

BA KW | Vorlesung

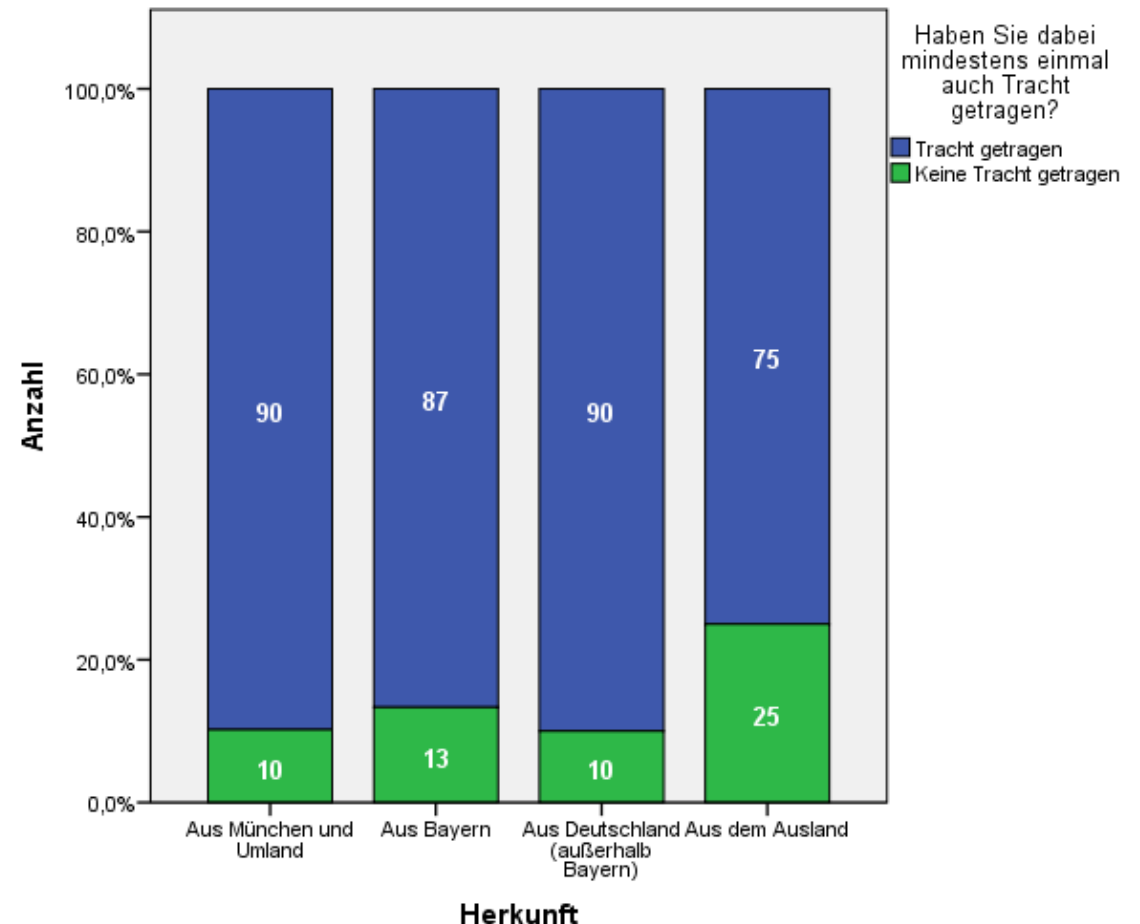
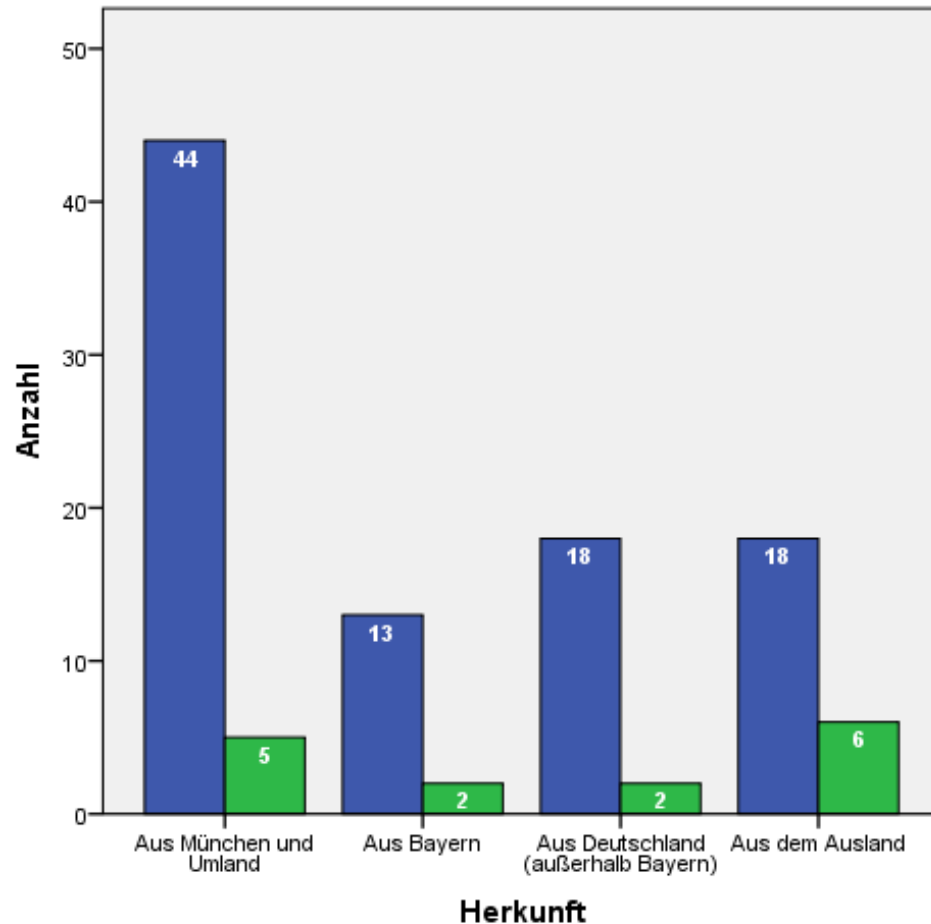
Einführung in die Statistik

