1 Allgemeine Symbole

Folgende Symbole werden in allen Abschnitten verwendet:

n Anzahl der Beobachtungen

K Anzahl der Gruppen oder Klassen eines Merkmals, $K \leq n$

x_i Merkmalsausprägung des Merkmals beim i-ten Element (i-ter Beobachtung), $i = 1, 2, \ldots, n$

h_j Absolute Häufigkeit der j-ten Gruppe eines Merkmals, $j = 1, 2, \ldots, K$, bzw. absolute Häufigkeit der j-ten Merkmalsausprägung einer Häufigkeitsverteilung

h_{ij} Absolute Häufigkeit der Kombination i, j von zwei Merkmalen mit $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, m$

f_j Relative Häufigkeit der j-ten Gruppe eines Merkmals, $j = 1, 2, \ldots, K$,

$$f_j = \frac{h_j}{n},$$

bzw. relative Häufigkeit der j-ten Merkmalsausprägung einer Häufigkeitsverteilung

$F(x)$ Empirische Verteilungsfunktion

$$F(x) = \frac{\text{Anzahl Werte } \leq x}{n}$$

x_j^o Obere Gruppengrenze der j-ten Gruppe, $j = 1, 2, \ldots, K$

x_j^u Untere Gruppengrenze der j-ten Gruppe, $j = 1, 2, \ldots, K$

x_j' Gruppenmittelwert oder Gruppenmitte der j-ten Gruppe eines Merkmals, $j = 1, 2, \ldots, K$, bzw. j-te Merkmalsausprägung einer Häufigkeitsverteilung

$x_{(j)}$ Merkmalsausprägung der j-ten Beobachtung in der aufsteigend sortierten Datenreihe (Rangliste), wobei $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$. j ist der Rang der Beobachtung

2 Lagemaße

Folgende Lagemaße gelten für nicht gruppierte Daten

- Modus

  $$x_{\text{mod}} = \text{die Merkmalsausprägung, die am häufigsten auftritt}$$

- Median

  $$x_{(0,5)} = \begin{cases} x_{(\frac{n+1}{2})} & : \text{ n ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & : \text{ n gerade} \end{cases}$$

- P-Quantil, wobei $0 < p < 1$

  $$x_{(QP)} = \begin{cases} x_{(k)} & : \text{ mit } k \text{ als der nächsten ganzen Zahl nach } n \cdot p \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & : \text{ falls } k = n \cdot p \text{ eine ganze Zahl ist} \end{cases}$$

- Arithmetischer Mittelwert

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

- Geometrischer Mittelwert

  $$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

Die Lagemaße gelten analog auch für gruppierte Daten und Häufigkeitsverteilungen, wobei dort die jeweiligen Häufigkeiten h_j bzw. f_j berücksichtigt werden

- Gruppenmitte (sofern nicht bekannt oder angegeben)

  $$x'_j = \frac{x_j^o + x_j^u}{2}$$

- Modus

  $$x_{\text{mod}} = \text{die Gruppe die am häufigsten autftritt}$$

- Median
  Zuerst muss die Gruppe gefunden werden, in der der Median liegt

  $$\text{Median in Gruppe i: } \sum_{j=1}^{i-1} f_j < 0,5 \text{ und } \sum_{j=1}^{i} f_j \geq 0,5$$

  Anschließend wird die Position des Medians innerhalb der Gruppe i gesucht

  $$x_{(0,5)} = x_i^u + \frac{0,5 - \sum_{j=1}^{i-1} f_j}{f_i}(x_i^o - x_i^u)$$

- Arithmetischer Mittelwert

  $$\bar{x} = \frac{1}{n}\sum_{j=1}^{K} x'_j h_j = \sum_{j=1}^{K} x'_j f_j$$

3 Streuungsmaße

Die folgenden Streuungsmaße gelten für nicht gruppierte Daten

- Spannweite

$$r = \max x_i - \min x_i$$

- Varianz

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

- Standardabweichung

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

- Quartilsabstand

$$IQR = Q3 - Q1 = x_{(0,75)} - x_{(0,25)}$$

Die Streuungsmaße gelten analog auch für gruppierte Daten bzw. Häufigkeitsverteilungen

- Spannweite

$$r = \max x_j^o - \min x_j^u$$

- Varianz

$$s^2 = \frac{1}{n}\sum_{j=1}^{K}(x_j' - \bar{x})^2 h_j = \sum_{j=1}^{K}(x_j' - \bar{x})^2 f_j$$

- Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\sum_{j=1}^{K}(x_j' - \bar{x})^2 f_j}$$

