Univariate Kennwerte

Definition

„Die in einem Datensatz für ein Merkmal enthaltene Information lässt sich zu *Kenngrößen* verdichten.

Diese charakterisieren das **Zentrum** oder die Variabilität des Datensatzes. Man hat also Kenngrößen zur Beschreibung der **„mittleren“ Lage** der Elemente des Datensatzes und solche zur Charakterisierung der Streuung.“

Mittag, Hans-Joachim (2015): Statistik: Eine Einführung mit interaktiven Elementen. Berlin, Heidelberg: Springer. S. 103.

- Datenverdichtung, Reduktion von Komplexität
- Interpretationshilfen, Vergleichsgrößen
- Kommunikation der Dateneigenschaften
- Maße der zentralen Tendenz, Lagemaße: Typische Werte
- Streumaße: Heterogenität, Unterschiedlichkeit der Werte

Skalenniveau und univariate Kennwerte

	Messniveau	Eigenschaften	mögliche Aussage	Beispiele
non-metrisch	Nominalskala	klassifizierend	gleich, ungleich	Farben, Geschlecht
	Ordinalskala	Rangordnung, keine gleichen Abstände	größer, kleiner	Bewertung von Kinofilmen
metrisch	Intervallskala	gleiche Abstände	Gleichheit von Differenzen	Temperatur in Grad C
	Ratioskala	absoluter Nullpunkt	Gleichheit von Verhältnissen	TV-Nutzung in Min/Tag

Skalenniveau	Lagemaß
nominal	Modalwert
ordinal	Median, Modalwert
metrisch	Arithmetisches Mittel, Modalwert, Median

Modalwert (Modus)

mindestens Nominalskalenniveau

der Wert (Merkmalsausprägung), der innerhalb einer Datenmenge am häufigsten vorkommt

Median (Md, $\tilde{x}$)

mindestens Ordinalskalenniveau

der Wert (Merkmalsausprägung), der in der Mitte steht, wenn alle Beobachtungswerte x_i der Größe nach geordnet sind.

nicht von Extremwerten beeinflusst

Ungrade Fallzahl

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Gerade Fallzahl

$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}$$

Arithmetisches Mittel (AM, $\bar{x}$)

