Statistik
03. Nichtparametrische Verfahren

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: `https://github.com/rsling/VL-Deutsche-Syntax`

Inhalt

Zähldaten

Übersicht

- Unterschiede in Zähldaten

Übersicht

- Unterschiede in Zähldaten
- Signifikanz und Effektstärke

Übersicht

- Unterschiede in Zähldaten
- Signifikanz und Effektstärke
- Unterschiede bei Ja/Nein-Experimenten

Literatur

- Gravetter & Wallnau 2007
- Bortz & Lienert 2008

Kreuztabelle

Beobachtungen von zwei kategorialen Variablen.
Auxiliarwahl beim Perfekt: haben, sein
Herkunft des Belegs: nord, sued

Fall	Aux	Region
1	haben	nord
2	haben	nord
3	sein	nord
4	sein	sued
5	sein	sued
6	haben	nord
7	haben	sued
8	haben	sued

Kreuztabelle

Beobachtungen von zwei kategorialen Variablen.
Auxiliarwahl beim Perfekt: haben, sein
Herkunft des Belegs: nord, sued

Fall	**Aux**	**Region**
1	haben	nord
2	haben	nord
3	sein	nord
4	sein	sued
5	sein	sued
6	haben	nord
7	haben	sued
8	haben	sued

	Aux	
Region	haben	sein
nord	3	1
sued	2	2

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	
sued	2	2	
Spaltensummen			

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	
Spaltensummen			

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	4
Spaltensummen			

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	4
Spaltensummen	5		

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	4
Spaltensummen	5	3	

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i: $\sum_k x_{ik}$

Zeilensumme für Zeile j: $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	4
Spaltensummen	5	3	**8**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
50 mal Norden, 50 mal Süden (= Zeilensummen)

	haben	sein	Zeilensummen
nord			50
sued			50
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
50 mal Norden, 50 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL
  = kein Zusammenhang zwischen Hilfsverb und Region?

	haben	sein	Zeilensummen
nord			50
sued			50
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
50 mal Norden, 50 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL
  = kein Zusammenhang zwischen Hilfsverb und Region?

	haben	sein	Zeilensummen
nord	25	25	50
sued	25	25	50
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord			
sued			
Spaltensummen			**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord			30
sued			
Spaltensummen			**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord			30
sued			70
Spaltensummen			**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord			30
sued			70
Spaltensummen	50		**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord			30
sued			70
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord	15		30
sued			70
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord	15	15	30
sued			70
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord	15	15	30
sued	35		70
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
50 mal *haben*, 50 mal *sein* (= Spaltensummen)
30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der NULL?

	haben	sein	Zeilensummen
nord	15	15	30
sued	35	35	70
Spaltensummen	50	50	**100**

Beobachtete vs. erwartete Häufigkeiten

n=100
30 mal Norden, 70 mal Süden
40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord			30
sued			70
Spaltensummen	40	60	**100**

Allgemein: erwartete Häufigkeit für Zellen: $\frac{Spaltensumme \cdot Zeilensumme}{n}$

bzw.: $EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$

Beobachtete vs. erwartete Häufigkeiten

n=100
30 mal Norden, 70 mal Süden
40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord	12		30
sued			70
Spaltensummen	40	60	**100**

Allgemein: erwartete Häufigkeit für Zellen: $\frac{Spaltensumme \cdot Zeilensumme}{n}$

bzw.: $EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$

Beobachtete vs. erwartete Häufigkeiten

n=100
30 mal Norden, 70 mal Süden
40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord	12	18	30
sued			70
Spaltensummen	40	60	**100**

Allgemein: erwartete Häufigkeit für Zellen: $\frac{Spaltensumme \cdot Zeilensumme}{n}$

bzw.: $EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$

Beobachtete vs. erwartete Häufigkeiten

n=100
30 mal Norden, 70 mal Süden
40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord	12	18	30
sued	28		70
Spaltensummen	40	60	**100**

Allgemein: erwartete Häufigkeit für Zellen: $\frac{Spaltensumme \cdot Zeilensumme}{n}$

bzw.: $EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$

Beobachtete vs. erwartete Häufigkeiten

n=100
30 mal Norden, 70 mal Süden
40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord	12	18	30
sued	28	42	70
Spaltensummen	40	60	**100**

