

TUTORIUM

Datenauswertung

ZUSAMMENHANG/UNTERSCHIED ZWEIER KATEGORIALER VARIABLEN

AGENDA

- Zwischenfazit
- Kreuztabellen
- Chi-Quadrat-Test
- Cramer's V
- Mosaikplots

HALBZEIT – ZWISCHENFAZIT

Was habe ich bisher gut gemacht?

Woran sollte ich noch arbeiten?

Was wünscht ihr euch für zukünftige Sitzungen?

ORGANISATORISCHES – FORSCHUNGSBERICHT

- Meine Aufgabe zu Schätzen und Testen kommt nicht in den Forschungsbericht
- Alle sonstigen bisher von mir am Ende der Folien gestellten Aufgaben kommen in den Forschungsbericht
- Alle in den folgenden Sitzungen gestellten Aufgaben von mir kommen in den Forschungsbericht
- Es sind nur diejenigen Aufgaben im Forschungsbericht zu behandeln, zu denen „(Script und) Text einfügen“ steht

KREUZTABELLEN I

Kreuztabellen

- Stellen die Merkmalskombinationen zweier kategorialer Variablen dar
- Können auf zwei Weisen benutzt werden:
 - Deskriptiv: Darstellung des Zusammenhangs zweier kategorialer Variablen in der Stichprobe
 - Inferenziell: Schluss auf den Zusammenhang zweier kategorialer Variablen in der Grundgesamtheit → Hypothesentests

KREUZTABELLEN II

Kreuztabellen

- Können auf verschiedene Arten dargestellt werden:
 - Die Prozentwerte werden zeilenweise berechnet
 - Die Prozentwerte werden spaltenweise berechnet (üblich!)
- Werden bei Kausalhypothesen so erstellt, dass
 - Die unabhängige Variable spaltenweise notiert wird
 - Die abhängige Variable zeilenweise notiert wird

KREUZTABELLEN – SCHRITT FÜR SCHRITT

1. Hypothese aufstellen.
2. Kreuztabelle für absolute Häufigkeiten der Merkmalskombinationen erstellen.
3. Kreuztabelle für relative Häufigkeiten (Spaltenprozent) der Merkmalskombinationen erstellen.
4. Kreuztabelle für die erwarteten Häufigkeiten bei statistischer Unabhängigkeit erstellen. deskriptiv

5. Chi-Quadrat-Test anwenden. inferenziell
6. Cramer's V berechnen (falls Chi-Quadrat-Test signifikant)

KREUZTABELLEN – HYPOTHESEN

Hypothesen

H_1 : Männer schauen mehr Fernsehen als Frauen.

H_0 : Frauen schauen gleich viel – wenn nicht sogar mehr – Fernsehen als Männer.

Variablen

UV: Geschlecht – männlich, weiblich

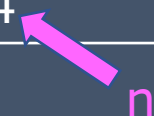
AV: Fernsehkonsum – niedrig, gewöhnlich, hoch

KREUZTABELLEN – ABSOLUTE HÄUFIGKEITEN

Geschlecht	Mann	Mann	Frau	Mann	Frau	Frau	Mann
Fernsehkonsum	niedrig	hoch	niedrig	niedrig	hoch	gewöhnlich	hoch
Geschlecht	Mann	Frau	Frau	Mann	Frau	Frau	Frau
Fernsehkonsum	niedrig	hoch	gewöhnlich	niedrig	hoch	niedrig	niedrig

Geschlecht

		Mann	Frau	Gesamt
Fernsehkonsum	niedrig	4	3	4+3=7
	gewöhnlich	0	2	0+2=2
	hoch	2	3	2+3=5
	Gesamt	4+0+2=6	3+2+3=8	14



KREUZTABELLEN – RELATIVE HÄUFIGKEITEN

		Geschlecht	
		Mann	Frau
Fernsehkonzum	niedrig	4	3
	gewöhnlich	0	2
	hoch	2	3
	Gesamt	6	8

