Block 2: Inferenz, Unsicherheit & Entscheidungslogik

Data Science and Strategy for Business

March 12, 2026

Lernziele Block 2

- Inferenz als Methode zur Messung von Unsicherheit erklären
- Bootstrapping und Permutationstests praktisch umsetzen
- Traditionelle Tests (t-Test, ANOVA) einordnen
- Fehler 1. und 2. Art als ökonomische Risiken interpretieren
- Gefahren von p-Hacking erkennen und FDR anwenden
- Zwischen statistischer Signifikanz und praktischer Relevanz unterscheiden

---

Statistische Inferenz quantifiziert Unsicherheit – essentiell für fundierte Entscheidungen.

Was ist statistische Inferenz?

Definition
Von einer Stichprobe auf die Population schliessen.

Kernfragen

- Wie präzise ist unsere Schätzung?
- Ist ein Effekt "echt" oder Zufall?
- Wie sicher können wir sein?

Signal vs. Rauschen
Inferenz hilft, echte Muster von zufälliger Variation zu unterscheiden.

Beispiel:
Stichprobe: 100 Kunden
Mittlere Zufriedenheit: 7.2

Fragen:

1. Wie nah ist 7.2 am wahren Wert?
2. Ist 7.2 "signifikant" hoeher als 7.0?
3. Welches Konfidenzintervall?

Jede Schätzung aus Daten ist mit Unsicherheit behaftet.

Das Stichprobenproblem

Problem
Wir sehen nur eine Stichprobe, nicht die Population.

Stichprobenvariabilität
Verschiedene Stichproben ergeben verschiedene Ergebnisse!

Beispiel:

- Stichprobe 1: $\bar{x} = 7.2$
- Stichprobe 2: $\bar{x} = 6.8$
- Stichprobe 3: $\bar{x} = 7.5$

Lösung: Standardfehler

$$SE = \frac{s}{\sqrt{n}}$$

Je größer n, desto kleiner SE, desto präziser die Schätzung.

Konfidenzintervall

$$\bar{x} \pm 1.96 \times SE$$

Mit 95% Wahrscheinlichkeit enthält das Intervall den wahren Wert.

Mehr Daten = weniger Unsicherheit = engere Konfidenzintervalle.

p-Werte richtig verstehen

Definition
Der p-Wert ist die Wahrscheinlichkeit, die beobachteten Daten (oder extremere) zu sehen, *wenn die Nullhypothese wahr ist*.

p = 0.03 bedeutet:

- 3% Chance für dieses Ergebnis unter H0
- NICHT: 3% Wahrscheinlichkeit, dass H0 wahr ist!

Häufige Fehlinterpretationen

- p = Wahrscheinlichkeit, dass H0 stimmt
- p = Wahrscheinlichkeit, dass Ergebnis repliziert
- $p < 0.05$ beweist Effekt
- $p > 0.05$ bedeutet kein Effekt

Korrekt:
p ist ein Mass für Kompatibilität mit H0, nicht mehr.

Der p-Wert beantwortet nicht die Frage, die wir eigentlich stellen wollen!

Konfidenzintervalle verstehen

Definition
Ein 95%-KI ist ein Bereich, in dem der wahre Parameter mit 95% Konfidenz liegt.

Interpretation
Bei wiederholter Stichprobenziehung wuerden 95% der berechneten Intervalle den wahren Wert enthalten.

Formel (grosses n):

$$CI = \bar{x} \pm z_{\alpha/2} \times SE$$

Warum KI besser als p?

- Zeigt Unsicherheit visuell
- Gibt Effektgröße und Präzision
- Erlaubt direkte Interpretation

Beispiel:
$\bar{x} = 7.2$, SE $= 0.3$
95%-KI: [6.61, 7.79]
"Wahre Zufriedenheit liegt zwischen 6.6 und 7.8"

Konfidenzintervalle sind informativer als p-Werte – sie zeigen Richtung und Unsicherheit.

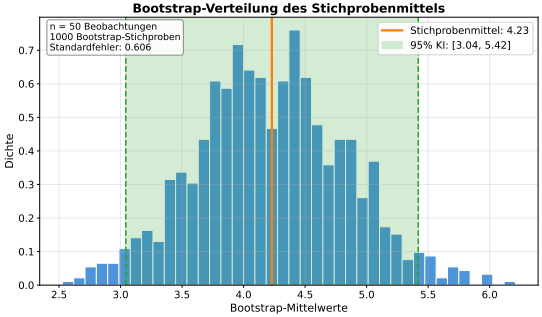
Only One Test: Bootstrapping

Idee
Aus einer Stichprobe viele "neue" Stichproben ziehen (mit Zurücklegen).

