

Biometrie

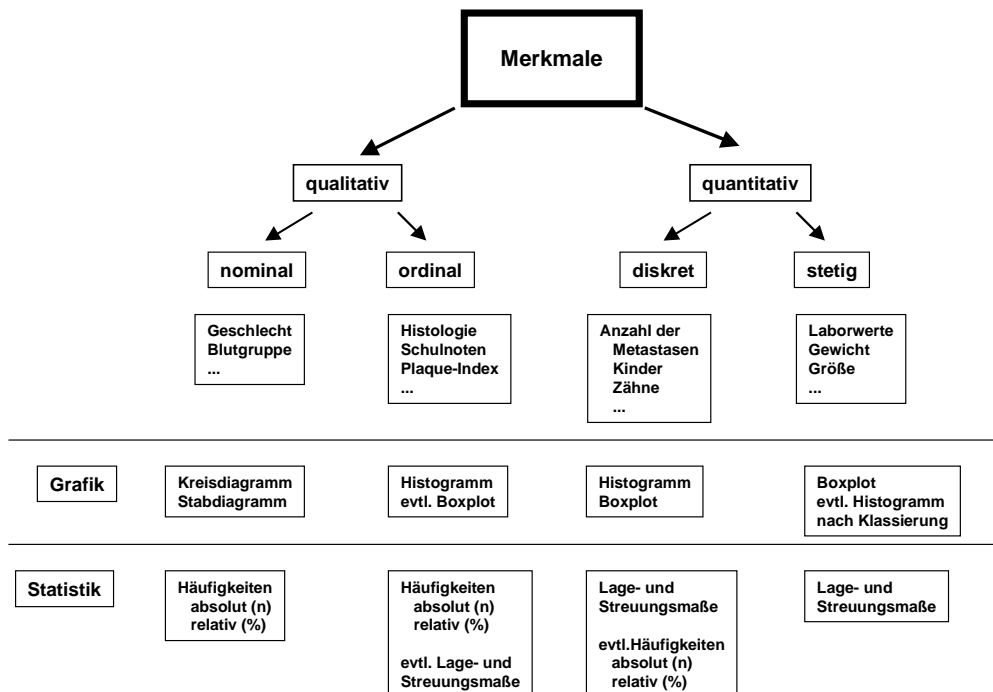
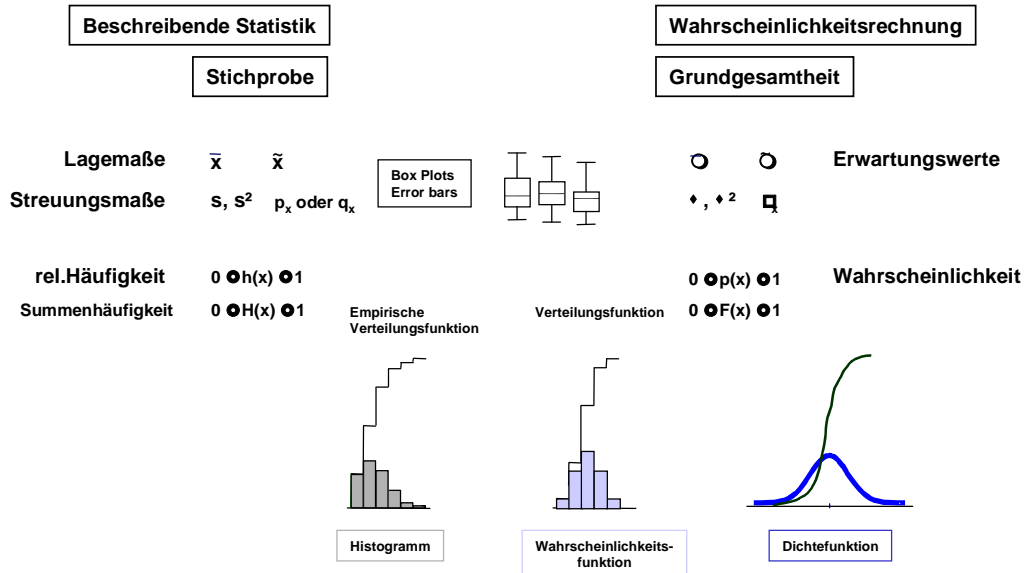
Werner Hopfenmüller
PD. Dr.rer.nat. Dr.med.

Biometrie / Biomathematik
Medizinische Statistik
Mathematische Modelle
Studienplanung
Auswertung klinischer Studien
Epidemiologie

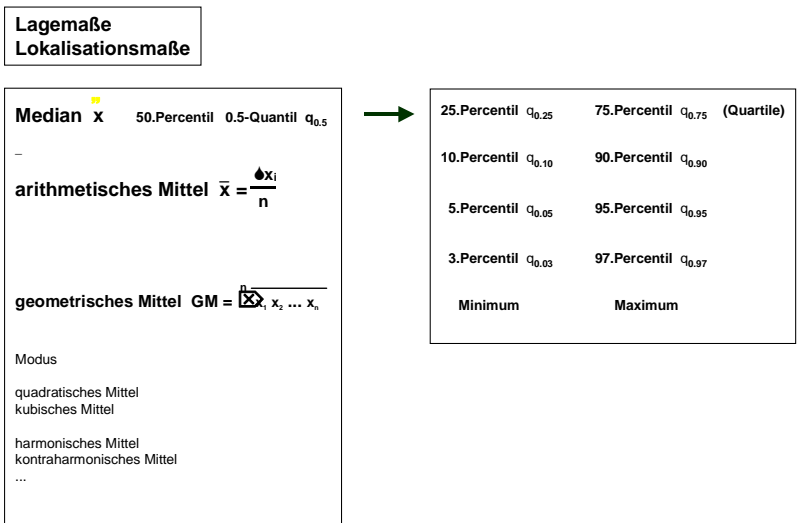
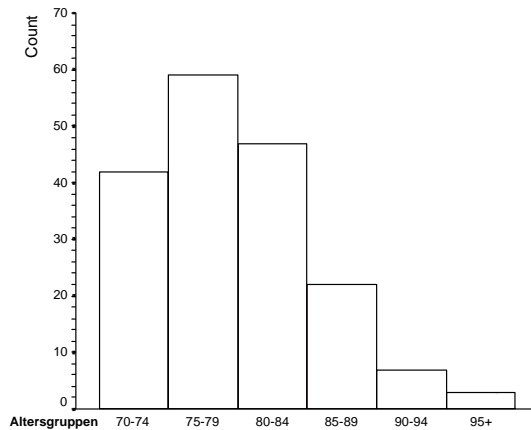
Institut für Biometrie und Klinische Epidemiologie
Universitätsklinikum Benjamin Franklin

Tel. 030 8445 2092 oder 030 8445 3512
FAX 030 8445 4471
E-mail werner.hopfenmueller@charite.de

Überblick Biomathematik



Häufigkeitsdiagramm / Histogramm



Lagemaße (Lokalisationsmaße)

Arithmetisches Mittel (mean) $\bar{x} = \frac{\sum x_i}{n}$

Er ist die Summe aller Beobachtungen, geteilt durch die Anzahl dieser Beobachtungen.

Der arithmetische Mittelwert stellt ein gutes Lagemaß für symmetrische Verteilungen dar, ist aber wenig aussagekräftig bei Verteilung mit einseitigen Ausreißern (schiefe Verteilungen). Bei normalverteilten Merkmalen gibt er die Lokalisation des mittleren Wertes (= Median s.u.) einer Meßreihe an. Bei nicht stetigen Parametern (z.B. Histologie etc) ist zur Beschreibung oft ein Histogramm aussagekräftiger als das arithmetische Mittel, da die Abstände zwischen den Werten nicht identisch sind

und der mathematisch korrekte Mittelwert als Merkmalsausprägung nicht vorkommen kann (z.B. 1,5 Kinder).

Median \tilde{x}

Der empirische *Median* (oder Zentralwert) teilt die Stichprobenwerte in 2 Hälften: Die eine Hälfte der Daten ist höchstens so groß wie der Median, die andere Hälfte ist größer oder gleich dem Median.

Demnach entspricht der Median dem 50. Perzentil. In einer rangierten Meßreihe steht er in der Mitte.

!: Bei gerader Anzahl von Werten ist der Median der Mittelwert der beiden in der Mitte stehenden Werte.

Bei symmetrischen Verteilungen sind \bar{x} und \tilde{x} nahezu identisch.

Differenz $\bar{x} - \tilde{x}$ und Quotient \bar{x} / \tilde{x} sind ein Maß für die Asymmetrie der Verteilung (s.a. Abb 2 übernächste Seite).

Der Median ist geeignet bei:

- nicht symmetrischen Verteilungen
- Verdacht auf Ausreißer
- bei ordinalskalierten Daten
- auch bei wenigen Beobachtungen

Bei mehrgipfligen Verteilungen ist der Median nicht zu empfehlen.

Percentile

In Anlehnung an das 50. Perzentil (Median) kann man in einer geordneten Meßreihe weitere Percentile (oder Quantile) bilden:

z.B. $q_{0.25} = 25.$ Perzentil, d.h. 25 % der Meßwerte sind kleiner als $q_{0.25}$

Üblicherweise werden bestimmt:

- $q_{0.25}$ und $q_{0.75}$ (auch Quartile genannt)
- $q_{0.10}$, $q_{0.90}$, $q_{0.05}$, $q_{0.95}$ oder $q_{0.03}$ und $q_{0.97}$ als Grenzen eines Normalbereichs
- Ergänzung $p_0 = \text{Min}$, $p_{100} = \text{Max}$

(Kombination von Median und Percentilen s.Abb nächste Seite *Box Whisker Plot*)

Modus (Modalwert)

Der *Modus* kann in allen Skalentypen ermittelt werden. Er entspricht bei diskreten Daten der Ausprägung mit der größten Häufigkeit. Bei klassierten Daten läßt sich die *modale Klasse* angeben – das ist die Klasse mit der größten Besetzungszahl. Deren Mitte gilt dann als der Modus.

Math. Definition: Der Modus ist derjenige Werte, an dem die Dichtefunktion einer stetigen Verteilung maximal wird.

Geometrisches Mittel

Das *geometrische Mittel* wird bei relativen Änderungen verwendet, bei denen sich der Unterschied zweier Merkmalsausprägungen besser durch einen Quotienten als durch eine Differenz beschreiben läßt. Dies ist der Fall bei Wachstumserscheinungen (Zunahme der Bevölkerung, Zunahme von Kosten), sowie bei Verdünnungsreihen (Antikörpertiter)

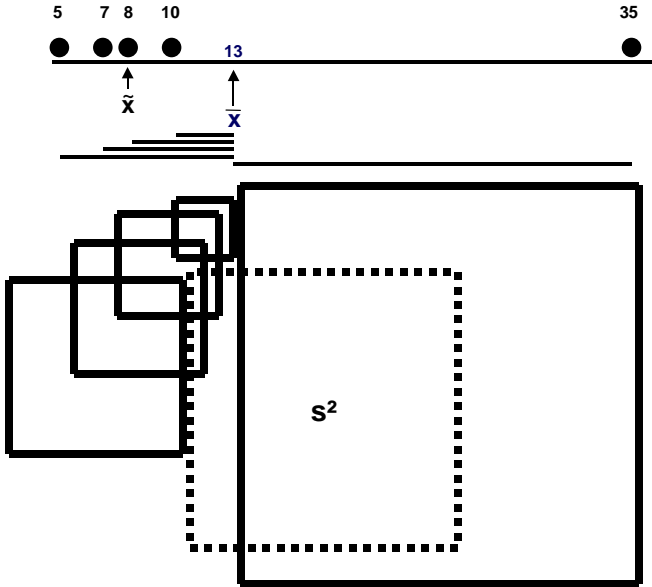
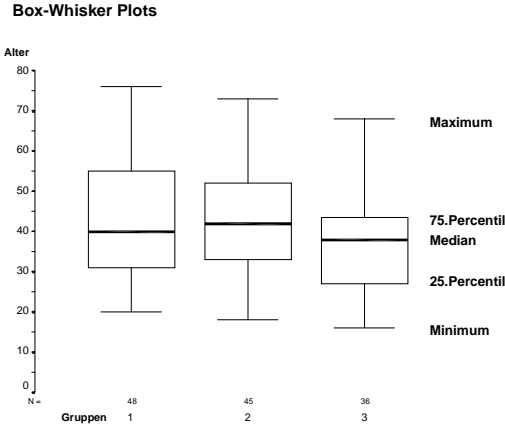
$$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Bemerkung:

$$\log x_G = \frac{1}{n} \log(x_1 \cdot x_2 \cdot \dots \cdot x_n) = \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n)$$

d.h. Der Logarithmus des geometrischen Mittels entspricht dem arithmetischen Mittel der Logarithmen

Abbildungen



Streuungsmaße (Dispersionsmaße)

Spannweite (Variationsbreite, Range)

Das einfachste Streuungsmaß ist die Spannweite vom Minimum zum Maximum (engl.: range) (wird bevorzugt bei kleinen Stichproben $n \leq 10$). Da die Spannweite nur die beiden Extremwerte berücksichtigt, ist sie weit mehr von Ausreißern beeinflusst als alle anderen Streuungsmaße.

Varianz s^2 Standardabweichung s

Das arithmetische Mittel ist ein häufig benutztes Lagemaß. Es liegt daher nahe, ein Streuungsmaß zu definieren, das die Abweichungen der Meßwerte vom Mittelwert quantifiziert. Ein solches Maß ist die *Varianz*, sie ist die mittlere quadratische Abweichung der Meßwerte vom Mittelwert

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

Die Varianz hat eine quadratische Dimension und ist deshalb schwer zu interpretieren (s.a Abb 2 oben). Um ein Streuungsmaß mit gleicher Dimension wie die Meßwerte zu erhalten, zieht man die Wurzel aus der Varianz und erhält die Standardabweichung s .

Es ist üblich, quantitative, annähernd symmetrisch verteilte Meßwerte durch Mittelwert und Standardabweichung unter Angabe des Stichprobenumfangs n zu charakterisieren.

Dabei gelten folgende Abschätzungen:

Im Bereich $x \pm s$	liegen etwa 68 % der Meßwerte
Im Bereich $x \pm 2s$	liegen etwa 95 % der Meßwerte
Im Bereich $x \pm 3s$	liegen etwa 99,7 % der Meßwerte

Variationskoeffizient (relative Standardabweichung) V_k

$$V_k = \frac{s}{\bar{x}} \cdot 100 [\%]$$

geeignet zur Bestimmung von Meßfehlern bei Meßwiederholungen,
z.B. bei Analysegeräten in der Labormedizin, bei ca. 20 Messungen
wenn kleiner als 5 %: gute Präzision (bei Tests in der Klinischen Chemie)
wenn kleiner als 10 %: akzeptabel

Standardfehler des arithmetischen Mittels (standard error of mean (SEM))

$$SEM = \frac{s}{\sqrt{n}}$$

SEM ist ein Streuungsmaß für den Mittelwert und nicht für die Einzelwerte. (s. Konfidenzbereiche)
z.B. $\bar{x} \pm SEM = 68\text{-Konfidenzintervall (unüblich)}$
 $\bar{x} \pm 2 SEM = 95\text{-Konfidenzintervall } (\bar{x} \pm 1.96 SEM)$

Interpercentilbereiche

Interpercentile sind Streuungsmaße die üblicherweise mit dem Median angegeben werden.
(s. Abb 1 nächste Seite)

Streuungsmaße
Dispersionsmaße

Median

Interpercentilbereiche

- $q_{0.25} - q_{0.75}$
- $q_{0.10} - q_{0.90}$
- $q_{0.05} - q_{0.95}$
- $q_{0.03} - q_{0.97}$
- Min - Max

Spannweite (range)

