

VO Statistik

Sitzung 6: Gut getestet: Inferenzstatistik I

Dominik Duell

Universität Innsbruck

Etwas Admin

- ▶ Hausaufgabe 1 ist am 14.5., Mitternacht, fällig
- ▶ etwas später, Pop Quiz, mit Link hier in diesem Video

Plan für heute

1. Wahrscheinlichkeit zu Ende bringen
2. Grundlagen der statistischen Inferenz

Grundlagen der statistischen Inferenz

Definitionen I

- ▶ Grundgesamtheit: Beobachtungen für jede mögliche, relevante Einheit mit Bezug zu unserer Forschungsfrage.
- ▶ Stichprobe: eine Teilmenge von Einheiten, die aus der zugrunde liegenden Grundgesamtheit gezogen wurden.
- ▶ Statistische Inferenz: Der Prozess, bei dem wir Rückschlüsse aus der Stichprobe über die Grundgesamtheit ziehen.
- ▶ Punktschätzung: Schätzung einer einzelnen Statistik der Grundgesamtheit aus unserer Stichprobe.
- ▶ Intervallschätzung: Schätzung einer Reihe von Statistik der Grundgesamtheit aus unserer Stichprobe.

Definitionen II

- ▶ Hypothesentest: Wahrscheinlichkeitsgestützte Aussagen über die Grundgesamtheit treffen.
- ▶ Stichprobenverteilung: hypothetische Verteilung der Stichprobenstatistik
- ▶ Standardfehler der Statistik: Standardabweichung der Stichprobenverteilung von der Stichprobenstatistiken
- ▶ Deskriptive Inferenz: Verwendung von vorhandenen Daten, um etwas über nicht vorhandene Daten zu lernen
- ▶ Kausale Inferenz: Lernen über die Ursachen des Beobachteten.

Stichprobenverteilung

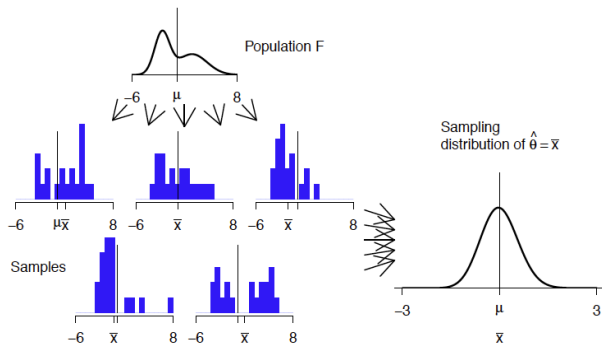


Figure 4: *Ideal world.* Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

Figure 1: Quelle: Hesterberg (2014), S.16

Definitionen III

Eine Stichprobe von n Beobachtungen einer oder mehreren Variablen, $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ ist eine **Zufallsstichprobe**, wenn die n Beobachtungen **unabhängig** voneinander aus derselben Grundgesamtheit mit der Wahrscheinlichkeitsverteilung $F(\mathbf{Y}, \theta)$ gezogen werden und die Wahrscheinlichkeit das ein bestimmter Wert gezogen wird die gleiche ist (**identisch**).

Definitionen IV

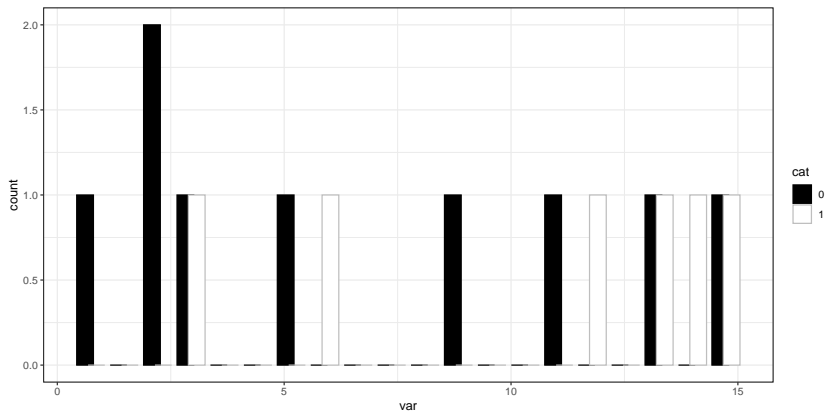
- ▶ Statistik: jede Funktion, die aus den Daten in einer Stichprobe berechnet wird – da diese Statistik eine Funktion einer oder mehrerer Zufallsvariablen ist, ist sie auch eine Zufallsvariable mit einer Wahrscheinlichkeitsverteilung.
- ▶ Stichprobenverteilung: Wahrscheinlichkeitsverteilung einer Statistik.
- ▶ Punktschätzung: Statistik, die einen einzelnen Wert für θ liefert.
- ▶ Intervallschätzung: Bereich von Werten, die θ enthalten, mit vorab zugewiesener Wahrscheinlichkeit.
- ▶ Standardfehler der Schätzung: Standardabweichung der Stichprobenverteilung.
- ▶ Schätzer: Regel zur Verwendung der Daten zur Schätzung von θ .