- Quartilsabstand

$$IQR = Q3 - Q1 = x_{(0,75)} - x_{(0,25)}$$

4 Weitere statistische Kennwerte

Einordnen der Variation

- Variationskoeffizient

$$v = \frac{s}{\bar{x}}$$

4.1 Konzentrationsmessung

- Gini Koeffizient
  - Einzeldaten

$$G = \frac{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2\bar{x}}$$

  - Gruppierte Daten

$$G = 1 - \sum_{i=1}^{m} f_i(Q_i + Q_{i-1})$$

    mit $Q_i = \sum_{j=1}^{i} q_j$ und g_j als den j-te kumulierte Anteil am Gesamtmerkmalsbetrag

4.2 Kennzahlen der Analyse zweier nominaler Merkmale

- Randhäufigkeit:
  Summe der i-ten Zeile: $h_{i\cdot} = \sum_{j=1}^{m} h_{ij}$
  Summe der j-ten Spalte: $h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$
- Erwartete Häufigkeit der Kombination i, j bei Unabhängigkeit

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

- Pearsonsche Chi-Quadrat

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

- Kontingenzkoeffizient

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

4.3 Kennzahlen der Analyse zweier metrischer Merkmale

- Kovarianz

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

- Korrelationskoeffizient nach Bravais-Pearson

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

4.4 Kennzahlen der Analyse zweier ordinaler Merkmale

Anstelle der einzelnen Beobachtungen (x_i, y_i) werden die jeweiligen Ränge (R_{xi}, R_{yi}) betrachtet

- Rangkorrelationskoeffizient nach Spearman

$$r_{sp} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_{xi} - \bar{R}_x)(R_{yi} - \bar{R}_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{xi} - \bar{R}_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{yi} - \bar{R}_y)^2}}$$

- Vereinfachte Form des Rangkorrelationskoeffizient nach Spearman, wenn alle x_i bzw. y_i verschieden sind

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n}(R_{xi} - R_{yi})^2}{n(n^2-1)}$$

5 Regression

Anpassen einer Regressionsgerade $y = a + bx + e$

- Steigung (Kleinste Quadrate)

$$\hat{b} = \frac{s_{xy}}{s_x^2}$$

- Achsenabschnitt (Kleinste Quadrate)

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

- Prognose für (neuen) Wert x_0

$$\hat{y_0} = \hat{a} + \hat{b}x_0$$

- Residuum (Prognosefehler)

$$\hat{e}_i = y_i - \hat{y}_i$$

- Bestimmtheitsmaß

$$R^2 = 1 - \frac{s_{\hat{e}}^2}{s_y^2} = r_{xy}^2$$

6 Wahrscheinlichkeitsrechnung und schließende Statistik

6.1 Elementare Formeln der Wahrscheinlichkeitsrechnung

- Wahrscheinlichkeit

$$P(A) = \frac{\text{Anzahl der Versuche, in denen A eingetreten ist}}{\text{Anzahl aller Versuche}}$$

- Für das sichere Ereignis gilt $P = 1$
- Für das unmögliche Ereignis gilt $P = 0$
- Es gilt immer $0 \leq P \leq 1$
- Für zwei sich ausschließende Ereignisse A, B gilt $P(A \text{ oder } B) = P(A) + P(B)$
- Für zwei beliebige Ereignisse A, B gilt $P(A \text{ oder } B) = P(A)+P(B)-P(A \text{ und } B)$
- Für das gegenteilige Ereignis gilt $P(\text{Gegenteil von } A) = 1 - P(A)$
- Für die bedingte Wahrscheinlichkeit $P(A|B)$, d.h. für die Wahrscheinlichkeit für A wenn B schon eingetreten ist gilt $P(A|B) = \frac{P(A \text{ und } B)}{P(B)}$. Bei Unabhängigkeit von A, B gilt $P(A \text{ und } B) = P(A)P(B)$
- Satz von der totalen Wahrscheinlichkeit:

$$P(A) = P(A|B) \cdot P(B) + P(A|\text{Gegenteil von } B) \cdot P(\text{Gegenteil von } B)$$

- Satz von Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\text{Gegenteil von } A) \cdot P(\text{Gegenteil von } A)}$$

6.2 Assoziationsanalyse

- $\text{Support}(A) = P(A)$
- $\text{Support}(A \to B) = P(A \text{ und } B)$
- $\text{Konfidenz}(A \to B) = P(B|A)$
- $\text{Lift}(A \to B) = \frac{\text{Konfidenz}_{(A \to B)}}{\text{Support}_{(B)}}$

