Martin-Luther-Universität Halle-Wittenberg

Juristische und Wirtschaftswissenschaftliche Fakultät

Wirtschaftswissenschaftlicher Bereich

Formelsammlung Statistik

für die Veranstaltungen Statistik I und Statistik II
im Grundstudium bzw. Bachelorstudium

Prof. Dr. Claudia Becker
Lehrstuhl für Statistik

Inhaltsverzeichnis

1 Summenzeichen

$$\sum(x_i + y_i) = \sum x_i + \sum y_i$$
$$\sum_{i=1}^{n} c\,x_i = c\sum_{i=1}^{n} x_i$$
$$\sum_{i=1}^{n} c = n \cdot c$$

2 Häufigkeitsverteilungen

2.1 Absolute Häufigkeit

$h_j = h(a_j) =$ Anzahl der Fälle in denen Ausprägung a_j auftritt
$a_j =$ j-te Merkmalsausprägung
mit $j = 1, ..., k$

Es gilt: $\sum_{j=1}^{k} h_j = n$

2.2 Relative Häufigkeit

$$f_j = f(a_j) = \frac{h(a_j)}{n} \qquad \text{Es gilt: } \sum_{j=1}^{k} f_j = 1$$

2.3 Histogramm

Klasseneinteilung: bei n Beobachtungen $\approx \sqrt{n}$
Klassenbreite (d_j) = obere Klassengrenze - untere Klassengrenze $= x_j^0 - x_j^u$

$$f_j^r = \text{Höhe} = \frac{f_j}{d_j}$$

2.4 Empirische Verteilungsfunktion

Für unklassierte Häufigkeitsverteilung (Urliste muss in Häufigkeitsverteilung überführt werden)

$$F(x) = \sum_{j:a_j \leq x} f(a_j) = \sum_{j:a_j \leq x} f_j \quad \text{(kumulierte relative Häufigkeit)}$$

Für klassierte Häufigkeitsverteilung

$$F(x) = \begin{cases} 0 & , x < x_1^u \\ kumf_{j-1} + \frac{x - x_j^u}{d_j} \cdot f_j & , x_1^u \leq x < x_k^o \\ 1 & , x_k^o \leq x \end{cases}$$

3 Lagemaße

3.1 Lagemaße I: Daten als Urliste

3.1.1 Arithmetisches Mittel

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(für Urliste)}$$

Bei linearer Transformation: $x_i \mapsto y_i = a \cdot x_i + b \Rightarrow \overline{y} = a \cdot \overline{x} + b$

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{r} n_j\,\overline{x}_j \left(\text{wobei} \quad n = \sum_{j=1}^{r} n_j\right), \text{Mittelwert aus Teilgesamtheiten (r Schichten)}$$

3.1.2 Geometrisches Mittel

Beobachtete Reihe des Merkmals X (Zeitreihe): $x_0, x_1, ..., x_n$

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} \qquad \text{(Wachstumsrate)}$$

$$w_t = 1 + r_t = \frac{x_t}{x_{t-1}} \qquad \text{(Wachstumsfaktor)}$$

- durchschnittlicher Wachstumsfaktor $\overline{w}_{geom}$

$$x_n = x_0 \cdot \overline{w}_{geom}^n$$

$$\overline{w}_{geom} = \sqrt[n]{\prod_{t=1}^{n} w_t} = \sqrt[n]{w_1 \cdot w_2 \cdot ... \cdot w_n}$$

$$\overline{w}_{geom} = \sqrt[n]{\frac{x_n}{x_0}} = \sqrt[n]{(1 + r_1) \cdot (1 + r_2) \cdot ... \cdot (1 + r_n)}$$

- durchschnittliche Wachstumsrate $\overline{r}_{geom}$

$$\overline{r}_{geom} = \overline{w}_{geom} - 1$$

3.1.3 Median (Zentralwert)

Ordnungsstatistiken $x_{(1)} \leq \ldots \leq x_{(n)}$

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

3.1.4 Modus (Modalwert)

Ausprägung mit größter relativer Häufigkeit.
Nicht bestimmbar, wenn mehrere Ausprägungen größte relative Häufigkeit besitzen.
Modalitätsgrad: relative Häufigkeit des Modus in Prozent $= f_{mod} \cdot 100\%$

3.1.5 p-Quantile $(0 < p < 1)$

$$x_p = \begin{cases} x_{([n \cdot p]+1)} & \text{, wenn } n \cdot p \text{ nicht ganzzahlig, wobei } [n \cdot p] \text{ die zu } n \cdot p \text{ nächst kleinere ganze Zahl} \\ \frac{1}{2}\left(x_{(n \cdot p)} + x_{(n \cdot p\,+1)}\right) & \text{, wenn } n \cdot p \text{ ganzzahlig} \end{cases}$$

Fünf-Punkte-Zusammenfassung:
Teilt den Wertebereich in 4 Intervalle die jeweils ca. ein Viertel der Werte enthalten.

$x_{(1)}$	...	kleinster Wert
$x_{0.25}$	...	unteres Quartil
x_{med}	...	Median
$x_{0.75}$	...	oberes Quartil
$x_{(n)}$	...	größter Wert

Arithmetisches Mittel

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	$\overline{x} = \frac{1}{n} \sum_{j=1}^{k} a_j \cdot h(a_j) = \sum_{j=1}^{k} a_j \cdot f(a_j)$	Nutze Klassenmitten $m_j = \frac{x_j^o + x_j^u}{2}$ $\overline{x} = \frac{1}{n} \sum_{j=1}^{k} m_j \cdot n_j = \sum_{j=1}^{k} m_j \cdot f_j \quad$ (Näherung)

p-Quantil

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
$x_p = \begin{cases} x_{([np]+1)} & , np \text{ nicht ganzzahl.} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & , np \text{ ganzzahlig} \end{cases}$	(1) Suche nach der Ausprägung a_j, bei der $kumf_j = p$ erstmals überschritten oder genau erreicht wird (2a) Wird p bei a_j überschritten: $x_p = a_j$ (2b) Wird p genau bei a_j erreicht: $x_p = \frac{a_j + a_{j+1}}{2}$	(1) Bestimme Klasse, in der $kumf_j = p$ erstmals überschritten wird (2) $x_p = x_j^u + (p - kumf_{j-1}) \cdot \frac{d_j}{f_j}$

