das Signifikanzniveau α wird das Risiko bei einem einzelnen Test begrenzt, einen Fehler 1. Art zu begehen, d.h. eine wahre Nullhypothese abzulehnen. Häufig werden aber zu einer Fragestellung simultan gleich mehrere Tests durchgeführt. Zum Beispiel wenn sehr viele Einflussvariablen erfasst werden, kann der Einfluss von all diesen Variablen einzeln getestet werden. Interpretiert man die einzelnen Testresultate wie im Fall eines einzelnen Tests, bleibt die Wahrscheinlichkeit, dass irgend einer der Tests einen Fehler 1. Art macht, nicht mehr durch α begrenzt.
Die Tabelle 4.4 zeigt die theoretischen Wahrscheinlichkeiten für mindestens ein falsch-signifikantes Ergebnis bei k unabhängigen Tests bei einem Signifikanzniveau der Einzeltests von α.

k	$P = 1 - 0,95^k$
1	0,05
2	0,10
3	0,14
5	0,23
10	0,40
20	0,64
50	0,92

Tabelle 4.4: theoretische Wahrscheinlichkeiten

Um dieses Problem in den Griff zu bekommen, kann man ein **Adjustierungsverfahren** anwenden. Bei nicht mehr als fünf simulatanen Tests bietet sich die Bonferroni-Korrektur an: Werden k Tests durchgeführt, wird bei den einzelnen Tests das Signifkanzniveau um den Faktor $\frac{1}{k}$ korrigiert. Statt α legt man also bei den Einzeltests das Signifikanzniveau $\frac{\alpha}{k}$ an. Dadurch wächst aber auch die Wahrscheinlichkeit β für den Fehler 2. Art und mehr echte Effekte werden übersehen. Insgesamt wird aber das Signifikanzniveau α sicher eingehalten. Ein solches Verfahren wird als konservativ bezeichnet. Eine Variante

zur Bonferroni-Korrektur ist die Methode von Holm, die etwas weniger konservativ ist und daher zu bevorzugen ist.
Wenn man eine sehr große Zahl von Tests parallel durchführen muss, werden andere Verfahren bevorzugt, weil die Bonferroni-Methode bei vielen Tests zu konservativ wird. Bekannt ist die Methode von Benjamini und Hochberg, bei der nicht mehr das gesamte Signifikanzniveau eingehalten wird, dafür aber der Anteil von falsch positiven Tests unter allen positiven Tests (die sogenannte false discovery rate FDR) unter einem vorgegebenem Anteil bleibt.

4.9 Ablaufschema eines statistischen Signifikanztests

An dieser Stelle wollen wir noch einmal den typischen Ablauf eines statistischen Tests skizzieren:

1. Formulierung der wissenschaftlichen Fragestellung
2. Aufstellen der Hypothesen H_0 und H_A
3. Wahl des Signifikanzniveaus α
4. Wahl des statistischen Tests und Prüfgröße gemäß
   - Ziel der Untersuchung
   - Merkmalsart der Zielgröße
   - Anzahl der Stichproben
   - Abhängige, unabhängige Stichproben
5. Erfassung der Stichproben
6. Prüfung der Testvoraussetzung
7. Ermittlung des Ablehnungs- und Annahmebereiches (kritischer Wert)
8. Berechnung der Testgröße und des p-Wertes
9. Testentscheidung
   - Vergleich der berechneten Testgröße mit kritischem Wert
   - Vergleich des p-Wertes mit vorgegebenem Signifikanzniveau α
10. Interpretation der Ergebnisse (mit Angabe des p-Wertes, der Fallzahl und des Testverfahrens)

4.10 Zusammenfassung

- Festlegung des Signifikanzniveaus α sollte eine inhaltliche Entscheidung sein („Welchen Fehler 1. Art nehme ich in Kauf?“)
- Testentscheidung:

Beibehalten von H_0		**Ablehnen von H_0**
Testgröße t liegt außerhalb des α-Ablehnungsbereichs		Testgröße t liegt innerhalb des α-Ablehnungsbereichs
$p > \alpha$		$p \leq \alpha$

- der p-Wert alleine genügt nicht für eine korrekte Interpretation des Testresultats
  - Der p-Wert ist **kein** Maß der Effektgröße
  - Der p-Wert hängt von Fallzahl und Testmethode ab
- Ein nicht signifikanter Unterschied heißt **nicht** unbedingt „Kein Unterschied vorhanden“: Ein vorhandener Unterschied könnte einfach auch zu klein sein als dass er mit der Stichprobengröße entdeckt werden könnte.
- Statistische Signifikanz ist **nicht** gleichbedeutend mit Relevanz
- Für Fallzahlberechnungen ist neben der Irrtumswahrscheinlichkeit α und der Power $1 - \beta$ der klinisch relevante Unterschied Δ anzugeben, der es verdient, aufgedeckt zu werden.

5 Spezielle Testverfahren

Im vorhergehenden Kapitel wurde allgemein das Prinzip des statistischen Signifikanztests erläutert. Der Mangel an konkreten Beispielen soll in diesem Kapitel behoben werden und einige der häufig benutzten klassischen statistischen Tests vorgestellt werden. Für jeden vorgestellten Test werden folgende Dinge angegeben:

1. die Fragestellung in Worten
2. die Hypothesen des Tests
3. die Voraussetzungen, die für die Durchführung des Test und die korrekte Interpretation des Testresultats erfüllt sein müssen
4. die Berechnungsvorschrift für die Testgröße und deren Verteilung unter der Nullhypothese H_0.
5. die Entscheidungsregel: wann wird die Nullhypothese des Tests abgelehnt.

5.1 Überblick

Es gibt eine ziemlich große Vielzahl an etablierten statistischen Tests. Folgende Aspekte können helfen, um den Überblick zu behalten:

Funktion des Tests

Was wird mit dem Test geprüft?

Lagetests machen Aussagen zur Lage, d.h. zum Mittelwert oder Median einer oder mehrerer Verteilungen. Wenn beispielsweise geprüft werden soll, ob ein Medikament im Mittel den Blutdruck senkt, ist ein Lagetest die richtige Wahl.

Häufigkeitstests geben Aufschluss über den Wert einer oder mehrerer Wahrscheinlichkeiten. Dabei kann es um eine oder mehrere Stichproben gehen. Häufigkeitstest sind weit verbreitet, weil Häufigkeiten bei jedem Merkmal (egal ob nominal, ordinal oder metrisch) erhoben werden können. So kann ein Häufigkeitstest eingesetzt werden, um zu prüfen, ob ein Würfel wirklich fair ist und alle sechs Augenzahlen mit gleicher Wahrscheinlichkeit auftreten. Ein weiteres Anwendungsgebiet für einen Häufigkeitstest wäre die Frage, ob der Anteil der Raucher bei Männern höher liegt als bei Frauen.

Anpassungstests gehen der Frage nach, ob Daten einer bestimmten Verteilung genügen. Solche Tests werden auch Goodness-of-fit-Tests genannt. Ein Einsatzgebiet ist etwa, dass die Daten in den Stichproben auf Normalverteilung getestet werden, bevor ein t-Test oder eine Varianzanalyse angewendet wird.

Anzahl der Stichproben

Wieviele Stichproben gehen in den Test ein?

Im Allgemeinen werden 1-*Stichprobentests*, 2-*Stichprobentests* und *Mehrstichprobentests* unterschieden. Dies hat seine Bedeutung bei der Auswahl des Tests und dessen Berechnung.

Art der Stichproben

Sind die Stichproben voneinander unabhängig?

Zwei oder mehr Stichproben können *abhängig* oder *unabhängig* voneinander sein. Eine Abhängigkeit liegt dann vor, wenn jeder Wert einer Stichprobe inhaltlich mit einem Wert der anderen Stichprobe zusammengehört. Diese Wertepaare kommen zum Beispiel dadurch zustande, dass ein Messwert vor und ein

Messwert nach einer Behandlung am selben Patienten erhoben wird. Abhängige Stichproben sind daher auch gleich groß, was im Allgemeinen bei unabhängigen Stichproben nicht der Fall sein muss.

Beispiel: 5.1.1

Abhängige Stichproben:

- Blutdruck:
  - 1. Messung vor Operation
  - 2. Messung nach Operation
- Visus:
  - linkes Auge
  - rechtes Auge

Unabhängige Stichproben:

- Ernährung
  - Messung von Cholesterin bei Vegetariern (Gruppe 1)
  - Messung von Cholesterin bei Nicht-Vegetariern (Gruppe 2)
- Vergleich des Geburtsgewichts bei:
  - männlichen Neugeborenen (Gruppe 1)
  - weiblichen Neugeborenen (Gruppe 2)

Testgröße

Welcher Verteilung folgt die Testgröße?

Die Verteilung der Testgröße prägt oft den Namen des statistischen Tests. Zum Beispiel wird häufig von einem t- oder F-Test gesprochen. Wird nicht mehr dazu erklärt, dann muss aus dem Kontext klar werden, welcher Test gemeint ist. So gibt es einen t-Test zum Mittelwertvergleich, aber auch einen zum Testen des Korrelationskoeffizienten.

Voraussetzungen

Welche Voraussetzungen stellt der Test an die Daten bezüglich ihrer Verteilung?

Tests können mehr oder weniger weitgehende Voraussetzungen an die Grundgesamtheiten stellen, aus denen die Stichproben gezogen wurden. Zum Beispiel sind viele Tests für normalverteilte Daten definiert. Ein Test, der eine bestimmte Verteilung der Merkmalswerte voraussetzt, wird auch *parametrischer Test* genannt. Das Gegenstück dazu bilden die *nicht-parametrischen Tests*.

5.2 Lagetests

Untersucht man eine metrische Messgröße in einer oder mehreren Gruppen, dann ist häufig die mittlere Lage (Erwartungswert) oder der Median in den Gruppen von besonderem Interesse. Zum Beispiel ist bei einem neuen Bluthochdruck-Medikament von großem Interesse, ob es tatsächlich *im Mittel* den Blutdruck senkt. Dazu vergleicht man die mittlere Lage des Blutdrucks in der Behandlungsgruppe mit dem in der Kontrollgruppe. Der Unterschied in den beiden Erwartungswerten der Gruppen ist ein Maß für den mittleren Effekt des Medikaments auf den Blutdruck (siehe Abbildung 5.1).

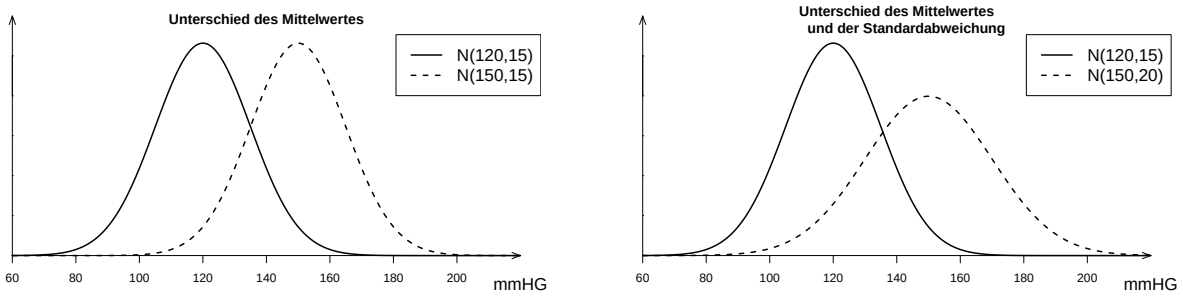


Abbildung 5.1: Vergleich von unterschiedlichen Normalverteilungen

Lagetests überprüfen genau diese Unterschiede des Erwartungswerts oder des Medians eines Merkmals in einer oder mehreren Gruppen. In diesem Abschnitt werden Tests für den Fall vorgestellt, dass eine Gruppe mit einem festen Sollwert oder zwei Gruppen miteinander verglichen werden. Des Weiteren wird ein Ausblick auf den Lagevergleich von mehr als zwei Gruppen gegeben.

5.2.1 t-Test: Lagetest bei Normalverteilung

Bei einer Normalverteilung der Daten fallen Median und Erwartungswert zusammen. Ein **t-Test** eignet sich bei Normalverteilung zum Vergleich der mittleren Lage bei ein oder zwei Stichproben. Dabei wird von der realistischen Situation ausgegangen, dass die Varianz in den Gruppen unbekannt ist und somit aus den Stichproben geschätzt werden muss.[1] Tabelle 5.1 gibt eine kurze Übersicht über die verschiedenen Arten von t-Tests.

Art des t-Tests	**Beispiel** Vergleich des mittleren Gewichts...
Eine Stichprobe	...bei Risikobabies mit einem Sollwert
Zwei verbundene Stichproben	...bei Patienten *vor* und *nach* einer Diät
Zwei unverbundene Stichproben	...in Behandlungsgruppe und Kontrollgruppe

Tabelle 5.1: Arten von t-Tests

Alle t-Tests haben einen gleichartigen Aufbau: Beispielhaft soll die Funktionsweise eines t-Tests in der Situation erklärt werden, wenn die Erwartungswerte aus zwei unverbundenen Gruppen x und y miteinander verglichen werden sollen. Als konkrete Situation könnte es um den Blutdruck in einer Behandlungs- und einer Kontrollgruppe gehen. Intuitiv ist klar, dass der Mittelwert $\bar{x}$ der Stichprobe x_i die beste Schätzung für den Erwartungswert μ_x der entsprechenden Gruppe ist. Gleiches gilt für $\bar{y}$ und μ_y. Falls die beiden Gruppen außer bezüglich der Behandlung gleich sind (etwa nach einer Randomisierung der Patienten auf die Gruppen), entspricht der Unterschied der Mittelwerte $\mu_y - \mu_x$ dem Effekt, der durch die Behandlung erreicht worden ist. Liegt der Mittelwert $\bar{y}$ „weit“ vom Mittelwert $\bar{x}$ entfernt, d.h. ist der beobachtete Effekt $\bar{y} - \bar{x}$ „groß“, so ist es wenig plausibel, dass die beiden Erwartungswerte μ_x und μ_y tatsächlich übereinstimmen. Die Frage ist nur, ab wann der beobachtete Effekt nicht mehr als Zufallsbefund durchgehen sollte. Es wird davon abhängen, wie genau die Mittelwertschätzungen $\bar{x}$ und $\bar{y}$ aus den beiden Stichproben sind.

[1] Falls die wahre Varianz bekannt ist und *nicht* aus den Daten geschätzt werden muss, ist der **Gauß-Test** anwendbar, der dem t-Test sehr ähnlich ist, aber als Referenz für die Testgröße statt der t-Verteilung die Normalverteilung verwendet.

Die Testgröße t des t-Tests ist darum eine *normierte* Effektgröße: der beobachtete Effekt $\bar{y} - \bar{x}$ wird durch den Standardfehler des Effekts geteilt. Ein t-Wert nahe Null bedeutet, dass der Unterschied [=Effekt] nicht groß ist und plausibel durch Zufall erklärt werden kann. Je weiter t von Null entfernt ist, desto größer ist der normierte Effekt. Das Vorzeichen von t bezeichnet die Richtung des Effekts.
Wie der Name der t-Tests nahelegt, folgt die Testgröße unter der Nullhypothese („kein Effekt“) einer t-Verteilung. Um die t-Verteilung genau festzulegen, muss ein Parameter mit angegeben werden, die sogenannten **Freiheitsgrade**[2]. Auf Abbildung 5.2 sieht man, dass die t-Verteilung symmetrisch um 0 herum verteilt ist. Im Vergleich zur Normalverteilung hat sie mehr Masse an den Rändern, aber mit zunehmenden Freiheitsgraden nähert sie sich der Normalverteilung an.
Die Freiheitsgrade, die für die t-Verteilung der Testgröße unter der Nullhypothese benötigt werden, hängen vom Stichprobenumfang der Stichprobe(n) ab.

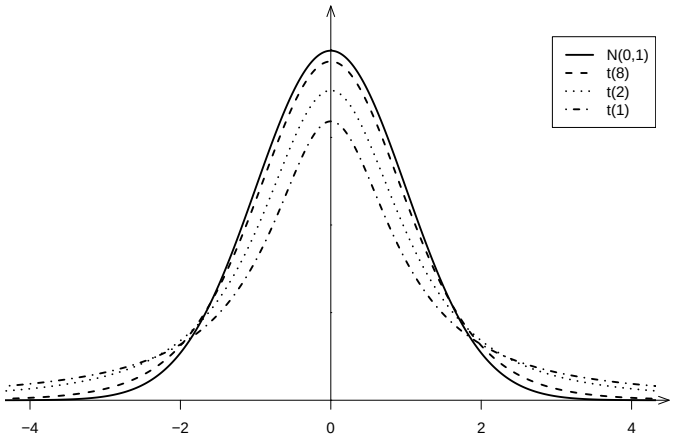


Abbildung 5.2: Vergleich von t-Verteilung mit verschiedenen Freiheitsgraden(1,2,8) mit Standardnormalverteilung

t-Test für eine Stichprobe

Wird ein normalverteiltes Merkmal in einer Gruppe anhand einer Stichprobe x_i untersucht, kann der Erwartungswert des Merkmals mittels eines 1-Stichproben t-Tests mit einem festen Wert μ_0 verglichen werden.

Fragestellung. Unterscheidet sich der Erwartungswert in einer Gruppe von dem festen Wert μ_0?

Hypothesen. Bei zweiseitiger Fragestellung stellt man die Nullhypothese $H_0 : \mu = \mu_0$ der Alternativhypothese $H_A : \mu \neq \mu_0$ gegenüber.[3]

Vorraussetzungen. Die Stichprobe x_i sollte einer normalverteilten Grundgesamtheit $\mathcal{N}(\mu, \sigma^2)$ entstammt. Prinzipiell reicht es für den Test, wenn das Stichprobenmittel einer Normalverteilung folgt. Bei größeren Stichproben (etwa $n \geq 25$) ist der Test aufgrund des zentralen Grenzwertsatzes (Kap. 3.2.2) auch dann durchführbar, wenn die Messwerte nicht normalverteilt sind. In der Praxis hat sich gezeigt, dass es bei Stichproben des Umfangs $n \geq 10$ genügt, wenn das Merkmal annähernd symmetrisch verteilt ist.

Testgröße. Der Unterschied zwischen Mittelwert in der Stichprobe $\bar{x}$ und dem Sollwert μ_0 wird zum Standardfehler des Mittelwerts ins Verhältnis gesetzt. Der Standardfehler des Mittelwerts ist ein Maß dafür, wie präzise die Mittelwertschätzung ist.

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} \quad \text{mit} \quad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.1}$$

Hierbei ist s^2 die empirische Varianz, die sich aus der Stichprobe errechnet. $\sqrt{\frac{s^2}{n}}$ ist der Standardfehler des Mittelwerts (siehe auch 3.1.2).

[2] Dies ist keine Besonderheit der t-Verteilung. Freiheitsgrade gibt es auch bei anderen Verteilungen wie der F- oder der χ^2-Verteilung.
[3] Die Alternativhypothese bei einseitiger Fragestellung lautet $H_A : \mu > \mu_0$ bzw. $H_A : \mu < \mu_0$ bei gleicher Nullhypothese.

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert[4], der durch eine Statistiksoftware errechnet wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.[5]

Im Kapitel zum Schätzen wurde das Konfidenzintervall für den Mittelwert angegeben(siehe (3.2.1)). Aus dem Konfidenzintervall kann die Testentscheidung abgelesen werden: Der Test lehnt die Nullhypothese genau dann ab, wenn μ_0 außerhalb dieses Intervalls liegt. Darüber hinaus liefert das Konfidenzintervall neben der Testentscheidung auch einen Hinweis darauf, wie genau die Mittelwertschätzung aufgrund der Stichprobe ist (Deswegen wird in der Situation, nur einer Stichprobe, das Konfidenzintervall dem t-Test vorgezogen).

Beispiel: 5.2.1

Vergleich des Gewichts von Neugeborenen nach einer Schwangerschaft unter Nikotinabusus mit dem mittleren Geburtsgewicht von 3300g.

Von 20 Neugeborenen nach einer Schwangerschaft unter Nikotinabusus liegt das Geburtsgewicht vor:

$$x_1, \ldots, x_{20} = (2281, 3083, 3183, 3343, 3077, 2829, 2749, 3687, 2810, 3651,\\ 2869, 3376, 1557, 3386, 3353, 2482, 3013, 2726, 1997, 2478)$$

Aus der Stichprobe ($n = 20$) berechnet sich ein mittleres Geburtsgewicht von $\bar{x} = 2896,5g$ und eine empirische Standardabweichung $s = 544,22g$.

Fragestellung: Senkt das Rauchen während der Schwangerschaft das mittlere Gewicht von Neugeborenen? Dies ist eine einseitige Fragestellung, d.h. die Richtung der Änderung (hier Senkung) ist vorgegeben.

Hypothesen: $H_0 : \mu = 3300$ versus $H_A : \mu < 3300$

Testgröße: Der 1-Stichproben t-Test soll zeigen, ob die Daten für die Gewichtsreduktion sprechen oder nicht.

$$t = \frac{\bar{x} - 3300}{\sqrt{\frac{s^2}{n}}} = \frac{2896,5 - 3300}{\frac{544,22}{\sqrt{20}}} = -3,32$$

Testentscheidung:
Bei einseitiger Fragestellung sollte man zunächst sicherstellen, dass „die Richtung stimmt“, d.h. in diesem Beispiel, dass das mittlere Geburtsgewicht in der Stichprobe tatsächlich *geringer* ist als 3300g. In der vorliegenden Stichprobe kommt es tatsächlich zu einer Senkung des Geburtsgewichts. Ob diese Senkung auch statistisch signifikant ist, ergibt sich durch den p-Wert, der sich aus der errechneten Testgröße ergibt. Im Beispiel ist der p-Wert $p = 0,00182$, und damit kleiner als das vorgegebene Signifikanzniveau $\alpha = 0,05$. Daher wird die Nullhypothese auf dem 5%-Signifikanzniveau abgelehnt. Das Geburtsgewicht von Kindern nach einer Schwangerschaft unter Nikotinabusus ist statistisch signifikant geringer als das mittlere Geburtsgewicht von 3300g.
Mittels der Stichprobe lässt sich auch die Obergrenze des einseitigen 95%-Konfidenzintervalls zum mittleren Geburtsgewicht nach Nikotinabusus in der Schwangerschaft abschätzen. Dazu benötigt man das 95%-Quantil der t-Verteilung mit $20 - 1 = 19$ Freiheitsgraden, $t_{0,95;19} = 1.729$, das man mittels Statistik-Software oder Quantil-Tabellen auslesen kann.

[4] Der p-Wert ist die Wahrscheinlichkeit unter gültiger Nullhypothese H_0, dass die Testgröße so wie in der Stichprobe errechnet oder noch weiter von Null entfernt ist („noch extremer“). Siehe Kapitel 4.

[5] Auch ohne p-Wert kann man zu einer Testentscheidung kommen: Unter H_0 folgt die Testgröße einer t-Verteilung mit $n - 1$ Freiheitsgraden. Die Untergrenze des Ablehnungsbereichs, d.h. der Bereich, in dem der Betrag der Testgröße zum Ablehnen der Nullhypothese führt, wird durch einen passenden Quantil-Wert einer t-Verteilung markiert. Ist der Betrag der errechneten Testgröße größer als dieser Quantil-Wert, wird die Nullhypothese abgelehnt. Diese Quantil-Werte können in der Regel in Statistikbüchern nachgeschlagen werden.

Fortsetzung des Beispiels

Das einseitige 95%-Konfidenzintervall ist nach oben durch

$$2896,5g + 1,729 \cdot \frac{544,22}{\sqrt{20}} = 3106,92g$$

begrenzt. Das signifikante Testergebnis zeigt sich beim Konfidenzintervall dadurch, dass der Wert 3300g größer als die Obergrenze ist und damit nicht im 95%-Konfidenzintervall enthalten ist.

Im Folgenden werden t-Tests vorgestellt, die Mittelwerte zweier Stichproben miteinander vergleichen. Konzeptuell ist wichtig zu unterscheiden, ob die beiden Stichproben unabhängig voneinander sind („unverbunden“) oder nicht („verbunden“).

t-Test für zwei verbundene Stichproben

Bei verbundenen Messungen können auf individuellem Niveau die Differenzen der Messwerte gebildet werden. Der t-Test für zwei verbundene Stichproben prüft, ob diese Differenzen im Mittel von Null verschieden sind, und ist damit rechnerisch dem 1-Stichproben t-Test sehr ähnlich.

Fragestellung. Unterscheiden sich die Erwartungswerte des Merkmals in beiden Gruppen, unter Berücksichtigung der Abhängigkeit der Messwerte?

Hypothesen. Die Nullhypothese besagt, dass im Mittel kein Unterschied in beiden Gruppen auftritt, also $H_0 : \mu_x = \mu_y$. Bei zweiseitiger Fragestellung ist die Alternativhypothese, dass es sehr wohl einen mittleren Unterschied gibt, also $H_A : \mu_x \neq \mu_y$. Sehr häufig werden beim t-Test für verbundene Stichproben die Hypothesen auch über den mittleren Unterschied ausgedrückt, weil gerade die individuellen Unterschiede bei der Berechnung dieses Tests die wesentliche Rolle spielen. Eine äquivalente Formulierung ist also (hier wieder bei zweiseitiger Fragestellung):

$$H_0 : \delta = 0 \text{ versus } H_A : \delta \neq 0, \quad \text{wobei } \delta = \mu_x - \mu_y$$

Vorraussetzungen. Betrachtet werden zwei *verbundene* Stichproben (Umfang n) mit Wertepaaren (x_i, y_i) aus den Grundgesamtheiten mit den Erwartungswerten μ_x bzw. μ_y, wobei die Differenzen $d_i = x_i - y_i$ **normalverteilt** sein sollten mit Erwartungswert $\delta = \mu_x - \mu_y$. Diese Normalverteilungs-Voraussetzung kann wie beim 1-Stichproben t-Test abgeschwächt werden. Für Stichproben mit $n \geq 10$ reicht es aus, wenn die Differenzen annähernd symmetrisch verteilt sind.

Testgröße. Wegen der Verbundenheit der Stichproben, lässt sich für jedes Beobachtungspaar aus beiden Gruppen die Differenz bilden. Die Testgröße prüft, ob diese Differenzen von Null verschieden sind und leitet sich analog wie beim 1-Stichproben t-Test her:

$$t = \frac{\overline{d}}{\sqrt{\frac{s_d^2}{n}}} \text{ mit } s_d^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (d_i - \bar{d})^2 \tag{5.2}$$

Hierbei ist s_d^2 die empirische Varianz der Differenzen d_i und $\sqrt{\frac{s_d^2}{n}}$ der Standardfehler der mittleren Differenz.

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware errechnet wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

Beispiel: 5.2.2

Anhand einer Stichprobe von n = 40 Patienten soll die blutdrucksenkende Wirkung eines Medikaments beurteilt werden. Dazu wurde bei diesen Patienten der Blutdruck $[mmHg]$ vor und nach Therapie bestimmt. Folgende Werte zur Beschreibung der Blutdruckdifferenzen wurden ermittelt:

$$\bar{d} = 9,275$$
$$s_d = 12,914$$

Prüfen Sie auf einem Signifikanzniveau von $\alpha = 5\%$, ob das Medikament im Mittel zu einer Veränderung des systolischen Blutdrucks führt!

Hypothesen. Dies ist eine zweiseitige Fragestellung. Die Null- und Alternativhypothesen lauten

$H_0 : \delta = 0$. Das Medikament führt in der Population im Mittel zu keiner Änderung des Blutdrucks.

$H_A : \delta \neq 0$. Das Medikament führt in der Population im Mittel zu einer Blutdruckänderung.

Testgröße.

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = \frac{\bar{d}}{s_d}\sqrt{n} = \frac{9,275}{12,914}\sqrt{40} = 4,54$$

Testentscheidung. Zur Testgröße $t = 4,54$ gehört der p-Wert $p = 0,0000528$. Da $p < \alpha = 0,05$, wird die Nullhypothese auf dem 5%-Signifikanzniveau abgelehnt.

Konkret heißt das, dass das Medikament in der Population im Mittel zu einer statistisch signifikanten Blutdruckänderung führt. Im Beispiel liegt eine Blutdrucksenkung vor, und die Absenkung kann mit 9,275mmHg abgeschätzt werden.

Auch bei zwei unverbundenen Stichproben kann man den Mittelwertvergleich mit einem geeigneten t-Test angehen. Es wird der Fall unterschieden, ob die Varianz in beiden Gruppen gleich ist oder nicht.

t-Test für zwei unabhängige Stichproben bei ungleicher Varianz (Welch-Test)

Fragestellung. Unterscheiden sich die Erwartungswerte eines Merkmals in zwei Gruppen? (zweiseitige Fragestellung)

Hypothesen.

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 \neq \mu_2 \text{ bei zweiseitiger Fragestellung.}$$

Vorraussetzungen. Die beiden unabhängigen Stichproben sollten **normalverteilten** Grundgesamtheiten entstammen, deren Varianzen aber nicht gleich sein müssen, d.h.: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Testgröße. Die Testgröße berechnet sich als die Differenz der Mittelwerte beider Stichproben, geteilt durch den Standardfehler dieser Mittelwertdifferenz. Dieser Standardfehler ist ein Maß dafür, wie gut dieser Mittelwertunterschied aus den vorliegenden Daten geschätzt werden konnte.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.3}$$

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware errechnet wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

Beispiel: 5.2.3

Im Rahmen einer Studie zu Lipidsenkern soll die Wirksamkeit eines neuen Medikaments im Vergleich zu einem etablierten Medikament geprüft werden. Als Zielgröße wird der Gesamtcholesterinspiegel in den Patienten gemessen. Dazu werden 23 zufällig ausgewählte Infarktpatienten mit dem neuen Medikament behandelt. Als Kontrollgruppe dienen weitere 20 zufällig ausgewählte Infarktpatienten, die nach der Standardtherapie behandelt werden.

In den Stichproben der Kontrollgruppe x und der Behandlungsgruppe y werden folgende Gesamtcholesterinspiegel-Werte ermittelt: $x_1 ... x_{20} = (277, 337, 329, 293, 253, 286, 256, 287, 303, 271, 338, 286, 302, 310, 339, 286, 273, 274, 294, 262)$
$y_1 ... y_{23} = (254, 179, 303, 201, 300, 208, 377, 251, 238, 307, 250, 249, 297, 269, 267, 254, 301, 246, 207, 247, 256, 224, 238)$

	n	$\bar{x}$	s
Kontrollgruppe	20	292,8	26,61
Behandlungsgruppe	23	257,52	43,15

Tabelle 5.2: Gruppenvergleich

Fragestellung: Unterscheidet sich der Erwartungswert der Behandlungsgruppe von dem der Kontrollgruppe ?
Hypothesen:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

Testgröße:

$$t = \frac{292,8 - 257,52}{\sqrt{\frac{26,61^2}{20} + \frac{43,15^2}{23}}} = 3,27$$

Testentscheidung Aus der Testgröße ($t = 3.27$) ergibt sich ein p-Wert $p = 0,0023 < \alpha = 0.05$. Somit darf die Nullhypothese auf dem 5%-Signifikanzniveau abgelehnt werden.

Liegen gleiche Varianzen in beiden Gruppen vor, vereinfacht sich die Berechnung der Testgröße (*t-Test für zwei Stichproben bei gleicher Varianz*). Auf diese Weise wird außerdem eine geringfügig höhere Power als beim Welch-Test erzielt (d.h. in diesem Fall findet er ein vorhandenen Unterschied häufiger).

Vergleich der t-Tests bei verbundenen und unverbundenen Stichproben

Sowohl der t-Test für zwei verbundene Stichproben als auch der t-Test für zwei unverbundene Stichproben vergleichen den Erwartungswert eines Merkmals in zwei Gruppen. Die Fragestellung ist also gleich. Der Unterschied zwischen beiden Tests ist, dass bei unverbundenen Stichproben die mittlere Differenz untersucht wird, während bei verbundenen Stichproben die „intra-individuellen“ Unterschiede betrachtet werden. Damit wird beim Test für verbundene Stichproben eine Varianzkomponente – die Variabilität zwischen den „Individuen“ – eliminiert und somit kann es sein, dass dieser Test Unterschiede aufdeckt, die der Test unter Vernachlässigung der Verbundenheit der Stichproben nicht findet.

Beispiel: 5.2.4

Vergleich einer Gruppe von sechs Patienten bezüglich eines Messwertes vor und nach einer Therapie.

Patient	I	II	III	IV	V	VI
Vor Therapie	18	20	21	23	17	22
Nach Therapie	23	23	21	23	20	21

Tabelle 5.3: Gruppenvergleich vor und nach der Therapie

Fortsetzung des Beispiels

Im Mittel steigt der Messwert von 20 *vor* Therapie auf 22 *nach* Therapie. Der mittlere Unterschied beträgt also 2. Ist dieser Unterschied auf dem 5%-Signifikanzniveau auch statistisch signifikant?
Ignoriert man die Paarung der Daten kann man einen Welch-Test durchführen. Der gepaarte t-Test berücksichtigt die „intra-individuelle“ Variabilität der verschiedenen Patienten. Die Resultate der beiden Tests sehen folgendermaßen aus:

	Welch-Test	gepaarter t-Test
Testgröße	-1,9365	-2,9277
Freiheitsgrade	8	5
p-Wert	0,0889	0,0327
Ergebnis	Nicht signifikant	Signifikant

Tabelle 5.4: Testvergleich

In diesem Beispiel führt die Vernachlässigung der Bindung in den Daten zu einem Verlust an „Power“, d.h. der vorhandene mittlere Unterschied wird mit dem Welch-Test nicht als statistisch signifikant nachgewiesen.

5.2.2 Nicht-parametrische Lagetests

Nicht alle Daten sind normalverteilt. und es gibt viele Situationen bei denen selbst die oben erwähnten abgeschwächten Voraussetzungen nicht erfüllt sind. Ein t-Test sollte hier nicht angewandt werden. Allerdings helfen nicht-parametrischen Tests weiter.
Die wichtigste Gruppe nicht-parametrischer Lagetests sind die **Rangsummentests**. Die Grundidee bei einem Rangsummentest ist, statt der eigentlichen Messwerte die Rangfolge (die *Ränge*) der Messwerte zu verwenden. Sie machen dabei Aussagen über die mittlere Lage der Verteilung der Messgröße in der Gruppe bzw. den Gruppen. Bei zwei Gruppen etwa geht es anschaulich um die Frage, ob eine Gruppe zu größeren Werten tendiert als die andere Gruppe. Konkret kann man die Aussage des Tests auf den Median $\tilde{\mu}$ der Messgröße beziehen, da er ein Maß für die mittlere Lage einer Verteilung ist. Bei nicht-symmetrischen Verteilungen fällt der Median im Allgemeinen nicht mit dem arithmetischen Mittel zusammen. Das heißt, dass in diesen Fällen auch die Fragestellung eines t-Tests und eines Rangsummentests unterschiedlich ist. Da nur die Rangfolge in den Test eingeht, können Rangsummentests neben metrisch skalierten Daten auch ordinal skalierte Daten untersuchen. Es ist klar, dass Ausreißer in der Stichprobe dadurch weniger starken Einfluss auf das Testergebnis haben.
Eine andere Familie nicht-parametrischer Lagetests sind die **Vorzeichentests**, die noch weniger Voraussetzungen an die Daten stellen als die Rangsummentests. Allerdings nutzen diese Tests auch noch weniger Informationen aus den Daten als die Rangsummentests, und haben daher eine geringere Power, um vorhandene Unterschiede als statistisch signifikant nachzuweisen. Vorzeichentests werden in diesem Kapitel nicht besprochen.

Wilcoxon-Test für eine Stichprobe

Der einfachste Rangsummentest befasst sich mit nur einer Stichprobe. Wir stellen hier die zweiseitige Fragestellung dar.

Fragestellung. Weicht der Median $\tilde{\mu}$ eines Merkmals von einem festen Wert $\tilde{\mu}_0$ ab?

Hypothesen. $H_0 : \tilde{\mu} = \tilde{\mu}_0$ versus $H_A : \tilde{\mu} \neq \tilde{\mu}_0$

Voraussetzungen. Der Test setzt eine symmetrische Verteilung der Messgröße in der Grundgesamtheit voraus. Auch kleine Stichproben sind zulässig. Ist die Symmetrie nicht gewährleistet, bietet sich der Vorzeichentest für eine Stichprobe als Alternative an.

Testgröße. Zunächst werden die Differenzen der n Stichprobenwerte mit dem festen Wert $\tilde{\mu}_0$ gebildet. Dann werden die Rangzahlen nach dem Betrag der Differenz vergeben, wobei die betragsmäßig kleinste Differenz den Rang 1 erhält. Werte aus der Stichprobe, die mit dem Sollwert übereinstimmen, werden nicht berücksichtigt und zählen bei den Freiheitsgraden auch nicht mit. Bei einer Rangbindung, d.h. treten mehr als eine Differenz vom gleichen Betrag auf, werden die Ränge gemittelt (siehe Beispiel 5.7). Die Summe der Ränge der positiven Differenzen R^+ und der negativen Differenzen R^- werden gebildet, wobei die kleinere der beiden Zahlen $R = \min(R^-, R^+)$ die Testgröße ist.

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware errechnet wird. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

Anders als bei den t-Tests sprechen hier *kleine* Werte der Teststatistik gegen die Nullhypothese: Je kleiner der Betrag der Testgröße, desto kleiner der p-Wert.

Beispiel: 5.2.5

Bei einer Stichprobe von 20 Patienten wurde der systolische Blutdruck ermittelt. Die mittlere Lage soll mit dem festen Referenzwert $\mu_0 = 120$mmHg verglichen werden. Von Normalverteilung der Blutdruckwerte kann nicht ausgegangen werden.

i	x_i	$d_i = x_i - 120$	Rangzahl für $d_i > 0$	Rangzahl für $d_i < 0$
1	155	35	17	
2	117	-3		2
3	133	13	6	
4	120	0	–	–
5	105	-15		8
6	178	58	18	
7	151	31	14	
8	114	-6		4,5
9	119	-1		1
10	137	17	10,5	
11	136	16	9	
12	143	23	13	
13	152	32	15,5	
14	139	19	12	
15	106	-14		7
16	116	-4		3
17	120	0	–	–
18	126	6	4,5	
19	137	17	10,5	
20	152	32	15,5	
			$R^+ = 145,5$	$R^- = 25,5$

Tabelle 5.5: systolischer Blutdruck bei 20 Patienten

Fortsetzung des Beispiels

Die Testgröße R ist die kleinere der beiden Rangsummen, hier also $R = \min(R^-, R^+) = 25, 5$.

