Einführung Inferenz kategoriale Werte

Karsten Lübke

Globaler Index der Religiösität und Atheismus

Im Jahre 2012 führte das WIN/Gallup International Institut in 57 Ländern eine Untersuchung zur Religiosität durch. Die Pressemitteilung zur Studie finden Sie hier.

Dabei wurde die Frage gestellt: *Unabhängig davon, ob Sie an Gottesdiensten teilnehmen oder nicht ("attend place of worship"), würden Sie sagen, dass Sie ein religiöser Mensch sind, eine nicht religiöse Person oder ein überzeugter Atheist?*

Die Befragten hatten dabei drei Antwortmöglichkeiten: Religiöser Mensch ("A religious person"), Nicht religiöser Mensch ("Not a religious person"), und Atheist ("A convinced atheist"). Die Befragten klassifizierten sich, es wurden als kategorielle (nominale) Daten (`factor`) erzeugt.

Übung:

1. Handelt es sich bei den im Bericht angegeben Kennzahlen um *Stichprobenstatistiken* oder um *Populationsparameter*?
2. Um die Ergebnisse der Studie zu verallgemeinern, also auf die Gesamtbevölkerung zu schließen, welche Annahmen müssen dafür erfüllt sein und klingen diese hier plausibel erfüllt?

Ein Teil der Daten kann direkt von OpenIntro als R Datensatz heruntergeladen werden, und anschließend in R eingelesen:

```
meine_url <- "http://www.openintro.org/stat/data/atheism.RData"
load(url(meine_url)) # Einlesen
```

Einen Überblick erhält man wie immer über:

```
str(atheism) # Datenstruktur
```

```
## 'data.frame':    88032 obs. of  3 variables:
##  $ nationality: Factor w/ 57 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ response   : Factor w/ 2 levels "atheist","non-atheist": 2 2 2 2 2 2 2 2 2 2 ...
##  $ year       : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
```

```
head(atheism) # Erste Beobachtungen
```

```
##   nationality    response year
## 1 Afghanistan non-atheist 2012
## 2 Afghanistan non-atheist 2012
## 3 Afghanistan non-atheist 2012
## 4 Afghanistan non-atheist 2012
## 5 Afghanistan non-atheist 2012
## 6 Afghanistan non-atheist 2012
```

```
tail(atheism) # letzte Beobachtungen
```

```
##       nationality    response year
## 88027     Vietnam non-atheist 2005
## 88028     Vietnam non-atheist 2005
## 88029     Vietnam non-atheist 2005
## 88030     Vietnam non-atheist 2005
## 88031     Vietnam non-atheist 2005
## 88032     Vietnam non-atheist 2005
```

Zur Analyse wird wieder das Paket mosaic verwendet:

```
library(mosaic)
```

Inferenz eines Anteilswerts

In Tabelle 6 der Pressemitteilung wird der Anteil der Atheisten für Deutschland mit 15% angegeben. Dies ist eine *Statistik* der Stichprobe, nicht der Parameter der *Population*. Es wird also die Frage beantwortet "Wie hoch ist der Anteil der Atheisten in der Stichprobe?". Um die Frage "Wie hoch ist der Anteil der Atheisten in der Population?" zu beantworten, muss von der Stichprobe auf die Population geschlossen werden, d. h., es wird z. B. der Anteilswert *geschätzt*.

Der folgende Befehl filtert den Datensatz auf das Ergebnis für Deutschland im Jahr 2012, d. h., es werden nur die gewünschten Zeilen im Datensatz belassen:

```
de12 <- filter(atheism, nationality == "Germany", year == "2012")
```

Die *Punktschätzung* des Anteilswertes der Atheisten für Deutschland im Jahr 2012 liegt dann bei

```
pdach <- tally(~response, data=de12, format='proportion')["atheist"]
pdach
```

```
##   atheist
## 0.1494024
```

also bei 15%.

