Statistische Grundlagen

Kausalanalyse: Economics meets Data Science

Marc Justin Schmidt

05. November 2026

Statistische Grundlagen - Zwischenstand

Begriff	Bedeutung	Notation
Wahrscheinlichkeit [1]	Häufigkeit eines Ereignisses	$\mathbb{P}(X = x)$
Erwartungswert	Mittelwert des Zufallsprozesses	$\mathbb{E}[X]$
Varianz	Streuung um den Mittelwert	$Var(X)$
Standardabweichung	Streuung um den Mittelwert	$SD(X) = \sqrt{Var(X)}$
Kovarianz	Zusammenhang zweier Variablen	$Cov(X, Y)$
Korrelation	Standardisierte Kovarianz	$CorrX, Y = \frac{Cov(X,Y)}{SD(X)SD(Y)}$
Schätzer	Annäherung an Populationsparameter	$\bar{Y} \approx \mathbb{E}[Y]$
Inferenz	Testen von Hypothesen über Parameter	$t(\mathbb{E}[Y]) = \frac{\bar{Y} - \mathbb{E}[Y]}{SD(\bar{Y})/\sqrt{n}}$

[1] Unterscheide die bedingten Wahrscheinlichkeiten $F_{X|Y}(x|y)$, $F_{Y|X}(y|x)$ und $F(x, y)$!

Wahrscheinlichkeiten in Daten

Theorie | Code | Output

Zufallsvariable und Verteilung

- $\mathbb{P}(X = x)$: Wahrscheinlichkeit eines Werts
- $F_X(x) = \mathbb{P}(X \leq x)$: Verteilungsfunktion

Gemeinsame Verteilung

- $\mathbb{P}(X = x, Y = y)$: gemeinsame Wahrscheinlichkeit zweier Werte
- $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$: gemeinsame Verteilungsfunktion

Vierfeldertafel

	$Y = \text{ja}$	$Y = \text{nein}$	Summe
$X = A$	$\mathbb{P}(X = A, Y = \text{ja})$	$\mathbb{P}(X = A, Y = \text{nein})$	$\mathbb{P}(X = A)$
$X = B$	$\mathbb{P}(X = B, Y = \text{ja})$	$\mathbb{P}(X = B, Y = \text{nein})$	$\mathbb{P}(X = B)$
Summe	$\mathbb{P}(Y = \text{ja})$	$\mathbb{P}(Y = \text{nein})$	1

Die Randwahrscheinlichkeiten / marginale Wahrscheinlichkeiten entstehen durch Summieren der jeweiligen Zeile oder Spalte.

Wahrscheinlichkeiten in Daten

Theorie | Code | Output

```
g <- c("A", "A", "A", "B", "B", "B")
y <- c("ja", "ja", "nein",
       "ja", "nein", "nein")

# Absolute und relative Häufigkeit
table(y)
prop.table(table(y))

# Kreuztabelle und Zeilenanteile
tab <- table(Gruppe = g, Bestanden = y)
tab
prop.table(tab)
```

Wahrscheinlichkeiten in Daten

Theorie | Code | Output

```
## y
##   ja nein
##    3    3

## y
##   ja nein
##  0.5  0.5

##       Bestanden
## Gruppe ja nein
##      A  2    1
##      B  1    2

##       Bestanden
## Gruppe   ja nein
##      A 0.33 0.17
##      B 0.17 0.33
```

Bedingte Wahrscheinlichkeit und Unabhängigkeit

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

Konditionieren

Die Bedingung $Y = y$ legt die betrachtete Teilpopulation fest.

$$F_{X|Y}(x \mid y) = \mathbb{P}(X \leq x \mid Y = y)$$

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x)$$

Die beiden Konditionierungsrichtungen sind im Allgemeinen verschieden.