Kontingenztabellen

Prof. Thomas Hanitzsch



Visualisierung: Gruppierte und gestapelte Balkendiagramme





Unbedingte und bedingte Wahrscheinlichkeiten

- **Unbedingte Wahrscheinlichkeit:**
 - Wie stehen die Chancen, dass eine Variable einen bestimmten Wert annimmt?
- **Bedingte Wahrscheinlichkeit:**
 - Wie stehen die Chancen, dass eine Variable einen bestimmten Wert annimmt, unter der Bedingung, dass das Eintreten eines anderen Wertes schon bekannt ist?
 - $$P(B|A) = \frac{P(B \cap A)}{P(A)}$$



Unabhängigkeit von Ereignissen

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

- Hängt Ereignis B vom Ausgang des Ereignisses A ab?
- Wenn die bedingten Wahrscheinlichkeiten für die verschiedenen Ausprägungen von Ereignis B gleich sind, sind die Ereignisse A und B vollständig voneinander unabhängig

Beispiel: Radiohören und Geschlecht

- Gegeben ist:
 - Ereignis A: Geschlecht der Person (männlich)
 - Ereignis B: Bewertung der Radiosendung (mögen)
 - $N=1000$
- Ergebnis:
 - Männer: $N=600$; davon mögen 300 die Sendung → $P(B|A)=0,5$
 - Frauen: $N=400$; davon mögen 200 die Sendung → $P(B|A^c)=0,5$
- Schlussfolgerung:
 - Die bedingten Wahrscheinlichkeiten unterscheiden sich nicht
→ die Ereignisse sind **voneinander unabhängig**

