LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Datenanalyse

Sitzung 10: Kontingenztab

Institut für Kommunikationswissenschaft und Medienforschung
Ludwig-Maximilians-Universität München

Ablauf der Sitzung

1. Wiederholung: Kontingenz- bzw. Kreuztabellen & Chi²-Koeffizient
   a) Übung 1: Wann wenden wir die Kreuztabelle an?
   b) Übung 2: Manuelle Berechnung (Hausaufgabe)
2. Kreuztabelle in R berechnen
   a) Übung 3: Berechnung der Übung 2 in R
   b) Übung 4: Kontingenztabelle in R am Beispiel WoJ

Kurze Wiederholung:
Kontingenz- bzw. Kreuztabellen & Chi^2-Koeffizient

Wozu gibt es Kreuztabellen? Wie sind sie aufgebaut?

- Darstellung des Zusammenhangs zwischen kategorialen Variablen
- *Grundsätzlich gilt:* Unabhängige (gruppierende) Variable in die Spalten, abhängige (Ziel-)Variable in die Zeilen

Tabelle X: Beispiel für eine gelungene Kreuztabelle.

		UV		
		Gruppe 1 (n = ?)	Gruppe 2 (n = ?)	Gruppe 3 (n = ?)
AV	Ausprägung 1			
	Ausprägung 2			

Wichtige Anmerkungen gehören in die Fußnote (z.B. Besonderheiten der Tabelle und statistische Kennzahlen).

Quadratische Kontingenz (Chi²- bzw. χ^2-Koeffizient)

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

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

- Basiert auf einem Vergleich von beobachteten und bei Unabhängigkeit theoretisch zu erwartenden Häufigkeiten
- Beschreibt, ob ein **Zusammenhang** zwischen den beiden (kategorialen) Variablen vorliegt – nicht aber die Stärke des Zusammenhangs!
- Wertebereich: 0 bis unendlich (∞); bei $\chi^2 = 0$ sind die beiden Merkmale vollständig unabhängig

Voraussetzungen für den χ^2-Test

- Die Beobachtungen sind voneinander unabhängig
- Maximal 20% aller Zellen der Tabelle weisen eine erwartete Häufigkeit kleiner als 5 auf (sonst droht eine Verfälschung von χ^2)
  - Ausweg: Zahl der Zellen (d. h. Kategorien) verringern, Ausprägungen zusammenfügen
- Keine der erwarteten Häufigkeiten ist Null

ÜBUNG 1 (HAUSAUFGABE)

Übung 1

- Entscheiden Sie, ob das Erstellen einer Kreuztabelle in den nachfolgenden Fällen sinnvoll ist.

  - 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

Lösung Übung 1

- Entscheiden Sie, ob das Erstellen einer Kreuztabelle in den nachfolgenden Fällen sinnvoll ist.

  - Zusammenhang zwischen Geschlecht und Fernsehnutzungsdauer in Minuten

Nein – Fernsehnutzungsdauer in Minuten ist metrisch skaliert; nur dann sinnvoll, wenn die Variable Fernsehnutzungsdauer im Vorfeld gruppiert wird.

Lösung Übung 1

- Entscheiden Sie, ob das Erstellen einer Kreuztabelle in den nachfolgenden Fällen sinnvoll ist.

  - Zusammenhang zwischen Bildung und Besitz eines Fernsehgeräts (ja/nein)

Ja – da sowohl Bildung als auch Besitz eines Fernsehgerätes kategoriale Variablen sind.

Lösung Übung 1

- Entscheiden Sie, ob das Erstellen einer Kreuztabelle in den nachfolgenden Fällen sinnvoll ist.

  - Zusammenhang zwischen Parteipräferenz und Alter

Nein – das Alter ist metrisch skaliert; nur dann sinnvoll, wenn die Variable Alter im Vorfeld gruppiert wird.

Lösung Übung 1

- Entscheiden Sie, ob das Erstellen einer Kreuztabelle in den nachfolgenden Fällen sinnvoll ist.

  - Zusammenhang zwischen Geschlecht und Augenfarbe

Ja – da sowohl Geschlecht als auch Augenfarbe kategoriale Variablen sind.

