Biometrie und Epidemiologie

Max Resch und Johannes Maschler

Sommersemester 2010

Inhaltsverzeichnis

1 Vorlesung am 4. März 2010

1.1 Überlebenszeiten und -funktionen

1.1.1 Bevölkerungsstatistik und Sterbetafeln

Abschnitt 6.4 (S.114 – 117)

$$\mathbb{E}\,X = \sum_i x_i p_i \ldots \text{ Erwartungswert, Lebenserwartung}$$

1.1.1.1 Sterbetafel[1]

Abschnitt 6.4.2

l_o	$\ldots$	nominelle Anzahl von Lebendgeborenen
l_x	$\ldots$	Anzahl der Personen, die den x-ten Geburtstag erleben
d_x	$:=$	$l_x - l_{x+1}$
x	$:=$	Anzahl der vollendeten Lebensjahre

Mit vollendetem Lebensalter ist das ganzzahlige Alter – die Anzahl der vollendeten Lebensjahre – gemeint, dies ist ein diskretes Merkmal. Das kontinuuierliche Alter hingegen ist stetig.

$$x \in \mathbb{N}_0 \quad \text{wobei} \quad x := \lfloor \text{kontinuuierliches Alter} \rfloor$$

Die Wahrscheinlichkeit, wenn ein Individuum den x-ten Geburtstag erlebt hat, im darauffolgenden Jahr stirbt:

$$q_x = \frac{d_x}{l_x}$$

Für die letzte Zeile, wird ω als Alter x angenommen, meistens ist das maximal berücksichtigte Alter $\omega = 100$, weiters wird angenommen, dass

$$l_{\omega+1} = 0 \qquad q_\omega = 1 \qquad e_\omega = \frac{1}{2}$$

1.1.1.2 Lebenserwartung

Für die Berechnung der Lebenserwartung e_0 (eines Neugeborenen)

[1] Lifetable

Zufallsvariable

$X \ldots$ Anzahl der Lebensjahre, die beim Tod *vollendet* sind

Wahrscheinlichkeit

$\mathrm{P}(X = x) = \frac{d_x}{l_0} \ldots$ Wahrscheinlichkeit im ganzzahligen Alter x zu sterben

Herleitung von e_0 Formel 6.16

$$
\begin{aligned}
e_0 &= \frac{1}{2} + \text{Anzahl der erwarteten vollendeten Lebensjahre} \\
&= \frac{1}{2} + \left[0 \cdot \frac{d_0}{l_0} + 1 \cdot \frac{d_1}{l_0} + 2 \cdot \frac{d_2}{l_0} + \cdots + (\omega - 1) \cdot \frac{d_{\omega-1}}{l_0} + \omega \cdot \frac{d_\omega}{l_0}\right] \\
&= \frac{1}{2} + \frac{1}{l_0}\left[0 \cdot (l_0 - l_1) + 1 \cdot (l_1 - l_2) + \cdots + (\omega - 1) \cdot (l_{\omega-1} - l_\omega) + \omega \cdot (l_\omega - l_{\omega+1})\right] \\
&= \frac{1}{2} + \frac{1}{l_0}\left[0 \cdot l_0 + l_1 + l_2 + \cdots + l_\omega + \omega \cdot 0\right] \\
&= \frac{1}{2} + \frac{1}{l_0} \cdot \sum_{x=1}^{\omega} l_x
\end{aligned}
$$

e_x ist die fernere Lebenserwartung (die Restliche Lebensdauer) eines x-Jährigen.

$$e_x = \frac{1}{2} + \frac{1}{l_x} \cdot \sum_{y=x+1}^{\omega} l_y \qquad \text{für } x \in [0 \ldots \omega]$$

Das erwartete Sterbealter ist dann $x + e_x$ Formel 6.18

$$F(X) = 1 - \frac{l_x}{l_0} \qquad \text{Verteilungsfunktion}$$

1.1.2 Verteilung von Überlebenszeiten

Abschnitt 8.4 (S. 162 – 166)

1.1.2.1 Überlebenszeit

Abschnitt 8.4.1

Die Überlebenszeit ist die Dauer vom Anfangsereignis (Geburt) bis zum Endereignis (Tod)

X (oder T) ... Überlebenszeit (Überlebensdauer)
$F(t) = \mathrm{P}(X \leq t)$
$S(t) = \mathrm{P}(X > t) = 1 - F(t)$
$S(t)$ ... Überlebensfunktion (survival function)
$f(t) = F'(t)$ ... Dichtefunktion
$r(t) = \frac{f(t)}{S(t)}$ ... Momentane Sterberate, Hazard-Funktion, Hazard-Rate, Ausfallrate

Bedingte Überlebenswahrscheinlichkeit[2]

$$\begin{aligned} \mathrm{P}(X > t + \Delta t \mid X > t) &= \frac{\mathrm{P}(X > t + \Delta t \wedge X > t)}{\mathrm{P}(X > t)} \\ \frac{\mathrm{P}(X > t + \Delta t)}{\mathrm{P}(X > t)} &= \frac{S(t + \Delta t)}{S(t)} \end{aligned}$$

Bedingte Sterbewahrscheinlichkeit

$$\begin{aligned} \mathrm{P}(t < X \leq t + \Delta t \mid X > t) &= \frac{\mathrm{P}(t < X \leq t + \Delta t \wedge X > t)}{P(X > t)} = \\ = \frac{\mathrm{P}(t < X \leq t + \Delta t)}{\mathrm{P}(X > t)} &= \frac{F(t + \Delta t) - F(t)}{S(t)} \end{aligned}$$

Momentane Sterberate Die Momentane Sterberate (Ausfallrate) ergibt sich aus der bedingten Sterbewahrscheinlichkeit folgendermaßen:

$$\begin{aligned} r &:= \lim_{\Delta t \to 0} \frac{\mathrm{P}(t < X \leq t + \Delta t \mid X > t)}{\Delta t} = \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{S(t)} = \\ &= F'(t) \cdot \frac{1}{S(t)} = \frac{f(t)}{S(t)} \end{aligned}$$

Zur Interpretation von $r(t)$

In linearer Approximation (für hinreichend kleines Δt) gilt bekanntlich folgendes:

$$F(t + \Delta t) \approx F(t) + F'(t) \cdot \Delta t$$

[2]Nach dem Bayestheorem für bedingte Wahrscheinlichkeiten $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$

Daher gilt für die Wahrscheinlichkeit, dass ein Individuum, nachdem es das Alter t erlebt hat im folgenden Zeitintervall der Δt stirbt:

$$\begin{aligned} \mathrm{P}\left(t < X \leq t + \Delta t \mid X > t\right) &= \frac{F(t+\Delta t) - F(t)}{S(t)} \approx \\ \approx \frac{F(t) + F'(t)\Delta t - F(t)}{S(t)} &= \frac{F'(t)\Delta t}{S(t)} = \\ = \frac{f(t)\Delta t}{S(t)} &= r(t)\Delta t \end{aligned}$$

Die Wahrscheinlichkeit ist also (in linearer Näherung) das Produkt aus momentaner Sterberate und der Länge des Zeitintervalls, somit proportional zu Δt

Mit $\Delta t := 1$ (Einheitszeit) ergibt sich daraus

$$r(t) \approx \mathrm{P}\left(t < X \leq t + 1 \mid X > t\right)$$

Das bedeutet, dass die Ausfallrate $r(t)$ folgendermaßen interpretiert werden kann: (näherungsweise in linearer Approximation) $r(t)$ ist die Wahrscheinlichkeit, dass ein Individuum, dass das Alter t erlebt hat während der folgenden (kleinen) Zeit*einheit* stirbt.

Die momentane Sterberate (Ausfallrate) ist charakterisiert durch

$$r(t) = \frac{f(t)}{S(t)}$$

wobei

$$\begin{aligned} f(t) &= F'(t) \\ S(t) &= 1 - F(t) \end{aligned}$$

Daraus folgt

$$f(t) = -S'(t)$$

und wegen

$$\begin{aligned} F(0) = 0 \qquad &\text{bzw.} \qquad S(0) = 1 \\ F(t) = \int_0^t f(\tau)\, d\tau \qquad &\text{bzw.} \qquad S(t) = 1 - \int_0^t f(\tau)\, d\tau \end{aligned}$$

Fragen

1. Wie kann $r(t)$ nur durch $S(t)$ oder $F(t)$ oder $f(t)$ ausgedrückt werden?

$$\begin{aligned} r(t) &= \frac{-S'(t)}{S(t)} \\ r(t) &= \frac{F'(t)}{-F(t)} \\ r(t) &= \frac{f(t)}{1 - \int_0^t f(\tau)\, d\tau} \end{aligned}$$

2. Wie können $S(t)$ bzw. $F(t)$ bzw. $f(t)$ *nur* durch $r(t)$ ausgedrückt werden?
Mit der Notation τ anstelle von t ist

$$r(\tau) = \frac{S'(\tau)}{S(\tau)}$$

Unter der Voraussetzung $S(t) > 0$ gilt auch $S(\tau) > 0$ für $0 \leq \tau \leq t$, daher gilt

$$\begin{aligned} -\int_0^t r(\tau)\, d\tau &= -\int_0^t \frac{S'(\tau)}{S(\tau)}\, d\tau \\ \ln S(\tau)\,|_0^t &= \ln S(t) - \ln S(0) \\ \ln S(t) - 0 &= \ln S(t) \end{aligned}$$

Somit ist

$$\begin{aligned} S(t) &= e^{-\int_0^t r(\tau)\, d\tau} \\ F(t) &= 1 - e^{-\int_0^t r(\tau)\, d\tau} \\ f(t) &= r(t) \cdot e^{-\int_0^t r(\tau)\, d\tau} \end{aligned}$$

Letzteres ergibt sich aus

$$f(t) = F'(t)$$

oder aus

$$r(t) = \frac{f(t)}{S(t)} \Longleftrightarrow f(t) = r(t) \cdot S(t)$$

2 Vorlesung am 11. März 2010

2.0.3 Verteilung von Überlebenszeiten

2.0.3.1 Exponentialverteilung

Abschnitt 8.4.2

Im Spezialfall wird die Exponentialverteilung

$$\operatorname{Exp}(\lambda)$$

Die Sterberate ist konstant

$$r(t) = \lambda$$

Also handelt es sich bei der Exponentialverteilung um eine gedächnislose Verteilung, da die Änderung unabhängig von der Zeit ist.

$$\int_0^t r(\tau)\, d\tau = \int_0^t \lambda\, d\tau = \lambda t$$

Damit ergeben sich auch auf diesem Weg die für die Exponentialverteilung bekannten Formeln

$$\begin{aligned} S(t) &= e^{-\lambda t} \\ F(t) &= 1 - e^{-\lambda t} \\ f(t) &= \lambda e^{-\lambda t} \end{aligned}$$

Der Erwartungswert und gefolgt von der Varianz, welche sich mit dem Verschiebesatz[1] errechnen lässt.

$$\begin{aligned} \mu &= \frac{1}{\lambda} \\ \sigma^2 &= \frac{1}{\lambda^2} \end{aligned}$$

[1] $\mathbb{E}\,(X - \mathbb{E}\,X)^2 = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$

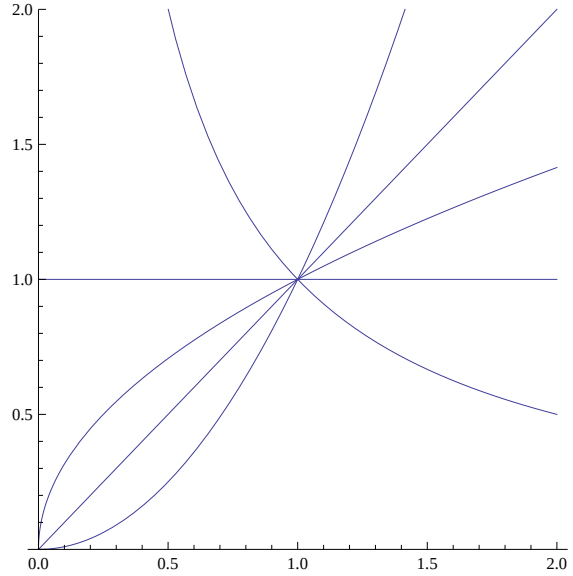


Abbildung 2.0.1: Das Verhalten der Potenenzfunktion $f(x) = x^a$ für $a \in \left\{-1, 0, \frac{1}{2}, 1, 2\right\}$

Die Schiefe[2] γ_1 ist immer 2.

2.0.3.2 Weibull Verteilung

Abschnitt 8.4.3

Die Weibullverteilung ist eine Verallgemeinerung der Exponentialverteilung

$$\mathrm{WB}\,(\lambda, \gamma)$$

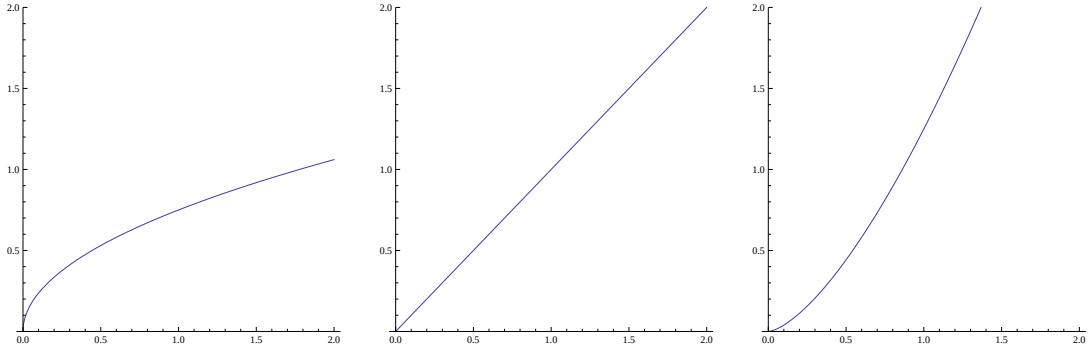
Es gelten dabei folgende Formeln

$$\begin{aligned} S(t) &= e^{-\lambda t^\gamma} \\ F(t) &= 1 - e^{-\lambda t^\gamma} \\ f(t) &= \lambda\gamma t^{\gamma-1} e^{-\lambda t^\gamma} \\ r(t) &= \frac{f(t)}{S(t)} = \lambda\gamma t^{\gamma-1} \end{aligned}$$

Die Exponentialverteilung kann als Ausprägung mit dem Parameter $\gamma = 1$ angesehen werden.

$$\mathrm{WB}\,(\lambda, 1) = \mathrm{Exp}\,(\lambda)$$

[2]Der Vollständigkeit: γ_2 gibt die Wölbung an.



(a) $\gamma = \frac{3}{2}$ (b) $\gamma = 2$ (c) $\gamma = \frac{5}{2}$

Abbildung 2.0.2: Die Änderungsrate für die Weibull Verteilung $r(t) = \lambda\gamma t^{\gamma-1}$ mit Parameter $\lambda = \frac{1}{2}$ im Bereich $0 < t < 2$ verhält sich wie die Potenzfunktion. Die Graphiken zeigen $\gamma \in \left\{\frac{3}{2}, 2, \frac{5}{2}\right\}$

2.0.4 Prognose Studien

2.0.4.1 Kaplan-Meier-Methode

Abschnitt 16.2.3 (S. 307 – 309)

Es wird für die unbekannte Funktion $S(t)$ eine geschätzte Funktion $\hat{S}(t)$ gesucht.

n	...	Anzahl der Personen in der Studie
$t_1 < t_2 < \cdots < t_k$	...	Zeitpunkte der kritischen Endereignisse
$t_0 := 0$	...	Beginn des Beobachtungszeitraumes
$t_{k+1} := t_\omega$	...	Ende des Beobachtungszeitraumes
c_i	...	Anzahl der zensierten Beobachtungen in $[t_{i-1}, t_i)$
d_i	...	Anzahl der Endereignisse zum Zeitpunkt t_i
n_i	...	Anzahl der Personen die unmittelbar vor t_i beobachtet werden

Daher ist

$$n_i = n_{i+1} - d_{i-1} - c_i \qquad \text{für } 1, \dots k$$

wobei definiert ist, dass

$$\begin{aligned} n_0 &:= n \\ d_0 &:= 0 \\ n_1 &= n - c_1 \end{aligned}$$

Daher gilt

$$\begin{aligned} \hat{S}(0) &= 1 \\ \hat{S}(t_i) &= \hat{S}(t_{i-1}) \cdot \frac{n_i - d_i}{n_i} \qquad \text{für } i = 1, 2, \ldots k \end{aligned}$$

Beispiel 16.1

Example 1. Nach einer Organtransplantation wurden bei 10 Patienten die Überlebenszeiten in Tagen ermittelt. Nach 160 Tagen wurde die Studie beendet. Bei 7 Patienten konnte der Zeitpunkt des Endereignisses ermittelt werden (nach 20, 35, 62, 91, 91, 128 und 148 Tagen). Ein Patient brach nach 98 Tagen die Studie ab; zwei Patienten lebten am Ende der Studie noch.