Median

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
Nutze Rechenvorschriften für p-Quantile mit p=0.5		

Modus

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
Die Merkmalsausprägung a_j mit der größten Häufigkeit $h(a_j)$ bildet den Modus		(1) Modalklasse: Klasse j mit größter Besetzungsdichte $f_j^r = f_j / d_j$ (2) Näherung für Modus: $x_{mod} = \frac{x_j^o + x_j^u}{2}$

4 Streuungsmaße

4.1 Spannweite (Range)

$$R = x_{(n)} - x_{(1)}$$

4.2 Interquartilsabstand

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

4.3 Mediane absolute Abweichung vom Median (MAD)

$$MAD = med\left\{\, |x_i - x_{med}|,\ i = 1, \ldots, n\right\}$$

4.4 Empirische Varianz I: Daten als Urliste

$$\widetilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad = \quad \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2 \quad \text{(Verschiebungssatz)}$$

$$\widetilde{s}^2 = \frac{1}{n}\left(\sum_{j=1}^{r} n_j \cdot \widetilde{s}_j^2 + \sum_{j=1}^{r} n_j \cdot (\overline{x}_j - \overline{x})^2\right) \text{,Varianz aus Teilgesamtheiten (r Schichten)}$$

Bei linearer Transformation: $x_i \mapsto y_i = a \cdot x_i + b \Rightarrow \widetilde{s}_y^2 = a^2 \cdot \widetilde{s}_x^2$

Ist X normalverteilt (großes n) gilt:

$\overline{x} \pm \widetilde{s}$ $\rightarrow$ ca. 68% aller Beobachtungen
$\overline{x} \pm 2 \cdot \widetilde{s}$ $\rightarrow$ ca. 95% aller Beobachtungen
$\overline{x} \pm 3 \cdot \widetilde{s}$ $\rightarrow$ ca. 99% aller Beobachtungen

4.5 Empirische Varianz II: Urliste, unklassierte und klassierte Häufigkeitsverteilung

Varianz

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
$\widetilde{s}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$	$\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}(a_j - \overline{x})^2 \cdot h(a_j)$ $= \sum\limits_{j=1}^{k}(a_j - \overline{x})^2 \cdot f(a_j)$	$\widetilde{s}^2 = \widetilde{s}^2_{ext} + \widetilde{s}^2_{int}$ Einzelwerte x_{ij} in den Klassen unbekannt; Klassenmittelwerte $\overline{x}_j$ können nicht berechnet werden; Verwende daher die Klassenmitten $m_j = \frac{x_j^o + x_j^u}{2}$ (a) Es liegen Informationen über Klassenvarianzen $\widetilde{s}^2_j$ vor: $\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}(m_j - \overline{x})^2 \cdot n_j + \frac{1}{n}\sum\limits_{j=1}^{k}\widetilde{s}^2_j \cdot n_j$ $= \sum\limits_{j=1}^{k}(m_j - \overline{x})^2 \cdot f_j + \sum\limits_{j=1}^{k}\widetilde{s}^2_j \cdot f_j$ (b) Keine Informationen über $\widetilde{s}^2_j$; Setze $\widetilde{s}^2_j = 0$: $\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}(m_j - \overline{x})^2 \cdot n_j = \sum\limits_{j=1}^{k}(m_j - \overline{x})^2 \cdot f_j$

Verschiebungssatz der Varianz

Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung
$\widetilde{s}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}x_i^2 - \overline{x}^2$	$\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}a_j^2 \cdot h(a_j) - \overline{x}^2$ $= \sum\limits_{j=1}^{k}a_j^2 \cdot f(a_j) - \overline{x}^2$	(a) Es liegen Informationen über Klassenvarianzen $\widetilde{s}^2_j$ vor: $\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}m_j^2 \cdot n_j - \overline{x}^2 + \frac{1}{n}\sum\limits_{j=1}^{k}\widetilde{s}^2_j \cdot n_j$ $= \sum\limits_{j=1}^{k}m_j^2 \cdot f_j - \overline{x}^2 + \sum\limits_{j=1}^{k}\widetilde{s}^2_j \cdot f_j$ (b) Keine Informationen über $\widetilde{s}^2_j$; Setze $\widetilde{s}^2_j = 0$: $\widetilde{s}^2 = \frac{1}{n}\sum\limits_{j=1}^{k}m_j^2 \cdot n_j - \overline{x}^2 = \sum\limits_{j=1}^{k}m_j^2 \cdot f_j - \overline{x}^2$

4.6 Stichprobenvarianz

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \; = \; \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2\right) \quad \text{(Verschiebungssatz)}$$

4.7 Standardabweichung

$$\widetilde{s} = \sqrt{\widetilde{s}^2}$$

4.8 Standardisierung

$$x_i \mapsto z_i \; = \; \frac{x_i - \overline{x}}{\widetilde{s}_x} \; = \; \underbrace{\frac{1}{\widetilde{s}_x}}_{a} \cdot x_i - \underbrace{\frac{1}{\widetilde{s}_x}}_{b} \cdot \overline{x}$$

Es gilt: $\overline{z} = 0$ und $\widetilde{s}_z^2 = 1$

4.9 Variationskoeffizient

$$v = \frac{\widetilde{s}}{\overline{x}}$$

5 Schiefemaße

5.1 Lageregeln

- $x_{mod} < x_{med} < \overline{x} \; \rightarrow$ rechtsschief
- $x_{mod} = x_{med} = \overline{x} \; \rightarrow$ symmetrisch
- $\overline{x} < x_{med} < x_{mod} \; \rightarrow$ linksschief

5.2 Schiefekoeffizient nach Pearson (Momentenkoeffizient)

$$g_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^3}}$$

- $g_m > 0 \; \rightarrow$ rechtsschief
- $g_m = 0 \; \rightarrow$ symmetrisch
- $g_m < 0 \; \rightarrow$ linksschief