Unabhängigkeit

Nur wenn X und Y unabhängig sind, liefert die Kenntnis von Y keine Information über X:

$$\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$$

Dann, und nur dann, gilt auch:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

Satz von Bayes

Bayes zeigt, was eine Beobachtung über einen uns interessierenden, unbekannten Zustand aussagt.

$$\mathbb{P}(Y \mid X) = \frac{\mathbb{P}(X \mid Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

BEISPIEL COVID-Test: Von 10.000 getesteten Personen sind 1 % infiziert, also 100 Personen.

Gruppe	Ergebnis
Infiziert und positiv getestet	90
Nicht infiziert, aber positiv getestet	495
Positive Tests insgesamt	**585**

$$\mathbb{P}(\text{infiziert} \mid \text{positiv}) = \frac{\mathbb{P}(\text{positiv} \mid \text{infiziert})\,\mathbb{P}(\text{infiziert})}{\mathbb{P}(\text{positiv})} = \frac{\left(\frac{90}{100}\right)\left(\frac{100}{10.000}\right)}{\frac{585}{10.000}} = \frac{0{,}90 \cdot 0{,}01}{0{,}0585} \approx 15\%$$

Zeit für die Übungen!

Datei: 02_1_Aufgabenblatt_Wahrscheinlichkeiten.pdf

Lösungsvorschlag: 02_1_Aufgabenblatt_Statistische_Grundlagen_Loesungen.pdf



Motivation für den zweiten Teil

Wir wollen (die wahren) statistischen Eigenschaften einer Population (**Populationsparameter**) kennen:

1. **Statistische Momente**
    - Erwartungswert, ggf. Varianz
    - z.B.: Effekt von Trainingseinheiten pro Woche auf die erwartete Toranzahl eines Spielers
2. **Bedingte Wahrscheinlichkeiten**
    - z.B.: Effekt von Menüwahl (Fleisch vs. vegetarisch) auf die Wahrscheinlichkeit, ein Dessert zu bestellen

Da wir nur **Stichproben** haben, sind unsere Schätzer selbst **Zufallsvariablen**.

Frage: Wie zuverlässig repräsentiert ein Schätzer den wahren Wert?

Statistische Grundlagen - Recap

Begriff	Bedeutung	Notation
Wahrscheinlichkeit [1]	Häufigkeit eines Ereignisses	$\mathbb{P}(X = x)$
Erwartungswert	Mittelwert des Zufallsprozesses	$\mathbb{E}[X]$
Varianz	Streuung um den Mittelwert	$Var(X)$
Standardabweichung	Streuung um den Mittelwert	$SD(X) = \sqrt{Var(X)}$
Kovarianz	Zusammenhang zweier Variablen	$Cov(X, Y)$
Korrelation	Standardisierte Kovarianz	$CorrX, Y = \frac{Cov(X,Y)}{SD(X)SD(Y)}$
Schätzer	**Annäherung an Populationsparameter**	$\bar{Y} \approx \mathbb{E}[Y]$
Inferenz	**Testen von Hypothesen über Parameter**	$t(\mathbb{E}[Y]) = \frac{\bar{Y} - \mathbb{E}[Y]}{SD(\bar{Y})/\sqrt{n}}$

➡️ Heutige Übung: **Simulationen** zum Verständnis der statistischen Konzepte mit Fokus auf Schätzern und Inferenz

[1] Unterscheide die bedingten Wahrscheinlichkeiten $F_{X|Y}(x|y)$, $F_{Y|X}(y|x)$ und $F(x, y)$!

Eigenschaften von Schätzern

Fokus in diesem Kurs auf Schätzern, da wir Stichproben statt Population(en) haben. **Schätzer sind bestmögliche Annäherungen** an wahre Parameter.