Aus dem Text schließt man $t_\omega = 160$ und $n = 10$. Desweiteren gibt es folgende Zeitpunkte mit kritischen Ereignissen: $t_1 = 20$, $t_2 = 35$, $t_3 = 62$, $t_4 = 92$, $t_5 = 128$ und $t_6 = 148$. Für alle Zeitpunkte gilt $d_i = 1$ außer für t_4 wo $d_4 = 2$ gilt, da an diesem Tag zwei Patienten verstorben sind.

t_i	c_i	n_i	d_i	$n_i - d_i$	$\hat{S}(t_i)$
0				10	1
20	0	10	1	9	$1 \cdot \frac{9}{10} = 0.9$
35	0	9	1	8	$0.9 \cdot \frac{8}{9} = 0.8$
62	0	8	1	7	$0.8 \cdot \frac{7}{8} = 0.7$
91	0	7	2	5	$0.7 \cdot \frac{5}{7} = 0.5$
128	1	4	1	3	$0.5 \cdot \frac{3}{4} = 0.375$
148	0	3	1	2	$0.375 \cdot \frac{2}{3} = 0.25$
160	0	2			

2.0.4.2 Logrank-Test

Abschnitt 12.2.7 (S. 242 – 243)

- Zum Vergleich (von Verteilungen) von Überlebenszeiten.

$$S_1(t) \qquad S_2(t)$$

- bei zwei unverbundenen Stichproben

- mit folgenden Hypothesen

$$\begin{array}{rlll} \text{Null Hyptohese} & N_0: & S_1(x) \equiv S_2(x) & \forall x: S_1(x) = S_2(x) \\ \text{Alternativ Hypothese} & N_1: & S_1(x) \not\equiv S_2(x) & \exists x: S_1(x) \neq S_2(x) \end{array}$$

Stichproben sind $j = 1, 2$ (manchmal auch $j = A, B$)
Die Zeitpunke sind mit t_i bezeichnet, wobei $i = 1, 2, \ldots k$ sowie $t_0 := 0$ gilt.

$$\begin{array}{rcl} d_{j,i} & \ldots & \text{Anzahl der Endereignisse bei der Stichprobe } j \text{ zum Zeitpunkt } i \\ d_j := \sum_{i=1}^{k} d_{j,i} & & j = 1, 2 \\ c_{j,i} & \ldots & \text{Anzahl zensierter Daten in Stichprobe } j \text{ zum Zeitpunkt } [t_{i-1}, t_i) \\ n_{j,0} := n_j & \ldots & \text{Anzahl der Beobachtungseinheiten bei Stichprobe } j \\ n_{j,i} := n_{j,i-1} - d_{j,i-1} - c_{j,i} & \text{bzw.} & n_{j,i} - d_{j,i} - c_{j,i+1} =: n_{j,i+1} \end{array}$$

$$\begin{aligned} e_j &:= \sum_{i=1}^{k} (d_{1,i} + d_{2,i}) \cdot \frac{n_{j,i}}{n_{1,i} + n_{2,i}} \quad \text{für } j = 1, 2 \\ e_1 + e_2 &= \sum_{i=1}^{k} (d_{1,i} + d_{2,i}) \cdot \frac{n_{1,i}}{n_{1,i} + n_{2,i}} + \sum_{i=1}^{k} (d_{1,i} + d_{2,i}) \cdot \frac{n_{2,i}}{n_{1,i} + n_{2,i}} = \\ &= \sum_{i=1}^{k} (d_{1,i} + d_{2,i}) \cdot \left[\frac{n_{1,i} + n_{2,i}}{n_{1,i} + n_{2,i}} \right] = d_1 + d_2 \\ e_2 &= d_1 + d_2 - e_1 \end{aligned}$$

Die zugehörige χ^2-Verteilung (mit einem Freiheitsgrad) ist somit definiert als

$$\chi^2 := \frac{(d_1 - e_1)^2}{e_1} + \frac{(d_2 - e_2)^2}{e_2}$$

Die Nullhypothese $H_0 : S_1(x) \equiv S_2(x)$ wird abgelehnt, falls $\chi^2 > \chi^2_{1,1-\alpha}$

Example 2. Gegeben sind zwei Studien $j = A, B$ mit $t_\omega = 160$

Therapie A $n_A = 10$
Zeiten der Endereignisse: $20, 35, 62, 91, 91, 128, 148$
Zeiten mit Abbruch: 98

Therapie B $n_B = 15$
Zeiten der Endereignisse: $1, 11, 3, 3, 20, 91$
Zeiten mit Abbruch: $2, 3, 15, 29, 29, 120$

t_i	$c_{A,i}$ $c_{B,i}$	$n_{A,i}$ $n_{B,i}$		$d_{A,i}$ $d_{B,i}$	$n_{A,i} \cdot d_{A,i}$ $n_{B,i} \cdot d_{B,i}$	$\hat{S}_A(t_i)$ $\hat{S}_B(t_i)$
0					10 15	
1	0 0	10 15	25	0 3	10 12	
3	0 1	10 11	21	0 2	10 9	
20	0 2	10 7	17	1 1	9 8	
35	0 2	9 4	13	1 0	8 4	
62	0 0	8 4	12	1 0	7 4	
91	0 0	7 4	11	2 1	5 3	
128	1 1	4 2	6	1 0	3 2	
148	0 0	3 2	5	1 0	2 2	
160	0 0	2 2	4	0 0		

Es ergibt sich aus der Tabelle als Summe

$$d_A = 7 \qquad \text{und} \qquad d_B = 7$$

$$\begin{aligned} e_A &= 3 \cdot \frac{10}{25} + 2 \cdot \frac{10}{21} + 2 \cdot \frac{10}{17} + 1 \cdot \frac{9}{13} + 1 \cdot \frac{8}{12} + 3 \cdot \frac{7}{11} + 1 \cdot \frac{4}{6} + 1 \cdot \frac{3}{5} = \\ &= 7.863\,583\,475 \end{aligned}$$

$$\begin{aligned} e_B &= 14 - e_A = 6.136\,416\,525 \\ \chi^2 &= 0.216\,372\,145 \end{aligned}$$

χ^2 Tabelle S. 322

$$\begin{aligned} \alpha = 10\% \quad & \chi^2_{1,0.9} = 2.706 > \chi^2 \quad \Rightarrow \text{Testentscheidung } H_0 \\ \alpha = 5\% \quad & \chi^2_{1,0.95} = 3.841 > \chi^2 \quad \Rightarrow \text{Testentscheidung } H_0 \\ \alpha = 1\% \quad & \chi^2_{1,0.99} = 6.635 > \chi^2 \quad \Rightarrow \text{Testentscheidung } H_0 \end{aligned}$$

3 Vorlesung am 18. März 2010

3.1 Wahrscheinlichkeiten in der Epidemiologie

Abschnitt 6.3 (S. 111)

Prävalenz genau gesagt, die Punktprävalenz gibt zu einem bestimmten Zeitpunkt folgendes an:

- Den relativen Krankenbestand
- Den Anteil der Erkrankten (einer bestimmten Krankheit) in der Population
- Die Wahrscheinlichkeit für eine beliebige Person *erkrankt zu sein*

Um die Prävalenz zu ermitteln werden oft Querschnittstudien verwendet.

Periodenprävalenz gibt an am Anfang, während, oder am Ende des Beobachtungszeitraumes erkrankt zu sein.

Für den Rest gilt dasselbe wie für die Prävalenz oben.

Inzidenz

- Die Wahrscheinlichkeit für eine beliebige Person während des Beobachtungszeitraumes neu *zu erkranken.*
- Falls der Beobachtunszeitraum eine Zeiteinheit (ein Jahr, ein Monat, ...) ist spricht man von der Neuerkrankungsrate.

$$\text{Prävalenz} = \text{Inzidenz} \cdot \text{durchschnittliche Krankheitsdauer}$$

Die Prävalenz wird mit $\mathrm{P}\left(K_T\right)$ bezeichnet und die Inzidenz mit $\mathrm{P}\left(K\right)$ in der Epidemiologie jedoch herrscht die Bezeichnung $\mathrm{P}\left(K\right)$ für die Prävalenz vor. Diese Bezeichnung wird auch im Buch und in Folge verwendet.

(krankheitspezifische) Mortalität auch „Todesrate“ genannt

- $\mathrm{P}\left(K \cap T\right)$
- Die Wahrscheinlichkeit (während des Beobachtungszeitraumes) an der Krankheit zu erkranken *und* zu sterben.

Letalität

- $\mathrm{P}\left(T \mid K\right)$
- Die Wahrscheinlichkeit (während des Beobachtungszeitraumes) zu sterben, falls das Individuum an der Krankheit K erkrankt ist.

- $\mathrm{P}(K \cap T) = \mathrm{P}(K) \cdot \mathrm{P}(T \mid K)$

$$\text{Mortalität} = \text{Inzidenz} \cdot \text{Letalität}$$

Kontragionindex

Manifestationsindex

3.1.1 Bevölkerungsstatistik

Abschnitt 6.4 (S. 114)

Natalität (auch: Geburtenziffer)

$$\frac{\|\text{Geborene Kinder pro Jahr}\|}{\text{Populationsgröße}}$$

Fertilitätsrate

$$\frac{\|\text{Lebend Geborene pro Jahr}\|}{\|\text{Frauen im gebärfähigen Alter}\|}$$

Sterbeziffer

$$\frac{\|\text{Gestorbene pro Jahr}\|}{\text{Populationsgröße}}$$

Pearl-Index

Anzahl der Frauen – von 100 Frauen, welche eine bestimmte Verhüttungsmethode anwenden – die in einem Jahr ungewollt schwanger werden.

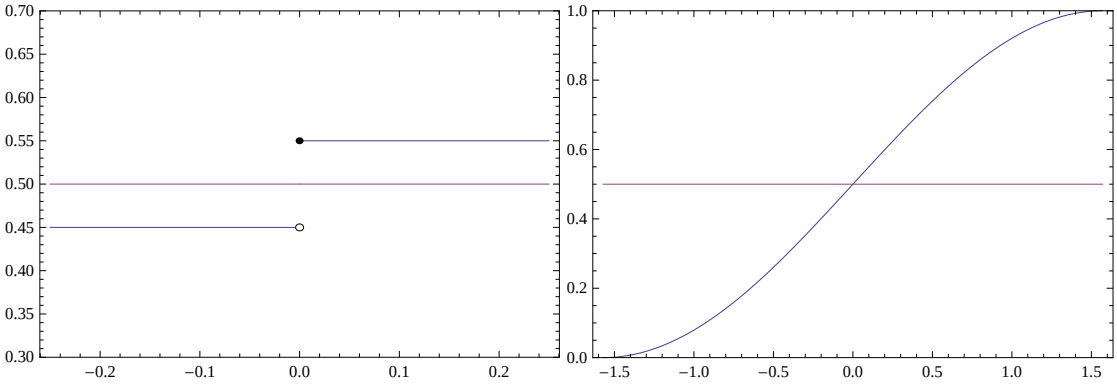
3.1.2 Diagnostische Tests

Abschnitt 6.5 (S. 119)

K	...	erkranktes Individuum
$\overline{K}$	...	gesundes Individuum
T_+	...	positives Testergebniss
T_-	...	negatives Testergebniss
$\mathrm{P}(T_+ \mid K)$	...	Sensitivität
$\mathrm{P}(T_- \mid \overline{K})$	...	Spezivität

3.1.2.1 Vorhersagewerte

$\mathrm{P}(K \mid T_+)$	...	positive Vorhersagewahrscheinlichkeit
$\mathrm{P}(\overline{K} \mid T_-)$	...	negative Vorhersagewahrscheinlichkeit



(a) diskrete Verteilungsfunktion (b) stetige Verteilungsfunktion

Abbildung 3.2.1: Median bei diskreten und stetigen Verteilungsfunktionen

$$\begin{aligned} \mathrm{P}(K \mid T_-) &= \frac{\mathrm{P}(K \cap T_-)}{\mathrm{P}(T_-)} = \frac{\mathrm{P}(K \cap T_-)}{\mathrm{P}(K \cap T_-) + \mathrm{P}(\overline{K} \cap T_-)} = \\ &= \frac{\mathrm{P}(K) \cdot \mathrm{P}(T_- | K)}{\mathrm{P}(K) \cdot \mathrm{P}(T_- \mid K) + \mathrm{P}(\overline{K}) \cdot \mathrm{P}(T_- \mid \overline{K})} \\ &= \frac{\mathrm{P}(K) \cdot [1 - \mathrm{P}(T+ \mid K)]}{\mathrm{P}(K) \cdot [1 - \mathrm{P}(T_+ \mid K)] + [1 - \mathrm{P}(K)] \cdot \mathrm{P}(T_- \mid \overline{K})} \end{aligned}$$

3.2 Diskrete Verteilungen

3.2.1 Diskrete Zufallsvariablen

Abschnitt 7.1 (S. 125)

Bezeichner

X	...	Zufallsvariable
x_i	...	Wert (Ausprägung)
p_i	...	Wahrscheinlichkeit für Wert x_i
$f(x_i) = p_i$	...	Dichtefunktion (Wahrscheinlichkeitsfunktion)

Erwartungswert

$$\begin{aligned}
\text{Verteilungsfunktion} \quad F(x) &= \mathrm{P}\left(X \leq x\right) \\
\text{Erwartungswert} \quad \mathbb{E}\, X &= \mu \\
\mathbb{E}\left(a \cdot X + b\right) &= a \cdot \mathbb{E}\, X + b \\
\mathbb{E}\left(\sum_i X_i\right) &= \sum_i \mathbb{E}\, X_i
\end{aligned}$$

Median und Quantill

$$\begin{aligned}
\text{Median} \quad \tilde{\mu} &= \min\left\{x \mid F(x) \geq 0.5\right\} \\
\alpha\text{-Quantill} \quad \tilde{\mu}_\alpha &= \min\left\{x \mid F(x) \geq \alpha\right\}
\end{aligned}$$

$\tilde{\mu}$ wird immer als 0.5-Quantill bezeichnet: $\tilde{\mu} = \tilde{\mu}_{0.5}$

Varianz und Standardabweichung

$$\begin{aligned}
\operatorname{Var} X = \sigma^2 &= \mathbb{E}\left(\left[X - \mu\right]^2\right) = \mathbb{E}\, X^2 - \left(\mathbb{E}\, X\right)^2 \\
&= \sum_i x_i^2 p_i - \left(\sum_i x_i p_i\right)^2 \\
\text{Standardabweichung} \quad \sigma &= \sqrt{\operatorname{Var} X} \\
\text{Variationskoeffizient} &= \frac{\sigma}{\mu}
\end{aligned}$$

Für das Rechnen mit Varianzen gilt,

$$\begin{aligned}
\operatorname{Var}\left(a \cdot X + b\right) &= a^2 \cdot \operatorname{Var} X \\
\operatorname{Var}\left(X + Y\right) &= \operatorname{Var} X + \operatorname{Var} Y + 2 \cdot \operatorname{Cov}\left(X, Y\right)
\end{aligned}$$

Wenn $X_{1,} \ldots X_n$ unabhängige Zufallsvariablen sind, so gilt,

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var} X_i$$

Kovarianz Wenn X, Y unabhängig[1] sind, so gilt für die Kovarianz

$$\operatorname{Cov}\left(X, Y\right) = 0$$

[1] Eigentlich muss es unkorreliert heißen, da es exotische Ausnahmen geben kann, wo dies für unabhängige Zufallsvariablen nicht nicht gilt.

3.2.2 Binomialverteilung

Abschnitt 7.2

3.2.2.1 Bernoulli-Experiment

Abschnitt 7.2.1

Das Bernoulli-Experiment geht davon aus, dass jeder durchgeführte Versuch zwei Ausgänge (Ereignisse) haben kann (Diese sind hier mit A und $\overline{A}$ bezeichnet).

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(X=1) = p \\ \mathrm{P}(\overline{A}) &= \mathrm{P}(X=0) = q = 1-p \end{aligned}$$

3.2.2.2 Bernoulli-Prozess

Abschnitt 7.2.2

$$\begin{aligned} X &\sim \mathrm{B}(n,p) \\ \mathbb{E}\,X &= n \cdot p \\ \mathrm{Var}\,X &= n \cdot p \cdot q \\ \mathrm{P}(X=k) &= \binom{n}{k} \cdot p^k \cdot q^{n-k} \quad \text{für } k = 0, \ldots n \\ &= \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot q^{n-k} \\ \text{Schiefe} \quad \gamma_1 &= \frac{q-p}{\sigma} \qquad \text{wenn} \begin{cases} > 0 & \text{rechtsscheif} \\ < 0 & \text{linksschief} \end{cases} \end{aligned}$$

Symetrische Binomialverteilung

$$\begin{aligned} p &= q = \frac{1}{2} \\ \mathbb{E}\,X &= \frac{n}{2} \\ \mathrm{Var}\,X &= \frac{n}{4} \end{aligned}$$

3.2.3 Poissonverteilung

Abschnitt 7.3.1

Die Poissonverteilung ist ein Spezialfall der Binomialverteilung, für die folgendes gilt:

$$\begin{aligned} \mathbb{E}\,X &= n \cdot p = \lambda \text{ (const)} \\ n &\to \infty \end{aligned}$$

In der Praxis wird die Poissonverteilung als Annäherung für die Binomialverteilung verwendet:

$$\mathrm{B}(n,p) \to \mathrm{P}(\lambda) \qquad \text{wenn } n \geq 30 \wedge p \leq 0.1$$

Es gilt also

$$X \sim \mathrm{P}(\lambda)$$

unter der Bedingung,

$$\begin{aligned}
\mathbb{E}\,X &= \lim_{n\to\infty} n \cdot p = \lim_{n\to\infty} \lambda = \lambda \\
\operatorname{Var} X &= \lim_{n\to\infty} n \cdot p \cdot (1-p) = \lim_{n\to\infty} \lambda \cdot \left(1 - \frac{1}{n}\right) = \lambda \\
\gamma_1 &= \lim_{n\to 0} \frac{(1-p)}{\sqrt{n \cdot p \cdot (1-p)}} = \lim_{n\to 0} \frac{1 - 2\frac{\lambda}{n}}{\sqrt{n \cdot p \cdot (1-p)}} = \frac{1}{\sqrt{\lambda}} > 0
\end{aligned}$$

Die Possionverteilung definiert sich also als

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

4 Vorlesung am 25. März 2010

Beispiel 7.5 (S. 137)

Example 3. In einer Geburtsklinik werden jährlich $n = 2\,000$ Kinder geboren. Die Wahrscheinlichkeit, dass ein Neugeborenes mit einem Down-Syndrom zur Welt kommt, beträgt $p = 0.001$. Unter der Annahme, dass die Ereignisse unabhängig sind, lässt sich die Anzahl der Neugeborenen mit Down-Syndrom durch eine Poisson-verteilte Zufallsvariable X beschreiben. Für die charakteristischen Parameter gilt: $\lambda = n \cdot p = 2\,000 \cdot 0.001 = 2$.

