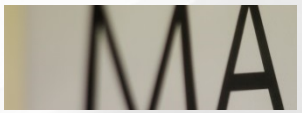
INSTITUT FÜR DEUTSCHE SPRACHE



Sandra Hansen-Morath
Sascha Wolfer

STATISTIK MIT R

Chi-Quadrat

Mitglied der Leibniz-Gemeinschaft

ZUSAMMENFASSUNG INFERENZSTATISTIK

- „schließende“ Statistik
- Prüfung von Hypothesen → von Parametern aus Stichproben wird auf die Werte von Grundgesamtheiten geschlossen
- Unterschiede in der Statistik werden als signifikant bezeichnet, wenn die Wahrscheinlichkeit, dass sie durch Zufall zustande gekommen sind, nicht über einer gewissen Schwelle liegt.
- Fehlertypen: Wahrscheinlichkeit für den α-Fehler = *p* (Irrtumswahrscheinlichkeit)
- Konfidenzintervalle geben einen Bereich an, bei dem wir uns zu x-prozentiger Wahrscheinlichkeit sicher sind, dass ein bestimmter Wert (z.B. Mittelwert) darin liegt.

HÄUFIGKEITEN

- Wir haben es oft mit Variablen zu tun, die nominalskaliert sind.
- Diese Variablen können lediglich bezüglich ihrer Häufigkeiten ausgewertet werden.
  - Aber: Logistische Regression berechnet Auftretenswahrscheinlichkeit einer binären Variable.
- **Deskriptiv** können Häufigkeiten über Kontingenztabellen beschrieben werden.

EIN NICHT-LINGUISTISCHES BEISPIEL

- Wir stellen uns zu zwei Zeitpunkten eine Stunde in die Stadt und notieren, wie viele Leute einen Regenschirm dabeihaben.
- Zeitpunkt 1: bei gutem Wetter; Zeitpunkt 2: bei Regen.
- 2 nominalskalierte Variablen:
  Wetter (Regen ja/nein) & Regenschirm dabei (ja/nein)

	kein Regen	**Regen**
Regenschirm dabei	13	30
kein Regenschirm dabei	62	35

EIN NICHT-LINGUISTISCHES BEISPIEL

	kein Regen	Regen	
Regenschirm dabei	13	30	*43*
kein Regenschirm dabei	62	35	*97*
	75	*65*	

- Aus dieser Kreuztabelle / Kontingenztabelle kann man mehrere Dinge ablesen:
  - Es sind mehr Leute unterwegs, wenn es nicht regnet (75 zu 65).
  - Es sind weniger Leute unterwegs, die einen Regenschirm dabeihaben (43 zu 97).
  - Bei beiden Wetterlagen sind die Leute in der Mehrheit, die **keinen** Regenschirm dabeihaben.
  - Diese Differenz ist aber kleiner, wenn es regnet (5 vs. 49).

CHI-QUADRAT-TEST

- Grundlegende Frage des Chi-Quadrat-Tests: "Weicht die beobachtete Verteilung von einer gleichmäßigen Verteilung ab?"
- Übertragen auf das Beispiel: "Weicht die Verteilung bzgl. Regenschirmtragen bei Regen oder schönem Wetter von einer gleichmäßigen Verteilung ab?"
- Oder: "Gibt es einen Zusammenhang zwischen dem Tragen eines Regenschirms und dem Wetter?"
- Es stellt sich konsequenterweise die nächste Frage: Wie sieht eine gleichmäßige Verteilung aus?!?

GLEICHMÄSSIGE VERTEILUNG

- Die gleichmäßige Verteilung in einer Kontingenztabelle bedeutet **nicht**, dass in jeder Zelle gleich viele Fälle sind.
- Es muss beachtet werden, wie wahrscheinlich das Auftreten in einer Zelle überhaupt ist.
  - Beispiel: Wir haben gesehen, dass bei schönem Wetter mehr Menschen auf der Straße sind und dass grundsätzlich weniger Leute einen Regenschirm dabeihaben. Das müssen wir beachten!
- Daher verrechnen wir die Zeilen- und Spaltensummen miteinander, um zu **erwarteten** Häufigkeiten in jeder Zelle zu kommen.
- Diese erwarteten Häufigkeiten repräsentieren dann die Werte, die wir – gegeben eine Gleichverteilung – in jeder Zelle erwarten.