- **Erwartungstreue (unbiased / unverzerrt)**: richtiger Erwartungswert, z.B. Mittelwertschätzer ($\mathbb{E}[\bar{Y}] = \mathbb{E}[Y]$)
- **Konsistenz**: Annäherung bei großer Stichprobe n ($\bar{Y} \to \mathbb{E}[Y]$ für $n \to \infty$)
- **Effizienz**: kleinste Varianz unter den unverzerrten Schätzern ($Var(\bar{Y}_1) < Var(\bar{Y}_2)$)

Monte Carlo Simulationen

Eine Monte-Carlo-Simulation ist ein Experiment mit Zufallsziehungen: wir wiederholen Zufallsprozesse **viele Male**, um zu sehen, **wie sich Ergebnisse im Durchschnitt verhalten**

1. Definiere eine Verteilung oder verwende einen realen Datensatz (Population)
2. Ziehe eine zufällige Stichprobe (z.B. 100 Beobachtungen)
3. Berechne den Schätzer von Interesse (z.B. den Mittelwert)
4. Wiederhole die Schritte 2–3 viele Male (z.B. 10.000 Mal)
5. Untersuche die Verteilung der Ergebnisse (z.B. der Mittelwerte)

Wir wollen verstehen, **wie sich der Mittelwertschätzer in verschiedenen Stichproben verhält** und warum die Eigenschaften (Erwartungstreue, Konsistenz, Effizienz) wichtig sind

Merksatz: Monte-Carlo-Simulation = wiederholtes Zufallsexperiment (mit künstlichen **oder echten** Daten), um Theorie durch Beobachtung zu verstehen

Zentraler Grenzwertsatz (CLT)

Das **CLT** (Central Limit Theorem) ist das theoretische Fundament unserer Simulationen für den Mittelwertschätzer

> Wenn die Stichprobengröße gegen unendlich geht, kann das standardisierte Stichprobenmittel $\frac{\bar{Y}-\mathbb{E}[Y]}{SD(Y)/\sqrt{n}}$ stetig besser mit einer Standardnormalverteilung $N(0,1)$ approximiert werden - unabhängig von der zugrundeliegenden Verteilung.

Warum betrachten wir hier nur den Mittelwert?

- Er ist der **einfachste und wichtigste Schätzer**, auf dem viele andere Methoden aufbauen
- Viele andere Schätzer (z.B. Regressionskoeffizienten) lassen sich **auf Mittelwerte zurückführen** oder folgen **ähnlichen Prinzipien**

Erwartungstreue

Erklärung | R-Code | Output

Idee: Ein Schätzer ist erwartungstreu, wenn sein Mittelwert dem wahren Parameter entspricht. Wir prüfen das am Beispiel des Mittelwertschätzers - empirisch (Monte Carlo Simulation)

$$\mathbb{E}[\bar{Y}] = \mathbb{E}[Y]$$

- Datensatz: `mtcars` (enthält den durchschnittlichen Benzinverbrauch (mpg) verschiedener Autos)
- Wir ziehen viele Zufallsstichproben und berechnen jeweils den Mittelwert der Verbrauchswerte
- Über alle Stichproben hinweg sollte der mittlere Schätzwert dem wahren Mittelwert entsprechen
- Hinweis: `mtcars` ist sehr klein - für „echte“ Monte-Carlo-Simulationen wären größere oder synthetische Datensätze (z. B. `rnorm()`) besser geeignet