Zur Approximation der Binomialverteilung durch die Poissonverteilung. Wenn X Binomial verteilt ist,

$$X \sim \mathrm{B}\left(n, \tfrac{\lambda}{n}\right) \qquad \text{wobei} \quad \tfrac{\lambda}{n} = p \text{ (const)}$$

Aus der Definition der Binomialverteilung ergibt sich:

$$\begin{aligned} \mathrm{P}(X = k) &= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{n!}{k! \cdot (n-k)!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} = \\ &= \frac{1}{k!} \cdot \underbrace{\frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}}_{\text{A}} \cdot \lambda^k \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{n}}_{\text{B}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\text{C}} = \end{aligned}$$

Diese Gleichung lässt sich vereinfachen, da der Grenzwert einiger Teilterme 1 ist. Der Übersicht genüge wird dies in mehreren Schritten dargestellt.

$$\begin{aligned} \text{A} \qquad & \lim_{n\to\infty} \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k} = \\ &= \lim_{n\to\infty} \underbrace{1 \cdot \overbrace{\left(1 - \frac{1}{n}\right)}^{\to 1} \cdots \overbrace{\left(1 - \frac{k-1}{n}\right)}^{\to 1}}_{k\text{-mal}} = 1 \end{aligned}$$

Da es im letzten Term nur k Faktoren von annähern 1 gibt, ergibt sich der Gesamtgrenzwert von 1.

$$\begin{aligned} \text{B} \qquad & \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \\ = \quad & \lim_{n\to\infty} \left(1 + \frac{-\lambda}{n}\right)^n = e^{-\lambda} \end{aligned}$$

$$\text{C} \qquad \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1^{-k} = 1$$

Folglich ist

$$\lim_{n\to\infty} \mathrm{P}\left(X = k\right) = \frac{1}{k!} \cdot 1 \cdot \lambda^k \cdot e^{-\lambda} \cdot 1 = \frac{1}{k!}\lambda^k e^{-\lambda}$$

Die Binomialverteilung kann also durch die Poissonverteilung für hinreichend große n und konstante p approximiert werden.

$$\mathrm{B}\left(n, p\right) \underset{\substack{n\to\infty \\ \frac{\lambda}{n} = p \text{ const}}}{\longrightarrow} \mathrm{P}(\lambda)$$

4.0.4 Polynomialverteilung

Abschnitt 7.3.2

Bei der Polynomialverteilung (oder auch Multinomialverteilung) werden n Zufallsexperimente mit k möglichen Ereignissen $A_1, A_2, \ldots A_k$ mit den zugehörigen Wahrscheinlichkeiten $p_1, p_2, \ldots p_k$ und Häufigkeiten $n_1, n_2, \ldots n_k$.

$$\mathrm{P}(n_1, n_2, \ldots n_k \mid p_1, p_2, \ldots p_k) = \frac{n!}{n_1! \cdots n_k!} \cdot p_1^{n_1} \cdot p_2^{n_2} \cdots p_k^{n_k}$$

Wobei gilt, dass

$$\sum_{i=1}^{n} n_i = n \quad \wedge \quad \sum_{i=1}^{n} p_i = 1 \quad \wedge \quad \forall p_i \geq 0 \quad \wedge \quad \forall n_i \in \{0, 1, 2, \ldots n\}$$

4.0.5 Negative Binomialverteilung

Abschnitt 7.3.3

Die Negative Binomialverteilung beschreibt die Wahrscheinlichkeit, dass bei einem Zufallsexperiment mit zwei Ausgängen ein Ereigniss A – mit der Wahrscheinlichkeit p – bei der j-ten Beobachtung zum r-ten Mal eintritt.

$$X \sim \mathrm{NB}\left(r, p\right)$$

4.0.5.1 Geometrische Verteilung

Die geometrische Verteilung ist ein Sonderfall der negativen Binomialverteilung mit $r = 1$.

$$X \sim \mathrm{NB}(1, p)$$

$\mathrm{P}(X = j)$ ist die Wahrscheinlichkeit, dass das Ereigniss bei der j-ten Beobachtung erstmals Eintritt.

$$\mathrm{P}(X = j) = q^{j-1}p \qquad \text{wobei } q = 1 - p \quad \wedge \quad j = 1, 2, 3, \ldots$$

4.0.5.2 Definition der negativen Binomialverteilung

$$\mathrm{P}(X = j) = \binom{j-1}{r-1} p^{r-1} q^{j-r} p \qquad \text{für } j = r, r+1, \ldots$$

Example 4. Eine Blutbank... Beispiel 7.8 (S. 139)

4.0.6 Hypergeometrische Verteilung

Abschnitt 7.3.4

Bei der Hypergeometrischen Verteilung nimmt man an, dass es bei einer Menge von N Objekten insgesamt M Objekte gibt, die eine gesuchte Eigenschaft A besitzen. Der Rest hat diese Eigenschaft nicht ($\overline{A}$). Der Anteilswert $p = \frac{M}{N}$ gibt an, wie groß der Anteil der gesuchten Objekte an der Gesamtmenge ist. Es werden nun zufällig n Objekte aus der Gesamtmenge zufällig ausgewählt und bezüglich ihrer Eigenschaft untersucht.

$$X \sim \mathrm{HG}(n; N, M)$$

Die Zufallsvariable X gibt nun an, wie häufig das gesuchte Merkmal in den gewählten n Beobachtungen auftritt.

$$\begin{aligned} \mathrm{P}(X = k) &= \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}} \\ \mathbb{E}\,X &= n \cdot p = n \cdot \frac{M}{N} \\ \mathrm{Var}\,X &= \frac{N-n}{N-1} \cdot n \cdot p \cdot (p-1) \end{aligned}$$

Bei der Varianz wird der Bruch $\frac{N-n}{N-1}$ als Endlichkeitskorrektur bezeichnet.

4.0.6.1 Annäherung durch die Binomialverteilung

Allgemien gesagt, ist die Approximation möglich, wenn n im Vergleich zu N sehr groß ist.

$$\begin{aligned} n = 1 \quad &\Rightarrow \quad \frac{N-n}{N-1} = 1 \\ N \to \infty \quad &\Rightarrow \quad \frac{N-n}{N-1} = \frac{1-\frac{n}{N}}{1-\frac{1}{N}} \to 1 \end{aligned}$$

4.0.7 Übersicht über die diskreten Verteilungen

Übersicht 6 (S. 142)

4.1 Stetige Verteilungen

Abschnitt 8

Die Schiefe der Verteilungsfunktion

$$\gamma_1 = \frac{\mathbb{E}\,(X - \mathbb{E}\,X)^3}{\sigma^3}$$

Die Wölbung, Krümmung der Verteilungsfunktion

$$\gamma_2 = \frac{\mathbb{E}\,(X - \mathbb{E}\,X)^4}{\sigma^4} - 3$$

Wobei für die Standardnormalverteilung gilt,

$$\gamma_1 = 0, \quad \gamma_2 = 0$$

4.1.1 Normalverteilung

Abschnitt 8.2.1

$$\begin{aligned} X &\sim \mathrm{N}\,(\mu, \sigma^2) \\ f(x) &= \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi} \cdot \sigma} \end{aligned}$$

Wobei μ den Mittelwert beschreibt, und σ^2 die Varianz, das Quadrat der Standardabweichung σ.

4.1.1.1 Standardnormalverteilung

Abschnitt 8.2.2

Die Standardnormalverteilung ist eine spezielle Form der Normalverteilung mit den Parametern

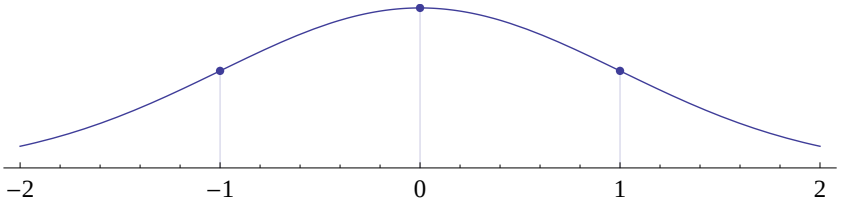


Abbildung 4.1.1: Dichtefunktion der Normalverteilung $X \sim \mathrm{N}(0,1)$, auch Standardnomalverteilung genannt. Das Maximum liegt bei 0, $\frac{1}{\sqrt{2\pi}\cdot\sigma}$. Die Wendepunkte sind bei $\mu - \sigma = -1$ und $\mu + \sigma = 1$.

$$\mu = 0 \quad \text{und} \quad \sigma^2 = 1$$

Die Zufallsvariable wird zur besseren Unterscheidung, mit Z bezeichnet.

$$Z \sim \mathrm{N}(0,1)$$

Es können alle Normalverteilungen auf die Standardnormalverteilung transformiert werden, so wird nur eine Nachschlagetabelle (Tabelle A auf Seite 317) benötigt.

$$\frac{X-\mu}{\sigma} := Z \sim \mathrm{N}(0,1)$$

Die Dichtefunktion wird ebenso nicht mit F sondern mit Φ bezeichnet.

$$\Phi(z) = \mathrm{P}(Z \leq z)$$

Example 5. Die Körpergröße einer Population... Beispiel 8.1 (S. 150)

$$\begin{aligned} \mathrm{P}(x_1 \leq X \leq x_2) &\sim \mathrm{N}(\mu, \sigma^2) \\ \mathrm{P}\left(\frac{x_1-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{x_2-\mu}{\sigma}\right) &\sim \mathrm{N}(0,1) \end{aligned}$$

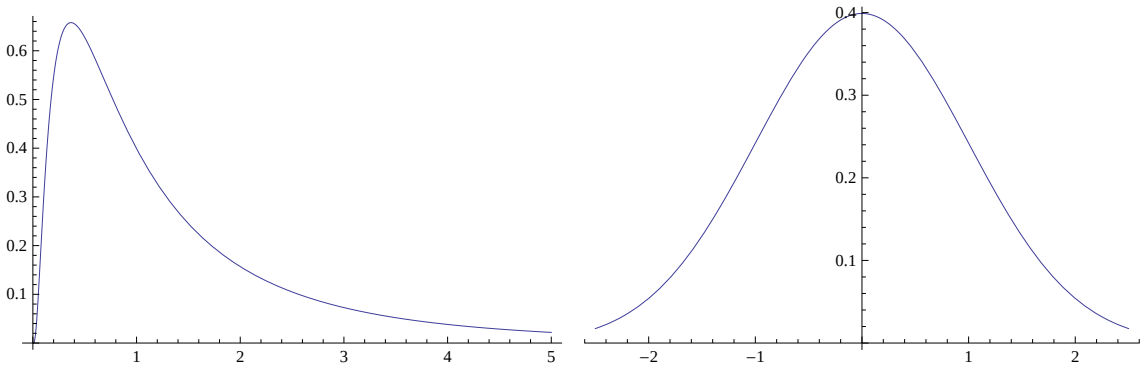
4.1.1.2 σ-Bereiche und Referenzbereiche

Abschnitt 8.2.3

Tabellen für Werte der Φ-Funktion konzentrieren sich auf den Bereich $\pm 3\sigma$ um μ herum, da innerhalb dieses Bereichs 99.7% aller Werte sind. Für medizinische Anwendungen sind vorallem Referenzberiche interresant, das sind bereiche, in denne genau 95% und 99% aller Werte liegen. Tabelle 8.1 (S. 152)

Bezeichung	Intervall	P
1σ-Bereich	$\mu - \sigma \leq X \leq \mu + \sigma$	0.6827
2σ-Bereich	$\mu - 2\sigma \leq X \leq \mu + 2\sigma$	0.9545
3σ-Bereich	$\mu - 3\sigma \leq X \leq \mu + 3\sigma$	0.9973
95%-Referenzbereich	$\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma$	0.95
99%-Referenzbereich	$\mu - 2.58\sigma \leq X \leq \mu + 2.58\sigma$	0.99

Tabelle 4.1: σ-Bereiche und Referenzbereiche



(a) $X \sim$ Nicht-Normal (b) $Y \sim \ln X$

Abbildung 4.1.2: Normalisierende Transformation

4.1.2 Normalisierende Transformationen

Abschnitt 8.2.4

Nicht alle Zufallsdaten für, die für eine Verteilung zugrunde liegen sind Normalverteilt. Daher muss die Verteilung transformiert werden, damit die Merkmale für die Normalverteilung geeignet sind.

In medizinischen Anwendungen besitzen viele Daten von Natur aus eine Rechtsschiefeverteilung, sie müssen mithilfe einer logarithmischen Transformation behandelt werden.

4.1.2.1 Lognormalverteilung

Die Logarithmische Transformation kann nur für $X > 0$ verwendet werden, da sonst das Ergebins komplex wird, bzw. für $X = 0$ nicht definiert ist, da es asymtotisch gegen negativ unendlich geht.

$$\begin{aligned} X \sim \mathrm{LN}(\mu, \sigma^2) &\Leftrightarrow Y \sim \mathrm{N}(\mu, \sigma^2) \\ X = e^Y &\Leftrightarrow Y = \ln X \end{aligned}$$

Daraus folgt:

$$\begin{aligned} \mathrm{P}(X \leq d) &= \mathrm{P}(\ln X \leq \ln d) = \mathrm{P}(Y \leq \ln d) = \Phi\left(\frac{\ln d - \mu}{\sigma}\right) \\ \mathrm{P}(Y \leq c) &= \mathrm{P}(e^Y \leq e^c) = \mathrm{P}(X \leq e^c) = \Phi\left(\frac{c - \mu}{\sigma}\right) \end{aligned}$$

Example 6. Sei, $\tilde{\mu}_X$ der Median von X und $\tilde{\mu}_Y$ der Median von Y.

$$\begin{aligned} \tilde{\mu}_Y = \mu &\Rightarrow \\ \frac{1}{2} = \mathrm{P}(Y \leq \tilde{\mu}_Y) &= \mathrm{P}(Y \leq \mu) = \\ = \mathrm{P}(X \leq e^{\mu}) &= \mathrm{P}(e^Y \leq e^{\mu}) \\ &\Rightarrow \tilde{\mu}_X = e^{\mu} \end{aligned}$$

Example 7. Die Konzentration von Serum IgM bei Kindern… Beispiel 8.3 (S. 154)

Anmerkungen zur Lognormalverteilung Bei extrem rechtsschiefen Verteilungen reicht oft das einfache Logarithmieren nicht aus. Solche Verteilungen müssen Doppellogarithmiert werden.

Grundsätzlich gilt also:

$$Y = \ln(X + a)$$

Wobei a der Versatz ist, der verwendet werden muss, damit die Werte im Bereich > 0 liegen.

Bei linksschiefen Verteilungsdichtefunktionen wir stattdessen potenziert.

$$Y = X^a \qquad \text{mit } a > 1$$

5 Vorlesung am 15. April 2010

5.0.3 Sätze der Wahrscheinlichkeitsrechnung

5.0.3.1 Tschebyscheff'sche Ungleichung

Abschnitt 8.3.1

Die Tschebyscheff'sche Ungleichung[1] dient zur Abschätzung der Wahrscheinlichkeit, dass die Zufallsvariable X außerhalb des Bereichs $k \cdot \sigma \pm \mu$. Wobei $k > 0$.

$$\mathrm{P}\left(|X - \mu| > k \cdot \sigma\right) \leq \frac{1}{k^2}$$

5.0.3.2 Gesetz der großen Zahlen

Abschnitt 8.3.2

Grundaussage des Gesetzes ist, dass bei einem größeren Stichprobenumfang die Schätzung des Erwartungswerts durch den Mittelwert genauer wird.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \forall X_i : \mathbb{E}\, X_i = \mu \ \wedge\ \mathrm{Var}\, X_i = \sigma^2$$

Nun gilt für den Erwartungswert und für die Varianz:

[1] Auch: Tschebescheff'sche Ungleichung, Tschebyschow-Ungleichung, eng.: Chebyshev, rus.: Чебышёв

$\mathrm{P}\left(\lvert X - \mu\rvert > k\sigma\right)$	$> \sigma$	$> 2\sigma$	$> 3\sigma$
allgemeine Verteilung		$\leq \frac{1}{4} = 0.25$	$\leq \frac{1}{9} \approx 0.11$
eingipfelige, symmetrische Verteilung		$\leq \frac{1}{9} \approx 0.11$	$\leq \frac{4}{81} \approx 0.049$
Gauß-Verteilung	$= 0.32 \approx \frac{1}{3}$	$= 0.046$	$= 0.003$

Tabelle 5.1: Anwendung der Tschebyscheff'schen Ungleichung

$$\begin{aligned} \mathbb{E}\,\bar{X} &= \mu \\ \operatorname{Var}\bar{X} &= \frac{\sigma^2}{n} \\ \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

5.0.3.3 Zentraler Grenzwertsatz

Abschnitt 8.3.3

Der zentrale Grenzwertsatz sagt aus, dass – unter recht allgemeinen Bedingungen – die Summe einer großen Anzahl von Zufallsvariablen normalverteilt ist.

Sind alle X_i unabhängig identisch verteilt

$$X_i \qquad \text{für } i = 1, \ldots, n$$

dann haben Erwartungswert und Varianz die gleichen Werte für alle X_i

$$\mathbb{E}\,X_i = \mu \quad \wedge \quad \operatorname{Var} X_i = \sigma^2$$

und die Summe aller Zufallsvariablen ist – asymptotisch – normalverteilt.

$$\sum_{i=1}^{n} X_i \underset{\text{asym}}{\sim} \mathrm{N}\left(n\mu, n\sigma^2\right)$$

Durch Transformation auf die Standardnormalverteilung ergibt sich

$$Z_n = \frac{\sum X_i - n\mu}{\sqrt{n} \cdot \sigma} \underset{\text{asym}}{\sim} \mathrm{N}\left(0, 1\right)$$

Verteilung von Mittelwerten Es geht aus dem Gesetz hervor, dass wenn der Stichprobenumfang n hinreichend groß ist – bei realistischen Fällen $n \geq 25$ – der Mittelwert normalverteilt ist.