Beispiel: Radiohören und Geschlecht

- Gegeben ist:
 - Ereignis A: Geschlecht der Person (männlich)
 - Ereignis B: Bewertung der Radiosendung (mögen)
 - $N=1000$
- Ergebnis:
 - Männer: $N=600$; davon mögen 300 die Sendung → $P(B|A)=0,5$
 - Frauen: $N=400$; davon mögen 350 die Sendung → $P(B|A^c)=0,875$
- Schlussfolgerung:
 - Die bedingten Wahrscheinlichkeiten unterscheiden sich
→ die Ereignisse sind **voneinander abhängig**

Kontingenztafeln: Grundlagen

- „Kreuztafeln“ bilden die gemeinsamen Häufigkeitsverteilungen von zwei Variablen ab
- Kann sowohl absolute als auch relative Häufigkeiten abbilden
- Geeignet für nominal- und ordinalskalierte Daten, aber auch für gruppierte metrische Variablen (Daten müssen dann in Klassen zusammengefasst werden)
- Kreuztafeln besitzen k Spalten und l Zeilen, sodass sich $k \times l$ Zellen ergeben
- Konvention: zur Untersuchung von gerichteten Zusammenhängen wird die unabhängige Variable (X) als Spaltenvariable (d.h. „oben“) und die abhängige Variable (Y) als Zeilenvariable (d.h. „links“) dargestellt
→ dies ist aber unerheblich für die statistische Unabhängigkeitsprüfung

Kontingenztafeln: Grundlagen

| Y | X | | | | | |
|----------|----------------|----------|----------------|----------|----------------|-----------------|
| | A_1 | ... | A_i | ... | A_k | |
| B_1 | n_{11} | ... | n_{i1} | ... | n_{k1} | $n_{\bullet 1}$ |
| \vdots | \vdots | \ddots | \vdots | \ddots | \vdots | \vdots |
| B_j | n_{1j} | ... | n_{ij} | ... | n_{kj} | $n_{\bullet j}$ |
| \vdots | \vdots | \ddots | \vdots | \ddots | \vdots | \vdots |
| B_l | n_{1l} | ... | n_{il} | ... | n_{kl} | $n_{\bullet l}$ |
| | $n_{1\bullet}$ | ... | $n_{i\bullet}$ | ... | $n_{k\bullet}$ | n |

