Mathematische und statistische Methoden I

Sprechstunde
jederzeit nach
Vereinbarung und
nach der Vorlesung

Wallstr. 3, 6. Stock,
Raum 06-206

Dr. Malte Persike


persike@uni-mainz.de


lordsofthebortz.de


twitter.com/methodenlehre


tinyurl.com/gplusmethodenlehre

WiSe 2011/2012

Fachbereich Sozialwissenschaften
Psychologisches Institut
Johannes Gutenberg Universität Mainz

Bortz, S. 153 – 158

Bivariate Daten

Grundlagen

- Bisher wurden Kennwerte für den **univariaten Fall** betrachtet, d.h. für Daten **einer** Variablen
- Mit geschachtelten Kontingenztabellen wurde eine kompakte Darstellungsmöglichkeit für den **multivariaten Fall** beschrieben, d.h. für Daten **mehrerer** Variablen
- In der Statistik sind weitere Verfahren gebräuchlich, die speziell den Zusammenhang zweier Variablen (also für den **bivariaten Fall**) beschreiben.
- **Beispiel:** Man weiß, dass die Nervenleitgeschwindigkeit am Unterarm und der im Intelligenztest gemessene IQ positiv zusammenhängen.
- **Frage:** Wie kann ein solcher Zusammenhang einfach grafisch/numerisch dargestellt werden?