Arithm. Mittel

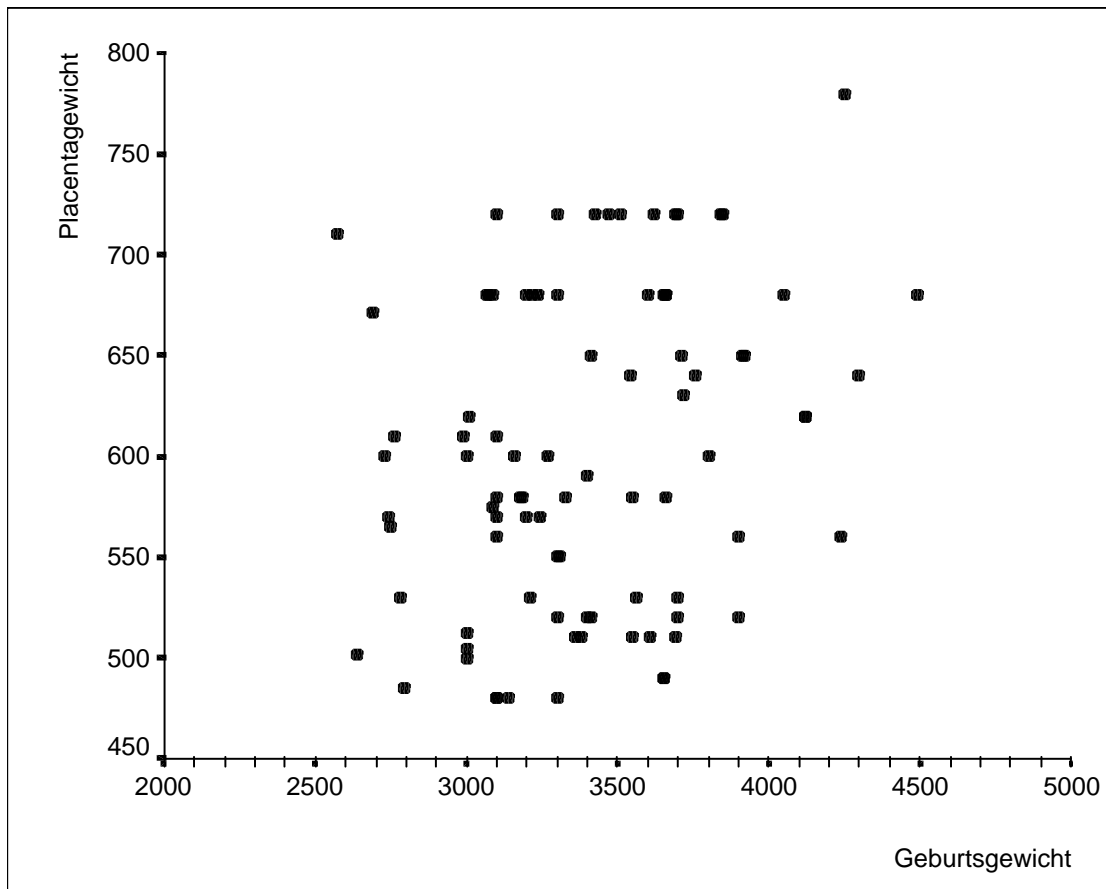
Standardabweichung $s = \sqrt{\frac{\sum (X_i - \bar{x})^2}{n - 1}}$

- $\bar{x} \pm s$ zentraler 68%- Bereich
- $\bar{x} \pm 2s$ zentraler 95%- Bereich (95.5%)
- $\bar{x} \pm 3s$ zentraler 99.7%- Bereich

Variationskoeffizient (rel. St.abw.) $V_k = \frac{s}{\bar{x}} \cdot 100 [\%]$

s^2 Varianz

$\frac{s}{\sqrt{n}}$ = SEM (standard error of mean) → **Konfidenzintervalle**



Korrelation und Regression

Beschrieben werden statistische Abhängigkeiten von stetigen Merkmalen. Bei 2 Merkmalen bietet es sich an, jeder Beobachtungseinheit ein Wertepaar (x_i / y_i) zuzuordnen und diese Punkte in ein Koordinatensystem einzutragen. Wenn beide Merkmale zufällige Werte sind, erhält man auf diese Weise eine Punktwolke (Korrelogramm) (s.o. Abb.2). Wie bei mathematischen Gleichungen üblich, sollte x das unabhängige und y das abhängige Merkmal sein.

Anhand der Punktwolke lassen sich charakteristische Eigenschaften erkennen:

- Stärke des Zusammenhangs:
Je dichter die Punkte beieinander liegen, desto stärker ist der Zusammenhang.
Mit Hilfe der Korrelationsanalyse lassen sich Maßzahlen errechnen (Korrelationskoeffizienten, Bestimmtheitsmaß), die Stärke des Zusammenhangs quantifizieren.
- Art des Zusammenhangs:
Diese wird durch die mathematische Funktion angegeben, die den Zusammenhang am besten beschreibt (Regressionsanalyse).
Kann der Zusammenhang durch eine Gerade charakterisiert werden, spricht man von einem linearen Zusammenhang. Die dazugehörige mathematische Funktion nennt man Regressionsgerade.

Linearer Korrelationskoeffizient r (nach Pearson)

Voraussetzungen:

- beide Merkmale x und y sind stetig und annähernd normalverteilt (keine Ausreißer)
- der Zusammenhang ist annähernd linear
- die Beobachtungspaare (x_i / y_i) sind unabhängig voneinander

Der Korrelationskoeffizient (r) ist ein Maß für die Stärke eines linearen Zusammenhanges. Ist der Zusammenhang deutlich, so ist es durchaus möglich, die Regressionsgerade in die Punktwolke zu legen. r beschreibt, wie gut die Punkte sich der Geraden annähern

$$r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Für r gilt $-1 \leq r \leq +1$

Ist $r > 0$, so handelt es sich um eine positive Korrelation (Regressionsgerade mit positiver Steigung)

Ist $r < 0$, so handelt es sich um eine negative Korrelation (Regressionsgerade mit negativer Steigung)

Je näher der Betrag von r ($|r|$) bei 1 liegt, desto stärker ist der Zusammenhang und desto dichter liegen die Punkte (x_i / y_i) an der Regressionsgerade.

Je näher r bei 0 liegt, desto schwächer ist der Zusammenhang und desto weiter ist die Punktwolke um die Regressionsgerade gestreut.

Qualitative Beurteilung von r (bei wenigstens 30 Wertepaaren)

$ r \leq 0,2$	kein statistische Bedeutung
$0,2 < r \leq 0,4$	schwache Korrelation
$0,4 < r \leq 0,75$	mittlere Korrelation
$ r > 0,75$	starke Korrelation

Korrelationskoeffizient r_s (nach Spearman)

Während r auf Mittelwerten und Standardabweichungen basiert, benutzt r_s die Rangfolgen der beiden Merkmale. r_s wird als Rangkorrelationskoeffizient bezeichnet und ist eine verteilungsfreie Variante für das Pearson r . Der Spearman'sche Koeffizient ist auch ein Maß für die Stärke eines monotonen Zusammenhangs.

Jeder Beobachtungseinheit kann eine Rangzahl für das x-Merkmal und eine für das y-Merkmal zugeordnet werden. Die Differenz dieser beiden Rangzahlen sei d_i . Aus diesen Differenzen wird der Spearman'sche Korrelationskoeffizient nach folgender Formel errechnet:

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

Vorteile von r_s :

- r_s kann auch bei Ausreißern benutzt werden
- r_s kann bei ordinalskalierten Merkmalen eingesetzt werden
- r_s erkennt auch nichtlineare monotone Zusammenhänge (Sättigungskurven, exponentieller Abfall)

Die qualitativen Beurteilungen von r und r_s sind identisch.

Bestimmtheitsmaß oder Determinationskoeffizient (r^2 oder r_s^2)

Dies ist ein weit verbreitetes Maß für die Güte einer Korrelation.

r^2 ist der relative Anteil der durch die Regression erklärten Varianz an der Gesamtvarianz der y-Werte. Die Gesamtvarianz wird üblicherweise mit 100% festgelegt.

Beispiel:

Wir wollen untersuchen, wie groß der Zusammenhang zwischen Geburtsgewicht und Placentagewicht ist.

Wir nehmen an, daß $r = 0,4$, d.h. $r^2 = 0,16$

D.h.: 16% der Streuung der y-Werte (Placentagewicht) ist durch die Abhängigkeit mit x (Geburtsgewicht) erklärt (erklärte Varianz s.o.)

84% der Streuung ist auf eine andere Einflußgröße zurückzuführen oder ist zufällig zustande gekommen (nicht erklärte Varianz)

Regression

Bei ausreichend großen Korrelationskoeffizienten kann der Punktwolke eine Ausgleichsgerade (Regressionsgerade) angepaßt werden.

Eine Gerade hat allgemein die Gleichung: $y=a+bx$

a ist der Achsenabschnitt der Geraden auf der Y-Achse

b ist die Steigung der Regressionsgerade und heißt auch Regressionskoeffizient.

Eine Regressionsgerade geht immer durch den Schwerpunkt (x, y) der Punktwolke

wenn $b>0$ ### $r>0$

wenn $b<0$ ### $r<0$

Wahrscheinlichkeitsrechnung

Der Begriff ‚wahrscheinlich‘ entstammt unserer Umgangssprache. Mit Sätzen ‚Wahrscheinlich werden sie in einem halben Jahr geheilt sein‘, drücken wir subjektive Wahrscheinlichkeiten aus, die auf Erfahrungen basieren, wobei es nicht möglich ist diese Wahrscheinlichkeiten exakt zu quantifizieren.

Auch die Prozesse und Entwicklungen in den Biowissenschaften unterliegen dem Zufall. Man bezeichnet sie als probabilistisch – im Gegensatz zu deterministischen Vorgängen, die sich exakt berechnen lassen. Praktisch alle Aussagen, die durch die analytische Statistik ermöglicht werden, sind mit Irrtumswahrscheinlichkeiten behaftet. Für wissenschaftliche Untersuchungen ist es deswegen notwendig, den Begriff der Wahrscheinlichkeit zu präzisieren und quantitativ zu beschreiben.

Für den praktischen Anwender sind Kenntnisse aus der Wahrscheinlichkeitsrechnung hilfreich und notwendig, um die statistischen Methoden zu verstehen und sinnvoll mit ihnen umgehen zu können.

Wahrscheinlichkeiten medizinischer Ereignisse (Krankheiten, Risiken etc) sind in der Regel nur schätzbar. Exakt berechenbar sind sie im Rahmen von Zufallsexperimenten, in denen die Wahrscheinlichkeiten der Einzelereignisse gegeben sind.

Das klassische Beispiel ist das Würfeln:

Jedermann weiß, auch ohne theoretisch-statistische Kenntnisse zu haben, daß die Wahrscheinlichkeit für das Würfeln einer der sechs möglichen Zahlen $1/6$ ist (falls der Würfel symmetrisch ist)

Man stützt sich dabei intuitiv auf folgende Formel nach Laplace:

$$P(E) = \text{Anzahl der günstigen Ereignisse} / \text{Anzahl der möglichen Ereignisse}$$

Für die relative **Häufigkeit** und für die **Wahrscheinlichkeit** gelten dieselben Rechenvorschriften. Dennoch ist darauf hinzuweisen, daß zwischen Wahrscheinlichkeit und Häufigkeit ein grundsätzlicher Unterschied besteht.

Häufigkeiten kennzeichnen Einzelversuche mit Stichprobencharakter. Sie können bei jedem Versuch andere Werte annehmen; die Wahrscheinlichkeit hingegen kennzeichnet Zufallsexperimente und ist damit eine fest zugeordnete Zahl.

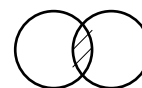
Bsp.: Die Wahrscheinlichkeit, eine „3“ zu würfeln, ist immer $1/6$. Die Häufigkeit bei n Versuchen, eine „3“ zu würfeln, kann dagegen sehr unterschiedlich ausfallen.

Rechenregeln für Wahrscheinlichkeiten (Axiome nach Kolmogoroff)

Vorbemerkung: Analog zur Mengenlehre definiert man

- Durchschnitt zweier Ereignisse $A \cap B$ („und“)

Bei Nichtüberschneidung spricht man von disjunkten Ereignissen



- Vereinigung zweier Ereignisse $A \cup B$ („oder“)



(Das mathematische „oder“ ist ein „nicht ausschließendes oder“)

Axiome A1, A2, A3 der Wahrscheinlichkeitsrechnung

- A1 legt den Definitionsbereich der Wahrscheinlichkeit fest: $0 \leq p(E) \leq 1$
D.h., jedem zufälligen Ereignis E wird eine Zahl p(E) zugeordnet, die zwischen 0 und 1 liegt.
p(E) wird als Wahrscheinlichkeit des Ereignisses E bezeichnet
- A2 legt die Richtung des Definitionsbereiches fest: $p(S) = 1$.
D.h., das sichere Ereignis hat die Wahrscheinlichkeit 1.
Daraus folgt unmittelbar, daß das unmögliche Ereignis die Wahrscheinlichkeit 0 hat.
Das unmögliche und das sichere Ereignis sind komplementäre Ereignisse, deren Wahrscheinlichkeiten sich zu 1 addieren.
Weiteres Beispiel für komplementäre Ereignisse: $p(\text{Tumor}) + p(\text{Non-Tumor}) = 1$.
oder allg.: $p(A) = 1 - p(\text{non}A)$ (genannt: "Komplement") **Abk.:** $p(\text{non}A) = p(\bar{A})$
- Beispiel.: Die Wahrscheinlichkeit eine ‚1‘ zu würfeln ist $1/6$
die Wahrscheinlichkeit keine ‚1‘ zu würfeln ist $1 - 1/6 = 5/6$.
- A3 heißt auch Additionssatz der Wahrscheinlichkeitsrechnung und berechnet die Wahrscheinlichkeit zweier oder mehrerer Ereignisse, die mit „oder“ verknüpft sind:
Es gilt: $p(E_1 \text{ \#\#\# } E_2) = p(E_1) + p(E_2)$
wenn E_1 und E_2 sich gegenseitig **ausschließen** (disjunkte Ereignisse).

Beispiel:

Die Wahrscheinlichkeit, mit einem regelmäßigen Würfel eine 3 oder 4 zu würfeln, beträgt: $1/6 + 1/6 = 1/3$,

Bei zwei Ereignissen, die sich **nicht ausschließen**, muß die Formel etwas verändert werden:

$$p(E_1 \text{ \#\#\# } E_2) = p(E_1) + p(E_2) - p(E_1 \text{ \#\#\# } E_2)$$

d.h. die Schnittmenge beider Ereignisse wird abgezogen, da sie sonst doppelt gezählt würde.

Beispiel:

Die Wahrscheinlichkeiten Windpocken (W) oder Masern (M) zu bekommen seien:

$$p(W) = 0,2$$

$$p(M) = 0,3.$$

$$\Rightarrow p(\text{Masern und Windpocken}): p(W \cap M) = 0,2 \times 0,3 = 0,06$$

$$\text{\#\#\# } p(\text{Masern oder Windpocken}): p(W \text{ \#\#\# } M) = 0,2 + 0,3 - 0,06 = 0,5 - 0,06 = 0,44$$

Aus den Axiomen abgeleitet wird der sog. Multiplikationssatz, der die Wahrscheinlichkeit zweier oder mehrerer Ereignisse, die mit „und“ verknüpft sind, berechnet.

Der Multiplikationssatz wird heute als Definition für die Unabhängigkeit zweier Ereignisse gebraucht:

Zwei Ereignisse A und B sind genau dann unabhängig, wenn gilt:

$$p(A \cap B) = p(A) \cdot p(B)$$

Beispiel:

Wirksamkeit einer Kombinationstherapie von 3 Antibiotika A, B, C, deren Wirkungen voneinander unabhängig sein sollen, d.h. sie verhalten sich weder synergistisch noch antagonistisch.

Wahrscheinlichkeit für die einzelnen Wirksamkeiten:

$$p(A) = 0,7 \quad p(B) = 0,6 \quad p(C) = 0,5$$

Frage: Wie groß ist die Wirksamkeit der Kombinationstherapie ABC?

Die Wahrscheinlichkeiten für die Nichtwirksamkeit (komplementäre Ereignisse) sind:

$$p(\bar{A}) = 1 - p(A) = 0,3$$

$$p(\bar{B}) = 1 - p(B) = 0,4$$

$$p(\bar{C}) = 1 - p(C) = 0,5$$

Die Wahrscheinlichkeit, daß alle 3 Antibiotika nicht wirksam sind, ist:

$$p(\bar{A} \text{ und } \bar{B} \text{ und } \bar{C}) = p(\bar{A}) \cdot p(\bar{B}) \cdot p(\bar{C}) = 0,3 \cdot 0,4 \cdot 0,5 = 0,06$$

Daraus folgt: Die Wahrscheinlichkeit für die Wirksamkeit der Kombinationstherapie berechnet sich zu

$$p = 1 - 0,06 = 0,94$$

Beispiel:

Die Wahrscheinlichkeit, bei einem Patienten mit Bronchialkarzinom Karzinomzellen im Sputum nachzuweisen, sei $p(K) = 0,8$.

Wie erhöht sich die diagnostische Wahrscheinlichkeit, daß Karzinomzellen gefunden werden, bei zweimaliger Untersuchung ?

Die Wahrscheinlichkeit, daß bei einmaliger Untersuchung keine Karzinomzellen gefunden werden, beträgt:

$$p(\bar{K}) = 1 - p(K) = 0,2$$

Die Wahrscheinlichkeit, daß bei zweimaliger Untersuchung keine Karzinomzellen gefunden werden, beträgt:

$$p(\bar{K}_1 \text{ und } \bar{K}_2) = 0,2 \cdot 0,2 = 0,04$$

Daraus folgt für die Wahrscheinlichkeit eines Nachweises von Karzinomzellen bei zweimaliger Untersuchung: $p = 1 - 0,04 = 0,96$

Bedingte Wahrscheinlichkeit

In gewissen Situationen ist es nicht sinnvoll, Wahrscheinlichkeiten anzugeben, die sich auf die Grundgesamtheit beziehen. Viele Krankheiten stehen im Zusammenhang mit Eigenschaften des Patienten (z.B. Genotyp) oder mit anderen Krankheiten oder mit Risikofaktoren. In diesen Fällen erscheint es sinnvoll, die Wahrscheinlichkeit für bestimmte Teilmengen anzugeben.