Beispiel

```
data <- read.csv('../data/gv300_data_fakeData.csv') %>%
  mutate(cat=factor(cat))
data %>% head()
```

##	var	cat	varCorr	varWeakCorr	varInd	varNonMon	varBiRaw	varOutlier
## 1	3	0	6.3203350	17.7186400	4.299648	8.75	0.0316617	-1.000000
## 2	5	0	6.8023790	8.6798990	5.445352	12.75	0.6922094	-1.666667
## 3	3	1	0.0886511	0.7677075	1.029927	8.75	0.1589295	-1.000000
## 4	15	0	12.8066200	18.4330800	9.987334	2.75	0.1257472	50.000000
## 5	12	1	11.5913100	8.5000760	7.071755	11.00	0.6499108	-4.000000
## 6	6	1	1.2054410	1.8113440	3.001570	14.00	0.6948789	-2.000000
##	varOutlierNoise	varBi	varExp					
## 1	11.278350	0	1.822119					
## 2	15.591280	1	2.718282					
## 3	16.707490	0	1.822119					
## 4	68.350900	0	20.085540					
## 5	-0.429595	1	11.023180					
## 6	13.922400	1	3.320117					

Beispiel



Beispiel

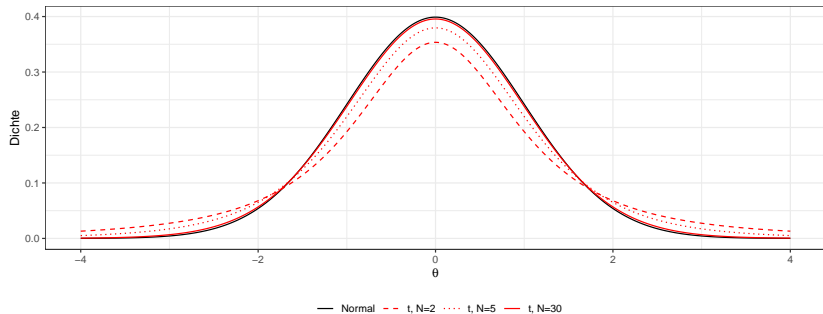
```
data %>% t.test(var~cat,data=.)
```

```
##  
## Welch Two Sample t-test  
##  
## data: var by cat  
## t = -1.4019, df = 11.571, p-value = 0.1872  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -9.531011 2.086567  
## sample estimates:  
## mean in group 0 mean in group 1  
## 6.777778 10.500000
```

Beispiel

$$t = \frac{|\overline{var}_1 - \overline{var}_0|}{\sqrt{\sigma_{var_1, var_0}^2 \left(\frac{1}{N_1} + \frac{1}{N_0} \right)}}$$

Beispiel



Beispiel

```
data %>% t.test(var~cat,data=.)
```

```
##  
## Welch Two Sample t-test  
##  
## data: var by cat  
## t = -1.4019, df = 11.571, p-value = 0.1872  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -9.531011 2.086567  
## sample estimates:  
## mean in group 0 mean in group 1  
## 6.777778 10.500000
```

Was gibt's nächste Woche

1. Inferenzstatistik: mehr hierzu, andere Statistiken, andere Stichprobenverteilungen, mehr Übung