Felix Baumann

Wichtige statistische Koeffizienten und Formeln

Skalenniveau	Art des Koeffizienten	Koeffizient	Formel
(ab) **Nominalskala**: (Unterscheidung nach Gleichheit/Ungleichheit; jeder Ausprägung eine Zahl zugeordnet, die keine quantitative Aussage hat) Beispiele: Geschlecht, Religion, Nationalität dichotome Variablen: genau 2 Ausprägungen	Lagemaß (Beschreibung von univariaten Verteilungen)	Modalwert	am häufigsten vorkommende Ausprägung
	Streuungsmaß (Beschreibung von univariaten Verteilungen)	Spannweite/range	Maximum-Minimum
	Zusammenhangsmaße (Kennwerte für bivariate Verteilungen) wichtige Werte für eine Kreuztabelle: beobachtete Häufigkeit (n_{ij}, wobei i = Zeile, j = Spalte), erwartete Häufigkeit ($n_{0\,ij}$: Zeilensumme mal Spaltensumme geteilt durch Gesamtzahl, auch: Indifferenztabelle), relative Häufigkeit (Anteil der Häufigkeit einer Zelle an der Gesamtanzahl), Spaltenprozente (Anteil der Häufigkeit einer Zelle an ihrer Spalte)	„Chi-Quadrat“ (χ) (der Wert steigt mit wachsendem N → Chi^2-Tabelle Freiheitsgrade/dF in einer Kreuztabelle: $(r-1)\cdot(c-1)$)	$Residuum = n_{ij} - n_{0ij}$ $stand.\,Residuum = \frac{Residuum}{\sqrt{n_{0ij}}}$ Vergleichbarkeit: wenn Wert nicht innerhalb [-2;2] → überzufällige Abweichung $Chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n_{0ij})^2}{n_{0ij}}$ doppeltes Summenzeichen: jede Kombination von r = row = Zeilenanzahl und c = column = Spaltenanzahl genau einmal
		Prozentsatzdifferenz (d%) (hauptsächlich für 2x2-Tabellen) 0 = kein Zusammenhang 100 = vollständiger Zusammenhang	d% = Spaltenprozent 1 - Spaltenprozent 2

		phi (Φ) (für 2x2-Tabellen)	ungerichtet (0 bis 1): $Phi = \sqrt{\frac{Chi^2}{N}}$ gerichtet (-1 bis 1): $Phi = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{\sqrt{n_{1.} \cdot n_{2.} \cdot n_{.1} \cdot n_{.2}}}$ $n_{1.}$: Randsumme der ersten Zeile (usw.)
		C (0 bis <1) (für beliebige Tabellengrößen, aber Werte von Tabellengröße abhängig! → C_{korr})	$C = \sqrt{\frac{Chi^2}{N + Chi^2}}$ $C_{max} = \sqrt{\frac{min(r-1\,;\,s-1)}{min(r\,;\,s)}}$ $C_{korr} = \frac{C}{C_{max}}$
		V (0 bis 1) (immer anwendbar)	$V = \sqrt{\frac{Chi^2}{Chi^2_{max}}} = \sqrt{\frac{Chi^2}{N \cdot min(r-1\,;\,c-1)}}$
	besonderes Zusammenhangsmaß: PRE („proportional reduction in error“) $PRE = \frac{E1 - E2}{E1}$ E1: Fehler nach erster Vorhersage (Kennwerte der AV), ohne Kenntnis der UV E2: Fehler mit Kenntnis der UV	lambda (λ) (Vorhersage anhand des Modus: E1 als alle Fälle, die vom Gesamt-Modus der AV nicht erfasst werden, E2 als alle Fälle, die von den partiellen Modi der AV in den einzelnen Ausprägungen der UV nicht erfasst werden)	Zeilenvariable (Randspalte) = AV: $lambda_r = \frac{\sum_{j=1}^{c} max\, n_j - max\, n_{i.}}{N - max\, n_{i.}}$ max n_j = Maxima der Spalten $n_{i.}$ = Randspalte symmetrisch: $lambda_s = \frac{\sum_{j=1}^{c} max\, n_j + \sum_{i=1}^{r} max\, n_i - max\, n_{i.} - max\, n_{.j}}{2 \cdot N - max\, n_{i.} - max\, n_{.j}}$