6 Konzentrationsmaße

6.1 Relative Konzentration

6.1.1 Gini-Koeffizient

Wertebereich: $0 \leq G \leq \frac{n-1}{n}$

Normierter Gini-Koeffizient

Wertebereich: $0 \leq G^* \leq 1$

$$G^* = \frac{n}{n-1}\, G$$

Relative Konzentration

	Urliste	unklassierte Häufigkeitsverteilung	klassierte Häufigkeitsverteilung (Merkmalssummen unbekannt)	klassierte Häufigkeitsverteilung (Merkmalssummen bekannt)
Gini-Koeffizient	$G = \sum_{i=1}^{n} u_i \cdot \widetilde{v}_i + \sum_{i=1}^{n} u_{i-1} \cdot \widetilde{v}_i - 1$	$G = \sum_{j=1}^{k} u_j \cdot \widetilde{v}_j + \sum_{j=1}^{k} u_{j-1} \cdot \widetilde{v}_j - 1$		
q	$\in \quad 1, 2, ..., n$	$\in \quad 1, 2, ..., k$		
relative Häufigkeit	$f_q = \frac{n_q}{n} = \frac{1}{n}$	$f(a_q) = \frac{h(a_q)}{n} = \frac{h(a_q)}{\sum_{j=1}^{k} h(a_j)}$	$f_q = \frac{n_q}{n} = \frac{n_q}{\sum_{j=1}^{k} n_j}$	$f_q = \frac{n_q}{n} = \frac{n_q}{\sum_{j=1}^{k} n_j}$
kumulierte rel. Häufigkeit	$u_q = \sum_{i=1}^{q} f_i = \frac{q}{n}$	$u_q = \sum_{j=1}^{q} f(a_j)$	$u_q = \sum_{j=1}^{q} f_j$	$u_q = \sum_{j=1}^{q} f_j$
relativer Merkmalsanteil	$\widetilde{v}_q = \frac{x_{(q)}}{\sum_{i=1}^{n} x_i}$	$\widetilde{v}_q = \frac{a_q \cdot h(a_q)}{\sum_{j=1}^{k} a_j \cdot h(a_j)} = \frac{a_q \cdot f(a_q)}{\sum_{j=1}^{k} a_j \cdot f(a_j)}$	$\widetilde{v}_q = \frac{m_q \cdot n_q}{\sum_{j=1}^{k} m_j \cdot n_j} = \frac{m_q \cdot f_q}{\sum_{j=1}^{k} m_j \cdot f_j}$	$\widetilde{v}_q = \frac{x_q}{\sum_{j=1}^{k} x_j}$
kumulierter rel. Merkmalsanteil	$v_q = \sum_{i=1}^{q} \widetilde{v}_i$	$v_q = \sum_{j=1}^{q} \widetilde{v}_j$	$v_q = \sum_{j=1}^{q} \widetilde{v}_j$	$v_q = \sum_{j=1}^{q} \widetilde{v}_j$

6.1.2 Lorenzkurve

Streckenzug durch
$(0,0) = (u_0, v_0), (u_1, v_1), ..., (u_n, v_n) = (1,1)$ (Urliste)
bzw.
$(0,0) = (u_0, v_0), (u_1, v_1), ..., (u_k, v_k) = (1,1)$ (unklassierte oder klassierte Häufigkeitsverteilung)

6.2 Absolute Konzentration

Index nach Hirschmann/Herfindahl. Beschreibt die absolute Konzentration.

Es muss gelten: $\sum_{i=1}^{n} x_i > 0$. Wertebereich: $\frac{1}{n} \leq H \leq 1$

$H = \sum\limits_{i=1}^{n} \widetilde{v}_i^2$ (Urliste)

$H = \frac{V^2+1}{n}$ mit $V = \frac{\widetilde{s}}{\bar{x}}$ (unklassierte Häufigkeitsverteilung)

$H = \frac{V^2+1}{n}$ mit $V = \frac{\widetilde{s}}{\bar{\bar{x}}}$ (klassierte Häufigkeitsverteilung)

7 Mehrdimensionale Merkmale

7.1 Kontingenztafeln

(k x m)-Kontingenztafel

a_i - Zeilen $i = 1, \ldots, k$
b_j - Spalten $j = 1, \ldots, m$

$$h_{ij} = h(a_i, b_j) \qquad \ldots \text{ absolute Häufigkeit der Kombination } (a_i, b_j)$$

$$f_{ij} = f(a_i, b_j) = \frac{h_{ij}}{n} \qquad \ldots \text{ relative Häufigkeit der Kombination } (a_i, b_j)$$

$$f_{i\bullet} = \sum_{j=1}^{m} f_{ij} = \frac{h_{i\bullet}}{n}, \; i = 1, \ldots, k \qquad \ldots \text{ relative Randhäufigkeiten von X}$$

$$f_{\bullet j} = \sum_{i=1}^{k} f_{ij} = \frac{h_{\bullet j}}{n}, \; j = 1, \ldots, m \qquad \ldots \text{ relative Randhäufigkeiten von Y}$$

7.2 Bedingte Verteilungen

7.2.1 Bedingte Verteilung von X

$$f_X(a_i|b_j) = \frac{f_{ij}}{f_{\bullet j}} = \frac{h_{ij}}{h_{\bullet j}}$$

$f_X(a_1|b_j), \ldots, f_X(a_k|b_j)$ heißt bedingte Verteilung von X geg. Y $= b_j$

Es gilt: $\sum_{i=1}^{k} f_X(a_i|b_j) = 1$ für jedes feste j, $j = 1, \ldots, m$

7.2.2 Bedingte Verteilung von Y

$$f_Y(b_j|a_i) = \frac{f_{ij}}{f_{i\bullet}} = \frac{h_{ij}}{h_{i\bullet}}$$

$f_Y(b_1|a_i), \ldots, f_Y(b_m|a_i)$ heißt bedingte Verteilung von Y geg. X $= a_i$

