Institut für Statistik
Friedrich Leisch

SS 08 und WS 08/09

Formelsammlung zur Vorlesung Statistik für Wirtschaftswissenschaftler

Diese Formelsammlung darf in der Klausur verwendet werden. Eigene Notizen und Ergänzungen dürfen eingefügt, aber keine zusätzlichen Blätter eingeheftet werden.

Inhaltsverzeichnis

1 Eindimensionale Merkmale

1.1 Häufigkeiten und Häufigkeitsverteilungen

Bezeichnungen:

- Urliste: $x_1, \ldots, x_n$
- geordnete Urliste: $x_{(1)} \leq \ldots \leq x_{(n)}$
- Merkmalsausprägungen: $a_1 < \ldots < a_k$
- absolute Häufigkeit der Ausprägung a_j: $h_j = h(a_j)$, wobei gilt: $\sum_{j=1}^{k} h(a_j) = n$
- relative Häufigkeit der Ausprägung a_j: $f_j = f(a_j) = h_j/n$, wobei gilt: $\sum_{j=1}^{k} f(a_j) = 1$

Häufigkeitsfunktion (-verteilung):

$$\text{absolut:} \quad h(x) = \begin{cases} h_j, & x = a_j, \quad j = 1, \ldots, k \\ 0 & \text{sonst} \end{cases}$$

$$\text{relativ:} \quad f(x) = \begin{cases} f_j, & x = a_j, \quad j = 1, \ldots, k \\ 0 & \text{sonst} \end{cases}$$

Empirische Verteilungsfunktion (kumulierte relative Häufigkeitsverteilung):

$$F(x) = \sum_{a_j \leq x} f(a_j)$$

Klassenbildung (Gruppierung):

- k Klassen der Form $(c_0, c_1], (c_1, c_2], \ldots, (c_{k-1}, c_k]$
- Klassenbreite: $d_j = c_j - c_{j-1} \qquad j = 1, \ldots, k$
- Klassenmitte: $m_j = (c_j + c_{j-1})/2$
- absolute Häufigkeit der Klasse j: $h_j = \sum_{a_i \in (c_{j-1}, c_j]} h(a_i)$
- relative Häufigkeit der Klasse j: $f_j = h_j/n$

Histogramm (flächentreue Häufigkeitsverteilung):

$$\text{„Blockhöhe“} = \hat{f}(x) = \begin{cases} f_j/d_j & x \in (c_{j-1}, c_j] \\ 0 & \text{sonst} \end{cases} \qquad \text{für} \quad j = 1, \ldots, k.$$

1.2 Lageparameter

Modus:

x_{mod}: Ausprägung mit der größten Häufigkeit

Median:

$$x_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}+1\right)} + x_{\left(\frac{n}{2}\right)}\right) & n \text{ gerade} \end{cases}$$

Arithmetisches Mittel (Durchschnittswert):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Gewichtetes arithmetisches Mittel:

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i \quad \text{wobei} \quad \sum_{i=1}^{n} w_i = 1 \quad \text{und} \quad 0 \leq w_i \leq 1 \quad \text{für alle } i$$

Spezialfall: arithmetisches Mittel für $w_i = 1/n$.

Arithmetisches Mittel bei Schichtenbildung:

$$\bar{x} = \frac{1}{n}(n_1\bar{x}_1 + \ldots + n_r\bar{x}_r) = \frac{1}{n}\sum_{j=1}^{r} n_j\bar{x}_j$$

1.3 Quantile

Jeder Wert x_p, mit $0 < p < 1$, für den mindestens ein Anteil p der Daten $\leq x_p$ und mindestens ein Anteil $1-p$ der Daten $\geq x_p$ ist, heißt *p-Quantil:*

$$\frac{\text{Anzahl}\,(x\text{-Werte} \leq x_p)}{n} \geq p \quad \text{und} \quad \frac{\text{Anzahl}\,(x\text{-Werte} \geq x_p)}{n} \geq 1-p.$$

Äquivalent dazu ist: x_p ist der kleinste x-Wert, für den $F(x) \geq p$ gilt, d.h.

$$F(x) < p \text{ für } x < x_p \quad \text{und} \quad F(x_p) \geq p.$$

Damit gilt für das *p-Quantil:*

$$\begin{aligned} x_p &= x_{([np]+1)} && \text{wenn } np \text{ nicht ganzzahlig,} \\ x_p &\in [x_{(np)}, x_{(np+1)}] && \text{wenn } np \text{ ganzzahlig.} \end{aligned}$$

Speziell: $x_{0.5}$ = Median, $x_{0.25}$ = unteres Quartil, $x_{0.75}$ = oberes Quartil

1.4 Streuungsparameter

Spannweite:

$$SP = x_{(n)} - x_{(1)} = x_{max} - x_{min}$$

Quantilsabstand:

$$x_{1-p} - x_p$$

Interquartilsabstand:

$$d_Q = x_{0.75} - x_{0.25}$$

Empirische Varianz (mittlere quadratische Abweichung):

Urliste: $$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$$

Häufigkeitsdaten: $$\tilde{s}^2 = \frac{1}{n}\sum_{j=1}^{k}(a_j - \bar{x})^2 h(a_j) \quad \text{bzw.} \quad \tilde{s}^2 = \sum_{j=1}^{k}(a_j - \bar{x})^2 f(a_j)$$

Empirische Standardabweichung:

$$\tilde{s} = +\sqrt{\tilde{s}^2}$$

Stichprobenvarianz:

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

Variationskoeffizient:

$$v = \frac{\tilde{s}}{\bar{x}} \quad (\bar{x} > 0)$$

Streuungszerlegung:

Für r disjunkte statistische Massen $E_1, \ldots, E_r$, deren jeweilige arithmetische Mittel bzw. mittlere quadratische Abweichungen mit $\bar{x}_1, \ldots, \bar{x}_r$ bzw. $\tilde{s}_1^2, \ldots, \tilde{s}_r^2$ bezeichnet sind, berechnet sich die mittlere quadratische Abweichung für die Gesamtmasse folgendermaßen:

$$\tilde{s}^2_{Ges} = \frac{1}{n}\sum_{j=1}^{r} n_j\tilde{s}_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x}_{Ges})^2$$

wobei $n_j = |E_j|$ und $\bar{x}_{Ges} = \frac{1}{n}\sum_{j=1}^{r} n_j\bar{x}_j$.

1.5 Konzentrationsmaße

Lorenzkurve:

Für die geordnete Urliste $x_1 \leq \ldots \leq x_n$ ergibt sich die *Lorenzkurve* als Streckenzug durch die Punkte

$$(0,0),(u_1,v_1),\ldots,(u_n,v_n)=(1,1)$$

mit

$$u_j = j/n \qquad \text{und} \qquad v_j = \frac{\sum_{i=1}^{j} x_i}{\sum_{i=1}^{n} x_i}.$$

Bei Häufigkeitsdaten:

$$u_j = \sum_{i=1}^{j} h_i/n = \sum_{i=1}^{j} f_i \qquad \text{und} \qquad v_j = \frac{\sum_{i=1}^{j} h_i a_i}{\sum_{i=1}^{k} h_i a_i} = \frac{\sum_{i=1}^{j} f_i a_i}{\sum_{i=1}^{k} f_i a_i} \quad j = 1,\ldots,k\,.$$

Gini-Koeffizient:

$$\begin{aligned} G &= \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Flächezwischen Diagonale und } u\text{-Achse}} \\ &= 2\cdot\text{Fläche zwischen Diagonale und Lorenzkurve} \end{aligned}$$

Geordnete Urliste:

$$G = \frac{2\sum_{i=1}^{n} i x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}.$$