Binomialverteilung Die Binomialverteilung – für hinreichend großes n – kann ebenso durch eine Normalverteilung approximiert werden.

$$X \sim \mathrm{B}\left(n, p\right)$$

Hinreichend groß kann mit der Erfüllung folgender Ungleichung angenommen werden.

$$n \cdot p \cdot q \geq 9$$

Nun gilt – approximativ – als Ersatz für die Binomialverteilung

$$X \underset{\text{approx}}{\sim} \mathrm{N}(np, npq)$$

5.0.4 Prüfverteilungen

Abschnitt 8.5

5.0.4.1 t-Verteilung

Abschnitt 8.5.1

Voraussetzungen

1. $X_1, \ldots, X_n$ sind unabhängig
2. $\mathbb{E}\, X_i = \mu \quad \wedge \quad \operatorname{Var} X_i = \sigma^2$ für $i = 1, \ldots, n$
3. *alle X_i sind normalverteilt*

$$\bar{X} \sim \mathrm{N}\left(\mu, \tfrac{\sigma^2}{n}\right)$$

$$\begin{aligned} Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} &\sim \mathrm{N}(0,1) \\ T = \frac{\bar{X} - \mu}{S/\sqrt{n}} &\sim \mathrm{t}_{n-1} \end{aligned}$$

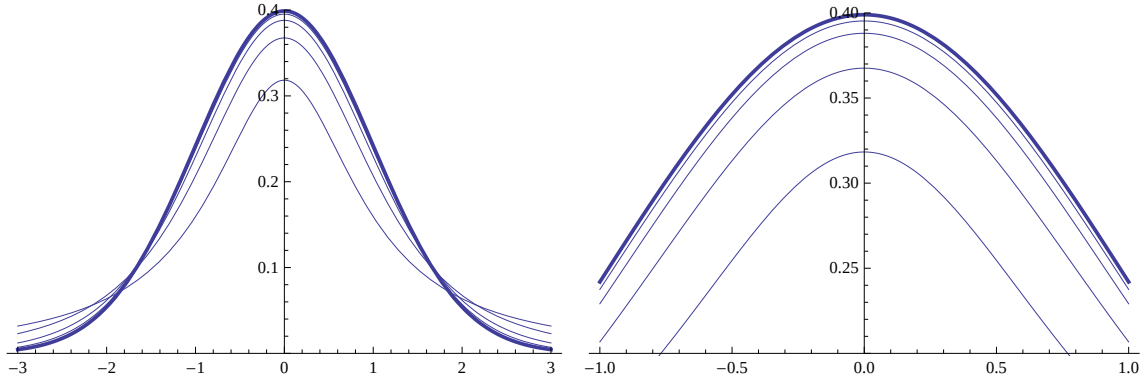
Der Index bei t gibt die Anzahl der Freiheitsgrade f der Verteilung an. Als Faustregel gilt hier, dass es immer $f = n - \|\text{Geschätzen Parameter}\|$ sind, da die geschätzten Parameter vorherbestimmt sind und sozusagen *Freiheit* wegnehmen.

Für die Schätzung der Standardabweichung S gilt

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

Für den Erwartungswert und die Varianz der t-Verteilung gilt nun:

$$\begin{aligned} T &\sim \mathrm{t}_f \\ \mathbb{E}\, T &= 0 \\ \operatorname{Var} T &= \tfrac{f}{f-2} \end{aligned}$$



(a) Im Bereich $-3 \leq x \leq +3$ (b) Vergrößerung für $-1 \leq x \leq +1$

Abbildung 5.0.1: Approximation der Standardnormalverteilung durch die t-Verteilung mit Freiheitsgraden $f \in \{1, 3, 9, 27\}$. Die Dichtefunktion der Normalverteilung ist dick eingezeichnet.

Die Varianz gilt insbesondere nur für $f \geq 3$ und kann daher immer als > 1 angenommen werden. Desweitern kann die t-Verteilung an die Standardnormalverteilung angenähert werden.

$$\mathrm{t}_f \underset{f\to\infty}{\longrightarrow} \mathrm{N}(0,1)$$

Ausgewähle γ-Quantille für die t-Verteilung können in der Tabelle B auf Seite 318 nachgeschlagen werden.

$$T \sim \mathrm{t}_f \Rightarrow \mathrm{P}(T \leq \mathrm{t}_{f,\gamma}) = \gamma$$

Example 8. Nachschlagen des Wertes ab dem 95% der Werte links von ihm liegen. (Suchen des 0.95-Quantills) bei einer t-Verteilung mit $f = 100$ und der Standardnormalverteilung.

$$0.95 = \mathrm{P}(T \leq \mathrm{t}_{100,0.95}) \Rightarrow \mathrm{t}_{100,0.95} = 1.660$$

$$0.95 = \mathrm{P}(Z \leq z_{0.95}) \Rightarrow z_{0.95} = 1.645$$

5.0.4.2 χ^2-Verteilung

Abschnitt 8.5.2

In der einfachsten Form – mit Freiheitsgrad $f = 1$ – beschreibt die χ^2-Verteilung die Verteilung des Quadrats einer standardnormalverteilten Zufallsvariablen $Z \sim \mathrm{N}(0,1)$. Nun gilt für Varianz und Erwartungswerte von $Z^2 \sim \chi^2_1$

$$\begin{aligned} \operatorname{Var} Z &= \mathbb{E}\, Z^2 - (\mathbb{E}\, Z)^2 \\ \mathbb{E}\, Z^2 &= \operatorname{Var} Z + (\mathbb{E}\, Z)^2 = 1 + 0 = 1 \end{aligned}$$

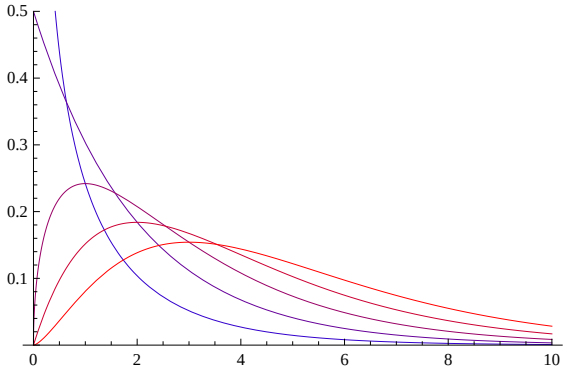


Abbildung 5.0.2: Dichtefunktion der χ^2-Verteilung mit Freiheitsgraden $1 \leq f \leq 5$

Unter den Vorraussetzungen

1. $Z_1, \ldots, Z_n$ sind unabhängig verteilt
2. $Z_i \sim \mathrm{N}(0,1) \quad \forall i = 1, \ldots, n$

gilt

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

mit

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^{n} Z_i^2\right) &= \sum \mathbb{E}\, Z_i^2 = \sum 1 = n \\ \mathrm{Var}\left(\sum_{i=1}^{n} Z_i^2\right) &= 2n \\ \gamma_1 &= \sqrt{\tfrac{8}{n}} \end{aligned}$$

Da die Schiefe γ_1 immer nur positiv sein kann – mit Annäherung an 0 bei $n \to \infty$ – ist die χ^2-Verteilung immer rechtsschief und wird mit steigendem n der Standardnormalverteilung immer ähnlicher.

Es gilt also, für Standardnormalverteilungen, für allgemeine Normalverteilungen $\mathrm{N}(\mu, \sigma^2)$ und für geschätzte Normalverteilungen mit Mittelwertschätzer $\bar{X}$.

$$\begin{aligned}\sum_{i=1}^{n} Z_i^2 &\sim \chi_n^2 \\ \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 &\sim \chi_n^2 \\ \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 &\sim \chi_{n-1}^2\end{aligned}$$

Da der Mittelwert $\bar{X}$ die Bedingungen der χ^2-Verteilung einschränkt hat die Verteilung nur $n-1$ Freiheitsgrade.

Für die Schätzung der Standardabweichung S gilt folgender Hinweis:

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{n-1}{n-1} \cdot \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \frac{n-1}{\sigma^2} \cdot S^2$$

$$\Rightarrow \quad \frac{n-1}{\sigma^2} \cdot S^2 \sim \chi_{n-1}^2$$

Spezielle Werte für γ-Quantillen von $\chi_{f,\gamma}^2$ sind in Tabelle E auf Seite 322 zu finden.

5.0.4.3 F-Verteilung

Abschnitt 8.5.3

Unter den Vorraussetzungen,

1. S_1^2, S_2^2 sind (empirische) Varianzen von zwei unabhängigen Stichproben, aus zwei normalverteilten Grundgesamtheiten mit derselben Varianz σ^2
2. n_1, n_2 sind die Stichprobenumfänge der zwei Stichproben

ist die F-Verteilung definiert als

$$\mathrm{F}_{n_1-1,n_2-1} := \frac{S_1^2}{S_2^2} = \mathrm{F}$$

Die F-Verteilung hat somit zwei Freiheitsgrade f_1, f_2 (manchmal auch als m, n bezeichnet).

$$f_1 = n_1 - 1 \qquad f_2 = n_2 - 1$$

Mit der vereinfachten Bezeichnung χ_f^2 für eine *Zufallsvariable*, die χ_f^2-verteilt ist, gilt auch

$$\mathrm{F}_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2} = \frac{\frac{\sigma^2}{n_1-1} \cdot \chi_{n_1-1}^2}{\frac{\sigma^2}{n_2-1} \cdot \chi_{n_2-1}^2} = \frac{\chi_{n_1-1}^2}{\chi_{n_2-1}^2} \cdot \frac{n_2-1}{n_1-1}$$

Beziehungsweise, mit alternativer Bezeichnung:

$$\mathrm{F}_{m,n} = \frac{\chi^2_m}{\chi^2_n} \cdot \frac{n}{m}$$

Im Speziellen gilt somit auch, das die t-Verteilung ein Speziallfall der F-Verteilung ist

$$\mathrm{F}_{1,n} = \mathrm{t}^2_n$$

5.1 Schätzverfahren

Abschnitt 9

5.1.1 Punktschätzungen

Abschnitt 9.2

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

5.1.1.1 Kriterien zur Güte einer Schätzung

Abschnitt 9.2.2

Erwartungstreue Schätzer sollten möglichst mit dem unbekannten Parameter übereinstimmen, so sollten Schätzvorschriften nicht systematisch einen zu hohen oder zu niedrigen Wert liefern. Eine erwartungstreue Schätzung nennt man unverzerrt.

Konsitenz Eine Schätzung sollte konsitent sein, dass heißt mit wachsender Stichprobengröße soll auch die Genauigkeit steigen. Für eine konsitente Schätzung gilt:

$$\operatorname{Var} X \underset{n\to\infty}{=} 0$$

Effizienz Für eine Schätzung ist eine möglichst kleine Varianz von großer Bedeutung. Eine Schätzung mit hoher Effizient liefert schon mit Stichproben geringeren Umfangs schon brauchbare Werte. Die Effizienz ist insbesondere dann wichtig, wenn zwei Schätzverfahren für einen Parameter miteinander verglichen werden sollen.

5.1.2 Spezielle Schätzfunktionen

Abschnitt 9.2.3

5.1.2.1 Erwartungswert

Der Mittelwert $\bar{X}$ schätzt den Erwartungswert μ der Grundgesamtheit. Es gilt daher

$$\begin{aligned} \mathbb{E}\,\bar{X} &= \mu \\ \operatorname{Var}\bar{X} &= \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

Diese Schätzung ist sowohl erwartungstreu und konsitent.

5.1.2.2 Median

Der Median $\tilde{X}$ ist ein erwartungstreuer Schätzer, falls die Verteilung stetig und symetrisch ist, in desem Fall gilt auch $\tilde{\mu} = \mu$. Daher ist bei der Normalverteilung der Median ein erwartungstreuer Schätzer für den Erwartungswert. Es gilt,

$$\operatorname{Var} \tilde{X} \quad = \quad \frac{\pi}{2} \cdot \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

Daher ist $\tilde{X}$ auch eine konsitente Schätzung, jedoch ist $\bar{X}$ ein effizineterer Schätzer für μ als $\tilde{X}$, da die Varianz größer ist.

5.1.2.3 Varianz

Herleitung von $\mathbb{E}\, S^2$

Nebenrechnung Zuerst wird gezeigt, dass

$$\begin{aligned}
\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 &= \sum \left(X_i - \mu\right)^2 - n(\bar{X} - \mu)^2 && (5.1.1)\\
\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 &= \sum \left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) \\
&= \sum X_i^2 - 2\left(\sum X_i\right)\bar{X} + n\bar{X}^2 \\
&= \sum X_i^2 - 2(n\bar{X})\bar{X} + n\bar{X}^2 \\
&= \sum X_i^2 - n\bar{X} && (5.1.2)\\
\sum_{i=1}^{n} \left(X_i - \mu\right)^2 &= \sum \left(X_i^2 - 2X_i\mu + \mu^2\right) \\
&= \sum X_i^2 - 2\left(\sum X_i\right)\mu + n\mu^2 \\
&= \sum X_i^2 - 2n\bar{X}\mu + n\mu^2 && (5.1.3)\\
-n\left(\bar{X} - \mu\right)^2 &= -n\bar{X}^2 + 2n\bar{X}\mu - n\mu^2 && (5.1.4)
\end{aligned}$$

Aus (5.1.2) sowie aus (5.1.3) und (5.1.4) ist ersichtlich, dass (5.1.1) stets stimmt.

Hauptrechnung Vorrausgesetzt wird

1. $\mathbb{E}\, X_i = \mu$
2. $\operatorname{Var} X_i = \sigma^2$
3. $X_1, \ldots, X_n$ sind unabhängig
4. Anmerkung: Normalverteilung ist nicht erforderlich

Daher ist,

$$\mathbb{E}\, \bar{X} = \mu \quad \wedge \quad \operatorname{Var} \bar{X} = \frac{\sigma^2}{n}$$

Allgemein gilt:

$$\sum \left(X_i - \bar{X}\right)^2 = \sum \left(X_i - \mu\right)^2 - n \left(\bar{X} - \mu\right)^2$$

somit ist

$$\begin{aligned} \mathbb{E}\, S^2 &= \mathbb{E}\left(\frac{1}{n-1} \cdot \sum \left(X_i - \bar{X}\right)^2\right) \\ &= \frac{1}{n-1} \cdot \mathbb{E}\left(\sum \left(X_i - \bar{X}\right)^2\right) \\ &= \frac{1}{n-1} \cdot \left[\mathbb{E}\left(\sum \left(X_i - \mu\right)^2\right) - n\mathbb{E}\left(\bar{X} - \mu\right)^2\right] \end{aligned} \tag{5.1.5}$$

Nun ist einerseits

$$\mathbb{E}\left(\sum \left(X_i - \mu\right)^2\right) = \sum \mathbb{E}\left(X_i - \mu\right)^2 = \sum \operatorname{Var} X_i = n\sigma^2 \tag{5.1.6}$$

Andererseits

$$\mathbb{E}\left(\bar{X} - \mu\right)^2 = \operatorname{Var} \bar{X} = \frac{\sigma^2}{n} \tag{5.1.7}$$

Aus (5.1.6) und (5.1.7) folgt, dass

$$\mathbb{E}\, S^2 = \frac{1}{n-1}\left[n\sigma^2 - \frac{n \cdot \sigma^2}{n}\right] = \sigma^2$$

daraus folgt

$$\mathbb{E}\, S^2 = \sigma^2$$

Dies ist ein erwartungstreuer Schätzer für σ^2.

Anmerkung: Aus (5.1.6) ist zu ersehen, dass $\frac{1}{n}\sum (X_i - \mu)^2$ eine erwartungstreue Schätzfunktion für σ^2 ist, soferne μ bekannt ist, denn

$$\mathbb{E}\left[\frac{1}{n}\cdot(X_i-\mu)^2\right] = \frac{\mathbb{E}\,(X_i-\mu)^2}{n} = \frac{n\cdot\sigma^2}{n} = \sigma^2$$

Alternative Herleitung von $\mathbb{E}\,S^2$ für unbekannte μ und bekannte $\bar{x}$, unter den Vorraussetzungen,

1. $X_i \sim \mathrm{N}(\mu, \sigma^2)$
2. $X_1, \ldots, X_n$ sind unabhängig

$$\frac{n-1}{\sigma^2}\cdot S^2 \sim \chi^2_{n-1}$$

Daraus folgt,

$$\mathbb{E}\left(\frac{n-1}{\sigma^2}\cdot S^2\right) = n-1$$

Daher gilt für den Erwartungswert:

$$\frac{n-1}{\sigma^2}\cdot\mathbb{E}\,S^2 = n-1 \quad\Leftrightarrow\quad \mathbb{E}\,S^2 = \sigma^2$$

Herleitung von $\mathrm{Var}\,S^2$ unter den Vorraussetzungen

1. $X_i \sim \mathrm{N}(\mu, \sigma^2)$
2. $X_1, \ldots, X_n$ sind unabhängig

$$\frac{n-1}{\sigma^2}\cdot S^2 \sim \chi^2_{n-1}$$

$$\begin{aligned}
\mathrm{Var}\left(\frac{n-1}{\sigma^2}\cdot S^2\right) &= 2(n-1)\\
\frac{(n-1)^2}{\sigma^4}\cdot\mathrm{Var}\,S^2 &= 2(n-1)\\
\mathrm{Var}\,S^2 &= \frac{2\sigma^4}{n-1} \underset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

Herleitung von $\mathbb{E}\, S < \sigma$ unter der Vorraussetzung, dass gilt

$$0 < \operatorname{Var} S = \mathbb{E}\, S^2 - (\mathbb{E}\, S)^2 = \sigma^2 - (\mathbb{E}\, S)^2$$

Diese Bedingung ist normalerweise bei allen realistischen Fällen erfüllt. Daraus folgt,

$$\begin{aligned} (\mathbb{E}\, S)^2 = \sigma^2 - \operatorname{Var} S &< \sigma^2 \\ \mathbb{E}\, S &< \sigma \end{aligned}$$

5.1.2.4 Schätzung der Wahrscheinlichkeit

Unter den Vorraussetzungen

1. $X_i \sim \mathrm{B}\,(1, p)$
2. $X_1, \ldots, X_n$ sind unabhängig

gilt,

$$\begin{aligned} X &:= \sum X_i \sim \mathrm{B}\,(n, p) \\ \hat{p} &:= \frac{X}{n} \end{aligned}$$

Wobei $\hat{p}$ in diesem Zusammenhang sowohl eine Zufallsvariable (so wie hier), als auch eine konkrete Zahl beschreiben kann.

$$\mathbb{E}\, \hat{p} = \frac{\mathbb{E}\, X}{n} = \frac{n \cdot p}{n} = p$$

Daher ist die Schätzung erwartungstreu, für die Varianz gilt,

$$\operatorname{Var} \hat{p} = \operatorname{Var} \left(\frac{X}{n} \right) = \frac{\operatorname{Var} X}{n^2} = \frac{n \cdot p \cdot (1-p)}{n^2} = \frac{p \cdot q}{n} \underset{n \to \infty}{\longrightarrow} 0$$

Daher ist die Schätzung auch konsitent.

