Statistik
07. Freiheitsgrade und Effektstärken

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

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

Inhalt

Freiheitsgrade

Freiheitsgrade „intuitiv“

- Beispiel: Schätzung eines Parameters (z. B. Mittel) auf Basis von 1000 gemessenen Werten
- Wenn 999 Werte bekannt sind, steht abhängig vom Mittel der 1000ste Wert fest.

- Für jedes Mittel μ einer Stichprobe mit n Messungen sind also nur $n - 1$ frei wählbar.

(Unintuitive) Erweiterung(en)

- generell: $df = n - |E|$
  wobei E die zu schätzenden Parameter sind. $|E|$ ist ihre Anzahl.
- Warum bei χ^2 dann $df = (Zeilenzahl - 1) \cdot (Spaltenzahl - 1)$?

- Bsp.: Tabelle mit 2×3 Feldern, also $df = (2-1)(3-1) = 1 \cdot 2 = 2$...
- Bei bekannten Randsummen sind aber tatsächlich nur 2 Felder frei wählbar!

	X1	X2	
Y1	⊕		ZS1
Y2	⊕		ZS2
Y3			ZS3
	SQ1	SQ2	

Mehr zu Zähldatentests

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

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

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:

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

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

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

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

Bernoulli-Experimente

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*."
  - „frei alternieren" = beide Kasus haben die gleiche Chance.
  - Grundgesamtheit per Hypothese: 50% Genitive und 50% Dative
- Korpusstichprobe: F(Genitiv)=41 und F(Dativ)=59
- Passt das zur Hypothese bei sig=0.05?

Binomialtest

- H0: Es gibt keine Abweichung von der erwarteten Wahrscheinlichkeit.
- H0: $p(Dativ) = 0.5$

Binomialtest im Einzelnen

Benötigte Größen:

- Stichproben der Größe n
- H0-Wahrscheinlichkeit p (hier $p = 0.5$)
- Anzahl der beobachteten Ereignisse: X (hier $X(Dativ) = 59$)

Unter Annahme der H0…

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

- Es gilt dann (unter Annahme der H0!) 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

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

- Der gemessene Wert liegt 1.8 Standardabweichungen vom H0-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und sig=0.05: $-1.96..1.96$
- Die H0 kann also nicht zurückgewiesen werden bei sig=0.05.

- Interpretation: Wir haben keine Evidenz dafür, dass die Variation in der Grundgesamtheit von einer 50:50-Verteilung abweicht.

- Falsche Interpretation: Wir haben Evidenz dafür, dass die Verteilung in der Grundgesamtheit 50:50 ist.

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
```

Effektstärken bei t-Test und ANOVA

Effektstärke Ein-Stichproben-t-Test

- Signifikanz ≠ starker Effekt
- Effektstärke beim t-Test für Stichprobe x:

$$\text{Cohens } d = \frac{\bar{x}-\mu}{s(x)}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

Erklärung der Varianz

- ähnlich der Effektstärke:
  Welcher Anteil der Varianz in den Daten
  wird durch die Unabhängige erklärt?

$$\text{Cohens } r^2 = \frac{t^2}{t^2+df}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

Effektstärke Zwei-Stichproben-t-Test

Effektstärke

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2}}$$

Erklärung der Varianz

$$r^2 = \frac{t^2}{t^2 + df}$$

Effektstärke einfaktorielle ANOVA

$$\eta^2 = \frac{SQ_{zwischen}}{SQ_{gesamt}}$$

(wieder ein r^2-Maß)

Effektstärken bei der zweifaktoriellen ANOVA

Entsprechend sind drei η^2 auszurechnen:

$$\eta^2_A = \frac{SQ_A}{SQ_{gesamt} - SQ_B - SQ_{A \times B}}$$

$$\eta^2_B = \frac{SQ_B}{SQ_{gesamt} - SQ_A - SQ_{A \times B}}$$

$$\eta^2_{A \times B} = \frac{SQ_{A \times B}}{SQ_{gesamt} - SQ_A - SQ_B}$$

Wir fragen jeweils, welchen Anteil an der Varianz,
die die anderen beiden Faktoren nicht erklären,
der jeweilige dritte Faktor hat.

Voraussetzungen für t-Test und ANOVA

Caveat

Bedingung für alle Tests:
Unabhängigkeit der Messungen

Wenn bei t-Test oder ANOVA also gepaarte Stichproben vorliegen
(Messung derselben Proband*innen unter Bedingung 1 und 2 usw.):
Besondere Versionen für geparte Stichproben nehmen!

Details hier nicht besprochen.

Voraussetzungen prüfen I

Die GGs müssen normalerverteilt sein:

```
shapiro.test(x)
```

Wenn $p \leq 0.05$ wird die Nullhypothese des Shapiro-Wilk-Tests verworfen.
H0: Die Werte stammen aus einer normalverteilten GG.

Die Varianzen müssen homogen sein:

```
var.test(x1, x2)
```

Auch hier: $p \leq 0.05$ weist die H0 zurück.
H0: Die Varianzen von x1 und x2 sind homogen.

Solche Tests sind umstritten, weil sie angeblich zu empfindlich reagieren.
Zuur u. a. 2009 empfehlen z. B. grafische Methoden. Ich nicht.

Voraussetzungen prüfen II

Wenn Voraussetzungen nicht erfüllt sind:

- steigt das Risiko für Typ 1-Fehler

- nicht-parametrische Alternative nehmen
- Daten transformieren (Logarithmus für Normalverteilung)

- sich über Robustheit des Test ggü. verletzten Annahmen informieren (oft schwer zugängliche und kontroverse Spezialliteratur)

Alternativen

Übersicht

- Alternativen, wenn Bedingungen für t-Test und ANOVA nicht erfüllt sind (Normalverteilung, Varianzhomogenität)

- Prinzip: Umrechnen von Werten in Ränge
- nicht-parametrische Tests

Literatur

- Bortz & Lienert 2008
- Gravetter & Wallnau 2007

Übersicht

- Mann-Whitney U-Test: Alternative zum t-Test mit zwei Stichproben
- Kruskal-Wallis H-Test: Alternative zur einfaktoriellen ANOVA

Wiederholung: Bedingungen für t-Test

- Intervallskalierung der Abhängigen
- Normalität der Abhängigen
- Varianzhomogenität der Abhängigen in den Gruppen
- Unabhängigkeit der Messungen

Alle bis auf die letzte entfallen beim Mann-Whitney U-Test.

Direkte Berechnung beim MWU

Gruppen/Stichproben (Messwerte):
$x_1 = [9, 8, 12, 16]$
$x_2 = [4, 11, 7, 13]$

Ränge in der zusammengelegten Stichprobe:
$X = [4, 7, 8, 9, 11, 12, 13, 16]$
$R(x_1) = [4, 3, 6, 8]$
$R(x_2) = [1, 5, 2, 7]$

Addiere für jeden Wert beider Gruppen die Anzahl der niedrigeren Ränge (=höhere Rangzahl!) in der anderen Gruppe:
$U(x_1) = 2 + 2 + 1 + 0 = 5$
$U(x_2) = 4 + 2 + 4 + 1 = 11$
$U = min(U_{x_1}, U_{x_2}) = U_{x_1} = 5$

Allgemeine Formel

$$U(x_\alpha) = n_1 \cdot n_2 + \frac{n_\alpha(n_\alpha+1)}{2} - \sum R(x_\alpha)$$

- $\sum R(x_1) = 4 + 3 + 6 + 8 = 21$
- $\sum R(x_2) = 1 + 5 + 2 + 7 = 15$
- $n_1 \cdot n_2 = 4 \cdot 4 = 16$
- $n_1(n_1 + 1) = n_2(n_2 + 1) = 4 \cdot 5 = 20$
- $U(x_1) = 16 + 10 - 21 = 5$
- $U(x_2) = 16 + 10 - 15 = 11$
- $U = 5$

Siginifikanz und Effektstärke

- Signifikanz für kleine Stichproben: Tabelle
- bei großen Stichproben: U ugf. normalverteilt, also z-Test
- in R:

```
> wilcox.test(x1,x2, paired = FALSE)
```

- Effektstärke: Punkt-biserielle Korrelation
- entspricht Pearson-Korrelation, aber Unabhängige ist dichotom
- In R: `cor(c(x1,x2), c(rep(0,4),rep(1,4)))`

- alternativ: „relativer Effekt“ (Bortz & Lienert, S. 142)

Probleme

- Bei sehr vielen gleichen Rängen ist der Mann-Whitney U-Test unzuverlässig.
- Bei gleichen Rängen generell: korrigierte Version (s. Bortz & Lienert, S. 146).

- Er ist daher nur begrenzt geeignet für Dinge wie 5-Punkt-Skalen.
- generell am stärksten bei gleich großen und gleich stark streuenden Stichproben

- letzter Ausweg: Mediantest (Bortz & Lienert, S. 137)

Mehr als zwei Gruppen

Wie vom t-Test zur ANOVA...

$x_1 = [9, 8, 12, 16]$
$x_2 = [4, 11, 7, 13]$
$x_3 = [13, 12, 5, 15]$

Gleiches Vorgehen wie bei Mann-Whitney über
Rang in der zusammengelegten Stichprobe:

<table>
<tr><td>X</td><td>4</td><td>5</td><td>7</td><td>8</td><td>9</td><td>11</td><td>12</td><td>12</td><td>13</td><td>13</td><td>15</td><td>16</td></tr>
<tr><td>R(X)</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td colspan="2">7.5</td><td colspan="2">9.5</td><td>11</td><td>12</td></tr>
</table>

$R(x_1) = [5, 4, 7.5, 12]$
$R(x_2) = [1, 6, 3, 9.5]$
$R(x_3) = [9.5, 7.5, 2, 11]$

Berechnung des Kruskal-Wallis H-Werts

$$H = \frac{12}{N(N+1)} \cdot \sum_i \frac{(\sum R(x_i))^2}{n_i} - 3(N+1)$$

Am Beispiel:

- Gruppen-Rang-Summen:
  - $R(x_1) = [5, 4, 7.5, 12]$, $\sum R(x_1) = 28.5$
  - $R(x_2) = [1, 6, 3, 9.5]$, $\sum R(x_2) = 19.5$
  - $R(x_3) = [9.5, 7.5, 2, 11]$, $\sum R(x_3) = 30$
- $H = \frac{12}{12 \cdot (12+1)} \cdot (\frac{28.5^2}{4} + \frac{19.5^2}{4} + \frac{30^2}{4}) - 3(12+1) =$
- $0.077 \cdot (203.06 + 95.06 + 225) - 39 = 1.28$

Signifikanztest

- Bei $n > 5$ ist H unter der H0 χ^2-verteilt.
- mit $df = k - 1$ (k ist die Anzahl der Gruppen)

- Effektstärke: tja...
- „relative Effekte“ sind rechenbar (Bortz & Lienert, S. 159)

```
> kruskal.test(c(x1,x2,x3) c(rep(0,4),rep(1,4),rep(2,4)))
```

Rechnen Sie bitte mal die U- und H-Tests von diese Folien
und vergleichen Sie die p-Werte mit denen von t-Test und ANOVA
über die gleichen Daten:

$x_1 = [9, 8, 12, 16]$
$x_2 = [4, 11, 7, 13]$
$x_3 = [13, 12, 5, 15]$

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.

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.

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