Formelsammlung Methodenlehre

Download der aktuellen Version unter https://johannes-titz.github.io/formelsammlung/main.pdf

Diese Formelsammlung wurde für Studierende der Psychologie entwickelt, dürfte aber auch für Studierende anderer Sozialwissenschaften nützlich sein. Sie basiert auf einer älteren Sammlung der Professur für Forschungsmethoden an der TU Chemnitz, die weder Autoren noch Lizenzangaben enthielt. Für eine sinnvolle Nutzung und Weiterentwicklung ist eine Lizenz erforderlich; deshalb wird das vorliegende Dokument unter der CC BY-SA 4.0 veröffentlicht. Die Formelsammlung orientiert sich am Lehrbuch von **sedlmeier2018<empty citation>**. Die erste Version wurde von Johannes Titz, Nils Heimhuber und Feline Baumgärtel erstellt. Weitere Mitwirkende waren Vivien Lungwitz und Annika Sternkopf. Zahlreiche Korrekturen und Ergänzungen erfolgten durch Johannes Titz. Für wertvolle Hinweise bedanken wir uns bei Lea Haase.

Einige Hinweise vorab: Lateinische Buchstaben ($\overline{x}, s^2, s$) werden für die Stichprobe verwendet, griechische Buchstaben (μ, σ^2, σ) für die Population. Ein Dach über einer Statistik ($\hat{\sigma}$) steht für eine Schätzung des Parameters. Abkürzungen:
MWU: Mittelwertsunterschied US: Unabhängige Stichproben AS: Abhängige Stichproben

Lagemaße

Modalwert

Der Modalwert ist der häufigste Wert aller Messwerte. Es kann mehrere Modal-Werte geben.

Mittelwert

Auch bezeichnet als arithmetisches Mittel, für den Mittelwert der Population wird der Buchstabe μ verwendet.

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

Median

Der Median ist der Wert in der Mitte aller in einer Rangreihe geordneten Messwerte.

$$\text{Tiefe}_{\text{Med}} = \frac{n+1}{2}$$

Quartile

Die Werte müssen in eine Rangreihe gebracht werden. Für das untere Quartil wird die Tiefe von unten gezählt, fürs obere Quartil von oben. Quartile werden manchmal auch als die Quantile Q_{25}, Q_{75} bezeichnet.

$$\text{Tiefe}_{\text{Quartil}} = \frac{\lfloor \text{Tiefe}_{\text{Med}} \rfloor + 1}{2}$$

Zäune bei Boxplots

Die Zäune werden benötigt um die Whiskers zu bestimmen. Die Whiskers sind durch tatsächlich vorkommende Werte repräsentiert. Von den Zäunen aus geht man in Richtung Box, bis man den ersten vorkommenden Wert findet.

$$\begin{aligned}\text{Zaun}_{\text{unten}} &= Q_{25} - 1{,}5 \cdot \text{IQA}\\ \text{Zaun}_{\text{oben}} &= Q_{75} + 1{,}5 \cdot \text{IQA}\end{aligned}$$

Streuungsmaße

Standardabweichung

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

Varianz

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

Schätzung für Populationsvarianz

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1} \cdot s^2$$

Hinweis: In der mathematischen Statistik wird die *Stichprobenvarianz* üblicherweise mit dem Nenner $n-1$ definiert, da dieser Schätzer erwartungstreu für die Populationsvarianz ist. In den Sozialwissenschaften wird hingegen oft die Variante mit n als *Varianz der Stichprobe* bezeichnet. Hat man dagegen die gesamte Population, so wird auch die *Populationsvarianz* mit n im Nenner berechnet.

Spannweite (Range)

$$R = x_{\max} - x_{\min}$$

Mittlerer absoluter Abstand

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

Interquartilsabstand

$$\text{IQA} = Q_{75} - Q_{25}$$

Transformation

z-Standardisierung

$$z_i = \frac{x_i - \overline{x}}{s}$$
$$\overline{z} = 0$$
$$s_z = 1$$

Zentrierung

$c_i = x_i - \overline{x}$

IQ-Standardisierung

$\mathrm{IQ_i} = z_i \cdot 15 + 100$

Standardfehler für Stichprobenverteilungen

Binomialverteilung

$$\sigma_{\text{Anzahl}} = \sqrt{np(1-p)}$$

$$\sigma_{\text{Anteil}} = \sqrt{\frac{p(1-p)}{n}}$$

Mittelwerte

$$\hat{\sigma}_{\overline{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

MWU für AS

$$\hat{\sigma}_{\overline{d}} = \frac{\hat{\sigma}_d}{\sqrt{n}}$$

MWU für US

wenn $n_A = n_B$:

$$\hat{\sigma}_{\overline{x}_\mathrm{A} - \overline{x}_\mathrm{B}} = \sqrt{\hat{\sigma}^2_{\overline{x}_\mathrm{A}} + \hat{\sigma}^2_{\overline{x}_\mathrm{B}}}$$

wenn $n_A \neq n_B$

$$\hat{\sigma}_{\overline{x}_\mathrm{A} - \overline{x}_\mathrm{B}} = \sqrt{\frac{(n_A - 1) \cdot \hat{\sigma}^2_\mathrm{A} + (n_\mathrm{B} - 1) \cdot \hat{\sigma}^2_\mathrm{B}}{(n_\mathrm{A} - 1) + (n_\mathrm{B} - 1)} (\frac{1}{n_\mathrm{A}} + \frac{1}{n_\mathrm{B}})}$$