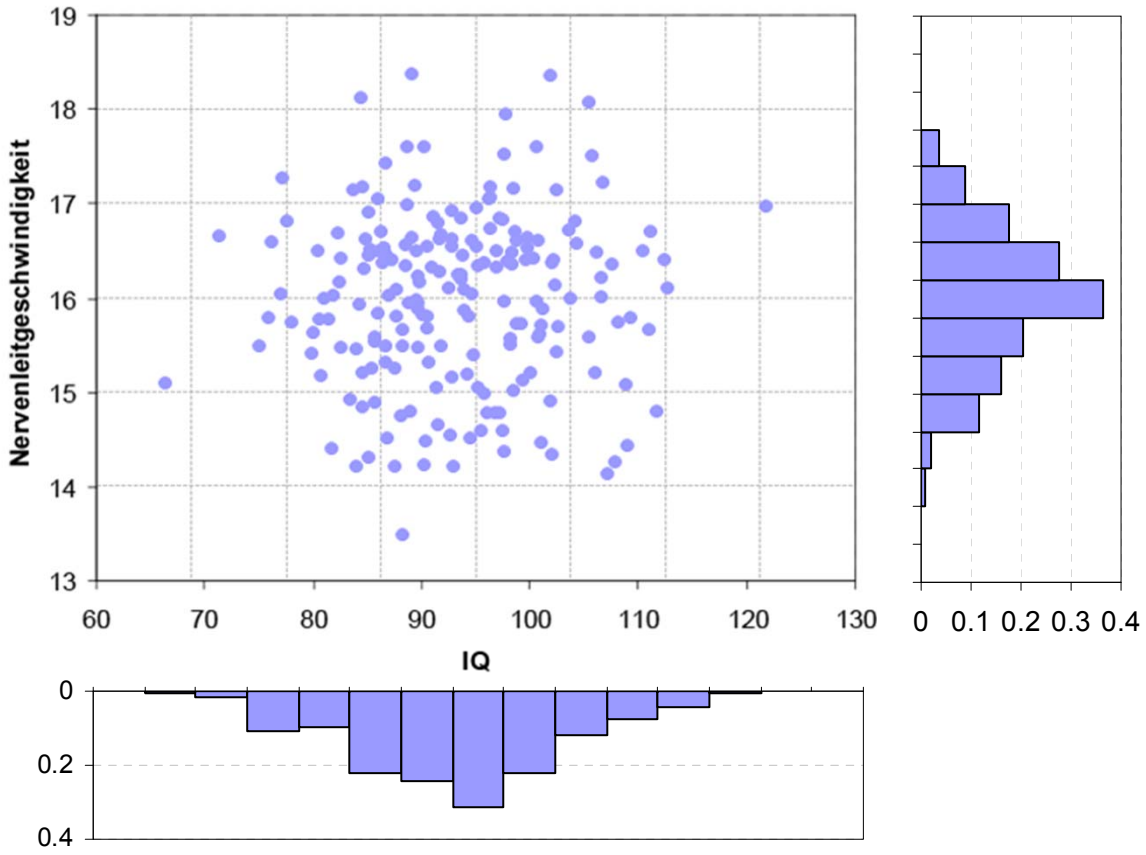
Bivariate Intervalldaten

Grundlagen

- Die Intervallskala trägt Informationen über die **Ordnung von Ausprägungen** und hat eine **feste Einheit** zwischen den Ausprägungen
- Die Werte einer intervallskalierten Variablen sind nicht direkt vergleichbar, wohl aber die **Unterschiede zwischen Werten**
- Weil die Ausprägungen einer festen Einheit folgen, kann man intervallskalierte Daten sowohl grafisch als auch numerisch sehr einfach behandeln.

Bivariate Intervalldaten

Grafische Beschreibung – Scatterplot



Bivariate Intervalldaten

Numerische Beschreibung - Kennwerte

Gewünschte Eigenschaften eines Zusammenhangskoeffizienten

- Sollte die Stärke eines Zusammenhangs numerisch ausdrücken
- Sollte die Richtung des Zusammenhangs anzeigen (sofern sinnvoll)
- Sollte invariant unter zulässigen Transformationen sein (z.B. m in cm)
- Sollte einfach interpretierbar sein

Bivariate Intervalldaten

Numerische Beschreibung - Kovarianz

- Für n Beobachtungen aus einem Zufallsexperiment $x_1 \ldots x_n$ und $y_1 \ldots y_n$ ist die **Kovarianz** definiert als

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

- Die Kovarianz ist Null, wenn **kein Zusammenhang** zwischen den Ausprägungen der Zufallsvariablen besteht
- Die Kovarianz ist positiv, wenn ein **gleichsinniger Zusammenhang** besteht
- Die Kovarianz ist negativ, wenn ein **gegensinniger Zusammenhang** besteht.

Bivariate Intervalldaten

Numerische Beschreibung - Kovarianz

- Die Kovarianz **erfüllt nicht die Forderung der Invarianz** gegenüber erlaubten Transformationen
- Addition einer Konstanten zu x und y:

$$s_{xy}(x+a, y+b) = s_{xy}(x, y)$$

- Aber: Multiplikation von x und y mit einer Konstanten

$$s_{xy}(a \cdot x, b \cdot y) = a \cdot b \cdot s_{xy}(x, y)$$

- Die Kovarianz ist also numerisch schwer zu interpretieren

Bivariate Intervalldaten

Numerische Beschreibung - Korrelation

- Für n Beobachtungen aus einem Zufallsexperiment $x_1 \ldots x_n$ und $y_1 \ldots y_n$ ist der **Korrelationskoeffizient** definiert als

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

- Für die **Richtungsinformation** gelten dieselben Regeln wie bei der Kovarianz

- Bei der Korrelation ist zudem die **Stärke** (der Betrag) des Zusammenhangs interpretier- und vergleichbar.

Bivariate Intervalldaten

Numerische Beschreibung - Korrelation

- Der so definierte Korrelationskoeffizient r_{xy} wird auch als **Produkt-Moment-Korrelation** oder **Korrelationskoeffizient nach Pearson** bezeichnet.
- Für Daten unterhalb Intervallskalenniveau gibt es andere Berechnungsformeln für die Korrelation
- Die Korrelation ist Null, wenn kein Zusammenhang zwischen den Ausprägungen der Zufallsvariablen besteht
- Die Korrelation liegt immer zwischen -1 und 1.
- Negative Werte zeigen einen gegensinnigen, positive Werte einen gleichsinnigen Zusammenhang an
- Die Korrelation ist **anfällig gegenüber Ausreißern**