Allgemein: erwartete Häufigkeit für Zellen: $\frac{Spaltensumme \cdot Zeilensumme}{n}$

bzw.: $EH(x_{ij}) = \frac{\sum\limits_k x_{ik} \cdot \sum\limits_k x_{kj}}{n}$

Beobachtete vs. erwartete Häufigkeiten

beobachtete Häufigkeiten für eine DeReKo-Stichprobe (*geschwebt*):

	haben	sein	Zeilensummen
nord	27	33	60
sued	3	34	37
Spaltensummen	30	67	**97**

Beobachtete vs. erwartete Häufigkeiten

beobachtete Häufigkeiten für eine DeReKo-Stichprobe (*geschwebt*):

	haben	sein	Zeilensummen
nord	27	33	60
sued	3	34	37
Spaltensummen	30	67	**97**

erwartete Häufigkeiten:

	haben	sein	Zeilensummen
nord			60
sued			37
Spaltensummen	30	67	**97**

Beobachtete vs. erwartete Häufigkeiten

beobachtete Häufigkeiten für eine DeReKo-Stichprobe (*geschwebt*):

	haben	sein	Zeilensummen
nord	27	33	60
sued	3	34	37
Spaltensummen	30	67	**97**

erwartete Häufigkeiten:

	haben	sein	Zeilensummen
nord	18.56	41.44	60
sued	11.44	25.56	37
Spaltensummen	30	67	**97**

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.
- NULL: kein Zusammenhang zwischen Region und Aux.

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.
- NULL: kein Zusammenhang zwischen Region und Aux.
- Ab wann ist der Unterschied „signifikant"?

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.
- NULL: kein Zusammenhang zwischen Region und Aux.
- Ab wann ist der Unterschied „signifikant“?

- Ein gemessener Unterschied ist siginifikant, wenn er angesichts der Stichprobengröße groß genug ist, dass wir das im Experiment gefundene Ergenbis nur sehr selten (typischwerweise in unter 5% der Fälle) erwarten würden, wenn er gar nicht bestünde.

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.
- NULL: kein Zusammenhang zwischen Region und Aux.
- Ab wann ist der Unterschied „signifikant“?

- Ein gemessener Unterschied ist siginifikant, wenn er angesichts der Stichprobengröße groß genug ist, dass wir das im Experiment gefundene Ergenbis nur sehr selten (typischwerweise in unter 5% der Fälle) erwarten würden, wenn er gar nicht bestünde.

- Diese 5% (als Anteil 0.05) sind das Signifikanzniveau.

Problem

- Beobachtete und erwartete Häufigkeit weichen ab.
- NULL: kein Zusammenhang zwischen Region und Aux.
- Ab wann ist der Unterschied „signifikant“?

- Ein gemessener Unterschied ist siginifikant, wenn er angesichts der Stichprobengröße groß genug ist, dass wir das im Experiment gefundene Ergenbis nur sehr selten (typischwerweise in unter 5% der Fälle) erwarten würden, wenn er gar nicht bestünde.

- Diese 5% (als Anteil 0.05) sind das Signifikanzniveau.
- In Fishers Philosophie abgekürzt SIG, nicht wie oft zu lesen „α-Niveau“.

χ^2-Unterschiedstest

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

χ^2-Unterschiedstest

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

χ^2-Unterschiedstest

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

bzw.: $$\chi^2 = \sum_{ij} \frac{(x_{ij} - EH(x_{ij}))^2}{EH(x_{ij})}$$

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \frac{(27-18.56)^2}{18.56}$$

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet-erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \frac{(27-18.56)^2}{18.56} + \frac{(33-41.44)^2}{41.44}$$

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \frac{(27-18.56)^2}{18.56} + \frac{(33-41.44)^2}{41.44} + \frac{(3-11.44)^2}{11.44}$$

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet-erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \frac{(27-18.56)^2}{18.56} + \frac{(33-41.44)^2}{41.44} + \frac{(3-11.44)^2}{11.44} + \frac{(34-25.56)^2}{25.56}$$