Bei Ungleichheit der Varianzen in der Population

$$\hat{\sigma}_{\overline{x}_A - \overline{x}_B} = \sqrt{\frac{\hat{\sigma}^2_A}{n_A} + \frac{\hat{\sigma}^2_B}{n_B}}$$

Zusammenhangsmaße

Kovarianz

$$\mathrm{cov}(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Korrelation

Pearson-Korrelationskoeffizient
(Produkt-Moment-Korrelation)

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{1}{n} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

Partialkorrelation

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r^2_{xz}} \cdot \sqrt{1 - r^2_{yz}}}$$

Phi-Koeffizient (für dichotome Variablen)

$$\Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

	nein	ja
nein	a	b
ja	c	d

Yules Q (für dichotome Variablen)

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

Interrater-Reliabilität (Kappa)

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$
$$p_e = \sum_{i=1}^{k} p_{i\cdot}\, p_{\cdot i}$$

p_o: beobachtete Übereinstimmung
p_e: erwartete Übereinstimmung unter Zufall
$p_{i\cdot}$: die Randwahrscheinlichkeit der Zeile i
$p_{\cdot i}$: die Randwahrscheinlichkeit der Spalte i
k: die Anzahl der Kategorien

Einfache lineare Regression

Regressionsgleichung

$$\hat{y}_i = a + b \cdot x_i$$
$$b = \frac{\mathrm{cov}(x,y)}{s_x^2} = r\frac{s_y}{s_x}$$
$$a = \overline{y} - b\overline{x}$$
$$y_i = \hat{y}_i + e_i$$

Determinationskoeffizient

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Standardschätzfehler

$$s_e = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = s_y \cdot \sqrt{1 - r^2}$$

$$\hat{\sigma}_e = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Multiple Regression

In den folgenden Formeln wird zur Veranschaulichung der Fall mit zwei Prädiktoren dargestellt. Für mehr als zwei Prädiktoren lassen sich die Regressionsgleichung, die Beziehung zwischen b und β, die Berechnung von b_0 sowie die Formeln für den Standardschätzfehler analog erweitern. Die Berechnung der β-Gewichte und von R^2 erfordert jedoch eine Matrixschreibweise, die nicht überall unterichtet wird.

Regressionsgleichung

für Originalwerte:

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i}$$

wobei:

$$b_1 = \beta_1 \cdot \frac{s_y}{s_{x_1}}$$
$$b_2 = \beta_2 \cdot \frac{s_y}{s_{x_2}}$$
$$b_0 = \overline{y} - b_1 \cdot \overline{x}_1 - b_2 \cdot \overline{x}_2$$

für z-standardisierte Werte:

$$\hat{z}_{y_i} = \beta_1 \cdot z_{x_{1i}} + \beta_2 \cdot z_{x_{2i}}$$

wobei:

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}$$

Gütemaße

Multipler Determinationskoeffizient

$$R^2 = \beta_1 \cdot r_{yx_1} + \beta_2 \cdot r_{yx_2}$$

Multipler Korrelationskoeffizient

$$R = \sqrt{R^2}$$

Standardschätzfehler für Originalwerte

$$s_e = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = s_y\sqrt{1 - R^2}$$

$$\hat{\sigma}_e = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$$

k ist die Anzahl der Prädiktoren

Standardschätzfehler für z-standardisierte Werte

$$s_e = \sqrt{1 - R^2}$$

Konfidenzintervalle

Binomialverteilung (Approximation)

Die folgende Approximation über die Normalverteilung ist nur nutzbar, wenn $\sigma^2_{\text{Anzahl}} = n \cdot p \cdot (1-p) > 9$

für Anteile

$$\hat{p} \pm \sigma_{\text{Anteil}} \cdot z_{\text{Konfidenz}}$$

für Anzahlen/Objekte

$$\hat{p} \cdot n \pm \sigma_{\text{Anzahl}} \cdot z_{\text{Konfidenz}}$$

Mittelwert

$$\overline{x} \pm \hat{\sigma}_{\overline{x}} \cdot t_{\text{df,Konfidenz}}$$

MWU, US

$$(\overline{x}_{\text{A}} - \overline{x}_{\text{B}}) \pm \hat{\sigma}_{\overline{x}_{\text{A}} - \overline{x}_{\text{B}}} \cdot t_{\text{df,Konfidenz}}$$