metrisches Skalenniveau

die Summe aller Werte, geteilt durch Anzahl der Fälle

„Gleichgewichtspunkt der Verteilung“

von Extremwerten beeinflusst

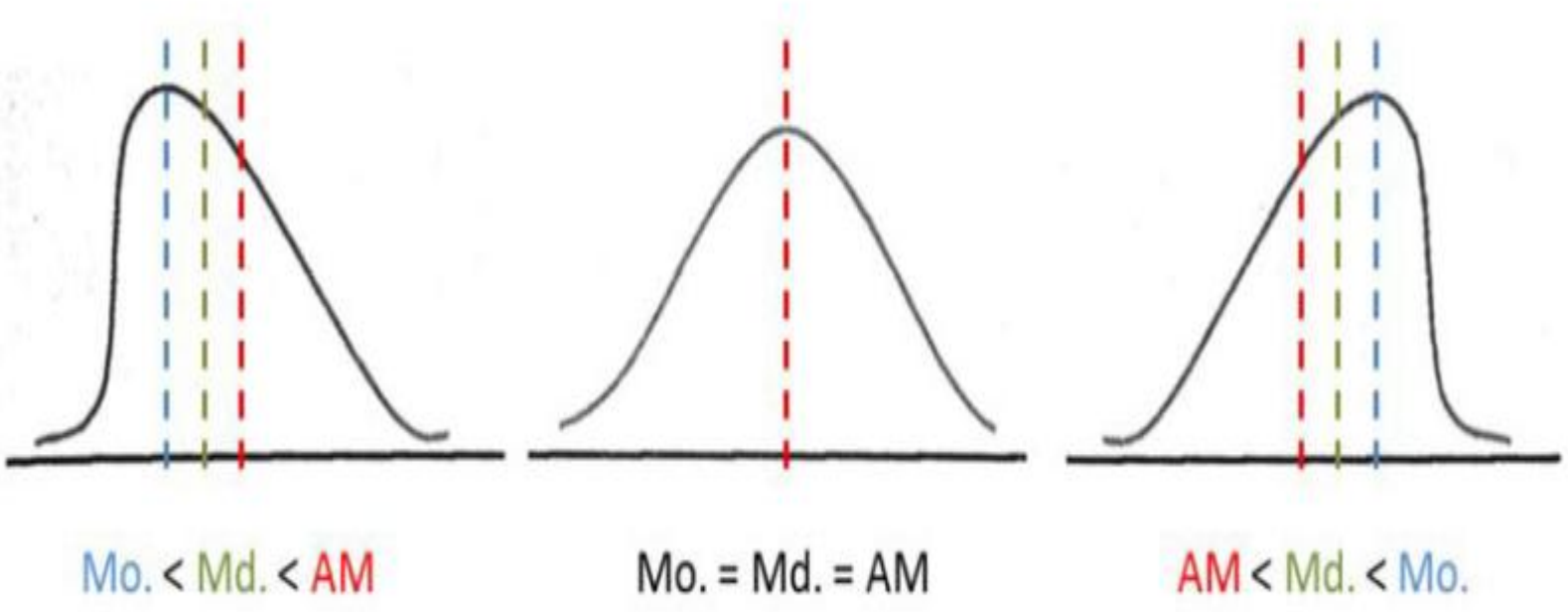
ohne Klassenbildung

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

mit Klassenbildung

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i * f_i$$

Lagemaße und Verteilung



linksschief

rechtsschief

Normalverteilung

Varianz (s^2)

Summe der quadrierten Abweichungen der Einzelfälle vom Arithmetischen Mittel

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

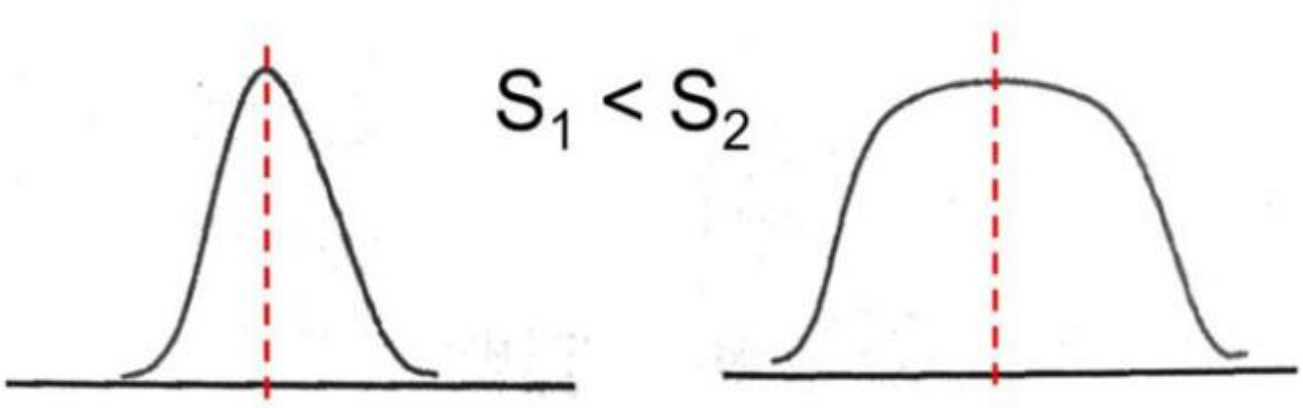
Standardabweichung (s)

Wurzel aus der Varianz

Aussagekraft nur im Vergleich

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

Standardabweichung und Verteilung



schmalgipflig

breitgipflig

Kleine Standardabweichung = homogene Verteilung

Große Standardabweichung = heterogene Verteilung

Diese Interpretation nur im Vergleich sinnvoll!