ÜBUNG 2 (HAUSAUFGABE)

Übung 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:
  a. Ergänzen Sie die Randhäufigkeiten der beiden Merkmale.
  b. Besteht ein Zusammenhang zwischen den beiden Merkmalen in der Grundgesamtheit (α = 5%)?
  c. Wie stark ist dieser Zusammenhang?

	Fernsehtyp		
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher
Öffentlich-rechtlich	16	10	3
Privat	5	9	16

Lösung Übung 2a

- 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:
  a. Ergänzen Sie die Randhäufigkeiten der beiden Merkmale.

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	16	10	3	29
Privat	5	9	16	30
	21	19	19	59

Lösung Übung 2b

- 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:
  b. Besteht ein Zusammenhang zwischen den beiden Merkmalen in der Grundgesamtheit (α = 5%)?

=> Frage nach Zusammenhang in der Grundgesamtheit mit Signifikanzniveau bedeutet: Signifikanztest!

Um die Frage nach dem Zusammenhang beantworten zu können, müssen wir einen **Chi²-Test** durchführen.

Lösung Übung 2b

- **Schritt 1:** Erwartungswert e_{ij} für jede Zelle berechnen: $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	16 $\frac{21 \times 29}{59} = 10{,}32$	10	3	**29**
Privat	5	9	16	30
	21	19	19	**59**

Lösung Übung 2b

- **Schritt 1:** Erwartungswert e_{ij} für jede Zelle berechnen: $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	16 $\frac{21 \times 29}{59} = 10{,}32$	10 $\frac{19 \times 29}{59} = 9{,}34$	3	**29**
Privat	5	9	16	30
	21	**19**	19	**59**

Lösung Übung 2b

- **Schritt 1:** Erwartungswert e_{ij} für jede Zelle berechnen: $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	16 $\frac{21 \times 29}{59} = 10{,}32$	10 $\frac{19 \times 29}{59} = 9{,}34$	3 $\frac{19 \times 29}{59} = 9{,}34$	29
Privat	5 $\frac{21 \times 30}{59} = 10{,}68$	9 $\frac{19 \times 30}{59} = 9{,}66$	16 $\frac{19 \times 30}{59} = 9{,}66$	30
	21	19	19	59

Lösung Übung 2b

- **Schritt 2:** Chi²-Koeffizient berechnen

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

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	n_{11}=16 e_{11}=10,32 $\frac{(16 - 10{,}32)^2}{10{,}32} =$ **= 3,13**	n_{21}=10 e_{21}=9,34	n_{31}=3 e_{31}=9,34	29
Privat	n_{12}=5 e_{12}=10,68	n_{22}=9 e_{22}=9,66	n_{32}=16 e_{32}=9,66	30
	21	19	19	59

Lösung Übung 2b

- **Schritt 2:** Chi²-Koeffizient berechnen

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

	Fernsehtyp			
Senderpräferenz	Wenigseher	Durchschnitts-seher	Vielseher	
Öffentlich-rechtlich	n_{11}=16 e_{11}=10,32 $5{,}68^2$ / 10,32 = **3,13**	n_{21}=10 e_{21}=9,34 $0{,}66^2$ / 9,34 = **0,05**	n_{31}=3 e_{31}=9,34 $(-6{,}34)^2$ / 9,34 = **4,30**	29
Privat	n_{12}=5 e_{12}=10,68 $(-5{,}68)^2$ / 10,68 = **3,02**	n_{22}=9 e_{22}=9,66 $(-0{,}66)^2$ / 9,66 = **0,05**	n_{32}=16 e_{32}=9,66 $6{,}34^2$ / 9,66 = **4,16**	30
	21	19	19	59

Lösung Übung 2b

- **Schritt 2:** Chi²-Koeffizient berechnen

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

Nun werden die soeben berechneten Werte aufsummiert:

$$\chi^2 = 3{,}13 + 0{,}05 + 4{,}30 + 3{,}02 + 0{,}05 + 4{,}16 = \mathbf{14{,}71}$$