Häufigkeitsdaten mit $a_1 < \ldots < a_k$:

$$G = \frac{\sum_{i=1}^{k}(u_{i-1}+u_i)h_i a_i}{\sum_{i=1}^{k} h_i a_i} - 1\,, \quad \text{wobei} \quad u_i = \sum_{j=1}^{i} h_j/n, \quad v_i = \sum_{j=1}^{i} h_j a_j / \sum_{j=1}^{k} h_j a_j.$$

Herfindahl-Index:

$$H = \sum_{i=1}^{n} p_i^2\,, \quad \text{wobei} \quad p_i = \frac{x_i}{\sum_{j=1}^{n} x_j}.$$

2 Zweidimensionale Merkmale

2.1 Gemeinsame Häufigkeiten, Randhäufigkeiten, bedingte Häufigkeiten

Bezeichnungen:

- Urliste: $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$
- Merkmalsausspägungen: $a_1,a_2,\ldots,a_k$ für X bzw. $b_1,b_2,\ldots,b_m$ für Y

Gemeinsame Häufigkeiten:

$$\begin{aligned} h_{ij} &= h(a_i,b_j) && \text{absolute Häufigkeiten} \\ f_{ij} &= f(a_i,b_j) = \frac{h_{ij}}{n} && \text{relative Häufigkeiten} \end{aligned}$$

Randhäufigkeiten:

$$h_{i\cdot} = \sum_{j=1}^{m} h_{ij}, \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij} \qquad \text{(absolut)}$$

$$f_{i\cdot} = \frac{h_{i\cdot}}{n} = f_X(a_i), \quad f_{\cdot j} = \frac{h_{\cdot j}}{n} = f_Y(b_j) \qquad \text{(relativ)}$$

Bedingte relative Häufigkeiten:

$$f_X(a_i|b_j) = \frac{f(a_i,b_j)}{f_Y(b_j)} = \frac{h_{ij}}{h_{\cdot j}}, \quad f_Y(b_j|a_i) = \frac{f(a_i,b_j)}{f_X(a_i)} = \frac{h_{ij}}{h_{i\cdot}}$$

2.2 Assoziation bei nominalen Merkmalen

χ^2-Koeffizient:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}} = n\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(f_{ij}-\tilde{f}_{ij})^2}{\tilde{f}_{ij}}$$

wobei $\tilde{h}_{ij} = \frac{h_{i\cdot}h_{\cdot j}}{n}, \tilde{f}_{ij} = f_{i\cdot}f_{\cdot j}$.

Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{\chi^2+n}}$$

korrigierter Kontingenzkoeffizient:

$$K^* = \frac{K}{K_{max}}$$

mit: $K_{max} = \sqrt{\frac{M-1}{M}}$, wobei $M = \min\{k;m\}$.

Spezialfall: Vierfeldertafel

χ^2-Koeffizient:

a	b	a+b
c	d	c+d
a+c	b+d	n

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

Kreuzproduktverhältnis (Odds-Ratio, relative Chance):

h_{11}	h_{12}
h_{21}	h_{22}

$$\gamma = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{12}h_{21}}.$$

2.3 Korrelationsrechnung für metrische und ordinale Merkmale

Bravais-Pearson-Korrelationskoeffizient

$$r = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}$$

Rangkorrelationskoeffizient von Spearman

$$r_{SP} = \frac{\sum_{i=1}^n (rg(x_i)-\bar{rg}_X)(rg(y_i)-\bar{rg}_Y)}{\sqrt{\sum_{i=1}^n (rg(x_i)-\bar{rg}_X)^2 \sum_{i=1}^n (rg(y_i)-\bar{rg}_Y)^2}} = 1 - \frac{6\sum_{i=1}^n (rg(x_i)-rg(y_i))^2}{(n^2-1)n},$$

wobei letzterer Ausdruck nur gilt, falls $x_i \neq x_j, y_i \neq y_j$ für alle i, j gilt.

2.4 Regressionsrechnung

Lineare Einfachregression:

Y abhängige (zu erklärende) Variable, Zielgröße, Regressand
X unabhängige (erklärende) Variable, Einflußgröße, Regressor

Regressionsansatz:

$$Y = f(X) + \epsilon = \alpha + \beta X + \epsilon$$

Bezeichnungen:

- geschätzte Regressionsgerade: $\hat{y} = \hat{\alpha} + \hat{\beta}x$
- Regressionskoeffizienten: $\hat{\alpha}$ und $\hat{\beta}$
- Residuen: $\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$

Kleinste-Quadrate-Schätzer:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Quadratsummenzerlegung:

$$\underbrace{\sum_{i=1}^n (y_i-\bar{y})^2}_{SQT} = \underbrace{\sum_{i=1}^n (\hat{y}_i-\bar{y})^2}_{SQE} + \underbrace{\sum_{i=1}^n (y_i-\hat{y}_i)^2}_{SQR}$$

SQT: Gesamtabweichungsquadratsumme in y-Richtung
SQE: Durch die Regression erklärter Teil von SQT
SQR: Trotz der Regression unerklärt bleibender Teil von SQT

Bestimmtheitsmaß:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}, \qquad \text{Berechnung:} \quad R^2 = \frac{\sum_{i=1}^n (\hat{y}_i-\bar{y})^2}{\sum_{i=1}^n (y_i-\bar{y})^2} = \frac{\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}$$

3 Elementare Wahrscheinlichkeitsrechnung

3.1 Rechenregeln für Wahrscheinlichkeiten

Axiome von Kolmogoroff:

- $P(A) \geq 0$ für jedes Ereignis A
- $P(\Omega) = 1$
- $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$
  für endlich oder abzählbar unendlich viele paarweise disjunkte Ereignisse, d.h. Ereignisse mit $A_i \cap A_j = \emptyset$ für alle $i \neq j$

Folgerungen:

- $P(A) \leq 1$
- $P(\emptyset) = 0$
- Aus $A \subset B$ folgt $P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$

Allgemeiner Additionssatz (für $n = 2$):

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

3.2 Bedingte Wahrscheinlichkeiten

Definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{falls} \quad P(B) > 0$$

Folgerungen:

- $P(A \cap B) = P(A|B)P(B), \quad$ falls $\quad P(B) > 0$
- $P(B \cap A) = P(B|A)P(A), \quad$ falls $\quad P(A) > 0$
- $P(A_1 \cap \ldots \cap A_m) = P(A_1)P(A_2|A_1) \cdot \ldots \cdot P(A_m|A_1 \cap \ldots \cap A_{m-1})$, falls $P(A_1 \cap \ldots \cap A_{m-1}) > 0$.

3.3 Satz von der totalen Wahrscheinlichkeit

Sei $A_1, \ldots, A_k$ eine disjunkte Zerlegung von Ω, dann gilt für jedes Ereignis $B \subset \Omega$

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

3.4 Formel von Bayes

Sei $A_1, \ldots, A_k$ eine disjunkte Zerlegung von Ω, wobei für mindestens ein $i, \ i = 1, \ldots, k$, $P(A_i) > 0$ und $P(B|A_i) > 0$ erfüllt ist, dann gilt

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} = \frac{P(B|A_j)P(A_j)}{P(B)} \quad \text{für jedes } j = 1, \ldots, k.$$

3.5 Unabhängigkeit zweier Ereignisse

Zwei Ereignisse A und B heißen (stochastisch) unabhängig, wenn gilt

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

bzw. $\quad P(A|B) = P(A)$, falls $P(B) > 0 \quad$ bzw. $\quad P(B|A) = P(B)$, falls $P(A) > 0$.

4 Eindimensionale Zufallsvariablen und ihre Verteilungen

4.1 Definition von diskreten und stetigen Zufallsvariablen und Dichten

Diskrete Zufallsvariable:

Eine ZV X heißt *diskret*, falls der Wertebereich von X nur endlich oder abzählbar unendlich viele Werte $x_1, x_2, \ldots, x_k, \ldots$ annehmen kann.