MWU, AS

$$\overline{d} \pm \hat{\sigma}_{\overline{d}} \cdot t_{\text{df,Konfidenz}}$$

t-Test

Mittelwert gegen Konstante

wobei c (constant) vorgegeben

$$t = \frac{\overline{x} - c}{\hat{\sigma}_{\overline{x}}}$$

$$\text{df} = n - 1$$

MWU, US

$$t = \frac{\overline{x}_A - \overline{x}_B}{\hat{\sigma}_{\overline{x}_A - \overline{x}_B}}$$

$$\text{df} = (n_A - 1) + (n_B - 1)$$

Hinweis: Berechnung der df gilt nur bei Gleichheit der Populationsvarianzen.

MWU, AS

$$t_{\text{AS}} = \frac{\overline{d}}{\hat{\sigma}_{\overline{d}}}$$

$$\text{df}_{\text{AS}} = n - 1$$

ANOVA US

F-Wert

$$F(\text{df}_{\text{zw}}, \text{df}_{\text{inn}}) = \frac{\hat{\sigma}^2_{\text{zw}}}{\hat{\sigma}^2_{\text{inn}}}$$

Schätzer der Populationsvarianzen

$$\hat{\sigma}^2_{\text{zw}} = \frac{\text{QS}_{\text{zw}}}{\text{df}_{\text{zw}}}$$

$$\hat{\sigma}^2_{\text{inn}} = \frac{\text{QS}_{\text{inn}}}{\text{df}_{\text{inn}}}$$

Hinweis: diese werden auch oft als Mean Squares (MS) bezeichnet.

Quadratsummenzerlegung

$$\text{QS}_{\text{ges}} = \text{QS}_{\text{zw}} + \text{QS}_{\text{inn}}$$

Quadratsummen

$$\text{QS}_{\text{zw}} = \sum_{j=1}^{k} n_j \cdot (\overline{x}_j - \overline{\overline{x}})^2$$

$$\text{QS}_{\text{inn}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

Freiheitsgrade

$$\text{df}_{\text{zw}} = k - 1$$

$$\text{df}_{\text{inn}} = \sum_{j=1}^{k} (n_j - 1) = N - k$$

Notation

k: Anzahl der Bedingungen
n_j: Anzahl der Messwerte pro Bedingung j
N: Gesamtzahl der Messwerte
$\overline{\overline{x}}$: Gesamtmittelwert

Hinweis: Gilt für balanciertes Design.

F-Wert

$$F = \frac{\hat{\sigma}^2_{\text{UV}}}{\hat{\sigma}^2_{\text{res}}}$$

Schätzer der Populationsvarianzen

$$\hat{\sigma}^2_{\text{UV}} = \frac{\text{QS}_{\text{UV}}}{\text{df}_{\text{UV}}}$$
$$\hat{\sigma}^2_{\text{res}} = \frac{\text{QS}_{\text{res}}}{\text{df}_{\text{res}}}$$

Hinweis: diese werden auch oft als Mean Squares (MS) bezeichnet.

Quadratsummen

$$\begin{aligned}
\text{QS}_{\text{ges}} &= \sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ji} - \overline{\overline{x}})^2 \\
\text{QS}_{\text{UV}} &= \sum_{j=1}^{k} n \cdot (\overline{x}_j - \overline{\overline{x}})^2 \\
\text{QS}_{\text{Person}} &= \sum_{i=1}^{n} k \cdot (\overline{x}_i - \overline{\overline{x}})^2 \\
\text{QS}_{\text{res}} &= \text{QS}_{\text{ges}} - \text{QS}_{\text{UV}} - \text{QS}_{\text{Person}}
\end{aligned}$$

Quadratsummenzerlegung

$$\text{QS}_{\text{ges}} = \text{QS}_{\text{UV}} + \text{QS}_{\text{Person}} + \text{QS}_{\text{res}}$$

Freiheitsgrade

$$\begin{aligned}
\text{df}_{\text{gesamt}} &= N - 1 \\
\text{df}_{\text{UV}} &= k - 1 \\
\text{df}_{\text{Person}} &= n - 1 \\
\text{df}_{\text{res}} &= (k-1)(n-1)
\end{aligned}$$

Notation

k: Anzahl der Bedingungen
n: Anzahl der Personen
N: Gesamtzahl der Messwerte
$\overline{\overline{x}}$: Gesamtmittelwert

Zweifaktorielle Varianzanalyse

Hinweis: gilt nur bei balancierten Designs (gleiche Stichprobengröße)