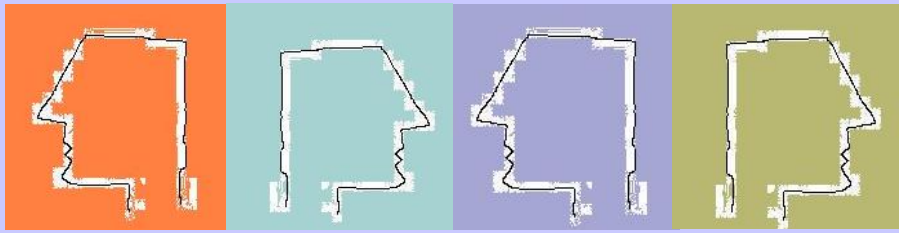
BIVARIATE STATISTIK

Zusammenhang zweier Messwerte / Variablen im Datensatz

DIFFERENZEN / UNTERSCHIEDE

ZUSAMMENHÄNGE

HYPOTHESEN-TESTS



Bivariate Häufigkeitsverteilung

Definition

„Hat man zwei diskrete Merkmale X und Y mit k bzw. m Ausprägungen, kann man die **absoluten oder relativen Häufigkeiten** für die k m Ausprägungskombinationen tabellarisch darstellen.

Diese auch als **Kontingenztafel** bezeichnete Tabelle definiert eine bivariate Häufigkeitsverteilung.

Ein Spezialfall der Kontingenztafel ist die **Vierfeldertafel**, bei der X und Y jeweils nur zwei Ausprägungen aufweisen.“

Mittag, Hans-Joachim (2015): Statistik: Eine Einführung mit interaktiven Elementen. Berlin, Heidelberg: Springer. S. 103.

Beispiel 1

Frage zur Präferenz von Filmen	Frauen (n=1.080)	Männer (n=1.090)	Gesamt (n=2.170)
Some like it Hot	48 %	8%	28 %
Der Sturm	3 %	7 %	5 %
Stirb langsam	10 %	40 %	25 %
Star Wars Episode IV	7 %	44 %	26 %
Anna Karenina	32 %	1 %	16 %
Gesamt	**100 %**	**100 %**	**100 %**

Konvention: Spalte = *unabhängige* (Einfluss-) Größe
Zeile = *abhängige* Größe

Beispiel 2

Frage zur Präferenz von Filmen	Frauen (n=1.080)	Männer (n=1.090)	Gesamt (n=2.170)
Some like it Hot	86 %	14%	100 %
Der Sturm	30 %	70 %	100 %
Stirb langsam	20 %	80 %	100 %
Star Wars Episode IV	14 %	88 %	100 %
Anna Karenina	97 %	3 %	100 %

→ Weniger Informationen als bei Spaltenprozentuierung

Bivariate Zusammenhangsmaße

Definition

Bivariate Zusammenhangsmaße beschreiben die **gemeinsame Verteilung** zweier Variablen. Sie lassen Aussagen über Zusammenhänge und Unterschiede zu.

Mit anderen Worten: sie sind Maße für die **Koinzidenz**zweier Merkmale.

Bivariate Zusammenhangsmaße gibt es für jedes Skalenniveau.