ERWARTETE UND BEOBACHTETE HÄUFIGKEITEN

- Die erwarteten Häufigkeiten werden mit den tatsächlich **beobachteten** Häufigkeiten verglichen.
- Sind die Abweichungen (**Residuen**) zwischen erwarteten und beobachteten Häufigkeiten insgesamt sehr groß, liegt offenbar eine Abweichung von der Gleichverteilung vor.
- Der Chi-Quadrat-Test zeigt uns dies durch einen signifikanten Chi-Quadrat-Wert an.
  - "**Anpassungstest**": Vergleicht beobachtete Verteilung mit erwarteter Verteilung.
- Achtung! Die erwarteten Häufigkeiten sollten nicht in über 20% der Zellen unter 5 fallen.

BERECHNUNG ERWARTETER HÄUFIGKEITEN

	kein Regen	Regen	
Regenschirm dabei	13	30	*43*
kein Regenschirm dabei	62	35	*97*
	75	*65*	*140*

- Die erwartete Häufigkeit einer Zelle ergibt sich durch Spaltensumme * Zeilensumme / Anzahl Fälle in Tabelle
- Erwartete Häufigkeiten ergeben sich aus den **Wahrscheinlichkeiten**, dass ein Fall in eine Zelle fällt.

BERECHNUNG ERWARTETER HÄUFIGKEITEN

	kein Regen	Regen	
Regenschirm dabei	43*75 / 140	43*65 / 140	*43*
kein Regenschirm dabei	97*75 / 140	97*65 / 140	*97*
	75	*65*	*140*

- Die erwartete Häufigkeit einer Zelle ergibt sich durch Spaltensumme * Zeilensumme / Anzahl Fälle in Tabelle
- Erwartete Häufigkeiten ergeben sich aus den **Wahrscheinlichkeiten**, dass ein Fall in eine Zelle fällt.

BERECHNUNG ERWARTETER HÄUFIGKEITEN

	kein Regen	Regen	
Regenschirm dabei	23,04	19,96	*43*
kein Regenschirm dabei	51,96	45,04	*97*
	75	*65*	*140*

- Die erwartete Häufigkeit einer Zelle ergibt sich durch Spaltensumme * Zeilensumme / Anzahl Fälle in Tabelle
- Erwartete Häufigkeiten ergeben sich aus den **Wahrscheinlichkeiten**, dass ein Fall in eine Zelle fällt.

BERECHNUNG ERWARTETER HÄUFIGKEITEN

	kein Regen	Regen
Regenschirm dabei	23,04	19,96
kein Regenschirm dabei	51,96	45,04

	kein Regen	Regen
Regenschirm dabei	13	30
kein Regenschirm dabei	62	35

BERECHNUNG ERWARTETER HÄUFIGKEITEN

	kein Regen	Regen
Regenschirm dabei	13 - 23,04	30 - 19,96
kein Regenschirm dabei	62 - 51,96	35 - 45,04

	kein Regen	Regen
Regenschirm dabei	-10,04	10,04
kein Regenschirm dabei	10,04	-10,04

OMNIBUS-TEST

- Ein signifikanter Chi-Quadrat-Test gibt uns an, ob die beobachtete Verteilung von der erwarteten Verteilung abweicht.
- Er gibt uns nicht unmittelbar an, **welche Zelle** in der Kontingenztabelle für diesen Effekt verantwortlich ist.
- Der Chi-Quadrat-Test wird daher als "**Omnibus-Test**" bezeichnet.
- Über die **standardisierten Residuen** können wir herausfinden, welche Zelle für den Effekt verantwortlich ist. → `residuals()`

EINBLICKE IN DIE INFERENZSTATISTIK

EIN BEISPIEL

- Wie ist das Vorkommen von *geil* in zwei Zeitungskorpora (St. Galler Tagblatt und Tages-Anzeiger)?
- Abfrage in COSMAS II:

	SG Tagblatt	Tages-Anzeiger
geil	131	170
Wörter TOTAL	103 644 782	60 065 707
Texte TOTAL	349 085	142 714

- Ist der Unterschied der Frequenzen von *geil* in den beiden Korpora signifikant? Kann mit genügend großer Sicherheit angenommen werden, dass der Frequenzunterschied in den beiden Korpora nicht zufällig zustande gekommen ist?
- Wie könnten die Hypothesen aussehen?