F-Werte

$$F_A = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_{\text{inn}}^2}$$

$$F_B = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_{\text{inn}}^2}$$

$$F_{A\times B} = \frac{\hat{\sigma}_{A\times B}^2}{\hat{\sigma}_{\text{inn}}^2}$$

Populationsvarianzen

$$\hat{\sigma}_A^2 = \frac{\text{QS}_A}{\text{df}_A}$$

$$\hat{\sigma}_B^2 = \frac{\text{QS}_B}{\text{df}_B}$$

$$\hat{\sigma}_{A\times B}^2 = \frac{\text{QS}_{A\times B}}{\text{df}_{A\times B}}$$

$$\hat{\sigma}_{\text{inn}}^2 = \frac{\text{QS}_{\text{inn}}}{\text{df}_{\text{inn}}}$$

Freiheitsgrade

$$\text{df}_{\text{inn}} = \sum_j^k \sum_l^m (n_{jl} - 1) = N - km$$

$$\text{df}_{\text{A}} = k - 1$$

$$\text{df}_{\text{B}} = m - 1$$

$$\text{df}_{\text{AxB}} = (k-1)(m-1)$$

Quadratsummen

$$\text{QS}_{\text{ges}} = \text{QS}_A + \text{QS}_B + \text{QS}_{A\times B} + \text{QS}_{\text{inn}}$$

$$\text{QS}_{\text{ges}} = \sum_j^k \sum_l^m \sum_i^n (x_{ijl} - \overline{\overline{x}})^2$$

$$\text{QS}_{\text{inn}} = \sum_j^k \sum_l^m \sum_i^n (x_{ijl} - \overline{x}_{jl})^2$$

$$\text{QS}_A = n \cdot m \sum_j^k (\overline{x}_{j\cdot} - \overline{\overline{x}})^2$$

$$\text{QS}_B = n \cdot k \sum_l^m (\overline{x}_{\cdot l} - \overline{\overline{x}})^2$$

$$\text{QS}_{A\times B} = n \sum_{j=1}^k \sum_{l=1}^m \left(\overline{x}_{jl} - \overline{x}_{j\cdot} - \overline{x}_{\cdot l} + \overline{\overline{x}}\right)^2$$

$$\text{QS}_{A\times B} = \text{QS}_{\text{ges}} - \text{QS}_{\text{A}} - \text{QS}_{\text{B}} - \text{QS}_{\text{inn}}$$

Notation

- N: Gesamtzahl der Messwerte der Untersuchung
- n: Anzahl der Messwerte pro Bedingung
- k: Anzahl der Stufen des Faktors A
- m: Anzahl der Stufen des Faktors B
- $\overline{\overline{x}}$: Gesamtmittelwert
- $\overline{x}_{j\cdot}$: Mittelwert für Faktor A, Bedingung j
- $\overline{x}_{\cdot l}$: Mittelwert für Faktor B, Bedingung l

Kontrastanalyse US

Voraussetzung: $\sum_i^k \lambda_i = 0$

$$F_{\text{Kontrast}} = \frac{\hat{\sigma}_{\text{Kontrast}}^2}{\hat{\sigma}_{\text{inn}}^2}$$

oder

$$F = t^2$$

$$|t_{\text{Kontrast}}| = \sqrt{F_{\text{Kontrast}}}$$

Geschätzte Populationsvarianz

$$\hat{\sigma}_{\text{Kontrast}}^2 = \frac{\text{QS}_{\text{Kontrast}}}{\text{df}_{\text{Kontrast}}}$$

$$\hat{\sigma}_{\text{inn}}^2 = \frac{\text{QS}_{\text{inn}}}{\text{df}_{\text{inn}}}$$

Siehe auch ANOVA US.

Quadratsummen

$$\text{QS}_{\text{Kontrast}} = \frac{\left(\sum_{i=1}^k \lambda_i \overline{x}_i\right)^2}{\sum_{i=1}^k \frac{\lambda_i^2}{n_i}}$$

Freiheitsgrade

$$\text{df}_{\text{Kontrast}} = 1$$

$$\text{df}_{\text{inn}} = N - k$$

Notation

- N: Gesamtzahl der Messwerte
- i: Bedingung
- k: Anzahl der Bedingungen (Gruppen)
- n_i: Anzahl der Messwerte pro Bedingung i

Kontrastanalyse für AS

Hinweis: Gilt für den Fall ohne Subgruppen.