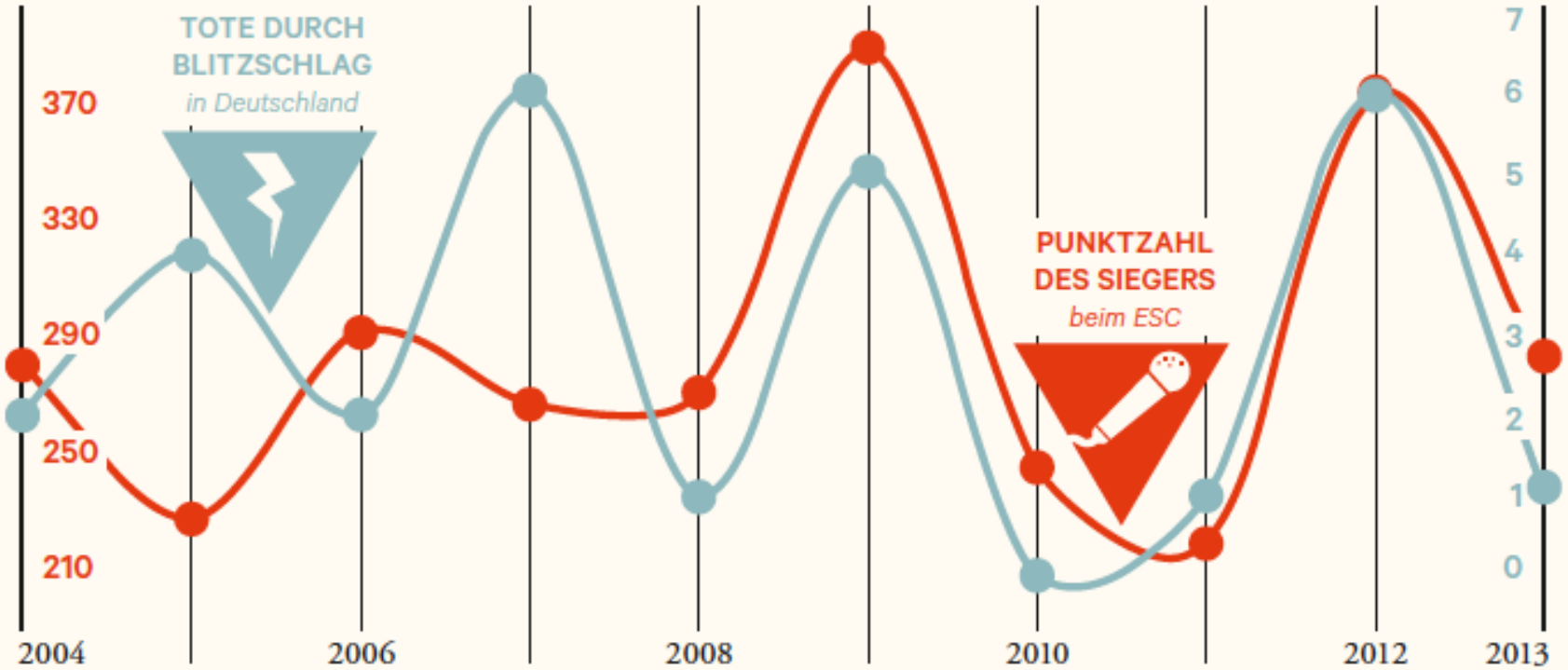
Skalenniveau	Beispiele	Zusammenhangsmaß	Aussage
nominal	Geschlecht, Parteipräferenz	**Chi²** **Cramer's V**	Zusammenhang Stärke
ordinal	Lieblingsfilme Person A und B	**Rangkorrelationskoeffizient Spearman's τ** (rho)	Übereinstimmung / Stärke
metrisch	Größe, Gewicht	**Kovarianz** **Korrelationskoeffizient**	Zusammenhang je-desto / Stärke

Zusammenhänge...