Der aus der Testgröße resultierende p-Wert von $p = 0.0095$ führt auf dem 5% Signifikanzniveau zum Ablehnen der Nullhypothese. Somit wird geschlussfolgert, dass sich der Median der Blutdruckmessungen in der Grundgesamtheit signifikant vom Vergleichswert (120mmHg) unterscheidet.

Wilcoxon-Test für zwei verbundene Stichprobe

Dieser Test ist die nicht-parametrische Entsprechung des t-Tests bei zwei verbundenen Stichproben. Wie bei den t-Tests ist der Wilcoxon-Test für zwei verbundene Stichproben dem Wilcoxon-Test für eine Stichprobe sehr ähnlich. Wir behandeln hier den Test bei zweiseitiger Fragestellung.

Fragestellung. Sind die Mediane einer Messgröße in zwei verbundenen Messungen verschieden?

Hypthesen. Bei zweiseitiger Fragestellung formuliert man $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ versus $H_A : \tilde{\mu}_1 \neq \tilde{\mu}_2$.

Vorraussetzungen. Die Verteilung der Messgröße in der Vorher- bzw. Nachher-Messung unterliegt keiner Voraussetzung. Allerdings sollten die Differenzen der Merkmalspaare symmetrisch verteilt sein. Oft ist dies gegeben, wenn die Messgröße zu beiden Messungen in etwa gleiche Verteilungsformen aufweist.

Testgröße. Die Differenzen eines jeden Merkmalspaares aus den verbundenen Stichproben werden nach der Größe ihres Betrags geordnet. Paare mit gleichem Messwert werden außer Betracht gelassen. Das Paar, mit den Messwerten, die am dichtesten aneinanderliegen, bekommt Rang 1 usw. Rangbindungen (d.h. mehrere Paare haben die gleiche Differenz zwischen den Messwerten) führen zur Mittelung der Ränge. Die Rangsumme der positiven und negativen Differenzen werden separat aufaddiert, und die kleinere der beiden Summen ist die Testgröße.

Testentscheidung. Die Nullhypothese wird auf dem Signifikanzniveau α abgelehnt, falls der zur Testgröße gehörende p-Wert kleiner als α ist. Je kleiner die Werte der Testgröße, desto kleiner der zugehörige p-Wert.

Beispiel: 5.2.6

Auf dem Signifikanzniveau $\alpha = 5\%$ soll geprüft werden, ob bei weiblichen Infarktpatienten der Blutdruck nach Behandlung (PSYSN) im Mittel niedriger ist als vorher (PSYS). Die Differenzen der Blutdruckwerte PSYS und PSYSN sind nicht normalverteilt.
In der folgenden Tabelle kommt es zu Rangbindungen. Fünf Differenzen haben einen Betrag von 5. Daher werden die ersten fünf Ränge gemittelt. $\frac{1+2+3+4+5}{5} = 3$. Für die anderen Ränge wird analog vorgegangen.

Fortsetzung des Beispiels

Patient (i)	PSYS (x_i)	PSYSN (y_i)	$d_i = x_i - y_i$	Rang von d_i für $d_i \neq 0$
1	130	120	10	7
2	140	135	5	3
3	160	150	10	7
4	165	165	0	–
5	165	160	5	3
6	170	170	0	–
7	170	175	-5	**3**
8	175	165	10	7
9	175	175	0	–
10	175	170	5	3
11	180	185	-5	**3**
12	190	160	30	9

Tabelle 5.6: Blutdruckbehandlung bei weiblichen Infarktpatienten

In diesem Beispiel liegt eine einseitigen Fragestellung vor. Die Testgröße wird wie beim zweiseitigen Test berechnet. Allerdings ändert sich der zugehöirge p-Wert.

Hypothesen: $H_0 : \tilde{\mu}_D \leq 0$
Das Medikament führt in der betrachteten Population weiblicher Infarktpatienten im Median zu keiner Blutdrucksenkung.
$H_A : \tilde{\mu}_D > 0$
Das Medikament führt in der Population im Median zu einer Blutdrucksenkung. Der Median der Differenzen ist positiv, was einer medianen Senkung des Blutdrucks entspricht. Ob die Senkung signifikant ist, soll der Wilcoxon-Test ergeben.

Testgröße:

$$\begin{aligned} R_+ &= 7+3+7+3+7+3+9 = 39 \\ R_- &= 3+3 = 6 \\ \min(R_+, R_-) &= 6 \end{aligned}$$

Testentscheidung: Die Nullhypothese wird auf dem 5% Signifikanzniveau abgelehnt, da der zur Testgröße R gehörende p-Wert von $p = 0,0264 < \alpha = 0,05$ ist. Das Medikament führt bei weiblichen Infarktpatienten mit einer Irrtumswahrscheinlichkeit von 5% im Median zu einer signifikanten Blutdrucksenkung.

U-Test nach Wilcoxon, Mann and Whitney

Dieser Rangsummentest stellt ein nicht-parametrisches Pendant zum 2-Stichproben t-Test dar. Wir stellen hier die zweiseitige Fragestellung dar.

Fragestellung. Unterscheiden sich die Mediane einer Messgröße in zwei unabhängigen Gruppen?

Hypothese. Entsprechend lauten die Hypothesen (bei zweiseitiger Fragestellung) folgendermaßen:

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 \text{ versus } H_A : \tilde{\mu}_1 \neq \tilde{\mu}_2.$$

Vorraussetzungen. Die beiden Stichproben sollten aus Verteilungen entstammen, die ungefähr gleiche Form aufweisen. Weder Normalverteilung noch Symmetrie werden vorausgesetzt (für eine Illustration siehe Abb. 5.3).

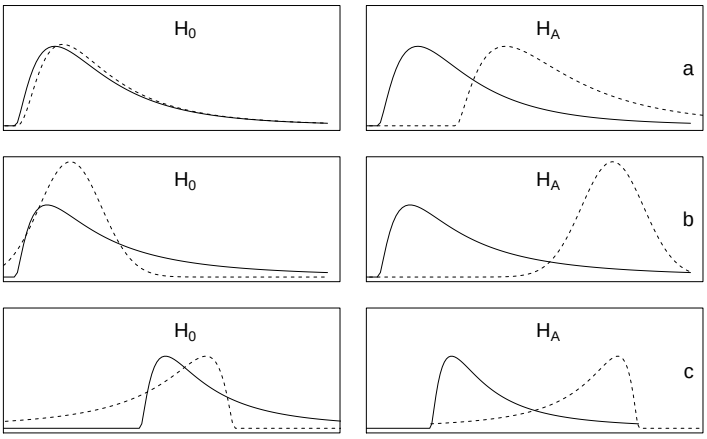


Abbildung 5.3: a: erfüllt Voraussetzung; b und c: erfüllen die Voraussetzungen nicht

Testgröße. Alle Werte aus den beiden Stichproben werden in aufsteigender Reihenfolge sortiert und mit Rangzahlen versehen. Durch Addition der Rangzahlen separat für Stichprobe 1 und Stichprobe 2 ergeben sich die Rangsummen R_1 und R_2. Aus diesen beiden Zahlen lassen sich U_1 und U_2 folgendermaßen berechnen:

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \text{und} \quad U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{5.4}$$

Hat man richtig gerechnet, gilt $U_1 + U_2 = n_1 \cdot n_2$. Dies kann zur Kontrolle herangezogen werden. Die Testgröße U ist das Minimum aus U_1 und U_2:

$$U = \min(U_1, U_2) \tag{5.5}$$

Testentscheidung. Die Nullhypothese wird auf dem Signifikanzniveau α abgelehnt, falls der zur Testgröße gehörende p-Wert kleiner als α ist. Bei diesem Test sprechen kleine Werte der Testgröße gegen die Nullhypothese: Je kleiner die Werte der Testgröße, desto kleiner der zugehörige p-Wert.

Beispiel: 5.2.7

Das Beispiel bezieht sich auf eine Myokard-Infarktstudie. Bei Frauen wurden folgende Cholesterinwerte $\left[\frac{mg}{dl}\right]$ erfaßt.
Kontrollgruppe (m=15):

Chol	140	145	150	160	170	175	175	180	180	185	190	190	195	200	205
Rang	1	2	3	4	5	6,5	6,5	8,5	8,5	11	14	14	16	17,5	19

Tabelle 5.7: Cholesterinwerte der Kontrollgruppe

Infarktpatienten (n=12):

Chol	185	185	190	200	210	210	210	290	295	380	395	485
Rang	11	11	14	17,5	21	21	21	23	24	25	26	27

Tabelle 5.8: Cholesterinwerte der Infarktgruppe

Prüfen Sie mit einer Irrtumswahrscheinlichkeit von 5%, ob sich die mittleren Cholesterinwerte der Infarktpatienten und der Kontrollgruppe unterscheiden! Normalverteilung in den Gruppen kann nicht vorausgesetzt werden.

Lösung:

H_0: Der mediane Cholesterinwert der Infarktpatientinnen stimmt mit dem der Kontrollen überein.

H_A: Die medianen Cholesterinwerte der Infarktpatientinnen und der Kontrollen unterscheiden sich.

$$\begin{aligned} R_1 &= 1+2+3+4+5+6,5+6,5+8,5+8,5+11+14+14+16+17,5+19 \\ &= 136,5 \\ R_2 &= 11+11+14+17,5+21+21+21+23+24+25+26+27 = 241,5 \\ U_1 &= 12\cdot 15+\frac{15\cdot 16}{2}-R_1 = 180+120-136,5 = 163,5 \\ U_2 &= 12\cdot 15+\frac{12\cdot 13}{2}-R_2 = 180+78-241,5 = 16,5 \\ U &= \min(U_1,U_2) = 16,5 < 50 = U_{2,5\%;15;12} \\ p &< 0,0004(errechnet durch Statistik-Software) \end{aligned}$$

Auf dem 5% Signifikanzniveau wird die Nullhypothese verworfen, da $p < 0.05$. Mit einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$ unterscheiden sich die medianen Cholesterolwerte von Infarktpatientinnen und Kontrollen signifikant.

5.2.3 Varianzanalyse (ANOVA - ANalysis Of VAriance)

In den vorherigen Abschnitten wurden Tests zur Lage (Mittelwert oder Median) in einer oder zwei Gruppen vorgestellt. Allerdings kommt es nicht selten vor, dass die mittlere Lage in mehr als zwei Gruppen verglichen werden soll (z.B. in einer vierarmigen Studie mit drei verschiedenen Behandlungsmethoden und einer Kontrollbehandlung). In dieser Situation kann immer auf paarweise Vergleiche zurückgefallen werden. Allerdings benutzt man dann jeweils nur einen Teil der Daten und insgesamt führt dies zum Problem des multiplen Testens (siehe Abschnitt 4.8).
Die Varianzanalyse (ANOVA) bietet einen globalen Test, der etwas darüber aussagt, ob es zwischen den Gruppen überhaupt einen Lageunterschied gibt oder nicht. Die Grundidee der Varianzanalyse ist, dass die Variabilität aller Beobachtungen aus allen Gruppen zu einem Teil durch die Gruppenzugehörigkeit

(„Varianz zwischen den Gruppen“) erklärt werden kann. Diese Varianz zwischen den Gruppen drückt sich in den Unterschieden zwischen den Mittelwerten der verschiedenen Gruppen aus. Die überbleibende Restvarianz wird auch als unerklärte Inner-Gruppen-Varianz bezeichnet. Ist die erklärte Varianz zwischen den Gruppen im Verhältnis zur unerklärten Varianz innerhalb der einzelnen Gruppen groß, spricht dies dafür, dass nicht alle Mittelwerte zwischen den Gruppen gleich sind. Im folgenden Beispiel 5.10 werden die Rechnungen zu den Varianzen vorgeführt.

Fragestellung. Gibt es irgend einen Unterschied in den Mittelwerten der Messgröße in den $k (k \geq 2)$ Gruppen?

Hypothesen. $H_0 : \mu_1 = \ldots = \mu_k$ versus H_A : Es gibt mindestens zwei Gruppen mit $\mu_i \neq \mu_j$

Voraussetzungen. Die Messgröße in den einzelnen Gruppen muss normalverteilt und die Varianz in den einzelnen Gruppen muss jeweils gleich sein. In jeder Gruppe sollten gleich viele Beobachtungen auftreten (balanciertes Design).

Testgröße. Der Quotient F aus erklärter zu unerklärter Varianz ist die Testgröße

$$F = \frac{\text{erklärte Varianz}}{\text{unerklärte Varianz}}. \tag{5.6}$$

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware errechnet wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

Häufig werden die Testgröße, die Freiheitsgrade und der p-Wert in einer sogenannten ANOVA-Tabelle wiedergegeben (siehe Beispiel 5.9).

Beispiel: 5.2.8

Vergleich der Patientenzufriedenheit in drei Krankenhäusern
In drei Krankenhäusern A, B und C wurden Patientenbefragungen durchgeführt und ein Zufriedenheitsscore ermittelt. In Krankenhaus A wurden 7 Patienten befragt, in Krankenhaus B 5 und in Krankenhaus C 6 Patienten. Besteht ein signifikanter Unterschied zwischen den Krankenhäusern hinsichtlich der Patientenzufriedenheit? Es kann davon ausgegangen werden, dass jeweils in den drei Städten die Scores normalverteilt sind, auch wenn für dieses fiktive Beispiel die Scores aus didaktischen Gründen bewusst einfach und ganzzahlig gehalten wurden.

Fortsetzung des Beispiels

i	A	B	C
1	3	4	5
2	3	5	6
3	3	3	5
4	4	5	4
5	2	4	5
6	4	-	5
7	2	-	-
$\bar{x}$	3	4.2	5

Tabelle 5.9: Patientenzufriedenheit

$\bar{\bar{x}}$ bezeichnet den globalen Mittelwert der Scores aller befragter Patienten. Er errechnet sich auch als gewichteter Mittelwert der Score-Mittelwerte in den drei Krankenhäusern A, B und C:

$$\begin{aligned}\bar{\bar{x}} &= \frac{n_A \cdot \bar{x}_A + n_B \cdot \bar{x}_B + n_C \cdot \bar{x}_C}{n_A + n_B + n_C} \\ &= \frac{7 \cdot 3 + 5 \cdot 4,2 + 6 \cdot 5}{7 + 5 + 6} \\ &= 4\end{aligned}$$

Hypothesen: $H_0 : \bar{x}_A = \bar{x}_B = \bar{x}_C$ versus H_A : Es gibt Mittelwert-Unterschiede **Testgröße:** Zunächst müssen die „erklärte“ und „unerklärte Varianz“ ausgerechnet werden:

$$\begin{aligned}\text{erklärte Var} &= \frac{n_A \cdot (\bar{x}_A - \bar{\bar{x}})^2 + n_B \cdot (\bar{x}_B - \bar{\bar{x}})^2 + n_C \cdot (\bar{x}_C - \bar{\bar{x}})^2}{k - 1} \\ &= \frac{7 \cdot (3-4)^2 + 5 \cdot (4,2-4)^2 + 6 \cdot (5-4)^2}{3-1} = 6,6\end{aligned}$$

$$\text{unerklärte Var} = \frac{\sum_{i=1}^{n_A}(x_{Ai} - \bar{x}_A)^2 + \sum_{i=1}^{n_B}(x_{Bi} - \bar{x}_B)^2 + \sum_{i=1}^{n_C}(x_{Ci} - \bar{x}_C)^2}{(n_A + n_B + n_C) - k}$$

$$\begin{aligned}\sum_{i=1}^{n_A}(x_{Ai} - \bar{x}_A)^2 &= (3-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2 + (2-3)^2 + (4-3)^2 + (2-3)^2 = 4 \\ \sum_{i=1}^{n_B}(x_{Bi} - \bar{x}_B)^2 &= (4-4,2)^2 + (5-4,2)^2 + (3-4,2)^2 + (5-4,2)^2 + (4-4,2)^2 = 2,8 \\ \sum_{i=1}^{n_C}(x_{Ci} - \bar{x}_C)^2 &= (5-5)^2 + (6-5)^2 + (5-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 = 2 \\ \text{unerklärte Var} &= \frac{4 + 2,8 + 2}{(7+5+6) - 3} = 0,587 \\ F &= \frac{\text{erklärte Varianz}}{\text{unerklärte Varianz}} = \frac{6,6}{0,587} = 11,25 \qquad p = 0,001\end{aligned}$$

Testentscheidung.
Da der p-Wert ($p = 0.001$) kleiner als $\alpha = 5\%$ ist, wird die Nullhypothese auf dem 5%-Signifikanzniveau abgelehnt. Es gibt einen signifikanten Unterschied in dem mittleren Zufriedenheitsscore zwischen den drei Krankenhäusern.
Wie die paarweisen Vergleiche aussehen, könnte nun in paarweisen Vergleichen (etwa t-Tests) weiter untersucht werden. Allerdings muss dabei beachtet werden, dass dann gleich mehrere Tests nötig sind und für das multiple Testen korrigiert werden muss.
In der folgenden Tabelle sehen Sie die klassische Darstellungsform einer ANOVA Auswertung:

Fortsetzung des Beispiels

	Freiheits-grade	Quadrat-summen	Mittel-wert	F	p-Wert
Zwischen den Gruppen	2	13,2	6,6	11,25	0,001
Innerhalb der Gruppen	15	8,8	0,587		

Tabelle 5.10: ANOVA Auswertung

Die hier vorgestellte ANOVA untersucht den Einfluss eines Faktors (nämlich Krankenhaus) auf die Patientenzufriedenheit. Deswegen heißt diese Analyse 1-faktorielle ANOVA. In einer 2-faktoriellen ANOVA wird der Einfluss zweier Faktoren auf die mittleren Lage der Messgröße in den Gruppen bestimmt.
Zur ANOVA existiert ein nicht-parametrisches Pendant – der nicht-parametrische Kruskal-Wallis Test. Im Grunde wird hier eine 1-faktorielle ANOVA auf Rangzahlen statt auf den orginalen Messwerten durchgeführt.

5.2.4 Übersicht zu Lagetests

Tests für Lagemaße dienen der Beantwortung der Frage, ob aus der Tatsache, dass sich zwei oder mehrere Messreihen hinsichtlich ihres arithmetischen Mittels bzw. Medians unterscheiden, geschlussfolgert werden kann, dass sich auch die Grundgesamtheiten, denen die Stichproben entnommen wurden, im Mittel unterscheiden. Eine Übersicht zu den verschiedenen Lagetests wird in Tabelle 5.11 vermittelt.

Anzahl der Stich-proben	**Art der Stich-proben**	**Merkmal normalverteilt**	**Test**
1		ja	t-Test (Normwertvergleich)
		nein	Wilcoxon-Test (Normwertvergleich)
2	unabhängig	ja	t-Test für unabhängige Stichproben
		nein	U-Test nach Mann-Whitney
	abhängig	ja	t-Test für Paardifferenzen
		nein	Wilcoxon-Test
> 2	unabhängig	ja	univariate Varianzanalyse
		nein	H-Test nach Kruskal-Wallis*
	abhängig	ja	multivariate Varianzanalyse*
		nein	Friedman-Test*

Tabelle 5.11: Übersicht zu Lagetests. Tests mit * werden im Skript nicht behandelt.

A_1	A_2	...	A_k	$\sum$
a	b	...		n

Tabelle 5.12: Binomialtest für eine Ereignis-Wahrscheinlichkeit 1

A	$\bar{A}$	$\sum$
a	b	n

Tabelle 5.13: Binomialtest für eine Ereignis-Wahrscheinlichkeit 2

5.3 Tests bezüglich relativer Häufigkeiten

Häufigkeiten können auf allen Skalenniveaus ermittelt werden, und deswegen sind Tests zum Vergleich von Häufigkeiten auch sehr breit anwendbar. Es genügt ein kategoriales Merkmal wie etwa Geschlecht oder Behandlungsart (Therapie A oder Therapie B), um Häufigkeiten zu ermitteln. Bei metrischen Merkmalen kommt man durch Kategorisierung zu Häufigkeiten, etwa wenn statt des systolischen Blutdruckwert nur notiert wird, ob der Wert über 160mmHg liegt oder nicht.
Beispiele für Hypothesen, die mit Häufigkeitstests geprüft werden können, sind:

- Liegt die Geburtenrate von Mädchen bei 50%?
- Hat Therapie A die gleiche Erfolgswahrscheinlichkeit wie Therapie B?
- Gibt es einen Zusammenhang zwischen Rauchen und dem Geschlecht, d.h. rauchen Männer genauso oft wie Frauen?
- Ist der Anteil der Patienten mit einem systolischen Blutdruck über 160mmHg in der Therapiegruppe niedriger als in der Kontrollgruppe?

5.3.1 Häufigkeiten bei einem Merkmal

Zu einem Merkmal mit k Ausprägungen erhält man in einer Stichprobe die entsprechenden Häufigkeitstabelle (siehe Tabelle 5.12), indem man für jede Merkmalsausprägung zählt, wie häufig es in der Stichprobe vorkommt. Teilt man diese Anzahlen durch die Größe der Stichprobe, erhält man die Anteile, d.h. die relativen Häufigkeiten der Merkmalsausprägungen.

Binomialtest für eine Ereignis-Wahrscheinlichkeit

Im einfachsten Fall hat das eine Merkmal lediglich zwei Ausprägungen und beschreibt eine Alternative, etwa Leben versus Tod oder Ja versus Nein. Oft benennt man die Ausprägungen dann A und $\bar{A}$, wobei das Eintreten von A dabei meist als „Erfolg“ gewertet wird. Zu beiden Ausprägungen bekommt man in einer Stichprobe der Größe n absolute Häufigkeiten a und b mit $a+b=n$. Die Häufigkeitstabelle hat dann die einfache Gestalt, wie in Tabelle 5.13. Die zugehörigen relativen Häufigkeiten $\frac{a}{n}$ und $\frac{b}{n}$ sind Schätzer für die Ereignis-Wahrscheinlichkeiten von A und $\bar{A}$. In dieser Situation eignet sich der Binomialtest, um die Ereignis-Wahrscheinlichkeit von A auf einen bestimmten Wert p_0 hin zu testen.

Fragestellung. Unterscheidet sich die Erfolgswahrscheinlichkeit p für die Ausprägung A von einem festen Wert p_0?

Hypothese. $H_0 : p = p_0$ versus $H_A : p \neq p_0$ (zweiseitige Fragestellung).

Voraussetzungen. Man braucht zu jeder der n Beobachtungseinheiten aus der Stichprobe die Information, ob das Ereignis eingetreten ist oder nicht. Damit die Normalapproximation gültig ist, sollte die Stichprobengröße hinreichend groß sein: oft wird in Abhängigkeit der zu zeigenden Wahrscheinlichkeit p_0 als Daumenregel $n \geq \frac{9}{p_0 \cdot (1-p_0)}$ angegeben. Bei $p_0 = 0,5$ beispielsweise ergibt sich eine Mindestanzahl von $n = 36$. Je näher p_0 an Null oder Eins herankommt, desto stärker wächst der minimale Stichprobenumfang. Ist die Mindestgröße nicht erfüllt, können sogenannte exakte Tests (z.B. Test nach Pearson-Clopper) zum Einsatz kommen, die hier aber nicht näher erläutert werden.

Testgröße. Ist X die Anzahl der Erfolge, dann ist $\hat{p} = \frac{X}{n}$ ein Schätzer für die Ereignis-Wahrscheinlichkeit p. Die Testgröße errechnet sich nach folgender Formel:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}. \tag{5.7}$$

Testentscheidung. Die Nullhypothese wird abgelehnt, wenn der aus der Testgröße berechnete p-Wert kleiner als das Signifikanzniveau α ist. Je größer der Betrag der Testgröße, desto kleiner der p-Wert.

Die Standardnormalverteilung ist nur näherungsweise die Verteiltung der Testgröße Z unter H_0. Oft wird eine sogenannte Stetigkeitskorrektur bei der Berechnung von Z empfohlen, bei der der Zähler betragsmäßig um $\frac{1}{2n}$ verkleinert wird. Insbesondere bei relativ kleinen Stichprobengrößen erhält man so eine bessere Näherung.
Neben der Normalapproximation gibt es auch die Möglichkeit, aufgrund der Wahrscheinlichkeiten $P(X = k \mid p_0)$ für k Erfolge bei gültiger Nullhypothese (d.h. bei gegebener Erfolgswahrscheinlichkeit p_0), einen Ablehnungsbereich zu bestimmen. Die bekannteste exakte Methode ist das Verfahren nach *Clopper-Pearson*. Wie auch beim 1-Stichproben t-Test sagt das Konfidenzintervall hier mehr aus als die Testentscheidung, da das Konfidenzintervall auch ausdrückt, was plausible Ereignis-Wahrscheinlichkeiten sind.

Beispiel: 5.3.1

Von n = 50 Patienten auf einer internistischen Station haben k = 15 Patienten einen Typ 2 Diabetes. Ist diese Häufigkeit auf dem 5%-Signifikanzniveau mit der Hypothese vereinbar, dass $p_0 = 45\%$ der internistischen Patienten einen Typ 2 Diabetes haben?

Lösung:
$H_0 : p = 0,45$ versus $H_A : p \neq 0,45$
Schätzer für die Häufigkeit: $\hat{p} = \frac{15}{50} = 0,3$
Die Stichprobengröße ist ausreichend für die Normalapproximation.

Testgröße (ohne Stetigkeitskorrektur):

$$z = \frac{\frac{15}{50} - 0,45}{\sqrt{\frac{0,45 \cdot (1-0,45)}{50}}} = -2,132$$

Dies entspricht einem p-Wert von

$$p = 0.0336$$

Da der p-Wert kleiner als das vorgegebene Signifikanzniveau ist, wird die Nullhypothese abgelehnt, d.h. die Daten der $n = 50$ Patienten sprechen auf dem 5%-Signifikanzniveau dagegen, dass 45% der internistischen Patienten an Typ 2 Diabetes leiden. Vielmehr scheint ein niedrigerer Anteil plausibel, da der Punktschätzer mit 30% unter $p_0 = 45\%$ bleibt.
Ohne Herleitung sei hier noch das Clopper-Pearson Konfidenz-Intervall angegeben: es reicht von 17,86% bis 44,6%. Also wird auch mit dieser exakten Methode die Nullhypothese abgelehnt, da 45% nicht im Konfidenzintervall enthalten ist.

χ^2-Test bei einem Merkmal

Gibt es mehr als zwei Ausprägungen eines Merkmals, kann man einen χ^2-Test anwenden. In der Nullhypothese werden für die einzelnen Ausprägungen Auftrittswahrscheinlichkeiten angenommen. Bei einer Stichprobengröße n entsprechen diese Wahrscheinlichkeiten p_i den erwarteten Häufigkeiten $n \cdot p_i$. Der Test überprüft, inwieweit die beobachteten Häufigkeiten von den erwarteten Häufigkeiten abweichen.

Fragestellung. Unterscheiden sich die Ereigniswahrscheinlichkeiten $p_1, \ldots, p_k$ für die k Ausprägungen eines Merkmals von den festen Wahrscheinlichkeiten $p_{01}, \ldots, p_{0k}$?

Hypothesen. $H_0 : p_1 = p_{01}, \ldots, p_k = p_{0k}$ versus $H_a : p_j \neq p_{0j}$ für mindestens ein $j = 1, \ldots, k$. Weil es um mehr als zwei Ausprägungen geht, gibt es bei diesem Test nur die zweiseitige Fragestellung.

Voraussetzungen. Die erwarteten Häufigkeiten zu den einzelnen Ausprägungen sollten alle mindestens 5 und keine der beobachteten Häufigkeiten sollte Null sein.

Testgröße. Wie bei allen χ^2-Tests werden auch hier die unter H_0 erwarteten Häufigkeiten e_i mit den beobachteten Häufigkeiten b_i verglichen.

$$X^2 = \sum_{i=1}^{n} \frac{(b_i - e_i)^2}{e_i}. \tag{5.8}$$

Testentscheidung. Aus der Testgröße resultiert ein p-Wert. Bei einem vorgegebenen Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn $p < \alpha$ gilt. Je größer der Betrag der Testgröße X^2, desto kleiner der p-Wert.

Beispiel: 5.3.2

Ein Freund hat zum Würfelpokern geladen. Nach einigen Runden hegen Sie den Verdacht, dass der grüne Würfel nicht fair ist und hohe Augen bevorzugt. Daher nehmen Sie diesen Würfel beiseite und werfen ihn 42 Mal. Folgende Tabelle zeigt das Resultat der Würfe.

⚀	⚁	⚂	⚃	⚄	⚅
2	3	6	7	11	13

Können Sie auf dem 5%-Signifikanzniveau schlussfolgern, dass der grüne Würfel tatsächlich nicht fair ist?

Lösung:

Bei einem fairen Würfel ist die Wahrscheinlichkeit für jede Augenzahl von ⚀ bis ⚅ genau $\frac{1}{6}$. Mit dem χ^2-Test bei einem Merkmal kann die Übereinstimmung geprüft werden. Hier wird der Test als Signifikanztest angewendet.

Auge	b_i	e_i	$\frac{(b_i-e_i)^2}{e_i}$
⚀	2	7	3,571
⚁	3	7	2,286
⚂	7	7	0
⚃	6	7	0,143
⚄	11	7	2,286
⚅	13	7	5,143

Tabelle 5.14: Augenzahl beim Würfeln

Testgröße $X^2 = \sum_i \frac{(b_i-e_i)^2}{e_i} = 13,429$ mit $p = 0.0197$.

Testentscheidung. Da $p < 0.05$ gilt, wird die Nullhypothese verworfen.

Auf dem Signifikanzniveau von 5% lässt sich schlussfolgern, dass der grüne Würfel nicht fair ist. Die Augen ⚀ und ⚁ sind unterrepräsentiert, während ⚄ und ⚅ häufiger als normal auftreten. (Sie verlassen die Spielrunde.)

Bei Signifikanztests geht es darum, die Nullhypothese abzulehnen, und im vorherigen Beispiel 2.12 wird der χ^2-Test bei einem Merkmal als Signfikanztest eingesetzt. Dieser χ^2-Test wird aber auch oft gerade dafür angewandt, die *Gültigkeit* der Nullhypothese zu untermauern. Dazu setzt man ein höheres Signifikanzniveau α an und wertet ein Nicht-Ablehnen der Nullhypothese als Hinweis auf deren Gültigkeit. Der

Test wird dann als *Anpassungstest* (*Goodness-of-fit Test*) und nicht als Signifikanztest verwendet. Mehr zu Anpassungstest steht in Abschnitt 5.4.

5.3.2 Häufigkeiten bei zwei Merkmalen

Kreuztabellen

Zur Darstellung der Häufigkeiten bei mehr als einem Merkmal eignen sich Kreuztabellen. Jedes Merkmal wird in einer eigenen Achse aufgetragen. Bei zwei Merkmalen genügt eine 2-dimensionale Kreuztabelle. Im folgenden wird der einfachste Fall dargestellt, wenn es zwei Alternativmerkmale sind, die also jeweils nur zwei Ausprägungen haben,wie etwa Erbrechen: ja/nein oder Leukopenie: ja/nein. Die Anzahlen lassen sich dann in einer 2×2-Kreuztabelle notieren.

	Erbrechen	kein Erbrechen
Leukopenie	a	b
keine Leukopenie	c	d

Tabelle 5.15: Kreuztabelle

Oft werden in den einzelnen Zellen neben der beobachteten absoluten Häufigkeit noch andere Kennzahlen angegeben, wie etwa die Zeilen-, Spalten- oder Gesamtprozente. Die absoluten Häufigkeiten in den einzelnen Zellen geben den Zusammenhang der beiden Merkmale wieder. Leiden Patienten mit einer Leukopenie viel häufiger unter Erbrechen als Patienten, die keine Leukopenie haben, dann werden in den Diagonalfeldern a und d der Kreuztabelle viel größere Anzahlen stehen als in b und c.
Es gibt statistische Tests, die überprüfen, ob ein Zusammenhang zwischen den beiden kategorischen Merkmalen besteht. Eine wichtige Erkenntnis ist, dass man die Häufigkeiten für die einzelnen Zellen ausrechnen kann, die man erwarten würde, wenn kein Zusammenhang zwischen den Merkmalen besteht. Dies sind die sogenannten *erwarteten Häufigkeiten*. In der Tabelle 5.15 wird der Anteil der Patienten mit Erbrechen mittels $r = \frac{a+c}{a+b+c+d}$ geschätzt. Sind Leukopenie und Erbrechen voneinander unabhängig, dann würde Erbrechen unter den Patienten mit Leukopenie ähnlich oft auftreten wie bei Patienten mit normaler Leukozytenzahl. Das heißt, ungefähr $r \cdot (a + b)$ der $a + b$ Patienten mit Leukopenie müssten unter Erbrechen leiden.
Die Grundidee der Tests ist, dass diese erwarteten Häufigkeiten mit den tatsächlich beobachteten Häufigkeiten in allen Zellen verglichen und zur Testgröße aufaddiert werden. In jeder einzelnen Zelle wird $\frac{(b-e)^2}{e}$ berechnet, wobei b die beobachtete Häufigkeit in der Zelle ist und e die unter H_0 erwartete Häufigkeit ist. Der Nenner e spiegelt die Tatsache wider, dass eine Abweichung um so mehr ins Gewicht fällt, je kleiner die erwartete Häufigkeit ist.
Die Testgröße X^2 errechnet sich dann durch Aufaddieren der Terme zu den einzelnen Zellen, $X^2 = \sum_i \frac{(b_i - e_i)^2}{e_i}$. Gilt die Nullhypothese H_0, dann folgt die Testgröße X^2 einer χ^2-Verteilung. Eine große Abweichung von den erwarteten und beobachteten Häufigkeiten und damit ein großer Wert von X^2 sprechen gegen die Nullhypothese, dass kein Zusammenhang vorliegt.

	A	$\overline{A}$	Randsummen
B	a	b	$n_1 = a + b$
$\overline{B}$	c	d	$n_2 = c + d$
Randsummen	a + c	b + d	$n = a + b + c + d$

Tabelle 5.16: allgem. Vierfeldertafel

Konzeptuell wird bei Kreuztabellen unterscheiden, ob lediglich nur eine Population (und damit nur eine Stichprobe) vorliegt, bei der zwei kategorische Merkmale gemessen werden oder ob es mehrere Populationen (und damit Stichproben) gibt, bei denen ein kategorisches Merkmal verglichen wird.
Liegt nur eine Population vor und zwei kategorische Merkmale wird mit diesem χ^2-Test geprüft, ob beide Merkmale unabhängig voneinander sind. In dieser Situation wird der Test **Unabhängigkeitstest** genannt. Bei mehreren Populationen prüft er, ob das kategorische Merkmal durch die Populationen hinweg

	beobachtete Häufigkeit B	unter H_0 erwartete Häufigkeit E	$\frac{(B-E)^2}{E}$
	a	$\frac{(a+b)(a+c)}{n}$	$\frac{(ad-bc)^2}{n(a+b)(a+c)}$
	b	$\frac{(a+b)(b+d)}{n}$	$\frac{(ad-bc)^2}{n(a+b)(b+d)}$
	c	$\frac{(c+d)(a+c)}{n}$	$\frac{(ad-bc)^2}{n(c+d)(a+c)}$
	d	$\frac{(c+d)(b+d)}{n}$	$\frac{(ad-bc)^2}{n(c+d)(b+d)}$
$\sum$	n	n	χ^2

Tabelle 5.17: beobachtete und erwartete Häufigkeit beim Vierfelder-Test

homogen verteilt ist, und man spricht vom **Homogenitätstest**. Die Rechnungen der beiden Tests sind allerdings identisch.
Beide Tests sollen im Folgenden für den einfachsten Fall demonstriert werden. Dies ist der Fall, wenn die Daten in eine 2×2 Kreuztabelle passen.
Sie sind aber auch an größeren $k \times l$-Kreuztabellen anwendbar.

χ^2 -Unabhängigkeitstest in 2×2 Tafel

In einer Stichprobe aus einer Grundgesamtheit sind zwei Alternativmerkmale A oder $\overline{A}$ und B oder $\overline{B}$ erhoben worden, und die Daten liegen in einer 2×2 Tafel vor.

Fragestellung. Gibt es Abhängigkeiten zwischen den beiden Alternativmerkmalen?

Hypothesen. Unter der Nullhypothese gilt die stochastische Unabhängigkeit, also $H_0 : P(B|A) = P(B)$. Die Alternativhypothese ist bei zweiseitigem Testen entsprechend $H_A : P(B|A) \neq P(B)$.

Voraussetzungen. Alle erwarteten Häufigkeiten in der Kreuztabelle sollten mindestens 5 sein, damit die Testgröße unter H_0 tatsächlich durch die kontinuierliche χ^2-Verteilung angenähert wird. Bei größeren Kreuztabellen werden üblicherweise 20% der Zellen mit Werten unter 5 toleriert. Darüber hinaus verlangen einige Statistiker, dass keine der *beobachteten* Häufigkeiten 0 sein sollte. Sind die Voraussetzungen nicht erfüllt und gibt es mehr als nur zwei Zeilen oder Spalten, kann man zwei Zeilen oder Spalten zu einer zusammenfassen, zumindest falls das auch inhaltlich vertretbar ist. Dadurch erhöhen sich dort die beobachteten und erwarteten Häufigkeiten.

Falls die erwarteten Häufigkeiten wirklich zu klein sind, bietet sich als Alternative Fishers exakter Test an (siehe in Abschnitt 5.3.2).