Erwartungstreue

Erklärung | R-Code | Output

```r
library(ggplot2) # Load package

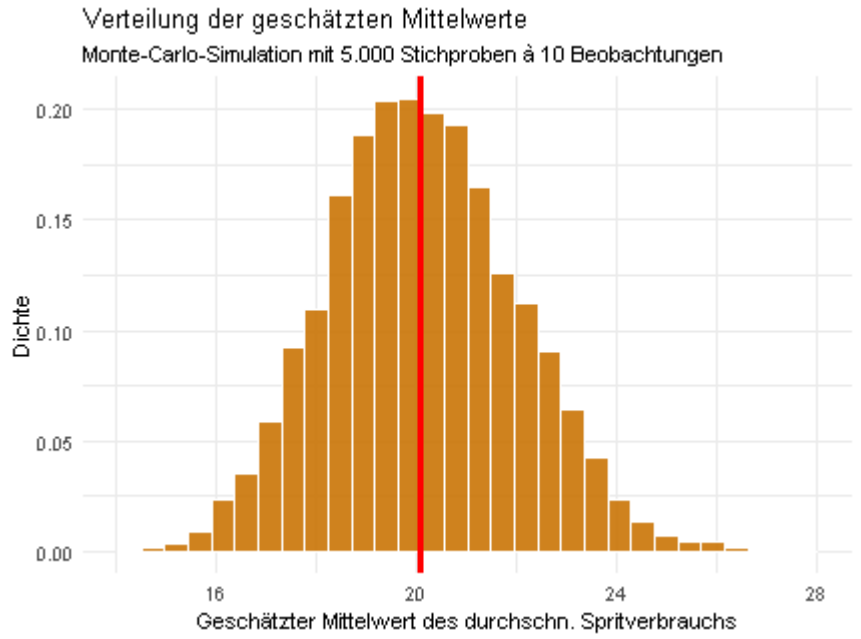
set.seed(123) # Fix randomness
mu_true <- mean(mtcars$mpg) # Compute the mean

# Monte Carlo draws of sample means (5,000 random samples of size 10 and compute mean)
means <- replicate(5000, mean(sample(mtcars$mpg, 10, replace=TRUE)))
df_means <- data.frame(means = means)

# Plot the distribution of simulated sample means
ggplot(df_means, aes(x = means)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "#CA7406", color = "white", alpha =
  geom_vline(xintercept = mu_true, color = "red", linewidth = 1.2) +
  labs(title = "Verteilung der geschätzten Mittelwerte",
       subtitle = "Monte-Carlo-Simulation mit 5.000 Stichproben à 10 Beobachtungen",
       x = "Geschätzter Mittelwert",
       y = "Dichte") +
  theme_minimal()
```

Erwartungstreue

Erklärung | R-Code | **Output**



- Die Verteilung der Mittelwerte zentriert sich um den wahren Mittelwert (rote Linie)
- Im Durchschnitt ist der Schätzer also „richtig“: erwartungstreu
- Erhöht man die Anzahl der Beobachtungen (n): Verteilung wird enger (präziser: Konsistenz)
- Erhöht man die Anzahl der Ziehungen (Wiederholungen): Verteilung wird glatter und stabiler

Konsistenz

Erklärung | R-Code | Output

Idee: Ein Schätzer ist konsistent, wenn er mit wachsendem Stichprobenumfang n immer näher an den wahren Wert heranrückt. Die Streuung der Schätzwerte nimmt ab, je größer n wird.

$$\bar{Y} \to \mathbb{E}[Y] \quad \text{für} \quad n \to \infty$$

- Datensatz: `diamonds` (enthält Preise von Diamanten, $n \approx 54{,}000$)
- Wir ziehen viele Zufallsstichproben mit unterschiedlicher Größe und vergleichen die Verteilungen der Mittelwerte
- Erwartung: mit größerem n wird die Streuung kleiner; der Schätzer wird stabiler