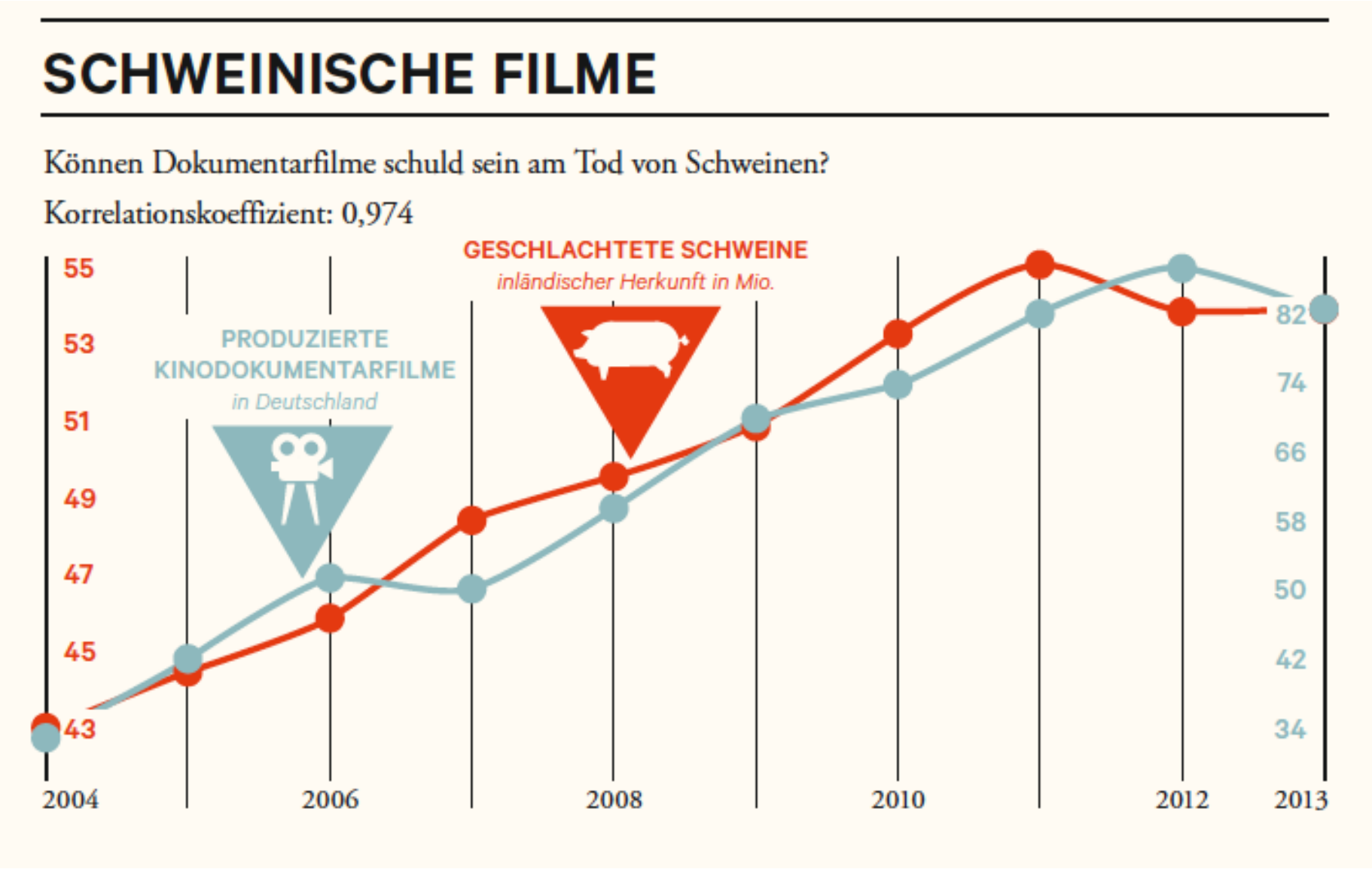
EINSCHLAGENDER ERFOLG

Was hat die Punktzahl des Siegers beim Eurovision Song Contest mit Toten durch Blitzschlag zu tun?

Korrelationskoeffizient: 0,571



Zusammenhänge...



Korrelation vs. Kausalität

Zusammenhangsmaß CHI² 1

Definition

Maßzahl für den Zusammenhang zweier nominalskalierter Variabeln.