n_{ij} : Häufigkeit der Merkmalskombinationen

$n_{i\bullet}$: Häufigkeit von Merkmal A_i

$n_{\bullet j}$: Häufigkeit von Merkmal B_j

Kontingenztafeln: Grundlagen

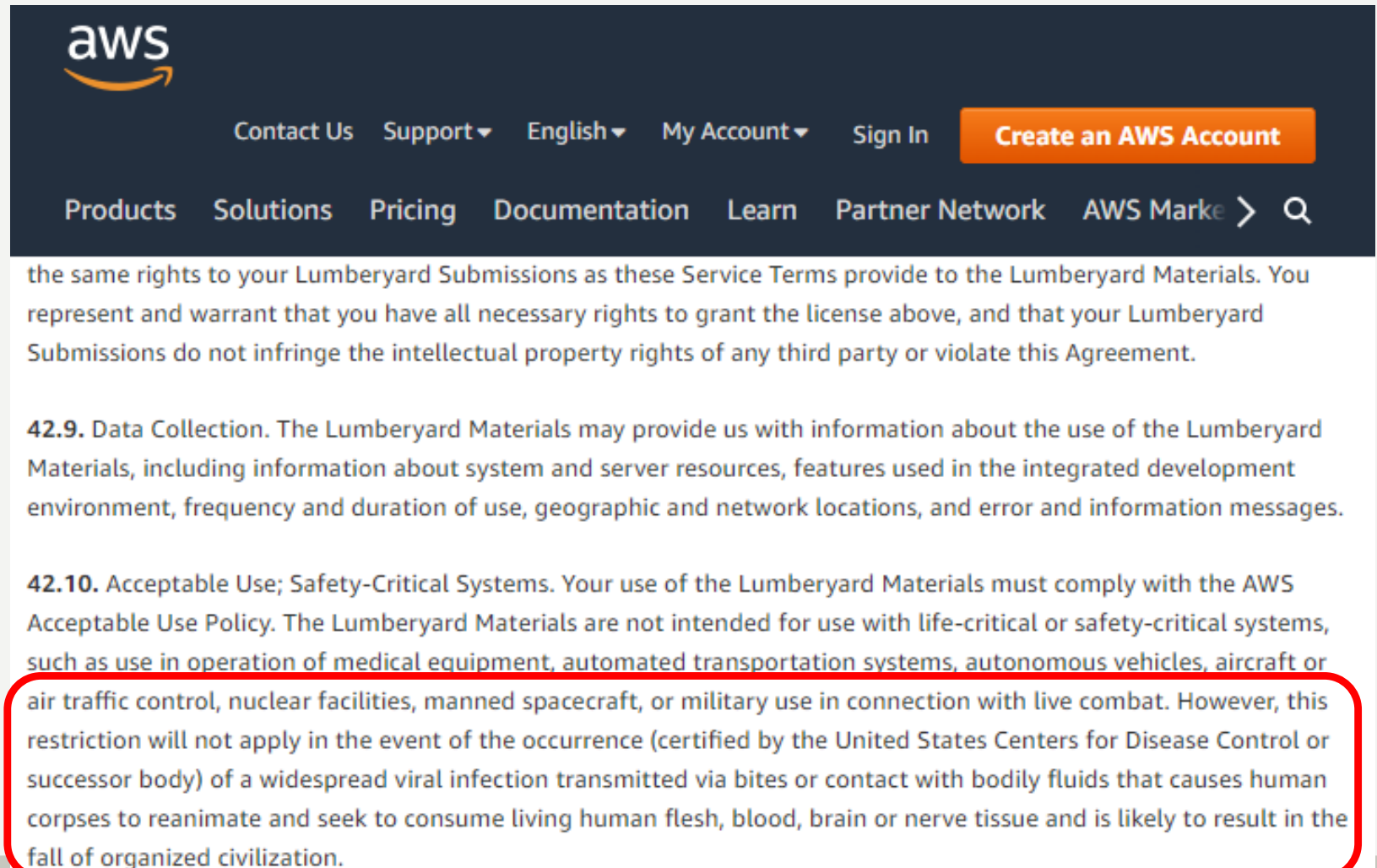
| Y | X | | | | | |
|----------|----------------|----------|----------------|----------|----------------|-----------------|
| | A_1 | ... | A_i | ... | A_k | |
| B_1 | n_{11} | ... | n_{i1} | ... | n_{k1} | $n_{\bullet 1}$ |
| \vdots | \vdots | \ddots | \vdots | \ddots | \vdots | \vdots |
| B_j | n_{1j} | ... | n_{ij} | ... | n_{kj} | $n_{\bullet j}$ |
| \vdots | \vdots | \ddots | \vdots | \ddots | \vdots | \vdots |
| B_l | n_{1l} | ... | n_{il} | ... | n_{kl} | $n_{\bullet l}$ |
| | $n_{1\bullet}$ | ... | $n_{i\bullet}$ | ... | $n_{k\bullet}$ | n |

$n_{i\bullet}$: Spaltensummen bzw. Randhäufigkeiten
 → eindimensionale Verteilung des Merkmals X

$n_{\bullet j}$: Zeilensummen bzw. Randhäufigkeiten
 → eindimensionale Verteilung des Merkmals Y

Kontingenztabellen: Schritt für Schritt

Beispiel:



The screenshot shows the AWS website header with the logo and navigation links: Contact Us, Support, English, My Account, Sign In, and a prominent orange 'Create an AWS Account' button. Below the header, a search bar contains the text 'Products Solutions Pricing Documentation Learn Partner Network AWS Marke' followed by a magnifying glass icon. The main content area displays a paragraph of text, with a red circle highlighting a specific section:

the same rights to your Lumberyard Submissions as these Service Terms provide to the Lumberyard Materials. You represent and warrant that you have all necessary rights to grant the license above, and that your Lumberyard Submissions do not infringe the intellectual property rights of any third party or violate this Agreement.