→ absolute Häufigkeiten

		Geschlecht	
		Mann	Frau
Fernsehkonzum	niedrig	$4/6 = 0,66$	$3/8 = 0,375$
	gewöhnlich	$0/6 = 0$	$2/8 = 0,25$
	hoch	$2/6 = 0,33$	$3/8 = 0,375$
	Gesamt	$0,66 + 0 + 0,33 = 1$	$2 \cdot 0,375 + 0,275 = 1$

→ relative Häufigkeiten

KREUZTABELLEN – ERWARTETE HÄUFIGKEITEN

Geschlecht

	Mann	Frau	Gesamt
niedrig	4	3	7
gewöhnlich	0	2	2
hoch	2	3	5
Gesamt	6	8	14

Fernsehkonsum

→ absolute Häufigkeiten

Geschlecht

	Mann	Frau	Gesamt
niedrig	$6 \cdot \frac{7}{14} = 3$	$8 \cdot \frac{7}{14} = 4$	7
gewöhnlich	$6 \cdot \frac{2}{14} = 0,86$	$8 \cdot \frac{2}{14} = 1,14$	2
hoch	$6 \cdot \frac{5}{14} = 2,14$	$8 \cdot \frac{5}{14} = 2,86$	5
Gesamt	6	8	14

Fernsehkonsum

→ erwartete Häufigkeiten

KREUZTABELLEN – ZWISCHENSTAND

	Mann	Frau
niedrig	4	3
gewöhnlich	0	2
hoch	2	3

→ Häufigkeitsverteilung aus der Stichprobe

	Mann	Frau
niedrig	3	4
gewöhnlich	0,86	1,14
hoch	2,14	2,86

→ Häufigkeitsverteilung bei statistischer Unabhängigkeit

KREUZTABELLEN – AUFGABE

Aufgabe: Stellt eine ungerichtete Forschungshypothese zu den unten angegebenen Daten auf und erstellt Kreuztabellen für die absoluten und relativen Häufigkeiten (spaltenweise) in der Stichprobe.

Abitur Eltern	Ja	Nein	Nein	Ja	Ja	Ja	Nein	Nein	Nein	Ja
Abitur Kind	Ja	Ja	Nein	Ja	Ja	Ja	Nein	Ja	Nein	Nein

KREUZTABELLEN – LÖSUNG: HYPOTHESEN

Hypothesen

H_1 : Es gibt einen Zusammenhang zwischen dem **Abitur der Eltern** und dem **Abitur der Kinder**.

H_0 : Es gibt keinen Zusammenhang zwischen dem **Abitur der Eltern** und dem **Abitur der Kinder**.

Variablen

Abitur der Eltern – ja, nein

Abitur der Kinder – ja, nein

KREUZTABELLEN – LÖSUNG: ABSOLUTE HÄUFIGKEITEN

Abitur Eltern	Ja	Nein	Nein	Ja	Ja	Ja	Nein	Nein	Nein	Ja
Abitur Kind	Ja	Ja	Nein	Ja	Ja	Ja	Nein	Ja	Nein	Nein

	Eltern haben Abitur	Eltern haben Abitur nicht	Gesamt
Kinder haben Abitur	4	2	4+2=6
Kinder haben Abitur nicht	1	3	1+3=4
Gesamt	4+1=5	2+3=5	10

KREUZTABELLEN – LÖSUNG: RELATIVE HÄUFIGKEITEN

	Eltern haben Abitur	Eltern haben Abitur nicht	Gesamt
Kinder haben Abitur	4	2	6
Kinder haben Abitur nicht	1	3	4
Gesamt	5	5	10

	Eltern haben Abitur	Eltern haben Abitur nicht
Kinder haben Abitur	$\frac{4}{5} = 0,8$	$\frac{2}{5} = 0,4$
Kinder haben Abitur nicht	$\frac{1}{5} = 0,2$	$\frac{3}{5} = 0,6$