Wahrscheinlichkeitsfunktion (Dichte):

$$f(x) = \begin{cases} P(X = x_i) = p_i & \text{für} \quad x = x_i \in \{x_1, x_2, \ldots, x_k, \ldots\} \\ 0 & \text{sonst} \end{cases}$$

Stetige Zufallsvariable:

Eine ZV X heißt *stetig*, wenn es eine Funktion $f(x) \geq 0$ gibt, so daß für jedes Intervall $[a, b]$ gilt:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Die Funktion $f(x)$ heißt **Dichte** von X
- Es gilt: $P(X = x) = 0$ für alle $x \in \mathbb{R}$.

4.2 Die Verteilungsfunktion einer Zufallsvariablen

Verteilungsfunktion einer Zufallsvariablen:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Rechenregeln:

- $P(X = x) = F(x) - P(X < x)$
- $P(X > x) = 1 - F(x)$
- $P(a < X \leq b) = F(b) - F(a)$, falls $a < b$
- $P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$
  speziell: $P(a \leq X \leq b) = F(b) - F(a)$ für stetige Verteilungen
- $P(a < X < b) = F(b) - F(a) - P(X = b)$
- $P(a \leq X < b) = F(b) - F(a) + P(X = a) - P(X = b)$

4.3 Zusammenhänge zwischen Dichten und Verteilungsfunktionen

Im diskreten Fall:

$$\begin{aligned} F(x) &= \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} P(X = x_i) \\ P(a \le X \le b) &= \sum_{a \le x_i \le b} f(x_i) = \sum_{a \le x_i \le b} P(X = x_i) \\ f(x_i) = P(X = x_i) &= F(x_i) - F(x_{i-1}) \end{aligned}$$

Im stetigen Fall:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)dt \\ P(a \le X \le b) &= \int_{a}^{b} f(t)dt \end{aligned}$$

$$f(x) = F'(x) = \frac{dF(x)}{dx}, \quad \text{falls } F(x) \text{ an der Stelle } x \text{ differenzierbar ist.}$$

4.4 Modus, Median und Quantile

Modus:

x_{mod}: Jeder Wert x, an dem $f(x)$ maximal ist.

Quantile:

- Jeder Wert x_p mit $0 < p < 1$, für den

$$P(X \ge x_p) \ge 1 - p \quad \text{und} \quad P(X \le x_p) \ge p$$

gilt, heißt *p-Quantil* einer diskreten Verteilung.

- Jeder Wert x_p mit $F(x_p) = p$ heißt *p-Quantil* einer stetigen Verteilung.

Median:

Jedes 50%-Quantil heißt Median (p = 0.5).

4.5 Erwartungswert, Varianz und Standardabweichung

Erwartungswert:

$$\mu = \mathrm{E}(X) = \begin{cases} \sum_{x} x f(x), & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{+\infty} x f(x) dx, & \text{falls } X \text{ stetig} \end{cases}$$

Varianz:

$$\sigma^2 = \mathrm{Var}(X) = \begin{cases} \sum_{x} (x - \mathrm{E}(X))^2 f(x), & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{+\infty} (x - \mathrm{E}(X))^2 f(x) dx, & \text{falls } X \text{ stetig} \end{cases}$$

Standardabweichung:

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\mathrm{Var}(X)}$$

4.6 Rechenregeln und Eigenschaften von Erwartungswerten

Transformation:

- Die Zufallsvariable $Y = g(X)$ besitzt den Erwartungswert

$$\mathrm{E}(Y) = \begin{cases} \sum_{x} g(x) f(x), & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{+\infty} g(x) f(x) dx, & \text{falls } X \text{ stetig} \end{cases}$$

- Spezialfall: lineare Transformation

$$\begin{aligned} \mathrm{E}(aX + b) &= a\mathrm{E}(X) + b \quad \text{für alle} \quad a, b \in \mathbb{R} \\ \mathrm{Var}(aX + b) &= a^2 \mathrm{Var}(X) \end{aligned}$$

Verschiebungssatz:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

4.7 Spezielle diskrete Verteilungen

Verteilung	Wahrscheinlichkeitsfunktion	E(X)	Var(X)
$X \sim B(n, \pi)$ Binomialverteilung	$f(x) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & \text{für } x = 0, 1, \ldots, n \\ 0, & \text{sonst} \end{cases}$	$n\,\pi$	$n\,\pi(1-\pi)$
$X \sim Po(\lambda)$ Poissonverteilung	$f(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{für } x = 0, 1, 2, \ldots \\ 0, & \text{sonst} \end{cases} \quad (\lambda > 0)$	λ	λ
$X \sim G(\pi)$ Geometrische Verteilung	$f(x) = (1-\pi)^{x-1}\pi \quad \text{für} \quad x = 1, 2, \ldots$	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$

4.8 Spezielle stetige Verteilungen

Verteilung	Dichte	E(X)	Var(X)
$X \sim U[a;b]$ Gleichverteilung	$f(x) = \begin{cases} \frac{1}{b-a}, & \text{für} \quad a \leq x \leq b \\ 0, & \text{sonst} \end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
$X \sim Ex(\lambda)$ Exponentialverteilung	$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für} \quad x \geq 0 \\ 0 & \text{sonst} \end{cases} \quad (\lambda > 0)$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
$X \sim N(\mu, \sigma^2)$ Normalverteilung	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(\dfrac{-(x-\mu)^2}{2\sigma^2}\right) \quad \text{für} \quad x \in \mathbb{R}$	μ	σ^2

Formeln zur Normalverteilung

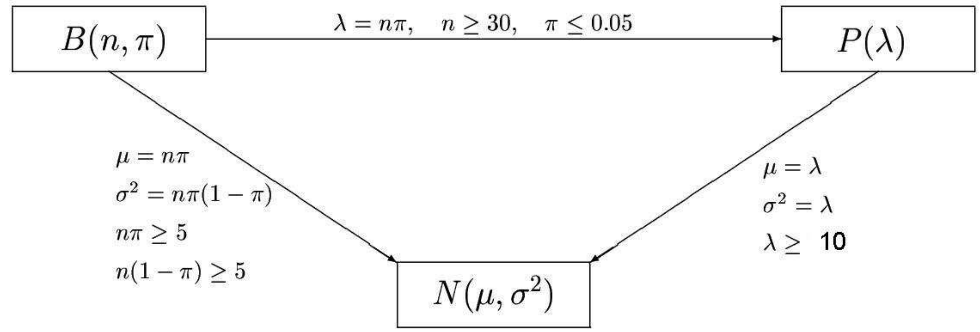
Seien $X \sim N(\mu, \sigma^2)$, $Z \sim N(0,1)$, und x_p bzw. z_p die p-Quantile von X bzw. Z. Ferner sei $\Phi(z) = P(Z \leq z)$ die Verteilungsfunktion von Z. Es gilt:

- $z_{1-p} = -z_p$
- $x_p = \mu + \sigma \cdot z_p$
- $\Phi(-z) = 1 - \Phi(z)$
- $\frac{X-\mu}{\sigma} \sim N(0,1)$ (Standardisieren einer Zufallsvariable)
- $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

Seien $X_1 \sim N(\mu_1, \sigma_1^2)$ und $X_2 \sim N(\mu_2, \sigma_2^2)$ unabhängig. Dann gilt:

- $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- $X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

4.9 Approximationsmöglichkeiten von Verteilungen



4.10 Die Chi-Quadrat-, Student- und Fisher-Verteilung

Chi-Quadrat-(χ^2-)Verteilung:

$$Z = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \qquad \text{d.h. } \chi^2\text{-verteilt mit } n \text{ Freiheitsgraden}$$

falls $X_1, \ldots, X_n$ unabhängige, standardnormalverteilte Zufallsvariablen sind

Student-(t-)Verteilung:

$$T = \frac{X}{\sqrt{Z/n}} \sim t(n) \qquad \text{d.h. } t\text{-verteilt mit } n \text{ Freiheitsgraden}$$

falls X standardnormalverteilt, Z $\chi^2(n)$-verteilt und X und Z unabhängig sind

Fisher-(F-)Verteilung:

$$Z = \frac{X/m}{Y/n} \sim F(m,n) \qquad \text{d.h. } F\text{-verteilt mit } m \text{ und } n \text{ Freiheitsgraden}$$

falls X $\chi^2(m)$- und Y $\chi^2(n)$-verteilt und unabhängig sind.