Es gilt: $\sum_{j=1}^{m} f_Y(b_j|a_i) = 1$ für jedes feste i, $i = 1, \ldots, k$

7.2.3 Rekonstruktion der gemeinsamen Häufigkeiten

$f_{ij} = f_Y(b_j|a_i) \cdot f_{i\bullet}$ bzw. $f_{ij} = f_X(a_i|b_j) \cdot f_{\bullet j}$

7.3 Zusammenhangsanalyse in Kontingenztafeln

7.3.1 Hypothetische absolute Häufigkeit (bei Unabhängigkeit der Merkmale)

$$e_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$$

7.3.2 Chi-Quadrat Koeffizent

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} \qquad , \chi^2 \in [0, \infty)$$

7.3.3 Kontingenzkoeffizent

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad , K \in \left[0, \sqrt{\tfrac{M-1}{M}}\right], \text{wobei } M = min\{k, m\}$$

7.3.4 Korrigierter Kontingenzkoeffizent

$$K^* = \frac{K}{\sqrt{\frac{M-1}{M}}} \qquad , K^* \in [0, 1]$$

$K^* \leq 0.2$	$\rightarrow$ kein wesentlicher Zusammenhang
$0.2 < K^* \leq 0.5$	$\rightarrow$ schwacher Zusammenhang
$0.5 < K^* < 0.8$	$\rightarrow$ deutlicher Zusammenhang
$0.8 \leq K^*$	$\rightarrow$ starker Zusammenhang

7.4 Zusammenhangsmaße bei metrischen Merkmalen

7.4.1 Korrelationskoeffizient nach Bravais-Pearson (linearer Zusammenhang)

Wertebereich: $-1 \leq r_{XY} \leq 1$

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\widetilde{s}_X \cdot \widetilde{s}_Y}$$

alternativ: $$r_{XY} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}}$$

Stärke des linearen Zusammenhangs: Betrachte $|r_{XY}|$, Einteilung wie in 7.3.4

7.4.2 Empirische Kovarianz von X und Y

Wertebereich: $-\infty \leq \widetilde{s}_{XY} \leq \infty$

$$\widetilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y}$$

7.4.3 Rangkorrelationskoeffizient nach Spearman (monotoner Zusammenhang)

Wertebereich: $-1 \leq r_{Sp} \leq 1$
Basiert auf den Rängen der beobachteten Werte.

1. Allgemein

$$r_{Sp} = \frac{\sum_{i=1}^{n}(rg(x_i) - \frac{n+1}{2}) \cdot (rg(y_i) - \frac{n+1}{2})}{\sqrt{\left(\sum_{i=1}^{n}(rg(x_i))^2 - \frac{n\cdot(n+1)^2}{4}\right) \cdot \left(\sum_{i=1}^{n}(rg(y_i))^2 - \frac{n\cdot(n+1)^2}{4}\right)}}$$
$$= \frac{\sum_{i=1}^{n} rg(x_i) \cdot rg(y_i) - \frac{n\cdot(n+1)^2}{4}}{\sqrt{\left(\sum_{i=1}^{n}(rg(x_i))^2 - \frac{n\cdot(n+1)^2}{4}\right) \cdot \left(\sum_{i=1}^{n}(rg(y_i))^2 - \frac{n\cdot(n+1)^2}{4}\right)}}$$

2. Ohne Bindungen

$$r_{Sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \text{ , wobei } d_i = rg(x_i) - rg(y_i)$$

Stärke des monotonen Zusammenhangs: Betrachte $|r_{Sp}|$, Einteilung wie in 7.3.4

8 Einfache lineare Regression

Sei Y eine interessierende Zielgröße mit den Beobachtungen y und x eine deterministische Einflussgröße. Modell:

$$y = a \cdot x + b + \varepsilon$$

8.1 Kleinste Quadrate Methode für die Regressionskoeffizienten

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2}$$

$$\widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x}$$

Die Werte $\widehat{y_i} = \widehat{a} \cdot x_i + \widehat{b}$ sind Vorhersagen oder Prognosen für die y_i.
Die Abweichungen $\widehat{\varepsilon_i} = y_i - \widehat{y_i}$ heißen **Residuen**.

8.2 Bestimmtheitsmaß

Güte der Anpassung der Daten an die berechnete Gerade.

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \qquad R^2 \in [0, 1]$$

Es gilt: $R^2 = r_{XY}^2$ (quadrierter Korrelationskoeffizient)

9 Analyse zeitlicher Verläufe

9.1 Komponentenmodelle für Zeitreihen

Trendkomponente (g) : langfristiges Verhalten

Saisonkomponente (s) : wiederkehrende zyklische Schwankungen

Irreguläre Komponente (ε) : Rest

1. Additives Modell:

$$y_t = g_t + s_t + \varepsilon_t\,, \qquad t = 1, \ldots, T$$

2. Multiplikatives Modell:

$$y_t = g_t \cdot s_t \cdot \varepsilon_t\,, \qquad t = 1, \ldots, T$$

   Rückführung auf Additives Modell mit $log(y_t) = log(g_t) + log(s_t) + log(\varepsilon_t)$ möglich.

9.2 Lineares Trendmodell

- Reines Trendmodell:

$$y_t = g_t + \varepsilon_t$$

- Trendmodell mit im zeitlichem Verlauf linearer Trendkomponente:

$$y_t = \alpha{\cdot}t{+}\beta{+}\varepsilon_t\,, \qquad t = 1, \ldots, T \quad \text{(Bestimmung mit KQ-Methode)}$$

9.3 Einfacher gleitender Durchschnitt der Ordnung p

Betrachtet wird eine Zeitreihe $y_1, \ldots, y_T$.
Ordnung p des gleitenden Durchschnitts gibt die Anzahl der in die Mittelwertberechnung eingehenden Zeitreihenwerte an. Trend g_t durch ein lokales arithmetisches Mittel der Zeitreihenwerte $y_{t-q}, \ldots, y_{t+q}$ approximieren:

- für ungerade Ordnung p: $q = \frac{p-1}{2}$

$$\widehat{g_t}^p = \frac{1}{2 \cdot q + 1} \sum_{j=-q}^{q} y_{t+j} = \frac{1}{p} \cdot (y_{t-q} + \ldots + y_t + \ldots + y_{t+q})$$

mit $t = q + 1, \ldots, T - q$

- für gerade Ordnungp: $q = \frac{p}{2}$

$$\widehat{g_t}^p = \frac{1}{p}(\frac{1}{2} \cdot y_{t-q} + \sum_{j=-q+1}^{q-1} y_{t+j} + \frac{1}{2} \cdot y_{t+q})$$

mit $t = q + 1, \ldots, T - q$

9.4 Indexzahlen

Bezeichnung:	Basiszeit 0	mit Preisen	$p_0(1), \ldots, p_0(n)$
		und Gütermengen	$q_0(1), \ldots, q_0(n)$
	Berichtszeit t	mit Preisen	$p_t(1), \ldots, p_t(n)$
		und Gütermengen	$q_t(1), \ldots, q_t(n)$

9.4.1 Umsatzindex

$$W_{0,t} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)} \cdot 100$$

9.4.2 Preisindex nach Laspeyres

$$P_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_0(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)} \cdot 100$$

9.4.3 Preisindex nach Paasche

$$P_{0,t}^{P} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_t(i)} \cdot 100$$

9.4.4 Mengenindex nach Laspeyres

$$Q_{0,t}^{L} = \frac{\sum_{i=1}^{n} q_t(i) \cdot p_0(i)}{\sum_{i=1}^{n} q_0(i) \cdot p_0(i)} \cdot 100$$

9.4.5 Mengenindex nach Paasche

$$Q_{0,t}^{P} = \frac{\sum_{i=1}^{n} q_t(i) \cdot p_t(i)}{\sum_{i=1}^{n} q_0(i) \cdot p_t(i)} \cdot 100$$

9.4.6 Index von March

$$I_M = \frac{\sum_{i=1}^{n} \frac{p_t(i)}{p_o(i)} \cdot q_t(i)}{\sum_{i=1}^{n} q_t(i)}$$

10 Wahrscheinlichkeitsrechnung

10.1 Mengenoperationen

Seien A und B Teilmengen einer Menge Ω

- **Schnittmenge:** $A \cap B$
- **Vereinigungsmenge:** $A \cup B$
- **Differenzmenge:** $A \backslash B$
- **Komplementärmenge** oder **Komplement:** A^C
- **Anzahl der Elemente von A:** $\mid A \mid$

10.2 Wahrscheinlichkeiten

10.2.1 Laplace-Wahrscheinlichkeiten

$P(A)$ ... Wahrscheinlichkeit des Ereignisses A
Gilt für $\Omega = \{\omega_1, \ldots, \omega_n\}$, dass $P(\{\omega_i\}) = \frac{1}{n}$, i = 1,...,n
dann gilt für A $\subseteq \Omega$, zusammengesetzt aus m Elementarereignissen:

$$P(A) = \frac{m}{n} \quad = \quad \frac{\text{Anzahl der Elementarereignisse in A}}{\text{Gesamtzahl der Elementarereignisse}}$$

10.2.2 Rechenregeln für Wahrscheinlichkeiten

Für eine Wahrscheinlichkeitsabbildung P und Ereignisse $A, B, A_1, \ldots, A_k$ sowie eine Grundmenge Ω von Ergebnissen gilt:

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- Falls $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A^C) = 1 - P(A)$
- Sind $A_1, \ldots, A_k$ paarweise disjunkt, dann gilt:

  $P(A_1 \cup \ldots \cup A_k) = P(A_1) + \ldots + P(A_k)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Ist Ω endlich mit Elementarereignissen $\{\omega_1\}, \ldots, \{\omega_n\}$, dann ist P(A) = $\sum_{w \in A} P(\{\omega\})$

10.2.3 Bedingte Wahrscheinlichkeit von A gegeben B

Seien $A, B \subset \Omega$ und $P(B) > 0$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B) \cdot P(B) \quad \textit{(Produktsatz)}$$

10.2.4 Satz von der totalen Wahrscheinlichkeit

Sei $B_1, \ldots, B_k$ eine disjunkte Zerlegung von Ω.
Dann gilt für $A \subset \Omega$:

$$P(A) = \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i)$$

10.2.5 Satz von Bayes

Sei $B_1, \ldots, B_k$ eine disjunkte Zerlegung von Ω, wobei $P(B_i) > 0$ und $P(A|B_i) > 0$ für mindestens ein i.
Dann gilt:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{k} P(A|B_j) \cdot P(B_j)} = \frac{P(A|B_i) \cdot P(B_i)}{P(A)}, \quad i = 1, \ldots, k$$

10.2.6 Unabhängigkeit von zwei Ereignissen

Seien $A, B \subset \Omega$ zwei Ereignisse.
A und B heißen (stochastisch) unabhängig, wenn gilt:

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

Alternativ: $P(A|B) = P(A)$ mit $P(B) > 0$ oder $P(B|A) = P(B)$ mit $P(A) > 0$
Falls P(B) = 0, so nennt man A und B stets unabhängig.

11 Zufallsstichproben

11.1 Allgemeines

Umfang Grundgesamtheit $\ldots N$
Umfang Stichprobe $\ldots n$
Einfache Zufallsstichprobe

Jede mögliche Stichprobe vom Umfang n aus der Grundgesamtheit hat die selbe Wahrscheinlichkeit realisiert zu werden.

11.2 Anzahl möglicher Stichproben

	ohne Zurücklegen	mit Zurücklegen
mit Beachtung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
ohne Beachtung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

12 Eindimensionale Zufallsvariablen

12.1 Dichte

1. **Diskrete Dichte** (f(x) Wahrscheinlichkeitsfunktion!)

   $$f(x_i) = P(X = x_i)$$

   Es gilt: $\forall i : 0 \leq f(x_i) \leq 1$ und $\sum_{i=1}^{\infty} f(x_i) = 1.$

2. **Stetige Dichte** (f(x) Dichtefunktion!)