Um jetzt ein 95% Konfidenzintervall für den Populationsparameter zu konstruieren (*Bereichsschätzung*) muss der Standardfehler *se* bestimmt werden, hier:

```
n <- nrow(de12) # Anzahl Beobachtungen
se <- sqrt( pdach * (1-pdach) / n)
se
```

```
##    atheist
## 0.01591069
```

Der Standardfehler, d. h., die Standardabweichung des Anteilswertes liegt hier also bei 1.59%. Zusammen mit dem 2,5% und 97,5% Quantil der Standardnormalverteilung ergibt sich folgendes Konfidenzintervall:

```
pdach + qnorm(c(0.025, 0.975)) * se
```

```
## [1] 0.1182180 0.1805868
```

Da der Populationsparameter unbekannt aber nicht zufällig ist, werden die 95% auch als *Überdeckungswahrscheinlichkeit* bezeichnet.

In `mosaic` besteht die Möglichkeit, Punkt- und Bereichsschätzungen mit dem Befehl `prop.test` durchzuführen. Sofern kein Wert für `p` angegeben wird lautet die Nullyhpothese $H_0 : p = 0.5$.

```
prop.test(~response, data=de12)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  de12$response  [with success = atheist]
## X-squared = 245.42, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1199815 0.1843172
## sample estimates:
##         p
## 0.1494024
```

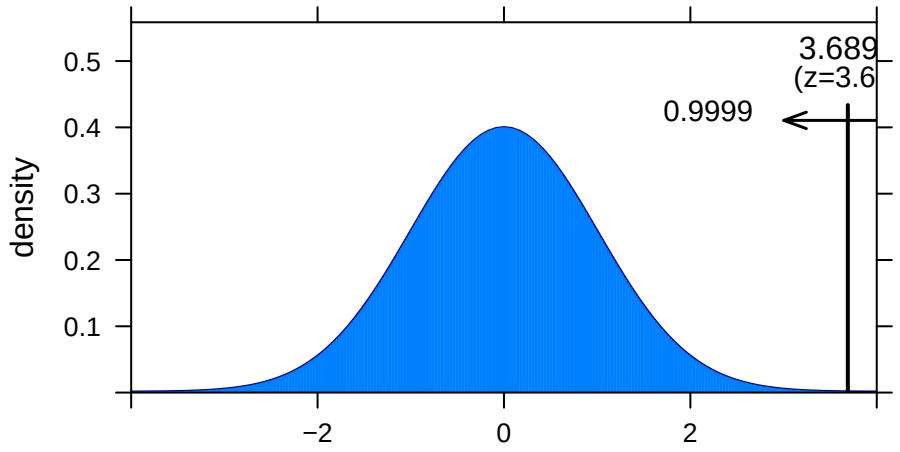
Übung:

3. Bei annähernd gleicher Stichprobengröße liegt der Anteil der Atheisten in Saudi Arabien bei 5%. Wie verändert sich der Standardfehler und damit die Breite des Konfidenzintervalls?
4. Der Anteil der Atheisten in Südkorea liegt in etwa ebenfalls bei 15%, allerdings liegen die Daten von 1523 Befragten vor. Wie verändert sich der Standardfehler und damit die Breite des Konfidenzintervalls?

Um für Deutschland die Nullhypothese "Der Anteil der Atheisten liegt nicht über 10%" gegen die Alternativhypothese (Forschungshypothese) "Der Anteil der Atheisten liegt über 10%" können entweder wieder Simulations- und Resamplingtechniken verwendet werden oder die Approximation durch die Normalverteilung :

```
se0 <- sqrt( (0.1 * (1-0.1)) /n)
z <- (pdach - 0.10) / se0
xpnorm(z, lower.tail = FALSE)
```

```
##
## If X ~ N(0, 1), then
##
##  P(X <= 3.689594) = P(Z <= 3.689594) = 0.9998877
##  P(X >  3.689594) = P(Z >  3.689594) = 0.0001123062
```




```
##      atheist
## 0.0001123062
```

Der *p-Wert* liegt also bei 0.0112%, die Nullhypothese wird also zum Signifikanzniveau von 5% verworfen.