42.9. Data Collection. The Lumberyard Materials may provide us with information about the use of the Lumberyard Materials, including information about system and server resources, features used in the integrated development environment, frequency and duration of use, geographic and network locations, and error and information messages.

42.10. Acceptable Use; Safety-Critical Systems. Your use of the Lumberyard Materials must comply with the AWS Acceptable Use Policy. The Lumberyard Materials are not intended for use with life-critical or safety-critical systems, such as use in operation of medical equipment, automated transportation systems, autonomous vehicles, aircraft or air traffic control, nuclear facilities, manned spacecraft, or military use in connection with live combat. However, this restriction will not apply in the event of the occurrence (certified by the United States Centers for Disease Control or successor body) of a widespread viral infection transmitted via bites or contact with bodily fluids that causes human corpses to reanimate and seek to consume living human flesh, blood, brain or nerve tissue and is likely to result in the fall of organized civilization.

Kontingenztabellen: Schritt für Schritt

Beispiel:

- Studie „*Mediatized Zombification*“
 - Untersucht wird die Wirkung von „The Walking Dead“ (TWD) auf das Publikum
 - Hypothese: die Rezeption von mindestens einer Folge TWD trägt dazu bei, dass sich TV-Zuschauer in Zombies verwandeln
 - Gegeben ist:
 - $X \rightarrow A_1 = \text{TWD-Seher}; A_2 = \text{kein TWD-Seher}$
 - $Y \rightarrow B_1 = \text{Mensch}; B_2 = \text{Zombie}$
- Es ergibt sich eine 2×2 Kontingenztafel (Vier-Felder-Tafel)





Kontingenztabellen: Schritt für Schritt

- Schritt 1: Kreuztabelle mit absoluten Häufigkeiten

| | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|-----------|----------------|--------------|
| Mensch | 20 | 40 | 60 |
| Zombie | 30 | 10 | 40 |
| Spaltensummen | 50 | 50 | 100 |



Kontingenztabellen: Schritt für Schritt

- Schritt 2: Kreuztabelle mit absoluten und relativen Häufigkeiten

| | | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|---|-----------|----------------|--------------|
| Mensch | n | 20 | 40 | 60 |
| | % | 20,0% | 40,0% | 60,0% |
| Zombie | n | 30 | 10 | 40 |
| | % | 30,0% | 10,0% | 40,0% |
| Spaltensummen | n | 50 | 50 | 100 |
| | % | 50,0% | 50,0% | |



Kontingenztabellen: Schritt für Schritt

- Schritt 3: Kreuztabelle mit Spaltenprozenten

| | | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|---|-----------|----------------|--------------|
| Mensch | n | 20 | 40 | 60 |
| | % | 40,0% | 80,0% | 60,0% |
| Zombie | n | 30 | 10 | 40 |
| | % | 60,0% | 20,0% | 40,0% |
| Spaltensummen | n | 50 | 50 | 100 |
| | % | 100,0% | 100,0% | |

Die quadratische Kontingenz (χ^2)

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

n_{ij} = beobachtete Häufigkeit in der
 i -ten Spalte und j -ten Zeile

e_{ij} = erwartete Häufigkeit in der
 i -ten Spalte und j -ten Zeile

k = Anzahl der Spalten

l = Anzahl der Zeilen

- Bezeichnet Größe des Zusammenhangs (*nicht* Stärke, denn χ^2 ist abhängig von N) zwischen den kreuztabellierten Merkmalen
- Beruht auf Vergleich von beobachteten und erwarteten Häufigkeiten
- Wenn X und Y unabhängig sind, dann muss das Unabhängigkeitskriterium für jedes n_{ij} erfüllt sein
- Wertebereich: 0 bis unendlich; bei $\chi^2 = 0$ sind Ereignisse absolut unabhängig

Die quadratische Kontingenz (χ^2)