Basis: Kreuz- bzw. **Kontingenztabelle**

Logik

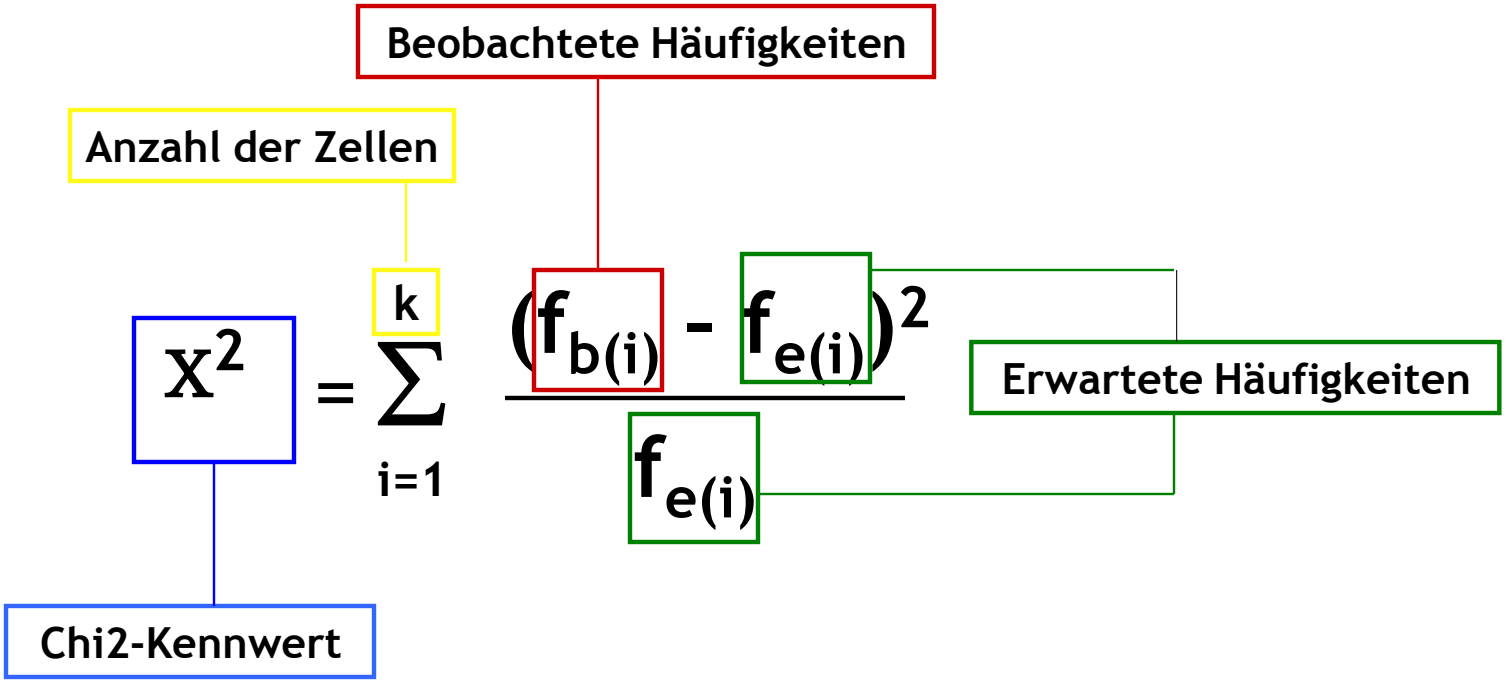
Berechnung einer zweiten sog. **Indifferenztabelle** unter der Annahme, dass **kein** Zusammenhang zwischen den Werten der beiden Variablen besteht.

Das Zusammenhangsmaß **Chi2** (χ^2) ist die **Summe der Werteabweichungen** zwischen empirischer Kontingenz- und berechneter Indifferenztabelle.

Chi2 (χ^2) hat einen Wertebereich von 0 (= kein Zusammenhang bis ∞ (= maximaler Zusammenhang).

Der Maximalwert ist abhängig von der Skalierung der Variablen, Tabellen unterschiedlicher Variablen lassen sich deshalb nicht ohne Weiteres vergleichen.