Wir wissen z.B., daß die Ereignisse Arteriosklerose (A) und Diabetes (D) nicht unabhängig sind. Neben einer allgemeinen Wahrscheinlichkeit für Arteriosklerose $p(A)$, kann man die Wahrscheinlichkeit für Arteriosklerose bei Diabetikern angeben (sie wird höher sein). Man spricht dann von einer bedingten Wahrscheinlichkeit (Bedingung ‚Diabetes‘).

Man bezeichnet diese als $p(A | D)$ (sprich Wahrscheinlichkeit für A unter der Bedingung D). Diese bedingte Wahrscheinlichkeit ist folgendermaßen definiert:

$$p(A / D) = \frac{p(A \cap D)}{p(D)}$$

Durch einfaches Umschreiben erhält man den Multiplikationssatz für abhängige Ereignisse (A und D):

$$p(A \cap D) = p(A / D) \cdot p(D)$$

Selbstverständlich kann die Bedingung vertauscht werden und wir erhalten:

$$p(A \cap D) = p(A) \cdot p(D / A)$$

Die Umkehrung der Bedingungen kann auch über das sog. Bayes-Theorem errechnet werden:

$$p(A / D) = \frac{p(A) \cdot p(D / A)}{p(A) \cdot p(D / A) + p(\bar{A}) \cdot p(D / \bar{A})}$$

Beispiel: Innerhalb einer Studie werden folgende Häufigkeiten bestimmt:

		Diabetes		∑
		Ja (D)	Nein (\bar{D})	
Arteriosklerose (A)	Ja	70	10	80
	(\bar{A}) Nein	30	90	120
∑		100	100	200

Von 200 Patienten sind 100 Diabetiker und 100 Nichtdiabetiker, d.h. $p(D) = 0,5$ (Wahrscheinlichkeit oder Prävalenz von Diabetes).

Von 200 Patienten haben 80 eine Arteriosklerose, d.h. $p(A) = 0,4$ (Wahrscheinlichkeit oder Prävalenz von Arteriosklerose).

Die Wahrscheinlichkeit für einen Diabetiker, eine Arteriosklerose zu haben, ist $70/100 = 0,7 = p(A/D)$ (bedingte Wahrscheinlichkeit)

Die Wahrscheinlichkeit für einen Nichtdiabetiker, eine Arteriosklerose zu haben, ist $10/100 = 0,1 = p(A / \bar{D})$ (bedingte Wahrscheinlichkeit)

Die Wahrscheinlichkeit für einen Arteriosklerotiker, einen Diabetes zu haben, ist $70/80 = 0,875 = p(D/A)$ (bedingte Wahrscheinlichkeit)

Die Wahrscheinlichkeit für einen Nichtarteriosklerotiker, einen Diabetes zu haben, ist $30/120 = 0,25 = p(D / \bar{A})$ (bedingte Wahrscheinlichkeit)

Diagnostische Tests

Von einem guten diagnostischen Test erwartet man:

- ein positives Ergebnis bei einem Erkrankten (\rightarrow Sensitivität)
- ein negatives Ergebnis bei einem Nichterkrankten (\rightarrow Spezifität)

Begriffe wie **Sensitivität** und **Spezifität** eines diagnostischen Testverfahrens sind ebenfalls bedingte Wahrscheinlichkeiten.

Betrachten wir einen Schwangerschaftstest mit folgenden Ergebnissen:

		Schwangerschaft		η
		S (ja)	\bar{S} (nein)	
Test	T_+	90	5	95
	T_-	10	95	105
η		100	100	200

Bei 100 Schwangeren ist der Test 90 mal positiv, d.h. $p(T_+ / S) = 90/100 = 0,9$

Diese Wahrscheinlichkeit ist eine bedingte Wahrscheinlichkeit; es ist die Wahrscheinlichkeit, daß der Test bei Vorliegen einer Schwangerschaft positiv wird. Man nennt diese Wahrscheinlichkeit auch **Sensitivität** des Schwangerschaftstests.

Man kann auch die bedingte Wahrscheinlichkeit $p(T_- / \bar{S}) = 95/100 = 0,95$ berechnen.

Es ist die Wahrscheinlichkeit, daß der Test negativ wird, wenn die Schwangerschaft nicht vorliegt. Diese Wahrscheinlichkeit nennt man **Spezifität** des Schwangerschaftstests.

Beide Wahrscheinlichkeiten Sensitivität und Spezifität geben die Qualität eines diagnostischen Testverfahrens an.

Vorhersagewerte (prädiktive Werte)

Für eine evtl. schwangere Frau ist es jedoch uninteressant, wie oft ein Test bei Schwangeren positiv wird (Sensitivität) oder wie oft ein Test bei Nichtschwangeren negativ wird (Spezifität).

Interessanter sind dann folgende Fragen:

- Mit welcher Wahrscheinlichkeit liegt tatsächlich eine Schwangerschaft vor, wenn der Test positiv ausfällt, also wie groß ist $p(S / T_+)$.
Diese bedingte Wahrscheinlichkeit heißt **positiv prädiktiver Wert**.
Für die obige Untersuchung sind 90 von 95 positiven Tests richtig positiv,
d.h. $p(S / T_+) = 90/95 = 0,947$.
- Mit welcher Wahrscheinlichkeit liegt keine Schwangerschaft vor, wenn der Test negativ ausfällt, also wie groß ist $p(\bar{S} / T_-)$.
Diese bedingte Wahrscheinlichkeit heißt **negativ prädiktiver Wert**.
Bei der obigen Untersuchung sind 95 von 105 negativen Testergebnissen richtig negativ,
d.h. $p(\bar{S} / T_-) = 95/105 = 0,905$.

! Die prädiktiven Werte sind von der Prävalenz der untersuchten Populationen abhängig !

Verteilungen

Es werden einige theoretische Verteilungen behandelt, die für die Biowissenschaften von Bedeutung sind. Die Verteilung von (medizinischen) Merkmalen, die in einer Stichprobe erhoben werden, lassen sich oft über solche theoretischen Verteilungen approximieren und in ihren wesentlichen Eigenschaften beschreiben.

Binomialverteilung

Dies ist eine diskrete Wahrscheinlichkeitsverteilung, die auf einem einfachen Zufallsexperiment beruht, bei dem nur zwei Ausgänge möglich sind. Nach Jakob Bernoulli (1654-1705) bezeichnet man dies auch als Bernoulli-Experiment.

Beispiele für Bernoulli-Experimente (2 Möglichkeiten):

- Münzwurfexperiment (Kopf / Zahl)
- Geschlechtsbestimmung (männlich / weiblich)
- Laborwert (normal / pathologisch)
- Histologie (maligne / benigne)
- Diabetes (ja / nein)

Beispiel: Eine Familie mit vier Kindern *Voraussetzung: $p(\text{Junge}) = p(\text{Mädchen}) = 0.5$*

Frage: Wie groß sind die Wahrscheinlichkeiten für die Geschlechtskonstellationen ?

Für eine Familie mit vier Kindern gibt es folgende fünf Möglichkeiten:

4J, kein M 3J, 1M 2J, 2M 1J, 3M kein J, 4M

Die Wahrscheinlichkeiten für diese Ereignisse sind:

$$p(JJJJ) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 1 \cdot \frac{1}{2^4}$$

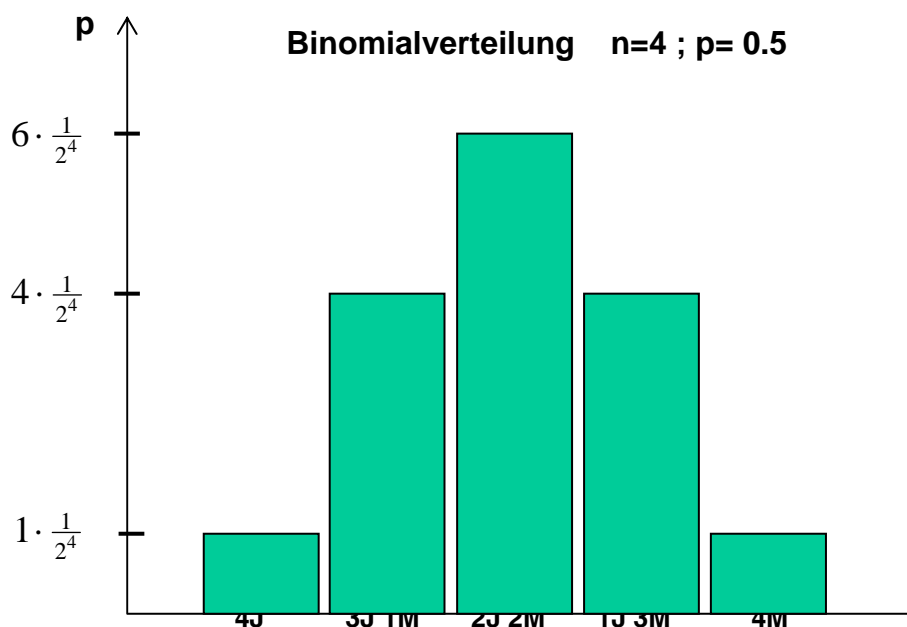
$$p(JJJM) + p(JJMJ) + p(JMJJ) + p(MJJJ) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 4 \cdot \frac{1}{2^4}$$

$$p(JJMM) + p(JMJM) + p(MJJM) + p(MJMJ) + p(JMMM) + p(MMJJ) = 6 \cdot \frac{1}{2^4}$$

$$p(JMMM) + p(MJMM) + p(MMJM) + p(MMMJ) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 4 \cdot \frac{1}{2^4}$$

$$p(MMMM) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 1 \cdot \frac{1}{2^4}$$

Überträgt man die Ereignisse (x-Achse) und die Wahrscheinlichkeiten (y-Achse) in ein x-y Diagramm, so erhält man das Bild einer Binomialverteilung für $n=4$. Diese ist symmetrisch wegen der Gleichheit der Wahrscheinlichkeiten für beide Geschlechter (s. Vorauss.).



Allgemein gilt

Führt man n unabhängige Versuche durch, bei denen ein Ereignis A jeweils mit einer Wahrscheinlichkeit p zu erwarten ist, dann ist die Wahrscheinlichkeit dafür, daß in der Versuchsserie das Ereignis A genau k -mal auftritt :

$$P(k|p,n) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \quad \text{heißt Binomial-Koeffizient}$$

wobei $n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \dots 3 \cdot 2 \cdot 1$

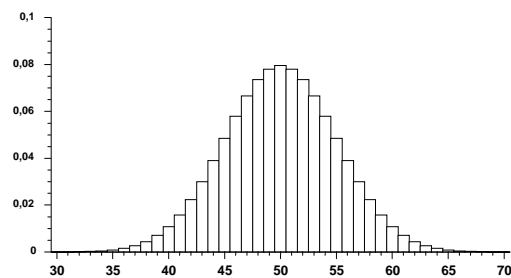
Die Binomial-Verteilung $B(k;p,n)$ beschreibt, wie sich die so berechneten Wahrscheinlichkeiten auf alle möglichen k verteilen.

Voraussetzungen für die Anwendung:

- 1) Es liegt eine **dichotome** Einteilung von möglichen Ereignissen vor (A, \bar{A})
- 2) Für alle n Versuche ist $P(A)$ konstant ($=p$)
- 3) Die Anzahl der Versuche muß im voraus festliegen

Vor037

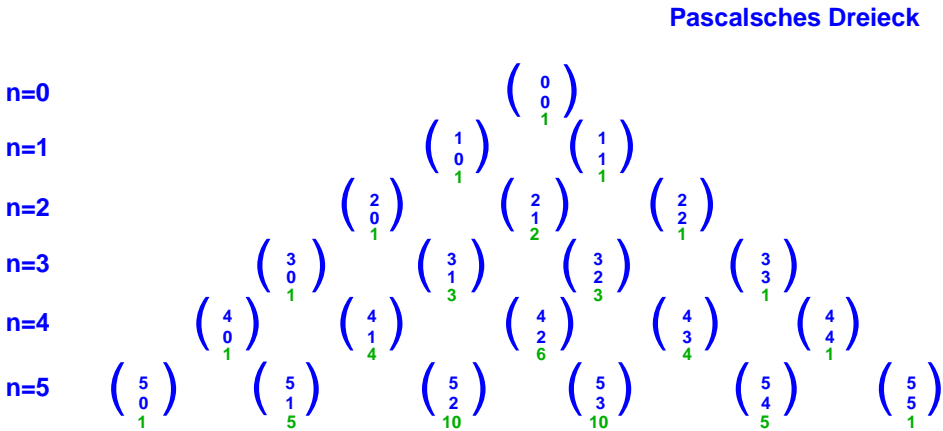
Binomialverteilung ($n=100, p=0.5$)



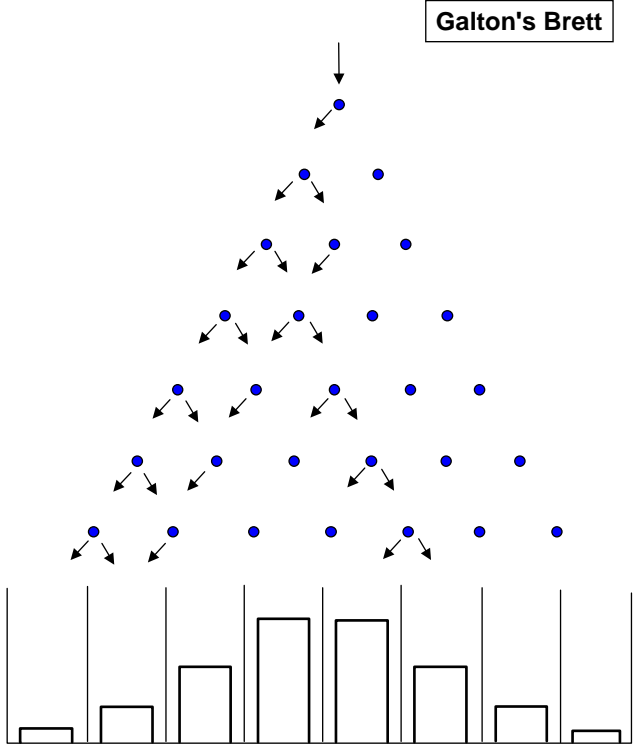
Prüfung einer Münze auf Symmetrie: Kopf / Zahl
Prüfung einer blutdrucksenkenden Wirkung: ja / nein

Kopf	100	90	80	70	65	60	55	50	...
Zahl	0	10	20	30	35	40	45	50	...

Die Binomialkoeffizienten erhält man auch über das Pascalsche Dreieck



Vor038



Vor040

Normalverteilung

Die Normalverteilung ist für die Statistik und deren praktische Anwendung von grundlegender Bedeutung. Während die Binomialverteilung zu den diskreten Wahrscheinlichkeitsverteilungen zählt (die Wahrscheinlichkeiten für die einzelnen Ereignisse kann man explizit ablesen), so ist die Normalverteilung eine Wahrscheinlichkeitsverteilung eines stetigen Merkmals, das theoretisch unendlich viele Merkmalsausprägungen zuläßt. Die Wahrscheinlichkeitsfunktion wird in diesem Fall als Dichte (oder Dichtefunktion) bezeichnet.

Die Dichte der Normalverteilung wird durch die bekannte Gauss'sche Glockenkurve dargestellt (s.a. 10-Mark-Schein). Diese Glockenkurve hat natürlich eine zugrunde liegende mathematische Funktion (s.u.). Deren Funktionswerte geben jedoch zunächst keinen Aufschluß über Wahrscheinlichkeiten von Ereignissen oder Merkmalsausprägungen.

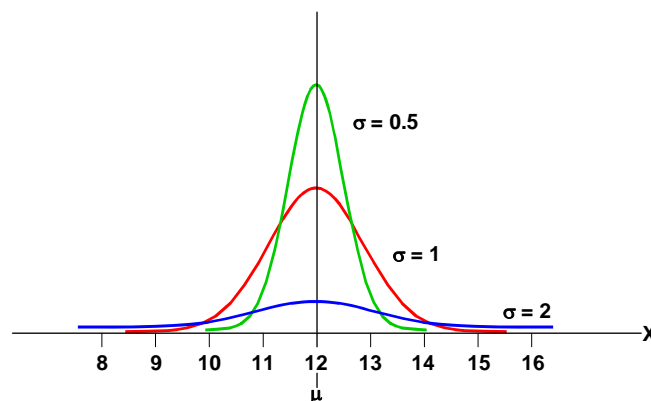
$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wahrscheinlichkeiten sind lediglich als Flächen unter der Dichtekurve zwischen zwei Meßwerten a,b (= Integral der obigen Formel von a nach b) bestimmbar.

Eine normalverteilte Größe ist durch den Erwartungswert μ und die Standardabweichung σ eindeutig charakterisiert. Sie wird deshalb allgemein als $N(\mu, \sigma^2)$ gegeben.