   $f(x) = F'(x)$, falls die Ableitung existiert

   Es gilt: $\forall x : f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(t)dt = 1.$ $(f(x) \geq 1$ ist möglich!)

12.2 Verteilungsfunktion

$$F(x) = P(X \leq x)$$

1. **Diskreter Wertebereich**

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

2. **Stetiger Wertebereich**

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

12.3 Rechnen mit Verteilungsfunktion und Dichte

1. **Diskrete Zufallsvariable** X
   - $P(a < X \leq b) = \sum_{x_i : a < x_i \leq b} P(X = x_i)$
   - Alternativ mit Hilfe der Verteilungsfunktion:
     $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$
   - $P(a \leq X \leq b) = F(b) - P(X < a)$
   - $P(a \leq X < b) = P(X < b) - P(X < a)$
   - $P(a < X < b) = P(X < b) - P(X \leq a) = P(X < b) - F(a)$
   - $P(X > a) = 1 - F(a)$
2. **Stetige Zufallsvariable** X
   - $P(a < X \leq b) = \int_{a}^{b} f(t)\,dt$
   - Alternativ mit Hilfe der Verteilungsfunktion:
     $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$
   - $P(X = x) = 0$ für jedes x, d.h. Wahrscheinlichkeit einen bestimmten Wert anzunehmen ist gleich Null.
   - $P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b)$
   - $P(X > a) = P(X \geq a) = 1 - F(a)$

12.4 Modus

Modus der Verteilung von X ist derjenige x-Wert x_{mod}, für den die Dichte $f(x)$ von X maximal wird.
Gibt es keinen eindeutigen x-Wert der dies erfüllt, so ist der Modus nicht definiert.

12.5 Erwartungswert

12.5.1 Definition

Betrachtet wird eine Zufallsvariable X mit Dichtefunktion $f(x)$.

1. Ist X **diskrete Zufallsvariable**:

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot f(x_i) = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \ldots$$

2. Ist X **stetige Zufallsvariable**:

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

12.5.2 Transformationen

1. **Lineare Transformation**

$$Y = a \cdot X + b \quad \rightarrow \quad E(Y) = E(a \cdot X + b) = a \cdot E(X) + b$$

2. **Transformation mit beliebiger Funktion**

$$Y = g(X)$$

- X ist **diskrete** Zufallsvariable

  $E(Y) = E(g(X)) = \sum_{i=1}^{\infty} g(x_i) \cdot f(x_i)$

- X ist **stetige** Zufallsvariable

  $E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$

12.6 Varianz und Standardabweichung

Sei X eine Zufallsvariable mit Dichtefunktion f und Erwartungswert $E(X)$:

- **Varianz**
  1. Ist X **diskret**:

$$Var(X) = E((X - E(X))^2) = \sum_{i=1}^{\infty} (x_i - E(X))^2 \cdot f(x_i)$$

$$= \sum_{i=1}^{\infty} x_i^2 \cdot f(x_i) - (E(X))^2 \qquad \text{(Verschiebungssatz)}$$

  2. Ist X **stetig**:

$$Var(X) = E((X - E(X))^2) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - (E(X))^2 \qquad \text{(Verschiebungssatz)}$$

- **Standardabweichung** $\quad \sigma_X = \sqrt{Var(X)}$
- **Verschiebungssatz allgemein** $\quad Var(X) = E(X^2) - (E(X))^2$
- **Lineare Transformation**
  - $Y = a \cdot X + b$
  - $Var(Y) = Var(a \cdot X + b) = a^2 \cdot Var(X)$
  - $\sigma_Y = |a| \cdot \sigma_X$

12.7 Quantile

1. Ist X **diskrete Zufallsvariable**:

   p-Quantil x_p ist die Zahl, für die

$$P(X < x_p) \leq p \quad \text{und} \quad P(X > x_p) \leq 1 - p$$

2. Ist X **stetige Zufallsvariable**:

   x_p ist die Zahl, für die

$$F(x_p) = p$$

Falls x_p nicht eindeutig bestimmbar, wähle jeweils die kleinste Zahl, die dies erfüllt.

13 Mehrdimensionale Zufallsvariablen

13.1 Gemeinsame Dichte und Randdichte

	diskrete Zufallsvariable	stetige Zufallsvariable
Gemeinsame Dichte	$f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$	$f_{X,Y}(x, y)$
Randdichten	$f_X(x_i) = P(X = x_i)$	$f_X(x)$
	$f_Y(y_j) = P(Y = y_j)$	$f_Y(y)$
	$f_X(x_i) = \sum_{j=1}^{\infty} f_{X,Y}(x_i, y_j)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$
	für f_Y analog	

13.2 Bedingte Dichte

- bedingte Dichte von X gegeben Y:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- bedingte Dichte von Y gegeben X:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

13.3 Unabhängigkeit von Zufallsvariablen

X und Y sind stochastisch unabhängig, wenn gilt:

$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$, für alle $x \in X(\Omega)$ und $y \in Y(\Omega)$

13.4 Kovarianz

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

13.4.1 Diskrete Zufallsvariablen

$$Cov(X, Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i - E(X)) \cdot (y_j - E(Y)) \cdot f_{X,Y}(x_i, y_j)$$

Zur vereinfachten Berechnung:

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

mit $$E(X \cdot Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i \cdot y_j \cdot f_{X,Y}(x_i, y_j)$$

13.4.2 Stetige Zufallsvariablen

$$Cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X)) \cdot (y - E(Y)) \cdot f_{X,Y}(x, y)\, dx\, dy$$

Zur vereinfachten Berechnung:

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

mit $$E(X \cdot Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y)\, dx\, dy$$

13.5 Rechenregeln Erwartungswert, Varianz, Kovarianz

- $E(X+Y) = E(X) + E(Y)$
- $E(X-Y) = E(X) - E(Y)$
- $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$
- $Var(X+Y) = Var(X) + Var(Y) + 2 \cdot Cov(X,Y)$
- $Var(X-Y) = Var(X) + Var(Y) - 2 \cdot Cov(X,Y)$
- $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$, falls $X_1, ..., X_n$ unabhängig
- $Cov(aX+b, cY+d) = a \cdot c \cdot Cov(X,Y)$