5 Zweidimensionale Zufallsvariablen und ihre Verteilungen

5.1 Definition zweidimensionaler Zufallsvariablen

Zweidimensionale diskrete Zufallsvariable:

Seien X und Y zwei diskrete ZV, wobei X die Werte $x_1, x_2, \ldots$ und Y die Werte $y_1, y_2, \ldots$ annehmen kann, so ist (X, Y) eine *zweidimensionale diskrete* Zufallsvariable mit Werten $(x_i, y_j),\ i = 1, 2, \ldots,\ j = 1, 2, \ldots$.

Gemeinsame Wahrscheinlichkeitsfunktion:

$$f(x,y) = \begin{cases} P(X = x, Y = y) & \text{für } (x,y) \in \{(x_1,y_1),(x_1,y_2),\ldots\} \\ 0 & \text{sonst} \end{cases}$$

Randverteilungen:

$$f_X(x) = P(X = x) = \sum_j f(x, y_j) \quad \text{und} \quad f_Y(y) = P(Y = y) = \sum_i f(x_i, y)$$

Zweidimensionale stetige Zufallsvariable:

Die Zufallsvariablen X und Y sind *gemeinsam stetig verteilt*, wenn es eine **zweidimensionale Dichtefunktion** $f(x,y) \geq 0$ gibt, so daß gilt

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dydx.$$

Randdichten:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{und} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Bedingte Wahrscheinlichkeitsfunktionen/Dichten:

$$f_X(x|y) = \frac{f(x,y)}{f_Y(y)} \quad \text{und} \quad f_Y(y|x) = \frac{f(x,y)}{f_X(x)}$$

Gemeinsame Verteilungsfunktion:

$$F(x,y) = P(X \leq x, Y \leq y) = \begin{cases} \displaystyle\sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j) & \text{(diskret)} \\ \displaystyle\int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dvdu & \text{(stetig)} \end{cases}$$

5.2 Unabhängigkeit, Kovarianz und Korrelation

Unabhängigkeit von zwei Zufallsvariablen:

Zwei Zufallsvariablen X und Y heißen unabhängig, wenn für alle x und y gilt

$$f(x,y) = f_X(x) f_Y(y)$$

Kovarianz:

$$\mathrm{Cov}(X,Y) = \begin{cases} \displaystyle\sum_i \sum_j f(x_i,y_j)(x_i - \mathrm{E}(X))(y_j - \mathrm{E}(Y)) & \text{(diskret)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)(x - \mathrm{E}(X))(y - \mathrm{E}(Y))dxdy & \text{(stetig)} \end{cases}$$

Verschiebungssatz:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(X \cdot Y) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$$

$$\text{mit} \quad \mathrm{E}(X \cdot Y) = \begin{cases} \displaystyle\sum_i \sum_j f(x_i,y_j) x_i y_j & \text{(diskret)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dydx & \text{(stetig)} \end{cases}$$

Lineare Transformation:

Für die Zufallsvariablen $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ gilt

$$\mathrm{Cov}(\tilde{X}, \tilde{Y}) = a_X a_Y \mathrm{Cov}(X,Y)$$

Korrelationskoeffizient:

$$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Unkorreliertheit:

Die Zufallsvariablen X und Y heißen unkorreliert, wenn gilt

$$\rho(X,Y) = 0 \quad \text{bzw.} \quad \mathrm{Cov}(X,Y) = 0$$

Varianz der Summe zweier Zufallsvariablen:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

6 Testen und Schätzen

Bezeichnungen:

- Merkmal X: metrisch oder dichotom (Bernoulliverteilt)
- Unbekannter Parameter der Verteilung von X: θ
- Stichprobenvariablen: $X_1, X_2, \ldots, X_n$
- Annahme: $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X
- Stichprobenmittelwert: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (entspricht der relativen Häufigkeit, falls X dichotom)
- Stichprobenvarianz: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Realisationen: $x_1, x_2, \ldots, x_n$
- Schätzfunktion/Schätzstatistik/Teststatistik: $T = g(X_1, \ldots, X_n)$
- Schätzwert: $\hat{\theta} = g(x_1, \ldots, x_n)$

Abkürzungen der Quantile:

z_p ... p-Quantil der Standardnormalverteilung (siehe S. 29)
$t_p(k)$ ... p-Quantil der t-Verteilung mit k Freiheitsgraden (siehe S. 30)
$\chi^2_p(k)$ ... p-Quantil der χ^2-Verteilung mit k Freiheitsgraden (siehe S. 31)

6.1 Punktschätzung

Erwartungstreue:

Eine Schätzstatistik T heißt erwartungstreu für θ, wenn gilt

$$\mathrm{E}_\theta(T) = \theta$$

Bias:

Eine nicht erwartungstreue Schätzstatistik heißt verzerrt. Die Stärke der Verzerrung wird durch den Bias angegeben

$$\mathrm{Bias}_\theta(T) = \mathrm{E}_\theta(T) - \theta$$

MSE (mittlere quadratische Abweichung):

$$\mathrm{MSE} = \mathrm{E}([T-\theta]^2) = \mathrm{Var}(T) + \mathrm{Bias}(T)^2$$

Maximum-Likelihood-Prinzip:

Der Maximum-Likelihood-Schätzwert $\hat{\theta}$ ist die Lösung der Gleichung

$$L(\hat{\theta}) = \max_\theta L(\theta)$$

6.2 Intervallschätzung

$(1-\alpha)$-Konfidenzintervall:

Die beiden Schätzstatistiken

$$G_u = g_u(X_1, \ldots, X_n) \quad \text{und} \quad G_o = g_o(X_1, \ldots, X_n)$$

bilden ein $(1-\alpha)$-Konfidenzintervall für θ, falls gilt

$$P(G_u \le G_o) = 1 \quad \text{und} \quad P(G_u \le \theta \le G_o) = 1 - \alpha$$

Das Konfidenzintervall besitzt dann die Gestalt

$$\mathrm{KI} = [g_u(x_1, \ldots, x_n), g_o(x_1, \ldots, x_n)]$$

Einseitige $(1-\alpha)$-Konfidenzintervalle:

Für $G_u = -\infty$ bzw. $G_o = \infty$ ergibt sich

$$P(\theta \le G_o) = 1 - \alpha \quad \text{bzw.} \quad P(G_u \le \theta) = 1 - \alpha$$

und man erhält die Konfidenzintervalle

$$\mathrm{KI} = (-\infty, g_o(x_1, \ldots, x_n)] \quad \text{bzw.} \quad \mathrm{KI} = [g_u(x_1, \ldots, x_n), \infty)$$

6.3 Spezielle Schätzprobleme

Verteilung	θ	$\hat{\theta}$	$(1-\alpha)$-Konfidenzintervall
$X \sim N(\mu, \sigma^2)$, σ^2 bekannt	μ	$\bar{X}$	$\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
$X \sim N(\mu, \sigma^2)$, σ^2 unbekannt	μ	$\bar{X}$	$\left[\bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right]$
$X \sim N(\mu, \sigma^2)$, μ beliebig	σ^2	S^2	$\left[(n-1)S^2\frac{1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)},\ (n-1)S^2\frac{1}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$

Verteilung	θ	$\hat{\theta}$	approximatives $(1-\alpha)$-Konfidenzintervall
X beliebig verteilt, $n \geq 30$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ bekannt	μ	$\bar{X}$	$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$
X beliebig verteilt, $n \geq 30$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ unbekannt	μ	$\bar{X}$	$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$
X dichotom $\Leftrightarrow \sum_{i=1}^{n} X_i \sim B(n,\pi),\ n \geq 30$	π	$\bar{X}$	$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$

6.4 Testen von Hypothesen

Statistisches Testproblem:

Nullhypothese H_0 und Alternative H_1 treffen Aussagen über θ

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

Die Entscheidung für oder gegen H_0 wird anhand einer Prüfgröße (Teststatistik) getroffen.