Zusammenhangsmaß CHI² 2



Vorgehensweise

Voraussetzungen

Kreuztabelle mit absoluten Zellen- und Randhäufigkeiten
nominalskalierte Variablen (u.U. auch ordinalskalierte)
Gesamtfallzahl mind. n = 60
alle Zellenwerte mind. n = 1
weniger als 20% aller Zellen mit einer Häufigkeit < n = 5

Schritte

(1) Kontingenztabelle erstellen
beobachtete Häufigkeiten (f_b) in absoluten Zahlen

(2) Indifferenztabelle berechnen
Erwartete Häufigkeiten (f_e) für alle Zellen berechnen

$$f_e = \frac{\textbf{Zeilen (n) * Spalten (n)}}{\textbf{Gesamt (n)}}$$

(3) Für jede Zelle die Abweichung zwischen Kontingenz- und Indifferenztabelle berechnen

$$\frac{(f_b - f_e)^2}{f_e}$$

(4) Aufsummieren zu **Chi2** (χ^2)

Standardisierungsmaße von CHI²

Warum?

Die Werte des Zusammenhangsmaßes Chi² hängen von der Anzahl n der Messwerte und der Grüße der Tabelle ab.

Chi2-Werte unterschiedlicher Tabellen können deshalb auch nicht miteinander verglichen werden.

Zur besseren Interpretation und Vergleichbarkeit stehen **standardisierte** Maße zur Verfügung:

> **Cramer's V** (für beliebige Kreuztabellen)
> **Kontingenzkoeffizient C** (für beliebige Kreuztabellen)
> **Phi** (für Vierfeldertabellen)

Wertebereiche:

0 (kein Zusammenhang) $\leq$ V/C/Phi $\leq$ 1 (perfekter Zusammenhang)

Stärke des Zusammenhangs, nicht Richtung!

Beispiel Cramer‘s V

Definition

Cramer‘s V ist ein **standardisiertes** Maß, das die **Stärke** des Zusammenhangs zweier **nominalskalierter** Variablen angibt.

$$V = \sqrt{\frac{x^2}{n * (R - 1)}}$$

$$0 \leq V \leq 1$$

x^2 = Chi²-Wert

i = Anzahl der Kategorien der Zeilenvariable

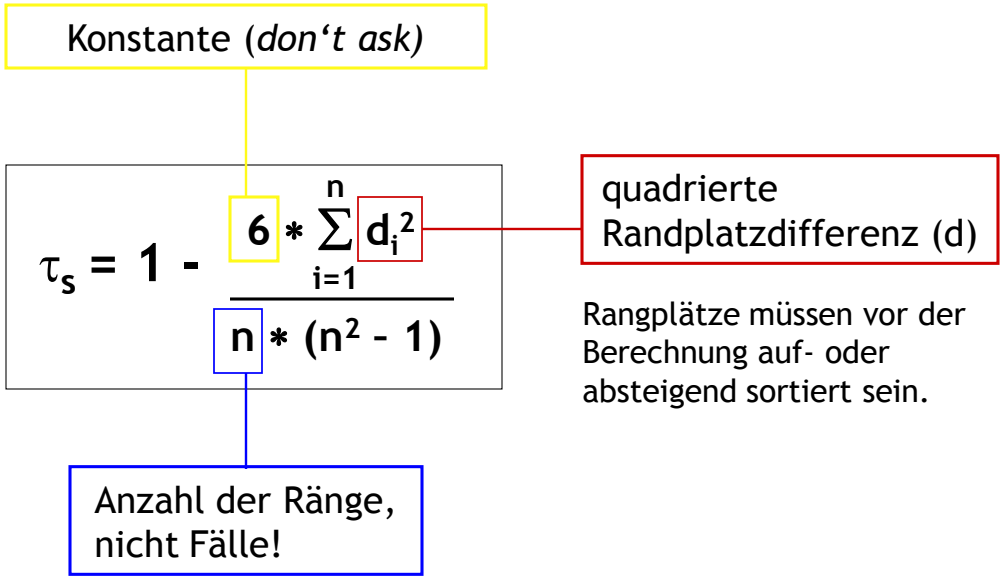
j = Anzahl der Spaltenvariable

R = min (i,j) -> ist die kleinere Zahl von beiden
(bei einer 3x4-Tabelle z.B. ist R = 3)

Rangkorrelationskoeffizient Spearman

Definition

Spearman‘s τ_s ist ein **standardisiertes** skalenunabhängiges Maß, das **Stärke** und **Richtung** des Zusammenhangs zweier mindestens **ordinalskalierter** Variablen angibt.