Lösung Übung 2b

- **Schritt 3:** Abgleich mit dem Quantil der Chi²-Verteilung

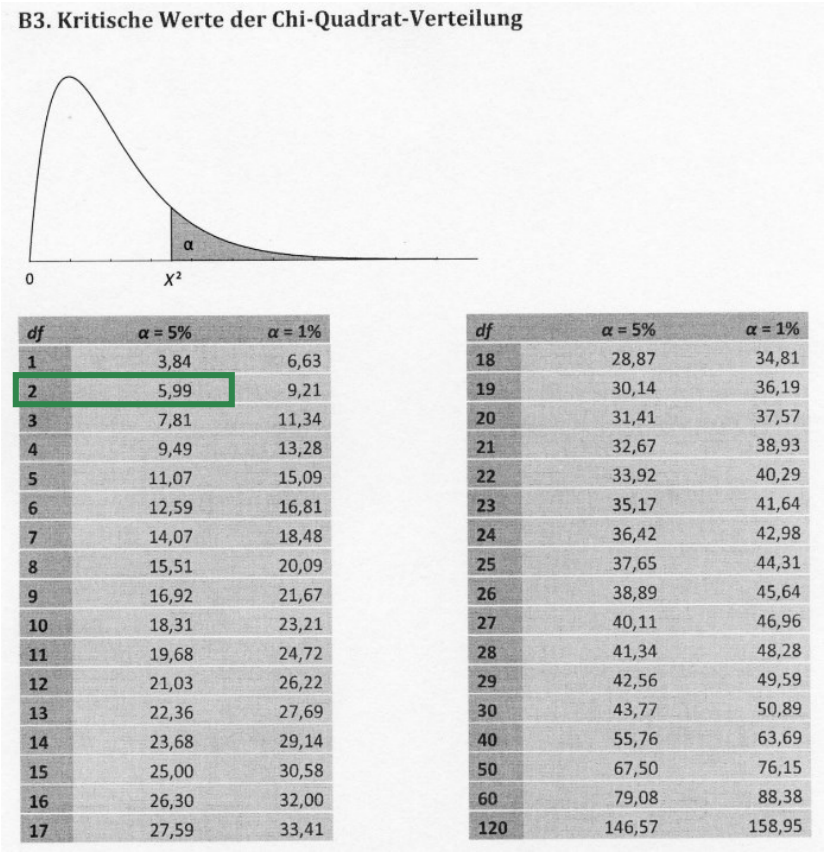
$\chi^2 = 14{,}71$

Anzahl der Spalten ↓ k — Anzahl der Zeilen ↓ l

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

$= (3-1) \cdot (2-1)$

$= 2$ →



B3. Kritische Werte der Chi-Quadrat-Verteilung

df	α = 5%	α = 1%
1	3,84	6,63
2	5,99	9,21
3	7,81	11,34
4	9,49	13,28
5	11,07	15,09
6	12,59	16,81
7	14,07	18,48
8	15,51	20,09
9	16,92	21,67
10	18,31	23,21
11	19,68	24,72
12	21,03	26,22
13	22,36	27,69
14	23,68	29,14
15	25,00	30,58
16	26,30	32,00
17	27,59	33,41

df	α = 5%	α = 1%
18	28,87	34,81
19	30,14	36,19
20	31,41	37,57
21	32,67	38,93
22	33,92	40,29
23	35,17	41,64
24	36,42	42,98
25	37,65	44,31
26	38,89	45,64
27	40,11	46,96
28	41,34	48,28
29	42,56	49,59
30	43,77	50,89
40	55,76	63,69
50	67,50	76,15
60	79,08	88,38
120	146,57	158,95

$\chi^2 = 14{,}71 > 5{,}99\ (\chi^2_{krit})$

→ H_0 (Unabhängigkeit) wird abgelehnt

Lösung Übung 2b

- **Schritt 4:** Beantwortung der Fragestellung

Der Test hat gezeigt, dass die beiden Merkmale „Fernsehtyp“ und „Senderpräferenz“ statistisch signifikant zusammenhängen ($\chi^2 = 14{,}71;\ df = 2; p < 0{,}05$). Vielseher scheinen eher private Sender zu schauen, Wenigseher eher öffentlich-rechtliche Sender.