Berechnung des χ^2-Werts

$$\chi^2 = \sum \frac{(beobachtet - erwartet)^2}{erwartet}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

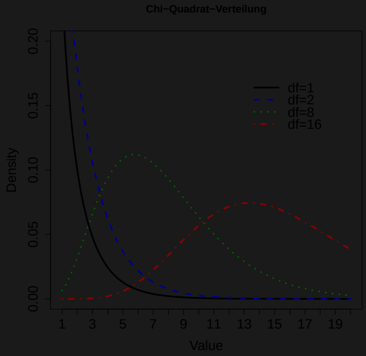
	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \frac{(27-18.56)^2}{18.56} + \frac{(33-41.44)^2}{41.44} + \frac{(3-11.44)^2}{11.44} + \frac{(34-25.56)^2}{25.56}$$

$$\chi^2 = 3.84 + 1.72 + 6.23 + 2.79 = 14.58$$

Die χ^2-Verteilung

Die χ^2-Verteilung für Stichproben
aus Grundgesamtheiten ohne Zusammenhang:



Freiheitsgrad?

Was sind „Freiheitsgrade“ oder *degrees of freedom (df)*?

- Das kommt später noch ausführlicher.

Freiheitsgrad?

Was sind „Freiheitsgrade“ oder *degrees of freedom (df)*?

- Das kommt später noch ausführlicher.
- Für n-Felder-Tests: (Zeilenzahl−1)·(Spaltenzahl−1)

Freiheitsgrad?

Was sind „Freiheitsgrade“ oder *degrees of freedom (df)*?

- Das kommt später noch ausführlicher.
- Für n-Felder-Tests: (Zeilenzahl−1)·(Spaltenzahl−1)
- Bei Vierfelder-Test also: $df = 1$

Die χ^2-Verteilung II

- Wahrscheinlichkeit eines bestimmten χ^2-Werts unter Annahme der NULL?
  VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.

Die χ^2-Verteilung II

- Wahrscheinlichkeit eines bestimmten χ^2-Werts unter Annahme der NULL?
  VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.

- In Fishers Philosophie Entscheidung nach Signifikanzniveau (SIG):
  Der χ^2-Wert muss in den extremen SIG-Anteilen liegen,
  um die NULL zu SIG zurückzuweisen.

Die χ^2-Verteilung II

- Wahrscheinlichkeit eines bestimmten χ^2-Werts unter Annahme der NULL?
  VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.

- In Fishers Philosophie Entscheidung nach Signifikanzniveau (SIG):
  Der χ^2-Wert muss in den extremen SIG-Anteilen liegen,
  um die NULL zu SIG zurückzuweisen.

Die χ^2-Verteilung II

- Wahrscheinlichkeit eines bestimmten χ^2-Werts unter Annahme der NULL?
  VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.

- In Fishers Philosophie Entscheidung nach Signifikanzniveau (SIG):
  Der χ^2-Wert muss in den extremen SIG-Anteilen liegen,
  um die NULL zu SIG zurückzuweisen.

In R ähnlich wie bei Normalverteilung:

```
> qchisq(0.95, df=1) ⇒ 3.84
```

Die χ^2-Verteilung II

- Wahrscheinlichkeit eines bestimmten χ^2-Werts unter Annahme der NULL?
  VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.

- In Fishers Philosophie Entscheidung nach Signifikanzniveau (SIG):
  Der χ^2-Wert muss in den extremen SIG-Anteilen liegen,
  um die NULL zu SIG zurückzuweisen.

  In R ähnlich wie bei Normalverteilung:

  ```
  > qchisq(0.95, df=1) ⇒ 3.84
  ```

- Also ist für $\chi^2 = 14.58$ auf jeden Fall $p < 0.05$ (weil $14.58 > 3.84$).

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten SIG.

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten SIG.
- Die Idee von „mehr oder weniger signifikant“ ist kompletter Schwachsinn.

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten Sig.
- Die Idee von „mehr oder weniger signifikant“ ist kompletter Schwachsinn.
- Entweder ist das gesetzte Niveau akzeptabel,
  und dann bringt ein kleineres p aber auch nicht mehr.