13.6 Korrelationskoeffizient

Wertebereich: $-1 \leq \rho(x,y) \leq 1$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

14 Diskrete Verteilungen

14.1 Bernoulli-Verteilung

Dichtefunktion:	$f(x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$	für $x_i = 0, 1$
Schreibweise:	$X \sim Bin(1,p)$	
Erwartungswert:	$E(X) = p$	
Varianz:	$Var(X) = p \cdot (1-p)$	

14.2 Binomialverteilung

Dichtefunktion:	$f(x_i) = \binom{n}{x_i} \cdot p^{x_i} \cdot (1-p)^{n-x_i}$	für $x_i = 0, \ldots, n$
Schreibweise:	$X \sim Bin(n,p)$	
Erwartungswert:	$E(X) = n \cdot p$	
Varianz:	$Var(X) = n \cdot p \cdot (1-p)$	

Eigenschaften

- Beschreibt Situation des Ziehens mit Zurücklegen.
- Die Bernoulli-Verteilung ist ein Spezialfall der Binomialverteilung mit n = 1.
- Sind $X_1, \ldots, X_n$ stochastisch unabhängig mit $X \sim Bin(1,p)$, $i = 1, \ldots, n$, dann ist $X = \sum_{i=1}^{n} X_i \sim Bin(n,p)$.
- Symmetrie: Sei $X \sim Bin(n,p)$ und $Y = n - X$, dann gilt: $Y \sim Bin(n, 1-p)$.

14.3 Die hypergeometrische Verteilung

Dichtefunktion: $f(x_i) = \frac{\binom{M}{x_i} \cdot \binom{N-M}{n-x_i}}{\binom{N}{n}}$ für $x_i = 0, \ldots, n$

Schreibweise: $X \sim Hyp(n, M, N)$

Erwartungswert: $E(X) = n \cdot \frac{M}{N}$

Varianz: $Var(X) = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}$

Beschreibt Situation des Ziehens ohne Zurücklegen.

14.4 Die Poisson-Verteilung

Dichtefunktion: $f(x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$ für $x_i = 0, 1, 2, \ldots$

Schreibweise: $X \sim Poi(\lambda)$

EW und Varianz: $E(X) = Var(X) = \lambda$

15 Stetige Verteilungen

15.1 Die stetige Gleichverteilung (Rechteckverteilung) auf $[a, b]$

Dichtefunktion: $f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$

Schreibweise: $X \sim G[a, b]$

Erwartungswert: $E(X) = \frac{a+b}{2}$

Varianz: $Var(X) = \frac{(b-a)^2}{12}$

15.2 Die Normalverteilung

Dichtefunktion: $f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \; exp \left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$

Schreibweise: $X \sim N(\mu, \sigma^2)$

Erwartungswert: $E(X) = \mu$

Varianz: $Var(X) = \sigma^2$

15.2.1 Eigenschaften

- Standardnormalverteilung:
  - spezielle Normalverteilung $N(0, 1)$ mit Parametern $\mu = 0$ und $\sigma^2 = 1$
  - Verteilungsfunktion: Φ
  - Speziell für die Verteilungsfunktion der Standardnormalverteilung gilt:

    $\Phi(-z) = P(Z \leq -z) = P(Z \geq z) = 1 - \Phi(z)$
- für p-Quantil z_p gilt: $z_{1-p} = -z_p$

- Standardisierung einer $N(\mu, \sigma^2)$-verteilten Zufallsvariable X, so dass Transformation Z $N(0,1)$-verteilt ist:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1), \text{ d.h. } \quad P(Z \le z) = \Phi(z).$$

- $X \sim N(\mu, \sigma^2)$, $Y = aX + b \Rightarrow Y \sim N(a\mu + b, a^2 \cdot \sigma^2)$
- $X_1, \ldots, X_n$ stochastisch unabhängig, $X_i \sim N(\mu, \sigma^2) \Rightarrow \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

15.2.2 Bestimmung von Wahrscheinlichkeiten $P(a \le X \le b)$

- Für eine $N(0,1)$-verteilte Zufallsvariable Z ist

$$P(Z \le z) = \Phi(z) \quad \text{und}$$

$$P(a \le Z \le b) = \Phi(b) - \Phi(a).$$

- Für eine $N(\mu, \sigma^2)$-verteilte Zufallsvariable X ist

$$P(X \le x) = P\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad \text{und}$$

$$P(a \le X \le b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

15.2.3 Bestimmung von Quantilen

- p-Quantil z_p der $N(0,1)$-Verteilung: z_p aus Tabelle
- p-Quantil x_p der $N(\mu, \sigma^2)$-Verteilung: $x_p = \sigma \cdot z_p + \mu$, $\quad z_p$ aus Tabelle

15.3 t-Verteilung mit n Freiheitsgraden (Student t-Verteilung)

Schreibweise: $\quad X \sim t_n$

- symmetrisch um 0
- für das p-Quantil gilt: $t_{n;p} = -t_{n;1-p}$
- $X \sim t_n$ und $n \ge 2 \Rightarrow E(X) = 0$
- $X \sim t_n$ und $n \ge 3 \Rightarrow Var(X) = \frac{n}{n-2}$
- Für $n \to \infty$ gilt $t_n \to N(0,1)$ $\quad$ (ca. ab $n \ge 30$)
- $X_1, \ldots, X_n$ unabhängig und identisch $N(\mu, \sigma^2)$-verteilt $\Rightarrow \sqrt{n} \cdot \frac{\overline{X} - \mu}{S} \sim t_{n-1}$

16 Schätzer

16.1 Schätzer für Erwartungswert und Varianz

$X_1, \ldots, X_n$ Zufallsvariablen mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

- Schätzer für μ:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{mit} \qquad E(\overline{X}) = \mu$$

zusätzlich ist $Var(\overline{X}) = \frac{\sigma^2}{n}$, falls die X_i unabhängig

- Schätzer für σ^2, falls die X_i unabhängig mit identischer Verteilung:

$$\widetilde{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad \text{mit} \qquad E(\widetilde{S}^2) = \frac{n-1}{n}\, \sigma^2$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad \text{mit} \qquad E(S^2) = \sigma^2$$

Hinweis: Verschiebungssatz siehe 4.4 und 4.6

16.2 Konfidenzintervalle für μ im Normalverteilungsmodell

Betrachte eine Zufallsvariable X mit $X \sim N(\mu, \sigma^2)$; seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X.
Gegeben sei weiter eine Irrtumswahrscheinlichkeit $\alpha, 0 < \alpha < 1$.