Bivariate Intervalldaten

Numerische Beschreibung - Vergleich

Kovarianz	Korrelation
$s_{xy}(x,y) = s_{xy}(y,x)$	$r(x,y) = r(y,x)$
$s_{xy}(x, a) = 0$	$r(x, a) =$ nicht def.
$s_{xy}(a, b) = 0$	$r(a, b) =$ nicht def.
$s_{xy}(x, x) = s^2_x(x)$	$r(x, x) = 1$
$s_{xy}(a \cdot x+b, c \cdot y+d)$ $= a \cdot c \cdot s_{xy}(x, y)$	$r(a \cdot x+b, c \cdot y+d) = r(x, y)$ Achtung: Ist a *oder* b negativ, verändert sich das Vorzeichen von r, sind beide negativ, bleibt r gleich.

Mit a, b, c, d = konstante Werte

Bivariate Intervalldaten

Numerische Beschreibung - Faustregeln

- Für die Bewertung der absoluten Höhe der Produkt-Moment-Korrelation existieren Faustregeln nach Cohen (1988)

  $r < \pm 0.10 \rightarrow$ keine Korrelation

  $r < \pm 0.30 \rightarrow$ kleine Korrelation

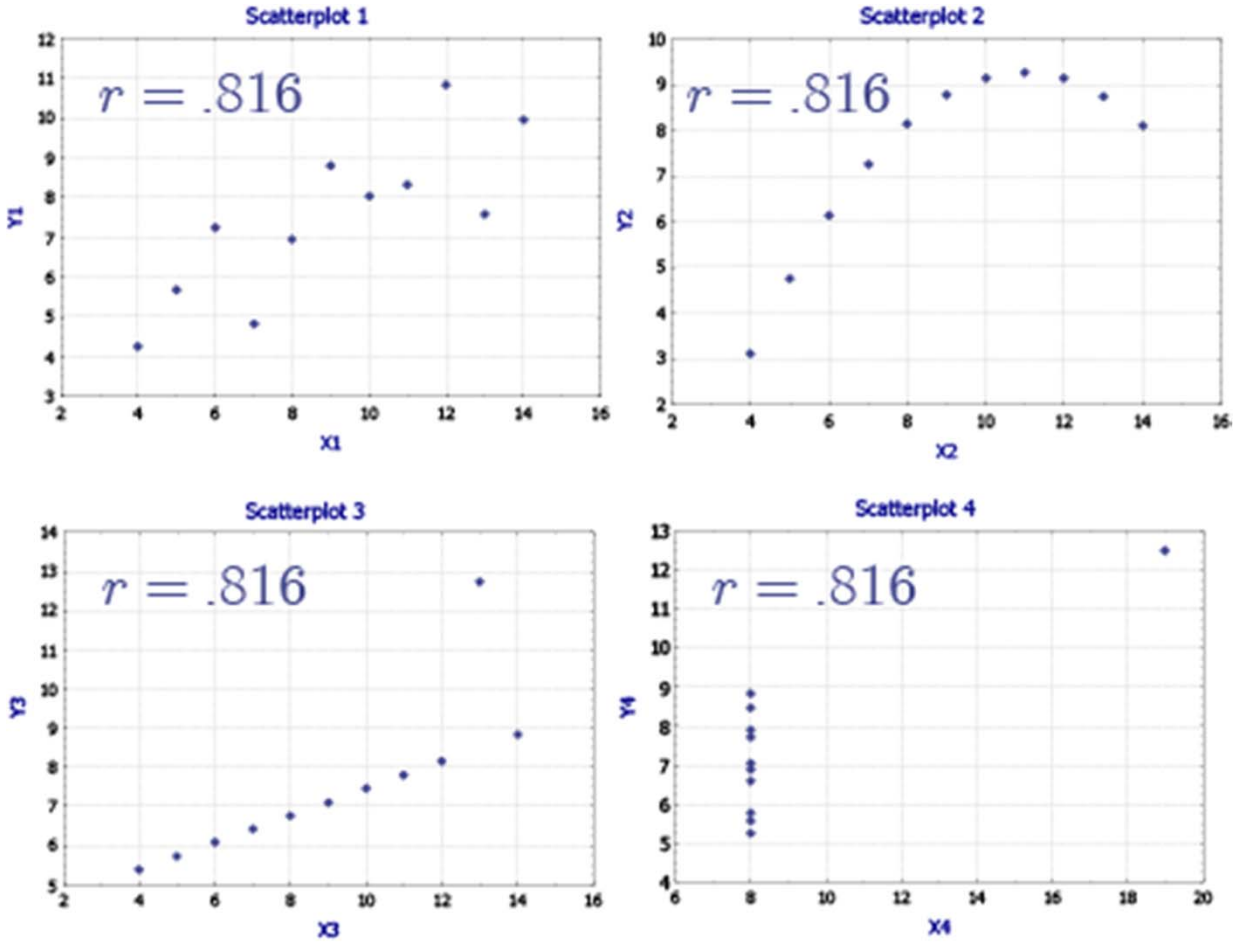
  $r < \pm 0.50 \rightarrow$ mittlere Korrelation