Eigenschaften der Normalverteilung:

- Symmetrie um den Erwartungswert μ
- 2 Wendepunkte bei $\mu - \sigma$ und $\mu + \sigma$
- Maximum bei μ
- Erwartungswert μ und Median stimmen überein
- Für $x \rightarrow \pm\infty$ nähert sie sich asymptotisch der x-Achse
- zwischen $\mu - \sigma$ und $\mu + \sigma$ liegen 68.3% (1 σ -Bereich)
- zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$ liegen 95.4% (2 σ -Bereich)
- zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$ liegen 99.7% (3 σ -Bereich)
- zwischen $\mu - 1.96\sigma$ und $\mu + 1.96\sigma$ liegen 95% (s. Konfidenzintervalle)



Normalverteilung $N(\mu, \sigma^2)$

Mittelwert μ
 Varianz σ^2
 Standardabweichung σ

Standardnormalverteilung

Durch die Transformation $z = \frac{x - \mu}{\sigma}$

geht jede Normalverteilung $N(\mu, \sigma^2)$ in eine sog. Standardnormalverteilung $N(0,1)$ über. Durch diese Transformation wird die Glockenkurve auf der x-Achse so verschoben, daß der Erwartungswert 0 wird. Außerdem wird die Kurve durch die Division durch σ in ihrer Form so angepaßt, daß die Standardabweichung den Wert 1 erhält (Normierung).

Die Standardnormalverteilung ist heute ausreichend tabelliert (s. *Tab. Normalverteilung*), so daß die Wahrscheinlichkeitsberechnungen für unterschiedliche Merkmale mit Hilfe dieser Tabellen und der dazugehörigen Transformation leicht durchgeführt werden können.

In der Medizin gibt es zwar einige relevante Merkmale, die angenähert normalverteilt sind (z.B. Körpergröße erwachsener Männer). Andere wichtige Verteilungen in der Medizin sind jedoch vollkommen anders geartet (z.B. Lebensdauer, Aufenthaltsdauer...) Schiefe Verteilungen lassen sich evtl. in eine Normalverteilung transformieren. (Logarithmierung bei rechtsschiefen Verteilungen; sog. Log-Normalverteilungen)

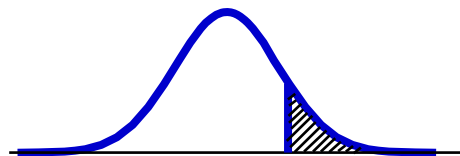


Tabelle Standardnormalverteilung N(0,1)

*	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0*	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
0,1*	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
0,2*	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
0,3*	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
0,4*	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
0,5*	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
0,6*	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
0,7*	0,242	0,239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
0,8*	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
0,9*	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
1,0*	0,159	0,156	0,154	0,152	0,149	0,147	0,145	0,142	0,140	0,138
1,1*	0,136	0,133	0,131	0,129	0,127	0,125	0,123	0,121	0,119	0,117
1,2*	0,115	0,113	0,111	0,109	0,107	0,106	0,104	0,102	0,100	0,099
1,3*	0,097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0,084	0,082
1,4*	0,081	0,079	0,078	0,076	0,075	0,074	0,072	0,071	0,069	0,068
1,5*	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
1,6*	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0,047	0,046	0,046
1,7*	0,045	0,044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
1,8*	0,036	0,035	0,034	0,034	0,033	0,032	0,031	0,031	0,030	0,029
1,9*	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
2,0*	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0,019	0,018
2,1*	0,018	0,017	0,017	0,017	0,016	0,016	0,015	0,015	0,015	0,014
2,2*	0,014	0,014	0,013	0,013	0,013	0,012	0,012	0,012	0,011	0,011
2,3*	0,011	0,010	0,010	0,010	0,010	0,009	0,009	0,009	0,009	0,008
2,4*	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,007	0,007	0,006
2,5*	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
2,6*	0,005	0,005	0,004	0,004	0,004	0,004	0,004	0,004	0,004	0,004
2,7*	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
2,8*	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
2,9*	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	0,001	0,001

Zentraler Grenzwertsatz

Werden aus einer beliebig verteilten Grundgesamtheit (mit Mittelwert μ und Varianz σ^2) zufällige Stichproben vom Umfang n gezogen, dann sind die Mittelwerte der Stichproben normalverteilt. Diese Normalverteilung hat denselben Erwartungswert μ wie die Grundgesamtheit. Die Varianz der Stichprobenmittel ist σ^2/n und ist verständlicherweise kleiner als die der Einzelwerte.

Daraus ergibt sich für die Standardabweichung der Mittelwerte (man spricht auch vom Standardfehler der Mittelwerte oder ‚standard error of mean‘ oder SEM) :

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Konfidenzintervall

a) Konfidenzintervall für den Erwartungswert

Nehmen wir an, wir hätten die Körpergröße von 39 männlichen Medizinstudenten gemessen und eine mittlere Körpergröße von $\bar{x}_m = 182.5$ cm (Standardabweichung der Stichprobe $s_m = 6.5$ cm) ermittelt. Wenn diese Gruppe als repräsentative Stichprobe für männlichen Medizinstudenten (in Deutschland) auffassen, dann handelt es sich bei diesem Mittelwert um eine Schätzung für den Erwartungswert der Grundgesamtheit. Man weiß, dieser Mittelwert ist zufallsbedingt – eine andere Stichprobe von $n=39$ würde andere Daten und einen anderen Mittelwert liefern.

Folgende Frage erscheint berechtigt:

Welcher Erwartungswert μ könnte den besagten Mittelwert von 182,5 erzeugt haben?

Es erscheint durchaus möglich, daß er aus einer Grundgesamtheit mit $\mu = 180$ cm oder auch mit $\mu = 185$ cm resultiert. Wir würden jedoch nicht annehmen, daß der Stichprobe eine Grundgesamtheit mit $\mu = 160$ cm zugrunde liegt – obwohl auch diese Möglichkeit nicht ganz ausgeschlossen werden kann. Um Anhaltspunkte bezüglich der Genauigkeit der Schätzung zu gewinnen, konstruiert man aus den Daten der Stichprobe ein sog Konfidenzintervall (bzw. Vertrauensintervall). Man hofft, ein Intervall zu erhalten, das den gesuchten Erwartungswert enthält. Es ist allerdings nicht auszuschließen, daß die Daten der Stichprobe ein Konfidenzintervall ergeben, das ‚daneben liegt‘ und den gesuchten Parameter μ nicht enthält. Die Wahrscheinlichkeit für das ‚Danebenliegen‘ sollte normalerweise $\alpha = 5\%$ sein. Für $\alpha=5\%$ erhält man 95%-Konfidenzintervalle (für $\alpha=1\%$ erhält man 99%-Konfidenzintervalle).

Das 95%-Konfidenzintervall für den Erwartungswert ist gegeben durch:

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Bei einem 99%-Konfidenzintervall ist der Wert 1.96 durch 2.58 zu ersetzen.

Beispiel (Körpergröße): (Teil 1)

Mit $\bar{x}_m = 182.5$ cm, $n=39$ (Stichprobe) und einer Standardabweichung der Grundgesamtheit $\sigma = 6.53$ cm

erhalten wir als Grenzen des 95%-Konfidenzintervalls: $182.5 \pm 1.96 \cdot \frac{6.53}{\sqrt{39}}$ oder ≈ 180.5 cm ; 184.5 cm

Bei der obigen Formel haben wir stillschweigend vorausgesetzt, daß die Standardabweichung der Grundgesamtheit σ bekannt ist. Dies ist aber bei (medizinischen) Untersuchungen fast niemals der Fall. Man könnte notgedrungen das σ durch die Standardabweichung der Stichprobe (s) ersetzen. Dies würde aber insbesondere bei kleinen Stichproben (eher die Regel als die Ausnahme) zu weiteren Ungenauigkeiten der Schätzung führen. Deswegen nimmt man nicht die Werte der Normalverteilung (1.96 oder 2.58 ...) für die Schätzung. Man benutzt eine für kleine Stichproben immer breiter werdende Verteilung (damit man auf der sicheren Seite bleibt), die auf William Gosset zurückgeht und Student-t-Verteilung heißt. (s.a. t.Tabelle)

Man erhält dadurch folgendes 95%-Konfidenzintervall für den Erwartungswert:

$$\bar{x} - \frac{t_{n-1;0.975} \cdot s}{\sqrt{n}}; \bar{x} + \frac{t_{n-1;0.975} \cdot s}{\sqrt{n}} \quad \text{oder} \quad \bar{x} - t_{n-1;0.975} \cdot SEM; \bar{x} + t_{n-1;0.975} \cdot SEM$$

Dabei ist $t_{n-1;0.975}$ das 0.975-Quantil der Student-t-Verteilung mit $n-1$ Freiheitsgraden (s.Tabelle t-Verteilung).

Beispiel (Körpergröße): (Teil 2)

Mit $\bar{x}_m = 182.5 \text{ cm}$, $n=39$ (Stichprobe) und einer Standardabweichung der Stichprobe $s = 6.5 \text{ cm}$ ergibt sich in der t-Tabelle ein Wert $t_{n-1;0.975} = t_{38;0.975} = 2.02$ und somit für das Konfidenzintervall:

$$182.5 \pm 2.02 \cdot \frac{6.5}{\sqrt{39}} \quad \text{oder} \quad \checkmark 180.4 \text{ cm} ; 184.6 \text{ cm}$$

Es zeigt sich daß die Schätzungen über die t-Verteilung geringfügig breiter sind (vgl. Beispiel Teil1 oben). Bei kleineren Stichprobenumfängen kommt dies noch wesentlich stärker zum Ausdruck.

t-Tabelle

Freiheitsgrad	<i>p</i>								
	0,5	0,2	0,1	0,05	0,02	0,025	0,01	0,001	0,0001
1	1,000	3,078	6,314	12,706	31,821	25,452	63,656	636,578	6370,544
2	0,817	1,886	2,920	4,303	6,965	6,2053	9,925	31,600	100,136
3	0,765	1,638	2,353	3,182	4,541	4,1766	5,841	12,924	28,014
4	0,741	1,533	2,132	2,776	3,747	3,4954	4,604	8,610	15,534
5	0,727	1,476	2,015	2,571	3,365	3,1634	4,032	6,869	11,176
6	0,718	1,440	1,943	2,447	3,143	2,9687	3,707	5,959	9,080
7	0,711	1,415	1,895	2,365	2,998	2,8412	3,499	5,408	7,888
8	0,706	1,397	1,860	2,306	2,896	2,7515	3,355	5,041	7,120
9	0,703	1,383	1,833	2,262	2,821	2,6850	3,250	4,781	6,594
10	0,700	1,372	1,812	2,228	2,764	2,6338	3,169	4,587	6,212
11	0,697	1,363	1,796	2,201	2,718	2,5931	3,106	4,437	5,923
12	0,695	1,356	1,782	2,179	2,681	2,5600	3,055	4,318	5,695
13	0,694	1,350	1,771	2,160	2,650	2,5326	3,012	4,221	5,513
14	0,692	1,345	1,761	2,145	2,624	2,5096	2,977	4,140	5,364
15	0,691	1,341	1,753	2,131	2,602	2,4899	2,947	4,073	5,239
16	0,690	1,337	1,746	2,120	2,583	2,4729	2,921	4,015	5,134
17	0,689	1,333	1,740	2,110	2,567	2,4581	2,898	3,965	5,043
18	0,688	1,330	1,734	2,101	2,552	2,4450	2,878	3,922	4,966
19	0,688	1,328	1,729	2,093	2,539	2,4334	2,861	3,883	4,899
20	0,687	1,325	1,725	2,086	2,528	2,4231	2,845	3,850	4,838
21	0,686	1,323	1,721	2,080	2,518	2,4138	2,831	3,819	4,785
22	0,686	1,321	1,717	2,074	2,508	2,4055	2,819	3,792	4,736
23	0,685	1,319	1,714	2,069	2,500	2,3979	2,807	3,768	4,694
24	0,685	1,318	1,711	2,064	2,492	2,3910	2,797	3,745	4,654
25	0,684	1,316	1,708	2,060	2,485	2,3846	2,787	3,725	4,619
26	0,684	1,315	1,706	2,056	2,479	2,3788	2,779	3,707	4,587
27	0,684	1,314	1,703	2,052	2,473	2,3734	2,771	3,689	4,557
28	0,683	1,313	1,701	2,048	2,467	2,3685	2,763	3,674	4,531
29	0,683	1,311	1,699	2,045	2,462	2,3639	2,756	3,660	4,505
30	0,683	1,310	1,697	2,042	2,457	2,3596	2,750	3,646	4,482
31	0,682	1,309	1,696	2,040	2,453	2,3556	2,744	3,633	4,461
32	0,682	1,309	1,694	2,037	2,449	2,3518	2,738	3,622	4,440
33	0,682	1,308	1,692	2,035	2,445	2,3483	2,733	3,611	4,421
34	0,682	1,307	1,691	2,032	2,441	2,3451	2,728	3,601	4,405
35	0,682	1,306	1,690	2,030	2,438	2,3420	2,724	3,591	4,389
36	0,681	1,306	1,688	2,028	2,435	2,3391	2,719	3,582	4,373
37	0,681	1,305	1,687	2,026	2,431	2,3363	2,715	3,574	4,359
38	0,681	1,304	1,686	2,024	2,429	2,3337	2,712	3,566	4,345
39	0,681	1,304	1,685	2,023	2,426	2,3313	2,708	3,558	4,333
40	0,681	1,303	1,684	2,021	2,423	2,3289	2,704	3,551	4,321
41	0,681	1,303	1,683	2,020	2,421	2,3267	2,701	3,544	4,310
42	0,680	1,302	1,682	2,018	2,418	2,3246	2,698	3,538	4,298
43	0,680	1,302	1,681	2,017	2,416	2,3226	2,695	3,532	4,289
44	0,680	1,301	1,680	2,015	2,414	2,3207	2,692	3,526	4,278
45	0,680	1,301	1,679	2,014	2,412	2,3189	2,690	3,520	4,269
46	0,680	1,300	1,679	2,013	2,410	2,3172	2,687	3,515	4,261
47	0,680	1,300	1,678	2,012	2,408	2,3155	2,685	3,510	4,251
48	0,680	1,299	1,677	2,011	2,407	2,3139	2,682	3,505	4,243
49	0,680	1,299	1,677	2,010	2,405	2,3124	2,680	3,500	4,235
50	0,679	1,299	1,676	2,009	2,403	2,3109	2,678	3,496	4,228
55	0,679	1,297	1,673	2,004	2,396	2,3045	2,668	3,476	4,196
60	0,679	1,296	1,671	2,000	2,390	2,2991	2,660	3,460	4,169
70	0,678	1,294	1,667	1,994	2,381	2,2907	2,648	3,435	4,127
80	0,678	1,292	1,664	1,990	2,374	2,2844	2,639	3,416	4,095
90	0,677	1,291	1,662	1,987	2,369	2,2795	2,632	3,402	4,072
100	0,677	1,290	1,660	1,984	2,364	2,2757	2,626	3,390	4,054
500	0,675	1,283	1,648	1,965	2,334	2,2482	2,586	3,310	3,922
1000	0,675	1,282	1,646	1,962	2,330	2,2448	2,581	3,300	3,907
10000	0,675	1,282	1,645	1,960	2,327	2,2417	2,576	3,292	3,892
∞	0,675	1,282	1,645	1,960	3,326	2,2414	2,576	3,290	3,891

b) Konfidenzintervall der relativen Häufigkeit

95% - Vertrauensbereich einer relativen Häufigkeit p in einer Stichprobe vom Umfang $n \geq 100$:

$$95\% \text{ KI} = \left[p - 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}, p + 1.96 \cdot \sqrt{\frac{p(1-p)}{n}} \right]$$

Für $n < 100$ und p nahe 0 oder 1

→ exakte Vertrauensgrenzen s. Tabelle

Beispiel: $p = 0.25$

	25%	95% - Vertrauensbereich
$n = 4$	1/4	[1%, 81%]
$n = 40$	10/40	[13%, 41%]
$n = 80$	20/80	[16%, 36%]
$n = 120$	30/120	[17%, 35%] (approximativ)

Konfidenz003

Umfangreichere Tabellen für $n=20$, $n=40$ und $n=80$ siehe nächste Seite
 Zusätzlich zu den 95%-Konfidenzintervallen sind die 99%-Konfidenzintervalle aufgelistet.