- Falls σ^2 bekannt, so ist

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

ein $(1-\alpha)$-Konfidenzintervall für μ.
Dabei bezeichnet $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der $N(0,1)$.

- Falls σ^2 unbekannt ist, ist

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

ein $(1-\alpha)$-Konfidenzintervall für μ.

Dabei ist $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$, und $t_{n-1;1-\alpha/2}$ bezeichnet das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

16.3 Approximative Konfidenzintervalle für μ

Betrachte eine Zufallsvariable X mit $E(X) = \mu$, $Var(X) = \sigma^2$;
seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X, sei $n \geq 30$.

- Falls σ^2 bekannt, so ist

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

ein approximatives $(1-\alpha)$-Konfidenzintervall für μ.

- Falls σ^2 unbekannt, so ist

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

ein approximatives $(1-\alpha)$-Konfidenzintervall für μ.
Dabei bezeichnet $t_{n-1;1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

17 Statistische Hypothesentests

17.1 Gauß-Test

Seien $X_1, \ldots, X_n$ unabhängige und identisch normalverteilte Zufallsvariablen,
$X_i \sim N(\mu, \sigma^2)$, und sei σ^2 **bekannt**.

Testproblem H_0 vs. H_1	Entscheidung Lehne H_0 ab, falls
$\mu = \mu_0$ vs. $\mu \neq \mu_0$	$\left\lvert \sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma} \right\rvert > z_{1-\alpha/2}$
$\mu \geq \mu_0$ vs. $\mu < \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma} < -z_{1-\alpha}$
$\mu \leq \mu_0$ vs. $\mu > \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma} > z_{1-\alpha}$

Dabei bezeichnet z_α das α-Quantil der Standardnormalverteilung.

17.2 t-Test

Seien $X_1, \ldots, X_n$ unabhängige und identisch normalverteilte Zufallsvariablen, $X_i \sim N(\mu, \sigma^2)$, und sei σ^2 **unbekannt**.

Testproblem H_0 vs. H_1	Entscheidung Lehne H_0 ab, falls
$\mu = \mu_0$ vs. $\mu \neq \mu_0$	$\left\lvert \sqrt{n}\,\frac{\overline{X}-\mu_0}{S} \right\rvert > t_{n-1;1-\alpha/2}$
$\mu \geq \mu_0$ vs. $\mu < \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S} < -t_{n-1;1-\alpha}$
$\mu \leq \mu_0$ vs. $\mu > \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S} > t_{n-1;1-\alpha}$

Dabei bezeichnet $t_{n-1,\alpha}$ das α-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

17.3 Approximativer Gauß-Test

Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen, die aber **nicht notwendig normalverteilt** sind, mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$.
Sei σ^2 **unbekannt** und $n \geq 30$.

Testproblem H_0 vs. H_1	Entscheidung Lehne H_0 ab, falls
$\mu = \mu_0$ vs. $\mu \neq \mu_0$	$\left\lvert \sqrt{n}\,\frac{\overline{X}-\mu_0}{S} \right\rvert > z_{1-\alpha/2}$
$\mu \geq \mu_0$ vs. $\mu < \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S} < -z_{1-\alpha}$
$\mu \leq \mu_0$ vs. $\mu > \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S} > z_{1-\alpha}$

17.4 Test auf einen Anteil

Ein Anteil p der Grundgesamtheit besitze eine interessierende Eigenschaft.
Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i = 1$, falls das i-te Element die Eigenschaft besitzt, $X_i = 0$ sonst.

Testproblem H_0 vs. H_1	Entscheidung Lehne H_0 ab, falls
$p = p_0$ vs. $p \neq p_0$	$\left\lvert \sqrt{n}\,\frac{\overline{X}-p_0}{\sqrt{p_0 \cdot (1-p_0)}} \right\rvert > z_{1-\alpha/2}$
$p \geq p_0$ vs. $p < p_0$	$\sqrt{n}\,\frac{\overline{X}-p_0}{\sqrt{p_0 \cdot (1-p_0)}} < -z_{1-\alpha}$
$p \leq p_0$ vs. $p > p_0$	$\sqrt{n}\,\frac{\overline{X}-p_0}{\sqrt{p_0 \cdot (1-p_0)}} > z_{1-\alpha}$

17.5 χ^2 Unabhängigkeitstest

Betrachtet werden zwei Zufallsvariablen X, Y. Die Beobachtungspaare (x_i, y_i) seien in einer $(k \times m)$-Kontingenztafel zusammengefasst.

- Gemeinsame absolute Häufigkeiten in der Tafel: h_{ij}
- Randhäufigkeiten: $h_{i.}$ bzw. $h_{.j}$
- Unter Unabhängigkeit von X und Y erwartete Häufigkeiten:

$$e_{ij} = \frac{h_{i.} \cdot h_{.j}}{n} \ , \ \ i = 1, \ldots, k \ , \ \ j = 1, \ldots, m$$

Testproblem: $\mathbf{H_0}$: X, Y **unabhängig** **vs.** $\mathbf{H_1}$: X, Y **abhängig**

Entscheidungsregel: H_0 wird zum Niveau α verworfen, falls

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} > \chi^2_{(k-1)\cdot(m-1);1-\alpha}$$

Dabei bezeichnet $\chi^2_{q;\alpha}$ das α-Quantil der χ^2-Verteilung mit q Freiheitsgraden.