INSTALLATION: TIGERSTATS

```
Console Terminal x
~/
> install.packages("tigerstats")
Installing package into 'C:/Users/vitus/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/tigerstats_0.3.zip'
Content type 'application/zip' length 947614 bytes (925 KB)
downloaded 925 KB

package 'tigerstats' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\vitus\AppData\Local\Temp\RtmpIpuCXj\downloaded_packages
> library(tigerstats)
```

INSTALLATION: LSR

```
Console Terminal x
~/
> install.packages("lsr")
Installing package into 'C:/Users/vitus/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/lsr_0.5.zip'
Content type 'application/zip' length 215351 bytes (210 KB)
downloaded 210 KB

package 'lsr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\vitus\AppData\Local\Temp\RtmpIpuCXj\downloaded_packages
> library(lsr)
```

KREUZTABELLEN (ABSOLUTE HÄUFIGKEITEN) IN R

#Laden und Auswahlen der Daten

```
load("daten_x")  
datensatz <- subset(daten_x, bedingung)  
library(knitr)
```

#Nichtgenannte Ausprägungen löschen

```
datensatz <- within(datensatz, {UV <- droplevels(UV)})  
datensatz <- within(datensatz, {AV <- droplevels(AV)})
```

#Erstellen der Kreuztabelle

```
table1 = addmargins(table(datensatz$AV, datensatz$UV))  
kable(round(table1,1))
```

KREUZTABELLEN (ABSOLUTE HÄUFIGKEITEN) IN R – BEISPIEL

#Laden und Auswahlen der Daten

```
load("daten_2019.RData")  
datensatz <- subset(daten_2019, jahr>2000)  
library(knitr)
```

#Nichtgenannte Ausprägungen löschen

```
datensatz <- within(datensatz, {altersgruppe <- droplevels(altersgruppe)})  
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
```

#Erstellen der Kreuztabelle

```
table1 = addmargins(table(datensatz$gem_radiop, datensatz$altersgruppe))  
kable(round(table1,1))
```

	15-24 Jahre	25-34 Jahre	35-44 Jahre	45-54 Jahre	55-64 Jahre	65-74 Jahre	Sum
ja	8	14	23	37	38	36	156
nein	41	50	41	55	36	23	246
Sum	49	64	64	92	74	59	402

KREUZTABELLEN (RELATIVE HÄUFIGKEITEN) IN R

#Laden und Auswahlen der Daten

```
load("daten_x")  
datensatz <- subset(daten_x, bedingung)  
library(knitr)
```

#Nichtgenannte Ausprägungen löschen

```
datensatz <- within(datensatz, {UV <- droplevels(UV)})  
datensatz <- within(datensatz, {AV <- droplevels(AV)})
```

#Erstellen der Kreuzbabelle

```
table1 = xtabs(~ AV + UV, data = datensatz)  
table2 = colPerc(table1)  
kable(round(table2,1))
```

KREUZTABELLEN (SPALTENPROZENT) IN R - BEISPIEL

#Laden und Auswahlen der Daten

```
load("daten_2019.RData")  
datensatz <- subset(daten_2019, jahr > 2000)  
library(knitr); library(tigerstats)
```

#Nichtgenannte Ausprägungen löschen

```
datensatz <- within(datensatz, {altersgruppe <- droplevels(altersgruppe)})  
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
```

#Erstellen der Kreuzbabelle

```
table1 = xtabs(~ gem_radiop + altersgruppe, data = datensatz)  
table2 = colPerc(table1)  
kable(round(table2,1))
```

	15-24 Jahre	25-34 Jahre	35-44 Jahre	45-54 Jahre	55-64 Jahre	65-74 Jahre
ja	16.3	21.9	35.9	40.2	51.4	61
nein	83.7	78.1	64.1	59.8	48.6	39
Total	100.0	100.0	100.0	100.0	100.0	100