Für vollständige Tabellen siehe „Exakte Vertrauensgrenzen für p (Binomialverteilung)
 in ‚Wissenschaftliche Tabellen Geigy‘ Teil 1 Statistik
 Herausgegeben von Ciba Geigy Basel

Bemerkung:

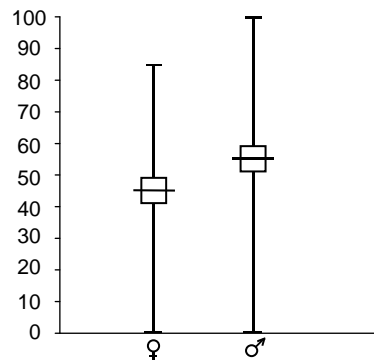
Die in den Tabellen errechneten Konfidenzbereiche entsprechen nur im Zentrum der Binomialverteilung den mit der obigen Formel errechneten Konfidenzintervallen. An den Rändern werden die Intervalle asymmetrisch korrigiert.

x	100 p _x	100(1 - 2α)-Grenzen	
		95% 100 p _l 100 p _r	99% 100 p _l 100 p _r
N = 80			
0	0,00	0,00- 4,51	0,00- 6,41
1	1,25	0,03- 6,77	0,01- 8,92
2	2,50	0,30- 8,74	0,13- 11,08
3	3,75	0,78- 10,57	0,43- 13,05
4	5,00	1,38- 12,31	0,85- 14,92
5	6,25	2,06- 13,99	1,37- 16,70
6	7,50	2,80- 15,61	1,96- 18,42
7	8,75	3,59- 17,20	2,61- 20,09
8	10,00	4,42- 18,76	3,31- 21,72
9	11,25	5,28- 20,28	4,04- 23,31
10	12,50	6,16- 21,79	4,81- 24,87
11	13,75	7,07- 23,27	5,61- 26,41
12	15,00	8,00- 24,74	6,43- 27,92
13	16,25	8,95- 26,18	7,28- 29,41
14	17,50	9,91- 27,62	8,14- 30,88
15	18,75	10,89- 29,03	9,03- 32,33
16	20,00	11,89- 30,44	9,94- 33,77
17	21,25	12,89- 31,83	10,86- 35,19
18	22,50	13,91- 33,21	11,80- 36,59
19	23,75	14,95- 34,58	12,75- 37,98
20	25,00	15,99- 35,94	13,72- 39,35
21	26,25	17,04- 37,29	14,70- 40,72
22	27,50	18,10- 38,62	15,70- 42,07
23	28,75	19,18- 39,95	16,70- 43,41
24	30,00	20,26- 41,28	17,72- 44,73
25	31,25	21,35- 42,59	18,75- 46,05
26	32,50	22,45- 43,89	19,79- 47,36
27	33,75	23,55- 45,19	20,84- 48,65
28	35,00	24,67- 46,48	21,90- 49,94
29	36,25	25,79- 47,76	22,97- 51,21
30	37,50	26,92- 49,04	24,05- 52,48
31	38,75	28,06- 50,30	25,14- 53,74
32	40,00	29,20- 51,56	26,24- 54,99
33	41,25	30,35- 52,82	27,35- 56,22
34	42,50	31,51- 54,06	28,46- 57,45
35	43,75	32,68- 55,30	29,59- 58,68
36	45,00	33,85- 56,53	30,72- 59,89
37	46,25	35,03- 57,76	31,87- 61,09
38	47,50	36,21- 58,98	33,02- 62,29
39	48,75	37,41- 60,19	34,18- 63,47
40	50,00	38,60- 61,40	35,35- 64,65
41	51,25	39,81- 62,59	36,53- 65,82
42	52,50	41,02- 63,79	37,71- 66,98
43	53,75	42,24- 64,97	38,91- 68,13
44	55,00	43,47- 66,15	40,11- 69,28
45	56,25	44,70- 67,32	41,32- 70,41
46	57,50	45,94- 68,49	42,55- 71,54
47	58,75	47,18- 69,65	43,78- 72,65
48	60,00	48,44- 70,80	45,01- 73,76
49	61,25	49,70- 71,94	46,26- 74,86
50	62,50	50,96- 73,08	47,52- 75,95
51	63,75	52,24- 74,21	48,79- 77,03
52	65,00	53,52- 75,33	50,06- 78,10
53	66,25	54,81- 76,45	51,35- 79,16
54	67,50	56,11- 77,55	52,64- 80,21
55	68,75	57,41- 78,65	53,95- 81,25
56	70,00	58,72- 79,74	55,27- 82,28
57	71,25	60,05- 80,82	56,59- 83,30
58	72,50	61,38- 81,90	57,93- 84,30
59	73,75	62,71- 82,96	59,28- 85,30
60	75,00	64,06- 84,01	60,65- 86,28
61	76,25	65,42- 85,05	62,02- 87,25
62	77,50	66,79- 86,09	63,41- 88,20
63	78,75	68,17- 87,11	64,81- 89,14
64	80,00	69,56- 88,11	66,23- 90,06
65	81,25	70,97- 89,11	67,67- 90,97
66	82,50	72,38- 90,09	69,12- 91,86
67	83,75	73,82- 91,05	70,59- 92,72
68	85,00	75,26- 92,00	72,08- 93,57
69	86,25	76,73- 92,93	73,59- 94,39
70	87,50	78,21- 93,84	75,13- 95,19
71	88,75	79,72- 94,72	76,69- 95,96
72	90,00	81,24- 95,58	78,28- 96,69
73	91,25	82,80- 96,41	79,91- 97,39
74	92,50	84,39- 97,20	81,58- 98,04
75	93,75	86,01- 97,94	83,30- 98,63
76	95,00	87,69- 98,62	85,08- 99,15
77	96,25	89,43- 99,22	86,95- 99,57
78	97,50	91,26- 99,70	88,92- 99,87
79	98,75	93,23- 99,97	91,08- 99,99
80	100,00	95,49-100,00	93,59-100,00

x	100 p _x	100(1 - 2α)-Grenzen	
		95% 100 p _l 100 p _r	99% 100 p _l 100 p _r
N = 40			
0	0,00	0,00- 8,81	0,00- 12,41
1	2,50	0,06- 13,16	0,01- 17,15
2	5,00	0,61- 16,92	0,26- 21,18
3	7,50	1,57- 20,39	0,86- 24,84
4	10,00	2,79- 23,66	1,73- 28,26
5	12,50	4,19- 26,80	2,80- 31,51
6	15,00	5,71- 29,84	4,02- 34,63
7	17,50	7,34- 32,78	5,37- 37,63
8	20,00	9,05- 35,65	6,82- 40,54
9	22,50	10,84- 38,45	8,36- 43,37
10	25,00	12,69- 41,20	9,98- 46,12
11	27,50	14,60- 43,89	11,68- 48,81
12	30,00	16,56- 46,53	13,44- 51,43
13	32,50	18,57- 49,13	15,26- 54,00
14	35,00	20,63- 51,68	17,13- 56,51
15	37,50	22,73- 54,20	19,06- 58,97
16	40,00	24,86- 56,67	21,05- 61,38
17	42,50	27,04- 59,11	23,08- 63,74
18	45,00	29,26- 61,51	25,16- 66,05
19	47,50	31,51- 63,87	27,29- 68,32
20	50,00	33,80- 66,20	29,46- 70,54
21	52,50	36,13- 68,49	31,68- 72,71
22	55,00	38,49- 70,74	33,95- 74,84
23	57,50	40,89- 72,96	36,26- 76,92
24	60,00	43,33- 75,14	38,62- 78,95
25	62,50	45,80- 77,27	41,03- 80,94
26	65,00	48,32- 79,37	43,49- 82,87
27	67,50	50,87- 81,43	46,00- 84,74
28	70,00	53,47- 83,44	48,57- 86,56
29	72,50	56,11- 85,40	51,19- 88,32
30	75,00	58,80- 87,31	53,88- 90,02
31	77,50	61,55- 89,16	56,63- 91,64
32	80,00	64,35- 90,95	59,46- 93,18
33	82,50	67,22- 92,66	62,37- 94,63
34	85,00	70,16- 94,29	65,37- 95,98
35	87,50	73,20- 95,81	68,49- 97,20
36	90,00	76,34- 97,21	71,74- 98,27
37	92,50	79,61- 98,43	75,16- 99,14
38	95,00	83,08- 99,39	78,82- 99,74
39	97,50	86,84- 99,94	82,85- 99,99
40	100,00	91,19-100,00	87,59-100,00

x	100 p _x	100(1 - 2α)-Grenzen	
		95% 100 p _l 100 p _r	99% 100 p _l 100 p _r
N = 20			
0	0,00	0,00- 16,84	0,00- 23,27
1	5,00	0,13- 24,87	0,03- 31,71
2	10,00	1,23- 31,70	0,53- 38,71
3	15,00	3,21- 37,89	1,76- 44,95
4	20,00	5,73- 43,66	3,58- 50,66
5	25,00	8,66- 49,10	5,83- 55,98
6	30,00	11,89- 54,28	8,46- 60,96
7	35,00	15,39- 59,22	11,39- 65,66
8	40,00	19,12- 63,95	14,60- 70,09
9	45,00	23,06- 68,47	18,06- 74,28
10	50,00	27,20- 72,80	21,77- 78,23
11	55,00	31,53- 76,94	25,72- 81,94
12	60,00	36,05- 80,88	29,91- 85,40
13	65,00	40,78- 84,61	34,34- 88,61
14	70,00	45,72- 88,11	39,04- 91,54
15	75,00	50,90- 91,34	44,02- 94,17
16	80,00	56,34- 94,27	49,34- 96,42
17	85,00	62,11- 96,79	55,05- 98,24
18	90,00	68,30- 98,77	61,29- 99,47
19	95,00	75,13- 99,87	68,29- 99,97
20	100,00	83,16-100,00	76,73-100,00

**95% - Konfidenzintervalle für den Mittelwert
der selbsteingeschätzten Kondition**

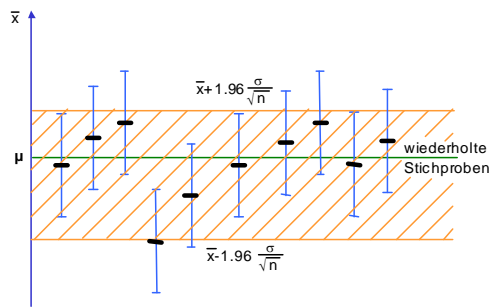


Deskriptive Statistiken zur Kondition

♀ (N = 127): Mittelwert = 45
 Median = 50
 Standardabweichung = 19.4
 Minimum = 0
 Maximum = 85
 95% Konfidenzintervall = [41.6 , 48.4]

♂ (N = 118): Mittelwert = 57
 Median = 60
 Standardabweichung = 21.3
 Minimum = 0
 Maximum = 100
 95% Konfidenzintervall = [53.1 , 60.9]

Konfidenz004



95% - Vertrauensbereiche
 für den Mittelwert μ einer Normalverteilung
 mit bekannter Varianz σ bei einer Serie von
 Zufallsstichproben vom Umfang n

Prinzip des statistischen Tests

Der Fortschritt in der Medizin beruht auf Beobachtungen, die die Ärzte bei der Patientenbehandlung oder im Labor machen. Möglicherweise werden dabei neue wissenschaftliche Erkenntnisse gewonnen oder es wird eine Therapie entwickelt, von der man glaubt, daß sie einer herkömmlichen Standardtherapie in irgendeiner Weise überlegen ist. Aus der Vielzahl von Beobachtungen gepaart mit fachlich-theoretischen Überlegungen entsteht so eine Vermutung – präzise formuliert eine Hypothese.

Die Überprüfung solcher Hypothesen sollte in zweifacher Hinsicht erfolgen:

- Zunächst sollte ein theoretischer Hintergrund erarbeitet werden, um die Hypothese mit sachlichen Argumenten zu untermauern. Dazu bedarf es überwiegend medizinischer Fachkenntnisse und Erfahrung
- Darüber hinaus ist es erforderlich, die Hypothese statistisch zu überprüfen. Dazu müssen relevante Daten aus einer oder mehrerer Stichproben mit einer geeigneten Testmethode analysiert werden.

Zunächst soll das Prinzip des statistischen Tests anhand eines einfachen Beispiels des Münzwurfexperimentes erläutert werden.

Dazu stelle man sich folgende Situation vor:

Ein Schiedsrichter (Mannschaftssport), der für die Seitenwahl eine bestimmte Münze benutzt, hat den Verdacht, diese Münze ist nicht symmetrisch, d.h. ‚Kopf‘ und ‚Zahl‘ haben unterschiedliche Häufigkeiten.

Generell sind dann zwei Möglichkeiten bezüglich der Ausgangssituation denkbar:

- **Die Münze ist symmetrisch** – d.h. man erhält ‚Kopf‘ und ‚Zahl‘ mit derselben Wahrscheinlichkeit von $p=1/2$. In diesem Fall würde man z.B. bei 100 Würfeln 50 mal ‚Kopf‘ und 50 mal ‚Zahl‘ erwarten. Kleinere Abweichungen wird man tolerieren; sie geben keinen Anlaß, an der Symmetrie der Münze zu zweifeln.
- **Die Münze ist nicht symmetrisch** – d.h. man erhält ‚Kopf‘ oder ‚Zahl‘ zu häufig.

Die beiden Aussagen sind komplementär, d.h. genau eine davon muß richtig sein. Ein statistischer Test hilft in solch einer Situation eine Entscheidung zu treffen. Dazu werden zunächst die beiden zu überprüfenden Hypothesen (Nullhypothese H_0 und Alternativhypothese H_A) vor der Durchführung des Tests so präzise wie möglich formuliert.

Üblicherweise wird die konservative Aussage zur Nullhypothese: Die Münze ist symmetrisch
Die sog. innovative Aussage wird zur Alternativhypothese: Die Münze ist nicht symmetrisch

Nach Übersetzung der inhaltlichen Aussagen in statistische Hypothesen erhält man:

$$H_0: p = 1/2 \quad (p \text{ sei die Wahrscheinlichkeit für ‚Kopf‘})$$

$$H_A: p \neq 1/2$$

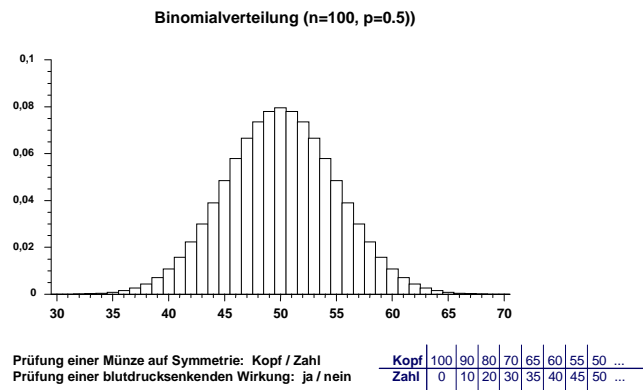
Die Nullhypothese beinhaltet das Gleichheitszeichen; sie ist eindeutig formuliert. Die Alternativhypothese ist dagegen sehr allgemein gehalten: sie vereinigt alle Hypothesen außer der Nullhypothese.

Gehen wir zum Beispiel zurück:

Nehmen wir an, unser Schiedsrichter möchte seine Münze auf Symmetrie überprüfen. Zunächst muß er sich dann Gedanken darüber machen, wie oft er die Münze werfen muß, um zu vernünftigen Aussagen zu gelangen. In der Biometrie nennt man dies ‚Stichprobenumfangschätzung (wieviel Patienten brauche ich?)‘.

Nehmen wir für unser Beispiel $n=100$ Münzwürfe.

Ist die Münze symmetrisch, so ist der Wahrscheinlichkeitsgipfel (z.B. für ‚Kopf‘) bei 50 zu erwarten. Da wir die Binomialverteilung kennen, wissen wir, wie die einzelnen Wahrscheinlichkeiten verteilt sind.. Die Abbildung gibt die Wahrscheinlichkeiten an, die unter der Voraussetzung der Nullhypothese (Symmetrie) mit Hilfe der Binomialverteilung errechnet wurden.



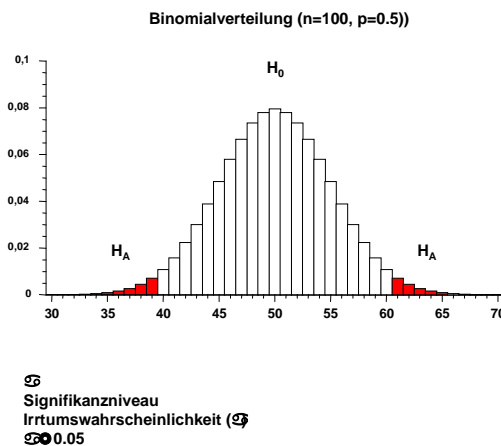
Die Zahlen unterhalb der Abbildung sind mögliche Ergebnisse des Experimentes. Einige dieser Ergebnisse sprechen eher nicht für die Nullhypothese (z.B. 100 mal ‚Kopf‘ und kein mal ‚Zahl‘ oder 90 mal ‚Kopf‘ und 10 mal ‚Zahl‘). Andere wiederum sprechen doch mehr für die Gültigkeit der Nullhypothese (50 mal ‚Kopf‘ und 50 mal ‚Zahl‘ oder evtl. auch 60 mal ‚Kopf‘ und 40 mal ‚Zahl‘).

Der statistische Test sollte mit einer Entscheidung enden:

Liegt das Ergebnis in einem Bereich um ‚50‘ (Annahmebereich), entscheidet man sich für die Nullhypothese.

Liegt das Ergebnis im kritischen Bereich, entscheidet man sich für die Alternativhypothese. Der kritische Bereich sollte wie in der Abbildung 5% der Gesamtfläche nicht überschreiten (α = 0.05).

(Da der Tests mit der Binomialverteilung arbeitet könnte man den Test ‚Binomialtest‘ nennen)



Eine **Entscheidung für die Alternativhypothese** ist ein sog ‚statistisch signifikantes‘ Testergebnis mit $\varphi=0.05$ als Signifikanzgrenze. φ kann jederzeit verkleinert werden, üblich sind dann Signifikanzgrenzen bei 0.01 oder 0.001.

Die Alternativhypothese wird also nur dann akzeptiert, wenn das Ergebnis mit der Nullhypothese nur schwer zu vereinbaren ist.