Lösung Übung 2c

- 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:
  c. Wie stark ist dieser Zusammenhang?

Um die Frage nach der *Stärke* des Zusammenhangs beantworten zu können, müssen wir **Cramers V** berechnen.

→ **Phi** ist nur für dichotome Merkmale (2x2 Kreuztabelle/Vierfeldertafel) sinnvoll

Lösung Übung 2c

- 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:
  c. Wie stark ist dieser Zusammenhang?

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

$$R = \min(k, l) = \min(3,2) = 2$$

$$V = \sqrt{\frac{14{,}71}{59 \cdot (2-1)}} = 0{,}50$$

Zwischen den Merkmalen „Fernsehtyp“ und „Senderpräferenz“ besteht gemäß Konventionen ein Zusammenhang von **hoher Stärke**.

Wichtige Take-Aways

- Der χ^2-Test prüft, ob ein **Zusammenhang** zwischen **zwei kategorialen** Variablen vorliegt – nicht aber die Stärke des Zusammenhangs
  - Wertebereich: 0 bis unendlich (∞)
  - bei $\chi^2 = 0$ sind die beiden Merkmale vollständig unabhängig
- Stärke des Zusammenhangs im Anschluss mittels Cramer‘s V

R: Test auf Zusammenhang zwischen zwei kategorialen Variablen

Kreuztabellen berechnen mittels `tidycomm`

Sie können Korrelationen mit dem `tidycomm` Package berechnen. Dafür können Sie u. a. die folgenden Funktionen nutzen:

- `crosstab()` erstellt die Kreuztabelle
- mit Argument `add_total = TRUE` für die Zeilensumme
- mit Argument `percentages = TRUE` für die Angaben in Prozenten (statt absoluter Werte)
- mit Argument `chi_square = TRUE` für den Signifikanztest mit Cramer‘s V

Struktur:

```
data %>%
   crosstab(variable1, variable2)


data %>%
   crosstab(variable1, variable2,
            add_total = TRUE,
            percentages = TRUE,
            chi_square = TRUE)
```

Übung 3

Übung 3: Umsetzung der Übung 2 in R

Nun berechnen wir die Übung 2 einmal mit Hilfe von R. Dafür legen wir als Erstes den Datensatz in R an (einfach ausführen):

```r
# Übung 3: Anlegen des Datensatzes
senderpraeferenz <- c("oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl",
                      "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl",
                      "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl",
                      "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl",
                      "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl", "oeff_rechtl",
                      "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat",
                      "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat",
                      "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat", "privat",
                      "privat", "privat", "privat"
                      )
fernsehtyp <- c("wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher",
                "wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher",
                "wenigseher", "wenigseher", "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt",
                "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt",
                "vielseher", "vielseher", "vielseher", "wenigseher", "wenigseher", "wenigseher", "wenigseher",
                "wenigseher", "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt",
                "durchschnitt", "durchschnitt", "durchschnitt", "durchschnitt", "vielseher", "vielseher", "vielseher",
                "vielseher", "vielseher", "vielseher", "vielseher", "vielseher", "vielseher", "vielseher",
                "vielseher", "vielseher", "vielseher", "vielseher", "vielseher", "vielseher"
                )
befragten_id <- c(1:59)
data <- cbind(befragten_id, senderpraeferenz, fernsehtyp)
data <- tidyr::as_tibble(data)
```

Übung 3: Umsetzung der Übung 2 in R

Dann lassen wir uns deskriptiv eine Kreuztabelle der beiden Variablen ausgeben:

```
30
31  #Kreuztabelle ausgeben lassen
32  data %>% crosstab(fernsehtyp, senderpraeferenz)
33
```

Ausgabe:

```
> #Kreuztabelle ausgeben lassen
> data %>% crosstab(fernsehtyp, senderpraeferenz)
# A tibble: 2 × 4
  senderpraeferenz durchschnitt vielseher wenigseher
* <chr>                   <dbl>     <dbl>       <dbl>
1 oeff_rechtl                10         3          16
2 privat                      9        16           5
>
```