  $r \geq \pm 0.50 \rightarrow$ hohe Korrelation

- In der nicht-experimentellen Psychologie liegen Korrelationen selten über 0.75.

Bivariate Intervalldaten

Nichtlineare Zusammenhänge und die Korrelation



Bortz, S. 171 – 173

Bivariate Intervalldaten

Spezielle Koeffizienten – Punktbiseriale Korrelation

- Gegeben seien zwei Variablen X und Y. X sei dichotom nominalskaliert (mit zwei Ausprägungen 0 und 1), Y intervallskaliert.
- Hier kann wie auch bei zwei intervallskalierten Variablen die **Produkt-Moment-Korrelation** berechnet werden.
- Die Formel lässt sich aber auch zur Formel für die **punktbiseriale Korrelation** vereinfachen

Mittelwert der Y-Werte, für die X=1

Mittelwert der Y-Werte, für die X=0

$$r_{pbis} = \frac{\bar{y}_{X=1} - \bar{y}_{X=0}}{s_y} \cdot \frac{\sqrt{n_{X=0} \cdot n_{X=1}}}{n}$$

← Anzahl der Fälle, für die X=0 bzw. X=1

Bivariate Intervalldaten

Spezielle Koeffizienten – Biseriale Korrelation

- Häufig werden in psychologischen Untersuchungen eigentlich (mindestens) intervallskalierte Merkmale **künstlich auf dichotome Variablen reduziert**.
- Beispiele: Alter (unter 25, über 25), Einkommen (niedrig, hoch), Depression (nein, ja), versetzungsfähig (nein, ja)
- Hier führt die konkrete **Setzung des impliziten Kriteriums**, welches die intervallskalierte Variable in zwei Gruppen teilt, zu beliebigen Ergebnissen, obwohl der „wahre“ Zusammenhang unverändert ist.

Bivariate Intervalldaten

Spezielle Koeffizienten – Biseriale Korrelation

- Die Korrektur dieser kriteriumsabhängigen Veränderung des Zusammenhangs leistet die **biseriale Korrelation**:

$$r_{bis} = r_{pbis} \cdot \frac{\sqrt{n_{X=0} \cdot n_{X=1}}}{n \cdot \omega}$$

- Dabei ist ω die **Ordinate der Standardnormalverteilung** für den z-Wert an der Stelle der Dichotomisierung (p).
- r_{pbis} und r_{bis} Korrelation haben dieselben Eigenschaften wie der Produkt-Moment-Korrelationskoeffizient
- r_{pbis} ist zumeist vorzuziehen, da hier **keine Normalverteilungsannahme** gemacht werden muss

Bortz, S. 176 – 177

Bivariate Intervalldaten

Spezielle Koeffizienten – Tetrachorische Korrelation

- Sind beide Variablen künstlich dichotomisiert und eigentlich normalverteilt, so kann der Zusammenhang durch die tetrachorische Korrelation ausgedrückt werden.
- Ausgegangen wird zunächst von einer 2×2 Kontingenztabelle
- Daraus berechnet sich die **tetrachorische Korrelation** als:

$$r_{tet} = \cos\frac{\pi}{1+\sqrt{n_{11}\cdot n_{22} \big/ n_{12}\cdot n_{21}}}$$

	x_1	x_2	
y_1	n_{11}	n_{12}	$n_{1\bullet}$
y_2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

- r_{tet} **überschätzt** die wahre Korrelation, wenn die Randverteilungen stark asymmetrisch sind oder ein $n_{XY}<5$ ist.