<table>
<tr><td rowspan="4">(ab) Ordinalskala:
(Einordnung in Rangfolge; Rangfolge an Zahlen ersichtlich; Abstände zwischen Ausprägungen nicht interpretierbar)

Beispiele:
Altersgruppen, (Einkommens-)Schichten, Einstellungsskala („stimme voll zu“ bis „stimme gar nicht zu“)</td><td>Lagemaß
(Beschreibung von univariaten Verteilungen)</td><td>Median
(der Wert des „mittleren“ Falls; 50%-Teiler)</td><td><u>Urliste:</u>
ungerades n: $\tilde{x} = x_{\frac{1}{2}\cdot(n+1)}$
gerades n: $\tilde{x} = \frac{1}{2}\cdot\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$
<u>gruppierte/klassierte Daten:</u>
$\tilde{x} = Q_2 = U + \frac{\frac{1}{2}\cdot N - F_u}{F_m}\cdot K_b$
U: untere exakte Intervallgrenze (des Intervalls, indem die 50% der Fälle erreicht werden)
N: Zahl der Fälle
F_u: kumulierte Häufigkeit unterhalb unterer Intervallgrenze
F_m: Häufigkeit im Intervall
K_b: Intervallbreite (meistens 1)</td></tr>
<tr><td>Streuungsmaß
(Beschreibung von univariaten Verteilungen)</td><td>Interquartilsabstand
(exakte Rechnung genau genommen nur bei > ordinalskalierten Variablen → Klassen)</td><td>$IQ = Q_3 - Q_1$
Q_1 ist der 25%-, Q_3 der 75%-Teiler
für Q_3 bzw. Q_1 in obiger Formel das ½ durch ¾ bzw. ¼ ersetzen und bei der Auswahl des Intervalls die 50% durch 75% bzw. 25%</td></tr>
<tr><td rowspan="3">Zusammenhangsmaße
(Kennwerte für bivariate Verteilungen):
Konkordanzmaße

konkordante Paare (N_C): die Ausprägungen der zwei Fälle sind in beiden Variablen unterschiedlich, und zwar in die gleiche Richtung</td><td>tau_a (τ)
(nur anwendbar ohne Ties, also fast nie)</td><td>$tau_a = \frac{N_C - N_D}{N\cdot(N-1)\cdot\frac{1}{2}}$</td></tr>
<tr><td>tau_b (-1 bis 1)
(für quadratische Tabellen)</td><td>$tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + T_X)\cdot(N_C + N_D + T_Y)}}$</td></tr>
<tr><td>tau_c (-1 bis 1)
(für rechteckige Tabellen)</td><td>$tau_c = \frac{N_C - N_D}{\frac{1}{2}\cdot N^2\cdot\frac{m-1}{m}}$ mit $m = min(r;c)$</td></tr>
</table>

	diskordante Paare (N_D): die Ausprägungen der zwei Fälle sind in beiden Variablen unterschiedlich, und zwar in entgegengesetzten Richtungen Ties auf X (T_X): dieselbe Ausprägung auf der x-Achse (Spalten), unterschiedliche auf der y-Achse (Zeilen) Ties auf Y (T_Y): dieselbe Ausprägung auf der y-Achse (Zeilen), unterschiedliche auf der x-Achse (Spalten) Ties auf X und Y (T_{XY}): dieselbe Ausprägung auf beiden Achsen alles addiert: Gesamtanzahl aller möglichen Paare = $N \cdot (N-1) \cdot \frac{1}{2}$	Somers' d (asymmetrisches Maß: hier wird gerichtet getestet, d.h. UV und AV bestimmen)	$d_{yx} = \frac{N_C - N_D}{N_C + N_D + T_Y}$ (Y=AV, X=UV)
		gamma (γ) (je weniger Ausprägungen, desto größer → steigt bei Recodierung, produziert Ungenauigkeiten)	$gamma = \frac{N_C - N_D}{N_C + N_D}$
	PRE-Maß $PRE = \frac{E1 - E2}{E1}$	gamma als PRE-Maß ($E1 = 0{,}5 \cdot (N_C + N_D)$, $E2 = min(N_C ; N_D)$)	$gamma(PRE) = \frac{N_C + N_D - 2 \cdot min(N_C ; N_D)}{N_C + N_D}$ $gamma(PRE) = \frac{\lvert N_C - N_D \rvert}{N_C + N_D}$