6 Vorlesung am 22. April 2010

6.0.3 Intervallschätzungen

Abschnitt 9.3

6.0.3.1 Bedeutung eines Konfidenzintervalls

Abschnitt 9.3.1

Punktschätzungen sind oft ungenügend, da ein einzelner Wert keine Informationen enthält, wie sehr er vom tatsächlichen Parameter abweichen kann. Es können bei einer Schätzung natürlich keine genauen Angaben getroffen werden, jedoch ist es wünschenswert, dass sich der tatsächliche – unbekannte – Wert in der *näheren Umgebung* der Schätzung liegt.

Es wird aus den Daten der Stichprobe ein *Konfidenzintervall* konstruiert, in der Hoffnung, dass bei dem passend gewählten Verfahren, der gesuchte Parameter innerhalb des Intervalls liegt.

Irrtumswahrscheinlichkeit α ist die Wahrscheinlichkeit, dass der Parameter *nicht* im Intervall liegt.

Konfidenzwahrscheinlichkeit γ oder Konfidenzniveau bezeichnet die Wahrscheinlichkeit, dass der Parameter im Intervall liegt

$$\gamma = 1 - \alpha$$

6.0.3.2 Konfidenzintervall für einen Erwartungswert

Abschnitt 9.3.2

Herleitung für beliebige Konfidenzintervalle für μ Vorraussetzung

$$\begin{aligned} \bar{X} \sim \mathrm{N}\left(\mu, \tfrac{\sigma^2}{n}\right) \quad &\Rightarrow \quad \frac{\bar{X} - \mu}{\sigma^2/\sqrt{n}} \sim \mathrm{N}(0,1) \\ \mathrm{P}\left(z_{\alpha_1} \leq \frac{\bar{X} - \mu}{\sigma^2/\sqrt{n}} \leq z_{1-\alpha_2}\right) \quad &= \quad 1 - (\alpha_1 + \alpha_2) \end{aligned}$$

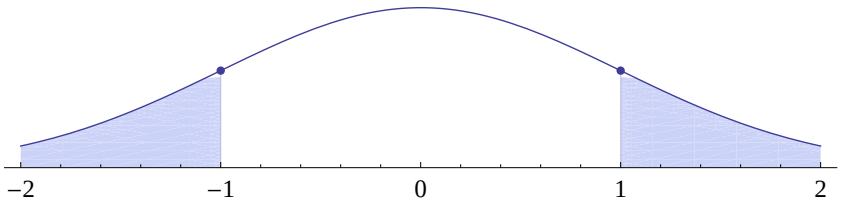


Abbildung 6.0.1: Intervallgrenzen, Links und Rechts eingezeichnet α_1 und α_2, Punkte zeigen jeweils z_{α_1} und $z_{1-\alpha_2}$

Wobei gilt:

$$z_0 = -\infty \qquad \wedge \qquad z_1 = +\infty$$

$$\alpha_1 + \alpha_2 = \alpha$$

Nun gilt ja:

$$\begin{aligned} z_{\alpha_1} \leq \quad & \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad & \leq z_{1-\alpha_2} \\ \frac{z_{\alpha_1} \cdot \sigma}{\sqrt{n}} \leq \quad & \bar{X} - \mu \quad & \leq \frac{z_{1-\alpha_2} \cdot \sigma}{\sqrt{n}} \\ \bar{X} - \frac{z_{\alpha_1} \cdot \sigma}{\sqrt{n}} \geq \quad & \mu \quad & \geq \bar{X} - \frac{z_{1-\alpha_2} \cdot \sigma}{\sqrt{n}} \end{aligned}$$

Daher

$$\mathrm{P}\left(\bar{X} - \frac{z_{1-\alpha_2} \cdot \sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{z_{\alpha_1} \cdot \sigma}{\sqrt{n}}\right) = 1 - \alpha$$

wobei $\alpha = \alpha_1 + \alpha_2$.

Konfidenzintervalle zu Konfidenzniveau $1 - (\alpha_1 + \alpha_2)$ sind allgemein alle Intervalle der Form

$$\left[\bar{x} - \frac{z_{1-\alpha_2} \cdot \sigma}{\sqrt{n}}; \bar{x} - \frac{z_{\alpha_1} \cdot \sigma}{\sqrt{n}}\right]$$

beziehungsweise, durch die Eigenschaft der Standardnormalverteilung: $z_{\alpha_1} = -z_{1-\alpha_1}$ kann das Intervall auch angescheiben werden als:

$$\left[\bar{x} - \frac{z_{1-\alpha_2} \cdot \sigma}{\sqrt{n}}; \bar{x} + \frac{z_{1-\alpha_1} \cdot \sigma}{\sqrt{n}}\right]$$

Für die Spezialfälle $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ ergibt sich

$$\left[\bar{x} - \frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\sqrt{n}}; \bar{x} + \frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\sqrt{n}}\right]$$

für $\alpha_1 = 0$ und $\alpha_2 = \alpha$ gilt

$$\left[\bar{x} - \frac{z_{1-\alpha} \cdot \sigma}{\sqrt{n}}; +\infty\right)$$

für $\alpha_1 = \alpha$ und $\alpha_2 = 0$ gilt

$$\left(-\infty; \bar{x} + \frac{z_{1-\alpha} \cdot \sigma}{\sqrt{n}}\right]$$

Grenzen zweiseitiger Konfidenzintervalle für μ unter der Vorraussetzung $\bar{X} \sim \mathrm{N}(\mu, \sigma^2)$ bzw. $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

Falls σ bekannt ist:

$$\bar{x} \pm \frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\sqrt{n}}$$

Falls σ unbekannt ist:

$$\bar{x} \pm \frac{\mathsf{t}_{n-1,1-\frac{\alpha}{2}} \cdot s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Wobei die Endlichkeitskorrektur $\sqrt{\frac{N-n}{N-1}}$ nur für Werte von $\frac{n}{N} > 0.01$ benötigt wird.

6.0.3.3 Konfidenzintervall für eine Wahrscheinlichkeit

Abschnitt 9.3.3

Unter der Vorraussetzung, dass X binomialverteilt ist,

$$X \sim \mathrm{B}(n, p)$$

ist die geschätzte Wahrscheinlichkeit definiert als,

$$\hat{p} := \frac{X}{n} \qquad \text{bzw.} \qquad \hat{p} := \frac{x}{n}$$

Weiters ist x approximativ normalverteilt

$$X \sim \mathrm{N}(np, npq) \qquad \text{für } npq \geq 9$$

daher

$$\frac{X}{n} \sim \mathrm{N}\left(p, \frac{p \cdot q}{n}\right)$$

Approximativ heißt dass,

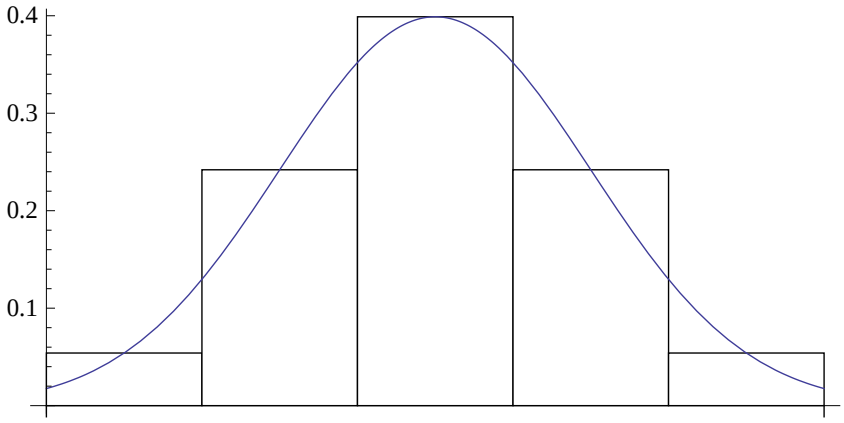


Abbildung 6.0.2: Annäherung der Binomialverteilung

$$\frac{X}{n} \sim \mathrm{N}\left(\hat{p}, \frac{\hat{p} \cdot \hat{q}}{n}\right)$$

wenn gilt,

$$n\hat{p} > 5 \quad \wedge \quad n\hat{q} > 5$$

Grenzen zweiseitiger Konfidenzintervalle für p

$$\hat{p} \pm \left[z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \cdot \sqrt{\frac{N-n}{N-1}} + \frac{1}{2n} \right]$$

Bei Bedarf wird die Endlichkeitskorrektur hinzugefügt, falls $\frac{n}{N} > 0.1$ und daher nicht vernachlässigbar ist.

Die Stetigkeitskorrektur $\frac{1}{2n}$ resultiert aus der Annäherung der Dichtefunktion mit über die Fläche. Siehe Abbildung 6.0.2.

6.0.4 Bedeutung des Stichprobenumfangs

Abschnitt 9.4.1

Die Breite des Konfidenzintervalls (BK) ist (etwa) indirektproportional zur Wurzel der Stichprobengröße n.

$$\mathrm{BK} = \frac{2 \cdot s \cdot \mathrm{t}_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}$$

$$n \geq 60: \quad \mathrm{t}_{n-1,0.975} \leq 2$$

Wenn die Breite des Konfidenzintervalls vorgegeben ist, gilt:

$$n = \frac{4 \cdot \mathrm{t}^2_{n-1,1-\frac{\alpha}{2}} \cdot s^2}{\mathrm{BK}^2} \approx \frac{16 \cdot s^2}{\mathrm{BK}^2}$$

Wenn $\alpha = 5\%$ für $n \gtrsim 60$

$$\mathrm{t}_{n-1,1-\frac{\alpha}{2}} \leq 2$$

7 Vorlesung am 29. April 2010

7.1 Das Prinzip eines statistischen Tests

Abschnitt 10

7.1.1 Durchführung eines Statistischen Tests

Abschnitt 10.1

H_0 Nullhypothese

H_1 Alternativhypothese

Sei nun μ_0 eine vorgegebene Zahl, so kann man nun unterscheiden

1. Einseitige Fragestellung

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu \neq \mu_0$$

2. Einseitige Fragestellung

   a)

$$\begin{aligned} H_0 : \mu = \mu_0 &\Leftrightarrow H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 &\Leftrightarrow H_1 : \mu < \mu_0 \end{aligned}$$

   b)

$$\begin{aligned} H_0 : \mu = \mu_0 &\Leftrightarrow H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 &\Leftrightarrow H_1 : \mu > \mu_0 \end{aligned}$$

In Kurzform $H_0 : \mu = \mu_0$

1. Zweiseitige Fragestellung

$$H_1 : \mu \neq \mu_0$$

2. Einseitige Fragestellung

   a)

$$H_1 : \mu < \mu_0$$

   b)

$$H_1 : \mu > \mu_0$$

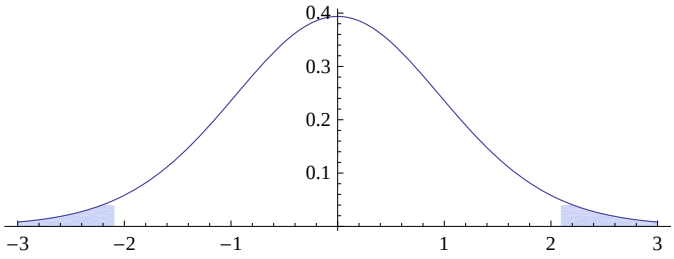


Abbildung 7.1.1: Verteilung von T beim Zutreffen von H_0, die beiden markierten Teile sind jeweils $\frac{\alpha}{2}$ in der Mitte befindet sich daher $(1-\alpha)$-Annahmebereich.

7.1.2 Arten von Fehlern

Fehler 1. Art α-Fehler

Fehler 2. Art β-Fehler

T ... Teststatistik
t ... Testgröße, Prüfgröße, Wert der Teststatistik

Angenommen die Teststatistik

$$T \sim \mathrm{t}_{n-1}$$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Fragestellung für Nullhypothese $H_0 : \mu = \mu_0$

1. Zweiseitige Fragestellung
$$H_1 : \mu \neq \mu_0$$
2. Einseitige Fragestellung
   a)
   $$H_1 : \mu < \mu_0$$
   b)
   $$H_1 : \mu > \mu_0$$

Kritischer Bereich

1. Zweiseitige Fragestellung

$$\left(-\infty; \mathrm{t}_{n-1,\frac{\alpha}{2}}\right) \cup \left(\mathrm{t}_{n-1,1-\frac{\alpha}{2}}; +\infty\right)$$

2. Einseitige Fragestellung

   a)
   $$(-\infty; \mathrm{t}_{n-1,\alpha})$$

   b)
   $$(\mathrm{t}_{n-1,1-\alpha}; +\infty)$$

Annahmebereich

1. Zweiseitige Fragestellung
   $$\left[\mathrm{t}_{n-1,\frac{\alpha}{2}}; \mathrm{t}_{n-1,1-\frac{\alpha}{2}}\right]$$

2. Einseitige Fragestellung

   a)
   $$[\mathrm{t}_{n-1,\alpha}; +\infty)$$

   b)
   $$(-\infty; \mathrm{t}_{n-1,1-\alpha}]$$

$$|t| > \mathrm{t}_{n-1,1-\frac{\alpha}{2}} \Leftrightarrow t < -\mathrm{t}_{n-1,1-\frac{\alpha}{2}} \vee t > \mathrm{t}_{n-1,1-\frac{\alpha}{2}}$$

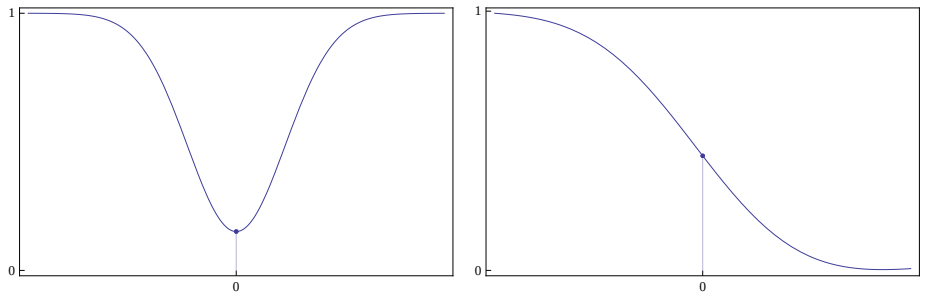
Falls H_0 zutrifft, aber – aufgrund der Stichprobenwerte – H_0 abgelehnt – und daher H_1 angenommen – wird, handelt es sich um einen Fehler 1. Art / α-Fehler.

Falls H_1 zutrifft, aber – aufgrund der Stichprobenwerte – H_1 nicht angenommen – und daher H_0 beibehalten – wird, handelt es sich um einen Fehler 2. Art / β-Fehler.

7.1.3 Testentscheidungen und Konsequenzen

Abschnitt 10.2

α	Ergebnis
10%	schwach signifikant
5%	signifikant
1%	hoch signifikant
1‰	höchst signifikant



(a) (b)

Abbildung 7.1.2: Gütefunktion $g(\mu)$ für Einseitige Fragestellungen (7.1.2a) und Zweiseitige Fragestellungen (7.1.2b) (Typ a).

Testentscheidung	Wirklichkeit			
	H_0 gilt		H_1 gilt	
für H_0	richtige Entscheidung	$1-\alpha$	Fehler 2. Art	β
für H_1	Fehler 1. Art	α	richtige Entscheidung	$1-\beta$
Summe	1		1	

Tabelle 7.1: Testentscheidungen und Fehler

$1-\beta$ . . . Güte, Teststärke, Trennschärfe, Macht, Power

Allgemein ist die Gütefunktion $g(\mu)$ die Wahrscheinlichkeit, H_0 abzulehnen, also H_1 anzunehmen, falls μ der tatsächlich (jedoch unbekannte) Parameterwert ist.

7.1.3.1 p-Wert und Konfidenzintervall

Abschnitt 10.2.2

$$\begin{aligned} p < \alpha &\Rightarrow \text{Testentscheidung } H_1 \\ p \geq \alpha &\Rightarrow \text{Testentscheidung } H_0 \end{aligned}$$

Interpretation des p-Werts Der p-Wert – zu einer aktuellen Stichprobe – ist jenes Signifikanzniveau, bei dem H_0 gerade noch beibehalten werden würde (maximaler α-Wert).
Oder, in einer Alternativen Definition, es ist jener Wert, bei dem H_1 gerade schon angenommen werden könnte (minimaler α-Wert).