DER CHI-QUADRAT-TEST

Der Chi-Quadrat-Test

- Ist ein Test für Hypothesen mit zwei kategorialen Variablen
- Basiert auf Kreuztabellen (absolute + erwartete Anzahl)
- Ist erst sinnvoll ab einer Stichprobengröße von $n > 100$
- Geht davon aus, dass die Variablen nicht zusammenhängen → statistische Unabhängigkeit
- Prüft, ob zwei Variablen zusammenhängen/sich unterscheiden, aber nicht, wie stark der Zusammenhang/Unterschied ist

DER EMPIRISCHE CHI-QUADRAT-WERT

Der empirische Chi-Quadrat-Wert (χ^2_{emp})

- Berechnet die quadrierten, durch den erwarteten Wert standardisierten Abstände zwischen gemessenem und erwartetem Wert → vgl. Varianz
- Formel:

Formal

$$\chi^2_{emp} = \sum_{i=1}^n \frac{(fo_i - fe_i)^2}{fe_i}$$

Nicht-formal

$$\text{empirischer Chi - Quadratwert} = \sum_{i=1}^n \frac{(\text{gemessener Wert}_i - \text{erwarteter Wert}_i)^2}{\text{erwarteter Wert}_i}$$

DER KRITISCHE CHI-QUADRAT-WERT

Der kritische Chi-Quadrat-Wert (χ^2_{crit})

- Ist der Wert, bis zu dem die Nullhypothese mit einer Irrtumswahrscheinlichkeit von α nicht verworfen wird
- Ist abhängig von zwei Angaben:
 - Irrtumswahrscheinlichkeit (normalerweise $\alpha=0,05$)
 - Freiheitsgrade: $df = (Spalten - 1) * (Zeilen - 1)$
- Kann von Chi-Quadrat-Tabellen abgelesen werden
- Wird so interpretiert:
 - $\chi^2_{emp} > \chi^2_{crit} \rightarrow$ signifikanter Zusammenhang
 - $\chi^2_{emp} < \chi^2_{crit} \rightarrow$ nicht-signifikanter Zusammenhang

DER EMPIRISCHE CHI-QUADRAT-WERT – BEISPIEL

Gemessene Werte:

	Mann	Frau
niedrig	4	3
gewöhnlich	0	2
hoch	2	3

Erwartete Werte:

	Mann	Frau
niedrig	3	4
gewöhnlich	0,86	1,14
hoch	2,14	2,86

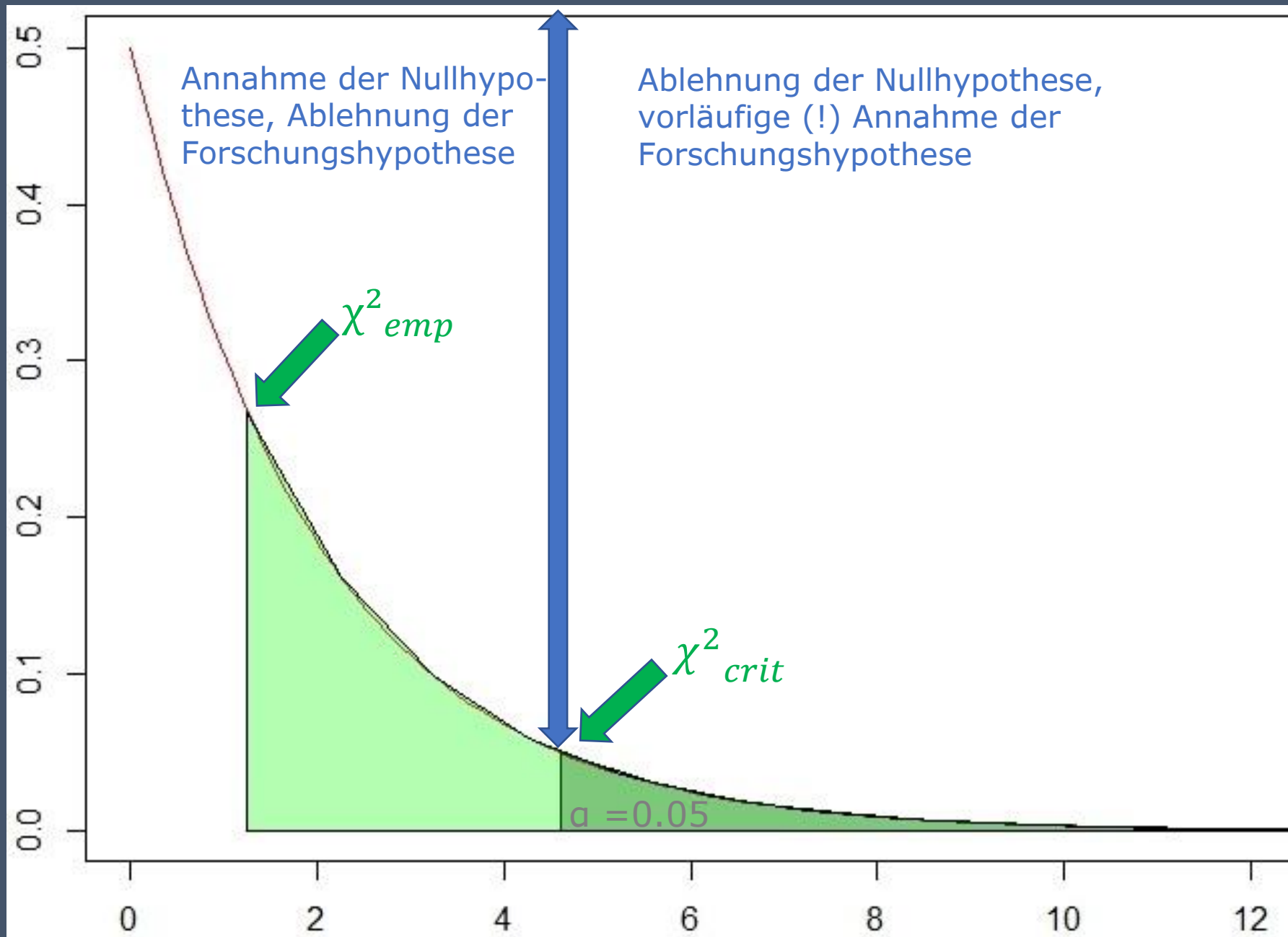
$(fo_i - fe_i)^2$	$\frac{(fo_i - fe_i)^2}{fe_i}$
$(4 - 3)^2 = 1$	$1/3 = 0,33$
$(0 - 0,86)^2 = 0,7396$	$0,7396/0,86 = 0,0086$
$(2 - 2,14)^2 = 0,0196$	$0,0196/2,14 = 0,009$
$(3 - 4)^2 = 1$	$1/4 = 0,25$
$(2 - 1,14)^2 = 0,7396$	$0,7396/1,14 = 0,6488$
$(3 - 2,86)^2 = 0,0196$	$0,0196/2,86 = 0,007$
$\chi^2_{emp} = \sum_{i=1}^n \frac{(fo_i - fe_i)^2}{fe_i} = 1,2534$	

DIE CHI-QUADRAT-TABELLE

	Right Tail Probability						
<i>df</i>	25%	10%	5%	2.5%	1%	0.5%	0.1%
1	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	2.773	4.605	5.991	7.378	9.21	10.597	13.816
3	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	5.385	7.779	9.488	11.143	13.277	14.86	18.467
5	6.626	9.236	11.07	12.833	15.086	16.75	20.515
6	7.841	10.645	12.592	14.449	16.812	18.548	22.458