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten SIG.
- Die Idee von „mehr oder weniger signifikant“ ist kompletter Schwachsinn.
- Entweder ist das gesetzte Niveau akzeptabel,
  und dann bringt ein kleineres p aber auch nicht mehr.
- Oder es müsste eigtl. ein strengeres SIG-Niveau gewählt werden,
  und dann ist $p < 0.05$ schlicht nicht ausreichend (s. Fishers Sensitivität).

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten Sɪɢ.
- Die Idee von „mehr oder weniger signifikant“ ist kompletter Schwachsinn.
- Entweder ist das gesetzte Niveau akzeptabel,
  und dann bringt ein kleineres p aber auch nicht mehr.
- Oder es müsste eigtl. ein strengeres Sɪɢ-Niveau gewählt werden,
  und dann ist $p < 0.05$ schlicht nicht ausreichend (s. Fishers Sensitivität).
- Die Entscheidung für ein bestimmtes Sɪɢ-Niveau muss
  auf Basis konzeptueller/inhaltlicher Gründe gefällt werden.

Mehr oder weniger signifikant?

- Oft liest man etwas von „α-Niveaus“ wie:
  - 5% („signifikant“)
  - 1%
  - 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten SIG.
- Die Idee von „mehr oder weniger signifikant“ ist kompletter Schwachsinn.
- Entweder ist das gesetzte Niveau akzeptabel, und dann bringt ein kleineres p aber auch nicht mehr.
- Oder es müsste eigtl. ein strengeres SIG-Niveau gewählt werden, und dann ist $p < 0.05$ schlicht nicht ausreichend (s. Fishers Sensitivität).
- Die Entscheidung für ein bestimmtes SIG-Niveau muss auf Basis konzeptueller/inhaltlicher Gründe gefällt werden.
- EIN signifikantes Testergebnis alleine sagt nicht viel aus!!!

Voraussetzungen für χ^2-Tests

1. Die Beoabachtungen sind voneinander unabhängig.

Voraussetzungen für χ^2-Tests

1. Die Beoabachtungen sind voneinander unabhängig.
2. In jeder Zelle ist die erwartete Häufigkeit mindestens 5.

Voraussetzungen für χ^2-Tests

1. Die Beoabachtungen sind voneinander unabhängig.
2. In jeder Zelle ist die erwartete Häufigkeit mindestens 5.
3. Keine Beschränkung auf vier Felder!

In R

Mit einer Matrix `my.matrix`:

```
> chisq.test(my.matrix)
```

Eingabe einer einfachen Vierfeldermatrix:

```
> my.matrix <- matrix(c(27,33,3,34), 2, 2, byrow=TRUE)
```

Ausgeben der erwarteten Häufigkeiten:

```
> chisq.test(my.matrix)$expected
```

Wann und wie Fisher-Exakt?

Der Fisher-Exakt-Test ist eine Alternative zum χ^2-Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit

Wann und wie Fisher-Exakt?

Der Fisher-Exakt-Test ist eine Alternative zum χ^2-Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit
- keine allgemein bessere Alternative zu χ^2

Wann und wie Fisher-Exakt?

Der Fisher-Exakt-Test ist eine Alternative zum χ^2-Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit
- keine allgemein bessere Alternative zu χ^2
- robuster bei sehr kleinen Stichproben

Wann und wie Fisher-Exakt?

Der Fisher-Exakt-Test ist eine Alternative zum χ^2-Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit
- keine allgemein bessere Alternative zu χ^2
- robuster bei sehr kleinen Stichproben
- aber nur für feststehende Randsummen geeignet!

Fisher-Exakt in R:

```
> fisher.test(my.matrix)
> fisher.test(my.vector.1, my.vector.2)
```

Wann und wie Fisher-Exakt?