- Die Prüfgröße χ^2 ist bei ausreichend großen Stichprobengrößen annähernd χ^2 -verteilt mit $df = (k - 1) \cdot (l - 1)$ Freiheitsgraden
- Voraussetzung für die Berechnung von χ^2 ist, dass maximal 20% aller Zellen der Tabelle eine erwartete Häufigkeit kleiner als 5 aufweisen (sonst droht Verfälschung von χ^2 bei kleinen Erwartungswerten)
 - **Ausweg**: Zahl der Zellen (d.h. Kategorien) verringern



Kontingenztabellen: Schritt für Schritt

- Schritt 4: Kreuztabelle mit Erwartungswerten (Indifferenztabelle)

| | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|-----------|----------------|--------------|
| Mensch | 30 | 30 | 60 |
| Zombie | 20 | 20 | 40 |
| Spaltensummen | 50 | 50 | 100 |

Werte werden für jede Zelle berechnet:

$$e_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N} = \frac{\text{Spaltensumme } i \cdot \text{Zeilensumme } j}{N}$$



Kontingenztafeln: Schritt für Schritt

- Schritt 5: Berechnung von χ^2

| | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|--------------------------|--------------------------|--------------|
| Mensch | $\frac{(20 - 30)^2}{30}$ | $\frac{(40 - 30)^2}{30}$ | 60 |
| Zombie | $\frac{(30 - 20)^2}{20}$ | $\frac{(10 - 20)^2}{20}$ | 40 |
| Spaltensummen | 50 | 50 | 100 |

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{100}{30} + \frac{100}{30} + \frac{100}{20} + \frac{100}{20} = \mathbf{16,67}$$



Der χ^2 -Unabhängigkeitstest

- Testet die **Nullhypothese** $H_0: P_{ij} = P_{i\cdot} \cdot P_{\cdot j}$
- H_0 wird zum Signifikanzniveau α abgelehnt, falls χ^2 größer dem $(1-\alpha)$ -Quantil der χ^2 -Verteilung mit $df = (k - 1) \cdot (l - 1)$ Freiheitsgraden ist
- Voraussetzung:
 - Beobachtungen sind voneinander unabhängig
 - Maximal 20% aller Zellen der Tabelle haben eine erwartete Häufigkeit von <5



Der χ^2 -Unabhängigkeitstest: Beispiel

- Schritt 6: Abgleich von χ^2 mit $(1-\alpha)$ -Quantil der χ^2 -Verteilung

| | | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|-----|-----------|----------------|--------------|
| Mensch | n | 20 | 40 | 60 |
| | e | 30 | 30 | |
| Zombie | n | 30 | 10 | 40 |
| | e | 20 | 20 | |
| Spaltensummen | n | 50 | 50 | 100 |

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \mathbf{16,67}$$

$$df = (k - 1) \cdot (l - 1) = 1$$

$$\alpha = 5\%$$

→ 0,95-Quantil der χ^2 -Verteilung

→ kritischer Wert $\chi_{krit}^2 = \mathbf{3,84}$

$\chi^2 = 16,67 > \chi_{krit}^2 \rightarrow \mathbf{H_0}$ wird abgelehnt



Der χ^2 -Unabhängigkeitstest: Beispiel

- **Schritt 7: Berichten**
- Die Studie konnte belegen, dass Personen, die mindestens eine Folge von „The Walking Dead“ gesehen haben, sich häufiger in Zombies verwandelt haben, als jene, die die Serie nicht gesehen haben. Der Zusammenhang ist signifikant ($\chi^2 = 16,67$; $df = 1$; $p < 0,05$) und von mittlerer Stärke ($V = 0,41$).