Testgröße. Bei den konkreten Daten der Stichprobe dienen die relativen Häufigkeiten als Abschätzung der Wahrscheinlichkeiten. Deswegen wird unter H_0 näherungsweise gelten, dass $\frac{a}{a+b} = \frac{a+c}{a+b+c+d}$ gilt. Durch Umstellen der Gleichung nach a sieht man, dass die erwartete Häufigkeit a^E für a den Wert $a^E = \frac{(a+c)\cdot(a+b)}{n}$ annimmt. Analog für die anderen Zellen. Für Zelle a ist $\frac{(a^E - a)^2}{a^E}$ die normierte Abweichung der beobachteten Anzahl und der erwarteten Anzahl. Die Testgröße errechnet sich als Summe dieser normierten Abweichungen aller Zellen der Kreuztabelle. Insgesamt ergibt sich (siehe auch Tabelle 5.16 und Tabelle 5.17)

$$X^2 = \frac{n \cdot (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}. \tag{5.9}$$

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware errechnet wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

	Therapie A	Therapie B	Zeilensummen
Erfolg	a	b	$a+b$
Nichterfolg	c	d	$c+d$
Spaltensummen	$n_2 = a+c$	$n_2 = b+d$	$n = a+b+c+d$

Tabelle 5.19: Vierfeldertafel

Beispiel: 5.3.3

χ^2-Unabhängigkeitstest
1695 Patienten mit fortgeschrittenem Morbus Hodgkin wurden im Rahmen einer Studie mit der Standard-Chemotherapie COPP-ABVD behandelt. Unter Therapie traten als akute Nebenwirkungen unter anderem Erbrechen und Leukopenie ($< 2,0 \cdot 10^3/mm^3$) auf.
Frage: Besteht eine Abhängigkeit zwischen Erbrechen und dem Auftreten einer Leukopenie?

	Erbrechen	**kein Erbrechen**	gesamt
Leukopenie	126	421	547
keine Leukopenie	249	899	1148
gesamt	375	1320	1695

Tabelle 5.18: χ^2-Unabhängigkeitstest

Lösung:
Nach Formel 5.9 wird die Teststatistik ausgerechnet:

$$\begin{aligned} X^2 &= \frac{n \cdot (ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} \\ &= \frac{1695 \cdot (126 \cdot 899 - 421 \cdot 249)^2}{(126+421) \cdot (126+249) \cdot (249+899) \cdot (421+899)} \\ &= 0,39 \\ p &= 0,5329 \end{aligned}$$

Testentscheidung: Da $p > 0.05$ wird H_0 auf dem 5%− Signifikanzniveau beibehalten. Die erhobenen Daten zeigen keinen signifikante Abhängigkeit zwischen Erbrechen und dem Auftreten einer Leukopenie unter o.g. Chemotherapie.

χ^2 -Homogenitätstest in 2 x 2 Kreuztabelle

Vergleicht man in zwei Stichproben die Verteilung eines Alternativmerkmals, passen die Daten ebenfalls in eine 2×2-Kreuztabelle. Allerdings ist diese Situation konzeptuell anders als bei dem χ^2-Unabhängigkeitstest, da zwei Grundgesamtheiten vorliegen, aus denen Stichproben gewonnen werden. An dieser Stelle kann der χ^2-Homogenitätstest gerechnet werden, der aber rechnerisch dem χ^2-Unabhängigkeitstest gleich ist.
Wird beispielsweise in einer randomisierten klinischen Studie das Behandlungsergebnis zweier Therapien A und B miteinander vergleichen, wobei jede Behandlung als Erfolg oder Misserfolg gewertet wird, so ergibt sich die in Tabelle 5.19 dargestellte 2×2-Kreuztabelle.

Fragestellung. Gibt es Unterschiede in der Erfolgswahrscheinlichkeit zwischen den beiden Gruppen (etwa Therapiearm A und B)?

Hypothesen. Die Nullhypothese unterstellt Gleichheit der Wahrscheinlichkeiten, also $H_0 : p_A = p_B$. Bei zweiseitiger Fragestellung lautet die Alternativhypothese entsprechend $H_A : p_A \neq p_B$.

Vorraussetzungen. Die *erwarteten* Häufigkeiten in der Kreuztabelle müssen jeweils ≥ 5 sein. Bei größeren Kreuztabellen werden üblicherweise 20% der Zellen mit Werten unter 5 toleriert. Keine der *beobachteten* Häufigkeiten sollte 0 sein. Sind die Voraussetzungen nicht erfüllt und gibt es mehr als zwei Zeilen oder Spalten, kann man zwei Zeilen oder Spalten zu einer zusammenfassen, zumindest falls das inhaltlich sinnvoll ist. Dadurch erhöhen sich dort die beobachteten und erwarteten Häufigkeiten. Bei verletzten Voraussetzungen bietet sich als alternativer Test Fishers exakter Test (siehe 5.3.2) an.

Testgröße. Unter H_0 gilt, dass die Erfolgs/Nichterfolgs-Verteilung in den Gruppen A und B gleich sind und der Erfolg unabhängig von der Gruppezugehörigkeit ist. D.h. $P(\text{Erfolg}|A) = P(\text{Erfolg})$ und damit gilt auch näherungsweise für die relativen Häufigkeiten in den Stichproben:

$$\frac{a}{a+c} = \frac{a+b}{n} \tag{5.10}$$

Durch das Umstellen nach a erhält man die erwartete Anzahl der Erfolge a^E in Gruppe A:

$$a^E = \frac{(a+b)(a+c)}{n} \tag{5.11}$$

Für jede einzelne Zelle a, b, c und d kann man die erwartete Häufigkeit ausrechnen und mit der beobachteten Häufigkeit vergleichen. Wie oben zur χ^2-Teststatistik beim Unabhängigkeitstest beschrieben wird pro Zelle die quadratische Differenz aus der beobachteten und erwarteten Häufigkeit berechnet und durch die erweiterte Häufigkeit geteilt (d.h. $\frac{(a-a^E)^2}{a^E}$).

$$X^2 = \frac{n \cdot (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}. \tag{5.12}$$

Testentscheidung. Ist der unter der Testgröße ermittelte p-Wert kleiner als das Signifikanzniveau α, so wird die Nullhypothese abgelehnt. Ein großer Betrag der Testgröße führt zu kleinen p-Werten.

Beispiel: 5.3.4

χ^2-Homogenitätstest
Patienten mit fortgeschrittenem Morbus Hodgkin wurden randomisiert zwischen der Standard-Chemotherapie COPP-ABVD und dem neuen Chemotherapie-Regime COPP/ABV/IMEP. Unter beiden Therapien traten als akute Nebenwirkung Leukopenien ($< 2,0 \cdot 10^3/mm^3$) auf.

Frage: Ist die Verteilung der Leukopenien unter den beiden Therapien gleich?

	COPP-ABVD	**COPP/ABV/IMEP**	**gesamt**
Leukopenie	547	675	1222
keine Leukopenie	1148	1013	2161
gesamt	1695	1688	3383

Tabelle 5.20: Verteilung der Leukopenie

Hier beantwortet der χ^2-Test die Frage nach der Homogenität zweier Kollektive bezüglich eines Merkmals.
Lösung:
Die Testgröße berechnet sich nach obiger Formel folgendermaßen

$$\begin{aligned} X^2 &= \frac{n \cdot (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)} \\ &= \frac{3383 \cdot (547 \cdot 1013 - 675 \cdot 1148)^2}{(547+675) \cdot (547+1148) \cdot (1148+1013) \cdot (675+1013)} \\ &= 21,83 \\ p &= 0,000003 \end{aligned}$$

Testentscheidung: Da $p < 0,05$ gilt, wird die Nullhypopthese auf dem 5%- Signifikanzniveau verworfen.
Es besteht ein statistisch signifikanter Unterschied in der Häufigkeit, mit der eine Leukopenie bei den beiden oben genannten Chemotherapien auftritt. Bei der Therapie COPP/ABV/IMEP tritt die Nebenwirkung häufiger auf.

Fishers exakter Test

Ist eine der beiden Voraussetzungen für den χ^2-Unabhängigkeits- bzw. Homogenitätstest nicht erfüllt (zu viele Zellen mit erwarteten Häufigkeiten kleiner 5 oder bleiben Zellen leer), kann man eventuell durch Zusammenlegen bestimmter Zeilen oder Spalten erreichen, dass die Voraussetzungen doch erfüllt werden. Das Verschmelzen von Zeilen bzw. Spalten ist aber nicht immer möglich oder gewünscht, etwa wenn es nur zwei Zeilen bzw. Spalten gibt. In solchen Fällen bietet sich an, statt des χ^2-Tests den exakten *Test von Fisher* zu benutzen.
Dieser Test ist in dem Sinn exakt, dass der p-Wert direkt ausgerechnet und nicht näherungsweise mit Hilfe einer Verteilung bestimmt wird.
Der Test wird hier an 2×2-Tabellen demonstriert, er kann aber prinzipiell auch bei größeren Tabellen benutzt werden, allerdings steigt der Rechenaufwand mit zunehmender Tabellengröße stark an.
Bei Fishers exaktem Test werden nun alle jene Tabellen betrachtet, die die gleichen Zeilen- und Spaltensummen wie die beobachtete Tabelle aufweisen. Die Grundidee für den Test ist, dass unter der Nullhypothese (also Unabhängigkeit bzw. Homogenität) zu jeder Tabelle die zugehörige Wahrscheinlichkeit unter H_0 ausgerechnet werden kann. Zur Testentscheidung wird der p-Wert ausgerechnet, d.h. die Wahrscheinlichkeit, die vorliegende 2×2-Tabelle oder unter H_0 noch unwahrscheinlichere 2×2-Tabellen zu beobachten.

Fragestellung. Besteht bei zwei Merkmalen einer Grundgesamtheit eine Abhängigkeit bzw. gibt es Un-

terschiede in einer Ereigniswahrscheinlichkeit in zwei Gruppen?

Hypothesen. Siehe χ^2-Unabhängigkeitstest bzw. χ^2-Homogenitätstest.

Testgröße. Folgende Formel berechnet die Wahrscheinlichkeit einer 2×2-Tabelle unter gültiger Nullhypothese (Unabhängigkeit bzw. Homogenität) bei fixierten Randsummen:

$$\frac{(a+b)! \cdot (a+c)! \cdot (b+d)! \cdot (c+d)!}{n! \cdot a! \cdot b! \cdot c! \cdot d!} \tag{5.13}$$

Der zweiseitige p-Wert ergibt sich durch Aufsummieren der Wahrscheinlichkeiten der beobachteten Tafel sowie aller noch unwahrscheinlicheren Tafeln mit gleichen Randsummen.

Testentscheidung. Ist der p-Wert kleiner als das vorgegebene Signifikanzniveau α, wird die Nullhypothese abgelehnt.

Beispiel: 5.3.5

In einer Studie wurden 4 Patienten mit Therapie A und weitere 4 Patienten mit Therapie B behandelt. Die Therapieergebnisse sind in der folgenden Tabelle dargestellt.

	lebt	tot	gesamt
Therapie A	3	1	4
Therapie B	2	2	4
gesamt	5	3	8

Tabelle 5.21: Therapieergebnisse

Überleben mit Therapie A mehr Patienten, oder ist dies ein Zufallsbefund (einseitige Fragestellung)? Mit den gegebenen Randsummen (4 Patienten mit Therapie A, 4 mit Therapie B, 5 leben, 3 gestorben) können vier verschiedene Tafeln auftreten:

I	lebt	tot	
A	4	0	4
B	1	3	4
	5	3	8

II	lebt	tot	
A	3	1	4
B	2	2	4
	5	3	8

III	lebt	tot	
A	2	2	4
B	3	1	4
	5	3	8

IV	lebt	tot	
A	1	3	4
B	4	0	4
	5	3	8

Für die Tabelle I errechnet sich die Wahrscheinlichkeit folgendermaßen

$$P(I) = \frac{4! \cdot 5! \cdot 3! \cdot 4!}{8! \cdot 4! \cdot 3!} = \frac{5! \cdot 4!}{8!} = \frac{4!}{6 \cdot 7 \cdot 8} = \frac{2 \cdot 3 \cdot 4}{6 \cdot 7 \cdot 8} = \frac{1}{14}.$$

Fortsetzung des Beispiels

Analog ergibt sich für die anderen Tabellen:

$$P(II) = \frac{3}{7},\ P(III) = \frac{3}{7},\ P(IV) = \frac{1}{14}$$

Zur Berechnung des p-Wertes werden die Wahrscheinlichkeiten der Tabellen aufaddiert, die genauso oder noch mehr für die Alternativhypothese sprechen wie die beobachtete Tabelle II.

$$\text{p-Wert} = P(II) + P(I) = \frac{3}{7} + \frac{1}{14} = \frac{1}{2} = 0,5$$

Die Daten geben also auf dem Signifikanzniveau von 5% keinen Anhalt dafür, dass mit Therapie A mehr Menschen überleben als mit Therapie B.

Häufigkeitstest bei verbundenen Stichproben (McNemar-Test)

Der Homogenitätstest prüft, ob ein Alternativmerkmal homogen zwischen zwei unabhängigen Stichproben verteilt ist. Sind beide Stichproben verbunden, dann sollte zur Überprüfung der Homogenität der **McNemar-Test** verwendet werden. Die Bindung der Stichproben kommt typischerweise durch Vorher/Nachher-Messung zustande.

	A_{SP1}	$\overline{A}_{\text{SP1}}$
A_{SP2}	a	b
$\overline{A}_{\text{SP2}}$	c	d

Tabelle 5.22: Vierfeldertafel beim McNemar-Test
SP: Stichprobe
A: ein beliebiges dichotomes Merkmal

Die Einträge a und d in Tabelle 5.22 sind die Übereinstimmungen des Alternativmerkmals in beiden Gruppen. Die Einträge zur Nicht-Übereinstimmung sind b und c. Grundidee des McNemar-Tests ist, dass bei gleicher Verteilung des Alternativmerkmals in beiden Gruppen (das ist H_0) die Abweichungen ungefähr gleich groß sein müssten, d.h. unter der Nullhypothese würde man erwarten, dass b ungefähr so groß ist wie c.

Fragestellung. Sind die Häufigkeitsverteilungen in beiden Gruppen verschieden?

Hypothesen. Bei zweiseitiger Fragestellung ergibt sich $H_0 : b = c$ versus $H_A : b \neq c$.

Voraussetzungen. Damit die χ^2-Verteilung eine gute Annäherung an die Verteilung der Testgröße unter H_0 ist, sollte die erwartete Häufigkeit für $b^E = c^E = \frac{b+c}{2}$ mindestens 5 sein.

Testgröße. Wie beim χ^2-Test üblich werden die beobachteten Werte b und c mit den unter H_0 zu erwartenden Größen verglichen. In der 2×2-Tabelle würde man unter H_0 für b und für c jeweils $\frac{b+c}{2}$ erwarten. Damit ergibt sich für die Testgröße

$$X^2 = \frac{(b - \frac{b+c}{2})^2}{\frac{b+c}{2}} + \frac{(c - \frac{b+c}{2})^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c} \tag{5.14}$$

Ist $b + c \leq 30$ wird eine sogenannte Stetigkeitskorrektur vorgenommen, um zu vermeiden, dass die Nullhypothese vorschnell abgelehnt wird. Mit Korrektur ist die Testgröße $X^2 = \frac{(|b-c|-1)^2}{b+c}$.

Testentscheidung. Die Nullhypothese wird auf dem Signifikanzniveau α abgelehnt, wenn der p-Wert kleiner als α ist.

Beispiel: 5.3.6

Eine neue Chemotherapie soll eingehender untersucht werden. Insbesondere ist von Interesse, ob die Leukozytenzahl durch die Therapie verändert wird. 59 Patienten werden vor und nach der Therapie auf ihre Leukozytenanzahl hin untersucht. Dabei wird lediglich vermerkt, ob die Leukozytenzahl unterhalb des Normbereichs (d.h. < 4/nl) liegt (d.h. Leukopenie) oder nicht. Folgende Tabelle gibt die Daten wieder:

		Leukozyten **nach** Therapie	
		niedrig	normal
Leukozyten	niedrig	23	7
vor Therapie	normal	17	12

Tabelle 5.23: Vierfeldertafel der Leukozyten

Ändert sich durch die Therapie die Häufigkeit von Leukopenien?
Lösung:
Der McNemar-Test ist geeignet, da es um Häufigkeiten eines Alternativmerkmals bei verbundenen Stichproben geht. Da $8+15 = 23 < 30$, sollte die Stetigkeitskorrektur angewandt werden. Die Testgröße errechnet sich als

$$X^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|7-17|-1)^2}{7+17} = \frac{81}{24} = 3,375$$

$$p = 0.066$$

Da $p > 0.05$ gilt, kann die Nullhypothese auf dem 5%-Signifikanzniveau nicht verworfen werden. Allerdings ist das Ergebnis grenzwertig mit einem p-Wert von 0,066. Ohne Stetigkeitskorrektur wäre der Test signifikant ausgefallen. Insgesamt geben die Daten Hinweise darauf, dass nach der Therapie eine Leukopenie vermehrt auftritt, wenn auch nicht signifikant auf dem 5%-Niveau.

Es gibt Erweiterungen des McNemar-Tests für mehr als nur zwei Merkmalsausprägungen bei zwei verbundenen Stichproben, beispielsweise den Symmetrietest von Bowker. Auch für mehr als zwei verbundene Stichproben gibt es Tests, etwa den Q-Test von Cochran. Auf beide wird hier nicht näher eingegangen.

5.4 Anpassungstests

Anpassungstests, auch „Goodness-of-fit"-Tests genannt, gehen der Frage nach, inwiefern konkrete Daten einer Stichprobe mit einer theoretischen Verteilung übereinstimmen. Die Nullhypothese besagt, dass die Daten der Stichprobe der vorgegebenen Verteilung entstammen. Ein Anwendungsbeispiel ist, dass vor Anwendung des t-Tests mit einem Anpassungstest auf Normalverteilung in den beiden Gruppen getestet wird.
Zwei Tests, die auch als Anpassungstest gebraucht werden können, wurden bereits im vorherigen Abschnitt zu Häufigkeitstest (5.3) vorgestellt: Der Binomialtest für die Ereignis-Wahrscheinlichkeit bei einem Alternativmerkmal überprüft, ob die Daten zu einer Binomialverteilung passen. Der χ^2-Test bei einem Merkmal mit mehreren Ausprägungen ist die Verallgemeinerung des Binomialtests von einem Alternativmerkmal auf ein Merkmal mit mehreren Ausgängen. Der χ^2-Test wird hier nochmal in seiner Rolle als Anpassungstest vorgestellt.
Anpassungstest unterscheiden sich von typischen Signifikanztests (etwa ein t-Test) in der Fragestellung. Während bei einem Signifikanztest „Signifikanzen" nachgewiesen werden sollen, geht es bei Anpassungstest darum, die Nullhypothese zu bestätigen, d.h. es soll gezeigt werden, dass die Daten mit der vorgegebenen Verteilung konsistent sind.
Weil bei Signifikanztests die Ablehnung der Nullhypothese das Ziel ist, wird ein niedriges Signifikanzniveau α für den Test festgesetzt, um ein falsches Ablehnen dieser Nullhypothese unwahrscheinlich zu machen. Wenn allerdings die Gültigkeit der Nullhypothese gezeigt werden soll, ist man daran interessiert die Nullhypothese nicht irrtümlich beizubehalten.
Durch Erhöhen des Signifikanzniveaus α wird erreicht, dass die Nullhypothese eher abgelehnt wird und man hofft, dass dadurch H_0 weniger häufig fälschlich beibehalten wird. Als grobe Faustregel wird etwa

$\alpha \geq 0,3$ verwendet. Das Nicht-Verwerfen der Nullhypothese wird dann als Bestätigung von H_0 gewertet. Dieses Ergebnis sollte dennoch nicht als Beweis für H_0 aufgefasst werden. Es ist besser vorsichtig zu sein und lediglich festzuhalten, dass die Daten *nicht gegen* die Verteilungsannahme aus H_0 sprechen.

χ^2-Anpassungstest

Insbesondere bei diskreten Verteilungen bietet sich der χ^2-Anpassungstest an. Aus der zu testenden diskreten Verteilung lässt sich für jeden möglichen Wert die erwartete Häufigkeit herleiten. Diese können in einer χ^2-Teststatistik mit den beobachteten Häufigkeiten verglichen werden. Aber auch stetige Verteilungen können mit dem χ^2-Anpassungstest geprüft werden. Dazu werden die metrischen Ausprägungen in Klassen zusammengefasst, damit die erwarteten Häufigkeiten auch groß genug sind, um den Test durchzuführen.

Fragestellung. Stimmt die beobachtete empirische Verteilung mit einer vorgegebenen Verteilung überein?

Hypothesen. $H_0 : F_x = \Phi$ versus $H_A : F_x \neq \Phi$, wobei F die empirische und Φ die vorgegebene Verteilungsfunktion ist.

Voraussetzungen. Die erwarteten Häufigkeiten sollten mindestens 5 sein und keine der beobachteten Häufigkeiten sollte Null sein, damit die χ^2-Verteilung eine gute Annäherung an die Verteilung der Testgröße unter H_0 sein kann.

Testgröße. Wie bei einem χ^2-Test üblich errechnet sich

$$X^2 = \sum_i \frac{(b_i - e_i)^2}{e_i}. \tag{5.15}$$

Testentscheidung. Die Nullhypothese wird abgelehnt, wenn der resultierende p-Wert kleiner als das Signifikanzniveau α ist. Je größer dabei der Betrag der Testgröße, desto kleiner der resultierende p-Wert.

Beispiel: 5.4.1

Überprüfen Sie, ob die Verteilung der Blutgruppen 0, A, B und AB in der Stichprobe mit der vorgegebenen Verteilung (40%, 45%, 10% und 5%) übereinstimmt.

Blutgruppe	Wahrscheinlichkeit	n_i	e_i	$\frac{(n_i-e_i)^2}{e_i}$
0	40%	61	61,2	0,00065
A	45%	64	68,85	0,342
B	10%	19	15,3	0,895
AB	5%	9	7,65	0,238

Tabelle 5.24: Verteilung der Blutgruppen

Lösung:
Da die Übereinstimmung mit den Wahrscheinlichkeiten aus H_0 gezeigt werden soll, wählt man als erstes einen erhöhtes Signifikanzniveau, etwa $\alpha = 0,3$.

Fortsetzung des Beispiels

Testgröße:

$$X^2 = \sum_i \frac{(b_i - e_i)^2}{e_i} = 1,475$$

Da es $k = 4$ Ausprägungsklassen gibt und keine Parameter aus der Stichprobe geschätzt werden ($r = 0$), ergibt sich für die Anzahl der Freiheitsgrade $f = k - 1 - r = 4 - 1 - 0 = 3$. Bei einem Signifikanzniveau von $\alpha = 0,3$ ist der kritische Wert folglich $\chi^2_{1-0,3;3} = 3,66$ mit $p = 0.688$.

Testentscheidung: Da $p > 0.3$ wird die Nullhypothese nicht verworfen. Dieses Ergebnis kann als Hinweis darauf gewertet werden, dass die theoretische Verteilung der Blutgruppen tatsächlich in der Grundgesamtheit gilt, aus der die Stichprobe entnommen wurde.

Kolmogorov-Smirnov-Test (KS-Test) und Shapiro-Wilk Test

Der KS-Test ist ein Anpassungstest, der sich insbesondere für stetige Verteilungen (etwa die Normalverteilung) eignet. Bei kleineren Stichproben ist er bei stetigen Verteilungen dem χ^2-Anpassungstest vorzuziehen. Der KS-Test vergleicht die empirische Verteilungsfunktion, die die Verteilung der Stichprobe mit der vorgegebenen Verteilungsfunktion beschreibt.
Da die Normalverteilung eine herausragende Bedeutung hat, wurden Anpassungstest speziell für die Normalverteilung entwickelt. Ein gängiger Normalitätstest ist der Shapiro-Wilk Test. Auch zu diesem Test wird auf die Herleitung verzichtet. In Abbildung 5.4 sind unter anderem Ergebnisse des KS-Tests und Shapiro-Wilk-Tests aufgeführt.

Grafischer Anpassungs-„Test“: der QQ-Plot

Eine nützliche Kontrolle, ob Daten einer Stichprobe einer vorgegebenen Verteilung folgen, kann auch mithilfe von Grafiken durchgeführt werden. Bei einem sogenannten Quantil-Quantil-Diagramm (QQ-Plot) werden die Quantile zweier Variablen gegeneinander abgetragen. Will man Daten einer Stichprobe mit einer gegebenen theoretischen Verteilung (etwa Normalverteilung) vergleichen, werden die empirischen Quantile der Daten auf der y-Achse abgetragen und das dazugehörige Quantil der theoretischen Verteilung auf der x-Achse. Zu jeder Beobachtung wird ein Quantilpaar als Punkt dargestellt. Passen die Daten zur vorgegebenen Verteilung, liegen die Punkte auf einer Geraden. Eine gewisse Abweichung an den äußeren Enden des Plots ist normal.
In Abbildung 5.4 wird dargestellt, wie QQ-Plots bei passenden Verteilungen und bei nichtpassenden Verteilungen aussehen.

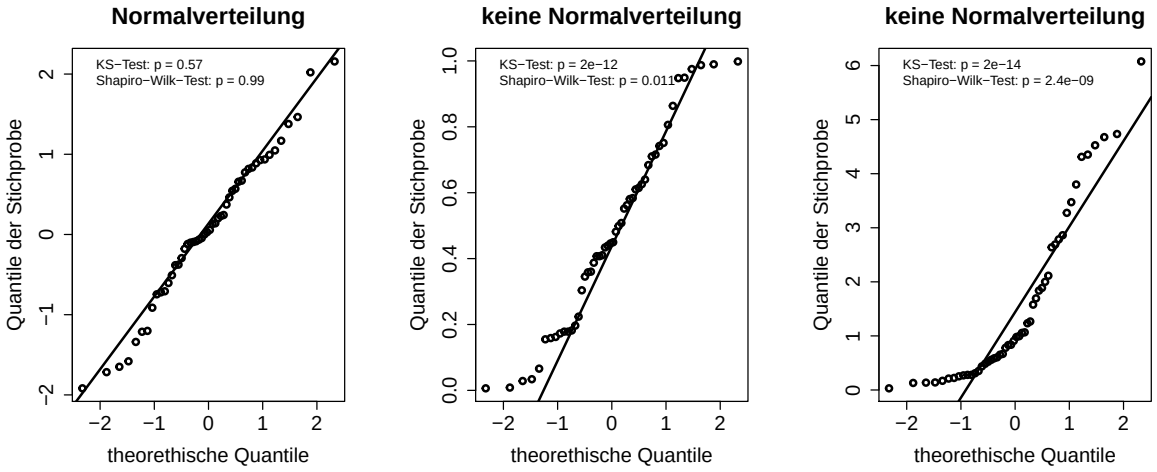


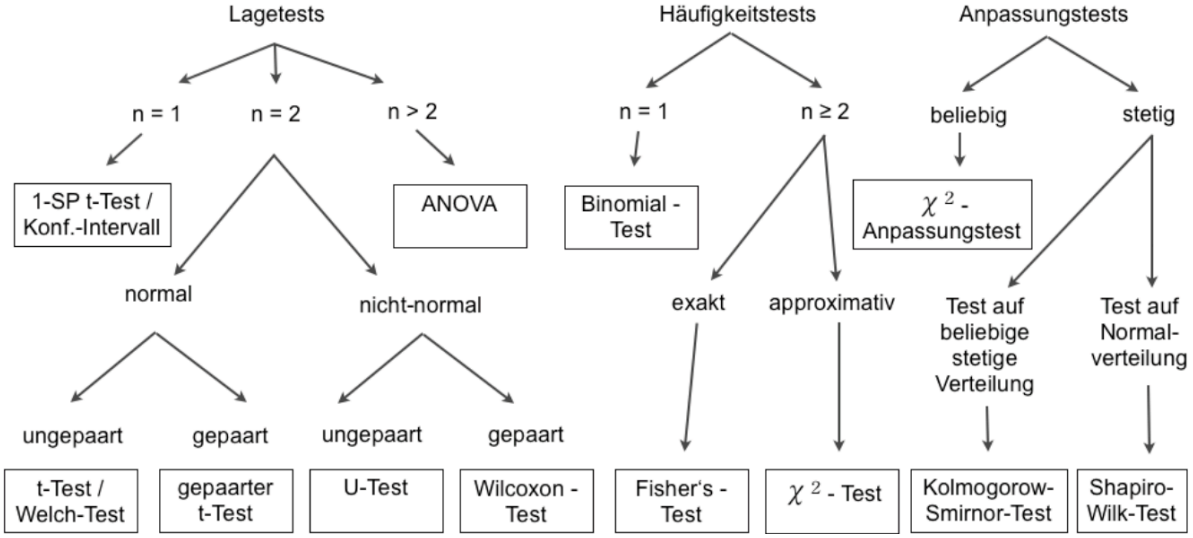
Abbildung 5.4: Verteilungsformen im QQ-Plot mit Angabe der Ergebnisse von KS-Test und Shapiro-Wilk-Test

Dieser „Test“ liefert zwar keine Testgröße und keinen p-Wert, dafür können die Daten der Stichprobe in ihrer Gesamtheit gesehen werden, so dass eventuell einzelne Beobachtungen identifiziert werden, die

nicht zur Verteilungen passen („Ausreißer“).

5.5 Systematisierung der vorgestellten Tests

Zum Abschluss dieses Kapitels soll eine grafische Übersicht zu den vorgestellten Tests gegeben werden.



6 Korrelation und Regression

In den bisher vorgestellten Verfahren der Deskription und des Testens wurde meist nur ein Merkmal betrachtet und ausgewertet. In vielen Studien ist man jedoch daran interessiert, Aussagen über zwei gleichzeitig beobachtete Merkmale zu machen. Bei den Kontingenztabellen haben Sie gesehen, wie der Zusammenhang zweier nominaler Merkmale (etwa Blutgruppe, Geschlecht) untersucht werden kann. Im folgenden Kapitel werden nun Methoden der Korrelationsanalyse vorgestellt, die den Zusammenhang zwischen zwei metrischen Merkmalen beschreiben.
Oftmals möchte man jedoch diesen Zusammenhang genauer charakterisieren. Die Regressionsanalyse gestattet, die Art des Zusammenhangs durch eine Funktion zu beschreiben. Am Ende des Kapitels wird insbesondere die lineare Regression näher beleuchtet.

In diesem Kapitel gehen wir von der Situation aus, dass zwei metrische Merkmale X und Y an den Beobachtungseinheiten gemessen worden sind. Zum Beispiel könnte es sich um eine Stichproben von Patienten handeln, zu denen jeweils Alter und systolischer Blutdruck bestimmt wurde.

6.1 Korrelation

6.1.1 Grafische Darstellung im Streuplot

Vollständige Wertepaare lassen sich graphisch als Punkte in einem (X,Y)-Koordinatensystem darstellen. Diese graphische Darstellung bezeichnet man als Scatterplot (auch: Streudiagramm).

Jede Beschreibung der Beziehung zweier metrisch skalierter Merkmale sollte mit einem Scatterplot beginnen, da dieses die verschiedenen Formen des Zusammenhangs (keiner, linearer, nichtlinearer) sichtbar macht.

Einige Muster sind nachfolgend in Abb. 6.1 dargestellt:

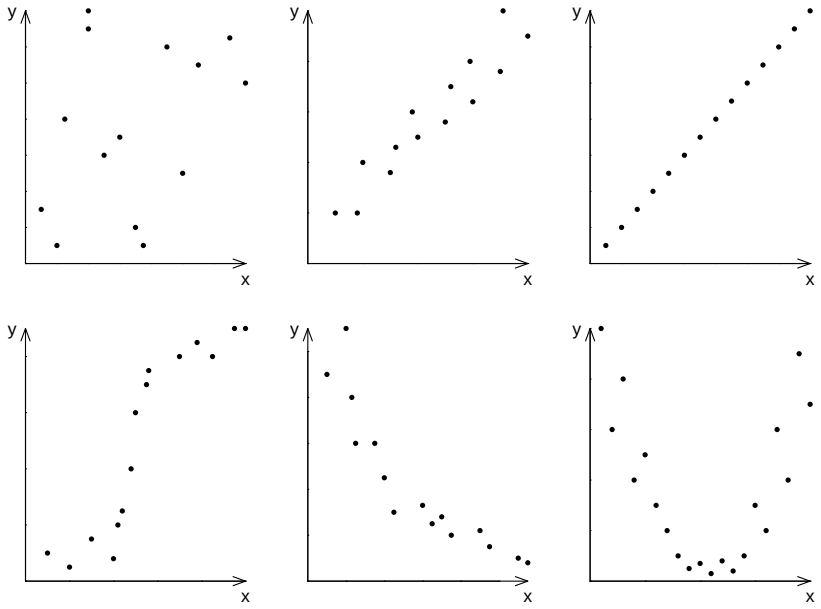


Abbildung 6.1: Verteilungformen

Beispiel: 6.1.1

Aufgabe: Bei einer Gruppe von 20 weiblichen Patienten im Alter zwischen 40 und 60 Jahren soll der Zusammenhang zwischen den Merkmalen X = Alter und Y = systolischer Blutdruck untersucht werden (vgl. Tab.: 6.1).

Nr. der Patientin	Alter	systolischer Blutdruck
1	41	130
2	42	125
3	42	140
4	43	135
5	45	140
6	47	140
7	48	155
8	49	150
9	49	160
10	50	155
11	51	165
12	52	160
13	54	170
14	54	165
15	55	170
16	56	160
17	57	175
18	57	165
19	58	170
20	59	175

Tabelle 6.1: systolischer Blutdruck

Aufgabe: Zeichnen Sie ein zweidimensionales Scatterplot!
Lösung:

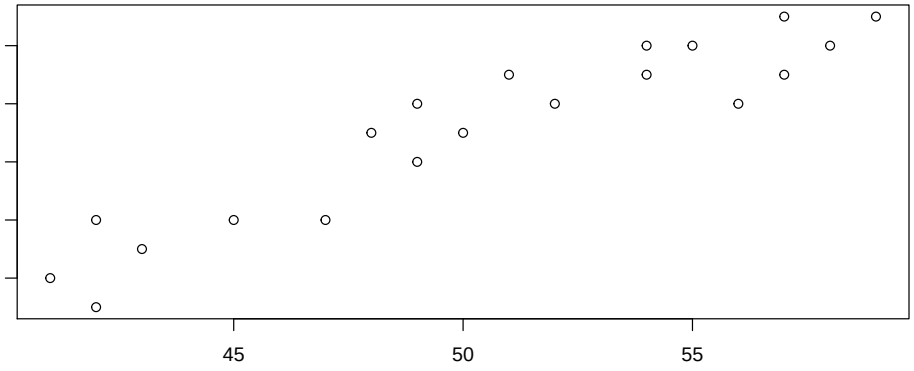


Abbildung 6.2: Scatterplot

Diese Darstellung in Abb. 6.2 deutet auf einen linearen Zusammenhang zwischen den Merkmalen X = Alter und Y = systolischer Blutdruck hin.

Ein Scatterplot dient u.a. zur Identifikation von Ausreißern, d.h. von Punkten, die sehr weit vom Rest der Punktwolke entfernt liegen. Es gilt an dieser Stelle abzuklären, ob es sich um einen Datenfehler handelt.

6.1.2 Korrelationsanalyse nach Pearson

Mit dem Begriff „Korrelationsanalyse“ wird meistens die Berechnung des Produktmoment-Korrelationskoeffizienten nach Pearson verbunden. Der Pearson'sche Korrelationskoeffizient r zwischen zwei metrischen Merkmalen X und Y macht eine Aussage über Richtung und Stärke eines *linearen* Zusammenhangs. Dabei wird eine gedachte Gerade so durch die Punkte gelegt, dass sie die Punktwolke möglichst gut beschreibt. Die Richtung dieser Geraden legt das Vorzeichen des Pearson'schen Korrelationskoeffizienten fest. Die Stärke des linearen Zusammenhangs drückt sich dadurch aus, wie eng die Punkte um die gedachte Gerade herum liegen.

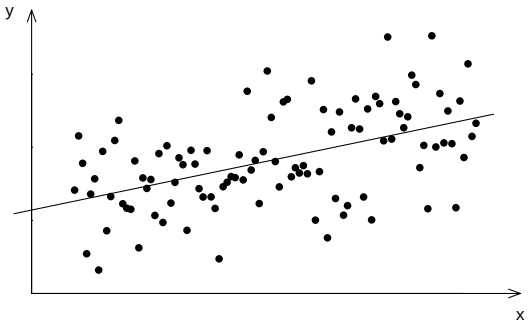


Abbildung 6.3: Punktwolke

Vorzeichen dieses Korrelationskoeffizienten
Der Pearson'sche Korrelationskoeffizient r kann positive und negative Werte annehmen. Man spricht dann von einem positiven (vgl. Abb. 6.4) oder negativen (vgl. Abb. 6.5) Zusammenhang.

Gerade	Vorzeichen
steigend	positiv
waagrecht	0
fallend	negativ

Wertebereich dieses Korrelationskoeffizienten
Der Pearson'sche Korrelationskoeffizienten nimmt Werte zwischen -1 und 1 an. Je größer der Betrag des Korrelationskoeffizienten ist, desto dichter schmiegt sich die Punktwolke um eine gedachte Gerade (vgl. Abb. 6.6). Wie die Stärke des Korrelationskoeffizient zu beurteilen ist, hängt immer davon ab, was gemessen wird und was dort zu erwarten ist. Als ungefähre Orientierungshilfe kann man folgende Einteilung verstehen:

$\|r\|$	Stärke des Zusammenhangs
unter 0,2	sehr schwach
0,2 bis 0,4	schwach
0,4 bis 0,6	mittel
0,6 bis 0,8	stark
über 0,8	sehr stark

Beispiel: 6.1.2

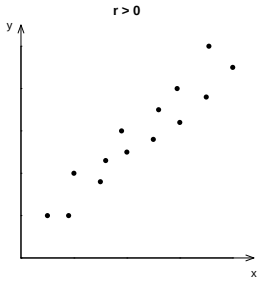


Abbildung 6.4: große X - Werte bedeuten große Y - Werte (gleichläufiger Zusammenhang) Bsp.: Körpertemperatur - Pulsschlag

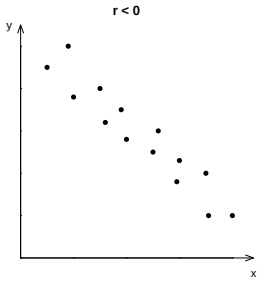


Abbildung 6.5: große X - Werte bedeuten kleine Y - Werte (gegenläufiger Zusammenhang) Bsp.: Akkomodationsfähigkeit des Auges - Alter

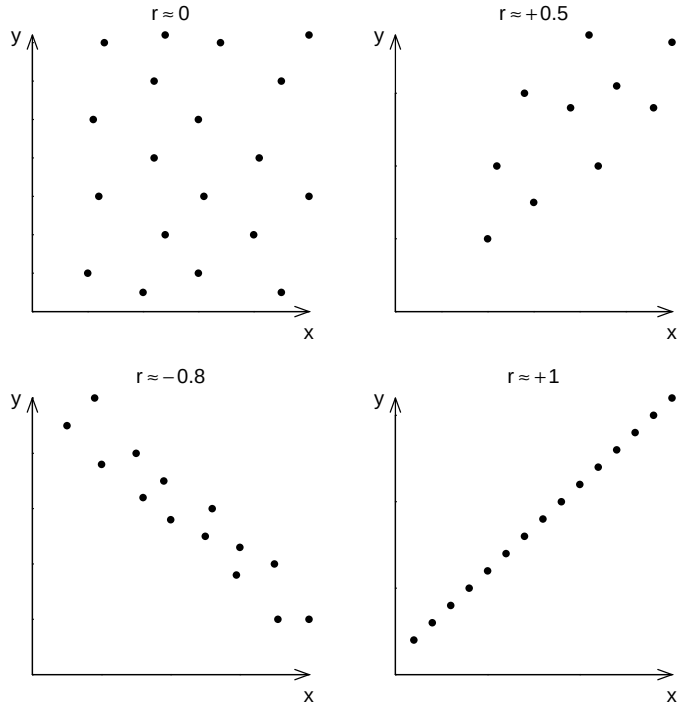


Abbildung 6.6: Verteilung in Abhangigkeit von r

Es ist wichtig, sich klar zu machen, dass durch die Korrelationsanalyse *nicht* geklärt wird, ob ein Zusammenhang auf eine direkte kausale Verknüpfung zurückzuführen ist. Der grundsätzliche kausale Zusammenhang muss durch andere sachlogische Überlegungen geprüft werden.