Voraussetzung: $\sum_{j}^{k} \lambda_j = 0$

$$t_{\text{Kontrast}} = \frac{\overline{L}}{\sqrt{\frac{\hat{\sigma}_L^2}{n}}}$$

$$L_i = \sum_{j=1}^{k} (x_{ij} \cdot \lambda_j)$$

$$\hat{\sigma}_L^2 = \frac{1}{n-1} \sum_{i=1}^{n} (L_i - \overline{L})^2$$

$$\text{df}_{\text{Kontrast}} = n - 1$$

Notation

- $\overline{L}$: Mittelwert $L-$Werte
- $\hat{\sigma}_L^2$: geschätzte Populationsvarianz der L-Werte
- i: Objekt/Person
- n: Anzahl der Objekte/Personen, für die mehrere Messungen durchgeführt wurden
- j: Bedingung
- k: Anzahl Bedingungen

Nonparametrische Verfahren

χ^2-Anpassungstests

für eine Variable

$$\chi^2 = \sum_{i}^{k} \frac{(f_{b,i} - f_{e,i})^2}{f_{e,i}}$$

$$f_{e,i} = N \cdot P_i$$

$$\text{df} = k - 1$$

für zwei Variablen

$$\chi^2 = \sum_{i}^{k} \sum_{j}^{m} \frac{(f_{b,ij} - f_{e,ij})^2}{f_{e,ij}}$$

$$f_{e,ij} = N \cdot P_{ij}$$

$$\text{df} = k \cdot m - 1$$

χ^2-Unabhängigkeitstest

$$\chi^2 = \sum_{i}^{k} \sum_{j}^{m} \frac{(f_{b,ij} - f_{e,ij})^2}{f_{e,ij}}$$

$$f_{e,ij} = \frac{Z_i \cdot S_j}{N}$$

$$\text{df} = (k-1) \cdot (m-1)$$

Bei 2 dichotomen Variablen: $\chi^2 = \phi^2 \cdot N$

U-Test nach Mann und Whitney (Wilcoxon Rangsummen-Test)

1. Messwerte insgesamt in Rangreihe bringen und jedem Messwert einen Rangplatz zuweisen
2. T_1: Summe der Rangplätze der Gruppe 1 T_2: Summe der Rangplätze der Gruppe 2
3. $U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - T_1$
   $U' = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - T_2 = n_1 \cdot n_2 - U$
4. Prüfgröße für die bei uns verwendete Tabelle ist der kleinere der beiden U-Werte

Wilcoxon-Test für AS (Vorzeichenrangtest)

1. Differenzen der Messwertpaare bilden
2. den Beträgen der Differenzen Rangplatz zuweisen (beim kleinsten Betrag mit Rangplatz 1 beginnen)
3. T_-: Summe der Rangplätze der negativen Differenzen
   T_+: Summe der Rangplätze der positiven Differenzen
4. Prüfgröße für die Tabelle ist der kleinere der beiden T-Werte

Notation

- $f_{b,i}$: beobachtete Häufigkeit
- $f_{e,i}$: erwartete Häufigkeit
- P_i: in der Merkmalsausprägung erwarteter Anteil
- N: Anzahl der Untersuchungsteilnehmer
- k, m: Anzahl der Merkmalsausprägungen beider Merkmale
- Z_i: Zeilenhäufigkeit
- S_j: Spaltenhäufigkeit
- ϕ: Phi-Koeffizient

Faktorenanalyse

Bei z-standardisierten Variablen, Hauptkomponenten-Analyse und orthogonaler Extraktion gilt:

1. Faktorladung: $a_{m,f} = r_{m,f}$
2. Eigenwert: $\lambda_f = \sum_{m=1}^{M} a_{m,f}^2$
3. Kommunalität: $h_m^2 = \sum_{f=1}^{F} a_{m,f}^2$
4. durch Faktor aufgeklärter Varianzanteil: $\frac{\lambda_f}{M}$

Notation

m:	Variable
M:	Anzahl der Variablen
f:	Faktor
F:	Anzahl der ausgewählten Faktoren
i:	Person/Objekt
r:	Korrelation

Clusteranalyse

Nominalskalierte Variablen

		Fall x	
		+	−
Fall y	+	a	c
	−	b	d

Tanimoto-Koeffizient

$$T = \frac{a}{a+b+c}$$

M-Koeffizient

$$M = \frac{a+d}{a+b+c+d}$$

Intervallskalierte Variablen

a,b:	Fälle
J:	Anzahl der Variablen (Dimensionen)
r:	Minkowski-Konstante

Minkowski-Metrik

$$d_{a,b} = \left[\sum_{j=1}^{J} |x_{aj} - x_{bj}|^r\right]^{\frac{1}{r}}$$

Euklidische Distanz

$$d_{a,b} = \sqrt{\sum_{j=1}^{J} |x_{aj} - x_{bj}|^2}$$

Manhattan-Distanz / City-Block-Metrik

$$d_{a,b} = \sum_{j=1}^{J} |x_{aj} - x_{bj}|$$

Konventionen nach Cohen

„These values are necessarily somewhat arbitrary, but were chosen so as to seem reasonable. The reader can render his own judgment as to their reasonableness.“ (**cohen1962**)

	d/g	r/w/ϕ	η^2
klein	± 0.2	± 0.1	0.01
mittel	± 0.5	± 0.3	0.06
groß	± 0.8	± 0.5	0.14