Auch hier direkt über `prop.test`:

```
prop.test(~response, p=0.1, alternative="greater", data=de12)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  de12$response  [with success = atheist]
## X-squared = 13.07, df = 1, p-value = 0.0001501
## alternative hypothesis: true p is greater than 0.1
## 95 percent confidence interval:
##  0.1241944 1.0000000
## sample estimates:
##         p
## 0.1494024
```

Je nach Alternativhypothese ergeben sich unterschiedliche Tests:

- `alternative="two.sided"`: ungerichtet, d. h. zweiseitig: $H_0 : p = 0.1$ gegen $H_A : p \neq 0.1$

- `alternative="less"`: gerichtet, d. h. einseitig: $H_0 : p \geq 0.1$ gegen $H_A : p < 0.1$
- `alternative="greater"`: gerichtet, d. h. einseitig: $H_0 : p \leq 0.1$ gegen $H_A : p > 0.1$

Differenz zweier Anteilswerte

In den Daten liegen außerdem die Ergebnisse aus 2005 vor:

```
de05 <- filter(atheism, nationality == "Germany" & year == "2005")
```

Im Jahre 2005 lag der Anteil der Atheisten in Deutschland bei

```
tally(~response, data=de05, format="proportion")
```

```
## response
##     atheist non-atheist
##  0.09960159  0.90039841
```

Der Aneil lag also bei unter 10% – in der *Stichprobe*! Können wir daraus auf eine Erhöhung des Anteils von 2005 zu 2012 in der *Population* schließen?

```
# 2012
a12 <- tally(~response, data=de12)["atheist"] # Anzahl Atheisten 2012
n12 <- nrow(de12) # Anzahl Studienteilnehmer 2012
p12 <- a12/n12 # Anteil Atheisten 2012
# 2005
a05 <- tally(~response, data=de05)["atheist"] # Anzahl Atheisten 2005
n05 <- nrow(de05) # Anzahl Studienteilnehmer 2005
p05 <- a05/n05 # Anteil Atheisten 2005
# Punktschätzer Differenz Population
pdiff <- p12-p05
pdiff
```

```
##   atheist
## 0.0498008
```

```
# Pooling zur Berechnen Standardfehler unter H_0
ppool <- (a12 + a05)/(n12+n05)
ppool
```

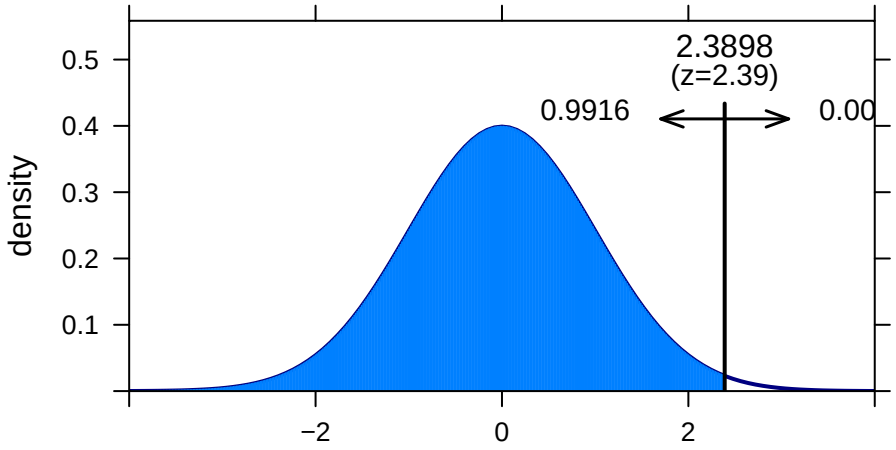
```
##  atheist
## 0.124502
```

```
# Standardfehler se
se0 <- sqrt( (ppool * (1-ppool) / n12) + (ppool * (1-ppool) / n05) )
se0
```

```
##   atheist
## 0.0208391
```

```
# z-Wert z
z <- (pdiff - 0)/se0
# p-Wert
xpnorm(z, lower.tail = FALSE)
```

```
##
## If X ~ N(0, 1), then
##
##  P(X <= 2.389777) = P(Z <= 2.389777) = 0.9915707
##  P(X >  2.389777) = P(Z >  2.389777) = 0.008429297
```




```
##     atheist
## 0.008429297
```

Der p-Wert ist klein und das Ergebnis damit *statistisch signifikant*. Die Wahrscheinlichkeit *zufällig* eine solche Erhöhung der Anteilswerte zu beobachten ist also gering – wenn die H_0 gilt! d. h. es wird auf eine Veränderung des Anteilswertes in der *Population* geschlossen.