Der Fisher-Exakt-Test ist eine Alternative zum χ^2-Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit
- keine allgemein bessere Alternative zu χ^2
- robuster bei sehr kleinen Stichproben
- aber nur für feststehende Randsummen geeignet!
- ohne feste Randsummen: Barnards Test

Fisher-Exakt in R:

```
> fisher.test(my.matrix)
> fisher.test(my.vector.1, my.vector.2)
```

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!

	haben	sein
nord	27	33
sued	3	34

$\chi^2 =$

	haben	sein
nord	54	66
sued	6	68

$\chi^2 =$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 =$

	haben	sein
nord	54	66
sued	6	68

$\chi^2 =$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 =$

	haben	sein
nord	54	66
sued	6	68

$\chi^2 =$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 = 12{,}89$

	haben	sein
nord	54	66
sued	6	68

$\chi^2 =$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 = 12{,}89$

	haben	sein
nord	54	66
sued	6	68

$\chi^2 = 27{,}46$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 = 12{,}89$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

	haben	sein
nord	54	66
sued	6	68

$\chi^2 = 27{,}46$

Effektstärke

Der χ^2-Wert sagt nichts über die Stärke eines Zusammenhangs!
Bei höheren absoluten Frequenzen wird auch der χ^2-Wert größer.

	haben	sein
nord	27	33
sued	3	34

$\chi^2 = 12{,}89$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

	haben	sein
nord	54	66
sued	6	68

$\chi^2 = 27{,}46$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

Effektstärke II

Pearsons ϕ: Maß für die Stärke des Zusammenhangs in 2×2-Tabellen

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Effektstärke II

Pearsons ϕ: Maß für die Stärke des Zusammenhangs in 2×2-Tabellen

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ ist eine Zahl zwischen 0 und 1:
Je größer, desto stärker der Zusammenhang zwischen den Variablen.

Effektstärke II

Pearsons ϕ: Maß für die Stärke des Zusammenhangs in 2×2-Tabellen

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ ist eine Zahl zwischen 0 und 1:
Je größer, desto stärker der Zusammenhang zwischen den Variablen.

Beispiel: $\phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{12.89}{97}} = 0.3648$

Cramérs v

Cramérs v für $n \times n$-Tabellen mit $n > 2$ oder $m > 2$

$$v = \sqrt{\frac{\frac{\chi^2}{n}}{min(s-1,z-1)}}$$

mit: s die Spaltenzahl und z die Zeilenzahl

Beachte: für 2×2-Tabellen: $s - 1 = 1$ und $z - 1 = 1$,

also $min(s-1, z-1) = 1$

daher: $v = \sqrt{\frac{\frac{\chi^2}{n}}{1}} = \sqrt{\frac{\chi^2}{n}} = \phi$

In R

Speichern des Test-Objekts:

```
> my.chi2.test <- chisq.test(my.matrix)
```

Speichern des χ^2-Werts mit:

```
> my.chi2.value <- as.numeric(my.chi2.test$statistic)
```

Speichern von n:

```
> my.n <- sum(my.matrix)
```

Also Effektstärke (mit Ausgabe):

```
> my.phi <- sqrt( my.chi2.value / my.n ); my.phi
```

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

und damit

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

und damit

$$p(E) = \frac{o(E)}{1+o(E)}$$

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

  $$o(E) = \frac{p(E)}{1-p(E)}$$

  und damit

  $$p(E) = \frac{o(E)}{1+o(E)}$$

- Ein Ereignis ist in Korpusstudien i. d. R. das Auftreten einer Variablenausprägung.

Chance (odds)

- Die Chance (odds) o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

und damit

$$p(E) = \frac{o(E)}{1+o(E)}$$

- Ein Ereignis ist in Korpusstudien i. d. R. das Auftreten einer Variablenausprägung.
- Die Information in den Maßen Wahrscheinlichkeit und Chance ist dieselbe (s. Umrechenbarkeit ineinander).