Berechnung dieses Korrelationskoeffizienten

Mathematisch gilt für den Pearson'schen Korrelationskoeffizienten r zu den metrischen Merkmalen X und Y folgende Definition:

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])\cdot(Y-\mathrm{E}[Y])\right]}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

Er berechnet sich als die Kovarianz zwischen X und Y, die noch durch das Produkt der Standardabweichungen von X und Y normiert wird.
In folgenden Rechenschritten wird der Pearson'sche Korrelationskoeffizient anhand einer konkreten Stichprobe zweier Messungen $X = (x_1, \ldots, x_n)$ und $Y = (y_1, \ldots, y_n)$ berechnet:

1. Summe der Produkte der Abweichungen:

$$SP(X,Y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{6.1}$$

Dieser Term entspricht bis auf den Vorfaktor $\frac{1}{n-1}$ der Kovarianz zwischen X und Y. Die Rechnung lässt sich weiter vereinfachen als $SP(X,Y) = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i$. Zur anschaulichen Bedeutung von $SP(X,Y)$ sei auf Abb. 6.7 verwiesen.

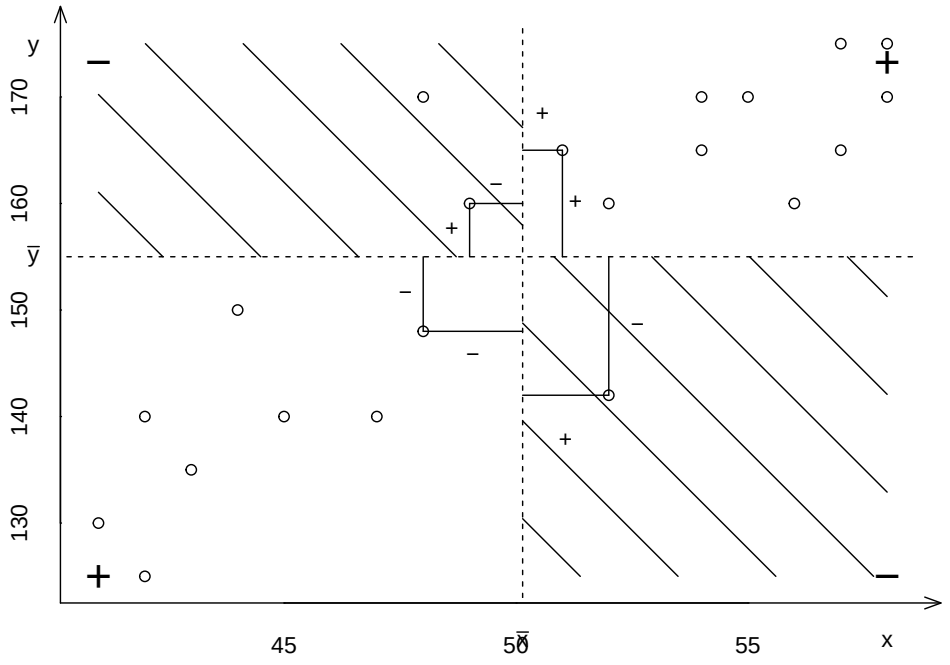


Abbildung 6.7: Veranschaulichung von SP(X,Y)

2. Summe der Quadrate der Abweichungen:

$$SQ(X) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{6.2}$$

$$SQ(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{6.3}$$

Diese Summen sind stark verwandt mit der Stichproben-Varianz von X und Y. Es fehlt lediglich der Faktor $\frac{1}{n-1}$.

3. Der Pearson'sche Korrelationskoeffizient berechnet sich schließlich durch

$$r = \frac{SP(X,Y)}{\sqrt{SQ(X)SQ(Y)}} \tag{6.4}$$

Das Vorzeichen wird alleine durch $SP(X,Y)$ bestimmt. Der Nenner normiert den Koeffizienten: die Summe der Produkte der Abweichungen $SP(X,Y)$ wird in Beziehung zu den Abweichungsquadraten von X und Y gesetzt.

Beispiel: 6.1.3

Aufgabe: Berechnen Sie den Korrelationskoeffizienten für obiges Beispiel! Welche Vermutung haben Sie beim Betrachten des Streuplots hinsichtlich des Vorzeichens und der Größe des Korrelationskoeffizienten?

Lösung:

Nr.	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	41	130	-9,45	-25,25	89,30	637,56	238,61
2	42	125	-8,45	-30,25	71,40	915,06	255,61
3	42	140	-8,45	-15,25	71,40	232,56	128,86
...	...	...	...	...	...	...	...
18	57	165	6,55	9,75	42,90	95,06	63,86
19	58	170	7,55	14,75	57,00	217,56	111,36
20	59	175	8,55	19,75	73,10	390,06	168,86
					SQ(X)=634,90	SQ(Y)=4473,70	SP(X,Y) = 1577,70

Tabelle 6.2: Korrelation

$$r = \frac{1577,70}{\sqrt{634,90 \cdot 4473,70}} = 0,94$$

Zwischen Alter und systolischem Blutdruck besteht eine positive Korrelation, d.h. mit zunehmendem Alter nimmt auch der systolische Blutdruck zu.

Transformation einer nichtlinearen Beziehung in eine lineare Beziehung

Wie bereits gesagt, beschreibt der Pearson'sche Korrelationskoeffizient die *lineare* Beziehung zwischen X und Y. Daher ist dieser Koeffizient nicht geeignet, nicht-lineare Zusammenhänge gut zu erfassen. Mitunter ist es aber möglich, durch geeignete Transformation (z.B. Logarithmieren, wenn eine exponentielle Abhängigkeit vorliegt) eine lineare Abhängigkeit herzustellen (vgl. Abb. 6.8), so dass der Pearson'sche Korrelationskoeffizient auf den transformierten Werten sinnvoll angegeben werden kann.

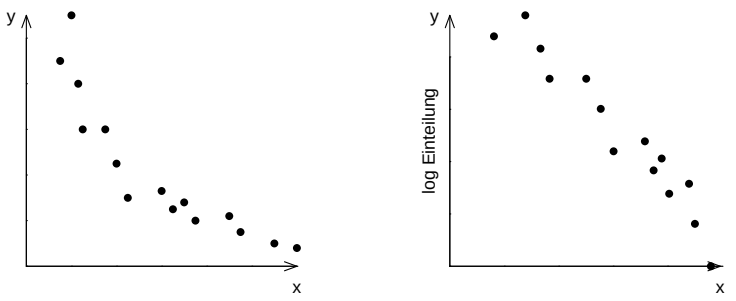


Abbildung 6.8: links: exponentiale Verteilung rechts: logarithmierte Darstellung der Exponentialverteilung

6.1.3 Fehlerquellen bei der Korrelationsanalyse

Ausreißer Datenpunkte (x, y) in einer Stichprobe, die stark von den sonstigen Punkten abweichen, werden Ausreißer genannt. Diese können u.U. eine Korrelation induzieren (vgl. Abb. 6.9).

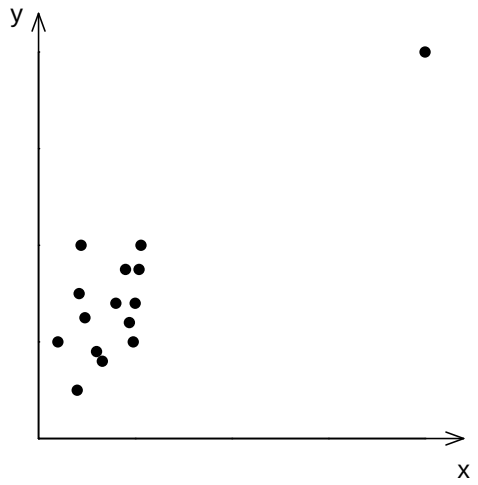


Abbildung 6.9: Ausreißer

Inhomogenität Dieser Fall tritt auf, wenn das Datenmaterial aus zwei oder mehreren Teilmengen besteht, die in verschiedenen Bereichen des Koordinatensystems liegen. Bei getrennter Behandlung der Teilmengen ergibt sich dann häufig keine Korrelation (vgl. Abb. 6.10).

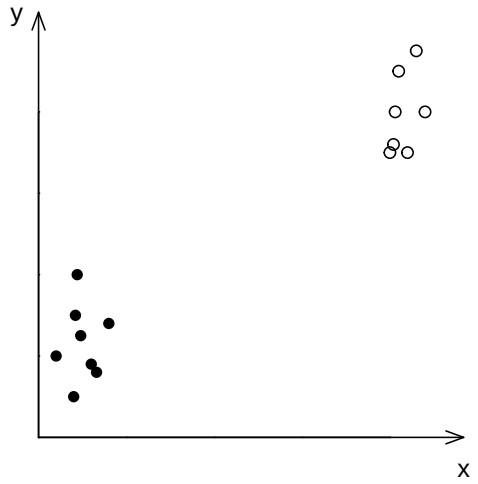


Abbildung 6.10: Inhomogenitat

Gemeinsamkeitskorrelation Dieser Fall tritt auf, wenn zwei Merkmale, die sich nicht gegenseitig bedingen, dem Einfluss einer dritten Größe gemeinsam unterliegen (s.u.: partielle Korrelation).

Selektionsbedingte Korrelation Dieser Fall tritt auf, wenn durch eine nicht-repräsentative Auswahl der Stichprobe ein Zusammenhang vorgetäuscht wird.

Formale Korrelation Dieser Fall tritt auf, wenn sich die Werte der Merkmale zu 100% ergänzen (komplementäre Variable).

6.1.4 Spearman'sche Korrelation

Der Pearson'sche Korrleationskoeffizient ist mit Blick auf zwei normalverteilte Größen entwickelt worden. Vor diesem Hintergrund ist auch seine starke Anfälligkeit gegenüber Ausreißern erklärbar. Es gibt eine ganze Reihe alternativer Korrelationskoeffizienten, die gegen Ausreißer robuster sind. Beispielhaft soll hier der Korrelationskoffizient von Spearman vorgestellt werden.
Die Grundidee ist, dass statt der eigentlichen Werte die Ränge untersucht werden. Es geht dadurch zwar Information aus den Daten verloren, aber mit den Rängen ist es immer noch möglich, zu beurteilen, ob zwisschen X und Y eine monotone Zuordnung herrscht, d.h. etwa für einen monoton steigenden Zusammenhang, ob größere Werte bei X mit größeren Werten von Y zusammengehen. Der Spearman'sche Korrelationskoeffizient errechnet sich als Pearson'scher Korrelationskoeffizient auf den Rängen von X und Y. Die Robustheit gegen Ausreißer und bei nicht normal verteilten Daten kommt daher, dass man sich sich auf die bekannte Verteilung der Ränge 1 bis n zurückzieht.

6.2 Lineare Regression

Die Korrelation machte Aussagen über Richtung und Stärke eines Zusammenhangs zweier Größen. Dabei ging es nicht primär um die Form der Abhängigkeit. Die Ermittlung der funktionalen Abhängigkeit zweier Merkmale nennt man Regression.

Erste Ansätze für Regressionsanalysen gab es im Zusammenhang mit Vererbungsstudien. *Francis Galton* untersuchte die Körperlänge von Vätern und ihren Söhnen und stellte im Jahre 1889 fest, dass „jede vom 'Normalen' abweichende Eigenschaft eines Menschen in der nachfolgenden Generation zwar übernommen wird, aber im Durchschnitt in einem geringeren Maße". Das heißt, es tritt ein Rückschritt (Regression) ein. Diese Beobachtung brachte der gesamten Methode den etwas negativen klingenden Namen ein.

Beispiel: 6.2.1

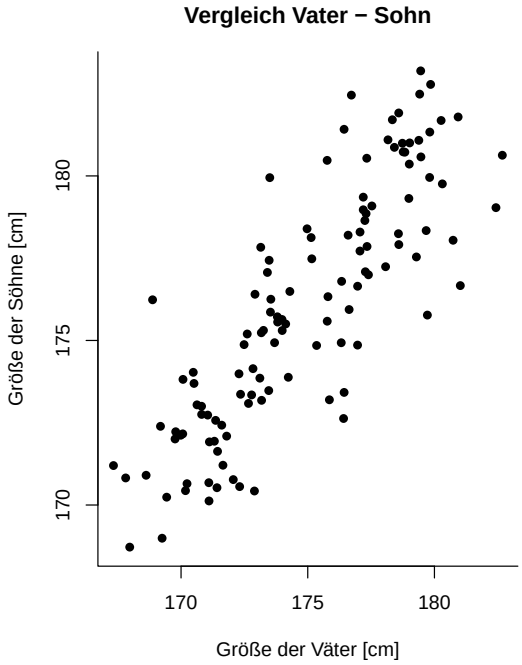


Abbildung 6.11: Größenvergleich bei Vätern und Söhnen

6.2.1 Vorbetrachtungen

Durch die Regression wird typischerwiese der Erwartungswert („Mittelwert“) von Y bei gegebenem Wert von X beschrieben. Anders als bei der Korrelation gibt es bei der Regression also eine „Richtung des Zusammenhangs“: X beschreibt bzw. beeinflusst Y. Die Variable X heißt entsprechend Prädiktor oder auch unabhängige Variable, während Y die Zielgröße oder auch abhängige Variable genannt wird.
Vor Beginn der eigentlichen Regressionsanalyse muss der Funktionstyp festgelegt werden. Dazu sind problembezogene Überlegungen notwendig; eine graphische Darstellung der Wertepaare im Streuplot ist hilfreich. Wird der Zusammenhang durch eine Gerade beschrieben, handelt es sich um eine lineare Regression. In diesem Kapitel soll es nur um diese lineare Regression gehen.

Die lineare Regression geht bei den (X,Y)-Daten von folgendem Zusammenhang zwischen X und Y aus:

$$Y = a + b \cdot X + e$$

wobei a der Achsenabschnitt (*intercept*) ist, b der Anstieg der Geraden (*slope*) und e eine Zufallsgröße ist, die *Fehlerterm* oder *Residuum* genannt wird. Das Residuum beschreibt, wie die Y-Werte von der Geraden abweichen und erfasst damit Messungenauigkeiten oder einfach (biologische) Variabilität in den Daten. Grundlegende Annahmen bei der linearen Regression sind:

1. Der wahre Zusammenhang zwischen dem Erwartungswert von Y bei gegebenem x ist linear, d.h. $E[Y|x] = a + b \cdot x$ und $E[e] = 0$.
2. Die Daten streuen gleichmäßig um die wahre Regressionsgerade, d.h. $\mathrm{Var}(e) = \sigma^2$ ist konstant.
3. Die Abweichungen von der wahren Regressionsgerade bei zwei Beobachtungen sind unkorreliert, d.h. wenn etwa ein Y-Wert oberhalb der Regressionslinie liegt, sollte das keinen Einfluss darauf haben, ob ein anderer Y-Wert auch oberhalb der Regressionslinie liegt. $Cor(e_i, e_j) = 0$.

Eine Regressionsgerade und deren Interpretation ist also nur sinnvoll, wenn ein linearer Zusammenhang (wenigstens innerhalb eines bestimmten Bereichs) gerechtfertigt ist. Ebenso kann eine Extrapolation über den beobachteten Bereich hinaus zu Fehlschlüssen führen. Das gleiche gilt für eine Interpolation zwischen zwei beobachteten Bereichen.
Damit es zu keiner Fehlinterpretation kommt, wie X den Erwartungswert von Y beeinflusst, sollte die Stichprobe X hinsichtlich unberücksichtigter Störgrößen homogen sein.

6.2.2 Bestimmung der Regressionsgerade

Eine lineare Beziehung zwischen zwei Merkmalen X und Y ist mathematisch gegeben durch $Y = a + bX$, wobei die Parameter a und b unbekannte Parameter sind, die die lineare Gerade eindeutig bestimmen. Aufgabe der linearen Regressionsanalyse ist es, die beiden Parameter der „besten“ Gerade zu berechnen, die durch die Daten gelegt werden kann. Dazu wird die **Methode der kleinsten Quadrate** (nach *Karl Friedrich Gauß*, 1777 - 1855) benutzt:
Von allen Geraden, die sich einer gegebenen Menge von Wertepaaren anpassen, wird diejenige Gerade als ”beste”Gerade bezeichnet, bei der die Summe der Quadrate der Abstände $({D_1}^2 + {D_2}^2 + ... + {D_n}^2)$ minimal ist, siehe auch Abb. 6.12. Die entsprechenden Parameter, die zur besten Geraden gehören, werden $\hat{a}$ und $\hat{b}$ bezeichnet.

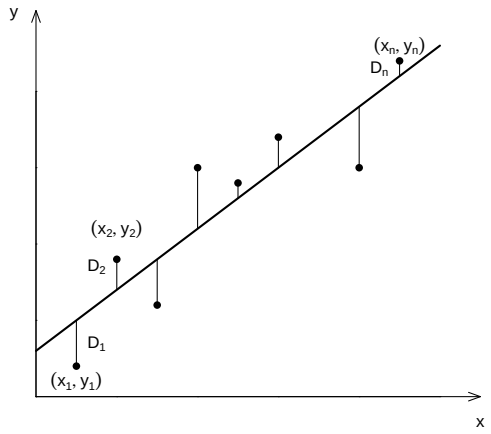


Abbildung 6.12: Regressionsgerade

In folgenden Rechenschritten lassen sich die Kleinste-Quadrate Schätzer für die Parameter a und b bei gegebenen Daten zu X und Y berechnen:

1. Mittelwerte $\bar{x}$ und $\bar{y}$
2. Summe der Quadrate der Abweichungen von X vom Mittelwert
$$\mathrm{SQ}(X) = \sum (x_i - \bar{x})^2$$
3. Summe der Produkte der Abweichungen von X und Y
$$\mathrm{SP}(X,Y) = \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \sum (x_i - \bar{x}) \cdot y_i$$
Die zweite Gleichung resultiert aus einer Vereinfachung der Summe.
4. Mit diesen Größen lassen sich die Parameter der Regressionsgerade (Anstieg und Achsenabschnitt) berechnen:
$$\hat{b} = \frac{\mathrm{SP}(X,Y)}{\mathrm{SQ}(X)} = \frac{\sum (x_i - \bar{x}) \cdot y_i}{\sum (x_i - \bar{x})^2} \text{ und } \hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

An der letzten Gleichung sieht man, dass die Regressionsgerade in jedem Fall durch den Punkt $(\bar{x}, \bar{y})$ der Mittelwerte von X und Y geht.

6.2.3 Interpretation des Regressionskoeffizienten b

b gibt an, um wie viel Einheiten Y im Mittel zunimmt bzw. abnimmt, wenn X um eine Einheit zunimmt. Siehe dazu auch Abb. 6.13. In Analogie zum Vorzeichen des Korrelationskoeffizienten ist der Regressionskoeffizient positiv, wenn er eine gleichläufige Beziehung beschreibt ("große X-Werte bedeuten große Y-Werte") und negativ, wenn er eine gegenläufige Beziehung beschreibt ("große X-Werte bedeuten kleine Y-Werte").

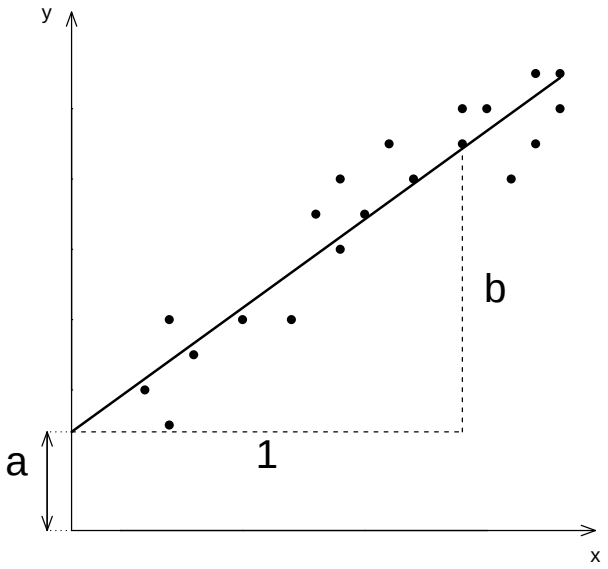


Abbildung 6.13: Regressionskoeffizient

Beispiel: 6.2.2

Aufgabe: Bestimmen Sie die Regressionsgerade für das obige Beispiel. Interpretieren Sie den Regressionskoeffizienten!

Lösung:

$$\begin{aligned} b &= \frac{1577,70}{634,50} = 2,485 \\ a &= 155,25 - 2,485 \cdot 50,45 = 29,882 \\ y &= 29,9 + 2,5x \end{aligned}$$

Wenn das Alter der Frauen um 1 Jahr zunimmt, steigt der systolische Blutdruck im Mittel um ungefähr 2,5 Einheiten (mmHg).

6.2.4 Residuenanalyse

Wenn Sie eine Regressionsgerade aus Daten (X, Y) geschätzt haben, dann haben Sie damit eine Berechnungsvorschrift in der Hand, mit der Sie zu jedem X-Wert ein vorhergesagten Y-Wert ausrechnen können. Bei den Wertepaaren aus der Stichprobe liegt Ihnen sogar der wahre Y-Wert vor. Den Abstand zwischen einem beobachteten Messpunkt y_i an der Stelle x_i und dem entsprechenden Punkt auf der Regressionsgeraden $\hat{y}_i$ ist das geschätzte Residuum $\hat{e}_i = y_i - \hat{y}_i$ (vgl. Abb. 6.14), das sich auf die geschätzte Regressionsgerade bezieht.

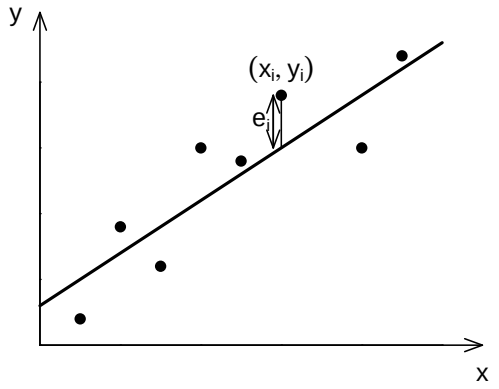


Abbildung 6.14: Residuenanalyse

Aus den geschätzten Residuen $\hat{e}_i = y_i - \hat{y}_i$ wird einerseits die Varianz σ^2 der wahren Fehler e geschätzt. Andererseits werden sie zur Überprüfung der Voraussetzungen des linearen Modells herangezogen. In beiden Fällen müsste man sich dazu eigentlich die echten Residuen e_i anschauen, aber diese Größen sind nicht beobachtbar, weil wir die wahren Regressionparameter a und b nicht kennen, sondern erst aus den Daten schätzen. Der Achsenabschnitt $\hat{a}$ der geschätzten Regressionsgeraden sorgt dafür, dass die beobachtbaren Residuen $\hat{e}_i$ einen Mittelwert von 0 haben. Das heißt, die Mittelwerte der wahren Fehler und der geschätzten Residuen stimmen in jedem Fall überein: $E[e] = E[\hat{e}] = 0$.
Man schätzt die Varianz der wahren Fehler e durch die Stichproben-Varianz der geschätzten Residuen ab.

$$\hat{\sigma}^2 = \frac{\sum(\hat{e}_i - \bar{\hat{e}})^2}{n-2} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} = \frac{\text{RSS}}{n-2}$$

Weil der Mittelwert der geschätzten Residuen Null ist, vereinfacht sich der Zähler zur Summe der Fehlerquadrate, die oft als RSS (*residual sum of squares*) bezeichnet wird: $\text{RSS} = \sum \hat{e}_i^2$. Im Nenner steht allerdings $n-2$ und nicht $n-1$ wie bei der Stichproben-Varianz sonst üblich, da bei der Berechnung der Residuen zwei Parameter $\hat{a}$ und $\hat{b}$ geschätzt worden sind.
Man sollte sich unbedingt den Streuplot der geschätzten Residuen $\hat{e}_i$ (auf der y-Achse) gegen die Modell-Vorhersagen $\hat{y}_i$ (auf der x-Achse) anschauen. Ist die lineare Form für den Mittelwert von Y bei gegebenem x gerechtfertigt, sollten diese Residuen zufällig um die 0 herum schwanken. Ist in dieser graphischen Darstellung irgendeine Struktur erkennbar (wie in Abb. 6.15 und Abb. 6.16), scheint ein lineares Modell nicht gerechtfertigt. Außerdem sollten bei konstanter Varianz des wahren Fehlerterms e auch die geschätzten Residuen $\hat{e}_i$ eine ungefähr konstante Streubreite aufweisen. Eine Trichterform etwa in diesem Residuenplot wie in Abb. 6.16 ist ein Hinweis darauf, dass die Voraussetzungen der linearen Regression verletzt sind.
Gibt es aufgrund des Residuenplots den Verdacht, dass nicht alle Voraussetzungen des linearen Modells erfüllt sind, sollte man vorsichtig sein. Es gibt einige Möglichkeiten (Transformation der X oder Y-Werte; Anwendung eines allgemeineren Regressionsmodells), wie man dennoch in dieser Situation valide Aussagen treffen kann. Allerdings sind diese Möglichkeiten Stoff eines fortgeschritteneren Statistik-Kurses und werden in diesem Skript nicht behandelt.

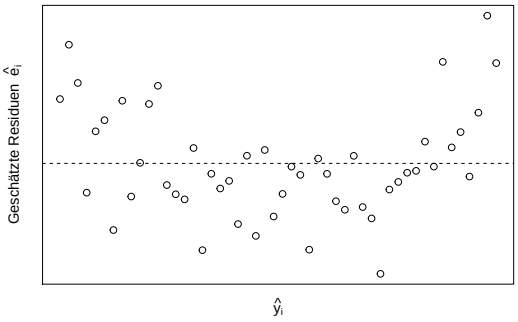


Abbildung 6.15: Muster in Residuen bei nicht-linearem Zusammenhang

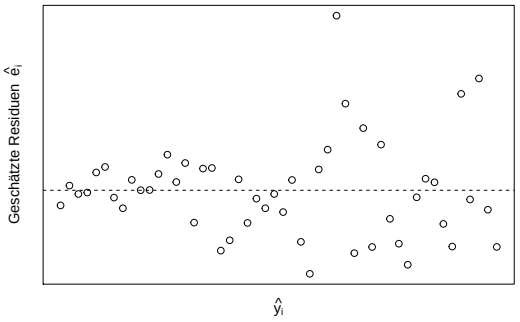


Abbildung 6.16: Trichterform der Residuen bei nicht konstanter Fehler-Varianz

Beispiel: 6.2.3

Aufgabe: Fertigen Sie für obiges Beispiel ein Residuenplot an!
Lösung:

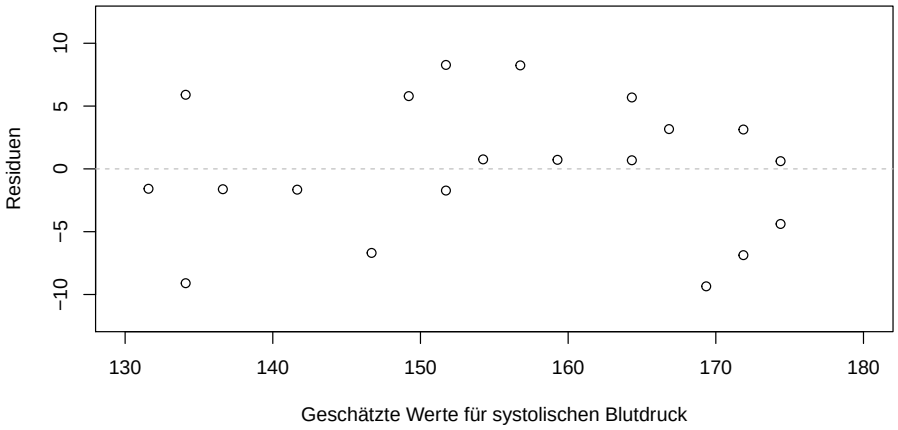


Abbildung 6.17: Residuenplot

6.2.5 Güte einer linearen Regression

Wie gut passt das lineare Modell zu den gegebenen Daten? Wenn das Modell gut passt, werden die einzelnen Residuen $\hat{e}_i = y_i - \hat{y}_i$, also die Abweichung der Beobachtungen y_i von den Modellvorhersagen $\hat{y}_i$, nahe Null sein. Damit wird dann auch die Summe der Residuenquadrate RSS $= \sum \hat{e}_i^2$ klein sein. Deswegen kann man die geschätzte Fehlervarianz $\hat{\sigma}^2 = \frac{\text{RSS}}{n-2}$ als Maß dafür auffassen, wie gut das Modell zu den Daten passt.
Allerdings hängt die Größe von $\hat{\sigma}$ von der Skala der Y ab. Eine unabhängigere Messgröße für die Güte der Modellanpassung ist das Bestimmtheitsmaß B (oder auch R^2, *coefficient of determination*). B ist folgendermaßen definiert:

$$B = 1 - \frac{\text{RSS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Hier wird RSS in Bezug zur Gesamtvariabilität gesetzt. Man kann zeigen, dass $B = \dfrac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ gilt. Insofern gibt B den Anteil der Gesamt-Variabilität der Zielgröße Y an, der durch die Einflussgröße X

erklärt wird. B liegt zwischen 0 und 1 und wird häufig in Prozent angegeben. Je näher B der 1 kommt, desto besser passt die Regressionsgerade zu den (X, Y)-Werten. Man kann zeigen, dass das Bestimmtheitsmaß das Quadrat des Pearson'schen Korrelationskoeffizienten zwischen X und Y ist, $B = r^2$.

Beispiel: 6.2.4

Aufgabe: Berechnen Sie das Bestimmtheitsmaß für obiges Beispiel und interpretieren Sie dieses!
Lösung: Der Pearson'sche Korrelationskoeffizient ist 0,94. Deswegen ist $B = 0,94^2 = 0,88$, d.h. 88% der Variabilität des systolischen Blutdrucks lassen sich durch das Alter erklären.

Abschließend noch ein Wort der Warnung: das Bestimmtheitsmaß B ist durch Ausreißer leicht zu beeinflussen. Aus der Definition wird ersichtlich, dass B davon abhängt, wie stark die Y-Werte in der Stichprobe streuen. Daher kann ein einzelner Daten-Punkt (x, y) mit einem extremen Y-Wert den Wert des Bestimmtheitsmaßes künstlich erhöhen.

6.2.6 Statistisches Prüfen des Regressionskoeffizienten

Der Steigungsparameter b ist bei der Regression meist wichtiger als der Achsenabschnitt a, weil sich durch b ausdrückt, wie sich Y mit Werten von X ändert. Ist b etwa betragsmäßig sehr klein, bedeutet dies, dass sich Y im linearen Modell kaum ändert, wenn X sich ändert. Mit einem t-Test kann man überprüfen, ob b von Null verschieden ist.

Fragestellung. Ist die Steigung der wahren Regressionsgeraden von Null verschieden?

Hypothesen. $H_0 : b = 0$, d.h. der wahre Regressionskoeffizient (Anstieg der Regressionsgeraden) ist Null. $H_A : b \neq 0$, d.h. der wahre Regressionskoeffizient (Anstieg der Regressionsgeraden) ist von Null verschieden.

Voraussetzungen. Die Voraussetzungen der linearen Regression müssen erfüllt sein, d.h. erstens, der Erwartungswert von Y hängt von X ab und verläuft auf einer Geraden und zweitens, die Varianz der Werte von Y um die Gerade herum ist konstant. Damit der Test funktioniert, wird zusätzlich verlangt, dass die Fehler normalverteilt sind, $e \sim N(0, \sigma^2)$. Zu gegebenem x, sind die Werte der Zielvariable Y damit normalverteilt um den Erwartungswert $a + b \cdot x$, siehe Abb. 6.18.

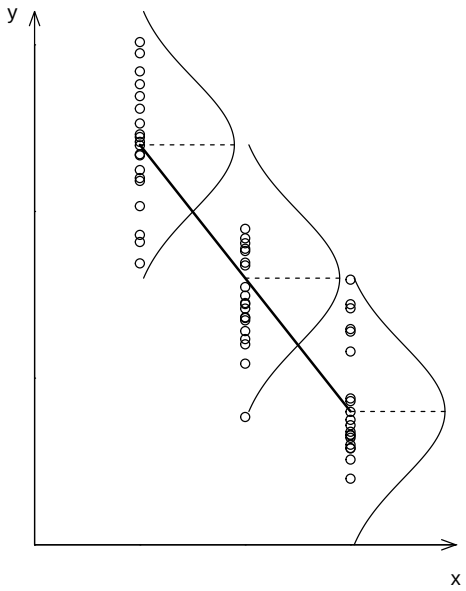


Abbildung 6.18: Normalverteilung der Y bei gegebenem x-Wert

Testgröße. Wie bei allen t-Tests wird der Schätzer ins Verhältnis zu seinem Standardfehler gesetzt

$$t = \frac{\hat{b} - 0}{s_{\hat{b}}} = \frac{\hat{b}}{s_{\hat{b}}},$$

wobei $s_{\hat{b}} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$ der Standardfehler des Schätzers $\hat{b}$ ist. Unter der Nullhypothese genügt die Testgröße einer t-Verteilung mit $n - 2$ Freiheitsgraden, wobei n die Anzahl der Punkte (x_i, y_i) in der Stichprobe ist.

Testentscheidung. Zur ausgerechneten Testgröße gehört ein p-Wert, der durch eine Statistiksoftware ausgegeben wird. Je größer der Betrag der Testgröße, desto kleiner der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese genau dann abgelehnt, wenn der p-Wert kleiner als α ist.

Konfidenzintervall. Dem Test des Anstiegs auf dem Signifikanzniveau α entspricht auch ein $(1 - \alpha)$-Konfidenzintervall zu b. Dieses Intervall ist gegeben durch:

$$\left(\hat{b} - t_{n-2;1-\frac{\alpha}{2}} \cdot s_{\hat{b}} \,;\, \hat{b} + t_{n-2;1-\frac{\alpha}{2}} \cdot s_{\hat{b}}\right),$$

wobei $t_{n-2;1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$-Quantil der t-Verteilung mit $n - 2$ Freiheitsgraden bezeichnet.

Wenn das 95%-Konfidenzintervall zu b die Null nicht enthält, wird der Test $H_0 : b = 0$ bei einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$ abgelehnt und umgekehrt.

Ganz analog kann man den wahren Achsenabschnitt a der Regressionsgeraden gegen einen festen Wert a_0 (etwa $a_0 = 0$) testen. Allerdings ist dieser Test selten von Interesse, da nur der Anstieg (und nicht der Achsenabschnitt) der Regressionsgeraden aussagt, wie der Erwartungswert von Y sich mit steigenden X-Werten entwickelt.

6.2.7 Vorhersagen

Wenn Sie aus einer Stichprobe (X, Y) eine Regressionsgerade geschätzt haben, können Sie auch zu jedem neuen Prädiktor-Werten x_0 eine Vorhersage $\hat{y}_0$ treffen. Zu diesem Punktschätzer lässt sich ebenfalls ein Konfidenzintervall ausrechnen. Genau genommen gibt es zwei Arten von Konfidenzintervallen. Zu einem gegebenem Prädiktor-Wert gibt es:

1. Konfidenzintervall für den geschätzten Mittelwert
2. Konfidenzintervall für eine geschätzte Beobachtung

Der Punktschätzer ist in beiden Fällen gleich. Den Mittelwert kann man sicherer schätzen, für eine einzelne Beobachtung muss man die Streuung σ des Fehlerterms e mit einkalkulieren. Deswegen ist das Konfidenzintervall für eine geschätzte Beobachtung breiter. Wenn zu einem gegebenem x_0 die Frage ist, was im Mittel für Y zu erwarten ist, dann kann man das schmalere Konfidenzintervall für den geschätzten Mittelwert nehmen. Geht es konkret zum Beispiel um einen Patienten mit Prädiktor-Wert x_0, dann muss man das breitere Konfidenzintervall für eine geschätzte Beobachtung heranziehen.
Beiden Konfidenzintervallen ist gemein, dass sie bei $\bar{x}$ am schmalsten sind, d.h. in der Nähe des Mittelwerts der Prädiktoren X aus der Stichprobe lässt sich am besten eine Vorhersage gemäß des Modells machen. Bei Konfidenzintervallen sollte man im Hinterkopf behalten, dass sie immer voraussetzen, dass das lineare Modell das richtige ist und dessen Voraussetzungen erfüllt sind.