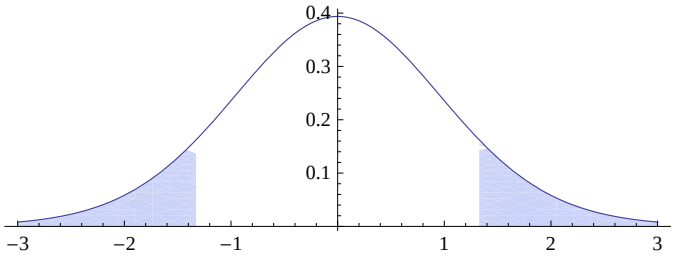


Abbildung 7.1.3: p_1 ist nur der linke Teil, p_2 beide Teile zusammen

Zusammenhang des p-Werts zu ein- und zweiseitigen Fragestellungen Bezeichnen wir mit

p_1 p-Wert eines einseitigen Tests – bei einseitiger Fragestellung

p_2 p-Wert des entsprechenden zweiseitigen Tests – bei entsprechender zweiseitiger Fragestellung

Dann gilt:

$$p_2 = 2 \cdot p_1 \qquad \text{bzw.} \qquad p_1 = \tfrac{p_2}{2}$$

Zusammenhang zwischen Testergebniss und Konfidenzintervall des t-Tests am Beispiel für eine Stichprobe

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \qquad \text{im Falle von } \mu = \mu_0$$

Die Wahrscheinlichkeit, dass $H_0 : \mu = \mu_0$ beibehalten wird, falls H_0 zutrifft – also tatsächlich $\mu = \mu_0$ ist, beträgt:

$$\mathrm{P}\left(\mathrm{t}_{n-1,\alpha_1} \leq T \leq t_{n-1,1-\alpha_2}\right) = 1 - (\alpha_1 + \alpha_2)$$

für den Fall dass $\mu = \mu_0$.

Wenn wir definieren $\alpha := \alpha_1 + \alpha_2$, dann gilt für

1. Zweiseitige Fragestellungen $\alpha_1, \alpha_2 = \frac{\alpha}{2}$
2. Einseitige Fragestellungen

   $H_1 : \mu < \mu_0 \; \alpha_1 = \alpha \quad \wedge \quad \alpha_2 = 0$

   $H_1 : \mu > \mu_0 \; \alpha_1 = 0 \quad \wedge \quad \alpha_2 = \alpha$

Es gilt

$$\begin{aligned}
\mathrm{t}_{n-1,\alpha_1} \leq \quad & T & \leq \mathrm{t}_{n-1,1-\alpha_2} \\
\mathrm{t}_{n-1,\alpha_1} \leq \quad & \tfrac{\bar{X}-\mu_0}{S/\sqrt{n}} & \leq \mathrm{t}_{n-1,1-\alpha_2} \\
\frac{\mathrm{t}_{n-1,\alpha_1} \cdot S}{\sqrt{n}} \leq \quad & \bar{X} - \mu_0 & \leq \frac{\mathrm{t}_{n-1,1-\alpha_2} \cdot S}{\sqrt{n}} \\
\bar{X} - \frac{\mathrm{t}_{n-1,\alpha_1} \cdot S}{\sqrt{n}} \geq \quad & \mu_0 & \geq \bar{X} - \frac{\mathrm{t}_{n-1,1-\alpha_2} \cdot S}{\sqrt{n}}
\end{aligned}$$

Daraus ergibt sich folgendes Intervall – mit Schreibweise $\mathrm{t}_{n-1,\alpha_1} = \mathrm{t}_{n-1,1-\alpha_1}$, da nur diese Werte sich in einer Tabelle finden.

$$\mu_0 \in \left[\bar{X} - \mathrm{t}_{n-1,1-\alpha_2} \cdot {}^{S}\!/\!\sqrt{n}; \bar{X} + \mathrm{t}_{n-1,1-\alpha_1} \cdot {}^{S}\!/\!\sqrt{n}\right]$$

$(1-\alpha)$-Konfidenzintervall μ_0 Somit gilt also, soferne die Nullhypothese $H_0 : \mu = \mu_0$ zutrifft, wird H_0 auf dem Signifikanzniveau gehalten, genau dann, wenn μ_0 von entsprechenden $(1-\alpha)$-Konfidenzintervall überdeckt wird.

Im Speziellen bedeutet dies

1. Bei der zweiseitigen Fragestellung, mit $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

   $$H_1 : \mu \neq \mu_0$$

   H_0 wird beibehalten – falls[1] $|t| \leq \mathrm{t}_{n-1,1-\frac{\alpha}{2}}$ – genau dann, wenn

   $$\mu_0 \in \left[\bar{x} - \mathrm{t}_{n-1,1-\frac{\alpha}{2}} \cdot {}^{s}\!/\!\sqrt{n}; \bar{x} + \mathrm{t}_{n-1,1-\frac{\alpha}{2}} \cdot {}^{s}\!/\!\sqrt{n}\right]$$

2. Bei einseitiger Fragestellung

   a) mit $\alpha_1 = \alpha$ und $\alpha_2 = 0$

   $$H_1 : \mu < \mu_0$$

   H_0 wird beibehalten – falls $t \geq -\mathrm{t}_{n-1,1-\alpha}$– genau dann, wenn

   $$\mu_0 \in (-\infty; \bar{x} + \mathrm{t}_{n-1,1-\alpha_1} \cdot {}^{s}\!/\!\sqrt{n}]$$

   b) mit $\alpha_1 = 0$ und $\alpha_2 = \alpha$

   $$H_1 : \mu > \mu_0$$

   H_0 wird beibehalten – falls $t \leq \mathrm{t}_{n-1,1-\alpha}$ – genau dann, wenn

   $$\mu_0 \in [\bar{x} - \mathrm{t}_{n-1,1-\alpha_2} \cdot {}^{s}\!/\!\sqrt{n}; +\infty)$$

[1] Wobei hier t – dies gilt auch für $\bar{x}$ und s im folgenden – anstelle von T steht, da es sich um eine konkrete Stichprobe handelt.

Example 9. Geburtsgewicht von Raucherkindern

Beispiel 10.1 (S. 198)

7.1.3.2 Multiples Testen

Bei k *unabhängig* durchgeführten Tests ist die Wahrscheinlichkeit H_0 beizubehalten, soferne H_0 auch tatsächlich zutrifft nur $(1-\alpha)^k$ hoch.

$$(1-\alpha)^k = 1 - \binom{k}{1}\alpha + \binom{k}{2}\alpha^2 - \binom{k}{3}\alpha^3 + \cdots + (-1)^k\alpha^k \approx 1 - k\alpha$$

Die Wahrscheinlichkeit H_0 zu verwerfen, obwohl H_0 tatsächlich zutrifft (ein Fehler 1. Art) ist

$$1 - (1-\alpha)^k \approx k\alpha$$

Der Fehler multipliziert sich also mit der Anzahl der durchgeführten Test

Wenn also $\frac{\alpha}{k}$ statt α als Signifikanzniveau für jeden einzelnen dieser k unabhängigen Tests gewählt wird (*Bonferroni-Korrektur*) beträgt die Wahrscheinlichkeit für einen Fehler 1. Art insgesammt nur noch

$$\left(1 - \left(1 - \frac{\alpha}{k}\right)^k\right) \approx \alpha$$

8 Vorlesung am 6. Mai 2010

8.1 Lagetests

Abschnitt 11

8.1.1 t-Tests

Abschnitt 11.1

8.1.1.1 t-Test für eine Stichprobe

Abschnitt 11.1.1

Unter der Vorraussetzung, dass alle Werte normalverteilt sind,

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

gilt für die Teststatistik:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s \cdot \sqrt{\frac{1}{n}}}$$

$H_0 : \mu = \mu_0$	H_0 wird abgelehnt, falls:
$H_1 : \mu \neq \mu_0$	$\lvert t \rvert > \mathrm{t}_{n-1,1-\frac{\alpha}{2}}$
$H_1 : \mu > \mu_0$	$t > \mathrm{t}_{n-1,1-\alpha}$
$H_1 : \mu < \mu_0$	$t < \mathrm{t}_{n-1,\alpha} = -\mathrm{t}_{n-1,1-\alpha}$

8.1.1.2 t-Test für zwei verbundene Stichproben

Abschnitt 11.1.2

Vorraussetzungen für die Zufallsvariablen X, Y:

$$\begin{aligned} D &:= X - Y \\ \delta &:= \mathbb{E}\, D \\ D &\sim \mathrm{N}(\delta, \sigma_d^2) \end{aligned}$$

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

wobei hier gilt:

$$d_i := x_i - y_i \qquad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})}$$

$H_0 : \delta = 0$	H_0 wird abgelehnt, falls:
$H_1 : \delta \neq 0$	$\lvert t\rvert > \mathrm{t}_{n-1,1-\frac{\alpha}{2}}$
$H_1 : \delta > 0$	$t > \mathrm{t}_{n-1,1-\alpha}$
$H_1 : \delta < 0$	$t < \mathrm{t}_{n-1,\alpha} = -\mathrm{t}_{n-1,1-\alpha}$

Konfidenzintervall, dass einem zweiseitigen Test entspricht:

$$\left[\bar{d} - \mathrm{t}_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}; \bar{d} + \mathrm{t}_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}\right]$$

8.1.1.3 t-Test für zwei unverbundene Stichproben

Abschnitt 11.1.3

Herleitung für die Prüfgröße Es ist

$$\bar{X} := \frac{1}{n_1} \cdot \sum_{i=1}^{n} X_i \qquad \wedge \qquad \bar{Y} := \frac{1}{n_2} \cdot \sum_{i=1}^{n} Y_i$$

wobei die Zufallsvariablen normalverteilt sind

$$X_i \sim \mathrm{N}\left(\mu_1, \sigma^2\right) \qquad \wedge \qquad Y_i \sim \mathrm{N}\left(\mu_2, \sigma^2\right)$$

und alle X_i, Y_i sind unabhängig, daher gilt

$$\bar{X} \sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right) \qquad \wedge \qquad \bar{Y} \sim \mathrm{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

So gilt nun für den Erwartungswert

$$\begin{aligned} \mathbb{E}\left(\bar{X} - \bar{Y}\right) &= \mathbb{E}\,\bar{X} + (-1)\mathbb{E}\,\bar{Y} \\ &= \mu_1 - \mu_2 \end{aligned}$$

und für die Varianz beider Verteilungen

$$\begin{aligned} \operatorname{Var}(\bar{X} - \bar{Y}) &= \operatorname{Var}\bar{X} + \operatorname{Var}(-\bar{Y}) \\ \operatorname{Var}\bar{X} + (-1)^2 \operatorname{Var}\bar{Y} &= \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \end{aligned}$$

Daher ist die Differenz normalverteilt mit:

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

Falls $H_0 : \mu_1 = \mu_2$ zutrifft, gilt daher:

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(0, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

Eigenschaften Vorraussetzungen

$$X \sim \mathrm{N}(\mu_1, \sigma^2) \qquad \wedge \qquad Y \sim \mathrm{N}(\mu_2, \sigma^2)$$

wobei X, Y unabhängig sind. Für die Teststatistik gilt nun (siehe Herleitung):

$$t = \frac{(\bar{x} - \bar{y})}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

wobei für s^2 gilt:

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

$H_0 : \mu_1 = \mu_2$	$\Leftrightarrow H_0 : \mu_1 - \mu_2 = 0$	H_0 wird abgelehnt, falls:
$H_1 : \mu_1 \neq \mu_2$	$\Leftrightarrow H_1 : \mu_1 - \mu_2 \neq 0$	$\lvert t \rvert > \mathrm{t}_{f,1-\frac{\alpha}{2}}$
$H_1 : \mu_1 > \mu_2$	$\Leftrightarrow H_1 : \mu_1 - \mu_2 > 0$	$t > \mathrm{t}_{f,1-\alpha}$
$H_1 : \mu_1 < \mu_2$	$\Leftrightarrow H_1 : \mu_1 - \mu_2 < 0$	$t < -\mathrm{t}_{f,1-\alpha}$

Wobei für den Freiheitsgrad f:=$n_1 + n_2 - 2$

Speziell für $n := n_1 = n_2$

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{2}{n}}} \qquad s^2 := \frac{s_1^2 + s_2^2}{2}$$

8.1.1.4 Welch-Test

Abschnitt 11.1.4

$$\begin{aligned} X_i &\sim \mathrm{N}(\mu_1, \sigma_1^2) & Y_i &\sim \mathrm{N}(\mu_2, \sigma_2^2) \\ \bar{X} &\sim \mathrm{N}(\mu_1, \tfrac{\sigma^2}{n_1}) & \bar{Y} &\sim \mathrm{N}(\mu_2, \tfrac{\sigma^2}{n_2}) \end{aligned}$$

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$(\bar{X} - \bar{Y}) \sim \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$t = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Für die Freiheitsgrade der t-Verteilung gilt:

$$f := \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \right\rfloor$$

8.1.1.5 Vorraussetzungen für t-Tests

Abschnitt 11.1.5

Der t-Test ist robust gegenüber Abweichungen von einer Normalverteilung, jedoch ist folgendes zusätzlich zu beachten.

t-Test für eine Stichprobe

$n \geq 10$	$\wedge$	Daten annähernd symetrisch
$n \geq 25$		sonst

t-Test für zwei verbundene Stichproben

$$n \geq 10 \qquad \wedge \qquad d_i \text{ annähernd symetrisch}$$

insbesondere, wenn X, Y eine gleiche Verteilungsform haben.

t-Test für zwei unverbundene Stichproben Vorrausgesetzt ist eine gleiche Verteilung der Zufallsvariablen

$$\begin{array}{lll} n_1, n_2 \geq 10 & \wedge & \text{symetrische Verteilung} \\ n_1, n_2 \geq 20 & & \text{sonst} \end{array}$$

und X, Y sind vom gleichen Verteilungstyp.

8.1.1.6 Andere Anwendungen des t-Tests

Abschnitt 11.1.6

Der t-Test kann auch nicht als Lagetest eingesetzt werden, um zu testen, ob sich ein empirische Korrelationskoeffizient nach Pearson ρ signifikant von 0 unterscheidet, kann folgende Prüfgröße verwendet werden. r steht hier für den empirisch ermittelten Wert.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$\begin{array}{ll} H_0 : \rho = 0 & \\ H_1 : \rho > 0 & t > \mathrm{t}_{n-2,1-\alpha} \\ H_1 : \rho < 0 & t < -\mathrm{t}_{n-2,1-\alpha} \end{array}$$

8.1.2 Rangsummentests

Abschnitt 11.2

Rangsummentests sind *verteilungsfreie*, also *nicht-prametrische*, Tests, da sie keine bestimmte Verteilungsform der Stichproben vorraussetzen. Die Prüfgrößen werden nicht aus den Messwerten, sondern aus deren Rangzahlen berechnet.

8.1.2.1 Wilcoxon-Test für eine Stichprobe

Abschnitt 11.2.1

Dieser Test überprüft, ob und in welchen Maß die Werte einer Stichprobe von einem vorgegebenen Sollwert $\tilde{\mu}_0$ abweichen. Vorraussetzung für diese Verfahren ist, eine symetrische Verteilung der Stichprobenwerte, also $\tilde{\mu} = \mu$.

Die Nullhypothese lautet:

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

Für die Durchführung wird folgender Ablauf eingehalten:

1. Die Differenz zum Sollwert wird bestimmt: $x_i - \tilde{\mu}_0$
2. Alle Werte mit $x_i = \tilde{\mu}_0$ werden eleminiert. Der Umfang der Stichprobe muss angepasst werden.
3. Die Differenzen werden nun nach der Größe ihres Betrags sortiert, die kleinste erhält die Rangzahl 1.
4. Bei gleichen Differenzen, wird eine mittlere Rangzahl zugeordnet: Rang $+ \frac{1}{\|\text{Gleiche Differenzen}\|}$ Hierbei spricht man von *verbundenen Rängen.*
5. Die Rangzahlen der positiven und negativen Differenzen werden addiert und mit R^-, R^+ bezeichnet.
6. Prüfgröße $R := \min\{R^-, R^+\}$
7. Nun muss in einer Tabelle[1] der kritische Wert R_{krit} in Abhängigkeit von Stichprobenumfang n und Irrtumswahrscheinlichkeit α nachgeschlagen werden.
8. H_0 wird angenommen, wenn $R \leq R_{krit} \Leftrightarrow R^- \leq R_{krit} \vee R^+ \leq R_{krit}$

Die Nullhypothese, und die Alternativhypothesen für zweiseitigen und einseitige Tests lauten wiefolgt:

$H_0 : \tilde{\mu} = \tilde{\mu}_0$	H_0 wird abgelehnt, falls:
$H_1 : \tilde{\mu} \neq \tilde{\mu}_0$	$R^- \leq R_{krit} \vee R^+ \leq R_{krit}$
$H_1 : \tilde{\mu} > \tilde{\mu}_0$	$R^- \leq R_{krit}$
$H_1 : \tilde{\mu} < \tilde{\mu}_0$	$R^+ \leq R_{krit}$

8.1.2.2 Wilcoxon-Test für zwei verbundene Stichproben

Abschnitt 11.2.2

Dieser Test verläuft wie sein Pendant zum t-Test mit verbundenen zwei Stichproben. Es werden wiederum die Mediane der beiden Stichproben $\tilde{\mu}_1, \tilde{\mu}_2$ verglichen.

Der Ablauf ist ähnlich dem des Tests für eine Stichprobe:

1. Es wird für jedes Wertepaar die Differenz gebildet: $d_i := x_i - y_i$
2. Alle $d_i = 0$ werden eliminiert.
3. Der Rest verläuft gleich.