Chance und Wahrscheinlichkeit und Zähldaten

Aux	**Anzahl**
haben	27
sein	33

Chance und Wahrscheinlichkeit und Zähldaten

Aux	**Anzahl**
haben	27
sein	33

$p(haben) = \frac{27}{27+33} = \frac{27}{60} = 0.45$ (Wahrscheinlichkeit)

Chance und Wahrscheinlichkeit und Zähldaten

Aux	Anzahl
haben	27
sein	33

$p(haben) = \frac{27}{27+33} = \frac{27}{60} = 0.45$ (Wahrscheinlichkeit)

$1 - p(haben) = p(\neg haben) = \frac{33}{27+33} = \frac{33}{60} = 0.55$ (Gegenwahrscheinlichkeit)

Chance und Wahrscheinlichkeit und Zähldaten

Aux	Anzahl
haben	27
sein	33

$p(haben) = \frac{27}{27+33} = \frac{27}{60} = 0.45$ (Wahrscheinlichkeit)

$1 - p(haben) = p(\neg haben) = \frac{33}{27+33} = \frac{33}{60} = 0.55$ (Gegenwahrscheinlichkeit)

Beachte: $p(haben) + p(\neg haben) = 1$

Chance und Wahrscheinlichkeit und Zähldaten

Aux	Anzahl
haben	27
sein	33

$p(haben) = \frac{27}{27+33} = \frac{27}{60} = 0.45$ (Wahrscheinlichkeit)

$1 - p(haben) = p(\neg haben) = \frac{33}{27+33} = \frac{33}{60} = 0.55$ (Gegenwahrscheinlichkeit)

Beachte: $p(haben) + p(\neg haben) = 1$

$o(haben) = \frac{\frac{27}{60}}{\frac{33}{60}} = \frac{27}{60} \cdot \frac{60}{33} = \frac{27}{33} = 0.82$

Chance und Wahrscheinlichkeit und Zähldaten

Aux	Anzahl
haben	27
sein	33

$p(haben) = \frac{27}{27+33} = \frac{27}{60} = 0.45$ (Wahrscheinlichkeit)

$1 - p(haben) = p(\neg haben) = \frac{33}{27+33} = \frac{33}{60} = 0.55$ (Gegenwahrscheinlichkeit)

Beachte: $p(haben) + p(\neg haben) = 1$

$o(haben) = \frac{\frac{27}{60}}{\frac{33}{60}} = \frac{27}{60} \cdot \frac{60}{33} = \frac{27}{33} = 0.82$

allgmein: $p(E) = \frac{Anzahl(E)}{Anzahl(E)+Anzahl(\neg E)}$ und $o(E) = \frac{Anzahl(E)}{Anzahl(\neg E)}$

Chancenverhältnis (odds ratio)

- Das Chancenverhältnis (odds ratio) gibt das Verhältnis an, wie sich die Chancen einer Variablenausprägung E unter Bedingung A – also $o(E|A)$ – und unter Bedingung B – also $o(E|B)$ – zueinander Verhalten:

$$r(E|A, E|B) = \frac{o(E|A)}{o(E|B)}$$

Beispiel zum Chancenverhältnis (1)

- Wir haben Texte aus Süddeutschland und Norddeutschland auf das Auftreten des Perfektauxiliars *haben* und *sein* bei bestimmten Verben untersucht.
- Die Kreuztabelle:

	nord	sued
haben	27	3
sein	33	34

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) =$
- $o(haben|sued) =$

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) =$

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen:

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen: $or = \frac{0.82}{0.09} = 9.11$

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen: $or = \frac{0.82}{0.09} = 9.11$
- D. h. die Chance von *haben* ist 9.11 mal größer, wenn die Region *nord* ist.

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen: $or = \frac{0.82}{0.09} = 9.11$
- D. h. die Chance von *haben* ist 9.11 mal größer, wenn die Region *nord* ist.
- Ersatz für Effektstärke bei Fisher-Test

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*."

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*."
  - „frei alternieren" = beide Kasus haben den gleichen Anteil.

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*."
  - „frei alternieren" = beide Kasus haben den gleichen Anteil.
  - Grundgesamtheit per Null-Hypothese: 50% Genitive und 50% Dative

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*."
  - „frei alternieren" = beide Kasus haben den gleichen Anteil.
  - Grundgesamtheit per Null-Hypothese: 50% Genitive und 50% Dative
- Korpusstichprobe: F(Genitiv)=41 und F(Dativ)=59

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*.“
  - „frei alternieren“ = beide Kasus haben den gleichen Anteil.
  - Grundgesamtheit per Null-Hypothese: 50% Genitive und 50% Dative
- Korpusstichprobe: F(Genitiv)=41 und F(Dativ)=59
- Stimmt das mit der Null überein bei $sig = 0.05$?