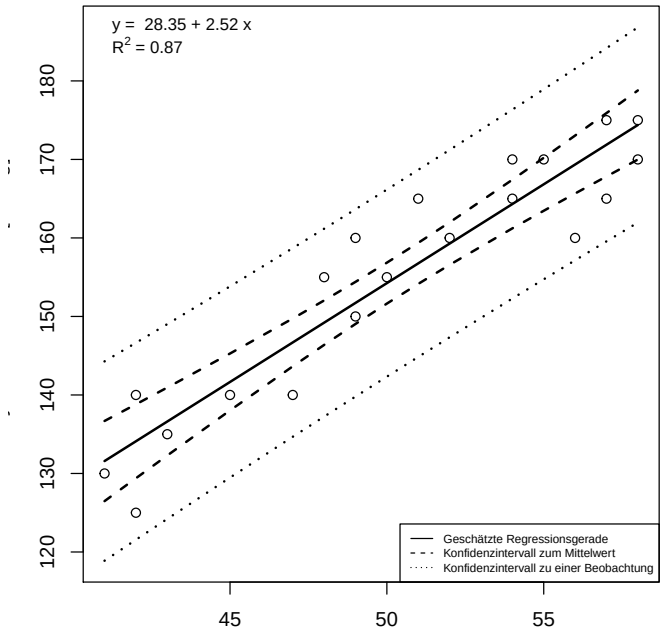


Abbildung 6.19: Konfidenzintervalle zur Vorhersage

6.2.8 Ausblick

Die lineare Regression mit einem Prädiktor ist die einfachste Art der Regression, an der aber gut die Prinzipien der Regression verdeutlicht werden können. Es gibt eine Vielzahl von Erweiterungen, von denen hier nur zwei besonders wichtige Konzepte kurz vorgestellt werden sollen:

Mehrere Prädiktoren. Bisher haben wir die lineare Regression bei nur einem Prädiktor (eine X-Variable) vorgestellt. In konkreten Studien werden aber meist mehr als ein Prädiktor erhoben, etwa Alter, Geschlecht, BMI, Blutdruck und Medikation (Standard versus neues Medikament), deren gemeinsamer Einfluss auf eine Zielvariable untersucht werden soll. Dabei ist auch denkbar, dass zwei Prädiktoren interagieren, das also etwa Blutdruck nur im hohen Alter einen starken Einfluss auf die Zielvariable ausübt. Die Regression kann erweitert werden, um mehr als einen Prädiktor zu berücksichtigen. Man spricht dann von *multipler* oder auch *multivariater Regression*. Auch Interaktionen zwischen zwei Prädkitoren können explizit modelliert werden. Die multiple lineare Regression etwa hat folgendes Modell

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_m X_m + e$$

Die Herleitung und Berechnung der multiplen Regressionsgeraden sind prinzipiell ähnlich wie bei der einfachen linearen Regression, allerdings werden die Rechnungen komplizierter, da nun Vektoren und Matrizen involviert sind. Multiple lineare Regression ist ein Standardverfahren, das Statistikprogramme (wie R oder SPSS) anbieten.

Nicht-lineare Funktion. Eine andere Erweiterung betrifft die funktionelle Form. Gibt man die Voraussetzung auf, dass der Erwartungswert von Y bei gegebenen X-Werten auf einer Geraden liegt, kommt man der nicht-linearen Regression, deren Modellgleichung allgemein mit Hilfe einer Funktion $f(\cdot)$ beschrieben wird:

$$Y = f(X_1, \ldots, X_m) + e.$$

Wie die Funktion f aussieht, hängt von der Natur der Sache ab: falls etwa ein exponentielles Wachstum modelliert werden soll, dann wäre etwa $f(x) = a \cdot \exp(b \cdot x)$ eine gute Wahl. Ein weiteres, wichtiges Beispiel für eine nicht-lineare Regression ist die *logistische Regression*. Sie gehört zu den *verallgemeinerten linearen Modellen*, die eine wichtige Untergruppe der nicht-linearen Regressionsmodelle ist. Bei der logistischen

Regression mit nur einem Prädiktor X hat die Funktion f die Form:

$$Y = \frac{\exp(a + b \cdot X)}{1 + \exp(a + b \cdot X)}$$

annimmt. Hiermit kann eine Eintrittswahrscheinlichkeit (zum Beispiel: Wahrscheinlichkeit, dass ein Patient verstirbt oder dass Nebenwirkungen auftreten) in Abhängigkeit von den Regressoren X modelliert werden.

7 Ereigniszeitanalyse

In medizinischen Anwendungen ist man häufig nicht nur daran interessiert, ob ein bestimmtes Ereignis eintritt oder nicht, sondern auch daran, wie lange es dauert, bis dieses Ereignis eintritt. Man interessiert sich also für die Zeitspanne zwischen einem definierten Anfangspunkt und dem Eintreten eines bestimmten Ereignisses. Diese Zeitspanne könnte man wie andere metrische Messgrößen behandeln, zum Beispiel die Verteilung der Zeitspannen durch ein Histogramm darstellen und Unterschiede der Zeitspannen zwischen zwei Gruppen mit einem t- oder U-Test vergleichen. Allerdings können Messungen von Zeitspannen unvollständig sein, so dass für die Analyse von Zeitspannen eigene Methoden entwickelt wurden, die in diesem Kapitel vorgestellt werden.

7.1 Einführung

Wenn die Zeit bis zum Auftreten eines Ereignisses die primäre Zielgröße ist, dann sollten diese Daten mit Methoden der Ereignisszeitanalyse betrachtet werden. Das Paradebeispiel für Ereigniszeitanalyse ist die Auswertung von Überlebenszeiten, das untersuchte Ereignis ist dann der Tod. Von daher ist zu verstehen, dass man statt Ereigniszeitanalyse oft auch von *Überlebenszeitanalyse* spricht. Ereigniszeiten werden aber nicht nur in der Medizin untersucht. Zum Beispiel ist die Betriebszeit bis zum Ausfall eines elektrischen Bauteils eine ganz wesentliche Messgröße in der industriellen Qualitätskontrolle.
Das Ereignis kann sowohl ein Erfolg (etwa Heilung) als auch ein Misserfolg (etwa Rezidiv oder Tod) sein. Wichtig ist in jedem Fall, dass der Anfangszeitpunkt als auch das Ereignis klar definiert sind. Um etwa eine bestimmte Krebsbehandlung zu untersuchen, kann der Anfangszeitpunkt als Therapiebeginn definiert werden. Als Zielereignis könnte *tumorbedingter* Tod oder aber Tod *jedweder Art* untersucht werden. Dies sind dann zwei unterschiedliche Fragestellungen.

Beispiel: 7.1.1

Beispiele für Ereigniszeiten im medizinischen Kontext:

- Bei Krebspatienten könnte von Interesse sein
  - Zeitdauer zwischen Therapiebeginn und Tod infolge der Krebserkrankung
  - Zeitdauer zwischen Therapiebeginn und Rezidiv eines Tumors
- Zeit bis zur Erholung der Neutrophilen $\leq 1000\text{mm}^3$ nach Hochdosis-Chemotherapie

Wichtige medizinische Fragestellungen bezüglich Ereigniszeiten liegen zum einen in der Beschreibung (Deskription) von Ereigniszeiten (etwa „Wie lange überleben 50% der Patienten nach Beginn der Therapie?“) und im Therapievergleich (etwa „Welche Therapie führt zu längerem tumorfreien Überleben?“).

7.1.1 Kalenderzeit und Ereigniszeit

Es wurde bereits gesagt, dass Anfangs- und Ereigniszeitpunkt klar definiert sein müssen. Im medizinischen Kontext ist der Anfangszeitpunkt für einen Patienten normalerweise dessen Aufnahme in die Studie. In den meisten Studien werden die Patienten nicht alle gleichzeitig zu Beginn der Studie rekrutiert, sondern im Laufe einer längeren Rekrutierungsphase, die sich über Monate oder sogar Jahre erstrecken kann. Die Anfangszeiten für die verschiedenen Studienteilnehmer sind dann individuell verschieden. Die Ereigniszeit ist die Zeit vom individuellen Anfangspunkt bis zum Eintreten des Ereignisses. Für alle Patienten, bei denen das Ereignis während der Studie nicht eingetreten ist, markiert das Ende der Studie das Ende der Nachbeobachtung (auch *Follow-up* genannt). Der Unterschied zwischen *individueller Ereigniszeit* und *Kalenderzeit* ist in der Abbildung 7.1 illustriert: In diesem Beispiel haben etwa die Patienten 1, 4 und 8 jeweils verschiedene Anfangszeiten, allerdings ist für die Ereigniszeit der individuelle Anfang immer $t = 0$.

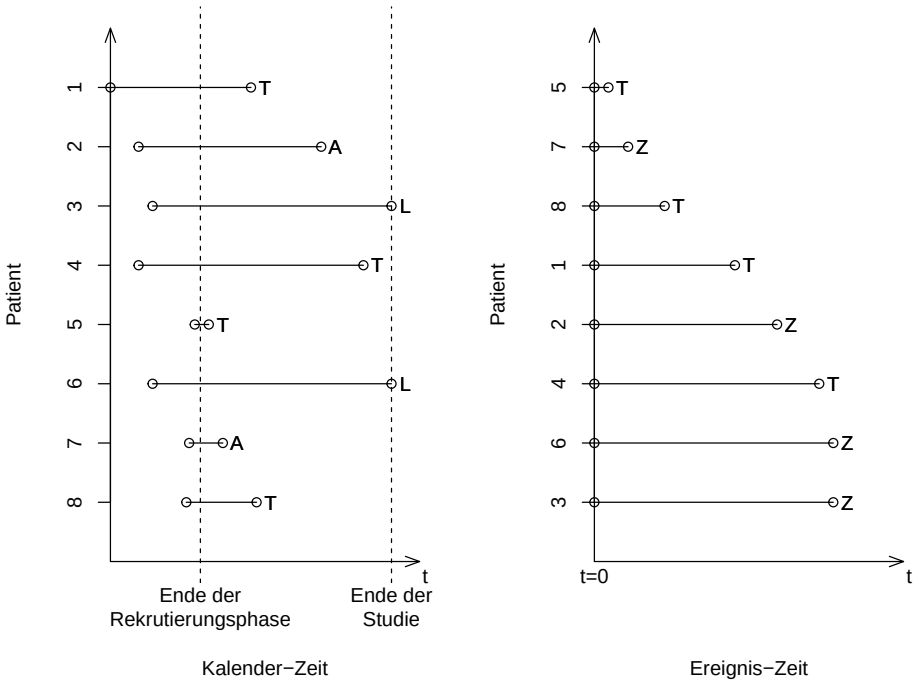


Abbildung 7.1: Kalenderzeit und Ereigniszeit. Tod ist das Zielereignis. (T = Tod, L = lebend, A = ausgeschieden). Für die Ereigniszeit unterscheidet man zwischen eingetretenem Ereignis (T=Tod) und unvollständiger Beobachtung (Z = „zensiert“)

7.1.2 Zensierte Daten

In der Einleitung wurde bereits angedeutet, dass Ereigniszeitdaten nicht mit statistischen Standardverfahren ausgewertet werden sollten. Zum Beispiel sollte man beim Vergleich von Überlebenszeitdaten in zwei Gruppen schon deswegen keinen t-Test verwenden, weil Ereigniszeitdaten meistens nicht normalverteilt sind. Oft ist deren Verteilung rechtsschief. Ein noch grundlegenderes Problem bei Ereigniszeiten sind aber unvollständige Messungen, sogenannte *Zensierungen*.
Bei Studien zur Untersuchung von Zeitspannen bis zum Eintreten eines Zielereignisses kann es vorkommen, dass

- ein Patient vor Eintritt des Zielereignisses aus der Studie ausscheidet (*drop out* oder auch *lost to follow up*), etwa weil er nicht mehr zu den Kontrolluntersuchungen erscheint oder weil er aus Gründen, die nicht mit der untersuchten Erkrankung im Zusammenhang stehen, verstirbt.
- zum Studienende ein Patient zwar noch unter Beobachtung steht, bei ihm aber das Zielereignis (noch) nicht eingetreten ist.

In beiden Fällen ist die Information, die man über den betreffenden Patienten hat, unvollständig: die genaue Zeit bis zum Eintreten des Zielereignisses ist nicht bekannt. Allerdings hat man eine untere Schranke für diese Zeitspanne: die Zeit zwischen Studienbeginn und dem letzten dokumentierten Patientenkontakt (bei Studienende oder vorher, falls der Patient vorzeitig aus der Studie ausscheidet). Die „tatsächliche“ Überlebenszeit ist zum Zeitpunkt des letzten Patientenkontaktes zensiert. Die Information ist zwar unvollständig, aber für die Schätzung der Überlebenszeiten trotzdem nützlich und sollte daher bei der Auswertung Berücksichtigung finden. In Abbildung 7.1 sind alle auftretenden Zensierungen bei der Ereigniszeit mit einem „Z“ markiert.
Vielleicht ist folgender Vergleich nützlich: Stellen wir uns die Überlebenszeit eines Patienten als eine Flüssigkeit vor. Ein Patient mit langer Überlebenszeit entspricht einer großen Flasche, in die viel Flüssigkeit hineinpasst. Während einer Studie beobachten wir die Überlebenszeit bei Patienten nicht auf ein Mal, sondern im Bild gesprochen füllen sich die Flaschen der Patienten langsam. Eine Zensierung heißt, dass das Füllen der Flasche abgebrochen wird, bevor die Flasche ganz voll ist. Auch wenn man nicht weiß, wie viel Flüssigkeit noch in die Flasche gepasst hätte, weiß man doch, dass mindestens die beobachtete Menge an Flüssigkeit hineingepasst hat. Ein Standardverfahren (wie etwa ein Boxplot zur

Beschreibung der Verteilung oder ein t-Test) kann diese unvollständige Information (die halb gefüllten Flaschen) nicht nutzen und würde Patienten mit zensierter Ereigniszeit aus der Auswertung komplett ausschließen. Die Verfahren, die hier vorgestellt werden, nutzen alle Flüssigkeit, die zur Verfügung steht. Welche Patienten als zensiert angesehen werden, hängt vom ausgewählten Zielereignis ab. Ist z.B. das Zielereignis „Rezidiv", dann werden alle ohne vorheriges Rezidiv verstorbenen Patienten (etwa Tod durch Herzversagen) als zensiert angesehen. Ist dagegen das Zielereignis Tod jedweder Art, dann sind alle verstorbenen Patienten unzensiert, aber die am Studienende noch nicht verstorbenen Patienten mit Rezidiv gelten als zensiert.
Bei den Analysemethoden, die hier vorgestellt werden, wird angenommen, dass die Zensierungen *nicht informativ* sind. Das heißt, dass eine Zensierung zufällig einen Patienten trifft und nicht mit dem Zielereignis zusammenhängt. Zensierungen von Patienten, bei denen das Zielereignis zum Zeitpunkt der Auswertung noch nicht aufgetreten ist, sind hier unproblematisch. Kritischer ist es mit Patienten, die vorzeitig aus der Studie ausscheiden. Hier könnte es sein, dass ein besonders guter oder ein besonders schlechter Gesundheitszustand zum Ausscheiden aus der Studie geführt haben. Solche Zensierungen bezeichnet man als informativ, und sie können zu einer systematischen Verzerrung der Ergebnisse der Analyse führen (siehe Beispiel 7.1.2). Daher ist es wichtig, insbesondere bei Therapiestudien, hohe Anforderungen an die Vollständigkeit der Nachbeobachtung und an die Datenqualität zu stellen.

Beispiel: 7.1.2

Informative Zensierungen

In einer klinischen Studie wird das tumorspezifische Überleben bei einer neuartigen Chemotherapie untersucht. Das Zielereignis ist deshalb der Tod infolge des Tumors.
Im Rahmen der Therapie kommt es hin und wieder vor, dass es bei Patienten zu einem Progress der Erkrankung kommt und ein verschlechterter Allgemeinzustand eine Fortführung der Therapie nicht mehr zulässt. Dies führt zu einer Zensierung solcher Patienten. Sollte die Patienten kurz darauf infolge ihres Tumorleidens versterben, dann sind diese Zensierungen informativ, denn sie treten nicht zufällig auf. Vielmehr sind solche Zensierungen bei diesen Patienten ein Hinweis darauf, dass der Tod infolge des Tumorleidens kurz bevor steht.
Behandelt man diese Zensierungen als nicht informativ, dann wird das Überleben höher eingeschätzt als es tatsächlich ist und damit die neue Therapie überschätzt.

7.1.3 Fehlerhafte Auswertung von Ereigniszeiten

Wenn den Besonderheiten der Ereigniszeit keine Bedeutung beigemessen wird, kann es zu Fehlern kommen. Als Warnung werden in diesem Abschnitt Auswerte-Strategien aufgezeigt, die suboptimale oder sogar falsche Eregnisse liefern können.

Ereignishäufigkeiten direkt auswerten Es ist möglich, lediglich die Häufigkeit zu analysieren, mit der das Zielereignis auftritt. D.h. man vernachlässigt die zeitliche Information, die einem zur Verfügung steht. Gerade in Studien mit langem Follow-up kann das ein erheblicher Informationsverlust sein. Wenn eine Therapie das Leben der Patienten um einige Jahre verlängert, heißt das bei langem Follow-up nicht automatisch, dass auch die Ereignishäufigkeit niedriger liegt. Noch drastischer wird es, wenn bei Therapievergleich die einzelnen Therapiearme unterschiedlich lange nachbeobachtet werden. Patienten mit längerem Follow-up haben auch eine höhere Chance, dass das Ereignis in der Studie beobachtet wird. In Therapien mit kürzerem Follow-up kommt es durch das frühe Studienende eher zu Zensierungen, so dass es tendenziell zu einer künstlich verringerten Ereignishäufigkeit kommt.

Beispiel: 7.1.3

Beobachtung von Ereignishäufigkeiten bei unterschiedlichen Beobachtungszeiten

Im Rahmen der Therapie von Depressionen soll untersucht werden wie häufig es nach einer erfolgreichen Therapie zu einem Rückfall (Rezidiv) der Erkrankung kommt.
Es existieren Daten aus einer Standardtherapie der letzten 10 Jahren. Parallel wurde in den letzten 3 Jahren schrittweise eine neue Therapie eingeführt. Die neue Therapie hat damit ein kürzeres Follow-up, woraus sich mehr Zensierungen und eine niedrigere Ereignisrate ergibt. Viele Rezidive erfolgen möglicherweise in den kommenden Jahren und konnten bisher nicht beobachtet werden.
Werte man diese Daten anhand der Ereignishäufigkeiten aus, dann kann es gut sein, dass die neue Therapie besser erscheint als sie vielleicht ist.

Ausschließen aller zensierter Patienten Es ist auch denkbar, die Ereigniszeiten zu untersuchen, allerdings nur die von den Patienten *mit* beobachtetem Zielereignis in der Studie. Die Ereigniszeiten aller zensierten Patienten, d.h. Patienten, bei denen in der Studie das Zielereignis nicht eingetreten ist, werden aus der Analyse ausgeschlossen. Dies kann zu erheblichen Verzerrungen führen, weil auch zensierte Patienten Informationen über die Ereigniszeit liefern, die in dieser Auswerte-Strategie vernachlässigt werden.

Beispiel: 7.1.4

In einer zweiarmigen Studie mit 1-jährigen Follow-up wird Tod als Zielereignis untersucht. Angenommen, im ersten Jahr versterben in der Gruppe mit Therapie A nur 5% der Patienten und im zweiten Studienarm mit Therapie B im gleichen Zeitraum allerdings 60% der Patienten. Dann kann es durchaus sein, dass durch das Ausschließen aller zensierten Patienten beide Therapien als gleich gut bewertet werden, weil mit dieser Auswerte-Strategie nur die beobachteten Überlebenszeiten der 5% bzw. der 60% der Patienten betrachtet werden. Die Tatsache, dass unter Therapie B viel mehr Patienten versterben als unter Therapie A (d.h. mehr Zensierungen in Therapie A) kommt nicht zum Tragen.

7.2 Deskription von Ereigniszeiten

In diesem Abschnitt soll gezeigt werden, wie die Verteilung von Ereigniszeiten in einer Stichprobe beschrieben werden kann. Im Kapitel 1 haben Sie die Verteilungsfunktion kennengelernt. Im Kontext von Ereigniszeiten benutzt man zur Beschreibung der Verteilung eine daran angelehnte Funktion, die *Überlebensfunktion*.
In diesem Abschnitt stellen wir ein Verfahren vor, wie diese Überlebensfunktion aus einer Stichprobe, die auch nicht informative Zensierungen enthalten darf, geschätzt werden kann. Aus der geschätzten Überlebensfunktion können dann auch Lagemaße wie etwa Median oder Mittelwert der Ereigniszeiten abgelesen werden.

7.2.1 Überlebensfunktion und Hazardfunktion

Die Überlebensfunktion (*Survivalfunktion*, manchmal auch *Survivorfunktion*) ist das wichtigste Hilfsmittel, um die Verteilung von Ereigniszeiten zu beschreiben. Der Name ist historisch bedingt und weist darauf hin, dass in den Anfängen der Ereigniszeitanalyse vor allem Tod als Ereignis betrachtet wurde. Die beobachtete Ereigniszeit t eines Individuums kann als Wert einer Zufallsvariable T betrachtet werden, die beliebige nicht-negative Werte annehmen kann. Negative Werte sind deswegen nicht möglich, weil ein Ereignis nicht vor Beobachtungsanfang geschehen kann. Wie in Kapitel 1 bereits erklärt, kann die Verteilung von T durch die Verteilungsfunktion F_T beschrieben werden.

$$F_T(t) = P(T \leq t) \tag{7.1}$$

Die Verteilungsfunktion F_T gibt zu einem Wert t die Wahrscheinlichkeit an, dass T einen Wert kleiner oder gleich t annimmt. Im Kontext von Ereigniszeiten etwa ist $F_T(10)$ die Wahrscheinlichkeit, dass ein Ereignis bis zum Zeitpunkt $t = 10$ eintritt. Die **Überlebensfunktion** $S(t)$ ist definiert als Wahrscheinlichkeit, dass die Ereigniszeit größer t ist und gibt damit für jedes t die Gegenwahrscheinlichkeit zur Verteilungsfunktion wieder: $S(10)$ ist etwa die Wahrscheinlichkeit, dass das Ereignis nach dem Zeitpunkt $t = 10$ eintritt. An der Überlebensfunktion kann also die Wahrscheinlichkeit dafür abgelesen werden, dass ein Individuum zum Zeitpunkt t noch lebt (kein Ereignis hatte).

$$S(t) = P(T > t) = 1 - F_T(t). \tag{7.2}$$

Die Hazardfunktion[1] $h(t)$ ist das Risiko, dass ein Individuum zum Zeitpunkt t stirbt, unter der Bedingung, dass es bis t gelebt hat. Die Hazardfunktion repräsentiert also die momentane Sterberate bzw. Ereignisrate (auch „instantane“ Sterberate) eines Individuums, das bis zum Zeitpunkt t gelebt (bzw. kein Ereignis) hat.

$$h(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{7.3}$$

Zum Beispiel ist die Sterberate für einen Mensch in Europa in den ersten Lebensjahren groß (hohes Hazard). Danach sinkt sie stark und verbleibt für einige Jahre auf einem niedrigen Niveau. Ungefähr ab dem 16. Lebensjahr steigt die Hazardfunktion immer stärker an. Damit drückt sich aus, dass ein 80-Jähriger mit viel höherer Wahrscheinlichkeit zu einem gegebenem Augenblick stirbt als ein 10-jähriges Kind.
Hazardfunktion und Survivalfunktion hängen zusammen. Man kann zeigen, dass die Hazardfunktion die negative Steigung der logarithmierten Survivalfunktion ist, also:

$$h(t) = -\frac{d}{dt} \ln(S(t)) = -\frac{S'(t)}{S(t)} \tag{7.4}$$

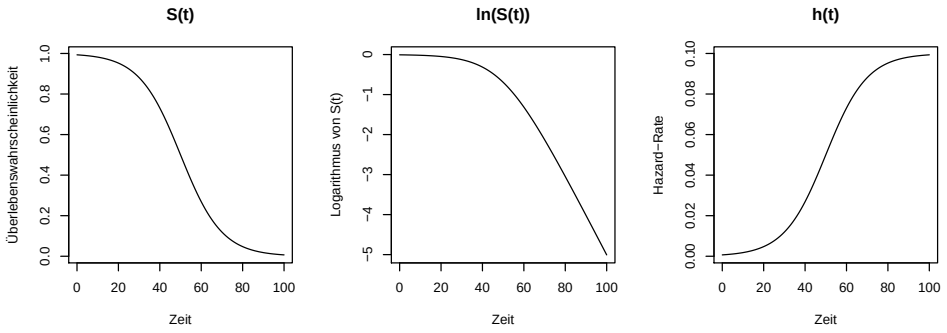


Abbildung 7.2: Beispiel für eine Überlebensfunktion S(t), Logarithmus von S(t) ln(S(t)) und Hazardfunktion h(t)

7.2.2 Überlebensfunktion schätzen nach Kaplan-Meier

Für die Schätzung der Überlebensfunktion, d.h. der Überlebenswahrscheinlichkeit in Abhängigkeit von der Zeit, wird meistens die von *Kaplan und Meier* (1958) vorgeschlagene Methode verwendet, da diese mit Zensierungen adäquat umgeht. In welcher Einheit dabei die Zeit gemessen wird (Tag, Woche, Monat,...) ist problemabhängig, für die Auswertung jedoch unerheblich.
Zur Schätzung der Überlebenszeit muss jeweils die Zeitdauer bis zum Zielereignis festgehalten sein. Bei Patienten ohne Zielereignis (Zensierung) wird stattdessen die letzte Zeitdauer notiert, bei der bekannt war, dass das Zielereignis noch nicht eingetreten ist. Praktisch wird das so gemacht, dass es eine Spalte für die Zeitdauer (Zielereignis oder Zensierung) gibt und eine weitere Spalte (oft *Status* genannt), die festhält, ob nach der angegebenen Zeitdauer bei dem Patienten eine Zensierung vorliegt oder tatsächlich ein Ereignis aufgetreten ist.
Die Überlebenszeitfunktion beginnt immer bei 100%, da am Anfang noch kein Ereignis eingetreten ist („alle leben noch“), $\hat{S}(0) = 1$. Der Kaplan-Meier-Schätzer zur Überlebensfunktion ist eine abfallende

[1] Das Wort *hazard* kommt ursprünglich aus dem Arabischen und bedeutet Würfel. Im übertragenen Sinn steht es für Chance und Risiko.

Treppenfunktion, die immer zu dem Zeitpunkt abfällt, wenn in der Stichprobe ein Ereignis vorgefallen ist. Bis kurz vor dem ersten Ereignis bleibt die geschätzte Überlebenskurve also auf 100%. Zum ersten Ereignis fällt die geschätzte Überlebenskurve auf einen niedrigeren Wert ab und bleibt bis kurz vor dem zweiten Ereignis konstant usw. Zensierungen führen zu keinem Abfall in der geschätzten Überlebenskurve, aber bei dem nächsten Ereignis ist der Abfall um so größer, da die zensierten Beobachtungen nicht mehr mitzählen, weil nichts mehr über sie bekannt ist.
Die Form der geschätzten Überlebenskurve ist also klar. Es fehlt noch, um wieviel die Überlebenskurve bei jedem Ereignis abfällt. Das soll jetzt hergeleitet werden. Dazu sortiert man alle Ereigniszeiten in der Stichprobe aufsteigend. Gibt es etwa m *verschiedene* Ereigniszeiten in der Stichprobe, dann bezeichnen wir mit $t_{(i)}$ die i-te Ereigniszeit und bekommen also:

$$t_{(1)} < t_{(2)} < ... < t_{(m)}.$$

Nun gehen wir folgendermaßen vor, um die Überlebensfunktion $S(t)$ zu schätzen:

- Wir zerlegen die Zeitachse in Intervalle, auf denen die Überlebenskurve konstant ist. Diese Intervalle beginnen *mit* einem Ereignis und enden *vor* dem nächsten Ereignis. Das erste Intervall I_0 reicht von $t = 0$ bis kurz vor das erste Ereignis $t_{(1)}$. Das nächste Intervall I_1 schließt sich daran an und beginnt mit der ersten Ereigniszeit $t_{(1)}$ und reicht bis kurz vor $t_{(2)}$, usw. Das letzte Intervall I_m startet mit dem letzten Ereignis $t_{(m)}$ und reicht bis ins Unendliche $t = \infty$.

  In mathematischer Notation bezeichnet $[a, b)$ das Intervall von der Zahl a bis zur Zahl $b > a$, wobei a zum Intervall dazugehört (eckige Klammer) und b nicht dazugehört (runde Klammer).

$$\begin{aligned} I_0 &= [0, t_{(1)}) \\ I_1 &= [t_{(1)}, t_{(2)}) \\ &\vdots \\ I_{m-1} &= [t_{(m-1)}, t_{(m)}) \\ I_m &= [t_{(m)}, \infty) \end{aligned}$$

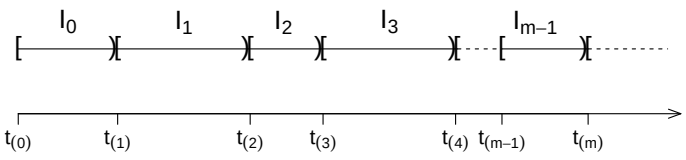


Abbildung 7.3: Zerlegung der Zeitachse in Intervalle

- Für jedes Intervall zählt man, wie viele Patienten am Anfang des Intervalls beobachtet wurden. Dazu gehören auch die Patienten, bei denen zu diesem Zeitpunkt das Ereignis eintritt. Man spricht von den „Patienten unter Risiko“. Außerdem zählt man die Anzahl der Ereignisse im Zeitintervall, d.h. wie viele Ereignisse zu Beginn des Intervalls auftreten. Die Anzahl der Patienten unter Risiko im Intervall I_j bezeichnen wir mit r_j, und mit d_j meinen wir die Anzahl der Ereignisse im Intervall I_j.

- Betrachten wir einen Patient, der bis zu einem Intervall I_j überlebt hat, dann können wir dessen Überlebenswahrscheinlichkeit p_j für das Intervall I_j abschätzen. Dazu betrachtet man die Patienten, die zu Anfang des Intervalls „unter Risiko“ standen und rechnet den Anteil der Patienten aus, die bis zum Ende des Intervalls *kein* Ereignis hatten. Also,

$$\hat{p}_j = \frac{r_j - d_j}{r_j} = 1 - \frac{d_j}{r_j}.$$

  Mathematisch betrachtet ist $\hat{p}_j$ eine Schätzung für die bedingte Wahrscheinlichkeit $P(T > t \,|\, T \geq t_{(j)})$ für einen Zeitpunkt t im Intervall I_j ($t \in I_j$). Es ist anders ausgedrückt die Wahrscheinlichkeit, das Intervall I_j zu überleben, wenn man das Intervall I_j lebend erreicht hat

$$P(T \geq t_{(j+1)} \,|\, T \geq t_{(j)}).$$

  Zensierungen in einem Intervall spielen für die Berechnung der dortigen Überlebenswahrscheinlichkeit keine Rolle. Allerdings wirken sie sich mittelbar aus, weil die Anzahl der Patienten unter Risiko in späteren Intervallen durch die Zensierung kleiner wird. Bezeichnet c_j die Anzahl der Zensierungen im Intervall I_j, dann ist nämlich $r_{j+1} = r_j - d_j - c_j$.

Im ersten Intervall I_0 gibt es per Definition keine Ereignisse ($d_0 = 0$). Daher wird die Überlebenswahrscheinlichkeit in I_0 immer geschätzt durch $\hat{p}_0 = P(T \geq t_{(1)} | T \geq 0) = 1$. Da alle Patienten eine Überlebenszeit ≥ 0 haben, kann man die Bedingung $T \geq 0$ auch weglassen, d.h. für das erste Intervall ist $\hat{p}_0 = P(T \geq t_{(1)}) = 1$. Die Wahrscheinlichkeit, bis mindestens zum ersten Ereignis in der Stichprobe zu überleben, wird immer mit 1 abgeschätzt.

- Um nun für einen beliebigen Zeitpunkt t die Überlebenswahrscheinlichkeit $S(t)$ zu schätzen, multipliziert man alle bedingten Überlebens-Wahrscheinlichkeiten $\hat{p}_j$ der Intervalle von I_0 bis zu dem Intervall I_k, das den Zeitpunkt t enthält. Also

$$\hat{S}(t) = P(T > t) = \prod_{j:t_{(j)} \leq t} \hat{p}_j = \hat{p}_0 \cdot \hat{p}_1 \cdot \ldots \cdot \hat{p}_{k-1} \cdot \hat{p}_k$$

Die Überlebensfunktion kann also schrittweise mit den früheren Werten ausgerechnet werden:

$$\hat{S}(t_{(i)}) = \hat{S}(t_{(i-1)}) \cdot \hat{p}_i.$$

Die unbedingte Wahrscheinlichkeit $P(T > t)$ errechnet sich als ein Produkt lauter bedingter Wahrscheinlichkeiten. Als Begründung, warum man das hier so machen kann, wird die Formel für einen Zeitpunkt $t \in I_1$ hergeleitet. Es ist also $t_{(1)} \leq t < t_{(2)}$. Dann ist:

$$S(t) = P(T > t) = P(T > t \cap T \geq t_{(1)}) = P(T > t \,|\, T \geq t_{(1)}) \cdot P(T \geq t_{(1)}) = \hat{p}_1 \cdot \hat{p}_0$$

Die Ereignisse $\{T > t\}$ und $\{T > t \cap T \geq t_{(1)}\}$ sind gleich: Wenn $T > t$ ist, dann gilt immer auch $T \geq t_{(1)}$, weil $t \geq t_{(1)}$ ist. Bei dem dritten Gleichheitszeichen wird die Definition von bedingten Wahrscheinlichkeiten benutzt. Für andere Zeitpunkte in späteren Intervallen geht die Herleitung analog.

Zum Verständnis des Kaplan-Meier Schätzers ist es sehr hilfreich, die Rechenschritte an einem kleinen Beispiel mit der Hand nachzuvollziehen.

Beispiel: 7.2.1

Eine Gruppe von Leukämie-Patienten wird mit dem Zytostatikum 6-Mercaptopurin behandelt. Anfangszeitpunkt ist der Therapiebeginn und Zielereignis ist Tod jedweder Art. Die Überlebensfunktion schätzt dann wirklich im wörtlichen Sinne das Überleben der Patienten zur Zeit t seit Therapiebeginn. In Tabelle 7.1 sind die fiktiven Daten wiedergegeben:

Patienten ID	Zeit bis zum Tod bzw. Zensierung (in Monaten)	Status 1=Tod, 0=zensiert)
1	10	1
2	19	0
3	6	1
4	28	0
5	32	1
6	11	1
7	13	0
8	16	1
9	6	1
10	24	1

Tabelle 7.1: Überlebenszeiten der 6-Mercaptopurin Gruppe

Fortsetzung des Beispiels

Die Stichprobe besteht aus $n = 10$ Patienten, bei denen innerhalb der Studie 7 Todesfälle an $m = 6$ verschiedenen Zeitpunkten beobachtet werden. Die restlichen Patienten sind zensiert. Zunächst werden die Patienten nach ihren beobachteten Zeiten (Ereignis oder Zensierung) in aufsteigender Reihenfolge geordnet. Dann können die $\hat{p}_i$ berechnet werden.

Ereigniszeit t_i	Anzahl unter Risiko r_i	Anzahl Zielereignisse zu t_i d_i	$\hat{p} = \frac{r_i - d_i}{r_i}$	geschätzte Überlebensfunktion $\hat{S}(t)$
0	10	0	1	1
6	10	2	0.8	0.8
10	8	1	0.875	0.7
11	7	1	0.857	0.6
16	5	1	0.8	0.48
24	3	1	0.667	0.32
32	1	1	0	0

Tabelle 7.2: geschätzten Überlebenswahrscheinlichkeiten

Abbildung 7.4 stellt die Überlebenszeitkurve grafisch dar. Zensierungen sind durch senkrechte Striche markiert. Ohne Zensierungen liefert der Kaplan-Meier Schätzer $\hat{S}(t)$ einfach den Anteil der lebenden Patienten in der Stichprobe zum Zeitpunkt t.

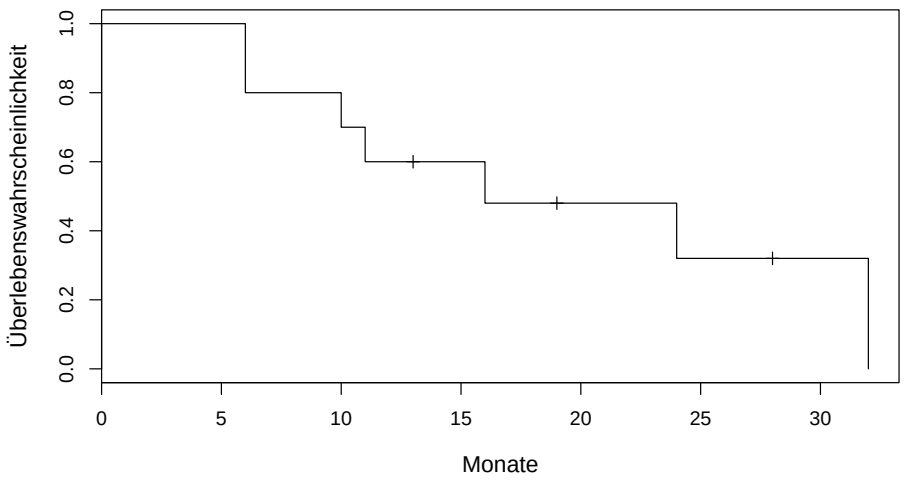


Abbildung 7.4: Geschätzte Überlebenszeitkurve

Fortsetzung des Beispiels

In der folgenden Abbildung 7.5 wird gezeigt, wie Zensierungen beim Kaplan-Meier Schätzer berücksichtigt werden. Die Größe der Figuren (Patienten) stellt deren Beitrag zur Überlebensrate dar. Beim Auftreten einer Zensierung verschwindet eine Figur. Der frei gewordene Platz wird unter den verbleibenden Figuren gleichmäßig aufgeteilt. Deswegen fällt beim nächsten Ereignis die Überlebenszeitkurve um einen größeren Betrag als vor dem Auftreten einer Zensierung.