Fehlentscheidung:

- Fehler 1.Art: H_0 wird verworfen, obwohl H_0 zutrifft
- Fehler 2.Art: H_0 wird beibehalten, obwohl H_1 zutrifft

Signifikanztest:

Falls gilt

$$P(H_0 \text{ verwerfen}|H_0 \text{ trifft zu}) \leq \alpha \qquad (\text{d.h. } P(\text{Fehler 1.Art}) \leq \alpha),$$

dann heißt der Test Signifikanztest, oder Test zum *Signifikanzniveau* α

p-Wert:

Der p-Wert ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu erhalten. Ist der p-Wert kleiner als ein vorgegebenes Signifikanzniveau α, so wird H_0 verworfen.

Gütefunktion:

$$g(\theta) = P(H_0 \text{ wird abgelehnt}|\theta \text{ liegt zugrunde})$$

6.5 Spezielle Testprobleme

6.5.1 Einstichproben-Testprobleme

Formulierung der Hypothesen:

- Zweiseitiges Testproblem:

$$(a) \qquad H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

- Einseitige Testprobleme:

$$(b) \qquad H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0$$

$$(c) \qquad H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

Verteilung	θ	Teststatistik	Ablehnbereiche
$X \sim N(\mu, \sigma^2)$, σ^2 bekannt	μ	$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$	(a) $\lvert Z\rvert > z_{1-\frac{\alpha}{2}}$ (b) $Z < -z_{1-\alpha}$ (c) $Z > z_{1-\alpha}$
$X \sim N(\mu, \sigma^2)$, σ^2 unbekannt	μ	$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$	(a) $\lvert T\rvert > t_{1-\frac{\alpha}{2}}(n-1)$ (b) $T < -t_{1-\alpha}(n-1)$ (c) $T > t_{1-\alpha}(n-1)$
X beliebig verteilt, $n \geq 30$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ bekannt	μ	$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$	(a) $\lvert Z\rvert > z_{1-\frac{\alpha}{2}}$ (b) $Z < -z_{1-\alpha}$ (c) $Z > z_{1-\alpha}$
X beliebig verteilt, $n \geq 30$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ unbekannt	μ	$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$	(a) $\lvert T\rvert > z_{1-\frac{\alpha}{2}}$ (b) $T < -z_{1-\alpha}$ (c) $T > z_{1-\alpha}$
X dichotom $\Leftrightarrow \sum_{i=1}^{n} X_i \sim B(n,\pi)$	π	$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}$	(a) $\lvert Z\rvert > z_{1-\frac{\alpha}{2}}$ (b) $Z < -z_{1-\alpha}$ (c) $Z > z_{1-\alpha}$

6.5.2 Zweistichproben-Mittelwertsvergleiche

Bezeichnungen:

- Metrische Merkmale X und Y
- Unbekannte Parameter: $\mathrm{E}(X) = \mu_X$ und $\mathrm{E}(Y) = \mu_Y$
- Stichprobenvariablen: $X_1, X_2, \ldots, X_n$ und $Y_1, Y_2, \ldots, Y_m$
- Annahmen: $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X
  $Y_1, \ldots, Y_m$ unabhängig und identisch verteilt wie Y
  $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ unabhängig

Formulierung der Hypothesen:

- Zweiseitiges Testproblem:

$$(a) \qquad H_0: \mu_X - \mu_Y = \delta_0 \quad \text{vs.} \quad H_1: \mu_X - \mu_Y \neq \delta_0$$

- Einseitige Testprobleme:

$$(b) \qquad H_0: \mu_X - \mu_Y \geq \delta_0 \quad \text{vs.} \quad H_1: \mu_X - \mu_Y < \delta_0$$

$$(c) \qquad H_0: \mu_X - \mu_Y \leq \delta_0 \quad \text{vs.} \quad H_1: \mu_X - \mu_Y > \delta_0$$

Verteilung	Teststatistik	Ablehnbereiche
$X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 bekannt	$Z = \dfrac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$	(a) $\lvert Z\rvert > z_{1-\frac{\alpha}{2}}$ (b) $Z < -z_{1-\alpha}$ (c) $Z > z_{1-\alpha}$
$X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$ unbekannt	$T = \dfrac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}}$	(a) $\lvert T\rvert > t_{1-\frac{\alpha}{2}}(n+m-2)$ (b) $T < -t_{1-\alpha}(n+m-2)$ (c) $T > t_{1-\alpha}(n+m-2)$
X, Y beliebig verteilt σ_X^2, σ_Y^2 unbekannt, $n, m > 30$	$T = \dfrac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$	(a) $\lvert T\rvert > z_{1-\frac{\alpha}{2}}$ (b) $T < -z_{1-\alpha}$ (c) $T > z_{1-\alpha}$

6.5.3 Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rang-Test:

- Annahmen: $X_1, \dots, X_n$ unabhängig und identisch verteilt wie X, X metrisch skaliert, Verteilungsfunktion stetig und symmetrisch.
- Hypothesen:

$$\begin{array}{lll} (a) & H_0: x_{med} = \delta_0 & H_1: x_{med} \neq \delta_0 \\ (b) & H_0: x_{med} \geq \delta_0 & H_1: x_{med} < \delta_0 \\ (c) & H_0: x_{med} \leq \delta_0 & H_1: x_{med} > \delta_0 \end{array}$$

- Teststatistik: $W^+ = \sum_{i=1}^n rg|D_i|Z_i$ mit $D_i = X_i - \delta_0$, $Z_i = \begin{cases} 1 & D_i > 0 \\ 0 & D_i \leq 0 \end{cases}$

  Für $n > 20$ ist W^+ approximativ verteilt nach $N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$.
- Ablehnungsbereich:

$$\begin{array}{ll} (a) & W^+ < w^+_{\alpha/2}(n) \quad \text{oder} \quad W^+ > w^+_{1-\alpha/2}(n) \\ (b) & W^+ < w^+_{\alpha}(n) \\ (c) & W^+ > w^+_{1-\alpha}(n) \end{array}$$

  wobei $w^+_\alpha(n)$ das tabellierte α-Quantil der Verteilung von W^+ ist.