Bivariate Ordinaldaten

Grundlagen

- Bei der Ordinalskala ist der numerische Abstand zwischen zwei Ausprägungen einer Variablen nicht interpretierbar.
- Die Ordinalskala trägt lediglich Information über die **Ordnung** der Ausprägungen.
- Damit sind mathematische Transformationen direkt auf den Werten einer ordinalskalierten Variablen nicht sinnvoll, also auch nicht die Produkt-Moment-Korrelation.
- **Ansatz:** Die Ordnung selbst muss genutzt werden, um Kennwerte zu berechnen.

Bortz, S. 178

Bivariate Ordinaldaten

Numerische Beschreibung - Rangbildung

- Bei der Rangbildung von k Ausprägungen $x_1 \ldots x_k$ einer Variablen X können maximal k Ränge vergeben werden.
- Per Konvention erhält die numerisch niedrigste Ausprägung von X den Rangplatz 1, die höchste den Rangplatz k (**kleinere Zahl = kleinerer Rang**).
- Bei gleichen mehreren gleichen Werten („**Ties**“) von X wird der mittlere Rangplatz vergeben nach der Regel:

  Es gebe m gleiche Werte von X. Wären sie unterschiedlich und direkt aufeinander folgend, erhielten sie die Rangplätze $rg_j \ldots rg_{j+m-1}$. Der **mittlere Rang** ist dann

$$rg_{Tie} = \frac{1}{m} \sum_{i=rg_j}^{rg_j+m-1} rg_i$$

Bivariate Ordinaldaten

Numerische Beschreibung – Spearman‘s r_s

- Nach der Rangbildung ordinalskalierter Daten für zwei Variablen X und Y kann die **Produkt-Moment-Korrelation der Ränge** $rg(X)$ und $rg(Y)$ berechnet werden
- Diese wird **Spearman‘s r_s** oder **Rangkorrelation** genannt und berechnet als

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

was nichts anderes ist als $r_s = \frac{s_{rg_x,rg_y}}{s_{rg_x} \cdot s_{rg_y}}$

Bivariate Ordinaldaten

Numerische Beschreibung – Spearman‘s r_s

- Wertebereich von –1 bis +1
- Vorzeichen gibt die Richtung des Zusammenhangs an
- Ist robust bezüglich Ausreißern
- Ist **invariant** bei streng monotonen Transformationen
- Liegen wenige Ties vor, gibt es vereinfachte näherungsweise Berechnungsformeln, die aber kaum mehr Anwendung finden.

Bivariate Ordinaldaten

Numerische Beschreibung – Weitere Kennwerte

- Neben Spearman's r_s existieren weitere Kennwerte für den Zusammenhang zweier ordinalskalierter Merkmale
- Die bekanntesten sind der **Konkordanzkoeffizient** γ („gamma“) nach Goodman-Kruskal und die daraus abgeleitete Weiterentwicklung **Kendall's** τ („tau“) für zwei ordinalskalierte Variablen
- Die Interpretation dieser Koeffizienten verläuft analog zu r und r_s

Bivariate Nominaldaten

Recap: Kontingenztabellen

- Wir haben Kontingenztabellen empirischer Verbundhäufigkeiten kennen gelernt.
- Schreibt man statt $h(x_i, y_j)$ kurz n_{ij}, so lautet die vereinfachte Notation für Kontingenztabellen:

	y_1	y_2	…	y_m	Σ
x_1	n_{11}	n_{12}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2m}	$n_{2\bullet}$
…	…	…	…	…	
x_k	n_{k1}	n_{k2}	…	n_{km}	$n_{k\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet m}$	$n_{\bullet\bullet}$