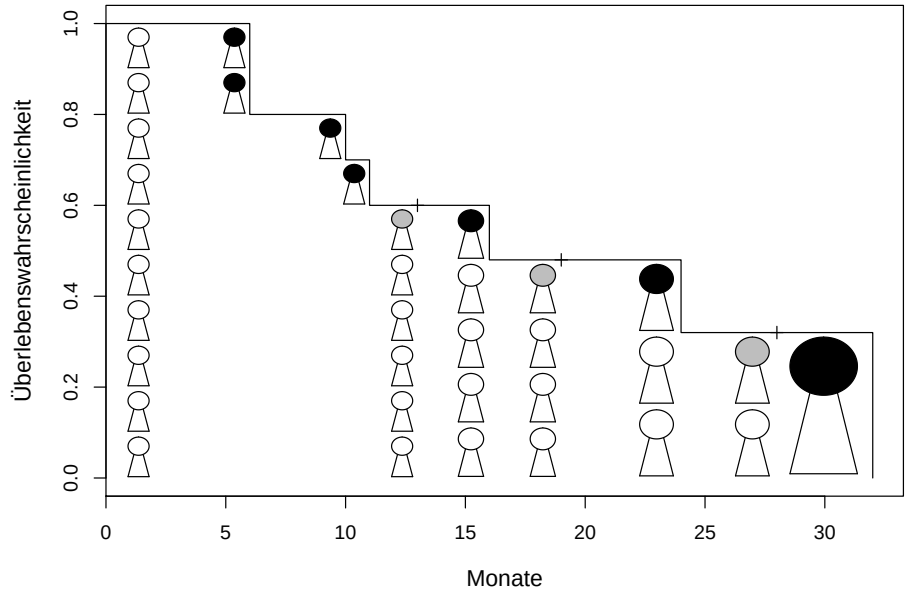


Abbildung 7.5: Veranschaulichung des Kaplan-Meier Schätzers (nach H. Tüchler)

7.2.3 Konfidenzintervalle für Überlebenszeiten

Die aus der Stichprobe der Beobachtungsdaten gewonnenen Überlebenswahrscheinlichkeiten sind Schätzwerte für die wahren Überlebenswahrscheinlichkeiten und somit fehlerbehaftet. Deswegen ist es sinnvoll, zur geschätzten Überlebenswahrscheinlichkeit $\hat{S}(t)$ zum Zeitpunkt t ein Konfidenzintervall anzugeben. Die *Formel von Greenwood* ist eine Berechnungsvorschrift für die Varianz des Kaplan-Meier Schätzers in Abhängigkeit von t.

$$\begin{aligned}\mathrm{Var}(\hat{S}(t)) &= \hat{S}(t)^2 \cdot \sum_{t_{(i)} \le t} \frac{d_i}{r_i(r_i - d_i)} \\ &= \hat{S}(t)^2 \cdot \Big(\frac{d_0}{r_0(r_0 - d_0)} + \frac{d_1}{r_1(r_1 - d_1)} + \frac{d_2}{r_2(r_2 - d_2)} + ... + \frac{d_j}{r_j(r_j - d_j)} \Big),\end{aligned}$$

wobei t im Intervall I_j liegt ($t \in I_j$). Für Zeiten vor dem ersten Ereignis $t \in I_0$ ergibt sich eine Varianz von 0. Die Wurzel aus dem Varianz-Term ist der Standard-Fehler (SE) zum Schätzer $\hat{S}(t)$. Damit kann das 95%-Konfidenzintervall zu einem festen Zeitpunkt t näherungsweise durch $\hat{S}(t) \pm 1,96 \cdot SE$ angegeben werden. Bei dieser Methode wird eine Normalverteilung des Schätzers $\hat{S}(t)$ vorausgesetzt.
Es gibt auch alternative Methoden, um das Konfidenzintervall zu bestimmen. Zum Beispiel kann man eine Logarithmus-Transformation benutzen. Diese Methoden vermeiden, dass das Konfidenzintervall über die Grenzen 0 und 1 hinausgehen kann. Jede Statistik-Software hat ihre eigene bevorzugte Standard-Methode für die Konfidenzintervalle. R zum Beispiel nutzt standardmäßig die Log-Transformation. Die Methode zur Abschätzung des Konfidenzintervalls sollte deswegen mit angegeben werden.

7.2.4 Mediane Überlebenszeit

Um die Verteilung einer Messgröße zu beschreiben, wird häufig der Mittelwert als Lagemaß angegeben. Weil die Verteilung von Überlebenszeitdaten oft hochgradig schief ist, ist bei Überlebenszeiten der Median als Lagemaß die oft besser geeignete Größe. Die geschätzte mediane Überlebenszeit ist der früheste Zeitpunkt, für den $\hat{S}(t) \leq 0,5$ gilt. Es kann jedoch sein, dass die aus der Stichprobe geschätzte Überlebensfunktion gar nicht unter den Wert 0,5 fällt. Dann ist die mediane Überlebenszeit mit den vorhandenen Daten nicht schätzbar.

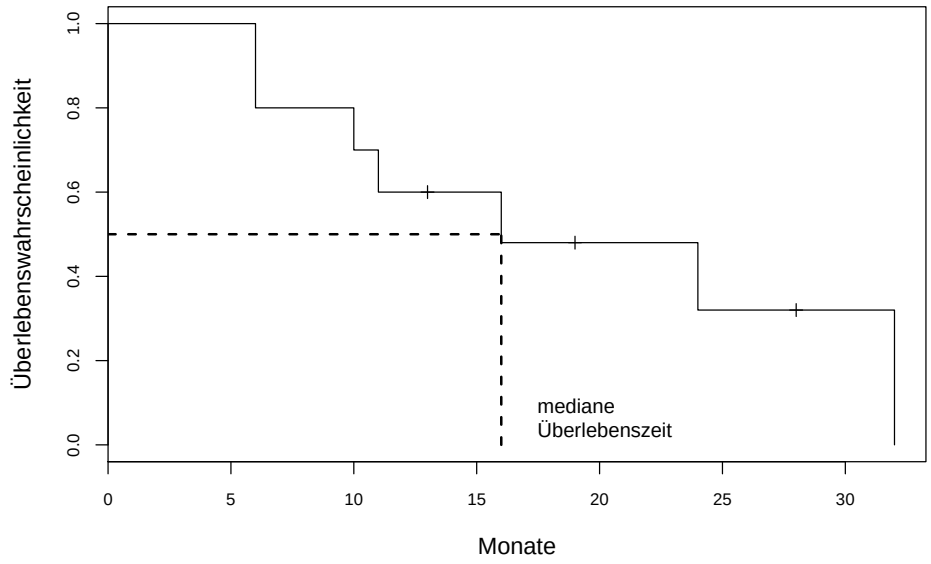


Abbildung 7.6: Mediane Überlebenszeit

Auch für den Punktschätzer zum medianen Überleben kann und sollte man ein Konfidenzintervall angeben. Diese Konfidenzintervall liegt dann auf der Zeitachse um die geschätzte mediane Ereigniszeit. Sehr häufig wird dazu die Methode von Brookmeyer & Crowley angewandt, die in zwei Schritten vorgeht:

1. Es wird ein Konfidenzintervall (auf der y-Achse) für die Überlebens-Wahrsscheinlichkeit 0,5 bestimmt. Das ist im obigen Abschnitt 7.2.3 beschrieben.

2. Die Grenzen dieses Konfidenzintervalls auf der y-Achse werden dann auf die x-Achse (Überlebens-Zeit) projiziert und bilden das Konfidenzintervall für die mediane Überlebenszeit.

Die mittlere Überlebenszeit kann natürlich auch ausgerechnet werden. Treten Zensierungen in der Stichprobe auf, ist dies allerdings nicht einfach das arithmetische Mittel der Ereigniszeiten, sondern ergibt sich nach einer komplizierteren Berechnungsvorschrift, auf die hier nicht weiter eingegangen wird.

7.3 Vergleich der Überlebenszeiten mehrerer Gruppen

Hat man in einer Studie zwei Gruppen (zum Beispiel neue Therapie versus Standard-Therapie), will man nicht nur das Überleben in beiden Gruppen beschreiben (etwa durch zwei Kaplan-Meier Überlebenskurven), sondern die entscheidende Frage wird sein, ob die neue Therapie ein längeres Überleben ermöglicht. Der bekannteste statistische Verfahren dazu ist der Log-Rank Test.

7.3.1 Der Log-Rank Test

Fragestellung. Unterscheidet sich die Verteilung der Überlebenszeiten zwischen zwei Gruppen?

Hypothesen. Die Nullhypothese H_0 beschreibt, dass beide Gruppen 1 und 2 sich bezüglich Überlebenszeit *nicht* unterscheiden. $H_0 : S_1 = S_2$. Die Alternativ-Hypothese H_A besagt entsprechend, dass die Überlebenszeiten in beiden Gruppen unterschiedlich verteilt sind.

Testgröße. Gilt H_0, dann sind beobachtete Unterschiede in der Überlebenszeit nur zufallsbedingt. Sind die beobachteten Unterschiede allerdings sehr groß, ist die Nullhypothese nicht plausibel und wird abgelehnt, d.h. man spricht dann von signifikanten Unterschieden zwischen den Überlebenszeiten beider Gruppen. Ab wann die beobachteten Unterschiede als „zu groß“ gelten, wird durch eine χ^2-verteilte Testgröße anhand der Daten bestimmt.
Im Kapitel zu speziellen Tests wurde bereits gesagt, dass χ^2-Testgrößen immer *beobachtete* Größen mit den entsprechenden, unter H_0 zu *erwartenden* Größen vergleichen. Unter H_0 („keine Unterschiede“) erwartet man, dass bei jeder Ereigniszeit die Ereignisse anteilig auf beide Gruppen verteilt sind. Sind etwa zu einer Ereigniszeit mit zwei Ereignissen in beiden Gruppen gleich viele Menschen unter Risiko, dann erwartet man unter H_0 je ein Ereignis in beiden Gruppen. Zu jedem Ereigniszeitpunkt kann man für beide Gruppe den Unterschied zwischen erwarteter Anzahl an Ereignissen und der Anzahl der tatsächlich beobachteten Ereignisse ermitteln. Für die Testgröße reicht es, die Unterschiede für eine der beiden Gruppen zu betrachten, da sie sich lediglich um ein Vorzeichen unterscheiden. Im folgenden nutzen wir die Unterschiede bei Gruppe 1.
Für die Testgröße des Log-Rank Tests werden die Differenzen für Gruppe 1 zu jedem Ereignis-Zeitpunkt in der Stichprobe gebildet, aufsummiert und quadriert. Anschließend wird diese Zahl noch durch einen Koeffizient geteilt, der ein Maß dafür ist, wie variabel die Anzahlen der beobachteten Ereignisse zu den verschiedenen Zeitpunkten in der Stichprobe sind.
Analog zur bisherigen Bezeichnung seien r_{i1} und r_{i2} die Anzahl der Patienten aus Gruppe 1 und Gruppe 2 unter Risiko zum Zeitpunkt t_i. d_{i1} und d_{i2} bezeichne die Zahl der Ereignisse in Gruppe 1 und Gruppe 2 zum Zeitpunkt t_i. Die Gesamtzahl der Patienten unter Risiko zum Zeitpunkt t_i ist dann $r_i = r_{i1} + r_{i2}$, und die Gesamtzahl der Ereignisse zum Zeitpunkt t_i ist $d_i = d_{i1} + d_{i2}$.
Der Anteil von Gruppe 1 an den Patienten unter Risiko zum Zeitpunkt t_i ist gerade $\frac{r_{i1}}{r_i}$. Unter H_0 ist deswegen die erwartete Anzahl von Zielereignissen in Gruppe 1 zum Zeitpunkt t_i

$$e_{i1} = d_i \cdot \frac{r_{i1}}{r_i}$$

Die Differenz für Gruppe 1 zur i-ten Ereigniszeit ist dann $d_{i1} - e_{i1}$ und die Testgröße errechnet sich damit als

$$X = \frac{\left[\sum_{i=1}^{m}(d_{i1} - e_{i1})\right]^2}{\sum_{i=1}^{m} v_i},$$

wobei die $v_i = \frac{d_i(r_i - d_i)r_{i1}r_{i2}}{{r_i}^2(r_i - 1)}$ ein Maß für die Varianz von d_i ist. v_i ist dann groß, wenn sich mehrere Ereignisse an einer Ereigniszeit häufen.

Testentscheidung. Man kann zeigen, dass unter H_0 die Testgröße X einer χ^2-Verteilung mit einem Freiheitsgrad genügt. So sind unter H_0 zum Beispiel große Werte von X unwahrscheinlich. Über die χ^2-Verteilung kann eine Statistik-Software zu X den zugehörigen p-Wert bestimmen. Je größer die Werte von X, desto kleiner wird der p-Wert. Bei vorgegebenem Signifikanzniveau α wird die Nullhypothese H_0 genau dann abgelehnt, wenn der p-Wert kleiner als α ist.
Wird H_0 abgelehnt, dann sieht man anhand der Kaplan-Meier Überlebenskurven für beide Gruppen, welche Gruppe ein signifikant längeres Überleben zeigt.

Voraussetzungen. Wie beim Kaplan-Meier Schätzer sollten Zensierungen nicht-informativ sein, d.h. unabhängig sein vom Auftreten von Ereignissen. Darüber hinaus sollten Zensierungen auch unabhängig von der Gruppenzugehörigkeit sein, d.h. es verzerrt den Test, wenn sich etwa in einer der beiden Gruppen die Zensierungen stark häufen. Um den p-Wert korrekt auszurechnen, ist es weiterhin wichtig, dass insgesamt die Summe an erwarteten Ereignissen groß ist.
Der Log-Rank Test setzt außerdem voraus, dass das Verhältnis des Ereignisrisikos (Hazard) zwischen beiden Gruppen über die Zeit hinweg konstant ist („proportionale Hazard“). Diese Annahme ist etwa dann verletzt, wenn sich beide Kaplan-Meier Überlebenskurven deutlich kreuzen.

Beispiel: 7.3.1

Zu dem Beispieldatensatz aus Tab. 7.1 gibt es eine Vergleichsgruppe (Patienten, die nicht mit 6-Mercatopurin behandelt wurden):

Patienten ID	Zeit bis zum Tod bzw. Zensierung (in Monaten)	Status 1=Tod, 0=zensiert)
11	5	1
12	1	1
13	8	1
14	12	1
15	8	1
16	2	1
17	22	1
18	17	1
19	2	1
20	10	0

Tabelle 7.3: Schätzung der Überlebenszeiten Gruppe 2

Um einen ersten Eindruck zu den Überlebenszeiten der beiden Gruppen zu bekommen, werden die geschätzten Überlebensfunktionen im Vergleich dargestellt:

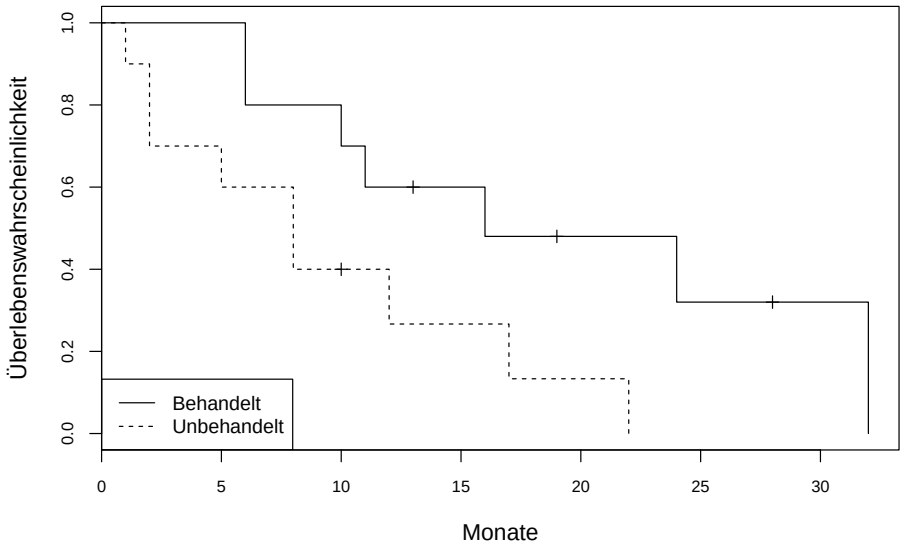


Abbildung 7.7: Überlebenszeiten der beiden Gruppen

Fortsetzung des Beispiels

Die Abbildung zeigt die geschätzte Überlebensfunktionen der beiden Gruppen aus dem Beispieldatensatz im Vergleich.

	t_i	r_i	d_i	r_{i1}	d_{i1}	r_{i2}	d_{i2}	e_{i1}	e_{i2}	$d_{i1} - e_{i1}$	$d_{i2} - e_{i2}$	v_i
	1	20	1	10	0	10	1	0.5	0.5	-0.5	0.5	0.5
	2	19	2	10	0	9	2	1.053	0.947	-1.053	1.053	0.895
	5	17	1	10	0	7	1	0.588	0.412	-0.588	0.588	0.412
	6	16	2	10	2	6	0	1.25	0.75	0.75	-0.75	0.7
	8	14	2	8	0	6	2	1.143	0.857	-1.143	1.143	0.791
	10	12	1	8	0	4	1	0.667	0.333	-0.667	0.667	0.333
	11	10	1	7	1	3	0	0.7	0.3	0.3	-0.3	0.3
	12	9	1	6	0	3	1	0.667	0.333	-0.667	0.667	0.333
	13	8	0	6	0	2	0	0	0	0	0	0
	16	7	1	5	1	2	0	0.714	0.286	0.286	-0.286	0.286
	17	6	1	4	0	2	1	0.667	0.333	-0.667	0.667	0.333
	19	5	0	4	0	1	0	0	0	0	0	0
	22	4	1	3	0	1	1	0.75	0.25	-0.75	0.75	0.25
	24	3	1	3	1	0	0	1	0	0	0	0
	28	2	0	2	0	0	0	0	0	0	0	0
	32	1	1	1	1	0	0	1	0	0	0	0
$\sum$			16		6		10	10.699	5.301	-4.699	4.699	5.133

Tabelle 7.4: Rechnung zum LogRank Test

$$X = \frac{\left[\sum_{i=1}^{m}(d_{i1} - e_{i1})\right]^2}{\sum_{i=1}^{m} v_i} = \frac{(-4.699)^2}{5.133} = 4.302$$

Die Testgröße X ist unter der Nullhypothese χ^2_1-verteilt, und ein Wert größer als 4 ist nicht sehr wahrscheinlich. Den zugehörigen p-Wert $p = 0,0363$ ermittelt eine Statistik-Software. Auf dem Signifikanzniveau von $\alpha = 0,05$ ist der Test signifikant, da $p < \alpha$. Die Nullhypothese wird also abgelehnt, d.h. wir gehen nicht davon aus, dass die Überlebenszeiten in beiden Gruppen gleich lang sind. An den Überlebenskurven sieht man, dass die zusätzliche Therapie in Gruppe 1 zu einer signifikanten Verbesserung führt.

7.3.2 Varianten des Log-Rank Tests

Beim Log-Rank Test werden alle Ereigniszeitpunkte gleich gewichtet. Es gibt Varianten zum Log-Rank Test, die die Differenzen $d_{1i} - e_{1i}$ zu den verschiedenen Zeitpunkten unterschiedlich stark bewerten. Stärkeres Gewicht auf die frühen Zeitpunkte legen der **Gehan Test** (auch **Breslow Test** genannt) und der **Peto-Peto Test**. Umgekehrt legt der **Fleming-Harrington Test** mehr Gewicht auf die späten Ereigniszeiten in den Stichproben.
Entsprechend können Unterschiede der Überlebenskurven in den frühen Zeitpunkten mit dem Gehan-Test und dem Peto-Peto Test besser aufgedeckt werden als mit dem Log-Rank Test. Umgekehrt findet der Fleming-Harrington Test eher Unterschiede zu späteren Zeitpunkten als der Log-Rank Test.
Welche Test-Variante man wählt, kann also einen großen Einfluss haben und sollte davon abhängen, ob man inhaltlich in der Studie eher frühe oder späte Unterschiede zwischen den Überlebenskurven aufdecken will. Insbesondere ist es wissenschaftlich nicht sauber, wenn man alle möglichen Test-Varianten ausprobiert und diejenige Variante wählt, die den kleinsten p-Wert liefert.

7.3.3 Erweiterungen zum Log-Rank Test

Alle Test-Varianten des Log-Rank Tests lassen sich auf den Vergleich von *mehr als zwei Gruppen* verallgemeinern. Bei mehr als zwei Gruppen bedeutet eine Ablehnung der Nullhypothese, dass nicht alle Überlebenszeitkurven übereinstimmen. Eine Aussage, welche Unterschiede im einzelnen vorliegen, erhält man damit jedoch nicht. Um dies zu untersuchen, sind paarweise Vergleiche erforderlich. Dabei ist zu beachten, dass bei wiederholten Tests im allgemeinen die Signifikanzschranke nach unten korrigiert werden muss (siehe auch 4.8).
Bei mehr als zwei Gruppen ist es auch möglich, dass man einen bestimmten *Trend* zwischen den Gruppen testet. Etwa, wenn man die Überlebenskurven von Patienten in vier Altersklassen unter einer bestimmten Therapie vergleichen will, und die Vermutung hat, dass das Überleben günstiger ist, je jünger die Patienten sind, so kann dieser Trend als spezielle Alternativ-Hypothese getestet werden.
Der Log-Rank Test lässt sich auch dahingehend erweitern, dass er mit *Schichten* innerhalb der Vergleichsgruppen umgehen kann. Ein Beispiel wäre etwa ein Therapievergleich zwischen zwei Therapien. In den beiden Studienarmen ist die Altersverteilung unterschiedlich. Wenn Alter einen Einfluss hat, sollte man für diesen Alterseffekt korrigieren. Dazu kann man Altersgruppen („jung“, „mittel“ und „alt“) als Schichtvariable definieren. Der geschichtete Log-Rank Test führt den Vergleich in den einzelnen Altersschichten durch und führt diese Teil-Ergebnisse in ein einziges Ergebnis zusammen. Annahme dabei ist allerdings, dass die Therapie in allen Schichten ähnlich wirkt. Zum Abschluss des Kapitels zur Überlebenszeitanalyse soll an einem Realbeispiel gezeigt werden, dass die Definition des Zielereignisses einen wesentlichen Einfluss auf die Form der Überlebenskurven und das Ergebnis von Log-Rank Tests hat.

Beispiel: 7.3.2

Beim fortgeschrittenen Morbus Hodgkin erreichen etwa 75% der Patienten mit einer Standardchemotherapie eine komplette Remission. Allerdings haben diese Patienten ein hohes Rezidivrisiko. Um das Rezidivrisiko zu senken, hat man versuchsweise in einigen europäischen Kliniken bei Patienten, die mit Standardchemotherapie eine komplette Remission erreicht hatten, eine Hochdosischemotherapie mit anschließender Kochenmarktransplantation durchgeführt. Diese Patienten wurden mit einer passend ausgewählten Gruppe von konventionell mit Standardchemotherapie behandelten Patienten der Deutschen Hodgkin Studie verglichen.
Es handelt sich hier um Daten einer veröffentlichten Analyse der Deutschen Hodgkin Studiengruppe (GHSG) in Zusammenarbeit mit der European Blood and Marrow Transplantation Group (EBMT) (Schmitz N. et al., Early high-dose to consolidate patients with high-risk Hodgkin's disease in first complete remission? Results of an EBMT-GHSG matched-pair analysis, ASH 1995, Abstract 1742).

Fortsetzung des Beispiels

1. Zielereignis: Rezidiv des Morbus Hodgkin
   Zensiert sind hier alle Verstorbenen ohne vorheriges Rezidiv sowie alle lebenden Patienten, die zum Zeitpunkt des Ausscheidens rezidivfrei waren.

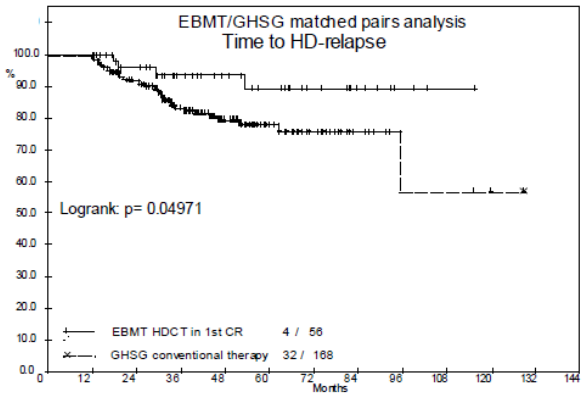


Abbildung 7.8: Rezidiv des Morbus Hodgkin

Es zeigt sich hier eine auf dem 5% Niveau signifikante Überlegenheit der mit zusätzlicher Hochdosis-Chemotherapie behandelten Gruppe. Aber: die Hochdosis-Chemotherapie ist eine riskante Therapie mit akuten Toxizitäten, die bisweilen tödlich sind. Daher ist es für den Vergleich der beiden Gruppen sicherlich verfälschend, wenn die rezidivfrei Verstorbenen zensiert werden. Diese Überlegung führt zum zweiten Vergleich:

2. Zielereignis: Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache, je nachdem, welches dieser beiden Ereignisse zuerst auftritt.
   Zensiert sind hier also alle zum Zeitpunkt des Ausscheidens rezidivfrei lebenden Patienten.

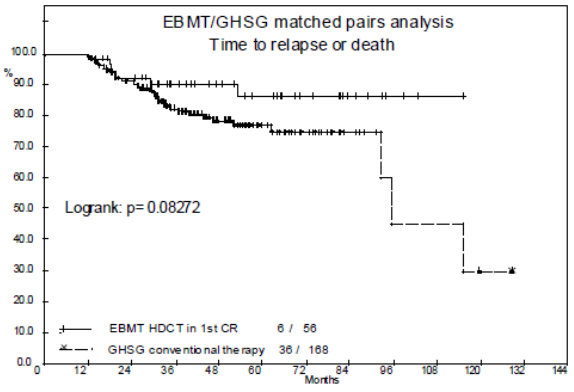


Abbildung 7.9: Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache

Der Unterschied zwischen den beiden Gruppen wird geringer, der Vergleich ist formal nicht mehr signifikant auf dem 5% Niveau.
Mit der Hochdosis-Chemotherapie ist auch ein zusätzlich (zu der Standard-Chemotherapie) erhöhtes Risiko einer Sekundärneoplasie (insbesondere Leukämien) verbunden. Berücksichtigt man die aufgetretenen Sekundärneoplasien zum Zeitpunkt ihres Auftretens, so ergibt sich:

Fortsetzung des Beispiels

3. Zielereignis: Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache oder Sekundärneoplasie, je nachdem, welches dieser Ereignisse zuerst auftritt.
   Zensiert sind hier alle zum Zeitpunkt des Ausscheidens rezidiv- und sekundärneoplasiefrei lebenden Patienten.

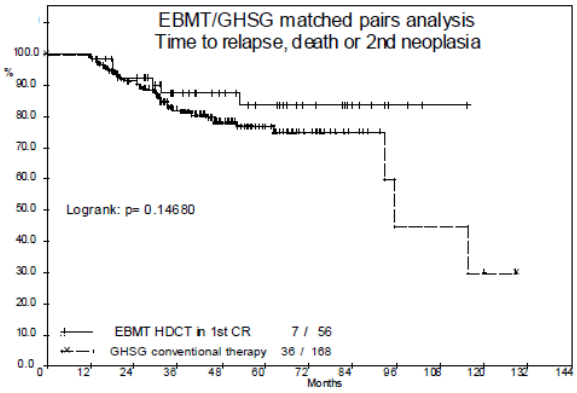


Abbildung 7.10: Rezidiv des Morbus Hodgkin oder Tod jeglicher Ursache oder Sekundärneoplasie

Der Unterschied zwischen beiden Gruppen wird durch die zusätzliche Betrachtung der Sekundärneoplasien noch geringer. Betrachtet man nun Tod jeglicher Ursache als Zielereignis, so ergibt sich folgendes Bild:

4. Zielereignis: Tod jeglicher Ursache
   Zensiert sind hier alle zum Zeitpunkt des Ausscheidens noch lebenden Patienten, auch diejenigen mit Rezidiv oder Sekundärneoplasie.

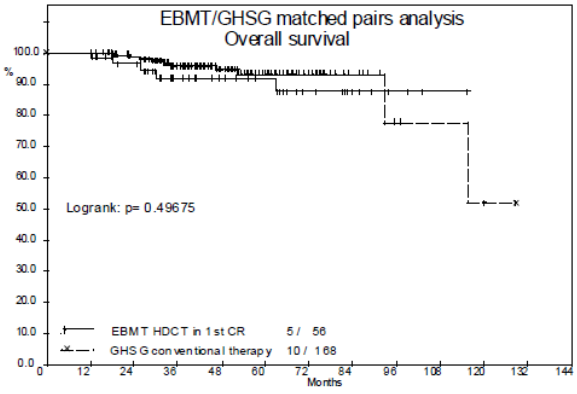


Abbildung 7.11: Tod jeglicher Ursache

Hier ist nun keinerlei Unterschied mehr zwischen den beiden Gruppen zu erkennen. Dies lässt sich damit erklären, dass Patienten, die mit der Standardchemotherapie behandelt wurden, im Falle eines Rezidivs noch eine gute Chance haben, erfolgreich therapiert zu werden (z.B. mit Hochdosischemotherapie), während für Patienten, die gleich mit Hochdosistherapie behandelt werden, dies nicht mehr zuzutreffen scheint.

Das Beispiel zeigt, dass man, je nach Wahl des Endpunktes, unterschiedliche Ergebnisse beim Vergleich zweier Gruppen bekommen kann. Es ist also sehr wichtig, bei der Betrachtung von Überlebenszeitkurven genau darauf zu achten, welcher Endpunkt gewählt wurde, und was für Patienten aufgrund dieser Wahl als zensiert angesehen wurde.

8 Appendix

8.1 Packages

```
library(gplots)
library(grid)
library(plotrix)
library(survival)
library(xtable)
```

8.2 Kapitel 1

8.2.1 Venn-Diagramme

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#Datenbereitstellung
r=2
a<-c(167:233)/100
b<-c(67:133)/100
area_x_coordinate <- c(5+r*cos(a*pi)-r/2, 5+r*cos(b*pi)+r/2)
area_y_coordinate <- c(5+1.8*r*sin(a*pi), 5+1.8*r*sin(b*pi))
#-------------------------------------------------------------------------------
#Plot
plot(0:10,0:10,type='n',xaxt='n',yaxt='n',xlab="",ylab="")
#Flächen/Linien
polygon(area_x_coordinate, area_y_coordinate, col="grey");
draw.circle(x=4,y=5,radius=2,lwd=2)
draw.circle(x=6,y=5,radius=2,lwd=2)
#Beschriftungen
text(x=3,y=5,"A")
text(x=7,y=5,"B")
text(x=1,y=1,expression(Omega))
mtext(text= expression(paste("A",intersect(B))),side=1,line=1)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#-------------------------------------------------------------------------------
#Plot
plot(0:10,0:10,type='n',xaxt='n',yaxt='n',xlab="",ylab="")
#Flächen/Linien
polygon(draw.circle(x=4,y=5,radius=2),col="grey")
polygon(draw.circle(x=6,y=5,radius=2),col="grey")
draw.circle(x=4,y=5,radius=2,lwd=2)
draw.circle(x=6,y=5,radius=2,lwd=2)
#Beschriftungen
text(x=3,y=5,"A")
text(x=7,y=5,"B")
text(x=1,y=1,expression(Omega))
mtext(text=expression(paste("A",union(B))),side=1,line=1)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#Datenbereitstellung
r=2
a<-c(167:233)/100
b<-c(67:133)/100
area_x_coordinate <- c(5+r*cos(a*pi)-r/2, 5+r*cos(b*pi)+r/2)
area_y_coordinate <- c(5+1.75*r*sin(a*pi), 5+1.75*r*sin(b*pi))
#------------------------------------------------------------------------------
#Plot
plot(0:10,0:10,type='n',xaxt='n',yaxt='n',xlab="",ylab="")
#Flächen/Linien
polygon(draw.circle(x=4,y=5,radius=2),col="grey",lwd=2)
draw.circle(x=6,y=5,radius=2,lwd=2)
polygon(area_x_coordinate, area_y_coordinate, col="white", lwd=2);
#Beschriftungen
text(x=3,y=5,"A")
text(x=7,y=5,"B")
text(x=1,y=1,expression(Omega))
mtext(text="A \\ B",side=1,line=1)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#------------------------------------------------------------------------------
#Plot
plot(0:10,0:10,type='n',xaxt='n',yaxt='n',xlab="",ylab="")
#Flächen/Linien
rect(par("usr")[1], par("usr")[3], par("usr")[2], par("usr")[4], col = "grey")
polygon(draw.circle(x=4,y=5,radius=2),col="white",lwd=2)
draw.circle(x=6,y=5,radius=2,lwd=2)
#Beschriftungen
text(x=3,y=5,"A")
text(x=7,y=5,"B")
text(x=1,y=1,expression(Omega))
mtext(text=expression(bar(A)),side=1,line=1)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#------------------------------------------------------------------------------
#Plot
plot(0:10,0:10,type='n',xaxt='n',yaxt='n',xlab="",ylab="")
#Flächen/Linien
draw.circle(x=2.5,y=5,radius=2,lwd=2)
draw.circle(x=7.5,y=5,radius=2,lwd=2)
#Beschriftung
text(x=1.5,y=5,"A")
text(x=8.5,y=5,"B")
text(x=1,y=1,expression(Omega))
mtext(text=expression(paste("A", intersect(B),"=",symbol("\306"))),side=1,line=1)
```

8.2.2 Wahrscheinlichkeitsfunktion vs. Verteilungsfunktion

```
#Grafikeinstellungen
par(mfrow = c(1,1),mar=c(3,3,3,2)+0.1)
#Datenbereitstellung
```

```
x1 <- 0:6
y1 <- c(0,rep(1/6,6))
#-------------------------------------------------------------------------------
#Plot
  plot(x1,y1,type="h",main="",ylab="",xlab="",ylim=c(0,0.5),axes=FALSE,lwd=2)
#Datenpunkte/Graph
  points(x1[-1],y1[-1],pch=19)
#Achsen
  axis(side=1,labels=TRUE,pos=0)
  axis(side=2,labels=FALSE,tck=0,pos=0)
  lines(x=c(-0.1,0),y=c(1/6,1/6))
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=0.4,x1=0,y1=0.5,angle=20,length=0.125)
    arrows(x0=6,y0=0,x1=6.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="Wahrscheinlichkeitsfunktion",side=1,adj=0.5,line=2,cex=1.2)
  mtext(text="P(X=x)",at=-0.5)
  mtext(text="1/6",side=2,at=1/6,las=1)
```

```
#Grafikeinstellungen
par(mfrow = c(1,1),mar=c(3,3,3,2)+0.1)
#Datenbereitstellung
x1 <- 0:6
y2 <- x1 / 6
#-------------------------------------------------------------------------------
#Plot
  plot(x1,y2,type="s",main="",ylab="",xlab="",ylim=c(0,1.2),axes=FALSE,lwd=2)
#Datenpunkte/Graph
  points(x1[-1],y2[-1],pch=19)
#Achsen
  axis(side=1, labels=TRUE, pos=0)
  axis(side=2, labels=FALSE,tck=0,pos=0)
    lines(x=c(-0.1,0), y=c(1/6,1/6))
    lines(x=c(-0.1,0), y=c(1,1))
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=1,x1=0,y1=1.2,angle=20,length=0.125)
    arrows(x0=6,y0=0,x1=6.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  lines(x=c(2,2),y=c(0,1/6), lty=2)
  lines(x=c(3,3),y=c(0,2/6), lty=2)
  lines(x=c(4,4),y=c(0,3/6), lty=2)
  lines(x=c(5,5),y=c(0,4/6), lty=2)
  lines(x=c(6,6),y=c(0,5/6), lty=2)
#Beschriftungen
  mtext(text="Verteilungssfunktion",side=1,adj=0.5,line=2,cex=1.2)
  mtext(text="P(X<=x)",at=-0.5)
  mtext(text="1/6",side=2,las=1,at=1/6)
  mtext(text="1",side=2,las=1,at=1)
```

8.2.3 Binomiale -Wahrscheinlichkeits- und Verteilungsfunktion

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(3,2,3,2)+0.1)
#Datenbereitstellung
x1 <- 0:4
y1 <- dbinom(x1,size=4,prob=.6)
x2 <- 0:6
y2 <- pbinom(q=x2,size=4,prob=.6)
#------------------------------------------------------------------------------
#Plot
  plot(x1,y1,type="h",
     main="",ylab="",xlab="",xlim=c(0,6),ylim=c(0,0.4),axes=FALSE,lwd=2)
#Datenpunkte/Graph
  points(x1,y1,pch=19)
#Achsen
  axis(side=1,labels=TRUE,pos=0)
  axis(side=2,labels=TRUE,pos=0)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=0.4,x1=0,y1=0.45,angle=20,length=0.125)
    arrows(x0=6,y0=0,x1=6.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="f",side=3,at=-0.5,line=1.)
  mtext(text="x",side=1,at=6.5,line=0.)
  legend("topleft",legend=c("WAHRSCHEINLICHKEITSFUNKTION",
                            "der Binomialverteilung Bin(4,06)"),bty="n",cex=0.8)
#------------------------------------------------------------------------------
#Plot
  plot(x2,y2,type="s",main="",ylab="",xlab="",ylim=c(0,1),axes=FALSE,lwd=2)
#Datenpunkte/Graph
  points(x2[0:5],y2[0:5],pch=19)
#Achsen
  axis(side=1,labels=TRUE,pos=0)
  axis(side=2,labels=TRUE,pos=0)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=1,x1=0,y1=1.1,angle=20,length=0.125)
    arrows(x0=6,y0=0,x1=6.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="F",side=3,at=-0.5,line=1.)
  mtext(text="x",side=1,at=6.5,line=0.)
  legend("topleft",legend=c("VERTEILUNGSFUNKTION","der Binomialverteilung",
                            "B(4,06)"),bty="n",cex=0.8)
```