Konsistenz

Erklärung | R-Code | Output

```r
library(ggplot2) # Load package

set.seed(123) # Fix randomness
mu_true <- mean(diamonds$price) # Compute the mean

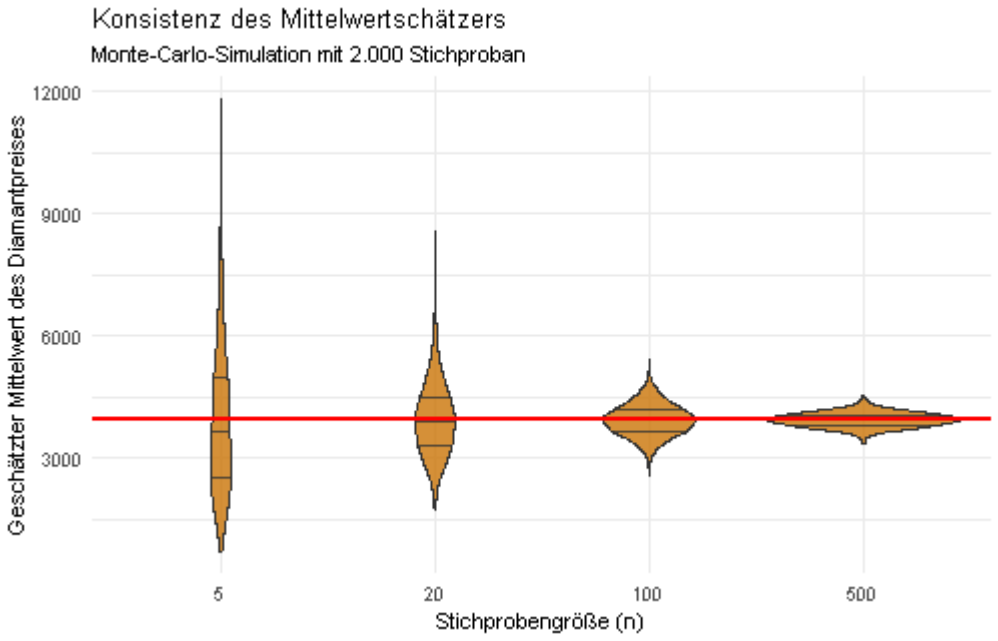
n_vals <- c(5, 20, 100, 500) # Define different sample sizes

# Means for all sizes
sim <- lapply(n_vals, function(n)
  replicate(2000, mean(sample(diamonds$price, n)))
)

# Violin Plot
ggplot(df, aes(x = Stichprobengröße, y = Mittelwert, fill = Stichprobengröße)) +
geom_violin(aes(fill = Stichprobengröße), alpha = 0.8, draw_quantiles = c(0.25, 0.5, 0.75)) +
geom_hline(yintercept = mu_true, color = "red", linewidth = 1) +
...
```

Konsistenz

Erklärung | R-Code | **Output**



- Bei kleinem n: hohe Streuung der Schätzwerte, unpräzise
- Bei großem n: enge Verteilung um den wahren Wert, präzise
- Konsistenz: Schätzer nähert sich mit wachsendem n dem wahren Erwartungswert an

Effizienz

Erklärung | R-Code | Output

Idee: Ein Schätzer ist effizient, wenn er unter allen unverzerrten Schätzern die kleinste Varianz hat. Wir vergleichen zwei Schätzer desselben Parameters (beide schätzen den Erwartungswert - einer nutzt alle, der andere nur die Hälfte der Stichprobe) und prüfen, welcher stabiler ist.

$$Var(\bar{Y}_1) < Var(\bar{Y}_2) \quad \Rightarrow \quad \text{Schätzer 1 ist effizienter}$$

- Datensatz: `faithful` - Dauer von Geysirausbrüchen in Minuten ($n = 272$ Beobachtungen)
- Wir ziehen viele Stichproben und vergleichen die Dichteverteilung der beiden Schätzer
- Erwartung: beide sind im Mittel korrekt, aber ein Schätzer ist effizienter (hat geringere Varianz)