Übung 3: Umsetzung der Übung 2 in R

Jetzt möchten wir noch den Signifikanztest sowie das Cramer‘s V berechnen. Zusätzlich lassen wir uns die Werte in Prozente angeben (optional) und die Zeilensummen hinzufügen (optional):

```
34  #Kreuztabelle mit Signifikanztest
35  data %>% crosstab(fernsehtyp, senderpraeferenz,
36                    add_total = TRUE,
37                    percentages = TRUE,
38                    chi_square = TRUE)
```

Ausgabe:

```
# A tibble: 2 × 5
  senderpraeferenz durchschnitt vielseher wenigseher Total
* <chr>                   <dbl>     <dbl>      <dbl> <dbl>
1 oeff_rechtl             0.526     0.158      0.762 0.492
2 privat                  0.474     0.842      0.238 0.508
# Chi-square = 14.697, df = 2, p < 0.001, V = 0.499
```

Unsere Berechnung in R zeigt uns dementsprechend die gleichen Ergebnisse (die minimalen Unterschiede ergeben sich durch die Rundungen bei der manuellen Berechnung)

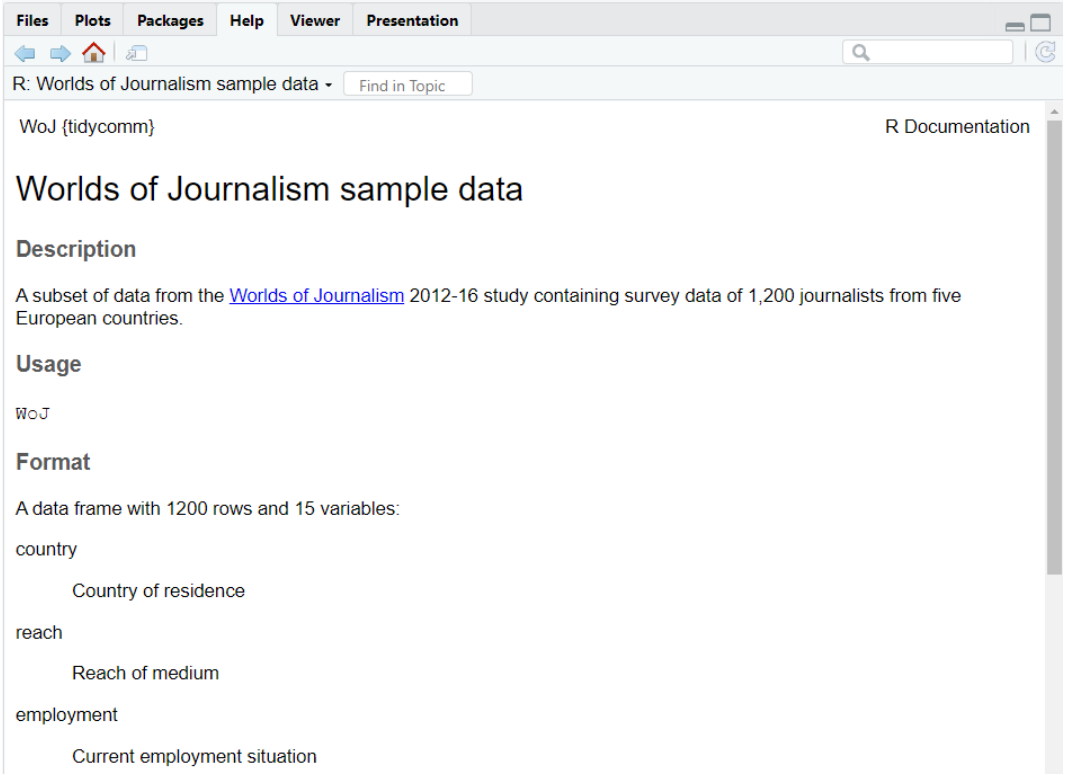
Übung 4

Übung 4

Laden Sie nun erneut den Datensatz „WoJ“ vom tidycomm-Package in Ihr RStudio-Environment.

```
41  # Übung 4: WoJ einlesen
42  WoJ <- WoJ
```

Erinnerung: Hilfeseite zum WoJ-Datensatz (`?WoJ`)



Files | Plots | Packages | Help | Viewer | Presentation

R: Worlds of Journalism sample data — Find in Topic

WoJ {tidycomm} — R Documentation

Worlds of Journalism sample data

Description

A subset of data from the Worlds of Journalism 2012-16 study containing survey data of 1,200 journalists from five European countries.