Ist nun

$$D := X - Y$$

die Verteilung der Differenzen, so ist der Median der Differenzen $\tilde{\delta} = \tilde{\mu}_D$

[1] Tabelle C Annhang Seite 319

Dies gilt natürlich nur unter der Vorraussetzung, dass X, Y annähernd die gleiche Verteilungsform haben. D ist daher eher symetrisch Verteilt.

Mit dieser Notation definiert sich folgende Nullhypothese und Alternativhypothesen für zweiseitigen und einseitige Tests. Die Bedingungen zur Annahme, oder Ablehnung von H_0 bleiben dieselben wie für den Test mit einer Stichprobe.

$$\begin{aligned} H_0: &\quad \tilde{\delta} = 0 \quad \Leftrightarrow H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 \\ H_1: &\quad \tilde{\delta} \neq 0 \\ H_1: &\quad \tilde{\delta} > 0 \\ H_1: &\quad \tilde{\delta} < 0 \end{aligned}$$

Example 10. Zehn Personen nehmen sechs Monate lang eine Diät ... Beispiel 11.3 (S. 217)

Mit Hilfe der kritischen Werte aus Tabelle C lässt sich dieses Beispiel sowohl in einseitiger als auch zweiseitiger Fragestellung lösen.

$H_1 : \tilde{\delta} \neq 0$	$R = 7$	Testentscheidung	Begründung
$\alpha = 10\%$	$R_{krit} = 10$	$H_1 : \tilde{\delta} \neq 0$	$R \leq R_{krit}$
$\alpha = 5\%$	$R_{krit} = 8$	$H_1 : \tilde{\delta} \neq 0$	$R \leq R_{krit}$
$\alpha = 2\%$	$R_{krit} = 5$	$H_0 : \tilde{\delta} = 0$	$R > R_{krit}$
$\alpha = 1\%$	$R_{krit} = 3$	$H_0 : \tilde{\delta} = 0$	$R > R_{krit}$

$H_1 : \tilde{\delta} > 0$	$R^- = 7$	Testentscheidung	Begründung
$\alpha = 5\ \%$	$R_{krit} = 10$	$H_1 : \tilde{\delta} > 0$	$R^- \leq R_{krit}$
$\alpha = 2.5\%$	$R_{krit} = 8$	$H_1 : \tilde{\delta} > 0$	$R^- \leq R_{krit}$
$\alpha = 1\ \%$	$R_{krit} = 5$	$H_0 : \tilde{\delta} = 0$	$R^- > R_{krit}$
$\alpha = 0.5\%$	$R_{krit} = 3$	$H_0 : \tilde{\delta} = 0$	$R^- > R_{krit}$

9 Vorlesung am 20. Mai 2010

9.0.2.3 U-Test (Wilcoxon-Test für unverbundene Stichproben)

Abschnitt 11.2.3

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$$

Die Nullhypothese nimmt an, dass die Mediane beider Stichproben gleich sind. Die Umfänge der Stichproben n_1, n_2 muss nicht gleich sein, jedoch müssen als Vorraussetzung beide Stichproben die gleiche Verteilungsform haben. Um die Tabelle D[1] aus dem Anhang zu verwenden muss zusätzlich gelten: $n_1 \geq n_2$.

Vorgehensweise:

1. Alle Werte beider Stichproben werden aufsteigend sortiert und mit Rangzahlen versehen. (Siehe Wilcoxon-Test)
2. Die Summe der Rangzahlen wird dann als R_1 und R_2 bezeichent.
3. Danach wird für die Testgröße berechnet:

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1 \\ U_2 &= n_1 n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2 \end{aligned}$$

   Dabei gilt:

$$\begin{aligned} U_1 + U_2 &= n_1 \cdot n_2 \\ R_1 + R_2 &= \frac{(n_1 + n_2) \cdot (n_1 + n_2 + 1)}{2} \end{aligned}$$

   Die Testgröße:

$$U = \min \{U_1, U_2\}$$

4. Die Nullhypothese wird abgelehnt,

$$U \leq U_{krit}$$

Falls nun gilt, $n_1 \geq 10 \wedge n_2 \geq 10$, so ist $U \sim \mathrm{N}\left(\frac{n_1 \cdot (n_1 + n_2 + 1)}{2}, \frac{n_1 n_2}{2}\right)$.

[1] Die Tabelle ist so konstruiert, dass n_1 vertikal ($\downarrow$) und n_2 horizontal ($\rightarrow$) angeordnet sind.

9.0.2.4 Vergleich von Rangsummentests und t-Tests

Abschnitt 11.2.4

9.0.3 Vorzeichentest

Abschnitt 11.3

9.0.3.1 Vorzeichentest für eine Stichprobe

Abschnitt 11.3.1

$$H_0 : \tilde{\mu}_0 = \tilde{\mu}_1$$

$$\forall x_i \begin{cases} x_i > \tilde{\mu}_0 & \text{gewertet als } + \\ x_i < \tilde{\mu}_0 & \text{gewertet als } - \end{cases}$$

Verallgemeinert lässt sich jetzt sagen, die Nullhypothese wird abgelehnt, wenn es nun *entscheident* mehr Plus als Minus gibt, oder vice versa. Im Detail heißt dass, zuerst wird die Prüfgröße errechnet

$$V_+ := \|+\|$$

Anmerkung: Es wird nur eine Größe benötigt, und nicht wie bei den Rangsummentests zwei, von denen dann die kleinere genommen wird, da in Tabelle F für die kritischen Werte ein Bereich angegeben ist. Wäre nur ein Wert angegeben, so wäre dies der kleinere, und man müsste als Testgröße das Minimum aus Plus und Minus Zahl nehmen, beziehungsweise kann die andere Zahl auch errechnet werden.

Falls nun die Nullhypothese zutrifft, gilt:

$$V_+ \sim \mathrm{B}\left(n, \tfrac{1}{2}\right)$$

Da nun die Binomialverteilung durch die Normalverteilung approximiert werden kann:

$$\mathrm{B}\left(n, p\right) \approx \mathrm{N}\left(np, npq\right) \qquad \text{wenn } npq \geq 9$$

so gilt nun

$$\mathrm{B}\left(n, \tfrac{1}{2}\right) \approx \mathrm{N}\left(\tfrac{n}{2}, \tfrac{n}{4}\right) \qquad \text{wenn } npq \geq 9 \Leftrightarrow n \geq 36$$

9.0.3.2 Vorzeichentest für zwei verbundene Stichproben

Abschnitt 11.3.2

Die Stichproben werden als X, Y und die Differenz als $D := X - Y$ bezeichnet. Der Median von D wird als $\tilde{\delta} := \tilde{\mu}_D$ bezeichnet.

$$
\begin{aligned}
H_0 &: \mathrm{P}(X < Y) = \mathrm{P}(X > Y) &&\Leftrightarrow \mathrm{P}(X - Y < 0) = \mathrm{P}(X - Y > 0) \Leftrightarrow &&\tilde{\delta} = 0 \\
H_1 &: \mathrm{P}(X < Y) \neq \mathrm{P}(X > Y) &&\Leftrightarrow \mathrm{P}(X - Y < 0) \neq \mathrm{P}(X - Y > 0) \Leftrightarrow &&\tilde{\delta} \neq 0 \\
H_1 &: \mathrm{P}(X < Y) > \mathrm{P}(X > Y) &&\Leftrightarrow \mathrm{P}(X - Y < 0) > \mathrm{P}(X - Y > 0) \Leftrightarrow &&\tilde{\delta} < 0 \\
H_1 &: \mathrm{P}(X < Y) < \mathrm{P}(X > Y) &&\Leftrightarrow \mathrm{P}(X - Y < 0) < \mathrm{P}(X - Y > 0) \Leftrightarrow &&\tilde{\delta} > 0
\end{aligned}
$$

$H_0 : \tilde{\delta} = 0$	H_0 wird abgelehnt, falls:
$H_1 : \tilde{\delta} \neq 0$	$V_+ < V_u \vee V_+ > V_o$
$H_1 : \tilde{\delta} < 0$	$V_+ < V_u$
$H_1 : \tilde{\delta} > 0$	$V_+ > V_o$

Ähnlich wie beim Vorzeichentest für eine Stichprobe wird hier jeder Differenz $d_i := x_i - y_i$ ein Vorzeichen zugeordnet. Die Zahl der positiven wird als Prüfgröße V_+ bezeichnet. Für die Entscheidung wird in Tabelle F die obere bzw. untere Grenze des Annahmebereichs V_o, V_u nachgeschlagen.

Falls nun $n \geq 36$, so gilt bei Zutreffen von H_0, approximativ

$$V_+ \sim \mathrm{N}\left(\tfrac{n}{2}, \tfrac{n}{4}\right)$$

Anmerkung: Für $\mu = n \cdot p = n \cdot \frac{1}{2}$ und $\sigma^2 = n \cdot p \cdot q = n \cdot \frac{1}{2} \cdot \frac{1}{2}$

Schranken für den Annahmebereich

Für eine zweiseitige Fragestellung

$$V_u, V_o := \frac{n}{2} \pm \left(z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n}}{2} + \frac{1}{2} \right)$$

Für eine einseitige Fragestellung

$$
\begin{aligned}
V_u &:= \frac{n}{2} - \left(z_{1-\alpha} \cdot \frac{\sqrt{n}}{2} + \frac{1}{2} \right) \\
V_o &:= \frac{n}{2} + \left(z_{1-\alpha} \cdot \frac{\sqrt{n}}{2} + \frac{1}{2} \right)
\end{aligned}
$$

9.1 Tests zum Vergleich von Häufigkeiten

Abschnitt 12

9.1.1 Binomialtest für eine Stichprobe

Abschnitt 12.1

Dieser Test überprüft die relative Häufigkeit der Ausprägung A mit der vorgegebenen Wahrscheinlichkeit p_0 vereinbar ist.

$$\begin{aligned} H_o &: p = p_0 \\ H_1 &: p \neq p_0 \end{aligned}$$

Falls H_0 zutrifft, ist die Stichprobe binomialverteilt:

$$X \sim \mathrm{B}(n, p_0)$$

$$\begin{aligned} H_1 &: p < p_0 \\ H_1 &: p > p_0 \end{aligned}$$

Wenn n hinreichend groß ist mit $np_0q_0 \geq 9$, lässt sich die Binomialverteilung durch eine Normalverteilung approximieren

$$\mathrm{B}(n, p_0) \approx \mathrm{N}(np_0, np_0q_0)$$

Wobei hier gilt $q_0 := 1 - p_0$

9.1.2 χ^2-Tests

Abschnitt 12.2

9.1.2.1 χ^2-Vierfelder-Test

Abschnitt 12.2.1

	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 9.1: Vierfeldertabelle für den χ^2-Vierfelder-Test

9.1.2.2 χ^2-Unabhängigheits-Test

Diesem Test liegt eine Stichprobe mit zwei alternativen Merkmalen und dem Umfang n. Die Häufigkeiten sind in der Vierfeldertafel (Tabelle 9.1) dargestellt.

$$H_0 : \mathrm{P}(A \mid B) = \mathrm{P}(A)$$

Wenn die Nullhypothese erfüllt ist, dann sind die Ereignisse unabhängig voneinander und es gilt:

$$\begin{aligned}
\mathrm{P}(A \mid B) = \mathrm{P}(A) \quad &\Leftrightarrow \quad \mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B) \\
&\Leftrightarrow \quad \mathrm{P}(A \mid B) = \mathrm{P}(A \mid \overline{B}) = \mathrm{P}(A) \\
&\Leftrightarrow \quad \mathrm{P}(\overline{A}|B) = \mathrm{P}(\overline{A}|\overline{B}) = \mathrm{P}(\overline{A}) \\
&\Leftrightarrow \quad \mathrm{P}(B \mid A) = \mathrm{P}(B \mid \overline{A}) = \mathrm{P}(B) \\
&\Leftrightarrow \quad \mathrm{P}(\overline{B} \mid A) = \mathrm{P}(\overline{B} \mid \overline{A}) = \mathrm{P}(\overline{B})
\end{aligned}$$

Die Idee des χ^2-Tests ist es nun, die beobachteten Häufigkeiten a, b, c, d mit den erwarteten Häufigkeiten –welche die Ereignisse unter der Nullhypothese annehmen würden – zu vergleichen.

$$\sum_{i \in \{a,b,c,d\}} \frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$$

Die Prüfgröße ist annähernd χ^2-verteilt mit $f = 1$. Für den Vierfeldertest berechnet sie sich als:

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

H_0 wird verworfen, falls $\chi^2 > \chi_{1,1-\alpha}$ (beim zweiseitigen Test mit $\neq$) (Siehe Tabelle F im Anhang)

Example 11.

$$\begin{aligned}
A &= \text{♂} \\
B &= \text{♀} \\
A, B &= \text{Rhesusfaktor positiv} \\
\overline{A}, \overline{B} &= \text{Rhesusfaktor negativ}
\end{aligned}$$

9.1.2.3 χ^2-Homogenitäts-Test

Eine andere Interpretation des Vierfelder-Tests ist eine Überprüfung der relativen Häufigkeit zweier unabhängiger Stichproben; es wird also untersucht, ob ein bestimmtes Merkmal identverteilt ist).

$$\begin{aligned} H_0: &\quad \mathrm{P}(A \mid B) &&= \mathrm{P}(A \mid \overline{B}) \\ H_1: &\quad \mathrm{P}(A \mid B) &&\neq \mathrm{P}(A \mid \overline{B}) \\ H_1: &\quad \mathrm{P}(A \mid B) &&> \mathrm{P}(A \mid \overline{B}) \\ H_1: &\quad \mathrm{P}(A \mid B) &&< \mathrm{P}(A \mid \overline{B}) \end{aligned}$$

Bei der einseitigen Fragestellung wird H_1 angenommen, falls $\chi^2 > \chi^2_{1,1-2\alpha} \wedge \frac{a}{n_1} > \frac{c}{n_2}$

Example 12. Medikamententest

$$\begin{aligned} B &= \text{Medikament} \\ \overline{B} &= \text{Placebo} \\ A &= \text{Besserung} \\ \overline{A} &= \text{keine Besserung} \end{aligned}$$

Daher gibt es n_1 Personen, die ein Medikament bekommen und n_2 Personen, die ein Placebo bekommen haben.

Überlegungen zum einseitigen χ^2-Homogenitätstest

$$H_0 : \mathrm{P}(A \mid B) = \mathrm{P}(A \mid \overline{B})$$

Falls H_0 zutrifft, gilt

$$\underbrace{\mathrm{P}\left(\chi^2 > \chi^2_{1,1-2\alpha}\right)}_{2\alpha} \cdot \underbrace{\mathrm{P}\left(\tfrac{a}{n_1} > \tfrac{c}{n_2} \mid \chi^2 > \chi^2_{1,1-2\alpha}\right)}_{\frac{1}{2}} = \underbrace{\mathrm{P}\left(\tfrac{a}{n_1} > \tfrac{c}{n_2} \wedge \chi^2 > \chi^2_{1,1-2\alpha}\right)}_{\alpha}$$

$H_0 : \mathrm{P}(A \mid B) = \mathrm{P}(A \mid \overline{B})$	H_0 wird abgelehnt, falls:
$H_1 : \mathrm{P}(A \mid B) \neq \mathrm{P}(A \mid \overline{B})$	$\chi^2 > \chi^2_{1,1-\alpha}$
$H_1 : \mathrm{P}(A \mid B) > \mathrm{P}(A \mid \overline{B})$	$\chi^2 > \chi^2_{1,1-2\alpha} \wedge \frac{a}{n_1} > \frac{c}{n_2}$
$H_1 : \mathrm{P}(A \mid B) < \mathrm{P}(A \mid \overline{B})$	$\chi^2 > \chi^2_{1,1-2\alpha} \wedge \frac{a}{n_1} < \frac{c}{n_2}$

Example 13. zum zwei- bzw. einseitigen χ^2-Test

$$\begin{aligned} M &\ldots \text{Medikament} \\ Pl &\ldots \text{Placebo} \\ B &\ldots \text{Besserung} \\ \overline{B} &\ldots \text{keine Besserung} \end{aligned}$$

Hierbei handelt es sich um einen einseitigen Test, da es die Placebowirkung sowohl beim Medikament als logischerweise auch beim Placebo existiert.

	B	$\overline{B}$	Randsummen
M	151	572	$n_1 = 723$
Pl	51	452	$n_2 = 502$
Randsummen	202	1024	$n = 1226$

$$
\begin{aligned}
H_0: &\quad \mathrm{P}(B \mid M) = \mathrm{P}(B \mid Pl) \\
H_1: &\quad \mathrm{P}(B \mid M) \neq \mathrm{P}(B \mid Pl) \\
H_1: &\quad \mathrm{P}(B \mid M) > \mathrm{P}(B \mid Pl)
\end{aligned}
$$

$$
\begin{aligned}
\chi^2 &= \frac{1226 \cdot (151 \cdot 452 - 51 \cdot 572)^2}{202 \cdot 1024 \cdot 503 \cdot 723} = 24.890\,978\,62 \\
\frac{a}{n_1} &= \frac{151}{723} = 0.208\,852\,006 \\
\frac{c}{n_2} &= \frac{51}{503} = 0.101\,391\,650
\end{aligned}
$$

$H_1 : \mathrm{P}(B \mid M) \neq \mathrm{P}(B \mid Pl)$	$\chi^2_{1,1-\alpha}$	Testentscheidung	Begründung
$\alpha = 10\%$	2.706	H_1	$\chi^2 > \chi^2_{1,0.9}$
$\alpha = 5\,\%$	3.841		$\chi^2 > \chi^2_{1,0.95}$
$\alpha = 2.5\%$	5.024		$\chi^2 > \chi^2_{1,0.975}$
$\alpha = 1\,\%$	6.635		$\chi^2 > \chi^2_{1,0.99}$
$\alpha = 0.5\%$	7.879		$\chi^2 > \chi^2_{1,0.995}$

$H_1 : \mathrm{P}(B \mid M) > \mathrm{P}(B \mid Pl)$	$\chi^2_{1,1-2\alpha}$	Testentscheidung	Begründung
$\alpha = 5\ \%$	2.706	H_1	$\chi^2 > \chi^2_{1,0.95} \wedge \frac{a}{n_1} > \frac{c}{n_2}$
$\alpha = 2.5\ \%$	3.841		$\chi^2 > \chi^2_{1,0.975} \wedge \frac{a}{n_1} > \frac{c}{n_2}$
$\alpha = 1.25\%$	5.024		$\chi^2 > \chi^2_{1,0.9875} \wedge \frac{a}{n_1} > \frac{c}{n_2}$
$\alpha = 0.5\ \%$	6.635		$\chi^2 > \chi^2_{1,0.995} \wedge \frac{a}{n_1} > \frac{c}{n_2}$
$\alpha = 0.25\%$	7.879		$\chi^2 > \chi^2_{1,0.9975} \wedge \frac{a}{n_1} > \frac{c}{n_2}$

Auf jeden dieser Signifikanzniveaus wirkt das Medikament besser als das Placebo.