6.5.4 Weitere Testprobleme

Korrelationstest:

- Annahmen: Unabhängige gemeinsam normalverteilte Stichprobenvariablen $(X_i, Y_i),\ i = 1, \dots, n$
- Hypothesen:

$$\begin{array}{llll} (a) & H_0: \rho_{XY} = 0 & \text{vs.} & H_1: \rho_{XY} \neq 0 \\ (b) & H_0: \rho_{XY} \geq 0 & \text{vs.} & H_1: \rho_{XY} < 0 \\ (c) & H_0: \rho_{XY} \leq 0 & \text{vs.} & H_1: \rho_{XY} > 0 \end{array}$$

- Teststatistik:

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}\sqrt{n-2}$$

- Ablehnungsbereiche:

$$\begin{array}{ll} (a) & |T| > t_{1-\frac{\alpha}{2}}(n-2) \\ (b) & T < -t_{1-\alpha}(n-2) \\ (c) & T > t_{1-\alpha}(n-2) \end{array}$$

χ^2-Unabhängigkeitstest:

- Annahmen: Unabhängige Stichprobenvariablen $(X_i, Y_i),\ i = 1, \dots, n$
- Hypothesen:

$$\begin{array}{lll} H_0: & P(X = i, Y = j) = P(X = i) \cdot P(Y = j) & \text{für alle } i, j \\ H_1: & P(X = i, Y = j) \neq P(X = i) \cdot P(Y = j) & \text{für mindestens ein Paar } (i, j) \end{array}$$

$$\begin{array}{cc|ccc|c} & & & Y & & \\ & & 1 & \dots & m & \\ \hline & 1 & h_{11} & \dots & h_{1m} & h_{1\cdot} \\ X & \vdots & \vdots & & \vdots & \vdots \\ & k & h_{k1} & \dots & h_{km} & h_{k\cdot} \\ \hline & & h_{\cdot 1} & \dots & h_{\cdot m} & n \end{array} \quad \xrightarrow{\text{unter } H_0} \quad \begin{array}{cc|ccc|c} & & & Y & & \\ & & 1 & \dots & m & \\ \hline & 1 & \frac{h_{1\cdot} h_{\cdot 1}}{n} & \dots & \frac{h_{1\cdot} h_{\cdot m}}{n} & h_{1\cdot} \\ X & \vdots & \vdots & & \vdots & \vdots \\ & k & \frac{h_{k\cdot} h_{\cdot 1}}{n} & \dots & \frac{h_{k\cdot} h_{\cdot m}}{n} & h_{k\cdot} \\ \hline & & h_{\cdot 1} & \dots & h_{\cdot m} & n \end{array}$$

- Teststatistik:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \quad \text{mit} \quad \tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$$

- Ablehnungsbereich:

$$\chi^2 > \chi^2_{1-\alpha}((k-1) \cdot (m-1))$$

χ^2-Homogenitätstest:

- Annahmen: Unabhängige Stichproben aus k Populationen mit den Stichprobenumfängen $n_1, \dots, n_k$
- Hypothesen:

$$\begin{array}{ll} H_0: & P(X_1 = j) = \ldots = P(X_k = j) \quad \text{für } j = 1 \ldots, m \\ H_1: & P(X_{i_1} = j) \neq P(X_{i_2} = j) \quad \text{für mindestens ein Tupel } (i_1, i_2, j) \end{array}$$

$$\begin{array}{c|ccc|c} & & X & & \\ & 1 & \ldots & m & \\ \hline 1 & h_{11} & \ldots & h_{1m} & n_1 \\ \vdots & \vdots & & \vdots & \vdots \\ \text{k} & h_{k1} & \ldots & h_{km} & n_k \\ \hline & h_{\cdot 1} & \ldots & h_{\cdot m} & n \end{array} \quad \overset{\text{unter } H_0}{\longrightarrow} \quad \begin{array}{c|ccc|c} & & X & & \\ & 1 & \ldots & m & \\ \hline 1 & \frac{n_1 h_{\cdot 1}}{n} & \ldots & \frac{n_1 h_{\cdot m}}{n} & n_1 \\ \vdots & \vdots & & \vdots & \vdots \\ \text{k} & \frac{n_k h_{\cdot 1}}{n} & \ldots & \frac{n_k h_{\cdot m}}{n} & n_k \\ \hline & h_{\cdot 1} & \ldots & h_{\cdot m} & n \end{array}$$

- Teststatistik:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \hat{h}_{ij})^2}{\hat{h}_{ij}} \quad \text{mit} \quad \hat{h}_{ij} = \frac{n_i h_{\cdot j}}{n}$$

- Ablehnungsbereich:

$$\chi^2 > \chi^2_{1-\alpha}((k-1) \cdot (m-1))$$

χ^2-Anpassungstest:

- Annahmen: $X_1, \dots, X_n$ unabhängig und identisch verteilt wie $X \in \{1, \dots, k\}$.
- Hypothesen:

$$\begin{array}{ll} H_0: & P(X = i) = \pi_i, \quad i = 1, \dots, k \\ H_1: & P(X = i) \neq \pi_i \quad \text{für mindestens ein } i \end{array}$$

- Teststatistik: $\chi^2 = \sum_{i=1}^{k} \frac{(h_i - n\pi_i)^2}{n\pi_i}$
- Verteilung unter H_0: approximativ $\chi^2(k-1)$. Approximation anwendbar, falls $n\pi_i \geq 1$ für alle i, $n\pi_i \geq 5$ für mindestens 80% der Zellen.
- Ablehnungsbereich: $\chi^2 > \chi^2_{1-\alpha}(k-1)$

7 Regressionsanalyse

7.1 Lineare Einfachregression

Regressionsansatz:

$$Y_i = \alpha + \beta x_i + \epsilon_i \ , \quad i = 1, \dots, n.$$

Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \quad \Leftrightarrow \quad Y_i \sim N(\alpha + \beta x_i \ , \sigma^2), \quad i = 1, \dots, n.$$

Kleinste-Quadrate-Schätzer:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \ , \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i \ , \quad i = 1, \dots, n.$$

Schätzer für die Varianz σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Verteilung der geschätzten Regressionskoeffizienten:

$$\hat{\alpha} \sim N(\alpha, \sigma^2_{\hat{\alpha}}) \quad \text{mit} \quad \text{Var}(\hat{\alpha}) = \sigma^2_{\hat{\alpha}} = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} = \sigma^2 \frac{\sum x_i^2}{n(\sum x_i^2 - n\bar{x}^2)}$$

$$\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}) \quad \text{mit} \quad \text{Var}(\hat{\beta}) = \sigma^2_{\hat{\beta}} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$$

Verteilung der standardisierten Schätzfunktionen:

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} \sim t(n-2) \quad \text{mit} \quad \hat{\sigma}_{\hat{\alpha}} = \hat{\sigma} \frac{\sqrt{\sum x_i^2}}{\sqrt{n \sum (x_i - \bar{x})^2}} = \hat{\sigma} \frac{\sqrt{\sum x_i^2}}{\sqrt{n(\sum x_i^2 - n\bar{x}^2)}}$$

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2) \quad \text{mit} \quad \hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2 - n\bar{x}^2}}$$

$(1-\alpha)$-Konfidenzintervalle für α und β:

$$\begin{array}{ll} \text{für } \alpha: & \left[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\alpha}{2}}(n-2), \ \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\alpha}{2}}(n-2)\right] \\ \text{für } \beta: & \left[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\alpha}{2}}(n-2), \ \hat{\beta} + \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\alpha}{2}}(n-2)\right] \end{array}$$

Teststatistiken:

$$T_{\alpha_0} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} \quad \text{und} \quad T_{\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

Hypothesen und Ablehnbereiche:

Hypothesen			Ablehnbereich
$H_0 : \alpha = \alpha_0$	vs.	$H_1 : \alpha \neq \alpha_0$	$\lvert T_{\alpha_0}\rvert > t_{1-\frac{\alpha}{2}}(n-2)$
$H_0 : \beta = \beta_0$	vs.	$H_1 : \beta \neq \beta_0$	$\lvert T_{\beta_0}\rvert > t_{1-\frac{\alpha}{2}}(n-2)$
$H_0 : \alpha \geq \alpha_0$	vs.	$H_1 : \alpha < \alpha_0$	$T_{\alpha_0} < -t_{1-\alpha}(n-2)$
$H_0 : \beta \geq \beta_0$	vs.	$H_1 : \beta < \beta_0$	$T_{\beta_0} < -t_{1-\alpha}(n-2)$
$H_0 : \alpha \leq \alpha_0$	vs.	$H_1 : \alpha > \alpha_0$	$T_{\alpha_0} > t_{1-\alpha}(n-2)$
$H_0 : \beta \leq \beta_0$	vs.	$H_1 : \beta > \beta_0$	$T_{\beta_0} > t_{1-\alpha}(n-2)$

Prognose:

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$$

Konfidenzintervall für Y_0:

$$\left[\hat{Y}_0 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}\,, \hat{Y}_0 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}\right]$$

7.2 Multiple lineare Regression

Regressionsansatz:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \ , \quad i = 1, \ldots, n.$$

Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \quad \Leftrightarrow \quad Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \ , \sigma^2), \quad i = 1, \ldots, n.$$

Gefittete Werte:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \ , \quad i = 1, \ldots, n.$$

Schätzer für die Varianz σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Verteilung der standardisierten Schätzfunktionen:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t(n-p-1) \ , \quad j = 0, \ldots, p$$

$(1-\alpha)$-Konfidenzintervalle für β_j:

$$\left[\hat{\beta}_j - \hat{\sigma}_j t_{1-\frac{\alpha}{2}}(n-p-1), \ \hat{\beta}_j + \hat{\sigma}_j t_{1-\frac{\alpha}{2}}(n-p-1)\right]$$

Teststatistiken:

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j} \ , \quad j = 0, \ldots, p$$

Hypothesen und Ablehnbereiche:

Hypothesen			Ablehnbereich
$H_0 : \beta_j = \beta_{0j}$	vs.	$H_1 : \beta_j \neq \beta_{0j}$	$\lvert T_j\rvert > t_{1-\frac{\alpha}{2}}(n-p-1)$
$H_0 : \beta_j \geq \beta_{0j}$	vs.	$H_1 : \beta_j < \beta_{0j}$	$T_j < -t_{1-\alpha}(n-p-1)$
$H_0 : \beta_j \leq \beta_{0j}$	vs.	$H_1 : \beta_j > \beta_{0j}$	$T_j > t_{1-\alpha}(n-p-1)$

Koeffizienten Tabelle einer multiplen Regression (R & Rcmdr):

	`Estimate`	`Std. Error`	`t value`	`Pr(>|t|)`
`(Intercept)`	$\hat{\beta}_0$	$\hat{\sigma}_0$	T_0	$P(T \geq T_0)$
X_1	$\hat{\beta}_1$	$\hat{\sigma}_1$	T_1	$P(T \geq T_1)$
X_2	$\hat{\beta}_2$	$\hat{\sigma}_2$	T_2	$P(T \geq T_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_p	$\hat{\beta}_p$	$\hat{\sigma}_p$	T_p	$P(T \geq T_p)$

Testprobleme:

- `t` ist der Wert der t-Statistik für

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \ ,$$

  d.h. $T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j}, \quad j = 0, \ldots, p$

- `Pr(>|t|)` ist der zugehörige p-Wert

8 Varianzanalyse

8.1 Einfaktorielle Varianzanalyse

Modell(I):

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Modell(II):

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\epsilon_{ij} \sim N(0, \sigma^2), \quad \sum_{i=1}^{I} n_i \alpha_i = 0$$

Die Schätzer für μ und α_i im Modell(II) sind gegeben als:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{\cdot\cdot}$$

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \quad \text{mit } \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Die Prüfgrösse für das Testproblem:

$$H_0 : \alpha_1 = \cdots = \alpha_I = 0 \quad \text{gegen} \quad H_1 : \text{ mindestens zwei} \alpha_i \neq 0$$

ist gegeben als

$$F = \frac{MQE}{MQR} = \frac{\sum_{i=1}^{I} n_i \cdot (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/(I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2/(n-I)},$$

wobei H_0 zu verwerfen ist, falls

$$F > F_{1-\alpha}(I-1, n-I).$$

Varianzanalyse Tabelle (R & Rcmdr):

`Analysis of Variance Table:`

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
var	p	SQE	MQE= $\frac{SQE}{p}$	$F_E = \frac{MQE}{MQR}$	$P(F \geq F_E)$
Residuals	$n-p-1$	SQR	MQR= $\frac{SQR}{n-p-1}$		

mit den Werten:

$$\texttt{SQE} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$\texttt{SQR} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$p = I - 1$$

`und Pr(>F) als zugehörigem p-Wert.`

9 Verteilungstabellen

9.1 Standardnormalverteilung

Tabelliert sind die Werte der Verteilungsfunktion $\Phi(z) = P(Z \leq z)$ für $z \geq 0$.
Ablesebeispiel: $\Phi(1.75) = 0.9599$
Funktionswerte für negative Argumente: $\Phi(-z) = 1 - \Phi(z)$
Die z-Quantile ergeben sich genau umgekehrt.
Beispielsweise ist $z(0.9599) = 1.75$ und $z(0.9750) = 1.96$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

9.2 Students *t*-Verteilung

Tabelliert sind die Quantile für n Freiheitsgrade.
Für das Quantil $t_{1-\alpha}(n)$ gilt $F(t_{1-\alpha}(n)) = 1 - \alpha$.
Links vom Quantil $t_{1-\alpha}(n)$ liegt die Wahrscheinlichkeitsmasse $1 - \alpha$.
Ablesebeispiel: $t_{0.99}(20) = 2.528$
Die Quantile für $0 < 1 - \alpha < 0.5$ erhält man aus $t_\alpha(n) = -t_{1-\alpha}(n)$

Approximation für $n > 30$:

$$t_\alpha(n) \approx z_\alpha \quad (z_\alpha \text{ ist das } (\alpha)\text{-Quantil der Standardnormalverteilung})$$

n	0.6	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	1.3764	3.0777	6.3138	12.706	31.821	63.657	318.31	636.62
2	0.2887	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.327	31.599
3	0.2767	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.215	12.924
4	0.2707	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.2576	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.2573	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.2571	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.2569	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.2567	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.2566	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.2564	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.2563	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.2562	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.2561	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.2560	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.2559	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.2558	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.2557	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.2556	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
∞	0.2533	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	3.0903	3.2906

9.3 χ^2-Verteilung

Tabelliert sind die Quantile für n Freiheitsgrade.
Für das Quantil $\chi^2_{1-\alpha}(n)$ gilt $F(\chi^2_{1-\alpha}(n)) = 1-\alpha$.
Links vom Quantil $\chi^2_{1-\alpha}(n)$ liegt die Wahrscheinlichkeitsmasse $1-\alpha$.
Ablesebeispiel: $\chi^2_{0.95}(10) = 18.307$

Approximation für $n > 30$:

$$\chi^2_\alpha(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2 \quad (z_\alpha \text{ ist das } \alpha\text{-Quantil der Standardnormalverteilung})$$