Zusammenhangsmaße

- Phi-Koeffizient (ϕ)
- Cramers V

| Betrag des Zusammenhangskoeffizienten | Stärke des Zusammenhangs |
|---------------------------------------|--------------------------|
| $0,0 \leq \text{Koeffizient} < 0,1$ | kein Zusammenhang |
| $0,1 \leq \text{Koeffizient} < 0,3$ | geringer Zusammenhang |
| $0,3 \leq \text{Koeffizient} < 0,5$ | mittlerer Zusammenhang |
| $0,5 \leq \text{Koeffizient} < 0,7$ | hoher Zusammenhang |
| $0,7 \leq \text{Koeffizient} < 1,0$ | sehr hoher Zusammenhang |



Zusammenhangsmaße

Phi-Koeffizient (ϕ ; 0 bis +1)

- Zusammenhangsmaß für zwei dichotome Variablen
 - Er eignet sich also nur für die **Vier-Felder-Tafel**
 - Wertebereich: 0 (kein Zusammenhang) ... +1 (maximaler Zusammenhang)
- Berechnung:

- aus χ^2 : $\phi = \sqrt{\frac{\chi^2}{N}}$

→ **Beispiel:** $\phi = \sqrt{\frac{16,67}{100}} = \mathbf{0,41}$

Zusammenhangsmaße

Cramers V (0 bis +1)

- Zusammenhangsmaß für zwei nominalskalierte Variablen
 - Variablen können beliebig viele Ausprägungen (d.h. Zeilen und Spalten in Kontingenztabelle) haben
 - Wertebereich: 0 (kein Zusammenhang) ... +1 (maximaler Zusammenhang)
- Berechnung:

$$V = \sqrt{\frac{\chi^2}{N \cdot (R-1)}} \quad \text{mit} \quad R = \min(k, l)$$

k = Anzahl der Kategorien der Spaltenvariablen

l = Anzahl der Kategorien der Zeilenvariablen



Zusammenhangsmaße

Cramers V (0 bis +1) → Beispiel

| | TWD-Seher | kein TWD-Seher | Zeilensummen |
|---------------|-----------|----------------|--------------|
| Mensch | 20 | 40 | 60 |
| Zombie | 30 | 10 | 40 |
| Spaltensummen | 50 | 50 | 100 |

$$R = \min(k, l) = \min(2, 2) = 2$$

$$V = \sqrt{\frac{\chi^2}{N \cdot (R - 1)}} = \sqrt{\frac{16,67}{100 \cdot (2 - 1)}} = \sqrt{0,167} = 0,41$$



Kontingenztabellen berichten

- Notwendige Informationen:
 - Tabellenbeschriftung und -Nummerierung
 - Erklärung der dargestellten Werte (absolute Häufigkeiten, Prozentwerte etc.)
 - In die Analyse eingegangene Fallzahl
 - ggf. zusätzliche Anmerkungen (z.B. Rundungsfehler, Beschreibung der Skala)
- Von SPSS generierte Tabellen sehen meist unschön aus
- Der Text sollte auf jede berichtete Tabelle rekurrieren
- Tabellen sollten auf den ersten Blick verständlich sein



Kreuztabellen: Aufgabe 1

- Kreuztabelle ja oder nein?
 - Zusammenhang zwischen Geschlecht und Fernsehnutzungsdauer in Minuten
 - Zusammenhang zwischen Bildung und Besitz eines Fernsehgeräts (ja/nein)
 - Zusammenhang zwischen Parteipräferenz und Alter
 - Zusammenhang zwischen Geschlecht und Augenfarbe

Kreuztabellen: Aufgabe 2

- In einer Studie zur Mediennutzung ($N = 59$) wurde neben dem Fernsehtyp auch die Senderpräferenz (öffentlich-rechtliche vs. private Sender) der Befragten erhoben. Es ergibt sich folgende Kontingenztabelle:

| | Fernsehtyp | | |
|-----------------------------|-------------------|---------------------------|------------------|
| Senderpräferenz | <i>Wenigseher</i> | <i>Durchschnittsseher</i> | <i>Vielseher</i> |
| <i>Öffentlich-rechtlich</i> | 16 | 10 | 3 |
| <i>privat</i> | 5 | 9 | 16 |

- Ergänzen Sie die Randverteilungen der beiden Merkmale.
- Besteht ein Zusammenhang zwischen den beiden Merkmalen?
- Wie stark ist dieser Zusammenhang?