Binomial-Test

Null: Es gibt keine Abweichung
von den erwarteten gleich großen Anteilen.

Binomial-Test

NULL: Es gibt keine Abweichung
von den erwarteten gleich großen Anteilen.

NULL: $p(Dativ) = 0.5$ (p für proportion)

Binomialtest im Einzelnen

Benötigte Größen:

- Stichproben der Größe n

Binomialtest im Einzelnen

Benötigte Größen:

- Stichproben der Größe n
- Proportion p (hier $p = 0.5$)

Binomialtest im Einzelnen

Benötigte Größen:

- Stichproben der Größe n
- Proportion p (hier $p = 0.5$)
- Anzahl der beobachteten Ereignisse: X (hier $X(Dativ) = 59$)

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:
  - Mittel: $\mu = p \cdot n$

Unter Annahme der NULL…

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:
  - Mittel: $\mu = p \cdot n$
  - Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:
  - Mittel: $\mu = p \cdot n$
  - Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$
  - Wir können für den gemessenen Wert den z-Wert ausrechnen.

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:
  - Mittel: $\mu = p \cdot n$
  - Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$
  - Wir können für den gemessenen Wert den z-Wert ausrechnen.

Unter Annahme der NULL...

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
  approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der NULL!) für die Normalverteilung:
  - Mittel: $\mu = p \cdot n$
  - Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$
  - Wir können für den gemessenen Wert den z-Wert ausrechnen.

$$z = \frac{X - \mu}{s} = \frac{X - p \cdot n}{\sqrt{n \cdot p \cdot (1 - p)}}$$

Ausrechnen des Beispiels und Signifikanz

$$z = \frac{59-(0.5\cdot100)}{\sqrt{100\cdot0.5\cdot0.5}} = \frac{59-50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

Ausrechnen des Beispiels und Signifikanz

$$z = \frac{59 - (0.5 \cdot 100)}{\sqrt{100 \cdot 0.5 \cdot 0.5}} = \frac{59 - 50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom Null-Mittel entfernt.

Ausrechnen des Beispiels und Signifikanz

$$z = \frac{59-(0.5\cdot 100)}{\sqrt{100\cdot 0.5\cdot 0.5}} = \frac{59-50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom Null-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und $sig = 0.05$: $-1.96..1.96$

Ausrechnen des Beispiels und Signifikanz

$$z = \frac{59-(0.5\cdot100)}{\sqrt{100\cdot0.5\cdot0.5}} = \frac{59-50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom Null-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und $sig = 0.05$: $-1.96..1.96$
- Die Null kann also nicht zurückgewiesen werden bei $sig = 0.05$.

Ausrechnen des Beispiels und Signifikanz

$$z = \frac{59-(0.5 \cdot 100)}{\sqrt{100 \cdot 0.5 \cdot 0.5}} = \frac{59-50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom NULL-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und $sig = 0.05$: $-1.96..1.96$
- Die NULL kann also nicht zurückgewiesen werden bei $sig = 0.05$.
- Interpretation: Entweder ist die Variation nicht genau gleich verteilt oder ein seltenes Ereignis ist eingetreten.

In R

```
> binom.test(59, 100, 0.5)

      Exact binomial test

data: 59 and 100
number of successes = 59, number of trials = 100, p-value = 0.08863
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
  0.4871442  0.6873800 sample estimates:
probability of success 0.59
```

Nächste Woche | Überblick

Einzelthemen

1. Inferenz
2. Deskriptive Statistik
3. Nichtparametrische Verfahren
4. z-Test und t-Test
5. ANOVA
6. Freiheitsgrade und Effektstärken
7. Power und Severity
8. Lineare Modelle
9. Generalisierte Lineare Modelle
10. Gemischte Modelle

Literatur I

Bortz, Jürgen & Gustav Lienert. 2008. *Kurzgefasste Statistik für die klinische Forschung*. Heidelberg: Springer.

Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.

Autor

Kontakt

Prof. Dr. Roland Schäfer
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena
Fürstengraben 30
07743 Jena

https://rolandschaefer.net
roland.schaefer@uni-jena.de

Lizenz