Auch dies kann direkt durch den Befehl `prop.test` gestestet werden. Dazu wird zunächst ein gemeinsamer Datensatz erzeugt:

```
de <- filter(atheism, nationality == "Germany")
```

und anschließend geteset:

```
prop.test(response~year, alternative="less", data=de)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tally(response ~ year)
## X-squared = 5.2633, df = 1, p-value = 0.01089
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.01362913
## sample estimates:
##     prop 1     prop 2
## 0.09960159 0.14940239
```

Hier wird die Alternativhypothese `less` verwendet: $H_0 : p_1 \geq p_2$, gegen $H_A : p_1 < p_2$.

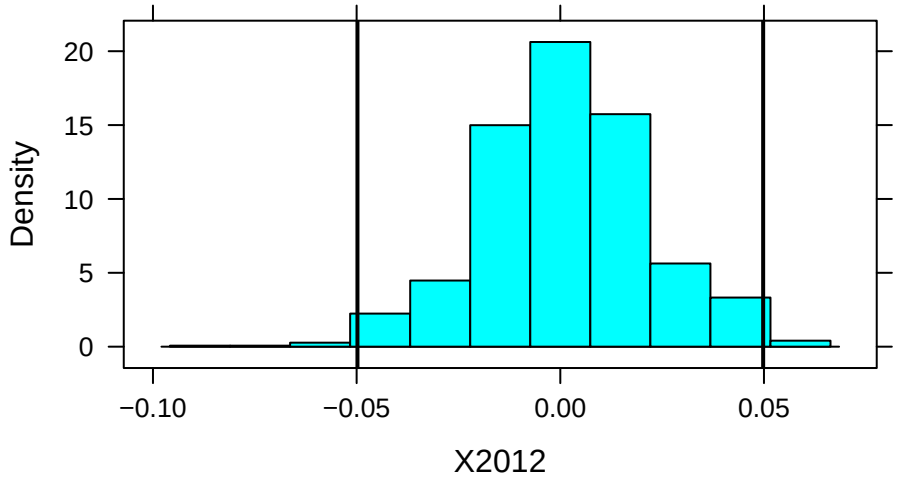
Exkurs: Mit dem Paket mosaic können Sie das auch einfach über Permutationen testen, indem das Erhebungsjahr zufällig gesampelt wird, wobei hier ungerichtet, d. h., zweiseitig getestet wird. Mit anderen Worten ist sowohl eine Erhöhung als auch ein Verringerung des Anteils in der Alternativhypothese möglich, daher werden die Absolutwerte der Differenz (`abs()`) verwendet:

```
# Beobachtete Differenz
pdiff <- diff(tally( ~ response | year, data=de, format="proportion")["atheist",])
pdiff
```

```
##      2012
## 0.0498008
```

```
# Zufallszahlengenerator setzen (Reproduzierbarkeit!)
set.seed(1896)
# 1000-mal das Jahr permutieren: Nullhypothese kein Unterschied
pdiff.null <- do(1000) * diff(tally( ~ response | sample(year), data=de,
                                     format="proportion")["atheist",])
```

```r
# Histogramm
histogram(~ X2012, data=pdiff.null, v=c(pdiff, -pdiff))
```




```r
# p-Wert
mean(abs(pdiff.null$X2012) >= abs(pdiff))
```

```
## [1] 0.02
```

Übung:

5. Überprüfen Sie für das Jahr 2012, ob es eine zum Niveau 5% signifikante Differenz zwischen den Anteil der Atheisten in Deutschland und den Niederlanden (`Netherlands`) in der Population gibt.
6. Überprüfen Sie für das Jahr 2012, ob es eine zum Niveau 5% signifikante Differenz zwischen den Anteil der Atheisten in Deutschland und Polen (`Poland`) in der Population gibt.

Chi-Quadrat Unabhängigkeitstest

Soll allgemein der Zusammenhang zwischen zwei kategoriellen (nominalen) Variablen untersucht werden, wird der Chi²-Unabhängigkeitstest verwendet. Diese testet die Nullhypothese der Unabhängigkeit, gegen die Alternativhypothese des Zusammenhangs. Im vorliegenden Datensatz können wir z. B. testen, ob die Verteilung (Anteil) der Atheisten in den teilnehmenden G7 Ländern gleich ist:

```r
G7 <- c("United States", "Canada", "Germany", "France", "Italy", "Japan")
G7.12 <- filter(atheism, nationality %in% G7 & year == 2012)
G7.12 <- droplevels(G7.12)
G7atheist <- tally(response ~ nationality, data = G7.12)
G7atheist
```

```
##              nationality
## response      Canada France Germany Italy Japan United States
##   atheist         90    485      75    79   372            50
##   non-atheist    912   1203     427   908   840           952
```

(Der Befehl `droplevels` sorgt dafür, dass die nicht mehr benötigten Ausprägungen der kategoriellen Variablen (`factor`) gelöscht werden.)