Theoretisch kann diese Entscheidung falsch sein, wenn trotz Gültigkeit der Nullhypothese ein Ergebnis im kritischen Bereich zustande kommt (trotz symmetrischer Münze erhält man 61 mal ‚Kopf‘ und 39 mal ‚Zahl‘).

Formulierung der Testentscheidung:

Die Nullhypothese wird verworfen, die Münze ist nicht symmetrisch ($p < 0.05$)

d.h. die Wahrscheinlichkeit für eine irrtümlich getroffene Entscheidung ist kleiner als 5%.
oder die Wahrscheinlichkeit, daß das Ergebnis zufällig entstanden ist, ist kleiner als 5%.

p wird auch als φ -Fehler bezeichnet (s.Abb unten)

Eine **Entscheidung für die Nullhypothese** ist ein sog ‚statistisch nicht signifikantes‘ Testergebnis und ist nur dann richtig, wenn die Aussage der Nullhypothese in Wirklichkeit zutrifft. Ansonsten ist man einem τ -Fehler erlegen. Dieser Fehler ist in der Regel im Gegensatz zu φ nicht abschätzbar; er kann insbes. bei kleinen Stichproben sehr groß sein. Eine Prüfgröße im Annahmebereich ist deshalb kein Beweis für die Richtigkeit der Nullhypothese, sondern lediglich ein Hinweis darauf, daß man aufgrund der vorhandenen Daten die Nullhypothese nicht ablehnen kann.

Deshalb die vorsichtige Formulierung:

*die Nullhypothese kann auf dem Signifikanzniveau φ nicht abgelehnt werden
oder es ergibt sich kein Widerspruch zur Nullhypothese*

Hier drückt sich aus, daß es unbefriedigend ist, wenn ein Test zur Nichtablehnung der Nullhypothese führt.

Cave: Die formale statistische Signifikanz (nur dies kann ein statistischer Test nachweisen!) und wirkliche medizinische Relevanz dürfen nicht miteinander verwechselt werden.

Einen Überblick über Entscheidungen und Wahrscheinlichkeiten bei einem statistischen Test gibt das folgende Diagramm:

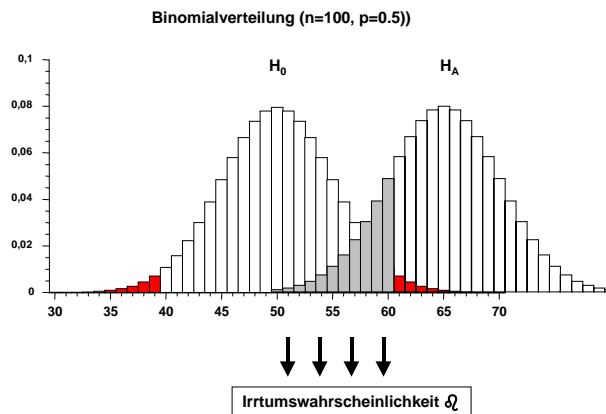
- 2 Stichproben A , B
- Hypothesen $H_0: A = B$ (Nullhypothese)
 $H_A: A \neq B$ (Alternativhypothese)
- Entscheidung für H_0 oder H_A

		Realität	
		H_0	H_A
Testentscheidung	H_0	richtig $1 - \alpha$	Fehler 2.Art falsch negative Entscheidung β
	H_A	Fehler 1.Art falsch positive Entscheidung α oder β	richtig $1 - \beta$ power (Macht des Tests)

$$\alpha \approx \beta^{-1}$$

$$\beta \approx \alpha^{-1} n^{-1}$$

Die Abhängigkeit der Irrtumswahrscheinlichkeiten α und β am Beispiel des Münzwurfexperiments und der Binomialverteilung zeigt die nächste Abbildung:



Klassifikation der Testmethoden

I

Nach der Anzahl der Stichproben unterscheidet man

- **1-Stichprobentests**
- **2-Stichprobentests**
- **Mehrstichprobentests**

II

Zwei Stichproben können paarig (verbunden) oder unverbunden sein.

Paarige Stichproben werden in der Medizin häufig untersucht:

- ein bestimmtes Merkmal wird vor und nach der Therapie erfaßt
- Patienten werden mit 2 unterschiedlichen Therapien gleichzeitig oder im ‚cross-over‘ behandelt

Jeder Wert x_i der einen Stichprobe bildet mit einem y_i der anderen Stichprobe inhaltlich ein Paar.

Paarige Stichproben haben immer denselben Stichprobenumfang.

Unverbundene Stichproben sind bezüglich ihrer Beobachtungseinheiten unabhängig voneinander (unterschiedliche Patienten). Die Stichprobenumfänge können unterschiedlich sein.

III

Unterscheidung nach der zu testenden Parameter

- **Lagetests** zur Prüfung von Erwartungswerten (in der medizinischen Forschung häufig)
- **Dispersionstests** zur Prüfung von Streuungsmaßen
- **Assoziationstests** zur Prüfung von Zusammenhängen
- **Wahrscheinlichkeitstests** und **Unabhängigkeitstests**
- **Homogenitätstests** zur Prüfung ob ein qualitatives Merkmal in mehreren Stichproben gleich verteilt ist
- **Anpassungstests** zum Vergleich einer empirischen Verteilung mit einer theoretischen Verteilung (z.B. Normalverteilung)

IV

Unterscheidung nach Verteilung der Prüfgrößen

- **t-Tests**
 - setzen theoretisch normalverteilte Grundgesamtheiten voraus
 - es werden nur bestimmte Parameter (z.B. Erwartungswerte) überprüft (parametrische Tests)
- **Rangsummentests**
 - im Gegensatz zum t-Test müssen die Grundgesamtheiten nicht normalverteilt sein
 - Prüfgrößen werden nicht aus den Original-Meßwerten, sondern aus deren Rangzahlen bestimmt
- **Binomialtests**
 - es wird überprüft, ob eine relative Häufigkeit mit einer vorgegebenen Wahrscheinlichkeit vereinbar ist (s.Beispiel - Prinzip des statistischen Tests)
- **χ^2 – Tests (CHI)**
 - dienen der Analyse von Häufigkeitsunterschieden (Kontingenztafeln) - vielseitig anwendbar

t-Test für 2 unverbundene Stichproben

Dies ist ein Lagetest, der zur Überprüfung der Gleichheit von 2 Erwartungswerten herangezogen wird.

Voraussetzungen

- 2 unverbundene Stichproben der Umfänge n_1 und n_2
- die Werte beider Stichproben sind normalverteilt mit ähnlichen Standardabweichungen

Es soll getestet werden, ob zwischen den beiden Stichproben statistisch signifikante Unterschiede bestehen.

Unter der Nullhypothese ('keine Unterschiede') hat die Differenz der beiden Mittelwerte $\bar{x}_1 - \bar{x}_2$ den

Erwartungswert Null. Die Prüfgröße berechnet sich dann zu
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

s^2 nennt man die 'gepoolte' Varianz; sie wird aus den Werten beider Stichproben errechnet:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Ist die Prüfgröße nahe bei '0', so spricht dies mehr für die Nullhypothese. Liegt die Prüfgröße weit genug vom Erwartungswert '0' entfernt, so wird die Nullhypothese abgelehnt und die Alternativhypothese akzeptiert. Explizit wird formuliert:

Die Nullhypothese wird abgelehnt, falls $|t| > t_{n_1+n_2-2; 1-\frac{\alpha}{2}}$ (bei zweiseitiger Fragestellung)

falls $|t| > t_{n_1+n_2-2; 1-\alpha}$ (bei einseitiger Fragestellung)

Beispiel: t-Test für 2 unverbundene Stichproben mit jeweils $n = 4$

Es soll gezeigt werden, daß männliche Studenten ein signifikant unterschiedliches Hämoglobin haben als weibliche (zweiseitige Fragestellung), dabei nehmen wir an, das Hämoglobin ist angenähert normalverteilt. Es wird das Blut von 4 Studenten und 4 Studentinnen untersucht:

Männer (m)	Frauen (f)
13.7	12.5
15.2	10.9
14.8	11.8
14.1	13.6
$\bar{x}_m = 14.45$	$\bar{x}_f = 12.2$
$s_m = 0.68$	$s_f = 1.14$

Die Prüfgröße t berechnet sich zu $t = 3.3356$ und ist größer als $t_{6; 0.975} = 3.1634$ (s. t-Tabelle)

Schlußfolgerung:

Die Nullhypothese kann mit einer Irrtumswahrscheinlichkeit kleiner 5% abgelehnt werden, d.h.:

Weibliche Studenten haben ein statistisch signifikant niedrigeres Hämoglobin als männliche ($p < 0.05$).

t-Test für 2 verbundene Stichproben (Paardifferenzen)

Dies ist ein Lagetest, der zur Überprüfung der Gleichheit von 2 Erwartungswerten herangezogen wird.

Voraussetzungen

- 2 paarige Stichproben des Umfangs n mit Wertepaaren (x_i, y_i)
- Differenzen $d_i = y_i - x_i$ (oder $x_i - y_i$), die normalverteilt sind

Es soll getestet werden, ob zwischen den Wertepaaren statistisch signifikante Unterschiede bestehen.

Unter der Nullhypothese ('keine Unterschiede') haben die Differenzen den Mittelwert $\bar{d} = 0$ und die

Standardabweichung $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$. Die Prüfgröße berechnet sich dann zu $t = \frac{\bar{d}}{s_{\bar{d}}}$

Ist die Prüfgröße nahe bei '0', so spricht dies mehr für die Nullhypothese. Liegt die Prüfgröße weit genug vom Erwartungswert '0' entfernt, so wird die Nullhypothese abgelehnt und die Alternativhypothese akzeptiert. Explizit wird formuliert:

Die Nullhypothese wird abgelehnt, falls $|t| > t_{n-1; 1-\frac{\alpha}{2}}$ (bei zweiseitiger Fragestellung)

falls $|t| > t_{n-1; 1-\alpha}$ (bei einseitiger Fragestellung)

Beispiel: t-Test für Paardifferenzen für $n = 6$

6 Probanden nehmen ein Blutdruck senkendes Medikament. Des systolische Blutdruck (mmHg) wird vor und nach der Applikation bestimmt (klassisches Beispiel für abhängige Stichproben).

Es wird getestet, ob sich Bd_{sys} vorher und nachher signifikant unterscheiden. Dabei können wir einseitig testen, da in der Regel ausgeschlossen ist, daß die Therapie den Blutdruck erhöht.

Proband	Bd_{sys} vorher	Bd_{sys} nachher	Differenz
1	132	128	4
2	144	137	7
3	129	132	-3
4	136	125	9
5	131	125	6
6	140	130	10

Mittelwert und Standardabweichung der Differenzen: $\bar{d} = 5.5$ $s_d = 4.68$

Falls die Nullhypothese gilt, so sind die Differenzen normalverteilt mit dem Erwartungswert = 0 und der aus der Stichprobe geschätzten Standardabweichung $s_{\bar{d}}$.

Die Prüfgröße berechnet sich zu $t = \frac{5.5}{\frac{4.68}{\sqrt{6}}} = 2.88$ und ist größer als $t_{5; 0.95} = 2.5706$ (s. t-Tabelle)

Schlußfolgerung:

Die Nullhypothese kann mit einer Irrtumswahrscheinlichkeit kleiner 5% abgelehnt werden, d.h.:

Der Blutdruck ist nach der Therapie statistisch signifikant niedriger als vorher ($p < 0.05$).

In diesem Beispiel testet der t-Test die Nullhypothese: **Differenz vor – nach = 0** (!).

Wahrscheinlich ist '0' aber keine klinisch relevante Differenz. Will man eine klinisch relevante Differenz testen, z.B. 10 mmHg, so kann diese in der Prüfgröße berücksichtigt werden.

Mann Whitney Test - Vertrauensgrenzen für die Rangsumme T, $\varphi = 0.05$
(Tabellenauszug)

		n_1											
n_2	4	5	6	7	8	9	10	11	12	13	14	15	16
4	10-26	16-34	23-43	31-53	40-64	49-77	60-90	72-104	85-119	99-135	114-152	130-170	147-189
5	11-29	17-38	24-48	33-58	42-70	52-83	63-97	75-112	89-127	103-144	118-162	134-181	151-201
6	12-32	18-42	26-52	34-64	44-76	55-89	66-104	79-119	92-136	107-153	122-172	139-191	157-211
7	13-35	20-45	27-57	36-69	46-82	57-96	69-111	82-127	96-144	111-162	127-181	144-201	162-222
8	14-38	21-49	29-61	38-74	49-87	60-102	72-118	85-135	100-152	115-171	131-191	149-211	167-233
9	14-42	22-53	31-65	40-79	51-93	62-109	75-125	89-142	104-160	119-180	136-200	154-221	173-243
10	15-45	23-57	32-70	42-84	53-99	65-115	78-132	92-150	107-169	124-188	141-209	159-231	178-254
11	16-48	24-61	34-74	44-89	55-105	68-121	81-139	96-157	111-177	128-197	145-219	164-241	183-265
12	17-51	26-64	35-79	46-94	58-110	71-127	84-146	99-165	115-185	132-206	150-228	169-251	189-275
13	18-54	27-68	37-83	48-99	60-116	73-134	88-152	103-172	119-193	136-215	155-237	174-261	195-285
14	19-57	28-72	38-88	50-104	62-122	76-140	91-159	106-180	123-201	141-223	160-246	179-271	200-296
15	20-60	29-76	40-92	52-109	65-127	79-146	94-166	110-187	127-209	145-232	164-256	184-281	206-306
16	21-63	30-80	42-96	54-114	67-133	82-152	97-173	113-195	131-217	150-240	169-265	190-290	211-317
17	21-67	32-83	43-101	56-119	70-138	84-159	100-180	117-202	135-225	154-249	174-274	195-300	217-327
18	22-70	33-87	45-105	58-124	72-144	87-165	103-187	121-209	139-233	159-257	179-283	200-310	223-337
19	23-73	34-91	46-110	60-129	74-150	90-171	107-193	124-217	143-241	163-266	184-292	205-320	228-348
20	24-76	35-95	48-114	62-134	77-155	93-177	110-200	128-224	147-249	167-275	188-302	211-329	234-358

Mann Whitney Test - Vertrauensgrenzen für die Rangsumme T, $\varphi = 0.01$
(Tabellenauszug)

		n_1											
n_2	4	5	6	7	8	9	10	11	12	13	14	15	16
4	-	-	21-45	28-56	37-67	46-80	57-93	68-108	81-123	94-140	109-157	125-175	141-195
5	-	15-40	22-50	29-62	38-74	48-87	59-101	71-116	84-132	98-149	112-168	128-187	145-207
6	10-34	16-44	23-55	31-67	40-80	50-94	61-109	73-125	87-141	101-159	116-178	132-198	149-219
7	10-38	16-49	24-60	32-73	42-86	52-101	64-116	76-133	90-150	104-169	120-188	136-209	154-230
8	11-41	17-53	25-65	34-78	43-93	54-108	66-124	79-141	93-159	108-178	123-199	140-220	158-242
9	11-45	18-57	26-70	35-84	45-99	56-115	68-132	82-149	96-168	111-188	127-209	144-231	163-253
10	12-48	19-61	27-75	37-89	47-105	58-122	71-139	84-158	99-177	115-197	131-219	149-241	167-265
11	12-52	20-65	28-80	38-95	49-111	61-128	73-147	87-166	102-186	118-207	135-229	153-252	172-276
12	13-55	21-69	30-84	40-100	51-117	63-135	76-154	90-174	105-195	122-216	139-239	157-236	177-287
13	13-59	22-73	31-89	41-106	53-123	65-142	79-161	93-182	109-203	125-226	143-249	162-273	181-299
14	14-62	22-78	32-94	43-111	54-130	67-149	81-169	96-190	112-212	129-235	147-259	166-284	186-310
15	15-65	23-82	33-99	44-117	56-136	69-156	84-176	99-198	115-221	133-244	151-269	171-294	191-321
16	15-69	24-86	34-104	46-122	58-142	72-162	86-184	102-206	119-229	136-254	155-279	175-305	196-332
17	16-72	25-90	36-108	47-128	60-148	74-169	89-191	105-214	122-238	140-263	160-288	180-315	201-343
18	16-76	26-94	37-113	49-133	62-154	76-176	92-198	108-222	125-247	144-272	164-298	184-326	206-354
19	17-79	27-98	38-118	50-139	64-160	78-183	94-206	111-230	129-255	148-281	168-308	189-336	211-365
20	18-82	28-102	39-123	52-144	66-166	81-189	97-213	114-238	132-264	152-290	172-318	193-347	216-376

Wilcoxon-Test für Paardifferenzen (Rangsummentest)

Dieser Test ist das verteilungsfreie Pendant zum t-Test für 2 verbundene Stichproben (s.a.Übersicht).

Prinzip des Tests:

- für je 2 zusammengehörige Merkmalswerte werden die Differenzen $d_i = x_i - y_i$ gebildet
- die Differenzen d_i werden nach der Größe ihres Betrages (d.h. unabhängig vom Vorzeichen) in aufsteigender Reihenfolge sortiert und mit Rangzahlen versehen
- man addiert die Rangzahlen der negativen (ϱ_-) und die Rangzahlen der positiven Differenzen (ϱ_+). Diese gelten als Testgrößen und können (z.B. für $n=6$) in der Abbildung unten eingeordnet werden.