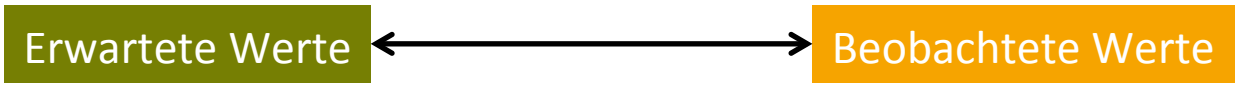
H0: Die Frequenzen des Wortes geil unterscheiden sich nicht signifikant in den beiden Korpora
H1: Die Frequenzen des Wortes geil unterscheiden sich statistisch bedeutsam in den beiden Korpora

(Bsp. nach Bubenhofer 2006-2013)

EINBLICKE IN DIE INFERENZSTATISTIK

EIN BEISPIEL

- Welche Frequenzen würde man erwarten, wenn man davon ausgeht, dass die Frequenz von *geil* gleichmäßig in den Korpora verteilt wäre?



- Wie groß ist der Abstand zwischen den beobachteten und den erwarteten Werten?

EINBLICKE IN DIE INFERENZSTATISTIK

	Korpus A	Korpus B	Total
Freq. Wort x	A	B	A+B
Alle anderen	C	D	C+D
Total	A+C	B+D	A+B+C+D

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	
Alle anderen			
Total	103 644 782	60 065 707	

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	
Alle anderen	103 644 651		
Total	103 644 782	60 065 707	

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	
Alle anderen	103 644 651	60 065 537	
Total	103 644 782	60 065 707	

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	301
Alle anderen	103 644 651	60 065 537	163 710 188
Total	103 644 782	60 065 707	163 710 489

- Wie müsste die Tabelle aussehen, wenn man von einer gleichmäßigen Verteilung des Wortes *geil* in den beiden Korpora ausgehen würden?
- Die erwarteten Werte können mit einem Dreisatz berechnet werden

erwartet	SG Tagblatt	Tages-Anzeiger	Total
geil	x = 301 / 163710489 * 103644782 x = 190,6	x = 301 / 163 710 489 * 60 065707 x = 110,4	301
Alle anderen	x = 103644782-190,6 x = 103 644 591,4	x = 60065707 – 110,4 x = 60 065 596,6	163 710 188
Total	103 644 782	60 065 707	163 710 489

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	301
Alle anderen	103 644 651	60 065 537	163 710 188
Total	103 644 782	60 065 707	163 710 489

- Wie müsste die Tabelle aussehen, wenn man von einer gleichmäßigen Verteilung des Wortes *geil* in den beiden Korpora ausgehen würden?
- Die erwarteten Werte können mit einem Dreisatz berechnet werden

erwartet	SG Tagblatt	Tages-Anzeiger	Total
geil	x = 301 / 163710489 * 103644782 x = 190,6	x = 301 / 163 710 489 * 60 065 707 x = 110,4	301
Alle anderen	x = 103644782-190,6 x = 103 644 591,4	x = 60065707 – 110,4 x = 60 065 596,6	163 710 188
Total	103 644 782	60 065 707	163 710 489

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	301
Alle anderen	103 644 651	60 065 537	163 710 188
Total	103 644 782	60 065 707	163 710 489

- Wie müsste die Tabelle aussehen, wenn man von einer gleichmäßigen Verteilung des Wortes *geil* in den beiden Korpora ausgehen würden?
- Die erwarteten Werte können mit einem Dreisatz berechnet werden

erwartet	SG Tagblatt	Tages-Anzeiger	Total
geil	x = 301 / 163710489 * 103644782 x = 190,6	x = 301 / 163 710 489 * 60 065707 x = 110,4	301
Alle anderen	x = 103644782-190,6 x = 103 644 591,4	x = 60065707 – 110,4 x = 60 065 596,6	163 710 188
Total	103 644 782	60 065 707	163 710 489

EINBLICKE IN DIE INFERENZSTATISTIK

beobachtet	SG Tagblatt	Tages-Anzeiger	Total
geil	131	170	301
Alle anderen	103 644 651	60 065 537	163 710 188
Total	103 644 782	60 065 707	163 710 489

- Wie müsste die Tabelle aussehen, wenn man von einer gleichmäßigen Verteilung des Wortes *geil* in den beiden Korpora ausgehen würden?
- Die erwarteten Werte können mit einem Dreisatz berechnet werden

erwartet	SG Tagblatt	Tages-Anzeiger	Total
geil	x = 301 / 163710489 * 103644782 x = 190,6	x = 301 / 163 710 489 * 60 065707 x = 110,4	301
Alle anderen	x = 103644782-190,6 x = 103 644 591,4	x = 60065707 – 110,4 x = 60 065 596,6	163 710 188
Total	103 644 782	60 065 707	163 710 489