Zeilen x Spalten

- Analoge Notation für relative Häufigkeiten (mit f_{ij} statt n_{ij})

Bortz, S. 158

Bivariate Nominaldaten

Zusammenhangsmaße für 2×2 Kontingenztabellen

- Viele psychologische Fragestellungen über Zusammenhänge von Variablen beziehen sich auf 2 Merkmale mit je 2 Ausprägungen.
- **Beispiele:** Auftreten von Schizophrenie bei Frauen/Männern
- In solchen 2x2 Situationen kann jeder beiden Variablen durch zwei Werte abgebildet werden.

$$X = \begin{cases} x_1: 0, \text{ wenn Gesund} \\ x_2: 1, \text{ wenn Schizophrenie} \end{cases}$$

$$Y = \begin{cases} y_1: 1, \text{ wenn Frau} \\ y_2: 2, \text{ wenn Mann} \end{cases}$$

Bivariate Nominaldaten

Zusammenhangsmaße für 2×2 Kontingenztabellen

- Weil hier de facto eine Intervallskala erzwungen wird (genau ein Abstand zwischen Skalenwerten = konstanter Abstand zwischen Skalenwerten), kann immer die **Produkt-Moment-Korrelation** r als Zusammenhangsmaß berechnet werden
- **Idee:** Der 2x2 Fall bei Nominaldaten kann immer auf den ja/nein bzw. 0/1 Fall zurückgeführt werden
- Die Berechnungsformel für r vereinfacht sich dadurch erheblich

X	Y
0	1
0	1
1	2
0	2
1	2
1	2
0	1
…	…

Bivariate Nominaldaten

Zusammenhangsmaße für 2×2 Kontingenztabellen

- Der **Phi-Koeffizient (ϕ)** beschreibt die Stärke des Zusammenhangs zweier dichotomer Variablen

- Der ϕ-Koeffizient lässt sich nach folgender Formel berechnen:

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

	x_1	x_2	
y_1	n_{11}	n_{12}	$n_{1\bullet}$
y_2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

- ϕ liegt zwischen -1 und 1.

Bivariate Nominaldaten

Zusammenhangsmaße für 2×2 Kontingenztabellen

- Problem: Bei schiefen Randverteilungen kann der ϕ-Koeffizient selbst bei maximalem Zusammenhang zwischen den Variablen die Grenze ±1 **nicht erreichen**
- Bei schiefen Randverteilungen sollte ϕ daher an der maximal möglichen Korrelation normiert werden.
- Diese berechnet sich als

$$\phi_{\max} = \sqrt{\frac{\min(n_{1\bullet}, n_{\bullet 1})\min(n_{2\bullet}, n_{\bullet 2})}{\max(n_{1\bullet}, n_{\bullet 1})\max(n_{2\bullet}, n_{\bullet 2})}}$$

- Und damit gilt für den normierten ϕ-Koeffizienten

$$\phi_{norm} = \frac{\phi}{\phi_{\max}}$$

Bortz, S. 158 – 159

Bivariate Nominaldaten

Zusammenhangsmaße für k×m Kontingenztabellen

- **Ansatz:** Man vergleicht die beobachteten Kontingenztabelle mit einer fiktiven Kontingenztabelle, die entstanden wäre, hätte **kein Zusammenhang** zwischen den Variablen bestanden.
- Abweichungen der beobachteten von den erwarteten Häufigkeiten sind dann als **Abweichungen von der Unabhängigkeit** aufzufassen
- Zur Konstruktion der Indifferenztabelle rechnet man für absolute Häufigkeiten aus n Beobachtungen

$$\tilde{h}(x_i, y_j) = \frac{h(x_i, \bullet) \cdot h(\bullet, y_j)}{n}$$

(~ = „erwartet“)