6.3 Rechnen mit Zufallsvariablen

- Wahrscheinlichkeitsrechnung und Verteilungsfunktion
  - $P(X \leq a) = F(a)$
  - $P(X > a) = 1 - F(a)$
  - $P(a < X \leq b) = F(b) - F(a)$

6.4 Spezielle Verteilungen: Normalverteilung

- Dichtefunktion Normalverteilung

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

- Erwartungswert Normalverteilung

$$E(X) = \mu$$

- Varianz Normalverteilung

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

- Verteilungsfunktion der Standardnormalverteilung mit $E(Z) = \mu = 0$ und $Var(Z) = \sigma^2 = 1$

$$\Phi(z) = F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)du$$

- Für die Standardnormalverteilung gilt

$$\Phi(-z) = 1 - \Phi(z)$$

- Überführung einer beliebigen normalverteilten Zufallsvariable X mit Parametern μ und σ in eine standardnormalverteilte Zufallsvariable Z

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

- Werttransformation Normalverteilungen
  - $F(a) = \Phi(\frac{a-\mu}{\sigma})$
  - Für das p-Quantil x_p mit $P(X \leq x_p) = p$ gilt $x_p = \mu + \sigma z_p$ mit $\Phi(z_p) = p$

6.5 Schätzen

- Punktschätzung
  - Schätzfunktion Erwartungswert μ (Stichprobenmittelwert)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (= \hat{\mu})$$

  - Schätzfunktion Varianz σ^2 (Stichprobenvarianz)

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

- Bereichsschätzung
  - Zweiseitiges $1-\alpha$ Konfidenzintervall für den Erwartungswert μ einer Normalverteilung

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}\right) = 1-\alpha$$

6.5.1 Testen

- Testen des Mittelwertes (Normalverteilung bzw. Approximation)[1]

Nullhypothese	$H_0 : \mu = \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu \leq \mu_0$
Alternativhypothese	$H_A : \mu \neq \mu_0$	$H_A : \mu < \mu_0$	$H_A : \mu > \mu_0$
Teststatistik	$T = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}}$		
Kritischer Wert Niveau α	$z_{1-\frac{\alpha}{2}}$	$-z_{1-\alpha}$	$z_{1-\alpha}$
H_0 ablehnen falls	$\lvert T\rvert > z_{1-\frac{\alpha}{2}}$	$T < -z_{1-\alpha}$	$T > z_{1-\alpha}$

Tabelle 1: Test auf Mittelwert bei $n \geq 30$ bzw. Normalverteilung

[1] Darstellung in Anlehnung an Oestrich, Romberg: Keine Panik vor Statistik!

7 Tabelle Standardnormalverteilung

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0,00	0,50	0,90	0,8159	1,80	0,9641	2,70	0,9965
0,05	0,5199	0,95	0,8289	1,85	0,9678	2,75	0,9970
0,10	0,5398	1,00	0,8413	1,90	0,9713	2,80	0,9974
0,15	0,5596	1,05	0,8531	1,95	0,9744	2,85	0,9978
0,20	0,5793	1,10	0,8643	2,00	0,9772	2,90	0,9981
0,25	0,5987	1,15	0,8749	2,05	0,9798	2,95	0,9984
0,30	0,6179	1,20	0,8849	2,10	0,9821	3,00	0,9987
0,35	0,6368	1,25	0,8944	2,15	0,9842	3,05	0,9989
0,40	0,6554	1,30	0,9032	2,20	0,9861	3,10	0,9990
0,45	0,6736	1,35	0,9115	2,25	0,9878	3,15	0,9992
0,50	0,6915	1,40	0,9192	2,30	0,9893	3,20	0,9993
0,55	0,7088	1,45	0,9265	2,35	0,9906	3,25	0,9994
0,60	0,7257	1,50	0,9332	2,40	0,9918	3,30	0,9995
0,65	0,7422	1,55	0,9394	2,45	0,9929	3,35	0,9996
0,70	0,7580	1,60	0,9452	2,50	0,9938	3,40	0,9997
0,75	0,7734	1,65	0,9505	2,55	0,9946	3,45	0,9997
0,80	0,7881	1,70	0,9554	2,60	0,9953	3,50	0,9998
0,85	0,8023	1,75	0,9599	2,65	0,9960	3,55	0,9998

Tabelle 2: Verteilungsfunktion Standardnormalverteilung

p	0,850	0,900	0,950	0,975	0,990	0,995	0,999
z_p	1,0364	1,2816	1,6449	1,9600	2,3263	2,5758	3,0902

Tabelle 3: Ausgewählte Quantile der Standardnormalverteilung