Effizienz

Erklärung | **R-Code** | Output

```r
library(ggplot2) # Load package

set.seed(123) # Fix randomness
x <- faithful$eruptions
mu_true <- mean(faithful$eruptions) # Compute the mean

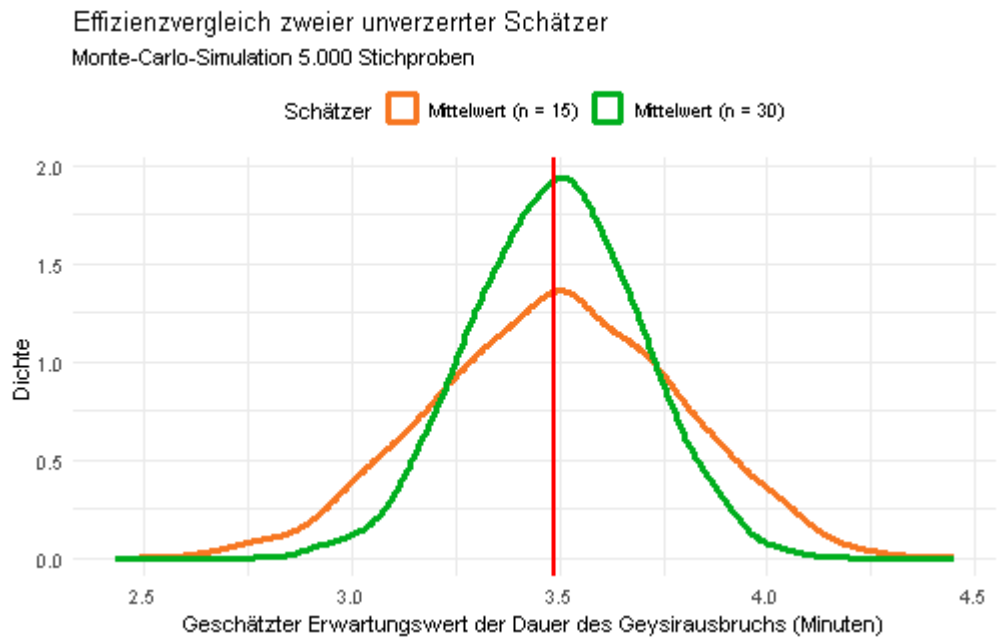
# Means for both estimators in a Monte Carlo simulation setting
sim_full <- replicate(5000, mean(sample(x, 30, replace = TRUE)))
sim_half <- replicate(5000, mean(sample(x, 15, replace = TRUE)))

# tidy DataFrame
df_eff <- data.frame(...)

# Density Plot
ggplot(df_eff, aes(x = Wert, color = Schätzer)) +
  geom_density(linewidth = 1.2, adjust = 1.1) +
  geom_vline(xintercept = mu_true, color = "red", linewidth = 1) +
  ...
```

Effizienz

Erklärung | R-Code | **Output**



- Beide Schätzer sind im Mittel unverzerrt (zentriert um den wahren Wert)
- Der Mittelwert-Schätzer in grün weist eine kleinere Streuung auf; effizienter Schätzer

Zeit für die Übungen!

Datei: 02_2_Aufgabenblatt_Statistische_Grundlagen.pdf

Aufgaben: 1, 2, A1, A2



Hypothesentests

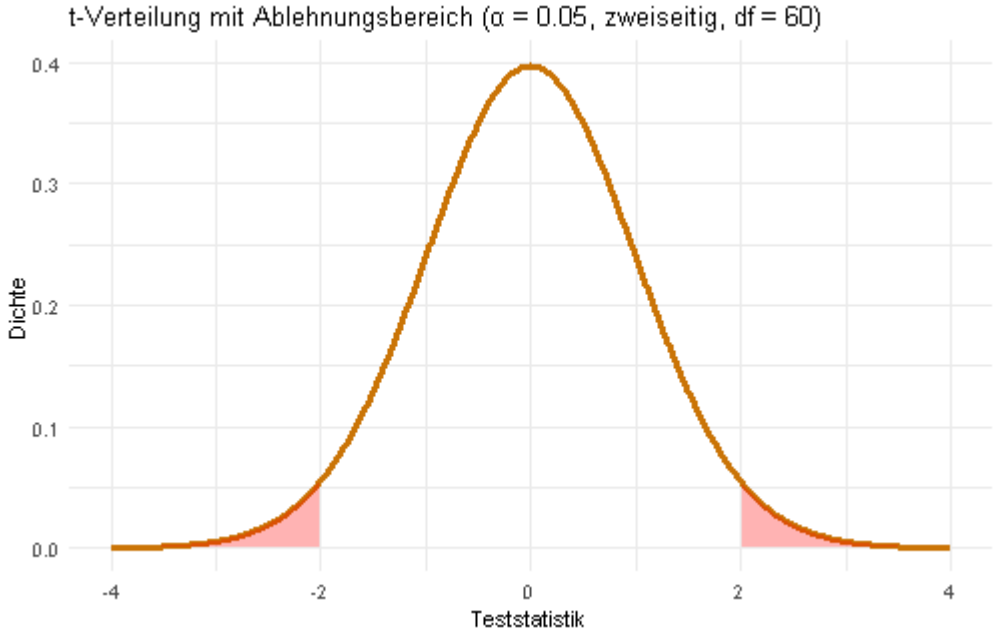
> "Inferenz" bedeutet, statistische Eigenschaften von Schätzern zu nutzen, um spezifische Aussagen über Populationsparameter zu überprüfen.