Beispiel (identisch zum t-Test): Wilcoxon für Paardifferenzen für $n = 6$

6 Probanden nehmen ein Blutdruck senkendes Medikament. Des systolische Blutdruck (mmHg) wird vor und nach der Applikation bestimmt (klassisches Beispiel für abhängige Stichproben). Es soll getestet werden, ob sich Bd_{sys} vorher und Bd_{sys} nachher signifikant unterscheiden.

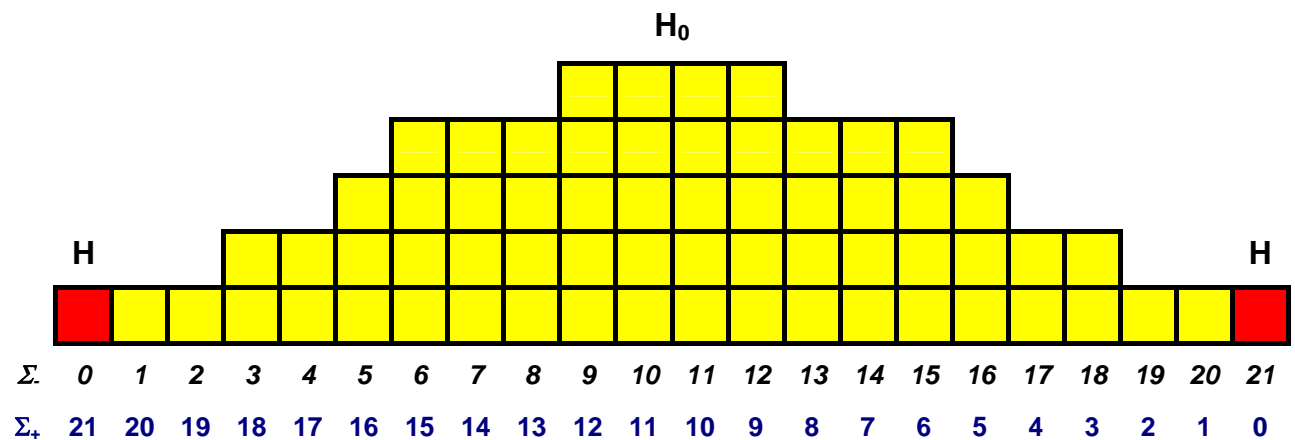
Proband	Bd_{sys} vorher	Bd_{sys} nachher	Differenz	Rangzahlen
1	132	128	4	2
2	144	137	7	4
3	129	132	-3	1
4	136	125	9	5
5	131	125	6	3
6	140	130	10	6

Rangzahlsumme der negativen Differenzen: $\varrho_- = 1$
 Rangzahlsumme der positiven Differenzen: $\varrho_+ = 20$

Beide Rangzahlsummen liegen im H_0 -Bereich, d.h. die Nullhypothese kann nicht abgelehnt werden (um bei $p=0.05$ zu H_A zu gelangen, müssen für $n=6$ alle Differenzen dasselbe Vorzeichen haben, d.h. der Blutdruck hätte bei Proband 3 ebenfalls sinken müssen).

*In diesem Beispiel testet der Wilcoxon-Test die Nullhypothese: **Differenz vor – nach = 0** (!). Wahrscheinlich ist ,0' keine klinische relevante Differenz. Will man eine klinisch relevante Differenz testen, z.B. 10 mmHg, so kann dies berücksichtigt werden und lediglich der Nr.6 erhält eine positive Rangzahl.*

Verteilung der Rangzahlsummen (Anzahl der Möglichkeiten 2^6)



Berechnung der beiden Randflächen: $p = 2/2^6 = 0,0325$

Die nächste Tabelle enthält die kritischen Schranken ($p=0.05$) für die Stichprobengrößen 6-25

Tabelle: Wilcoxon-Test für Paardifferenzen (n=Anzahl der Paare)

n	• 0.10	• 0.05	• 0.01
5	0 – 15		
6	2 – 19	0 – 21	
7	3 – 25	2 – 26	
8	5 – 31	3 – 33	0 – 36
9	8 – 37	5 – 40	1 – 44
10	10 – 45	8 – 47	3 – 52
11	13 – 53	10 – 56	5 – 61
12	17 – 61	13 – 65	7 – 71
13	21 – 70	17 – 74	9 – 82
14	25 – 80	21 – 84	12 – 93
15	30 – 90	25 – 95	15 – 105
16	35 – 101	29 – 107	19 – 117
17	41 – 112	34 – 119	23 – 130
18	47 – 124	40 – 131	27 – 144
19	53 – 137	46 – 144	32 – 158
20	60 – 150	52 – 158	37 – 173
21	67 – 164	58 – 173	42 – 189
22	75 – 178	66 – 187	48 – 205
23	83 – 193	73 – 203	54 – 222
24	91 – 209	81 – 219	61 – 239
25	100 – 225	89 – 236	68 – 257

CHI²-Tests (χ²-Tests)

χ²-Tests dienen zur Analyse von Häufigkeitsunterschieden bezüglich eines oder mehrerer Merkmale. Im einfachsten Fall untersucht der χ²-Tests als ‚Vierfeldertest‘ die Unabhängigkeit von 2 Alternativmerkmalen.

Zwei Merkmale A, D seien dichotom, d.h. für die Ausprägungen ‚A‘ bzw. ‚ \bar{A} ‘ und ‚D‘ bzw. ‚ \bar{D} ‘ gibt es 4 Kombinationsmöglichkeiten mit den Häufigkeiten a,b,c,d, die sich in einer Vierfeldertafel darstellen lassen:

	A	\bar{A}	∅
D	a	b	a+b
\bar{D}	c	d	c+d
∅	a+c	b+d	n

Bei Gültigkeit der Nullhypothese (Unabhängigkeit der beiden Merkmale A und D) gilt nach dem Multiplikationssatz der Wahrscheinlichkeitsrechnung:

$$p(A/D) = p(A)$$

d.h. wenn die Nullhypothese gilt dann gilt für die beobachteten Häufigkeiten: $\frac{a}{a+b} \approx \frac{a+c}{n}$

Dagegen besagt die Alternativhypothese, daß eine Abhängigkeit besteht

Die Idee des χ²-Tests ist:

Die beobachteten Häufigkeiten (B) a,b,c,d werden verglichen mit den Häufigkeiten, die bei Gültigkeit der Nullhypothese zu erwarten sind (E).

Nach einfachem Dreisatz berechnen sich die Erwartungswerte: $E = \frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{\text{Gesamtsumme}}$

Dann berechnet man für jede Häufigkeit den Quotienten

$$\frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}} = \frac{(B - E)^2}{E}$$

In einer Vierfeldertafel hat man 4 solche Quotienten und die Summe dieser Quotienten ist die Prüfgröße χ²

$$\chi^2 = \sum \frac{(B - E)^2}{E}$$

Ist die Prüfgröße nahe bei ‚0‘, so spricht dies mehr für die Nullhypothese. Liegt die Prüfgröße weit genug vom Erwartungswert ‚0‘ entfernt, so wird die Nullhypothese abgelehnt und die Alternativhypothese akzeptiert. Explizit wird formuliert:

Falls der Wert der Prüfgröße innerhalb des Intervalls $[0, \chi^2_{v=1; 1-\alpha}]$ liegt, wird die Nullhypothese auf dem ∅-Niveau beibehalten. Für eine Vierfeldertafel (Freiheitsgrad ∅ = 1) und ∅=5% ist $\chi^2_{1; 0.95} = 3.841$ (s.a. Tabelle)

Zur Vermeidung irrtümlich berechneter Signifikanzen sollten beim χ²-Test wenigstens 80% der zu erwartenden Häufigkeiten mindestens 5 betragen (d.h. bei einer Vierfeldertafel müssen alle 4 zu erwartenden Häufigkeiten mindestens 5 betragen).

Sind diese Anforderungen nicht erfüllt, gilt als Alternative der ‚**Exakte Fishertest**‘.

Den ‚Exakten Fishertest‘ verwendet man hauptsächlich bei Vierfeldertafeln mit kleinen Randsummen.

Theoretisch kann er aber auch bei ∅ m-Kontingenztafeln und beliebig großen Randsummen berechnet werden, allerdings ist der Rechenaufwand so groß, daß man sich besser mit der Approximation über das χ² begnügt.

Mit dem χ²-Test läßt sich die Existenz eines Zusammenhangs zwischen zwei Merkmalen nachweisen, über dessen Stärke macht das Testergebnis jedoch keine Angabe. Diese kann quantifiziert werden durch den **Kontingenzkoeffizienten**, der wie der Korrelationskoeffizient r von Pearson entwickelt wurde und in gewisser Weise mit diesem vergleichbar ist.

Sind die Merkmale nicht dichotom, (d.h. die Anzahl der Ausprägungen ist größer als 2), so erhält man anstatt einer Vierfeldertafel eine Mehrfeldertafel ($n \times m$ -Kontingenztafel). Der Test heißt dann χ^2 -Kontingenztest.

Beispiel (Vierfeldertafel):

In einer kontrollierten klinischen Studie sollte die antidepressive Wirkung von ‚Hypericum‘ im Vergleich zu Placebo untersucht werden. 200 Patienten mit depressiven Verstimmungen wurden in zwei Gruppen zu je 100 randomisiert. Mit der Hamilton-Depressivitätsskala (HAMD) wurden Responderraten bestimmt. Unter Hypericum sprachen 40 Patienten, in der Placebo-Gruppe nur 25 Patienten auf die Therapie an.

Nullhypothese:

Gruppe (Spaltenmerkmal) und Responderrate (Zeilenmerkmal) sind unabhängig - oder

zwischen den beiden Gruppen (Hypericum / Placebo) bestehen keine statistisch signifikanten Unterschiede bezüglich der Responderraten.

	Hypericum	Placebo	φ
Responder	40 (32.5)	25 (32.5)	65
Nonresponder	60 (67.5)	75 (67.5)	135
φ	100	100	200

Neben den Beobachtungswerten stehen in Klammern die berechneten Erwartungswerte (unter Gültigkeit der Nullhypothese).

Bei Durchführung des χ^2 – Vierfeldertests ergibt sich $\chi^2 = 5.13 > \chi^2_{1;0.95} = 3.841$

Schlußfolgerung:

Die Nullhypothese kann mit einer Irrtumswahrscheinlichkeit kleiner 5% abgelehnt werden, d.h.:

Die Responderrate ist in der Hypericum-Gruppe statistisch signifikant höher als bei Placebo ($p < 0.05$).

χ^2 -Tabelle

FG	φ	0.05	0.01	0.001
1		3.841	6.635	10.828
2		5.991	9.210	13.816
3		7.815	11.345	16.266
4		9.488	13.277	18.467

Übersicht häufiger statistischer Testverfahren in der Medizin

		2 Stichproben		mehr als 2 Stichproben	
		unverbunden	verbunden	unverbunden	verbunden
stetig	normalverteilt	t-Test	t-Test paarig	Varianzanalyse	Varianzanalyse
	nicht normalverteilt	Mann-Whitney Test	Wilcoxon Test	Kruskal-Wallis Test	Friedman Test
diskret		χ^2 -Test Exakter Fisher Test	χ^2 -Test nach McNemar	χ^2 -Kontingenztest	

Problem des multiplen Testens

In einer klinischen Studie oder im medizinischen Alltag entstehen jede Menge von Daten, die im Zeitalter schneller Computer kein nennenswertes Problem bei der Analyse darstellen. So ist man leider allzu oft geneigt, den statistischen Test für jedes Merkmal und jede Merkmalsverknüpfung durchzuführen. Man macht einen Test nach dem anderen, in der Hoffnung statistisch signifikante Ergebnisse zu erhalten. Genau diese werden dann publiziert, der Rest wird verschwiegen.

Vorsicht:

Bei mehrmaligem Testen steigt der φ -Fehler in einen Bereich, der für wissenschaftliche Studien nicht zu akzeptieren ist.

Nehmen wir an, bei einem Test würde tatsächlich die Nullhypothese gelten (*es gibt in Wirklichkeit keinen Unterschied zwischen A und B*). Die Wahrscheinlichkeit, daß sich der Test für die Nullhypothese entscheidet beträgt $1 - \varphi$, also in der Regel (bei $\varphi=5\%$) 0.95 .

Bei 10 Tests beträgt diese Wahrscheinlichkeit nur noch $(1 - \varphi)^{10}$, also etwa 60% und das bedeutet, daß der φ -Fehler auf 40% angestiegen ist.

Entschärfen des Problems:

- nicht wahllos jeden statistischen Test durchführen, der theoretisch denkbar ist
- gründliche Versuchsplanung und Präzisieren der konkreten Fragestellung vor Studienbeginn
- überlegen, welche Tests dem inhaltlichen Problem angemessen sind
- evtl. multivariate Verfahren anstatt mehrerer einzelner Tests (Varianzanalyse statt mehrerer t-Tests)
- evtl. Zusammenlegung der Merkmale (z.B. AUC, c_{\max} , t_{\max} statt Werte einzelner Meßzeitpunkte)
- Bonferroni-Korrektur (sog. φ -Adjustierung)

Interpretation eines Testergebnisses

- Die formale statistische Signifikanz (nur dies kann ein statistischer Test nachweisen!) und die medizinische Relevanz eines Ergebnisses sind zwei verschiedene Dinge und sollten deswegen nicht verwechselt werden.
- Ein signifikantes Ergebnis sagt lediglich aus, daß ein Unterschied existiert, der mit einer Wahrscheinlichkeit von kleiner φ auch zufällig entstanden sein könnte.
- Ein signifikantes Testergebnis sagt nichts über die Größe des Unterschiedes.
- Ein signifikantes Testergebnis sagt nichts aus über die Ursachen des Unterschiedes.
- Aus einem nicht signifikanten Testergebnis kann nicht geschlossen werden, daß tatsächlich kein Unterschied besteht. Möglicherweise gibt es ja einen bedeutsamen (und auch klinisch relevanten) Unterschied, der aber wegen zu kleiner Stichprobengrößen nicht nachzuweisen war. (bei klinischen Studien macht man aus diesem Grund vor der Studie sog. 'Stichprobenumfangschätzungen')
- Der Anwender eines statistischen Tests ist nicht nur für die korrekte Anwendung des Tests verantwortlich, sondern auch für die Interpretation des Testergebnisses.

Überlebensanalyse (Life-tables, Survival Analyse) - Kaplan-Meier-Methode

Unter Überlebenszeiten verstehen wir die Zeit zwischen zwei Zeitpunkten, zum Beispiel die Zeit

- von Geburt bis Tod
- von Diagnosestellung bis Tod
- von Diagnosestellung bis Rezidiv
- von Operationszeitpunkt bis Tod
- von Operationszeitpunkt bis Funktionsausfall

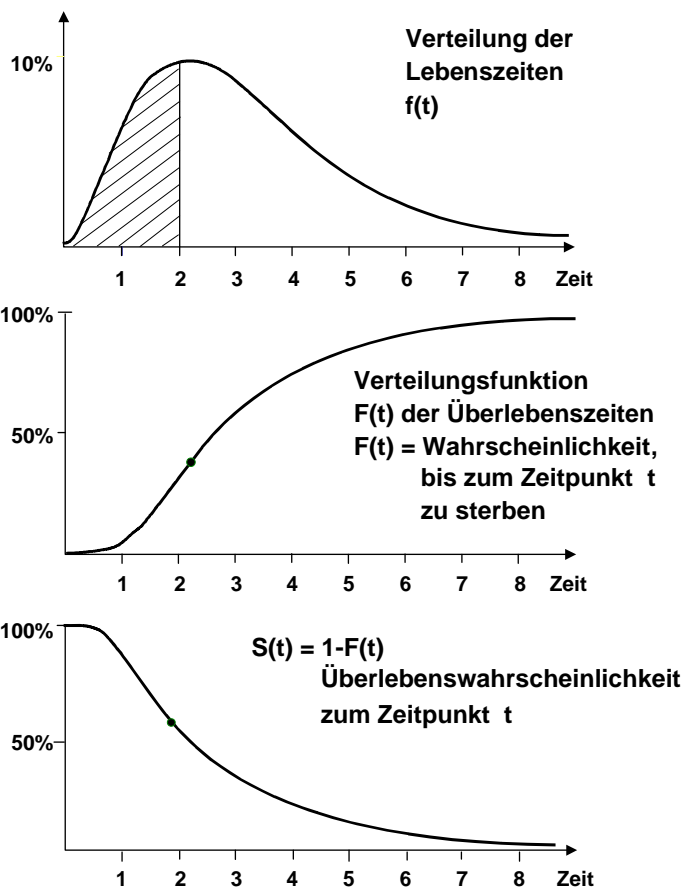
Mögliche Ziele einer Überlebensstudie sind u.a.

- Beschreibung des zeitlichen Verlaufs des Überlebens
- Bestimmung der Überlebensrate zu einem bestimmten Zeitpunkt
- Vergleich von Therapien bezüglich des Überlebens

Würde man in einer Studie sämtliche Überlebenszeiten der Patienten, so könnte man wie in der nächsten Abbildung gezeigt, die Verteilung der Lebenszeiten erstellen. Diese Verteilung ist i.d.R. nicht normalverteilt, sondern rechtsschief (obere Grafik).