DER KRITISCHE CHI-QUADRAT-WERT – BEISPIEL

- Freiheitsgrade berechnen: $df = (3 - 1) * (2 - 1) = 2$
- χ^2_{crit} von Tabelle ablesen: $\chi^2_{crit}(2, 0.10) = 4,605$
- χ^2 -Werte vergleichen: $1,2534 < 4,605 \rightarrow$ nicht signifikant!
- Ergebnis deuten:
 - Mit einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese nicht verworfen
 - H_1 – Männer schauen mehr Fernsehen als Frauen – ist falsifiziert



CHI-QUADRAT-TEST IN R

#Daten laden

```
load("daten_x.RData")
```

```
datensatz <- subset(daten_x, Bedingung)
```

#Nicht genannte Ausprägungen löschen

```
datensatz <- within(datensatz, {UV <- droplevels(UV)})
```

```
datensatz <- within(datensatz, {AV <- droplevels(AV)})
```

#Chi-Quadrat-Test durchführen

```
chisq.test(datensatz$AV, datensatz$UV)
```


CHI-QUADRAT-TEST IN R – BEISPIEL

#Daten laden

```
load("daten_2019.RData")
```

```
datensatz <- subset(daten_2019, jahr>2000)
```

#Nicht genannte Ausprägungen löschen

```
datensatz <- within(datensatz, {altersgruppe <-  
droplevels(altersgruppe)})
```

```
datensatz <- within(datensatz, {gem_radiop <-  
droplevels(gem_radiop)})
```

#Chi-Quadrat-Test durchführen

```
chisq.test(datensatz$gem_radiop, datensatz$altersgruppe)
```

```
````{r}
#Daten laden
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr>2000)
#Nicht genannte Ausprägungen löschen
datensatz <- within(datensatz, {altersgruppe <- droplevels(altersgruppe)})
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
#Chi-Quadrat-Test durchführen
chisq.test(datensatz$gem_radiop, datensatz$altersgruppe)
....
```

### Pearson's Chi-squared test

data: datensatz\$gem\_radiop and datensatz\$altersgruppe  
X-squared = 35.613, df = 5, p-value = 1.135e-06

$\chi^2_{emp}$

Freiheitsgrade

p - Wert:  $p < \alpha \rightarrow$  signifikant!

# CRAMER'S V

## Cramer's V

- Gibt an, wie stark der Zusammenhang zwischen zwei kategorialen Variablen ist
- Variiert zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang)
- Gibt nicht an, wie der Zusammenhang gerichtet ist (positiv oder negativ)
- Ist berechenbar ab einer Kreuztabelle von 2x3 bzw. 3x2
- Ist erst gut interpretierbar mit größerer Stichprobe ( $n > 100$ )

# BERECHNUNG UND INTERPRETATION

Formal	Nicht-formal
$V = \sqrt{\frac{\chi^2_{emp}}{n * (c - 1)}}$	$V = \sqrt{\frac{\text{empirischer Chi - Quadrat - Wert}}{\text{Anzahl der Fälle} * (\text{kleinste Anzahl Ausprägungen} - 1)}}$

<b>V-Wert</b>	<b>Stärke des Einflusses</b>
$V < 0,1$	Nicht vorhanden
$0,1 \leq V < 0,3$	gering
$0,3 \leq V < 0,5$	mittel
$V \geq 0,5$	stark

# CRAMER'S V – BEISPIEL

Formel:  $V = \sqrt{\frac{\chi^2_{emp}}{n*(c-1)}}$

gegeben:  $\chi^2_{emp} = 1,2534, n = 14, c = 2$

Rechnung:  $V = \sqrt{\frac{1,2534}{14*2}} = 0,21 \rightarrow$  geringer Zusammenhang!

# CRAMERS V IN R