- Prüfen, ob ein **beobachteter Zusammenhang real oder zufällig ist**
- Schätzer unterliegen zufälligen Schwankungen
- Wir brauchen ein **systematisches Entscheidungsverfahren**

Ablauf

1. Formuliere Hypothesen:
   - Nullhypothese H_0: kein Effekt (z.B. Mittelwert = 0)
   - Alternativhypothese H_1: Effekt vorhanden (z.B. Mittelwert $\neq 0$)
2. Berechne Teststatistik (z.B. t-Wert): $t = \frac{\bar{Y} - \mathbb{E}[Y]}{SD(\bar{Y})/\sqrt{n}}$ und lege ggf. Signifikanzniveau α fest (1%, 5%, 10%)
3. Vergleiche mit kritischem Wert oder berechne p-Wert (wenn p-Wert < Signifikanzniveau α, dann H_0 ablehnen)

Hypothesentests



$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})}$$

- Die roten Bereiche zeigen die Werte, die unter H_0 sehr unwahrscheinlich sind
- Liegt unsere Teststatistik dort: H_0 verwerfen, ansonsten wird H_0 toleriert

p-Wert

> Der p-Wert zeigt, wie ungewöhnlich unsere Daten wären, wenn die Nullhypothese H_0 tatsächlich gilt. Anders: Der p-Wert misst nicht, wie wahrscheinlich H_0 wahr ist, sondern wie plausibel unsere Daten unter H_0 sind.

$$p = \mathbb{P}(\text{Ergebnis so extrem oder extremer wie beobachtet} \,|H_0)$$

- Kleiner p-Wert: Ergebnis ist sehr unwahrscheinlich, falls H_0 stimmt: H_0 verwerfen
- Großer p-Wert: Ergebnis ist plausibel unter H_0: H_0 behalten
- Daumenregel: p-Wert $\leq \alpha$ bedeutet statistisch signifikant, z.B. p ≤ 0.05 bei $\alpha = 0.05$

Konfidenzintervalle

> Ein Konfidenzintervall (KI) zeigt den Bereich, in dem der wahre Parameter mit hoher Wahrscheinlichkeit liegt.

$$KI = \bar{Y} \pm z \cdot SE(\bar{Y}) = \bar{Y} \pm z \cdot \frac{SD(Y)}{\sqrt{n}}$$

- Liefert **plausiblen Wertebereich** für den wahren (aber unbekannten) Wert in der Grundgesamtheit
- Breite des Intervalls zeigt die Präzision der Schätzung: schmales Intervall bedeutet hohe Präzision, breites Intervall hohe Unsicherheit
- Werte innerhalb des Intervalls gelten als "kompatibel" mit unseren Daten; Werte außerhalb gelten als statistisch unwahrscheinlich

Zeit für die Übungen!

Datei: 02_2_Aufgabenblatt_Statistische_Grundlagen.pdf

Aufgaben: 3, 4



Vielen Dank!

Fragen?

marcjustin.schmidt@tu-dortmund.de

Zurück zur Hauptseite