<table>
<tr>
<td rowspan="5">(ab) <u>Metrische Skalen</u>:

Intervallskala:
(„um wie viele Einheiten stärker ist eine Ausprägung?“; willkürlicher Nullpunkt, willkürliche Skaleneinheit; Abstände interpretierbar; gleiche Intervallgrößen zwischen den Ausprägungen)

Beispiele:
(additive) Indizes (z.B. Gewaltbereitschaft), Temperatur in °C

Verhältnis-/Ratioskala:
(„um wieviel größer ist eine Ausprägung?“; natürlicher Nullpunkt, willkürliche Skaleneinheit; Quotienten interpretierbar; Größenverhältnisse zwischen den Zahlen entsprechen Stärke der Merkmalsausprägungen)

Beispiele: Körpergröße, Einkommen (in €), Temperatur in Kelvin</td>
<td>Lagemaß
(Beschreibung von univariaten Verteilungen)</td>
<td>arithmetisches Mittel
(„Mittelwert“)</td>
<td>Grundformel (ungruppierte Daten):
$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$</td>
</tr>
<tr>
<td>Streuungsmaß
(Beschreibung von univariaten Verteilungen)</td>
<td>Varianz/Standardabweichung</td>
<td>$Var = s^2$
Grundformel (in der Stichprobe):
$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$</td>
</tr>
<tr>
<td rowspan="2">Gestaltmaße</td>
<td>Schiefe (v)
(größer 0 → linkssteil/rechtsschief, kleiner 0 → rechtssteil/linksschief)</td>
<td>$v = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$</td>
</tr>
<tr>
<td>Wölbung/Kurtosis (w)
(größer 0 → steiler als Normalverteilung, kleiner 0 → flacher als NV)</td>
<td>$w = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4$</td>
</tr>
<tr>
<td rowspan="2">Zusammenhangsmaße
(Kennwerte für bivariate Verteilungen)</td>
<td>Korrelationskoeffizient r (-1 bis 1)
(gibt den <u>linearen</u> Zusammenhang an)
r² beschreibt den Anteil der Variation der AV, der durch die UV erklärt wird (siehe auch: r² als PRE-Maß)</td>
<td>$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2}}$
aus z-Werten:
$r = \frac{1}{N} \cdot \sum_{i=1}^{N} z_{xi} \cdot z_{yi}$</td>
</tr>
<tr>
<td>Regression
(Berechnung der Geraden, die den Verlauf einer Punktwolke möglichst gut beschreibt, Minimierung des Vorhersagefehlers; auch hier nur <u>linearer</u> Zusammenhang erfasst)</td>
<td>Allgemeine Formel für die Gerade:
$y'_i = a_{yx} + b_{yx} \cdot x_i$

Schnittpunkt mit der y-Achse:
$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$
Steigung:</td>
</tr>
</table>