Mittelwertsunterschiede

$$d = \frac{\overline{x} - c}{s}$$
$$d = \frac{\overline{x}_A - \overline{x}_B}{s_{AB}}$$
$$g = \frac{\overline{x} - c}{\hat{\sigma}}$$
$$g = \frac{\overline{x}_A - \overline{x}_B}{\hat{\sigma}_{AB}}$$

$$s_{AB} = \sqrt{\frac{s_A^2\, n_A + s_B^2\, n_B}{n_A + n_B}}$$
$$\hat{\sigma}_{AB} = \sqrt{\frac{(n_A - 1)\, \hat{\sigma}_A^2 + (n_B - 1)\, \hat{\sigma}_B^2}{n_A + n_B - 2}}$$

Das normale g ist nach oben verzerrt. Für eine unverzerrte Schätzung in der Population (**hedges1981**):

$$g^U \approx g\left(1 - \frac{3}{4(n_A + n_B) - 9}\right)$$

ANOVA

US

$$\eta^2 = \frac{\mathrm{QS}_{\mathrm{zw}}}{\mathrm{QS}_{\mathrm{ges}}}$$
$$\eta_p^2 = \frac{\mathrm{QS}_{\mathrm{zw}}}{\mathrm{QS}_{\mathrm{zw}} + \mathrm{QS}_{\mathrm{Fehler}}}$$

AS

$$\eta^2 = \frac{\mathrm{QS}_{\mathrm{zw}}}{\mathrm{QS}_{\mathrm{ges}}}$$
$$\eta_p^2 = \frac{\mathrm{QS}_{\mathrm{zw}}}{\mathrm{QS}_{\mathrm{zw}} + \mathrm{QS}_{\mathrm{res}}}$$

mehrfaktoriell

$$\eta^2 = \frac{\mathrm{QS}_{\mathrm{Effekt}}}{\mathrm{QS}_{\mathrm{ges}}}$$
$$\eta_p^2 = \frac{\mathrm{QS}_{\mathrm{Effekt}}}{\mathrm{QS}_{\mathrm{Effekt}} + \mathrm{QS}_{\mathrm{inn}}}$$

Kontrast-Analyse

$$r_{\mathrm{effectsize}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(\lambda_i - \overline{\lambda})}{s_x \cdot s_\lambda}$$
$$r_{\mathrm{contrast}} = \frac{\mathrm{QS}_{\mathrm{Kontrast}}}{\mathrm{QS}_{\mathrm{Kontrast}} + \mathrm{QS}_{\mathrm{inn}}}$$
$$r_{\mathrm{alerting}}^2 = \frac{\mathrm{QS}_{\mathrm{Kontrast}}}{\mathrm{QS}_{\mathrm{zw}}}$$
$$g_{\mathrm{contrast}} = \frac{\overline{L}}{\hat{\sigma}_L}$$

Effektgrößen aus Signifikanztests

t-Test: Einstichprobenfall

$$g = \frac{t}{\sqrt{n}}$$

$$d = \frac{t}{\sqrt{\text{df}}}$$

t-Test: US

$$g = t_{\text{US}} \cdot \sqrt{\frac{n_{\text{A}} + n_{\text{B}}}{n_{\text{A}} \cdot n_{\text{B}}}}$$

$$d = t_{\text{US}} \cdot \frac{n_{\text{A}} + n_{\text{B}}}{\sqrt{\text{df}} \cdot \sqrt{n_{\text{A}} \cdot n_{\text{B}}}}$$

$$r = \sqrt{\frac{t_{\text{US}}^2}{t_{\text{US}}^2 + \text{df}}}$$

Letzteres gilt auch für F, wenn df_{zw} 1 sind, da in diesem Fall $F = t^2$.

Vereinfachung, wenn $n_{\text{A}} = n_{\text{B}}$:

$$d = \frac{2 \cdot t_{\text{US}}}{\sqrt{\text{df}}}$$

$$g = \frac{2 \cdot t_{\text{US}}}{\sqrt{n}}$$

t-Test: AS

$$g = \frac{t_{\text{AS}}}{\sqrt{n}}$$

$$d = \frac{t_{\text{AS}}}{\sqrt{\text{df}}}$$

Kontrastanalyse US

aus F-Test

$$r_{\text{effectsize}} = \sqrt{\frac{F_{\text{Kontrast}}}{F_{\text{zw}} \cdot \text{df}_{\text{zw}} + \text{df}_{\text{inn}}}}$$

$$r_{\text{contrast}} = \sqrt{\frac{F_{\text{Kontrast}}}{F_{\text{Kontrast}} + \text{df}_{\text{inn}}}} = \sqrt{\frac{t_{\text{Kontrast}}^2}{t_{\text{Kontrast}}^2 + \text{df}}}$$