```
#Daten laden
```

```
load("daten_x.RData")
```

```
datensatz <- subset(daten_x, Bedingung)
```

```
library(lsr)
```

```
#Nicht genannte Ausprägungen löschen
```

```
datensatz <- within(datensatz, {UV <- droplevels(UV)})
```

```
datensatz <- within(datensatz, {AV <- droplevels(AV)})
```

```
#Cramer's V berechnen lassen
```

```
cramersV(datensatz$AV, datensatz$UV))
```

# CRAMERS V IN R – BEISPIEL

```
#Daten laden
```

```
load("daten_2019.RData")
```

```
library(lsr)
```

```
datensatz <- subset(daten_2019, jahr>2000)
```

```
#Nicht genannte Ausprägungen löschen
```

```
datensatz <- within(datensatz, {altersgruppe <- droplevels(altersgruppe)})
```

```
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
```

```
#Cramer's V berechnen lassen
```

```
cramersV(datensatz$gem_radiop, datensatz$altersgruppe)
```

```
````{r}
#Daten laden
load("daten_2019.RData")
library(lsr)
datensatz <- subset(daten_2019, jahr>2000)
#Nicht genannte Ausprägungen löschen
datensatz <- within(datensatz, {altersgruppe<- droplevels(altersgruppe)})
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
#Cramer's V berechnen lassen
cramersV(datensatz$altersgruppe, datensatz$gem_radiop)
```

```
[1] 0.2976397
```

← 0,1 ≤ V < 0,3 → geringer bis mittelstarker Zusammenhang

MOSAIKPLOTS

Mosaikplots

- Stellen Kreuztabellen graphisch dar
- Sind nur sinnvoll für Variablen mit überschaubar vielen Ausprägungen
- Stellen jede Merkmalskombination als Balken dar
 - Breite eines Balkens: Anteil der Ausprägung an der UV
 - Länge eines Balkens: Anteil der Ausprägung an der AV

MOSAIKPLOTS IN R

#Laden und Auswahlen der Daten

```
load("daten_x.RData")  
datensatz <- subset(daten_x, bedingung)
```

#Nichtgenannte Ausprägungen löschen

```
datensatz <- within(datensatz, {UV <- droplevels(UV)})  
datensatz <- within(datensatz, {AV <- droplevels(AV)})
```

#Erstellen des Mosaikplots

```
library(graphics)  
table2 = data.frame(unclass(table(datensatz$UV, datensatz$AV)))  
table2 = as.table(as.matrix(table2))  
mosaicplot(table2, shade = F, las = 1, color = c("dark grey", "light grey", "grey", "black"),  
            main = "Titel")
```

MOSAIKPLOTS IN R – BEISPIEL

```
#Laden und Auswahlen der Daten
```

```
load("daten_2019.RData")
```

```
datensatz <- subset(daten_2019, jahr > 2000)
```

```
#Nichtgenannte Ausprägungen löschen
```

```
datensatz <- within(datensatz, {altersgruppe <- droplevels(altersgruppe)})
```

```
datensatz <- within(datensatz, {gem_radiop <- droplevels(gem_radiop)})
```

```
#Erstellen des Mosaikplots
```