τ_s berücksichtigt die **Rangreihenfolge**, nicht deren Höhe, und ist dadurch robust gegenüber Ausreißern. es kann ab n > 5 berechnet werden.

Wertebereich

-1 (perfekter negativer Zusammenhang) $\leq \tau_s \leq 1$ (perfekter positiver Zusammenhang)

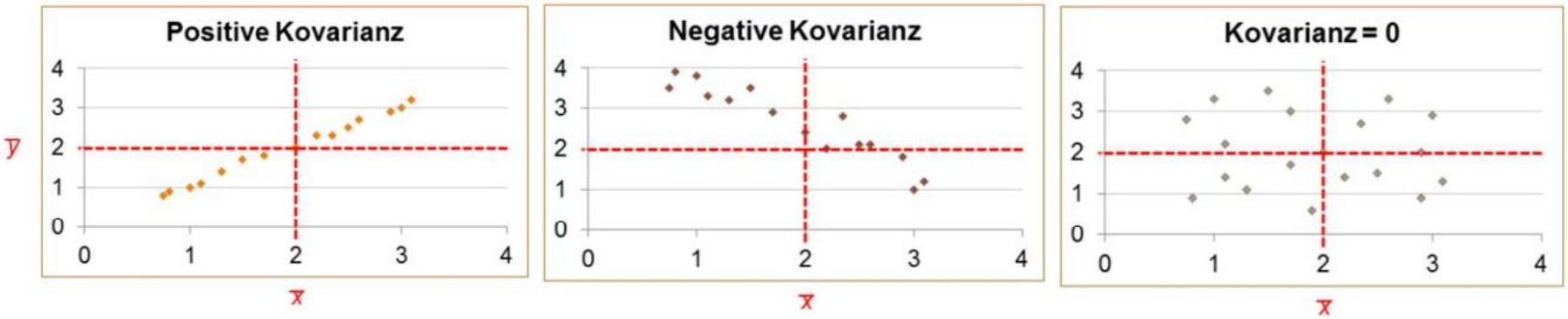
Bei $\tau_s = 0$ sind die Variablen unabhängig voneinander.

Definition

Die Kovarianz (cov_{xy}) ist ein **nicht-standardisiertes** Zusammenhangsmaß zur Beschreibung **linearer Zusammenhänge** zwischen zwei mindestens **metrisch** skalierten Variablen X und Y.

Die Kovarianz ist das durchschnittliche Abweichungsprodukt aller Messwertepaare von ihrem jeweiligen Mittelwert.

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



Korrelationskoeffizient Pearson‘s r

Definition

Pearson‘s r ist ein **standardisiertes** skalenunabhängiges Maß, das die **Stärke** und **Richtung** des **linearen** Zusammenhangs zweier **metrisch** skalierter Variablen angibt.

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

Kovarianz

Produkt der Standardabweichungen von X und Y

Wertebereich

-1 (perfekt negativer linearer Zusammenhang) $\leq r_{x,y} \leq 1$ (perfekt positiver linearer Zusammenhang)

Bei $r_{x,y}$ = besteht kein **linearer** Zusammenhang.

Pearson‘s r: ein bisschen Geformel

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

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

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

Überblick standardisierte bivariate Zusammenhangsmaße

Y- Variable → X-Variable ↓	Nominal	Ordinal	Metrisch
nominal	Cramer's V		
ordinal		Spearman's Rho	
metrisch			Pearson's r

Wertebereich: $0 \leq V \leq 1$ $-1 \leq r_s \leq 1$ $-1 \leq r_p \leq 1$