Kontrastanalyse AS

aus t-Test

$$g = \frac{t}{\sqrt{n}}$$

χ^2-Tests

$$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{b,i} - P_{e,\,i})^2}{P_{e,i}}}$$

aus Ergebnis χ^2-Test:

$$w = \sqrt{\frac{\chi^2}{N}}$$

aus 2 dichotomen Variablen

$$w = \sqrt{\frac{\chi^2}{N}} = \phi$$

ANOVA

US

$$\eta^2 = \frac{F \cdot \text{df}_{\text{zw}}}{F \cdot \text{df}_{\text{zw}} + \text{df}_{\text{inn}}}$$

Odds Ratio (OR)

$$\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

	Krankheit	Keine Krankheit
Risikofaktor	a	b
Ohne Risikofaktor	c	d

Effektgrößen aus Effektgrößen

Abstandsmaße aus Abstandsmaßen

$$d = g \cdot \sqrt{\frac{n}{\mathrm{df}}}$$

$$g = d \cdot \sqrt{\frac{df}{n}}$$

Korrelationen aus Abstandsmaßen

$$r = \frac{d}{\sqrt{d^2 + \frac{1}{p \cdot q}}}$$

$$r = \sqrt{\frac{g^2(n_A \cdot n_B)}{g^2(n_A \cdot n_B) + (n_A + n_B)\mathrm{df}}}$$

Abstandsmaße aus Korrelationen

$$d = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{\frac{1}{p \cdot q}}$$

$$g = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{\frac{(n_A + n_B) \cdot df}{n_A \cdot n_B}}$$

wobei

$$p = \frac{n_A}{n_A + n_B}$$

und

$$q = \frac{n_B}{n_A + n_B}$$

Meta-Analyse

Mittlere, an der Stichprobe gewichtete, Effektgröße: $\overline{r} = \frac{\sum_{i=1}^{n}(N_i r_i)}{\sum_{i=1}^{n} N_i}$

Wahrscheinlichkeitstheorie

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit von A unter der Bedingung B wird notiert als $P(A|B)$ und berechnet sich durch:

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

Hierbei ist $P(A \cap B)$ die Wahrscheinlichkeit, dass sowohl A als auch B eintreten.

Additionssatz der Wahrscheinlichkeit

Die Wahrscheinlichkeit, dass mindestens eines der Ereignisse A oder B eintritt, ergibt sich zu:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Sind A und B *disjunkt* (d. h. $A \cap B = \emptyset$), so gilt die *Summenregel*:

$$P(A \cup B) = P(A) + P(B)$$

Allgemeines Inklusions-Exklusions-Prinzip

Für drei Ereignisse A, B und C gilt:

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - [P(A \cap B) + P(A \cap C) + P(B \cap C)] \\ & + P(A \cap B \cap C) \end{aligned}$$

Das Prinzip verallgemeinert sich für n Ereignisse abwechselnd mit Addition und Subtraktion der Schnittwahrscheinlichkeiten aller k-fachen Überlappungen.

Produktregel

Aus der Definition der bedingten Wahrscheinlichkeit folgt:

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

Alternativ kann die Produktregel auch in umgekehrter Reihenfolge geschrieben werden:

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

Stochastische Unabhängigkeit

Zwei Ereignisse A und B heißen *stochastisch unabhängig*, wenn das Eintreten des einen keinen Einfluss auf die Wahrscheinlichkeit des anderen hat:

$$P(A|B) = P(A) \quad \text{bzw.} \quad P(B|A) = P(B)$$

Dann folgt unmittelbar:

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

Satz der totalen Wahrscheinlichkeit

Sei $\{B_1, B_2, \ldots, B_n\}$ eine Zerlegung des Ergebnisraums in paarweise disjunkte Ereignisse mit $P(B_i) > 0$. Dann gilt:

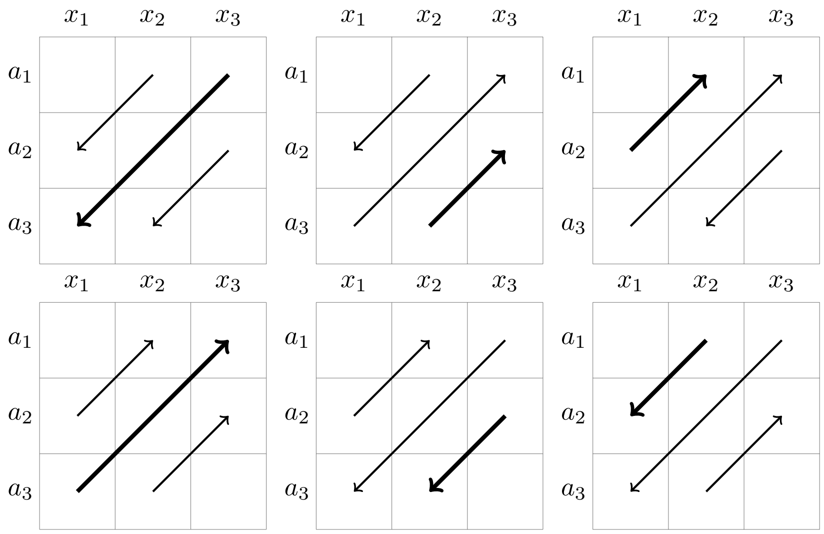
$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