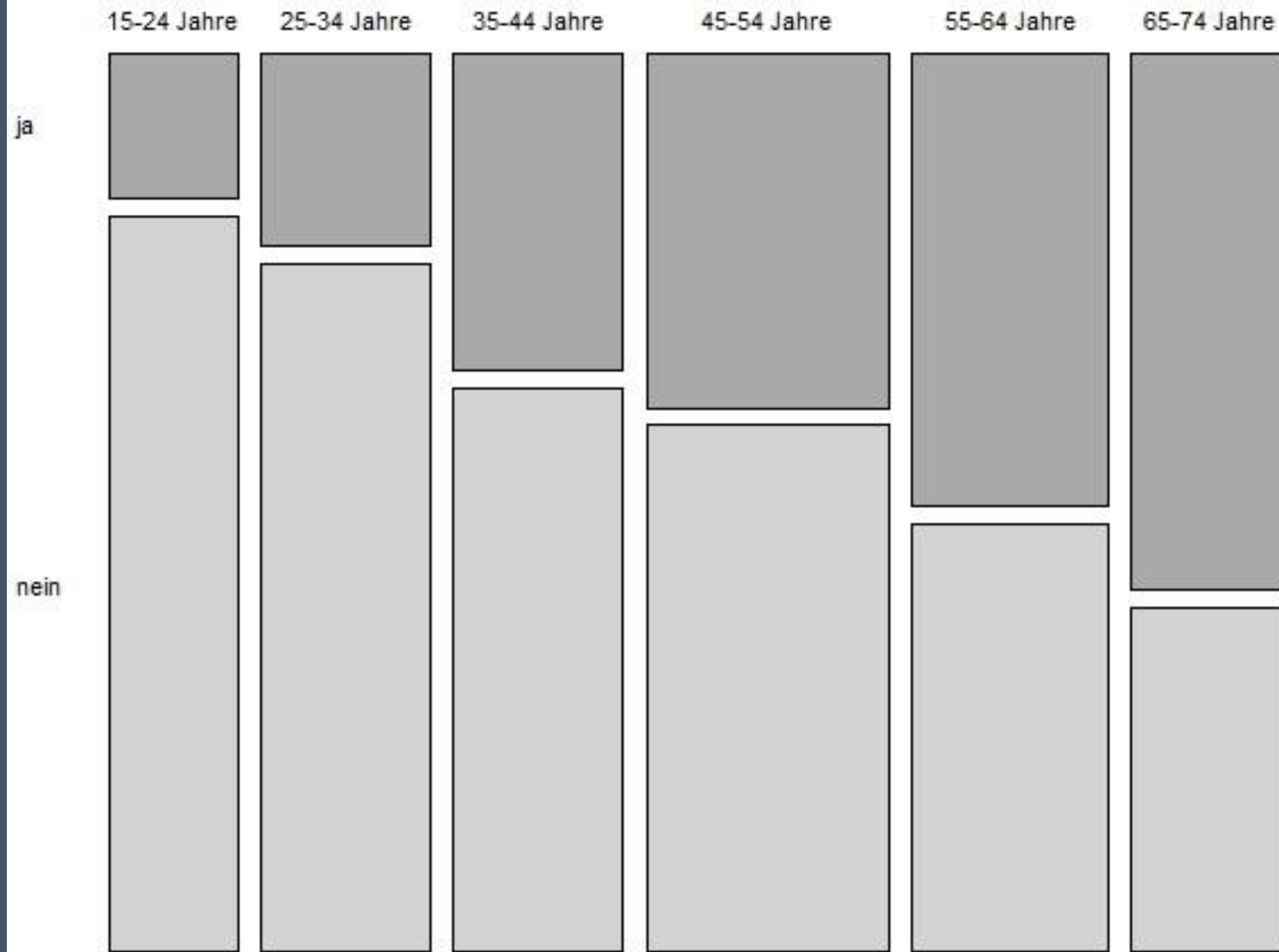
```
library(graphics)
```

```
table2 = data.frame(unclass(table(datensatz$altersgruppe, datensatz$gem_radiop)))
```

```
table2 = as.table(as.matrix(table2))
```

```
mosaicplot(table2, shade = F, las = 1, color = c("dark grey", "light grey", "grey", "black"),  
            main = "TV-Nutzung mit Partner nach Altersgruppe")
```

TV-Nutzung mit Partner nach Altersgruppe



FRAGEN?

SELBSTSTUDIUM

- Forschungsbericht (EFB und ZFB)
 - Zusammenhang zwischen den Variablen `daten_2019$altersgruppe` und `daten_2019$gem_wwwp`
 - Kreuztabelle Spaltenprozent
 - Chi-Quadrat-Test
 - Cramer's V
 - Mosaikplot
 - Text zur Interpretation der Ergebnisse
- Wiederholung: Experiment

BIS NÄCHSTE WOCHEN!