EINBLICKE IN DIE INFERENZSTATISTIK
DER CHI-QUADRAT-TEST

- Ein Standardverfahren für den vorliegenden Fall ist der Chi-Quadrat-Test

Beobachtet (erwartet)	SG Tagblatt	Tages-Anzeiger	Total
geil	131 (190,6)	170 (110,4)	301
Alle anderen	103 644 651 (103 644 591,4)	60 065 537 (60 065 596,6)	163 710 188
Total	103 644 782	60 065 707	163 710 489

- Die Formel für den Chi-Quadrat-Test lautet:

beobachteter Wert (observed)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

erwarteter Wert (expexted)

EINBLICKE IN DIE INFERENZSTATISTIK

DER CHI-QUADRAT-TEST

Beobachtet (erwartet)	SG Tagblatt	Tages-Anzeiger	Total
geil	131 (190,6)	170 (110,4)	301
Alle anderen	103 644 651 (103 644 591,4)	60 065 537 (60 065 596,6)	163 710 188
Total	103 644 782	60 065 707	163 710 489

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$X^2 = ((131 - 190{,}6)^2/190{,}6)$
$+ ((170 - 110{,}4)^2/110{,}4)$
$+ ((103\,644\,651 - 103\,644\,591{,}4)^2/103\,644\,591{,}4)$
$+ ((60\,065\,537 - 60\,065\,596{,}6)^2/60\,065\,596{,}6)$

= **50,74**

EINBLICKE IN DIE INFERENZSTATISTIK
DER CHI-QUADRAT-TEST

- Der Chi-Quadrat-Wert ist ermittelt: **50,74**
- Ablesen in einer Tabelle, in der die sog. **kritischen Werte** für X^2 aufgeführt sind, ob der berechnete Wert signifikant ist
- Diese Tabellen sind in Statistikbüchern zu finden oder aber im Web
- Auszug aus einer Tabelle:

df	p = 0,05	p = 0,01	p = 0,001
1	3,84	6,64	10,83
2	5,99	9,21	13,82
3	7,82	11,35	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46

df → Freiheitsgrad
df = (Reihenzahl – 1) * (Spaltenzahl – 1)
df = 1

Ergebnis:
- Wenn X^2 größer als 3,84 ist, dann sind die Frequenzunterschiede mit 95%iger Sicherheit signifikant , also nicht zufällig
- In dem vorliegenden Fall kann mit einer Wahrscheinlichkeit von 99,9% davon ausgegangen werden, dass die Frequenzverteilungen nicht zufällig sind

Die H0 kann abgelehnt werden!
Die H1 ist bestätigt!

DER CHI-QUADRAT-TEST

- In Kreuztabellen ergeben sich die erwarteten Häufigkeiten aus den Zeilen- und Spaltensummen.
- Die Prüfgröße Chi-Quadrat ergibt sich aus der Summe der quadrierten standardisierten Abweichungen von erwarteten und beobachteten Häufigkeiten.
- Die Prüfgröße Chi-Quadrat gemeinsam mit den assoziierten Freiheitsgraden ergibt die Irrtumswahrscheinlichkeit p.

  → Freiheitsgrade (df) = (Reihenzahl – 1) * (Spaltenzahl – 1)
- Wenn die Nullhypothese (=Verteilung ist gleichmäßig) abgelehnt werden kann, wissen wir aber nicht, wo genau (also in welchen Zellen) die beobachteten von den erwarteten Häufigkeiten abweichen.