Wahrscheinlichkeitsrevision (Bayes)

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c)}$$

Conjoint-Measurement-Theory

Messaxiome fürs Intervallskalenniveau lassen sich teilweise durch ein Conjoint-Measurement-Experiment prüfen. Hierbei geht man davon aus, dass $P = f(A,X)$ wobei f eine nicht-interaktive Funktion ist. Geprüft werden müssen die einfache Aufhebung (single cancellation) und die doppelte Aufhebung (double cancellation). Single cancellation erfordert, dass innerhalb jeder Zeile und innerhalb jeder Spalte die Ordnung der Präferenzen konsistent ist. Die Doppelaufhebung beschränkt die möglichen Ordnungen der Matrix. Wenn alle Spalten und Zeilen bereits geordnet sind, lässt sich über folgende Visualisierung die Doppelaufhebung prüfen. Hierbei entsprechen die Pfeile dem $\leq$-Zeichen. Dünne Pfeile zwischen den Zellen repräsentieren die Antezedenz-Ordnungsbeziehungen, dicke Pfeile repräsentieren die Konsequenz-Ordnungsbeziehung.



R Verteilungsfunktionen

Funktion	**Verteilung**	**Parameter**
`dnorm`	Normalverteilung	`x`, `mean`, `sd`
`pnorm`	Normalverteilung	`q`, `mean`, `sd`, `lower.tail`, `log.p`
`qnorm`	Normalverteilung	`p`, `mean`, `sd`, `lower.tail`, `log.p`
`rnorm`	Normalverteilung	`n`, `mean`, `sd`
`dbinom`	Binomialverteilung	`x`, `size`, `prob`
`pbinom`	Binomialverteilung	`q`, `size`, `prob`, `lower.tail`, `log.p`
`qbinom`	Binomialverteilung	`p`, `size`, `prob`, `lower.tail`, `log.p`
`rbinom`	Binomialverteilung	`n`, `size`, `prob`
`dt`	t-Verteilung	`x`, `df`, `ncp`
`pt`	t-Verteilung	`q`, `df`, `ncp`, `lower.tail`, `log.p`
`qt`	t-Verteilung	`p`, `df`, `ncp`, `lower.tail`, `log.p`
`rt`	t-Verteilung	`n`, `df`, `ncp`
`df`	F-Verteilung	`x`, `df1`, `df2`, `ncp`
`pf`	F-Verteilung	`q`, `df1`, `df2`, `ncp`, `lower.tail`, `log.p`
`qf`	F-Verteilung	`p`, `df1`, `df2`, `ncp`, `lower.tail`, `log.p`
`rf`	F-Verteilung	`n`, `df1`, `df2`, `ncp`

Präfix	**Erklärung**
`d`	Dichtefunktion: Berechnet die Wahrscheinlichkeitsdichte (z. B. $f(x)$)
`p`	Verteilungsfunktion: Berechnet die kumulierte Wahrscheinlichkeit (z. B. $P(X \leq x)$)
`q`	Quantilfunktion: Berechnet den Wert x, sodass $P(X \leq x) = p$
`r`	Zufallszahlengenerator: Generiert Zufallswerte aus der Verteilung

Parameter	**Erklärung**
`x`	Werte, an denen die Dichte berechnet wird
`q`	Quantile
`p`	Wahrscheinlichkeiten
`n`	Anzahl der zu generierenden Zufallswerte
`mean`	Mittelwert der Verteilung
`sd`	Standardabweichung
`size`	Anzahl der Versuche
`prob`	Erfolgswahrscheinlichkeit
`df`	Freiheitsgrade
`df1`	Freiheitsgrade des Zählers (F-Verteilung)
`df2`	Freiheitsgrade des Nenners (F-Verteilung)
`ncp`	Nichtzentralitätsparameter (optional)
`lower.tail`	Logisch; wenn TRUE (Standard), werden Wahrscheinlichkeiten $P[X \leq x]$ berechnet
`log.p`	Logisch; wenn TRUE, werden Wahrscheinlichkeiten als Log-Werte ausgegeben

Literatur