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
1	0.0002	0.0010	0.0039	0.0158	0.4549	2.7055	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	0.2107	1.3863	4.6052	5.9915	7.3778	9.2103
3	0.1148	0.2158	0.3518	0.5844	2.3660	6.2514	7.8147	9.3484	11.345
4	0.2971	0.4844	0.7107	1.0636	3.3567	7.7794	9.4877	11.143	13.277
5	0.5543	0.8312	1.1455	1.6103	4.3515	9.2364	11.070	12.833	15.086
6	0.8721	1.2373	1.6354	2.2041	5.3481	10.645	12.592	14.449	16.812
7	1.2390	1.6899	2.1674	2.8331	6.3458	12.017	14.067	16.013	18.475
8	1.6465	2.1797	2.7326	3.4895	7.3441	13.362	15.507	17.535	20.090
9	2.0879	2.7004	3.3251	4.1682	8.3428	14.684	16.919	19.023	21.666
10	2.5582	3.2470	3.9403	4.8652	9.3418	15.987	18.307	20.483	23.209
11	3.0535	3.8157	4.5748	5.5778	10.341	17.275	19.675	21.920	24.725
12	3.5706	4.4038	5.2260	6.3038	11.340	18.549	21.026	23.337	26.217
13	4.1069	5.0088	5.8919	7.0415	12.340	19.812	22.362	24.736	27.688
14	4.6604	5.6287	6.5706	7.7895	13.339	21.064	23.685	26.119	29.141
15	5.2293	6.2621	7.2609	8.5468	14.339	22.307	24.996	27.488	30.578
16	5.8122	6.9077	7.9616	9.3122	15.338	23.542	26.296	28.845	32.000
17	6.4078	7.5642	8.6718	10.085	16.338	24.769	27.587	30.191	33.409
18	7.0149	8.2307	9.3905	10.865	17.338	25.989	28.869	31.526	34.805
19	7.6327	8.9065	10.117	11.651	18.338	27.204	30.144	32.852	36.191
20	8.2604	9.5908	10.851	12.443	19.337	28.412	31.410	34.170	37.566
21	8.8972	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932
22	9.5425	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892

9.4 Wilcoxon-Vorzeichen-Rang-Test

Kritische Werte $w^+_{n;\gamma}$ des Vorzeichen-Rang-Tests von Wilcoxon

n	$w^+_{n;0.01}$	$w^+_{n;0.025}$	$w^+_{n;0.05}$	$w^+_{n;0.10}$	$w^+_{n;0.90}$	$w^+_{n;0.95}$	$w^+_{n;0.975}$	$w^+_{n;0.99}$
4	0	0	0	1	8	9	10	10
5	0	0	1	3	11	13	14	14
6	0	1	3	4	16	17	19	20
7	1	3	4	6	21	23	24	26
8	2	4	6	9	26	29	31	33
9	4	6	9	11	33	35	38	40
10	6	9	11	15	39	43	45	57
11	8	11	14	18	47	51	54	57
12	10	14	18	22	55	59	62	66
13	13	18	22	27	63	68	72	77
14	16	22	26	32	72	78	82	88
15	20	26	31	37	82	88	93	99
16	24	30	36	43	92	99	105	111
17	28	35	42	49	103	110	117	124
18	33	41	48	56	114	122	129	137
19	38	47	54	63	126	135	142	151
20	44	53	61	70	139	148	156	165

10 R & Rcmdr Ausgaben

10.1 Kennzahlen

Allgemeine Zusammenfassung mit einer numerischen Variable `A` und einer kategorialen (mit k Ausprägungen) Variable `B`:

```
> summary(daten)
         A                        B
Min.   :   'Minimum'   a1:'Anzahl der Auspraegung 1'
1st Qu.:'1. Quantil'   a2:'Anzahl der Auspraegung 1'
Median :    'Median'   ...
Mean   :'Mittelwert'   ak:'Anzahl der Auspraegung k'
3rd Qu.:'3. Quantil'
Max    :   'Maximum'
```

Zusammenfassung numerischer Variablen:

```
> numSummary(daten, statistics=c("mean", "sd", "quantiles")
          mean                   sd                    0%
  'Mittelwert' 'Standardabweichung' '0%-Quantil (Minimum)'

           25%           50%           75%                    100%
 '25%-Quantil' '50%-Quantil' '75%-Quantil' '100%-Quantil (Maximum)'

                          n
 'Anzahl der Beobachtungen'
```

Einzelne Kennzahlen:

```
> mean(daten)
[1] 'Mittelwert'

> sd(daten)
[1] 'Standardabweichung'

> var(daten)
[1] 'Varianz'

> median(daten)
[1] 'Median'

> IQR(daten)
[1] 'Interquartilsabstand'

> skewness(daten)
[1] 'Schiefe'

> kurtosis(daten)
[1] 'Woelbung'
```


Lorenzkurve:

```
> Lc(daten)
$p
[1] 'Anteile der Merkmalstraeger u'

$L
[1] 'kumulierten relativen Merkmalssummen v'

$L.general
[1] 'Werte fuer die generalisierte Lorenzkurve'

attr(,'class')
[1] 'Lc'
```

Konzentrationsmaße:

```
> Gini(daten)
[1] 'Ginikoeffizient'

> Herfindahl(daten)
[1] 'Herfindahlindex'
```

10.2 Korrelation

```
> cor(var1, var2, method='pearson')
[1] 'Pearson-Korrelationskoeffizient'

> cor(var1, var2, method='spearman')
[1] 'Spearmans Korrelationskoeffizient'
```

10.3 Regression

```
> reg <- lm(abh. Variable ~ var1 + var2 + ... + varP, data=daten)
> summary(reg)

Call:
lm(formula = abh. Variable ~ var1 + var2 + ... + varP, data = daten)

Residuals:
      Min            1Q    Median            3Q        Max
'Minimum'  '1. Quantil'  'Median'  '3. Quantil'  'Maximum'

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat{\beta}_0$	$\hat{\sigma}_0$	T_0	$P(T \geq T_0)$
var_1	$\hat{\beta}_1$	$\hat{\sigma}_1$	T_1	$P(T \geq T_1)$
var_2	$\hat{\beta}_2$	$\hat{\sigma}_2$	T_2	$P(T \geq T_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
var_P	$\hat{\beta}_p$	$\hat{\sigma}_p$	T_p	$P(T \geq T_p)$

```
---
Signif. codes:  0 ***0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 'RSE' on 'DF' degrees of freedom
Multiple R-squared: 'Bestimmtheitsmass R^2', Adjusted R-squared: 'Angepasstes R^2'
F-statistic: 'F' on 'DF1' and 'DF2' DF,  p-value: 'P-Wert'
```

10.4 Varianzanalyse

```
Analysis of Variance Table:
```

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
var	p	SQE	MQE=$\frac{\text{SQE}}{p}$	$F_E = \frac{\text{MQE}}{\text{MQR}}$	$P(F \geq F_E)$
Residuals	$n-p-1$	SQR	MQR=$\frac{\text{SQR}}{n-p-1}$		

mit den Werten:

$$\text{SQE} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$\text{SQR} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$p = I - 1$$

und `Pr(>F)` als zugehörigem p-Wert.

10.5 Tests

Tests in R befolgen in den meisten Fällen folgendes Ausgabeschema:

```
      Name des Tests

data: Datensatz
Teststatistik = T, ... weitere Informationen ..., p-value = P
alternative hypothesis: Beschreibung der Alternativhypothese
```

Ja nach Test kommen zu dieser Minimalausgabe noch weitere nützliche Informationen hinzu.
Als Beispiel dafür die Ausgabe eines gepaarten zweiseitigen t-Tests mit der Berechnung des 95%-Konfidenzintervalls:

```
> t.test(data$x, data$y, alternative='two.sided',
+        conf.level=.95, paired=TRUE)

      Paired t-test

data: data$x and data$y
t = T, df = DF, p-value = P
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  'untere Grenze' 'obere Grenze'
sample estimates:
mean of the differences
  'MD'
```

mit der t-Statistik `T`, der Anzahl der Freiheitsgraden `DF` und dem p-Wert `P`. Die Alternativhypothese ist die, dass die Differenz der Mittelwerte ungleich 0 ist. Zusätzlich werden noch die Grenzen des 95%-Konfidenzintervalls und der Mittelwert der paarweisen Differenzen `MD` berechnet.