8.2.4 Zusammenhang von Verteilungs- und Dichtefunktion

```
#Grafikeinstellungen
par(mfrow=c(2,1),mar=c(2,2,2,2)+0.1)
#Datenbereitstellung
a <- seq(from=-4, to=4, by=0.001)
fx <- dnorm(a)
Fx <- pnorm(a)
cord.x <- seq(from = -1.5, to = -0.5)
cord.y <- dnorm(cord.x)
#------------------------------------------------------------------------------
#Plot
```

```
  plot(a,Fx,type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=0,pos=0 )
  axis(side=2,labels=FALSE,tck=0, pos=-4)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=-4,y0=1,x1=-4,y1=1.1,angle=20,length=0.125)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  lines(c(-4.2,-0.5),c(pnorm(-0.5),pnorm(-0.5)),lty=2)
  lines(c(-4.2,-1.5),c(pnorm(-1.5),pnorm(-1.5)),lty=2)
  par(xpd=TRUE)
    lines(x=c(-1.5,-1.5),y=c(1,-7),lty=2)
    lines(x=c(-0.5,-0.5),y=c(1,-7),lty=2)
  par(xpd=FALSE)
#Beschriftungen
  mtext("Verteilungsfunktion",side=4,line=0)
  mtext(text="F(x)",side=3,at=-4.2,line=0)
  mtext(text="x",side=1,at=4.3,line=0.)
  mtext(text="F(x2)",side=2,at=pnorm(-0.5),line=0,las=1)
  mtext(text="F(x1)",side=2,at=pnorm(-1.5),line=0,las=1)
#-------------------------------------------------------------------------------
#Plot
  plot(a,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=0,pos=0 )
  axis(side=2,labels=FALSE,tck=0, pos=-4)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=-4,y0=0.4,x1=-4,y1=0.45,angle=20,length=0.125)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  arrows(x0=2,y0=0.28,x1=-1,y1=0.05,0.45,angle=20,length=0.125)
  par(xpd=TRUE)
    lines(x=c(-1.5,-1.5),y=c(2,0),lty=2)
    lines(x=c(-0.5,-0.5),y=c(2,0),lty=2)
  par(xpd=FALSE)
  polygon(x = c(-1.5, cord.x, -0.5),y = c(0, cord.y, 0),
          col = "black",density=10)
#Beschriftungen
  mtext("Dichtefunktion",side=4,line=0)
  mtext(text="f(x)",side=3,at=-4.2,line=0)
  mtext(text="x",side=1,at=4.3,line=0.)
  mtext(text="x2",side=1,at=-0.5,line=0,las=1)
  mtext(text="x1",side=1,at=-1.5,line=0,las=1)
  text(x=2,y=0.32, "P(x1<X<x2)",cex=0.8)
```

8.2.5 Variation des Erwartungswertes

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x1 <- seq(from=-10,to=10,by=0.01)
y1 <- dnorm(x1,mean=0,sd=1)
y2 <- dnorm(x1,mean=0,sd=1.5)
```

```
y3 <- dnorm(x1,mean=5,sd=1)
y4 <- dnorm(x1,mean=5,sd=1.5)
v1 <- c(0,0.1,0.2,0.3,0.4,0.5)
v2 <- c("","0.1","0.2","0.3","0.4","0.5")
#-------------------------------------------------------------------------------
#Plot
  plot(x1,y1,type="l",xlab="",ylab="",xlim=c(-5,10),ylim=c(0,0.5),axes=F,lwd=2)
#Datenpunkte/Graph
  lines(x1,y2,lty=2,lwd=2)
  lines(x1,y3,lwd=2)
  lines(x1,y4,lty=2,lwd=2)
#Achsen
  axis(side=2,pos=0,las=2,at=v1,labels=v2,cex.axis=0.8)
  axis(side=1,pos=0,cex.axis=0.8)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=0.5,x1=0,y1=0.55,angle=20,length=0.125)
    arrows(x0=10,y0=0,x1=10.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="f(x)",side=3,at=-0.5,line=0,cex=1.2)
  mtext(text="x",side=1,at=10.5,line=0,cex=1.2)
  text(x=-4.8,y=0.06,"N(0,1.5)")
  text(x=-1.8,y=0.36,"N(0,1)")
  text(x=6.8,y=0.36,"N(5,1)")
  text(x=8.8,y=0.06,"N(5,1.5)")
```

8.2.6 Intervalgrenzen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,0.1,3,0.1)+0.1)
#Datenbereitstellung
x1 <- seq(from=-10,to=10,by=0.01)
y1 <- dnorm(x1,mean=0,sd=1)
v1 <- c(0,0.1,0.2,0.3,0.4)
v2 <- c("","0.1","0.2","0.3","0.4")
#-----------------------------------------------
#Plot
  plot(x1,y1,type="l",xlab="",ylab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=FALSE,tck=0)
  axis(side=2,pos=-4,las=2,at=v1,labels=v2,cex.axis=0.8)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=-4,y0=0.4,x1=-4,y1=0.45,angle=20,length=0.125)
    arrows(x0=4,y0=0,x1=4.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  #vertikale
    lines(c(1,1),c(0,dnorm(1)),lty=2)
    lines(c(-1,-1),c(0,dnorm(-1)),lty=2)
    lines(c(2,2),c(0,dnorm(2)),lty=2)
    lines(c(-2,-2),c(0,dnorm(-2)),lty=2)
    par(xpd=TRUE)
      lines(c(3,3),c(-0.02,dnorm(3)),lty=2)
      lines(c(-3,-3),c(-0.02,dnorm(3)),lty=2)
    par(xpd=FALSE)
```

```
  #horizontale
    arrows(x0=0.5,y0=dnorm(1),x1=1,y1=dnorm(1),angle=20,length=0.125)
    arrows(x0=-0.5,y0=dnorm(-1),x1=-1,y1=dnorm(-1),angle=20,length=0.125)
    arrows(x0=0.5,y0=dnorm(2),x1=2,y1=dnorm(2),angle=20,length=0.125)
    arrows(x0=-0.5,y0=dnorm(-2),x1=-2,y1=dnorm(-2),angle=20,length=0.125)
    par(xpd=TRUE)
      arrows(x0=0.5,y0=-0.02,x1=3,y1=-0.02,angle=20,length=0.125)
      arrows(x0=-0.5,y0=-0.02,x1=-3,y1=-0.02,angle=20,length=0.125)
    par(xpd=FALSE)
#Beschriftungen
  mtext(text="f(x)",side=3,at=-4.4,line=0.5,cex=1.2)
  mtext(text="x",side=1,at=4.4,line=0,cex=1.2)
  text(x=0,y=dnorm(1),"68.3%")
  text(x=0,y=dnorm(2),"95.5%")
  par(xpd=TRUE)
    text(x=0,y=-0.02,"99.7%")
  par(xpd=FALSE)
```

8.2.7 Logarithmische Normalverteilung

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
n2 <- 4725
meanlog2 <- 2.5
sdlog2 <- .4
v1 <- c("","200","400","600","800","1000")
v2 <- c(0,0.02,0.04,0.06,0.08,0.1)
v3 <- c("","200","400","600","800","1000")
v4 <- c(0,0.2,0.4,0.6,0.8,1)
#-------------------------------------------------------------------------------
#Plot
  hist(rlnorm(n=n2,meanlog=meanlog2,sdlog=sdlog2),probability=T,col="grey",
       main="",xlab="Leukozyten im Blut [G/l]",ylab="",xlim=c(0,50),
       ylim=c(0,0.1),axes=FALSE)
#Datenpunkte/Graph
  lines(density(rlnorm(n=n2,meanlog=meanlog2,sdlog=sdlog2),adjust=4))
#Achsen
  axis(side=1,pos=0,labels=T,cex.axis=0.8)
  axis(side=2,pos=0,las=2,at=v2,labels=v1,cex.axis=0.8)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=0,y0=0.1,x1=0,y1=0.11,angle=20,length=0.125)
    arrows(x0=50,y0=0,x1=54,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="Patienten mit fortgeschrittenem Morbus Hodgkin",side=3,line=2)
  legend("topright",c(paste("N=",n2),paste("Mittel=",round(exp(meanlog2),1)),
                      paste("SD=",round(exp(sdlog2),1))),bty="n")
#-------------------------------------------------------------------------------
#Plot
  hist(log(rlnorm(n=n2,meanlog=meanlog2,sdlog=sdlog2)),probability=T,col="grey",
       main="",xlab="log(Leukozyten im Blut) [G/l]",ylab="",axes=F)
#Datenpunkte/Graph
  lines(density(log(rlnorm(n=n2,meanlog=meanlog2,sdlog=sdlog2)),adjust=4))
#Achsen
  axis(side=1,pos=0,labels=T,cex.axis=0.8)
```

```
  axis(side=2,pos=1,las=2,at=v4,labels=v3,cex.axis=0.8)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=1,y0=0,x1=1,y1=1.1,angle=20,length=0.125)
    arrows(x0=1,y0=0,x1=4.5,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext(text="Patienten mit fortgeschrittenem Morbus Hodgkin",side=3,line=2)
  legend("topright",c(paste("N=",n2),paste("Mittel=",meanlog2),
                      paste("SD=",sdlog2)),bty="n")
```

8.3 Kapitel 2

8.3.1 Balken- und Kreisdiagramm

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(3,3,3,4)+0.1)
#Datenbereitstellung
x <- c(677,1613,183,8)
labels <- c("oberfl.<=5mm","tief>5mm","penetrierend","fehlend")
#-------------------------------------------------------------------------------
#Plot
barplot(height=x,names.arg=labels,main="Balkendiagramm",col="grey",ylim=c(0,2000))
#-------------------------------------------------------------------------------
#Plot
pie(x=x,labels=labels,main="Kreisdiagramm",col=c("grey","white","grey30","black"))
```

8.3.2 absolute Häufigkeiten

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,2,1,1)+0.1)
#Datenbereitstellung
x <- c(rep(10,2),rep(15,5),rep(20,15),rep(25,25),rep(30,45),rep(35,75),
       rep(40,120),rep(45,170),rep(50,180),rep(55,230),rep(60,350),rep(65,450),
       rep(70,455),rep(75,280),rep(80,150),rep(85,60),rep(90,10))
#-------------------------------------------------------------------------------
#Plot
hist(x,main="",xlab="Alter",ylab="absolute Häufigkeit",xlim=c(0,100),
     ylim=c(0,500),col="grey",axes=FALSE)
  #Achsen
    axis(side=1,pos=0,)
    axis(side=2,pos=0,las=2)
```

8.3.3 typische Verteilungstypen

```
#Grafikeinstellungen
par(mfrow=c(2,2),mar=c(2,1,1,1)+0.1)
#Datenbereitstellung
set.seed(123456)
x1 <- seq(from=0,to=10,by=0.01)
y1 <- rnorm(x1)
x2 <- c(3.4,3.8,3.3,3.5,3,3.5,3.8,3.7,3.2)
```

```
x3 <- c(60,80,100,90,75,60,50,44,30,25,14,10,8,5,3,2,1.5,1)
x4 <- c(1,1.5,2,3,5,8,10,14,25,30,44,50,60,75,90,100,80,60)
#-------------------------------------------------------------------------------
#Plot
hist(y1,col="grey",main="",xlab="",ylab="",axes=FALSE)
#Beschriftungen
mtext(text="symmetrisch (normal)",side=1,line=0.5)
#-------------------------------------------------------------------------------
#Plot
barplot(x2,space=F,axes=F,xlim=c(0,9),ylim=c(0,4))
#Beschriftungen
mtext(text="symmetrisch (uniform)",side=1,line=0.5)
#-------------------------------------------------------------------------------
#Plot
barplot(x3,space=F,axes=F)
#Beschriftungen
mtext(text="rechtsschief",side=1,line=0.5)
#-------------------------------------------------------------------------------
#Plot
barplot(x4,space=F,axes=F)
#Beschriftungen
mtext(text="linksschief",side=1,line=0.5)
```

8.3.4 Fehlerbalken

```
#Datenbereitstellung
x <- c(677,1613,183,8)
labels <- c("oberfl.<=5mm","tief>5mm","penetrierend","fehlend")
#-------------------------------------------------------------------------------
#Plot
barplot(height=x,names.arg=labels,main="Balkendiagramm",col="grey",ylim=c(0,2000))
#-------------------------------------------------------------------------------
#Plot
pie(x=x,labels=labels,main="Kreisdiagramm",col=c("grey","white","grey30","black"))
```

8.3.5 Boxplot

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(1,4,4,15)+0.1)
#Datenbereitstellung
  x1<-c(rep(9,2), rep(11,2), rep(13,15),
      rep(15,55), rep(17,95), rep(19,145),
      rep(21,140), rep(23,100), rep(25,40),
      rep(27,15), rep(29,2))
  a <- max(x1);c <- quantile(x=x1,probs=0.75);e <- quantile(x=x1,probs=0.25)
  b <- ((c-e)*0.7)+c;d <- quantile(x=x1,probs=0.5);f <- e-((c-e)*0.7)
  g <- min(x1);h <- mean(x1)
#-------------------------------------------------------------------------------
#Plot
boxplot(x1,ylim=c(9,30),range=0.9, main="",col="grey")
#Datenpunkte/Graph
points(x=1,y=mean(x1),pch=5)
#Flächen/Linien
par(xpd=TRUE)
arrows(x0=1.6,y0=a,x1=1.2,y1=a,angle=20,length=0.125)
```

```
arrows(x0=1.6,y0=b,x1=1.2,y1=b,angle=20,length=0.125)
arrows(x0=1.9,y0=c,x1=1.3,y1=c,angle=20,length=0.125)
arrows(x0=1.9,y0=h,x1=1.3,y1=h,angle=20,length=0.125)
arrows(x0=1.6,y0=d,x1=1.3,y1=d,angle=20,length=0.125)
arrows(x0=1.9,y0=e,x1=1.3,y1=e,angle=20,length=0.125)
arrows(x0=1.6,y0=f,x1=1.2,y1=f,angle=20,length=0.125)
arrows(x0=1.6,y0=g,x1=1.2,y1=g,angle=20,length=0.125)
par(xpd=FALSE)
#Beschriftungen
mtext(text="Ausreißer",side=4,las=2,at=a,line=2)
mtext(text="Maximum",side=4,las=2,at=b,line=2)
mtext(text="75% Quantil",side=4,las=2,at=c,line=6)
mtext(text="Median",side=4,las=2,at=d,line=2)
mtext(text="Mittelwert",side=4,las=2,at=h,line=6)
mtext(text="25% Quantil",side=4,las=2,at=e,line=6)
mtext(text="Minimum",side=4,las=2,at=f,line=2)
mtext(text="Ausreißer",side=4,las=2,at=g,line=2)
```

8.3.6 Darstellung Boxplot und Histogramm

```
#Grafikeinstellungen
par(mfrow=c(2,2),mar=c(2,1,2,1)+0.1)
#Datenbereitstellung
x1<-c(rep(9,2), rep(11,2), rep(13,15),rep(15,55), rep(17,95), rep(19,145),
     rep(21,140), rep(23,100), rep(25,40),rep(27,15), rep(29,2))
x2<-c(rep(13,2), rep(15,4), rep(16,24),rep(17,42), rep(18,76), rep(19,80),
      rep(20,62), rep(21,40), rep(22,52),rep(23,60), rep(24,88), rep(25,42),
      rep(26,18), rep(27,8))
#-------------------------------------------------------------------------------
#Plot
  boxplot(x1,ylim=c(9,30),range=0.9, main="Boxplot",col="grey")
#-------------------------------------------------------------------------------
#Plot
  hist(x1,breaks=11,xlab="",main="Histogramm",xlim=c(8,30),axes=F,ylab="",col="grey")
  #Achsen
  axis(side=1,labels=TRUE,pos=0)
  axis(side=2,labels=TRUE,pos=8)
    #Pfeilspitzen
    par(xpd=TRUE)
      arrows(x0=8,y0=150,x1=8,y1=156,angle=20,length=0.125)
      arrows(x0=30,y0=0,x1=31,y1=0,angle=20,length=0.125)
    par(xpd=FALSE)
#-------------------------------------------------------------------------------
#Plot
  boxplot(x2,ylim=c(9,30), main="Boxplot",col="grey")
#-------------------------------------------------------------------------------
#Plot
  hist(x2,breaks=11,xlab="",main="Histogramm",xlim=c(13,30),axes=F,ylab="",col="grey")
  #Achsen
  axis(side=1,labels=TRUE,pos=0)
  axis(side=2,labels=TRUE,pos=13)
    #Pfeilspitzen
    par(xpd=TRUE)
      arrows(x0=13,y0=80,x1=13,y1=85,angle=20,length=0.125)
      arrows(x0=30,y0=0,x1=31,y1=0,angle=20,length=0.125)
    par(xpd=FALSE)
```

8.4 Kapitel 3

8.4.1 Verteilung der Einzelwerte

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
x1 <- c(rep(1,10),rep(2,10),rep(3,10),rep(4,10),rep(5,10),rep(6,10),rep(7,10),
        rep(8,10),rep(9,10),rep(10,10))
y1 <- rnorm(10*10, mean=175, sd=10)
m <- matrix(y1,ncol=10,byrow=FALSE)
#-------------------------------------------------------------------------------
#Plot
plot(x1,m,ylab="Körpergröße",xlab="Stichprobe",ylim=c(150,200))
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
y1 <- rnorm(10*10, mean=175, sd=10)
#-------------------------------------------------------------------------------
#Plot
boxplot(y1,col="grey",xlab="Verteilung der Einzelwerte",ylim=c(150,200))
```

8.4.2 Verteilung der Mittelwerte

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
x2 <- c(1:10)
y1 <- rnorm(10*10, mean=175, sd=10)
m <- matrix(y1, ncol=10,byrow=FALSE)
y2 <- means <- colMeans(m)
#-------------------------------------------------------------------------------
#Plot
plot(x2,y2,ylab="Körpergröße",xlab="Stichprobe",ylim=c(160,200))
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
y1 <- rnorm(10*10, mean=175, sd=10)
m <- matrix(y1, ncol=10,byrow=FALSE)
y2 <- means <- colMeans(m)
#-------------------------------------------------------------------------------
#Plot
boxplot(y2,col="grey",xlab="Verteilung der Einzelwerte",ylim=c(150,200))
```

8.4.3 Beispiel Für Intervallschätzung

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,3,3)+0.1)
#Datenbereitstellung
set.seed(6789)
m <- matrix(rnorm(20*25, mean=55.1, sd=11.7), ncol=20)
means <- colMeans(m)
sds <- apply(m, 2, sd)
a <- rep(1,18)
b <- c(1,3,a)
#-------------------------------------------------------------------------------
#Plot
plotCI(x=means,uiw=qnorm(.975)*sds/5,ylab="",xlab="",lty=b)
#Flächen/Linien
abline(h=55.1,lty=2)
#Beschriftung
mtext(text="Stichprobe",side=1,line=2,font=2)
mtext(text="95%KI HDL[mg/dl] bei Frauen",side=2,line=2,font=2)
```

8.5 Kapitel 4

8.5.1 Ablehnungsbereich

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
cord.x <- seq(from = 2, to = 4,by=0.01)
cord.y <- dnorm(cord.x)
#-------------------------------------------------------------------------------
#Plot
  plot(x,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
  axis(side=2,labels=FALSE,tck=0, pos=-4)
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=-4,y0=0.4,x1=-4,y1=0.45,angle=20,length=0.125)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  lines(x=c(2,2),y=c(0,dnorm(2)),lwd=2)
  lines(x=c(1.5,1.5),y=c(0,-0.2),lwd=2)
  lines(x=c(3,3),y=c(0,-0.2),lwd=2)
  arrows(x0=3,y0=0.28,x1=2.4,y1=0.01,angle=20,length=0.125,lwd=1.2)
  polygon(x = c(2, cord.x, 4),y = c(0, cord.y, 0),col = "black",density=10)
#Beschriftungen
  mtext(text="f(x)",side=3,at=-4.2,line=0,cex=1.2)
  mtext(text="x",side=1,at=4.3,line=0,cex=1.2)
  mtext(text="b",side=1,at=3,line=0,las=1,cex=1.2)
  mtext(text="a",side=1,at=1.5,line=0,las=1,cex=1.2)
  text(x=3,y=0.32, "Ablehnungs-
       bereich",cex=1.2)
```

8.5.2 ein- und zweisseitige Fragestellung

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(1,1,1,1)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
cord.x1 <- seq(from = -4, to = -2,by=0.01)
cord.y1 <- dnorm(cord.x1)
#-------------------------------------------------------------------------------
#Plot
  plot(x,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
  #Pfeilspitzen
  par(xpd=TRUE)
   arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
    arrows(x0=-3,y0=0.2,x1=-2.4,y1=0.01,angle=20,length=0.125)
    polygon(x = c(-4, cord.x1, -2),y = c(0, cord.y1, 0),col = "black",density=10)
#Beschriftungen
  mtext(text="x",side=1,at=4.3,line=0,cex=1.4)
  mtext(text="-1.645",side=1,at=-2,line=0,las=1,cex=1.4)
  text(x=-3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=0,y=0.05,"Annahmebereich",cex=1.4)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(1,1,1,1)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
cord.x2 <- seq(from = 2, to = 4,by=0.01)
cord.y2 <- dnorm(cord.x2)
#-------------------------------------------------------------------------------
#Plot
plot(x,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  arrows(x0=3,y0=0.2,x1=2.4,y1=0.01,angle=20,length=0.125)
  polygon(x = c(2, cord.x2, 4),y = c(0, cord.y2, 0),col = "black",density=10)
#Beschriftungen
  mtext(text="x",side=1,at=4.3,line=0,cex=1.4)
  mtext(text="1.645",side=1,at=2,line=0,las=1,cex=1.4)
  text(x=3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=0,y=0.05,"Annahmebereich",cex=1.4)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(1,1,1,1)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
```

```
cord.x1 <- seq(from = -4, to = -2,by=0.01)
cord.x2 <- seq(from = 2, to = 4,by=0.01)
cord.y1 <- dnorm(cord.x1)
cord.y2 <- dnorm(cord.x2)
#------------------------------------------------------------------------------
#Plot
plot(x,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
#Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  arrows(x0=3,y0=0.2,x1=2.4,y1=0.01,angle=20,length=0.125)
  polygon(x = c(-4, cord.x1, -2),y = c(0, cord.y1, 0),col = "black",density=10)
  arrows(x0=-3,y0=0.2,x1=-2.4,y1=0.01,angle=20,length=0.125)
  polygon(x = c(2, cord.x2, 4),y = c(0, cord.y2, 0),col = "black",density=10)
#Beschriftungen
  mtext(text="x",side=1,at=4.3,line=0,cex=1.4)
  mtext(text="-1.96",side=1,at=-2,line=0,las=1,cex=1.4)
  mtext(text="1.96",side=1,at=2,line=0,las=1,cex=1.4)
  text(x=-3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=0,y=0.2,"Annahmebereich",cex=1.4)
```

8.5.3 Ablehnungsbereich a=0.05

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(1,1,1,1)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
cord.x1 <- seq(from = -4, to = -2,by=0.01)
cord.x2 <- seq(from = 2, to = 4,by=0.01)
cord.y1 <- dnorm(cord.x1)
cord.y2 <- dnorm(cord.x2)
#------------------------------------------------------------------------------
#Plot
plot(x,fx, type="l",main="",ylab="",xlab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
  #Pfeilspitzen
  par(xpd=TRUE)
    arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  arrows(x0=3,y0=0.2,x1=2.4,y1=0.01,angle=20,length=0.125)
  polygon(x = c(-4, cord.x1, -2),y = c(0, cord.y1, 0),col = "black",density=10)
  arrows(x0=-3,y0=0.2,x1=-2.4,y1=0.01,angle=20,length=0.125)
  polygon(x = c(2, cord.x2, 4),y = c(0, cord.y2, 0),col = "black",density=10)
#Beschriftungen
  mtext(text="x",side=1,at=4.3,line=0,cex=1.4)
  mtext(text="-1.96",side=1,at=-2,line=0,las=1,cex=1.4)
  mtext(text="1.96",side=1,at=2,line=0,las=1,cex=1.4)
```

```
  text(x=-3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=3,y=0.25, "Ablehnungs-
     bereich",cex=1.4)
  text(x=0,y=0.05,"Annahmebereich",cex=1.4)
```

8.5.4 signifikantes Testergebnis

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(5,2,2,2)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
#li,grau
cord.x1 <- seq(from=-4,to=-2.25,by=0.01);cord.y1 <- dnorm(cord.x1)
#li,lang
cord.x11 <- seq(from=-4,to=-1.75,by=0.01);cord.y11 <- dnorm(cord.x11)
#re,grau
cord.x2 <- seq(from=2.25,to=4,by=0.01);cord.y2 <- dnorm(cord.x2)
#re,lang
cord.x21 <- seq(from=1.75,to=4,by=0.01);cord.y21 <- dnorm(cord.x21)
#------------------------------------------------------------------------------
#Plot
plot(x,fx, type="l",main="signifikantes Testergebnis (zweiseitig)",ylab="",xlab="",
     xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Flächen/Linien
#grau
polygon(x=c(1.75,cord.x21, 4),y=c(0,cord.y21,0),col="grey")
polygon(x=c(-4,cord.x11,-1.75),y=c(0,cord.y11,0),col="grey")
#schraffiert
polygon(x=c(-4,cord.x1,-2.25),y=c(0,cord.y1,0),col="black",density=10)
polygon(x=c(2.25,cord.x2,4),y=c(0,cord.y2,0),col="black",density=10)
#border
lines(x=c(-2.25,-2.25),y=c(-0.006,dnorm(-2.25)),lwd=2)
lines(x=c(-1.75,-1.75),y=c(-0.006,dnorm(-1.75)),lwd=2)
lines(x=c(2.25,2.25),y=c(-0.006,dnorm(2.25)),lwd=2)
lines(x=c(1.75,1.75),y=c(-0.006,dnorm(1.75)),lwd=2)
#Beschriftungen
mtext(text="T",side=1,at=4.3,line=0,cex=1.4)
mtext(text="0",side=1,at=0,line=0,cex=1.4)
mtext(text=expression(-t),side=1,at=-2.35,line=0,las=1,cex=1.4)
mtext(text=expression(-K),side=1,at=-1.85,line=0,las=1,cex=1.4)
mtext(text=expression(+t),side=1,at=2.15,line=0,las=1,cex=1.4)
mtext(text=expression(+K),side=1,at=1.65,line=0,las=1,cex=1.4)
mtext(text="K: kritischer Wert",side=1,line=2,cex=1.4)
mtext(text="t: beobachteter Wert",side=1,line=3.5,cex=1.4)
legend("topright",legend=c(expression(alpha),"p-Wert"),fill=c("grey","black"),
       density=c(NA,25,75),inset = .02,cex=1.4)
```

8.5.5 nicht signifikantes Testergebnis

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(5,2,2,2)+0.1)
#Datenbereitstellung
x <- seq(from=-4, to=4, by=0.01)
fx <- dnorm(x,mean=0,sd=1)
#li,grau
cord.x1 <- seq(from=-4,to=-1.25,by=0.01);cord.y1 <- dnorm(cord.x1)
#li,lang
cord.x11 <- seq(from=-4,to=-1.75,by=0.01);cord.y11 <- dnorm(cord.x11)
#re,grau
cord.x2 <- seq(from=1.25,to=4,by=0.01);cord.y2 <- dnorm(cord.x2)
#re,lang
cord.x21 <- seq(from=1.75,to=4,by=0.01);cord.y21 <- dnorm(cord.x21)
#------------------------------------------------------------------------------
#Plot
plot(x,fx, type="l",main="nicht signifikantes Testergebnis (zweiseitig)",
     ylab="", xlab="",
     xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
axis(side=1,labels=FALSE,tck=-0.015,pos=0 )
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=4,y0=0,x1=4.3,y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Flächen/Linien
#grau
polygon(x=c(1.75,cord.x21, 4),y=c(0,cord.y21,0),col="grey")
polygon(x=c(-4,cord.x11,-1.75),y=c(0,cord.y11,0),col="grey")
#schraffiert
polygon(x=c(-4,cord.x1,-1.25),y=c(0,cord.y1,0),col="black",density=10)
polygon(x=c(1.25,cord.x2,4),y=c(0,cord.y2,0),col="black",density=10)
#border
lines(x=c(-1.25,-1.25),y=c(-0.006,dnorm(-1.25)),lwd=2)
lines(x=c(-1.75,-1.75),y=c(-0.006,dnorm(-1.75)),lwd=2)
lines(x=c(1.25,1.25),y=c(-0.006,dnorm(1.25)),lwd=2)
lines(x=c(1.75,1.75),y=c(-0.006,dnorm(1.75)),lwd=2)
#Beschriftungen
mtext(text="T",side=1,at=4.3,line=0,cex=1.4)
mtext(text="0",side=1,at=0,line=0,cex=1.4)
mtext(text=expression(-t),side=1,at=-1.35,line=0,las=1,cex=1.4)
mtext(text=expression(-K),side=1,at=-1.85,line=0,las=1,cex=1.4)
mtext(text=expression(+t),side=1,at=1.15,line=0,las=1,cex=1.4)
mtext(text=expression(+K),side=1,at=1.65,line=0,las=1,cex=1.4)
mtext(text="K: kritischer Wert",side=1,line=2,cex=1.4)
mtext(text="t: beobachteter Wert",side=1,line=3.5,cex=1.4)
legend("topright",legend=c(expression(alpha),"p-Wert"),fill=c("grey","black"),
       density=c(NA,25,75),inset = .02,cex=1.4)
```

8.5.6 Verteilung der Prüfgröße

```
#Grafikeinstellungen
par(mfrow=c(2,1),mar=c(2,2,2,2)+0.1)
#Datenbereitstellung
  #Parameter
  meanx1 <- 0;meanx2 <- 2;sdx1 <- 1;sdx2 <- 1;krit <- 1;krit2 <- 1.5
```

```
  #Funktion
  x1 <- seq(from=meanx1-4, to=meanx1+4, by=0.01)
  fx1 <- dnorm(x1,mean=meanx1,sd=sdx1)
  x2 <- seq(from=meanx2-4, to=meanx2+4, by=0.01)
  fx2 <- dnorm(x2,mean=meanx2,sd=sdx2)
  #alpha
  cord.x1 <- seq(from=krit,to=krit+4,by=0.01)
  cord.y1 <- dnorm(cord.x1,mean=meanx1,sd=sdx1)
  cord.x3 <- seq(from=krit2,to=krit2+4,by=0.01)
  cord.y3 <- dnorm(cord.x3,mean=meanx1,sd=sdx2)
  #beta
  cord.x2 <- seq(from=krit-4,to=krit,by=0.01)
  cord.y2 <- dnorm(cord.x2,mean=meanx2,sd=sdx1)
  cord.x4 <- seq(from=krit2-4,to=krit2,by=0.01)
  cord.y4 <- dnorm(cord.x4,mean=meanx2,sd=sdx2)
#-------------------------------------------------------------------------------
#Plot
  plot(x1,fx1, type="n",main="",ylab="",xlab="",ylim=c(0,0.5),
       xlim=c(meanx1-4,meanx2+4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=0,pos=0 )
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=meanx2+3.5,y0=0,x1=meanx2+4,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  #alpha
  polygon(x=c(krit,cord.x1,krit+4),y=c(0,cord.y1,0),col="grey")
  #beta
  polygon(x=c(krit-4,cord.x2,krit),y=c(0,cord.y2,0),col="black",density=10)
  par(xpd=TRUE)
    lines(x=c(meanx1,meanx1),y=c(-4,dnorm(meanx1,mean=meanx1,sd=sdx1)),lty=2)
    lines(x=c(meanx2,meanx2),y=c(-4,dnorm(meanx1,mean=meanx1,sd=sdx2)),lty=2)
  par(xpd=FALSE)
#Datenpunkte/Graph
  lines(x1,fx1,lwd=2)
  lines(x2,fx2,lwd=2)
#Beschriftung
  mtext(text="Prüfgröße",side=1,at=meanx2+4,line=0)
  legend("topright",legend=c(expression(alpha),expression(beta)),
         fill=c("grey","black"),
         density=c(NA,25,75),cex=1.2,bty='o')
  text(x=meanx1,y=0.45,expression(H[0]),cex=1.2)
  text(x=meanx2,y=0.45,expression(H[a/speziell]),cex=1.2)
  par(xpd=TRUE)
  arrows(x0=krit,y0=-0.03,x1=krit,y1=0,angle=20,length=0.125)
  text(x=krit,y=-0.05,"krit. Wert")
  par(xpd=FALSE)
#-------------------------------------------------------------------------------
#Plot
  plot(x1,fx1, type="n",main="",ylab="",xlab="",ylim=c(0,0.5),
     xlim=c(meanx1-4,meanx2+4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,labels=FALSE,tck=0,pos=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=meanx2+3.5,y0=0,x1=meanx2+4,y1=0,angle=20,length=0.125) #x-Achse
  par(xpd=FALSE)
```

```
#Flächen/Linien
  #alpha
  polygon(x=c(krit2,cord.x3,krit2+4),y=c(0,cord.y3,0),col="grey")
  #beta
  polygon(x=c(krit2-4,cord.x4,krit2),y=c(0,cord.y4,0),col="black",density=10)
  par(xpd=TRUE)
  lines(x=c(meanx1,meanx1),y=c(0,4),lty=2)
  lines(x=c(meanx2,meanx2),y=c(0,4),lty=2)
  par(xpd=FALSE)
#Datenpunkte/Graph
  lines(x1,fx1,lwd=2)
  lines(x2,fx2,lwd=2)
#Beschriftung
  mtext(text="Prüfgröße",side=1,at=meanx2+4,line=0)
  par(xpd=TRUE)
  arrows(x0=krit2,y0=-0.03,x1=krit2,y1=0,angle=20,length=0.125)
  text(x=krit2,y=-0.05,"krit. Wert")
  par(xpd=FALSE)
```

8.6 Kapitel 5

8.6.1 Vergleich von unterschiedlichen Normalverteilungen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,1,2,1)+0.1)
#Datenbereitstellung
x1 <- seq(from=60,to=220,by=0.01)
y1 <- dnorm(x1,mean=120,sd=15)
y2 <- dnorm(x1,mean=150,sd=15)
lab <- c(60,80,100,120,140,160,180,200)
#------------------------------------------------------------------------------
#Plot
  plot(x1,y1,type="l",main="Unterschied des Mittelwertes",xlab="",ylab="",
       axes=FALSE,lwd=2)
#Datenpunkte/Graph
  lines(x1,y2,lty=2,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=lab,at=lab)
  axis(side=2,pos=60,labels=FALSE,tck=0,)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=60,y0=0.020,x1=60,y1=0.03,angle=20,length=0.125) #y-Achse
  arrows(x0=200,y0=0,x1=220,y1=0,angle=20,length=0.125) #x-Achse
  par(xpd=FALSE)
#Beschriftungen
  legend("topright",c("N(120,15)","N(150,15)"),lty=c(1,2),lwd=2,cex=1.4)
  mtext(text="mmHG",side=1,line=1,at=220,cex=1.4)
```

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,1,2,1)+0.1)
#Datenbereitstellung
x1 <- seq(from=60,to=220,by=0.01)
y1 <- dnorm(x1,mean=120,sd=15)
y2 <- dnorm(x1,mean=150,sd=20)
lab <- c(60,80,100,120,140,160,180,200)
```

```
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="l",main="Unterschied des Mittelwertes
     und der Standardabweichung",xlab="",ylab="",axes=FALSE,lwd=2)
#Datenpunkte/Graph
  lines(x1,y2,lty=2,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=lab,at=lab)
  axis(side=2,pos=60,labels=FALSE,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=60,y0=0.020,x1=60,y1=0.03,angle=20,length=0.125)
  arrows(x0=200,y0=0,x1=220,y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  legend("topright",c("N(120,15)","N(150,20)"),lty=c(1,2),cex=1.4,lwd=2)
  mtext(text="mmHG",side=1,line=1,at=220,cex=1.4)
```

8.6.2 Vergleich von t-Verteilung mit verschiedenen Freiheitsgraden

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,2,2,2)+0.1)
#Datenbereitstellung
x1 <- seq(from=-10,to=10,by=0.01)
y1 <- dnorm(x1,mean=0,sd=1)
y2 <- dt(x1,df=8)
y3 <- dt(x1,df=2)
y4 <- dt(x1,df=1)
#-----------------------------------------------
#Plot
  plot(x1,y1,type="l",xlab="",ylab="",xlim=c(-4,4),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=TRUE)
  axis(side=2,pos=0,labels=FALSE,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0.4,x1=0,y1=0.45,angle=20,length=0.125) #y-Achse
  arrows(x0=4,y0=0,x1=4.5,y1=0,angle=20,length=0.125) #x-Achse
  par(xpd=FALSE)
#Flächen/Linien
  lines(x1,y2,lty=2,lwd=2)
  lines(x1,y3,lty=3,lwd=2)
  lines(x1,y4,lty=4,lwd=2)
#Beschriftungen
  legend("topright",c("N(0,1)","t(8)","t(2)","t(1)"),lty = c(1,2,3,4),lwd=2)
```

8.6.3 Voraussetzungen

```
#Grafikeinstellungen
par(mfrow=c(3,2),mar=c(1,1,1,1)+0.1)
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x <- seq(-4,4,length=200)
  ydf <- df(x,df1=5,df2=10)
  ydf2 <- df(x,df1=8,df2=8)
```

```
#Plot
  plot(x,ydf,type="l",xlim=c(0,4),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x,ydf2,lty=2)
#Beschriftungen
  legend("top",legend=expression(H[0]),cex=2,bty="n")
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x1 <- seq(-4,4,length=200)
  x2 <- seq(-4,6,length=200)
  ydf <- df(x,df1=5,df2=10)
#Plot
  plot(x1,ydf,type="l",xlim=c(0,4),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x2,ydf,lty=2)
#Beschriftung
  legend("right",legend="a",cex=2,bty="n")
  legend("top",legend=expression(H[A]),cex=2,bty="n")
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x <- seq(-4,4,length=200)
  ydf <- df(x,df1=5,df2=4)
  ydnorm <- dnorm(x,mean=0.7,sd=0.4)
#Plot
  plot(x,ydf,type="l",xlim=c(0,4),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x,ydnorm,lty=2)
#Beschriftung
legend("top",legend=expression(H[0]),cex=2,bty="n")
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x <- seq(-4,4,length=200)
  ydf <- df(x,df1=5,df2=4)
  ydnorm <- dnorm(x,mean=3,sd=0.4)
#Plot
  plot(x,ydf,type="l",xlim=c(0,4),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x,ydnorm,lty=2)
#Beschriftungen
  legend("right",legend="b",cex=2,bty="n")
  legend("top",legend=expression(H[A]),cex=2,bty="n")
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x1 <- seq(-4,4,length=200)
  x2 <- seq(-3,6,length=200)
  ydf <- df(x,df1=5,df2=4)
  ydnorm <- dnorm(x,mean=3,sd=0.4)
#Plot
  plot(x1,ydf,type="l",xlim=c(-2,3),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x2,rev(ydf),lty=2)
#Beschriftung
  legend("top",legend=expression(H[0]),cex=2,bty="n")
#-------------------------------------------------------------------------------
#Datenbereitstellung
  x1 <- seq(-4,4,length=200)
  x2 <- seq(0,8,length=200)
  ydf <- df(x,df1=5,df2=4)
```