Der Test selber erfolgt in mosaic über `xchisq.test`, d. h.:

```r
xchisq.test(G7atheist)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  x
## X-squared = 504.04, df = 5, p-value < 2.2e-16
##
##     90      485       75       79      372       50
## ( 180.40) ( 303.91) (  90.38) ( 177.70) ( 218.21) ( 180.40)
## [ 45.30] [107.91] [  2.62] [ 54.82] [108.39] [ 94.26]
## <-6.73>  <10.39>  <-1.62>  <-7.40>  <10.41>  <-9.71>
##
##    912     1203      427      908      840      952
## ( 821.60) (1384.09) ( 411.62) ( 809.30) ( 993.79) ( 821.60)
## [  9.95] [ 23.69] [  0.57] [ 12.04] [ 23.80] [ 20.70]
## < 3.15>  <-4.87>  < 0.76>  < 3.47>  <-4.88>  < 4.55>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

Der Wert der Teststatistik Chi² liegt bei 504.04, die Anzahl der Freiheitsgrade ("degrees of freedom", df) bei 5, der p-Wert ist sehr klein, die Nullhypothese der Unabhängigkeit von Nationalität und Verteilung Atheismus wird für die *Population* verworfen.

Die Formel zur Berechnung der χ^2 Teststatistik lautet:

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

mit $e_{ij} = \frac{h_{i\cdot}h_{\cdot j}}{n}$ wobei für die Zeilensumme $h_{i\cdot} = \sum_{j=1}^{m} h_{ij}$ und für die Spaltensumme $h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$ gilt.

Mit diesen Daten kann auch über den Befehl `pchisq` der p-Wert berechnet werden:

```
pchisq(504, 5, lower.tail = FALSE)
```

```
## [1] 1.093548e-106
```

Übung:

7. Gibt es einen Zusammenhang zwischen der Verteilung des Atheismus und der Nationalität im Jahr 2012 innerhalb der afrikanischen Länder `c("Nigeria","Kenya", "Tunisia", "Ghana", "Cameroon", "South Sudan")`?

Übung:

Wir werden jetzt den *tips* Datensatz aus *Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing* weiter analysieren.

Sofern noch nicht geschehen, können Sie diesen als `csv` Datei herunterladen:

```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
```

Das Einlesen erfolgt, sofern die Daten im aktuellen Verzeichnis liegen, über:

```
tips <- read.csv2("tips.csv")
```

Tipp: Wenn Sie nicht mehr wissen wo die Daten liegen: statt `tips.csv` den Befehl `file.choose()` als Argument für die Funktion `read.csv2` verwenden.

1. Bestimmen Sie ein 90% Konfidenzintervall für den Anteil der Raucher (`smoker`) in der Population.

2. Testen Sie zum Niveau 5%, ob sich der Anteil der Raucher in der Population beim Lunch von dem beim Dinner (`time`) unterscheidet.
3. Gibt es einen Zusammenhang zwischen Rauchen und Wochentag (`day`)?

Literatur

- David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel (2014): *Introductory Statistics with Randomization and Simulation*, https://www.openintro.org/stat/textbook.php?stat_book=isrs, Kapitel 3
- Nicholas J. Horton, Randall Pruim, Daniel T. Kaplan (2015): Project MOSAIC Little Books *A Student's Guide to R*, https://github.com/ProjectMOSAIC/LittleBooks/raw/master/StudentGuide/MOSAIC-StudentGuide.pdf, Kapitel 4.2, 4.3, 6.3
- Maike Luhmann (2015): *R für Einsteiger*, Kapitel 18.1
- Andreas Quatember (2010): *Statistik ohne Angst vor Formeln*, Kapitel 3.4, 3.6, 3.8
- Daniel Wollschläger (2014): *Grundlagen der Datenanalyse mit R*, Kapitel 10.2

Versionshinweise:

- Datum erstellt: 2017-03-10
- R Version: 3.3.3
- `mosaic` Version: 0.14.4