Bivariate Nominaldaten

Zusammenhangsmaße für k×m Kontingenztabellen

- Die Indifferenztabelle konstruiert sich also durch

	y_1	y_2	…	y_m	Σ
x_1	$\tilde{n}_{11}$	$\tilde{n}_{12}$	…	$\tilde{n}_{1m}$	$n_{\bullet}$
x_2	$\tilde{n}_{21}$	$\tilde{n}_{22}$	…	$\tilde{n}_{2m}$	$n_{\bullet}$
…	…	…	…	…	
x_k	$\tilde{n}_{k1}$	$\tilde{n}_{k2}$	…	$\tilde{n}_{km}$	$n_{k\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet m}$	$n_{\bullet\bullet}$

- Mit $\tilde{h}(x_i, y_j) = \frac{h(x_i, \bullet) \cdot h(\bullet, y_j)}{n}$ bzw. $\tilde{n}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n_{\bullet\bullet}}$

Bivariate Nominaldaten

Zusammenhangsmaße für k×m Kontingenztabellen

- Aus den beobachteten und unter der Annahme keines Zusammenhangs (Indifferenz) erwarteten Häufigkeiten berechnet sich nun:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} \qquad \frac{(beob - erw)^2}{erw}$$

- χ^2 ist Null bei perfekter Unabhängigkeit, ansonsten größer Null

- χ^2 kann beliebig große Werte annehmen, abhängig von der Anzahl der Ausprägungen und der Beobachtungen

Bortz, S. 180

Bivariate Nominaldaten

Zusammenhangsmaße für k×m Tabellen – Cramérs V

- Um aus dem nicht normierten χ^2-Koeffizienten ein als Korrelationskoeffizient interpretierbares Maß zu berechnen, wird folgende Formel verwendet:

$$V = \sqrt{\frac{\chi^2}{n_{\bullet\bullet} \min(k-1, m-1)}}$$

- Cramérs V ist wie χ^2 Null bei perfekter Unabhängigkeit, ansonsten größer Null
- V schwankt zwischen 0 und 1

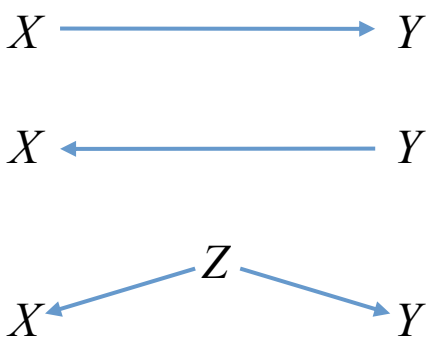
Bortz, S. 159 – 160

Zusammenhangsmaße

Interpretation von Korrelationen

- Eine vorhandene (hohe) Korrelation zwischen zwei Zufallsvariablen X und Y darf **nicht ohne weiteres als Kausalität** zwischen den Variablen interpretiert werden.

- Eine signifikante Korrelation zeigt zunächst nur eine **Assoziation** an. Diese kann viele Ursachen haben, z.B.



Assoziation (Korrelation) ist nicht Kausalität

Zusammenhangsmaße

Interpretation von Korrelationen

- **Frage:** Wann darf in einer psychologischen Untersuchung auf Kausalität geschlossen werden?

1. Die betrachteten Variablen müssen **kovariieren**
   → die Korrelation muss ungleich Null sein

Probleme:

– Standards („wann ist eine Korrelation ungleich Null“) sind normativ

– Je kleiner n, desto größere Korrelationen können per Zufall auftreten

Zusammenhangsmaße

Interpretation von Korrelationen

- **Frage:** Wann darf in einer psychologischen Untersuchung auf Kausalität geschlossen werden?

1. Die betrachteten Variablen müssen **kovariieren**
   → die Korrelation muss ungleich Null sein
2. Die Ursache muss der Wirkung **zeitlich vorausgehen** (z.B. Pretest – Treatment – Posttest)
3. Andere plausible Erklärungen für die Kovariation müssen ausgeschlossen werden können
4. Die Kovariation muss raum-zeitlich indifferent sein
   → Generalisierung auf eine Population zu jeder Zeit

Relevante Excel Funktionen

- Zusammenhangsmaße
  - KOVAR()
  - KORREL()
  - NORMINV()
  - COS(), PI()
  - RANG.MITTELW()