Absolutskala: (natürlicher Nullpunkt, natürliche Skaleneinheit) Beispiel: Anzahl von irgendetwas (z.B. Anzahl der Teilnehmer einer Studie) **Z-Transformation** (Standardisierung von Verteilungen): $z_{xi}=\frac{x_i-\bar{x}}{s_x}$			$b_{yx}=\frac{\sum_{i=1}^{N}(y_i-\bar{y})\cdot(x_i-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$
	PRE-Maß	Regression/r² als PRE-Maß (E1: Gesamtvariation der AV $E1=\sum_{i=1}^{N}(y_i-\bar{y})^2$ E2: Abweichungen von den Regressionswerten; nicht erklärte Variation $E2=\sum_{i=1}^{N}(y_i-y'_i)^2$)	Gesamtvariation = erklärte Variation + n. e. Variat. $\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}(y'_i-\bar{y})^2+\sum_{i=1}^{N}(y_i-y'_i)^2$ y'_i: Wert bei x_i auf der Regressionsgeraden tatsächlich gilt: $\frac{E1-E2}{E1}=r^2$ r² ist also der Anteil der erklärten Variation an der Gesamtvariation und der Wert für das PRE-Maß
Kombination aus nominalem und metrischem Skalenniveau	Zusammenhangsmaß	eta (η) (0 bis 1) (die Ausprägungen der nominalen UV werden Kolonnen genannt, es gibt k Kolonnen)	$eta=\sqrt{\sum_{j=1}^{k}n_j\cdot(\bar{y}_j-\bar{y})}$ n_j: Anzahl der Fälle in der j-ten Kolonne $\bar{y}_j$: Mittelwert der j-ten Kolonne
	PRE-Maß	eta² (dieselbe Aufteilung in erklärte und nicht erklärte Variation wie bei der Regression: E1 ist die Gesamtvariation, E2 die nicht erklärte Variation erklärte Variation = externe Variation → zwischen Kolonnen nicht erklärte Variation = interne Variation → in den Kolonnen)	Gesamtvariation = erklärte Variation + n. e. Variat. $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2=\sum_{j=1}^{k}n_j\cdot(\bar{y}_j-\bar{y})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$ doppeltes Summenzeichen: in jeder Kolonne wird jeder Wert einmal betrachtet (=alle Werte) $y_{ij}-\bar{y}_j$: Differenz jedes Wertes von seinem jeweiligen Kolonnenmittelwert es gilt: $\frac{E1-E2}{E1}=eta^2$ eta² ist also der Anteil der erklärten Variation an der Gesamtvariation und der Wert für das PRE-Maß

<table>
<tr><td rowspan="5">Schließende Statistik

Versuch, von den Werten einer Stichprobe auf die der ihr zugrunde liegenden Grundgesamtheit (die gesamte Population, über die eine Aussage gemacht wird) zu schließen</td><td rowspan="2">Standard-/Stichprobenfehler</td><td>Anteilswerte (p)
(Bsp.: 30% aller Jugendlichen in Bayern haben Erfahrung mit Cannabis)</td><td>$SF = \sqrt{\frac{p \cdot (1-p)}{n}}$
n: Stichprobenumfang</td></tr>
<tr><td>arithmetisches Mittel</td><td>$SF = \frac{s}{\sqrt{n-1}}$</td></tr>
<tr><td rowspan="2">Konfidenzintervall
(in diesem Intervall um den wahren Wert der GG liegen bei gegebener Irrtumswahrscheinlichkeit die Werte aller möglichen Stichproben; durch Umdrehen der Logik testet man rückwärts auf die GG)</td><td>Anteilswerte</td><td>$KI = p \pm t_{dF;\,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$
t: Wert aus der Tabelle der Quantile der t-Verteilung
$dF = n-1$; bei n > 800 → z-Werte
α: Irrtumswahrscheinlichkeit bei Ablehnung von H_0, standardmäßig 0,05, also 5%</td></tr>
<tr><td>arithmetisches Mittel</td><td>$KI = \bar{x} \pm \frac{s}{\sqrt{n-1}} \cdot t_{dF;\,1-\frac{\alpha}{2}}$</td></tr>
<tr><td rowspan="2">notwendige Stichprobengröße von Zufallsstichproben</td><td>Variante 1: aus Anteilswert und Konfidenzintervall</td><td>$n = \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot p \cdot (1-p)}{d^2}$
$d = 2 \cdot e$, e: Stichprobenfehler, d: Breite des KI
α ist standardmäßig 0,05</td></tr>
<tr><td>Variante 2: aus Größe der GG</td><td><u>Größe der GG > 100.000:</u>
$n = \left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2 \cdot p \cdot (1-p)$
α wird standardmäßig auf 0,05 gesetzt, e auf 0,03, p auf 0,5 (ungünstigster Fall) → $n = 1067{,}11$ als standardisierter Wert
<u>Größe der GG < 100.000:</u>
$n = \frac{N \cdot z_{\frac{\alpha}{2}}^2 \cdot p \cdot (1-p)}{z_{\frac{\alpha}{2}}^2 \cdot p \cdot (1-p) + N \cdot e^2}$</td></tr>
</table>