Durch Integration erhält man die Verteilungsfunktion und damit die Wahrscheinlichkeiten bis zu einem bestimmten Zeitpunkt zu sterben (mittlere Grafik).

Durch Umkehrung erhält man Überlebenswahrscheinlichkeiten bis zu einem bestimmten Zeitpunkt (untere Grafik). Die Kurve wird als Überlebenskurve (Survival Kurve, Kaplan-Meier Kurve) bezeichnet.



Zunächst sind einige Begriffe zu klären die innerhalb einer Überlebenszeitanalyse notwendig sind:

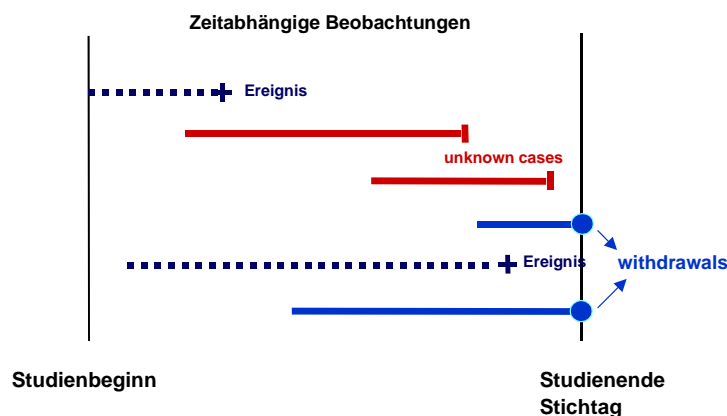
- Überleben:** Nichteintreffen eines (genau definierten) Zielereignisses
z.B. Tod, Rezidiv, Krankheit, Schmerzen, Infektion, vollständige Heilung, ... im Beobachtungszeitraum
- Überlebenszeit:** Zeiteinheit zwischen einem festgelegten Zeitpunkt (Operation, Erstdiagnose) und dem Eintreffen des Ereignisses
- Zensierte Überlebenszeit:** Ereignis wird nicht beobachtet
 - unknown case* Ausscheiden wegen anderer Ursachen (Umzug, Tod nicht im Zshg. mit dem untersuchten Ereignis, ...)
 - withdrawal* Beobachtungszeit zu kurz

Die zu untersuchende Patientengruppe kann zum Studienbeginn bereits vollständig gegeben sein oder sie bildet sich im Laufe der Studie, zum Beispiel, wenn die Diagnose gestellt wird, oder wenn die Operation erfolgt.

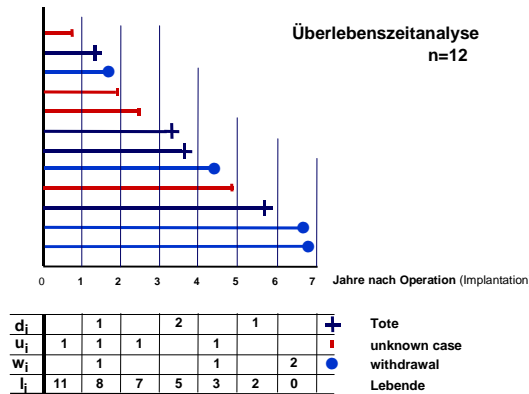
Der Patient wird solange beobachtet, bis

1. Das Ereignis (Tod, Rezidiv,...) eintritt (Abk.: $d_i = \text{'dead'}$) **oder**
- 2a die Studienzeit zu Ende ist $I_i = \text{lebend, am Ende } w_i = \text{withdrawal}$
- b der Patient die Kriterien der Studie nicht mehr erfüllt I_i , dann $u_i = \text{unknown case}$
- c der Patient nicht mehr beobachtet werden kann (lost to follow-up) I_i , dann $u_i = \text{unknown case}$
- d ein konkurrierendes Ereignis eintritt (Tod aus anderen Ursachen...) I_i , dann $u_i = \text{unknown case}$

Im ersten Fall wird die gesamte Überlebenszeit, im zweiten Fall nur ein Teil der Überlebenszeit beobachtet. Im zweiten Fall weiß man nur, daß der Patient mindestens diesen beobachteten Teil der Lebenszeit überlebt hat. Aus diesem Grund nennt man die letzteren Beobachtungen **zensiert**. Zudem können die Patienten zu verschiedenen Zeiten in die Studie aufgenommen werden, was im nachfolgenden zeitlichen Schema deutlich wird.



Um einen besseren Überblick zu gewinnen werden die Überlebenszeiten auf den Zeitpunkt ‚Null‘ (Eintritt in die Studie) synchronisiert (s. nächste Abbildung).



In den meisten Überlebensstudien ist ein Teil der Beobachtungen zensiert. Die Zensierung macht eine biometrische Analyse der Studie mit konventionellen statistischen Mitteln unmöglich. Deshalb wurden spezielle Verfahren entwickelt (Aktuars-Methode, Verfahren nach Kaplan und Meier u.a.). Prinzipiell ist es möglich, den mathematischen Teil einer Überlebensanalyse einem Computerprogramm zu überlassen, was allerdings wenig zum Verständnis beiträgt und eine überzeugende inhaltliche Interpretation der Ergebnisse gefährdet.

In einem Beispiel (Tab. unten) wird versucht, den korrekten Umgang mit den bisher definierten Größen zu veranschaulichen und die rechnerische Ermittlung der Überlebenswahrscheinlichkeiten leicht nachvollziehbar zu machen. Die Zeitachse ist dabei in gleich lange Intervalle eingeteilt, was im Rechenweg der sog. *Aktuarmethode nach Cutler und Ederer* entspricht.

Berechnung der Überlebensrate

Zeit	Leben unter Beob.	Tote	with-drawal	unkn cases	Leben unter Risiko	Korrig. Sterbewahrsch HAZARD	Überl.-wahrsch.	kum. Erfolgsrate
t	l	d	w	u	l'	q	p	S
t ₀ - t ₁	126	47	15	4	116,5	0,4	0,6	0,6
t ₁ - t ₂	60	5	11	6	51,5	0,1	0,9	0,54
t ₂ - t ₃	38	2	15	0	30,5	0,07	0,93	0,5
t ₃ - t ₄	21	2	7	2	16,5	0,12	0,88	0,44
t ₄ - t ₅	10	0	6	0	7,0	0,0	1,0	0,44

Möglichkeiten der Berechnung der Sterbewahrscheinlichkeit im ersten Intervall:

$$\begin{array}{ccc}
 \frac{47}{126} & \frac{47}{116,5} & \frac{47}{107} \\
 0,373 & 0,403 & 0,439 \\
 \text{optimistisch} & \uparrow & \text{pessimistisch} \\
 & \text{korrigiert} &
 \end{array}$$

Das Beispiel beginnt im ersten Zeitintervall mit 126 Lebenden und 47 Toten (bzw. Mißerfolgen, Rezidiven). Eine übliche Schätzung der Sterbewahrscheinlichkeit mit $47/126 = 0.37$ wäre zu optimistisch, denn es liegen neben den Toten im Zeitintervall 19 sog zensierte Überlebende vor (15 ‚withdrawals‘ und 4 ‚unknown cases‘).

Ignoriert man die zensierten Beobachtungen, so würde das mit $47/(126-15-4) = 0.43$ eine zu pessimistische Schätzung zur Folge haben.

Um die zensierten Beobachtungen optimal in die Wahrscheinlichkeitsberechnungen einzubeziehen (statistisch hochzurechnen), werden die Toten auf die ‚effektiv unter Risiko stehenden Lebenden‘ bezogen, die man dadurch erhält, daß man von den ursprünglichen 126 Lebenden die Hälfte der zensierten Beobachtungen ($\frac{1}{2}$ (withdrawals + unknown cases) = 9.5) subtrahiert. Der mathematische Hintergrund für diese Arithmetik ist in verständlicher Kurzform in der nächsten Abbildung zusammengefaßt.

Mit $47/116.5 = 0.40$ erhält man schließlich eine korrigierte Sterbewahrscheinlichkeit, die als *Hazard* bezeichnet wird. Da Sterbe- und Überlebenswahrscheinlichkeit komplementär sind, ergibt eine Überlebenswahrscheinlichkeit von 0.6, was zur Folge hat, daß die ‚Überlebenskurve‘ am Ende des ersten Zeitintervalls von 100% auf 60% abfällt (s übernächste Abbildung)

Mathematische Grundlagen der Life-tables

Sterbewahrscheinlichkeit = $\frac{\text{Tote}}{\text{Lebende}}$ allgemeiner Ansatz

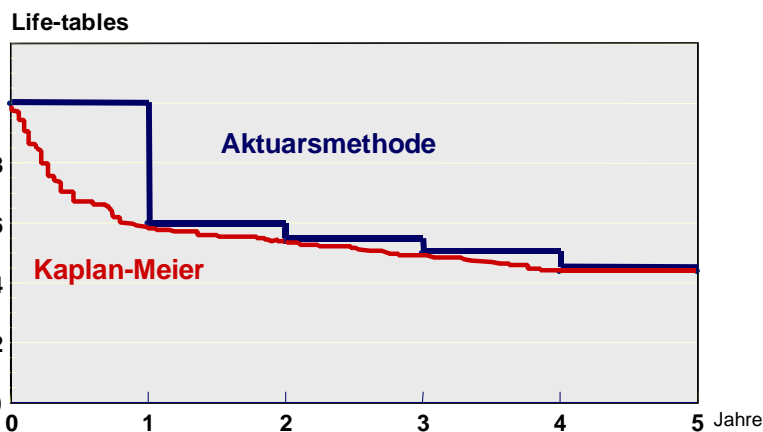
$q = \frac{d}{l}$ zu optimistisch: **q zu klein**
(withdrawals und unknown in bei Lebenden l enthalten)

$q = \frac{d + q \cdot 1/2 (w+u)}{l}$ Idee: w und u stehen im Mittel bis zur Intervallmitte unter Beobachtung

mathematische Umformung

$= \frac{d}{l - 1/2(w+u)}$ Korr. Sterbewahrsch. (HAZARD)
Tote bezogen auf die "effektiv Risiko"-Lebenden

$q = \frac{d}{l - (w+u)}$ zu pessimistisch: **q zu groß**
(withdrawals und unknown sind unberücksichtigt)



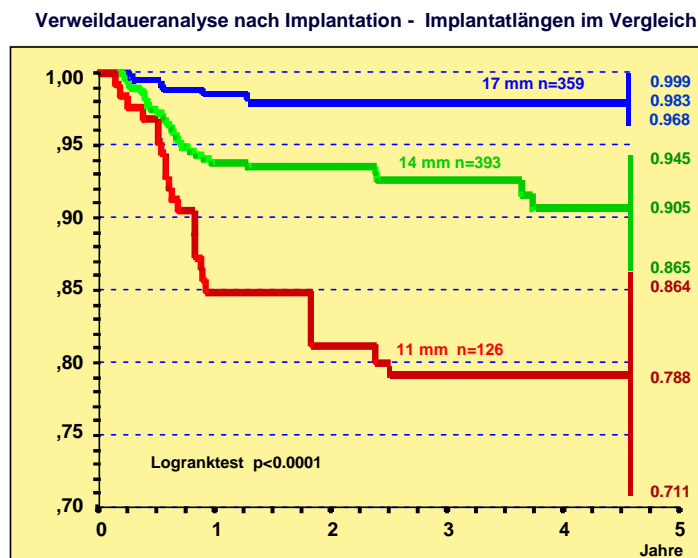
Die Berechnungsprozedur wird für das zweite Intervall wiederholt. Als Startwert stehen dann 60 Lebende unter Beobachtung ($126 - 47 - 15 - 4 = 60$). Bei 5 Toten, 11 withdrawals und 6 unknown cases ergibt sich für das zweite Intervall eine korrigierte Sterbewahrscheinlichkeit (Hazard) von $5/51.5 = 0.1$ und damit eine Überlebenswahrscheinlichkeit von 0.9. Da die Wahrscheinlichkeit ‚zwei Intervalle zu überleben‘ gleich der Wahrscheinlichkeit ist, das erste als auch das zweite Intervall zu überleben, errechnet sich die kumulierte Überlebenswahrscheinlichkeit zu $0.6 \cdot 0.9 = 0.54$, was wiederum bedeutet, daß die Überlebenskurve am Ende des zweiten Intervalls von 60% auf 54% abfällt.

Diese Berechnung kann entsprechend intervallweise fortgeführt werden. Natürlich sollte man dabei beachten, daß mit zunehmend kleineren Zahl der Lebenden die Schätzungen der Wahrscheinlichkeiten immer unzuverlässiger werden. Um dieser Tatsache Rechnung zu Tragen, sollten entsprechende Konfidenzintervalle die Ungenauigkeiten dokumentieren und ggf. in die Abbildung aufgenommen werden (s. z.B. nächste Abbildung).

Die beschriebene Aktuarsmethode nach *Cutler* und *Ederer* ist für beliebig klein werdende Zeitintervalle (z.B. Tage) mit der Methode nach *Kaplan* und *Meier* identisch. Die Kaplan-Meier Kurve ändert also ihre Höhe nicht erst am Ende eines Intervalls, sondern dort, wo Ereignisse (Tote) auftreten.

Überlebenskurven unterschiedlicher Gruppen können über den gesamten Beobachtungszeitraum auf statistische Signifikanz überprüft werden, z.B. mit Hilfe des ‚Logrank-Tests‘ (=generalisierter χ^2 -Test). Das Prinzip des Logrank-Test ist in der übernächsten Abbildung skizziert.

Beispiele für verschiedene Verweildauerkurven bei Zahnimplantaten (entspricht Überlebenskurven nach Kaplan und Meier) sind inklusive Konfidenzintervalle und Ergebnis des Logrank-Tests in der nächsten Abbildung gegeben.



Berechnung von signifikanten Unterschieden zwischen zwei Überlebenskurven am Beispiel des Logrank-Tests

Zeitpunkt	Gruppe 1		Gruppe 2		Gesamt		Erwartungswerte bei H_0	
	$l1$	$d1$	$l2$	$d2$	$l1 + l2$	$d1 + d2$	$e1$	$e2$
1	30	0	50	1	80	1	0,38	0,63
2	30	0	46	1	76	1	0,39	0,61
3	29	4	41	0	70	4	1,66	2,34
4	25	3	41	2	66	5	1,89	3,11
5	22	2	37	3	59	5	1,86	3,14
6	20	4	33	4	53	8	3,01	4,97
7	15	3	26	3	41	6	2,21	3,81
8	12	2	22	0	34	2	0,71	1,29
9	10	1	18	2	28	3	1,07	1,93
9	6	2	15	0	21	2	0,56	1,43
10	3	1	10	1	13	2	0,46	1,55
11	1	0	9	1	10	1	0,11	0,89
	22		18				14,31	25,7

$$\begin{aligned}
 \text{logrank } \chi^2_{FG=1; p=0.05} &= \frac{(\sum d1 - \sum e1)^2}{\sum e1} + \frac{(\sum d2 - \sum e2)^2}{\sum e2} = \frac{(22-14,31)^2}{14,31} + \frac{(18-25,7)^2}{25,7} \\
 &= 4,13 + 2,3 = 6,43
 \end{aligned}$$

($\chi^2_{FG=1; p=0.05} = 3,84$)

Beispiele für häufige Arten von Studien in der Medizin

Fall-Kontroll-Studie

Case control study

- Design retrospektiv, vergleichend
- Prinzip erkrankte Personen
Bildung einer geeigneten Kontrollgruppe
(matched pairs)
- Ziel Klärung ätiologischer Fragestellungen
- ! Erhebungsschwierigkeiten
Selektion in der Stichprobe der Erkrankten
- Auswertung explorative Datenanalyse
relatives Risiko, odd's ratio

Kohortenstudie

Anwendungsbeobachtung
follow-up study

- Design prospektiv
- Prinzip exponierte Personen
Folgebeobachtungen
- Ziel Erfassung von unerwünschten Ereignissen (Nebenwirkungen)
bei Arzneimitteln, Genußmitteln, ...
- ! Zeitaufwand
drop-out Probleme
unterschiedliche Beobachtungsintensität in den Kohorten
- Auswertung explorative Datenanalyse
Erkrankungsinzidenzen
Überlebensraten (survival, life-table)
relatives Risiko, odd's ratio

Kontrollierte klinische Therapiestudie

controlled clinical study

- Design prospektiv, kontrolliert (vergleichend), randomisiert
- Prinzip 2 (oder mehr) Gruppen von Patienten
,offen' oder ,verblindet'
- Ziel Untersuchung von Wirksamkeit und Verträglichkeit
- ! Ein- und Ausschlusskriterien
Dosierungsschema, Behandlungs- bzw. Untersuchungsplan
Haupt- und Nebenzielgrößen
Stichprobenumfangschätzung, drop-outs
Patientenaufklärung, Patientenversicherung
Compliance
Monitoring
...
- Auswertung statistische Testverfahren
Konfidenzbereiche
explorative Datenanalyse