10 Vorlesung am 27. Mai 2010

10.0.2.4 χ^2-Test für $k \cdot l$ Felder

Abschnitt 12.2.3

Wie Vierfelder-Test nur mit Freiheitsgrad

$$f = (k-1) \cdot (l-1)$$

für

$$\text{erwartete Häufigkeiten} > 5$$

10.0.2.5 Assiotiationsmaße für qualitative Merkmale

Abschnitt 12.2.4

ϕ-Koeffizient

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Zusätzlich gilt folgendes:

$$0 \leq \phi \leq 1 \quad \wedge \quad b = c = 0 \Rightarrow \phi = 1$$

Der ϕ-Koeffizient ist 0 bei vollkommener Unabhängigkeit der Merkmale. Im Falle von $\phi = 1$ lässt sich aufgrund einers Merkmals das andere präziese vorhersagen. Der ϕ-Koeffizient ist signifikant größer als 0, falls das Ergebnis des Vierfelder-Tests signifikant ist.

Abschnitt 3.4.2
Seite 52

Assoziationskoeffizient nach Yule

$$Q = \frac{ad - bc}{ad + bc}$$

$$ad = bc \Leftrightarrow a : b = c : d \Leftrightarrow Q = 0$$

$$-1 \leq Q \leq 1$$

10.0.2.6 McNemar Test

Abschnitt 12.2.5

Häufigkeitstest für 2 verbundene Stichproben hinsichtlich eines Alternativmerkmals

		Stichprobe 1	
		A	$\overline{A}$
Stichprobe 2	A	a	b
	$\overline{A}$	c	d

$$X \quad \ldots \quad \text{Stichprobe 1} \mathrel{\hat{=}} \text{Therapie 1}$$
$$Y \quad \ldots \quad \text{Stichprobe 2} \mathrel{\hat{=}} \text{Therapie 2}$$

$$H_0: \quad \mathrm{P}(X = A) = \mathrm{P}(Y = A)$$
$$H_1: \quad \mathrm{P}(X = A) \neq \mathrm{P}(Y = A)$$

Für die Testgröße gilt

$$\chi^2 = \frac{(b-c)^2}{b+c} \qquad f = 1$$

Unter der Bedingung

$$b + c \leq 30$$

gilt

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

Vorraussetzung

$$\frac{b+c}{2} \geq 5$$

Einseitiger McNemar Test

$$H_0: \quad \mathrm{P}(X = A) = \mathrm{P}(Y = A)$$
$$H_1: \quad \mathrm{P}(X = A) > \mathrm{P}(Y = A)$$

H_1 wird angenommen, falls $\chi^2 > \chi^2_{1,1-2\alpha} \wedge c > b$

Example 14. Zweiseitig

Beispiel 12.5 (S. 239)

10.0.2.7 Mittelwert

Minimum Eigenschaft des Mittlwerts

$$\sum (x_i - \bar{x})^2 < \sum (x_i - c)^2 \quad \forall c \in \mathbb{R} \setminus \{\overline{x}\}$$

Minimum Eigenschaft des Medians

$$\sum |x_i - \tilde{x}| \leq \sum |x_i - c| \quad \forall c \in \mathbb{R} \setminus \{\tilde{x}\}$$

Abschnitt 4.2.6/7

Geometrisches Mittel und Harmonisches Mittel

Das geometrische Mittel x_i sind relative Änderungen (Änderungensfaktoren)

$$\bar{x}_G = \sqrt[n]{\prod x_i}$$

Das harmonische Mittel x_i Quotienten, die sich nur im Nenner unterscheiden

$$\bar{x}_H = \frac{n}{\sum \frac{1}{x_i}}$$

Die Mittel ergeben sich aus dem arithmetischen Mittel, allerdings von entsprechend transformierenden Stichprobenwerten

$$\begin{aligned} \ln \bar{x}_G = \frac{1}{n} \sum_{i=1}^{n} \ln x_i \quad &\Leftrightarrow \quad \bar{x}_G = e^{\frac{1}{n} \sum \ln x_i} \\ \sqrt[n]{\prod e^{\ln x_i}} \quad &= \quad \sqrt[n]{\prod x_i} \end{aligned}$$

Analog gilt dies auch für ^{10} log und jede andere Basis.

Für das harmonische Mittel gilt:

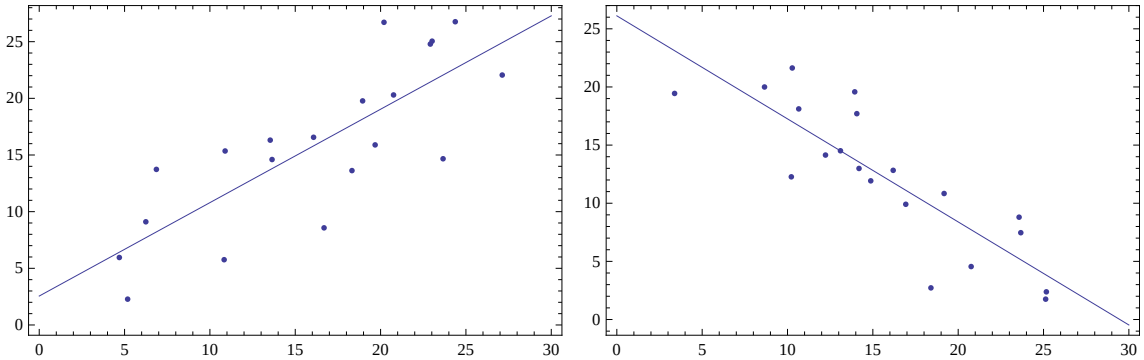
$$\begin{aligned} \frac{1}{\bar{x}_H} &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \\ \bar{x_H} &= \frac{1}{\frac{1}{n} \cdot \sum \frac{1}{x_i}} \\ \bar{x}_H &= n \cdot \left(\sum x_i^{-1} \right)^{-1} \end{aligned}$$

11 Vorlesung am 11. Juni 2010

11.0.3 Korrelationsanalyse

Abschnitt 5.2

11.0.3.1 Punktwolke



(a) Positive Kovarianz

(b) Negative Kovarianz

Abbildung 11.0.1: Punktwolke mit Regressionsgerade

Die Punktwolke ist eine quantitative graphische Darstellung der Wertepaare (x_i, y_i) zweier Stichproben X, Y. Hierbei wird für jedes Wertepaar ein Punkt (x_i, y_i) im kartesischen Koordinatensystem gezeichnet.

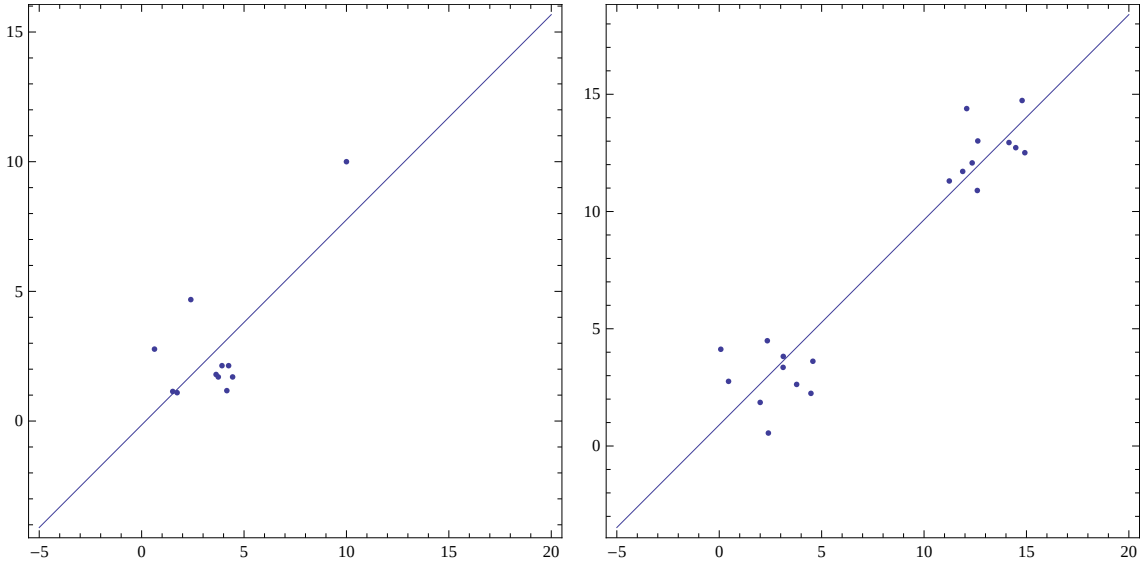
11.0.3.2 Kovarianz

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Eine positive Kovarianz heißt die Regressionsgerade steigt und es herrscht ein gleichsinniger Zusammenhang.

Eine negative Kovarianz bedeutet die Regressionsgerade fällt, es herrscht daher ein gegensinniger Zusammenhang

Im Falle, dass die Kovarianz 0 ergibt, so ergibt sich kein linearer Zusammenhang zwischen den beiden Stichproben X und Y.



(a) Korrelation durch Ausreißer (b) Inhomogenitätskorrelation

Abbildung 11.0.2: Arten der Scheinkorrelation

11.0.4 Pearson Koeffizient

Abschnitt 5.2.4

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Für den Pearsonkoeffizient r gilt,

$$\begin{aligned} -1 &\leq r \leq 1 \\ 0 &\leq r^2 \leq 1 \end{aligned}$$

11.0.4.1 Interpretation des Korrelationskoeffizienten

Wird der Korrelationskoeffizient falsch, das heißt von dem Kontext der Werte getrennt, interpretiert, so kann es zu Scheinrelationen (Nonsensrelationen) kommen.

Formale Korrelation können entstehen, wenn sich zwei Merkmale immer zu einem Ganzen (100%) addieren lassen. Dieser Zusammenhang ist Mathematisch gegeben und es ergibt sich immer ein Korrelationskoeffizient von $r = -1$.

Selektions Korrelation In einer Stichprobe müssen alle Variationen von Merkmalen mit ähnlicher Wahrscheinlichkeit wie in der Grundgesamtheit vorkommen, damit die Stichprobe repräsentativ ist. Wird mit nicht repräsentativen Stichproben eine Korrelation errechnet, so muss diese nicht auch für die Grundgesamtheit gelten.

Korrelation durch Ausreißer Eine Ausreißer in der Stichprobe kann einen stark abweichenden Korrelationskoeffizienten verursachen. Ausreißer lassen sind in Punktwolken meist leicht erkennbar, siehe Abbildung 11.0.2a.

Inhomogenitätskorrelation ergibt sich, wenn für zwei inhomogene Gruppen ein gemeinsamer Korrelationskoeffizient ermittelt wird. Auch hier lässt sich dass bei einer Punktwolke leicht erkennen, siehe Abbildung 11.0.2b.

Gemeinsamkeitskorrelation liegt vor, wenn zwei Merkmale durch ein drittes beeinflusst werden. Wird – zum Beispiel – der Storchenbestand eines Gebiets mit der Geburtenrate verglichen, so ergibt sich ein positiver Zusammenhang, obwohl jedem bekannt sein sollte, dass der Storch nicht die Kinder bringt. Der Grund ist eine dritte Größe, nämlich die allgemeine zeitliche Tendenz.

11.0.5 Regressionsanalyse

Abschnitt 5.3

Die Regressionsgerade in der Form

$$y = a + bx$$

Mit den Werten

$$b = \frac{s_{xy}}{s_{x^2}} \qquad \wedge \qquad a = \bar{y} - b\bar{x}$$

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}}$$

11.0.6 Spearman'scher Korrelationskoeffizient

Abschnitt 5.4.1

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}$$

Für die Berechnung des Spearman Koeffizienten r_S werden alle x-Werte sortiert und mit Rangzahlen versehen. Ebenso wird so mit den y-Werten verfahren. Danach wird die Differenz d_i der Rangzahlen für jedes Wertepaar (x_i, y_i) ermittelt.

11.0.7 Klassifikation nach formalen Aspekten

Abschnitt 13.3

11.0.7.1 Deskriptiv vs. Analytisch

Deskriptive Studien sind rein beschreibend, sie werden zum Beispiel für Register (Krebsreg., Geburdenreg., Sterbereg.) sowie Fallberichte, Fallserien und Querschnittsstudien verwendet.

Analytische Studien werden durchgeführt, um den Zusammenhang mehrerer Einflussgrößen auf eine Zielgröße zu ermitteln. Bei sogenannten ökologischen[1] Studien werden verschiedene Register miteinander verknüpft, mit dem Ziel eine deskriptive Studie in eine analytische zu überführen.

11.0.7.2 Transversial vs. Longitudinal

Transversale Studien sind Querschnittsstudien, eine Momentaufnahme einer Population bei der eine oder Mehrere Eigenschaften der Studienteilnehmer erfasst werden. Eine einfache Form ist die Fallserie. Die Prävalenzstudie ist eine transversale Studie bei der die Prävalenz einer Krankheit zu einem bestimmten Zeitpunkt festgestellt wird. Transversale Studien sind meist deskriptiv.

Longitudinale Studien sind Längsschnittstudien, sie haben das Ziel, einen zeitlichen Verlauf zu beschreiben oder einen zeitlichen Zusammenhang zu ermitteln.

11.0.7.3 Retrospektiv vs. Prospektiv

Retrospektive Studien betrachten zunächst die Zielgröße und dann die Einflussgrößen. Ein Beispiel sind die Fall-Kontroll-Studien. Der Vorteil ist, dass diese Studien mit vorhandenen Daten schnell durchgeführt werden kann, jedoch der Nachteil liegt auf der Hand: oft schlechte Datenqualität bzw. Informationbias.

Prospektive Studien betrachten zunächst die Einflussgrößen und dann die Zielgrößen, hier gibt es eine Kontrollmöglichkeit bezüglich der Stichprobe. Der Vorteil ist die bessere Datenqualität, der Nachteil, der erhöhte Zeit- und Kostenaufwand, da die Daten hier durch Experimente (randomisierte klinische Studien) oder Kohorten Studien ermittelt werden.

11.0.7.4 Beobachtend vs. Experimentell

Beobachtend

Experimentell

11.0.7.5 Monozentrisch vs. Multizentrisch

Monozentrisch

Multizentrisch

[1] vom Griechischen οἶκος (oikos) *Haus, Haushalt* und λόγος (logos) *Lehre*

11.0.7.6 Typische Studien

Risikostudien

Diagnosestudien

Therapiestudien

Präventionsstudien

11.0.8 Fehlerquellen und -möglichkeiten

11.0.8.1 Zufällige Fehler

11.0.8.2 Systematische Fehler

Bias als Systematischer Erfassungsfehler

Selektionsbias

Bias durch Confounder

11.0.9 Fall-Kontroll-Studien

Abschnitt 14.3

11.0.9.1 Matching

Abschnitt 14.3.3

11.0.9.2 Biasquellen

Abschnitt 14.3.4

11.0.9.3 κ-Koeffizient nach Cohen

Abschnitt 15.1.4

Der κ-Koeffizient wird verwendet um die interindividuelle[2] Variabilität und die intraintividuelle[3] Variabilität zu messen.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o ... Beobachtete Wahrscheinlichkeit
p_e ... Erwarteter Anteil übereinstimmender Urteile

[2] Grad der Übereinstimmung zwischen zwei Beobachtern
[3] Grad der Übereinstimmung der z.B. Beurteilung desselben Beobachter zu zwei verschiedenen Zeitpunkten

Beobachter B	Beobachter A: 1	2	3
1	n_{11}	n_{12}	n_{13}
2	n_{21}	n_{22}	n_{23}
3	n_{31}	n_{32}	n_{33}
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$

$$
\begin{aligned}
p_o = & \quad \sum_{i=1}^{k} \frac{n_{ii}}{n} & = \frac{1}{n} \sum_{i=1}^{k} n_{ii} \\
e_{ii} = & \quad \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot i}}{n} \cdot n & = \frac{n_{\cdot i} \cdot n_{i\cdot}}{n} \\
p_e = & \quad \sum_{i=1}^{n} \frac{e_{ii}}{n} & = \frac{1}{n^2} \sum_{i=1}^{n} n_{\cdot i} \cdot n_{i\cdot}
\end{aligned}
$$

Interpretation des κ-Wertes Wenn $\kappa \leq 1$ gilt, so gibt es einen Zusammenhang:

$\kappa \leq 0.40$	schwache Übereinstimmung
$0.40 \leq \kappa \leq 0.60$	erkennbare Übereinstimmung
$0.60 \leq \kappa \leq 0.80$	gute Übereinstimmung
$0.80 \leq \kappa \leq 0.90$	exzellente Übereinstimmung
$0.90 \leq \kappa \leq 1.00$	perfekte Übereinstimmung

Bei $\kappa = 0$ ist kein Zusammenhang vorhanden. Es sind aber auch Werte mit $\kappa < 0$ möglich.