Usage

```
WoJ
```

Format

A data frame with 1200 rows and 15 variables:

country
: Country of residence

reach
: Reach of medium

employment
: Current employment situation

Übung 4

Journalist*innen im WoJ-Datensatz wurden u.a. gefragt, …

- in welchem Land sie wohnen («country») und
- welche Anstellungssituation sie haben («employment»)

Mara interessiert sich in ihrer Forschung vor allem für die Situation der Journalist*innen in Dänemark, Großbritannien und der Schweiz. Sie möchte sich nur diese drei Länder anschauen und vermutet, dass die Stellensituation der Journalist*innen in diesen drei Ländern signifikant mit ihrem Land zusammen hängt.

a) Filtern Sie den Datensatz nach den drei gefragten Ländern («Denmark», «UK», «Switzerland»)*

b) Was lässt sich zu Maras Überlegung sagen: Lässt sich ein signifikanter Zusammenhang zwischen dem Land und der Anstellung der Journalist*innen bei diesen drei Ländern feststellen? (Signifikanzniveau 5%)

*Tipp: Die Filter-Funktion haben wir in Sitzung 2 kennengelernt

Lösung Übung 4

Schritt 1: Filterung des Datensatzes

Variante 1:

```
44  #Filterung nach Land: Variante 1
45  data <- WoJ %>%
46    filter(country == "Denmark" | country == "UK" | country == "Switzerland")
```

Variante 2:

```
48  #Filterung nach Land: Variante 2
49  data <- WoJ %>%
50    filter(country %in% c("Denmark","UK","Switzerland"))
51
```

Lösung Übung 4

Schritt 2: Kreuztabelle mit absoluten Werten

Befehl:

```
52  #Kreuztabelle ausgeben lassen
53  data %>% crosstab(country, employment)
54
```

Ausgabe:

```
# A tibble: 3 × 4
  employment Denmark Switzerland    UK
* <chr>        <dbl>       <dbl> <dbl>
1 Freelancer      85          10    32
2 Full-time      275         154   169
3 Part-time       16          69    10
```

Lösung Übung 4

Schritt 3: Signifikanztest

Befehl:

```
55  #Kreuztabelle mit Signifikanztest
56  data %>% crosstab(country, employment,
57                    add_total = TRUE,
58                    percentages = TRUE,
59                    chi_square = TRUE)
60
```

Ausgabe:

```
# A tibble: 3 × 5
  employment Denmark Switzerland     UK Total
* <chr>        <dbl>       <dbl>  <dbl> <dbl>
1 Freelancer  0.226       0.0429 0.152  0.155
2 Full-time   0.731       0.661  0.801  0.729
3 Part-time   0.0426      0.296  0.0474 0.116
# Chi-square = 125.495, df = 4, p < 0.001, V = 0.277
```

Maras Vermutung kann bestätigt werden: Der Test hat gezeigt, dass die beiden Merkmale „country“ und „employment“ statistisch signifikant zusammenhängen ($\chi^2 = 125{,}50;\ df = 4; p < 0{,}05$). Der Zusammenhang hat allerdings nur eine geringe Stärke.

Wichtige Take-Aways

- **Kreuztabelle und χ^2-Test in R**: berechnen mit
  - `crosstab(variable1, variable2)` für die deskriptive Kreuztabelle
  - mit dem Argument `chi_square = TRUE` wird der χ^2-Test automatisch mitberechnet, mitsamt Cramer‘s V
  - Optional lassen sich folgende Argumente anfügen:
    - `add_total = TRUE` für die Zeilensummen
    - `percentages = TRUE` für die Angabe der Zellenwerte als Prozent