→ Lösung: Berechnung und Visualisierung der Pearson-Residuen

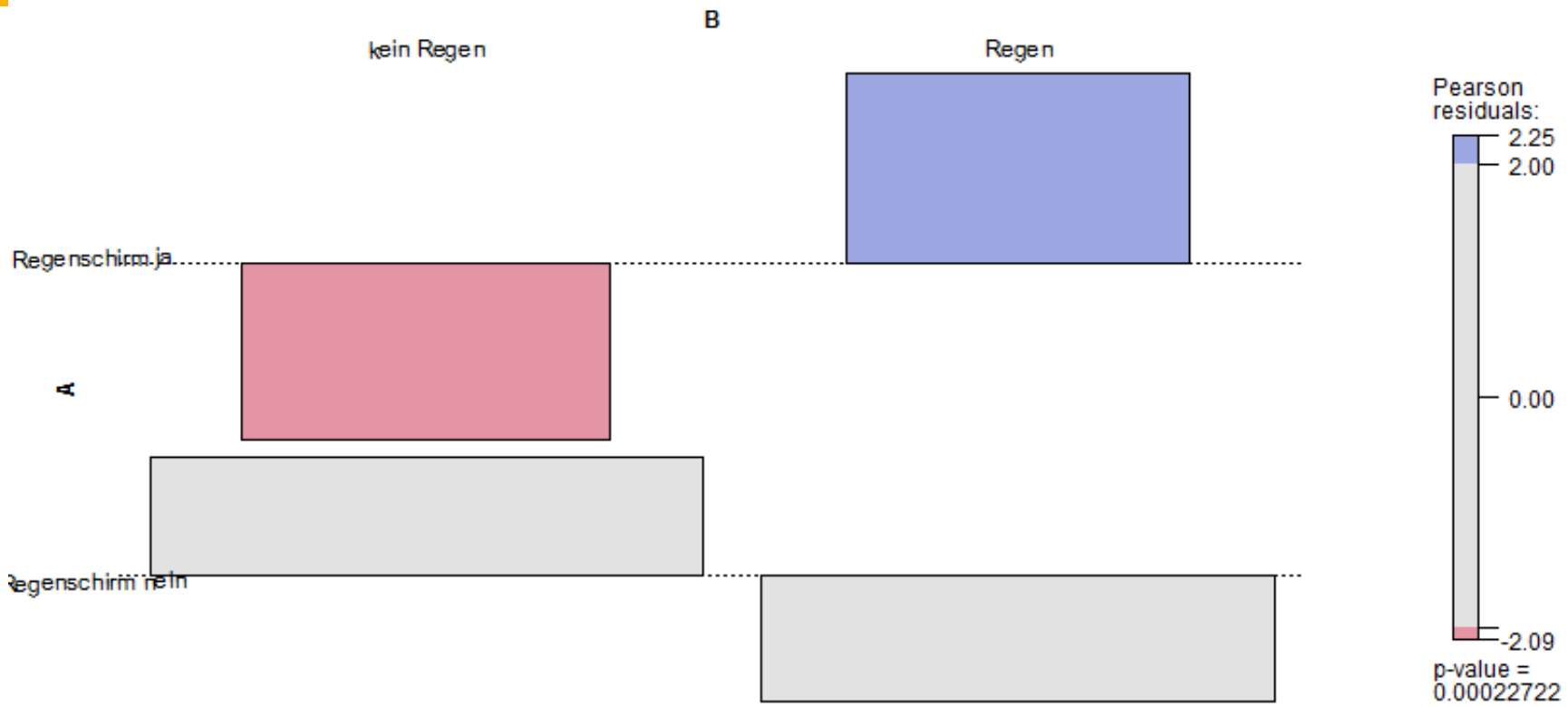
→ Immer wenn Betrag des Pearson-Residuums > 2: Signifikante Abweichung.

BEISPIEL VISUALISIERUNG: REGEN

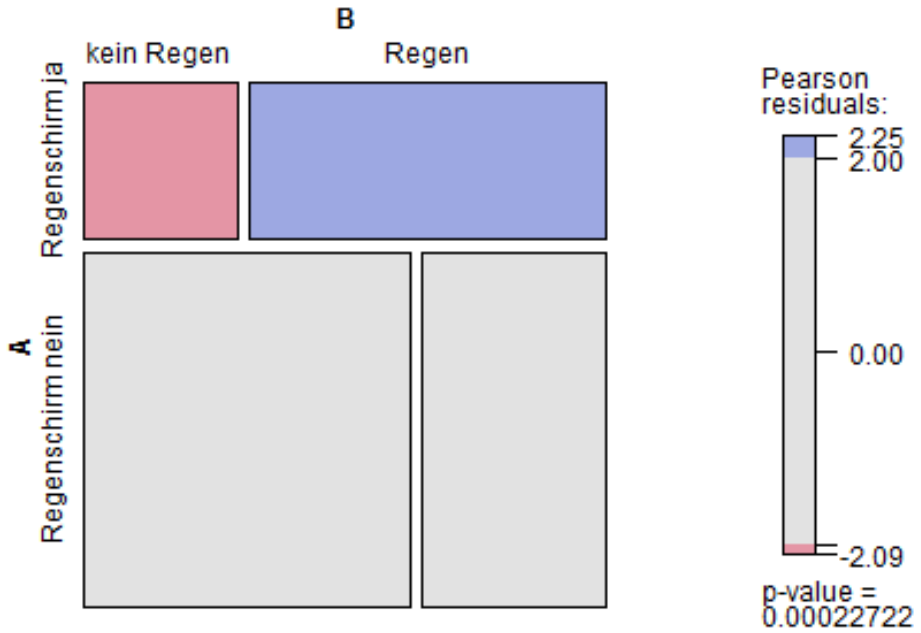
- Wir stellen uns zu zwei Zeitpunkten eine Stunde in die Stadt und notieren, wie viele Leute einen Regenschirm dabeihaben.
- Zeitpunkt 1: bei gutem Wetter; Zeitpunkt 2: bei Regen.
- 2 nominalskalierte Variablen:
  Wetter (Regen ja/nein) & Regenschirm dabei (ja/nein)