```
  ydnorm <- dnorm(x,mean=3,sd=0.4)
#Plot
  plot(x1,ydf,type="l",xlim=c(-1,5),ylim=c(0,1),xlab="",ylab="",xaxt="n",yaxt="n")
#Datenpunkte/Graph
  lines(x2,rev(ydf),lty=2)
#Beschriftungen
  legend("right",legend="c",cex=2,bty="n")
  legend("top",legend=expression(H[A]),cex=2,bty="n")
```

8.6.4 QQ-Plot

```
#Grafikeinstellungen
par(mfrow=c(1,3),mar=c(4,4,4,4)+0.1)
#Datenbereitstellung
  set.seed(1234567)
  x1 <- rnorm(50);x2 <- runif(50);x3 <- rlnorm(50)
  x <- data.frame(rnorm(50),runif(50),rlnorm(50))
  ks1 <- ks.test(x1,"pnorm");sw1 <- shapiro.test(x1)
  ks2 <- ks.test(x2,"pnorm");sw2 <- shapiro.test(x2)
  ks3 <- ks.test(x3,"pnorm");sw3 <- shapiro.test(x3)
  ks <- c(signif(ks1$p.value,2),signif(ks2$p.value,2),signif(ks3$p.value,2))
  sw <- c(signif(sw1$p.value,2),signif(sw2$p.value,2),signif(sw3$p.value,2))
  main <- c("Normalverteilung",rep("keine Normalverteilung",2))
#-------------------------------------------------------------------------------
#Plot
for (i in 1:3){
  qqnorm(x[[i]], main=main[[i]], xlab="theorethische Quantile",
         ylab="Quantile der Stichprobe",lwd=2,cex=1,cex.lab=1.5,
         cex.main=1.8,cex.axis=1.5)
#Datenpunkte/Graph
  qqline(x[[i]],lwd=2)
#Beschriftungen
  legend("topleft",legend=c(paste("KS-Test: p =",ks[[i]]),
         paste("Shapiro-Wilk-Test: p =",sw[[i]])),bty="n")
}
```

8.7 Kapitel 6

8.7.1 Verteilungformen

```
#Grafikeinstellungen
par(mfrow=c(2,3),mar=c(1.5,1.5,1.5,1.5)+0.1)
#Datenbereitstellung
data1 <- data.frame(x1=c(1,2,2.5,4,5,6,7,7.5,9,10,11,13,4,14),
                    y1=c(3,1,8,14,6,7,2,1,12,5,11,12.5,13,10))
data2 <- data.frame(x2=c(1,1.8,2,3,3.2,3.8,4,5,5.2,6,5.9,7,7.1,8),
                    y2=c(1,1,2,1.8,2.3,3,2.5,2.8,3.5,3.2,4,3.8,5,4.5))
data3 <- data.frame(x3=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14),
                    y3=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14))
data4 <- data.frame(x4=c(1,2,3,4,4.2,4.4,4.8,5,5.5,5.6,7,7.8,8.5,9.5,10),
                    y4=c(1,0.5,1.5,0.8,2,2.5,4,6,7,7.5,8,8.5,8,9,9))
data5 <- data.frame(x5=c(1,2,2.3,2.5,3.5,4,4.5,6,6.5,7,7.5,9,9.5,11,11.5),
                    y5=c(9,11,8,6,6,4.5,3,3.3,2.5,2.8,2,2.2,1.5,1,0.8))
data6 <- data.frame(x6=c(1,3,2,5,4,6,7,8,10,9,11,13,12,14,16,15,18,17,20,19),
```

```
                    y6=c(10,8,6,5,4,3,2,1,0.7,0.5,0.3,0.4,0.8,1,2,3,4,6,7,9))
a <- list(data1$x1,data2$x2,data3$x3,data4$x4,data5$x5,data6$x6)
b <- list(data1$y1,data2$y2,data3$y3,data4$y4,data5$y5,data6$y6)
#-------------------------------------------------------------------------------
#Plot
for (i in 1:6){
  plot(x=a[[i]],y=b[[i]],type="p",pch=16,xlab="",ylab="",xlim=c(0,max(a[[i]])),
       ylim=c(0,max(b[[i]])),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=FALSE,tck=0)
  axis(side=2,pos=0,labels=FALSE,tck=0)
  #Pfeilspitzen
    par(xpd=TRUE)
    arrows(x0=0,y0=0,x1=0,y1=(max(b[[i]])),angle=20,length=0.125)
    arrows(x0=0,y0=0,x1=(max(a[[i]])),y1=0,angle=20,length=0.125)
    par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(a[[i]]))
  mtext("y",side=2,line=0,at=max(b[[i]]),las=2)
}
```

8.7.2 Scatterplot

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,1,2,1)+0.1)
#Datenbereitstellung
Alter <- c(41,42,42,43,45,47,48,49,49,50,51,52,54,54,55,56,57,57,58,59)
sysBL <- c(130,125,140,135,140,140,155,150,160,155,165,160,170,165,170,160,175,
           165,170,175)
#-------------------------------------------------------------------------------
#Plot
plot(Alter,sysBL,xlab="Alter [Jahre]",ylab="systolischer Blutdruck [mmHg]",
     xlim=c(min(Alter),max(Alter)),ylim=c(min(sysBL),max(sysBL)))
```

8.7.3 Punktwolke

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
N <- 101
x1 <- seq(from=3,to=31, length.out=N)
y1 <- 6 + 0.2 * x1 + rnorm(n=N, mean=0, sd=3)
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="p",pch=16,main="",xlab="",ylab="",
     xlim=c(0,max(x1)),ylim=c(0,max(y1)),axes=FALSE,lwd=2)
#Achsen
axis(side=1,pos=0,labels=F,tck=0)
axis(side=2,pos=0,labels=F,tck=0)
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=0,y0=0,x1=0,y1=max(y1)+(1/10*max(y1)),angle=20,length=0.125)
arrows(x0=0,y0=0,x1=max(x1)+(1/10*max(x1)),y1=0,angle=20,length=0.125)
par(xpd=FALSE)
```

```
#Flächen/Linien
abline(lm(y1~x1))
#Beschriftungen
mtext("x",side=1,line=0,at=max(x1)+(1/10*max(x1)))
mtext("y",side=2,line=0,at=max(y1)+(1/10*max(y1)),las=2)
```

8.7.4 positive Korrelation

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,4,4)+0.1)
#Datenbereitstellung
x1 <- c(1,1.8,2,3,3.2,3.8,4,5,5.2,6,5.9,7,7.1,8)
y1 <- c(1,1,2,1.8,2.3,3,2.5,2.8,3.5,3.2,4,3.8,5,4.5)
x2 <- x1
y2 <- rev(y1)
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="p",pch=16,main="r > 0",xlab="",ylab="",
     xlim=c(0,max(x1)),ylim=c(0,max(y1)),axes=FALSE,lwd=2)
#Achsen
axis(side=1,pos=0,labels=F,tck=0)
axis(side=2,pos=0,labels=F,tck=0)
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=0,y0=0,x1=0,y1=max(y1)+(1/10*max(y1)),angle=20,length=0.125)
arrows(x0=0,y0=0,x1=max(x1)+(1/10*max(x1)),y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Beschriftungen
mtext("x",side=1,line=0,at=max(x1)+(1/10*max(x1)))
mtext("y",side=2,line=0,at=max(y1)+(1/10*max(y1)),las=2)
```

8.7.5 negative Korrelation

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,4,4)+0.1)
#Datenbereitstellung
x1 <- c(1,1.8,2,3,3.2,3.8,4,5,5.2,6,5.9,7,7.1,8)
y1 <- c(1,1,2,1.8,2.3,3,2.5,2.8,3.5,3.2,4,3.8,5,4.5)
x2 <- x1
y2 <- rev(y1)
#-------------------------------------------------------------------------------
#Plot
plot(x2,y2,type="p",pch=16,main="r < 0",xlab="",ylab="",
     xlim=c(0,max(x2)),ylim=c(0,max(y2)),axes=FALSE,lwd=2)
#Achsen
axis(side=1,pos=0,labels=F,tck=0)
axis(side=2,pos=0,labels=F,tck=0)
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=0,y0=0,x1=0,y1=max(y2)+(1/10*max(y2)),angle=20,length=0.125)
arrows(x0=0,y0=0,x1=max(x2)+(1/10*max(x2)),y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Beschriftungen
mtext("x",side=1,line=0,at=max(x2)+(1/10*max(x2)))
mtext("y",side=2,line=0,at=max(y2)+(1/10*max(y2)),las=2)
```

8.7.6 Verteilung in Abhangigkeit von r

```
#Grafikeinstellungen
par(mfrow=c(2,2),mar=c(1.5,1.5,1.5,1.5)+0.1)
#Datenbereitstellung
data1 <- data.frame(x1=c(1,1.2,1.1,1.3,2,2.4,2.2,2.2,2.4,3,3.1,3,3.4,4.5,4,4.1,
                         4.5,5,5),
                    y1=c(1,3,5,7,0.5,2,4,6,7.2,1,3,5,7,0.5,2,4,6,7.2,3))
data2 <- data.frame(x2=c(1,1.1,1.4,1.5,1.9,2.1,2.2,2.4,2.8,3),
                    y2=c(1,2,3,1.5,2.8,3.8,2,3.1,2.8,3.7))
data3 <- data.frame(x3=c(1,1.8,2,3,3.2,3.8,4,5,5.2,6,5.9,7,7.1,8),
                    y3=c(4.5,5,3.8,4,3.2,3.5,2.8,2.5,3,2.3,1.8,2,1,1))
data4 <- data.frame(x4=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15),
                    y4=c(2,3,4,5,6,7,8,9,10,11,12,13,14,15,16))
a <- list(data1$x1,data2$x2,data3$x3,data4$x4)
b <- list(data1$y1,data2$y2,data3$y3,data4$y4)
c <- list(expression(r%~~%0),expression(r%~~%+0.5),
          expression(r%~~%-0.8),expression(r%~~%+1))
#-------------------------------------------------------------------------------
for (i in 1:4){
  plot(x=a[[i]],y=b[[i]],type="p",pch=16,xlab="",ylab="",xlim=c(0,max(a[[i]])),
       ylim=c(0,max(b[[i]])),axes=FALSE,lwd=2)
  #Achsen
  axis(side=1,pos=0,labels=FALSE,tck=0)
  axis(side=2,pos=0,labels=FALSE,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=(max(b[[i]])),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=(max(a[[i]])),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
  #Beschriftungen
  mtext("x",side=1,line=0,at=max(a[[i]]))
  mtext("y",side=2,line=0,at=max(b[[i]]),las=2)
  mtext(c[[i]],side=3,line=0)
}
```

8.7.7 Veranschaulichung von SP(X,Y)

```
#Grafikeinstellungen
par(mfrow = c(1,1),mar=c(2,2,2,2)+0.1)
#Datenbereitstellung
x <- c(52, 41, 42, 42, 43, 45, 47, 48, 44, 49, 48,
       51, 52, 54, 54, 55, 56, 57, 57, 58, 58)
y <- c(142,130,125,140,135,140,140,170,150,160,
       148,165,160,170,165,170,160,175,165,170,175)
mx <- mean(x)
my <- mean(y)
#-------------------------------------------------------------------------------
#Plot
  plot(x,y,xlab="",ylab="",xlim=c(min(x),max(x)),ylim=c(min(y),max(y)),axes=F)
#Achsen
  axis(side=1,pos=min(y)-(1/50*min(y)),labels=T,tck=0)
  axis(side=2,pos=min(x)-(1/50*min(x)),labels=T,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=min(x)-(1/50*min(x)),y0=min(y)-(1/50*min(y)),x1=min(x)-(1/50*min(x)),
         y1=max(y)+(1/50*max(y)),angle=20,length=0.125)
```

```
  arrows(x0=min(x)-(1/50*min(x)),y0=min(y)-(1/50*min(y)),x1=max(x)+(1/50*max(x)),
         y1=min(y)-(1/50*min(y)),angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  abline(h=my,lty=2)
  abline(v=mx,lty=2)
  rect(xleft=min(x),ybottom=my,xright=mx,ytop=max(y),density=2,angle=-45,border=F)
  rect(xleft=mx,ybottom=min(y),xright=max(x),ytop=my,density=2,angle=-45,border=F)
#Quadrant 1
q1 <- which(x>mx&y>my,arr.ind=F)
  q1x <- x[min(q1)]
  q1y <- y[min(q1)]
  lines(x=c(mx,q1x),y=c(q1y,q1y))
  lines(x=c(q1x,q1x),y=c(my,q1y))
  #text
    #drüber
    text(x=(mx+q1x)/2,y=q1y+((q1y-my)/3),"+")
    #daneben re
    text(x=q1x+((q1x-mx)/3),y=(my+q1y)/2,"+")
#Quadrant 2
q2 <- which(x<mx&y>my,arr.ind=F)
  q2x <- x[max(q2)]
  q2y <- y[max(q2)]
  lines(x=c(mx,q2x),y=c(q2y,q2y))
  lines(x=c(q2x,q2x),y=c(my,q2y))
  #text
    #drüber
    text(x=(mx+q2x)/2,y=q2y+((q2y-my)/3),"-")
    #daneben li
    text(x=q2x+((q2x-mx)/3),y=(my+q2y)/2,"+")
#Quadrant 3
q3 <- which(x<mx&y<my,arr.ind=F)
  q3x <- x[max(q3)]
  q3y <- y[max(q3)]
  lines(x=c(mx,q3x),y=c(q3y,q3y))
  lines(x=c(q3x,q3x),y=c(my,q3y))
  #text
    #drüber
    text(x=(mx+q3x)/2,y=q3y+((q3y-my)/3),"-")
    #daneben li
    text(x=q3x+((q2x-mx)/3),y=(my+q3y)/2,"-")
#Quadrant 4
q4 <- which(x>mx&y<my,arr.ind=F)
  q4x <- x[min(q4)]
  q4y <- y[min(q4)]
  lines(x=c(mx,q4x),y=c(q4y,q4y))
  lines(x=c(q4x,q4x),y=c(my,q4y))
  #text
    #drunter
    text(x=(mx+q4x)/2,y=q4y+((q4y-my)/3),"+")
    #daneben re
    text(x=q4x+((q4x-mx)/3),y=(my+q4y)/2,"-")
#Beschriftung
mtext(text="x",side=1,line=1,at=max(x))
mtext(text="y",side=2,line=1,at=max(y),las=2)
mtext(text=expression(bar(x)),side=1,line=1,at=mx)
mtext(text=expression(bar(y)),side=2,line=1,at=my, las=2)
mtext(text="-",side=1,at=max(x),cex=2,line=-1)
```

```
mtext(text="+",side=1,at=min(x),cex=2,line=-1)
mtext(text="+",side=3,at=max(x),cex=2,line=-2.2)
mtext(text="-",side=3,at=min(x),cex=2,line=-2.2)
```

8.7.8 Exponentiale Verteilung

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(4,4,3,2)+0.1)
#Datenbereitstellung
x1 <- c(1.5,2,2.3,2.5,3.5,4,4.5,6,6.5,7,7.5,9,9.5,11,12)
y1 <- c(9,11,8,6,6,4.5,3,3.3,2.5,2.8,2,2.2,1.5,1,0.8)
x2 <- log(x1)
y2 <- log(y1)
a <- list(x1,x2)
b <- list(y1,y2)
c <- list("","log Einteilung")
#-------------------------------------------------------------------------------
#Plot
for (i in 1:2){
  plot(x=a[[i]],y=b[[i]],type="p",pch=16,xlab="",ylab="",xlim=c(0,max(a[[i]])),
       ylim=c(0,max(b[[i]])),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=FALSE,tck=0)
  axis(side=2,pos=0,labels=FALSE,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=(max(b[[i]])),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=(max(a[[i]])),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(a[[i]]))
  mtext("y",side=2,line=0,at=max(b[[i]]),las=2)
  mtext(c[[i]],side=2,line=0)
}
```

8.7.9 Ausreißer

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x1 <- c(1,2,2.1,2.2,2.4,3,3.3,4,4.5,4.7,4.9,5,5.2,5.3,20)
y1 <- c(2,1,3,4,2.5,1.8,1.6,2.8,3.5,2.4,2,2.8,3.5,4,8)
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="p",pch=16,main="",xlab="",ylab="",xlim=c(0,max(x1)),
     ylim=c(0,max(y1)),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=F,tck=0)
  axis(side=2,pos=0,labels=F,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=max(y1)+(1/10*max(y1)),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(x1)+(1/10*max(x1)),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
```

```
  mtext("x",side=1,line=0,at=max(x1)+(1/10*max(x1)))
  mtext("y",side=2,line=0,at=max(y1)+(1/10*max(y1)),las=2)
```

8.7.10 Inhomogenitat

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x1 <- c(1,2,2.1,2.2,2.4,3,3.3,4)
y1 <- c(2,1,3,4,2.5,1.8,1.6,2.8)
x2 <- c(20,20.2,20.3,20.5,21,21.5,22)
y2 <- c(7,7.2,8,9,7,9.5,8)
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="p",pch=16,main="",xlab="",ylab="",xlim=c(0,max(x2)),
     ylim=c(0,max(y2)),axes=FALSE,lwd=2)
#Datenpunkte/Graph
  points(x=x2,y=y2,pch=1)
#Achsen
  axis(side=1,pos=0,labels=F,tck=0)
  axis(side=2,pos=0,labels=F,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=max(y2)+(1/10*max(y2)),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(x2)+(1/10*max(x2)),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(x2)+(1/10*max(x2)))
  mtext("y",side=2,line=0,at=max(y2)+(1/10*max(y2)),las=2)
```

8.7.11 Größenvergleich bei Vätern und Söhnen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(5,4,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
n=20;sd=1.5
mean1 <- c(170,172,174,176,178,180)
mean2 <- c(171,173,175,177,179,181)
x1 <- c(rnorm(n=n,mean=mean1[1],sd=sd),
        rnorm(n=n,mean=mean1[2],sd=sd),
        rnorm(n=n,mean=mean1[3],sd=sd),
        rnorm(n=n,mean=mean1[4],sd=sd),
        rnorm(n=n,mean=mean1[5],sd=sd),
        rnorm(n=n,mean=mean1[6],sd=sd))
y1 <- c(rnorm(n=n,mean=mean2[1],sd=sd),
        rnorm(n=n,mean=mean2[2],sd=sd),
        rnorm(n=n,mean=mean2[3],sd=sd),
        rnorm(n=n,mean=mean2[4],sd=sd),
        rnorm(n=n,mean=mean2[5],sd=sd),
        rnorm(n=n,mean=mean2[6],sd=sd))
#-------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="p",pch=16,main="Vergleich Vater - Sohn",xlab="Größe der Väter [cm]",
     ylab="Größe der Söhne [cm]",axes=T,bty="L",lwd=1)
```

8.7.12 Regressionsgerade

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x1 <- c(0,1,2,3,4,5,6,7,8,9,10)
y1 <- 2*x1+3
#------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="l",main="",xlab="",ylab="",xlim=c(0,max(x1)+1),
     ylim=c(0,max(y1)),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=F,tck=0)
  axis(side=2,pos=0,labels=F,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=max(y1)+(1/10*max(y1)),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(x1)+(1/10*max(x1)),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
#point 2
  points(x=c(x1[2]),y=c(y1[2]-3),pch=16)
  lines(x=c(x1[2],x1[2]),y=c(y1[2],y1[2]-3))
  text(expression((list(x[1],y[1]))),x=x1[2],y=y1[2]-4)
  text(expression(D[1]),x=(x1[2]+x1[3])/2,y=y1[2]-1.5)
#point 3
  points(x=c(x1[3]),y=c(y1[3]+2),pch=16)
  lines(x=c(x1[3],x1[3]),y=c(y1[3],y1[3]+2))
  text(expression((list(x[2],y[2]))),x=x1[3],y=y1[3]+4)
  text(expression(D[2]),x=(x1[2]+x1[3])/2,y=y1[3]+1.5)
#point 4
  points(x=c(x1[4]),y=c(y1[4]-3),pch=16)
  lines(x=c(x1[4],x1[4]),y=c(y1[4],y1[4]-3))
#point 5
  points(x=c(x1[5]),y=c(y1[5]+4),pch=16)
  lines(x=c(x1[5],x1[5]),y=c(y1[5],y1[5]+4))
#point 6
  points(x=c(x1[6]),y=c(y1[6]+1),pch=16)
  lines(x=c(x1[6],x1[6]),y=c(y1[6],y1[6]+1))
#ponit 7
  points(x=c(x1[7]),y=c(y1[7]+2),pch=16)
  lines(x=c(x1[7],x1[7]),y=c(y1[7],y1[7]+2))
#point 9
  points(x=c(x1[9]),y=c(y1[9]-4),pch=16)
  lines(x=c(x1[9],x1[9]),y=c(y1[9],y1[9]-4))
#point 10
  points(x=c(x1[10]),y=c(y1[10]+1),pch=16)
  lines(x=c(x1[10],x1[10]),y=c(y1[10],y1[10]+1))
  text(expression((list(x[n],y[n]))),x=x1[10],y=y1[10]+2)
  text(expression(D[n]),x=(x1[9]+x1[10])/2,y=y1[10]+0.5)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(x1)+(1/10*max(x1)))
  mtext("y",side=2,line=0,at=max(y1)+(1/10*max(y1)),las=2)
```

8.7.13 Regression

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x1 <- c(2,3,4,5,6,7,9,10)
y1 <- c(2,9,6,15,14,17,15,22)
meanx1 <- mean(x1)
meany1 <- mean(y1)
SQX <- sum((x1-meanx1)^2)
SQY <- sum((y1-meany1)^2)
SPXY <- sum((x1-meanx1)*(y1-meany1))
bx <- SPXY/SQX
ax <- meany1-bx*meanx1
by <- SPXY/SQY
ay <- meanx1-by*meany1
X <- 0:max(x1)
X1 <- 0:max(y1)
Y <- ax+bx*X
Y1 <- ay+by*X1
#-------------------------------------------------------------------------------
#Plot
plot(X,Y,type="l",main="",xlab="",ylab="",xlim=c(0,max(X)),ylim=c(0,max(Y)),
     axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=T,tck=0)
  axis(side=2,pos=0,labels=T,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=max(Y)+(1/10*max(Y)),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(X)+(1/10*max(X)),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(X)+(1/10*max(X)))
  mtext("y",side=2,line=0,at=max(Y)+(1/10*max(Y)),las=2)
#empirische Datenpunkte
  points(x1,y1,pch=16)
#Linien
for (i in 1:8){
  lines(x=c(x1[[i]],x1[[i]]),y=c(y1[[i]],Y[x1[[i]]+1]))
  }
#-------------------------------------------------------------------------------
#Plot
plot(X1,Y1,type="l",main="",xlab="",ylab="",xlim=c(0,max(X1)),
     ylim=c(0,max(Y1)),axes=FALSE,lwd=2)
#Achsen
  axis(side=1,pos=0,labels=T,tck=0)
  axis(side=2,pos=0,labels=T,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=max(Y1)+(1/10*max(Y1)),angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(X1)+(1/10*max(X1)),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=0,at=max(X1)+(1/10*max(X1)))
  mtext("y",side=2,line=0,at=max(Y1)+(1/10*max(Y1)),las=2)
#empirische Datenpunkte
  points(y1,x1,pch=16)
#Linien
```

8.7.14 Regressionskoeffizient

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(5,4,4,4)+0.1)
#Datenbereitstellung
x <- c(41,42,42,43,45,47,48,49,49,50,51,52,54,54,55,56,57,57,58,58)
y <- c(130,125,140,135,140,140,155,150,160,155,165,160,170,165,170,160,175,165,170,175)
meanx <- mean(x)
meany <- mean(y)
SQX <- sum((x-meanx)^2)
SPXY <- sum((x-meanx)*(y-meany))
bx <- SPXY/SQX
ax <- meany-bx*meanx
X <- 38:max(x)
Y <- ax+bx*X
#------------------------------------------------------------------------------
#Plot
plot(X,Y,type="l",main="",xlab="",ylab="",xlim=,ylim=,axes=F,lwd=2)
#Datenpunkte/Graph
  points(x=x,y=y,pch=16)
#Achsen
  axis(side=1,pos=110,labels=F,tck=0)
  axis(side=2,pos=38,labels=F,tck=0)
  #Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=38,y0=110,x1=38,y1=max(y)+(1/20*max(y)),angle=20,length=0.125)
  arrows(x0=38,y0=110,x1=max(x)+(1/40*max(x)),y1=110,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  lines(x=c(X[1],x[14]),y=c(Y[1],Y[1]),lty=2)
  lines(x=c(x[14],x[14]),y=c(Y[1],y[14]),lty=2)
  par(xpd=TRUE)
  lines(x=c(38,37),y=c(110,110),lty=3)
  lines(x=c(38,37),y=c(Y[1],Y[1]),lty=3)
  arrows(x0=37,y0=110,x1=37,y1=Y[1],angle=20,length=0.125)
  arrows(x0=37,y0=Y[1],x1=37,y1=110,angle=20,length=0.125)
  par(xpd=FALSE)
#Beschriftungen
  mtext("x",side=1,line=4,at=max(x)+(1/40*max(x)))
  mtext("y",side=2,line=0,at=max(y)+(1/20*max(y)),las=2)
  mtext(text="a",side=2,line=1,las=2,at=120,cex=2)
  par(xpd=TRUE)
  text("1",x=(X[1]+x[14])/2,y=Y[1]-(1/40*Y[1]),cex=2)
  par(xpd=FALSE)
  text("b",x=x[14]+(1/40*x[14]),y=(Y[1]+y[14])/2,cex=2)
```

8.7.15 Residuenanalyse

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(2,2,2,2)+0.1)
#Datenbereitstellung
x1 <- c(0,1,2,3,4,5,6,7,8,9,10)
y1 <- 2*x1+3
#------------------------------------------------------------------------------
#Plot
plot(x1,y1,type="l",main="",xlab="",ylab="",xlim=c(0,max(x1)+1),
     ylim=c(0,max(y1)),axes=FALSE,lwd=2)
```

```
#Achsen
axis(side=1,pos=0,labels=F,tck=0)
axis(side=2,pos=0,labels=F,tck=0)
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=0,y0=0,x1=0,y1=max(y1)+(1/10*max(y1)),angle=20,length=0.125)
arrows(x0=0,y0=0,x1=max(x1)+(1/10*max(x1)),y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Datenpunkte/Graph
#point 2
points(x=c(x1[2]),y=c(y1[2]-3),pch=16)
#point 3
points(x=c(x1[3]),y=c(y1[3]+2),pch=16)
#point 4
points(x=c(x1[4]),y=c(y1[4]-3),pch=16)
#point 5
points(x=c(x1[5]),y=c(y1[5]+4),pch=16)
#point 6
points(x=c(x1[6]),y=c(y1[6]+1),pch=16)
#ponit 7
points(x=c(x1[7]),y=c(y1[7]+4),pch=16)
lines(x=c(x1[7],x1[7]),y=c(y1[7],y1[7]+4))
arrows(x0=x1[7]-0.2,y0=y1[7],x1=x1[7]-0.2,y1=y1[7]+4,angle=20,length=0.125)
arrows(x0=x1[7]-0.2,y0=y1[7]+4,x1=x1[7]-0.2,y1=y1[7],angle=20,length=0.125)
#point 9
points(x=c(x1[9]),y=c(y1[9]-4),pch=16)
#point 10
points(x=c(x1[10]),y=c(y1[10]+1),pch=16)
#Beschriftungen
mtext("x",side=1,line=0,at=max(x1)+(1/10*max(x1)))
mtext("y",side=2,line=0,at=max(y1)+(1/10*max(y1)),las=2)
text(expression((list(x[i],y[i]))),x=x1[7],y=y1[7]+6)
text(expression(e[i]),x=(x1[7]+x1[6])/2,y=y1[7]+2)
```

8.7.16 Quadratischer Zusammenhang der Residuen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,1,.1)+0.1,mgp=c(1,1,0))
#Datenbereitstellung
set.seed(1234567)
x <- seq(from=-2,to=2, length.out=50)
y <- x^2 + rnorm(length(x), sd=2.5)
yx <- lm(y ~ x)
#-------------------------------------------------------------------------------
#Plot
plot(x=predict(yx),y=resid(yx),
     ylab=expression(paste("Geschätzte Residuen  ",hat(e[i]))),
     xlab=expression(hat(y[i])),xaxt="n",yaxt="n")
#Flächen/Linien
abline(h=0,lty=2)
```

8.7.17 Trichterform der Residuen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,1,.1)+0.1,mgp=c(1,1,0))
#Datenbereitstellung
set.seed(1234)
x2 <- seq(from=-2, to=3, length.out=50)
y2 <- -1.1+.3*x+rpois(length(x),lambda=1:length(x))/2+rnorm(length(x),mean=0,sd=1)
yx2 <- lm(y2 ~ x2)
#-------------------------------------------------------------------------------
#Plot
plot(x=predict(yx2),y=resid(yx2),
     ylab=expression(paste("Geschätzte Residuen  ",hat(e[i]))),
     xlab=expression(hat(y[i])),xaxt="n",yaxt="n")
#Flächen/Linien
abline(h=0,lty=2)
```

8.7.18 Residuenplot

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,2,2)+0.1)
#Datenbereitstellung
x <- c(41,42,42,43,45,47,48,49,49,50,51,52,54,54,55,56,57,57,58,58)
y <- c(130,125,140,135,140,140,155,150,160,155,165,160,170,165,170,160,175,165,170,175)
yx <- lm(y ~ x)
#-------------------------------------------------------------------------------
#Plot
plot(x=predict(yx), y=resid(yx),ylab="Residuen",
     xlab="Geschätzte Werte für systolischen Blutdruck",ylim=c(-12,12),
     xlim=c(130,180))
#Flächen/Linien
abline(h=0, col="grey", lty=2)
```

8.7.19 Normalverteilung um den Erwartungswert

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
set.seed(1234567)
x1 <- rep(.2,20)
x2 <- rep(.4,20)
x3 <- rep(.6,20)
y1 <- rnorm(n=20,mean=12,sd=2)
y2 <- rnorm(n=20,mean=8,sd=2)
y3 <- rnorm(n=20,mean=4,sd=2)
x4 <- seq(from=8,to=16,by=.1)
x5 <- seq(from=4,to=12,by=.1)
x6 <- seq(from=0,to=8,by=.1)
y4 <- dnorm(x=x4,mean=12,sd=2)
y5 <- dnorm(x=x5,mean=8,sd=2)
y6 <- dnorm(x=x6,mean=4,sd=2)
#-------------------------------------------------------------------------------
#Plot
plot(x=c(0,.8),y=c(0,16),type="n",axes=F,xlab="",ylab="")
#Datenpunkte/Graph
points(x1,y1)
points(x2,y2)
```

```
points(x3,y3)
lines(x=c(.2,.6),y=c(mean(x4),mean(x6)),lwd=2)
points(x=.18+y4,y=x4,type='l')
points(x=.38+y5,y=x5,type='l')
points(x=.58+y6,y=x6,type='l')
lines(x=c(.2,.18+max(y4)),y=c(mean(x4),mean(x4)),lty=2)
lines(x=c(.4,.38+max(y5)),y=c(mean(x5),mean(x5)),lty=2)
lines(x=c(.6,.58+max(y6)),y=c(mean(x6),mean(x6)),lty=2)
#Achsen
axis(side=1,pos=0,labels=F,tck=0)
axis(side=2,pos=0,labels=F,tck=0)
#Pfeilspitzen
par(xpd=TRUE)
arrows(x0=0,y0=0,x1=0,y1=16,angle=20,length=0.125)
arrows(x0=0,y0=0,x1=.8,y1=0,angle=20,length=0.125)
par(xpd=FALSE)
#Beschriftungen
mtext("x",side=1,line=0,at=.8)
mtext("y",side=2,line=0,at=16,las=2)
```

8.7.20 Konfidenzband

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(3,3,3,3)+0.1)
#Datenbereitstellung
x <- c(41,42,42,43,45,47,48,49,49,50,51,52,54,54,55,56,57,57,58,58)
y <- c(130,125,140,135,140,140,155,150,160,155,165,160,170,165,170,160,175,165,170,175)
yx <- lm(y ~ x) # Modell y= a*x + b
x.neu <- data.frame(x = x)
yx.conf <- predict(yx, interval="confidence")
yx.pred <- predict(yx, newdata=x.neu, interval="prediction")
b <- round(coefficients(yx)[1],2)
a <- round(coefficients(yx)[2],2)
yx.sum <- summary(yx)
#-------------------------------------------------------------------------------
#Plot
matplot(x, cbind(yx.conf, yx.pred[,-1]), type="l",main="",xlab="Alter",
        ylab="Systol. Bluthochdruck [mmHg]",lty=c(1,2,2,3,3),col="black",lwd=2)
#Datenpunkte/Graph
points(x,y)
#Beschriftungen
legend("bottomright",
       legend=c("Geschätzte Regressionsgerade", "Konfidenzintervall zum Mittelwert",
                "Konfidenzintervall zu einer Beobachtung"),
       lty=c(1,2,3),col="black",cex=0.6)
legend("topleft",legend=c(paste("y = ", b, if (a>=0) "+", a,"x"),
                expression(paste(R^2," = 0.87"))),
                cex=0.8,bty="n")
#0.87 <- round(yx.sum$adj.r.squared,2)
```

8.8 Kapitel 7

8.8.1 Studienzeit und Patientenzeit

```
#Grafikeinstellungen
par(mfrow=c(1,2),mar=c(4,4,2,2)+0.1)
#Datenbereitstellung
data <- data.frame(x8=c(0,5),y8=c(8,8),x7=c(1,7.5),y7=c(7,7),
                   x6=c(1.5,10),y6=c(6,6),x5=c(1,9),y5=c(5,5),
                   x4=c(3,3.5),y4=c(4,4),x3=c(1.5,10),y3=c(3,3),
                   x2=c(2.8,4),y2=c(2,2),x1=c(2.7,5.2),y1=c(1,1))
a <- list(data$x8,data$x7,data$x6,data$x5,data$x4,data$x3,data$x2,data$x1)
a1 <-list(data$x8-(min(data$x8)),data$x7-(min(data$x7)),data$x6-(min(data$x6)),
          data$x5-(min(data$x5)),data$x4-(min(data$x4)),data$x3-(min(data$x3)),
          data$x2-(min(data$x2)),data$x1-(min(data$x1)))
b  <- list(data$y8,data$y7,data$y6,data$y5,data$y4,data$y3,data$y2,data$y1)
b1 <- list(data$y5,data$y4,data$y1,data$y3,data$y8,data$y2,data$y7,data$y6)
c  <- c("T","A","L","T","T","L","A","T")
c1 <- c("T","Z","Z","T","T","Z","Z","T")
x <- 0:11; y.max <- 9; #y <- 0:9
#-------------------------------------------------------------------------------
#Plot
  plot(x, y=seq(from=0, to=y.max, length.out=length(x)),
       type="n",axes=F, xlab="Kalender-Zeit",ylab="Patient")
#Achsen
  axis(side=2, pos=0,at=1:8,labels=8:1)
  mtext("t",side=1, at=11.2,line=-0.8)
#Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=y.max,angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(x),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Flächen/Linien
  abline(v=3.2,lty=2);#mtext("Ende der Rekrutierungsphase",side=1,at=2)
  mtext("Ende der",side=1,at=3.2); mtext("Rekrutierungsphase",side=1,at=3.2, line=1)
  abline(v=10,lty=2);#mtext("Ende der Studie",side=1,at=10)
  mtext("Ende der",side=1,at=10); mtext("Studie",side=1,at=10, line=1)
#Datenpunkte/Graph
for (i in 1:8){
  points(x=a[[i]],y=b[[i]]); lines(x=a[[i]],y=b[[i]])
  text(c[i], x=max(a[[i]])+0.5, y=b[[i]])
}
#-------------------------------------------------------------------------------
#Plot
  plot(x, y=seq(from=0, to=y.max, length.out=length(x)),
       type="n",axes=F,xlab="Ereignis-Zeit", ylab="Patient")
#Achsen
  axis(side=2,pos=0,at=8:1,labels=c(5,7,8,1,2,4,6,3))
  mtext("t=0",side=1,at=0,line=-0.8)
  mtext("t",side=1,at=11.2,line=-0.8)
#Pfeilspitzen
  par(xpd=TRUE)
  arrows(x0=0,y0=0,x1=0,y1=y.max,angle=20,length=0.125)
  arrows(x0=0,y0=0,x1=max(x),y1=0,angle=20,length=0.125)
  par(xpd=FALSE)
#Datenpunkte/Graph
for (i in 1:8){
  points(x=a1[[i]],y=b1[[i]]); lines(x=a1[[i]],y=b1[[i]])
  text(c1[i], x=max(a1[[i]])+0.5, y=b1[[i]])
}
```

8.8.2 mediane Ueberlebenszeit

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,2,2)+0.1)

data <- data.frame(x1=count,x2=time,x3=event)
fit<-survfit(Surv(time=data$x2,event=data$x3)~1,data=data)
median <- summary(fit)$table["median"]
#-------------------------------------------------------------------------------
#Plot
  plot(fit, conf.int=F, xlab="Monate",
       ylab="Überlebenswahrscheinlichkeit", cex.axis=0.8, cex.lab=1.2)
#Flächen/Linien
  lines(x=c(0,median),y=c(0.5,0.5),lty=2,lwd=2)
  lines(x=c(median,median),y=c(0,0.5),lty=2,lwd=2)
#Beschriftung
  legend(x=1 * median, y=0.15,
         legend=c("mediane", "Überlebenszeit"), bty="n")
```

8.8.3 Überlebenszeiten der beiden Gruppen

```
#Grafikeinstellungen
par(mfrow=c(1,1),mar=c(4,4,2,2)+0.1)
#Datenbereitstellung
count_both <- c(count, count_nt)
time_both <- c(time, time_nt)
event_both <- c(event, event_nt)
group <- c(rep("Treated",10),rep("NonTreaded",10))
data <- data.frame(x1=time_both, x2=event_both, x3=group)
fit <- survfit(Surv(time=x1, event=x2)~x3, data=data)
#-------------------------------------------------------------------------------
#Plot
  plot(fit,conf.int=F,xlab="Monate",
       ylab="Überlebenswahrscheinlichkeit",
       cex.axis=0.8, cex.lab=1.2,lty=c(2,1))
#Beschriftung
  legend("bottomleft",legend=c("Behandelt","Unbehandelt"),lty=c(1,2))
```