	kein Regen	Regen
Regenschirm dabei	13	30
kein Regenschirm dabei	62	35

BEISPIEL ASSOZIATIONSPLOT: REGENSCHIRME



BEISPIEL MOSAIKPLOT REGENSCHIRME



DER CHI-QUADRAT-TEST IN R

Schritt 1: Kreuztabelle erstellen, 2 Möglichkeiten:

1. **aus einem Dataframe** mit der Funktion
   `table(<Spalte1>[, <Spalte2>, ...])`
2. **manuelles Erstellen** einer Tabelle mit `rbind()` oder `cbind()`

`x <- table(dat$spalte1, dat$spalte2)` (Variante **1**)

Schritt 2: Berechnung des Chi-Quadrat-Wertes und der Irrtumswahrscheinlichkeit

- `chisq.test(x)`
- Das Ergebnis kann zur Weiterverarbeitung in einer Variable gespeichert werden:
  `chi <- chisq.test(x)`
- Extraktion der Residuen: `residuals(chi)`

DER CHI-QUADRAT-TEST IN R

Schritt 3: Visualisierung der Residuen mit einem Assoziationsplot oder einem Mosaikplot.

- Die Plots können mit Hilfe des Pakets `vcd` erzeugt werden.

```
> install.packages("vcd")
> library(vcd)
```

- `x` enthält noch immer eine Tabelle.
- Assoziationsplot: `assoc(x)`
- Mosaikplot: `mosaic(x)`

Schritt 4: Berechnung der **Effektstärke**.

EFFEKTSTÄRKE

- Achtung! Der Chi-Quadrat-Test ist **extrem sensitiv für die Anzahl der Fälle** in der zugrundeliegenden Tabelle.
- Fallbeispiel:
  - In Tabelle 1 ist das Vorkommen eines linguistischen Merkmals auf *n* in 1000 Fällen normiert.
  - In Tabelle 2 verändern wir lediglich die Normierung auf *n* in 10000 Fällen.

Tab. 1

31	48
22	53

$p = 0{,}2611$

* 10

310	480
220	530

Tab. 2

$p = 0{,}000054$

Lösung: Berechnung der Assoziationsstärke, die **nicht** von der Anzahl der Fälle beeinflusst wird.

EFFEKTSTÄRKE

- Die Assoziationsstärke für Kontingenz-Tabellen variiert zwischen 0 und 1.
- Als Maß verwenden wir Cramérs *V*.
  (für 4-Felder-Tabellen = ϕ-Koeffizient)
- 0: kein Zusammenhang, 1: perfekter Zusammenhang
- Für unser Beispiel gilt ϕ = V = 0,104
- Schwellenwerte:
  - 0,1 – 0,3: schwacher Zusammenhang
  - 0,4 – 0,5: mittlerer Zusammenhang
  - > 0,5: starker Zusammenhang

EFFEKTSTÄRKE

Funktion in R:

```
cv.test <- function(x)
{
  CV <- sqrt(chisq.test(x, correct = FALSE)$statistic /
            (sum(x) * min(dim(x) - 1 )))
  print("Cramer V / Phi:")
  as.numeric(CV)
}

cv.test(x)
```

Die Funktion `phi()` aus dem Paket `psych` berechnet den ϕ-Koeffizienten für 4-Felder-Tabellen.

ÜBUNG: ANAPHORISCHE BEZIEHUNG, TYP DER REFERENZ (AUS WOLFER (I.V.))

	Volle NP	Proform
Innerhalb Satz	178	126
Über Satzgrenze	879	112