Algorithmus

1. Ziehe n Beobachtungen mit Zurücklegen
2. Berechne Statistik (z.B. Mittelwert)
3. Wiederhole 1000+ Mal
4. Analysiere Verteilung

Vorteil: Keine Annahmen über Verteilung nötig!



Bootstrap liefert eine empirische Stichprobenverteilung ohne theoretische Annahmen.

Bootstrap: Konfidenzintervall

Perzentil-Methode
Das 95%-KI ist das 2.5%- und 97.5%-Perzentil (= Wert, unter dem x% der Daten liegen) der Bootstrap-Verteilung.

Beispiel in R (infer-Paket):

```
library(infer)
bootstrap_dist <- data %>%
  specify(response = umsatz) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

ci <- bootstrap_dist %>%
  get_confidence_interval(level = 0.95)
```

Das infer-Paket implementiert die "Only one test"-Philosophie konsistent.

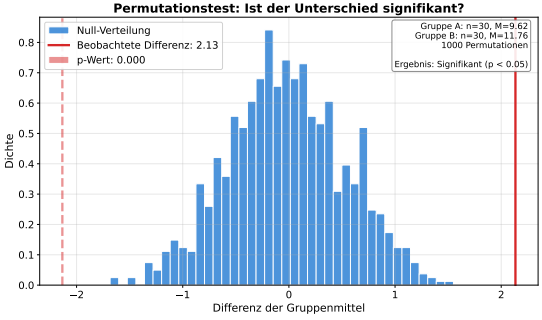
Permutationstest

Frage
Ist der beobachtete Unterschied zwischen zwei Gruppen “echt” oder Zufall?

Idee
Wenn es keinen echten Unterschied gibt (H0), ist die Gruppenzugehoerigkeit irrelevant.

Algorithmus

1. Berechne beobachtete Differenz
2. Mische Gruppenlabels zufällig
3. Berechne Differenz
4. Wiederhole 1000+ Mal
5. Vergleiche mit **Null-Verteilung** (= Verteilung der Teststatistik unter H0)



p-Wert = Anteil der Permutationen mit extremerem Ergebnis als beobachtet.

Permutationstest in R

Mit dem infer-Paket:

```
library(infer)

# Beobachtete Differenz
obs_diff <- data %>%
  specify(umsatz ~ gruppe) %>%
  calculate(stat = "diff in means", order = c("B", "A"))

# Null-Verteilung durch Permutation
null_dist <- data %>%
  specify(umsatz ~ gruppe) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("B", "A"))

# p-Wert berechnen
null_dist %>%
  get_p_value(obs_stat = obs_diff, direction = "two-sided")
```

Konsistente Syntax für Bootstrap und Permutation im infer-Paket.

Traditionelle Tests

t-Test

- Vergleich zweier Mittelwerte
- Annahme: Normalverteilung
- `t.test(gruppe1, gruppe2)`

ANOVA

- Vergleich mehrerer Gruppen
- F-Statistik
- `aov(y ~ gruppe, data)`

Chi-Quadrat-Test

- Kategoriale Variablen (z.B. Geschlecht, Ja/Nein)
- Unabhängigkeit prüfen
- `chisq.test(table)`

Shapiro-Wilk-Test

- Normalitätsprüfung
- `shapiro.test(x)`

Einordnung: Alle Tests sind Spezialfaelle des gleichen Grundprinzips – Bootstrap und Permutation sind oft flexibler!

Traditionelle Tests basieren auf mathematischen Annahmen, die oft verletzt sind.

t-Test: Varianten

Ein-Stichproben t-Test
Testet ob Mittelwert = hypothetischer Wert

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

```
t.test(x, mu = 100)
```

Zwei-Stichproben t-Test
Testet ob zwei Gruppen gleichen Mittelwert haben

```
t.test(x, y)
```

Gepaarter t-Test
Für abhängige Messungen (vorher/nachher)

```
t.test(x, y, paired = TRUE)
```

Welch vs. Student

- Student: gleiche Varianz angenommen
- Welch: verschiedene Varianzen erlaubt
- R nutzt Welch als Default

Welch t-Test ist robuster und sollte bevorzugt werden.

ANOVA: Mehrere Gruppen vergleichen

Frage
Unterscheiden sich mehr als zwei Gruppen?

Hypothesen

- H0: Alle Gruppenmittelwerte gleich
- H1: Mindestens ein Unterschied

F-Statistik

$$F = \frac{\text{Varianz zwischen Gruppen}}{\text{Varianz innerhalb Gruppen}}$$

In R:

```
model <- aov(y ~ gruppe, data)
summary(model)
```

Post-hoc Tests (nachtraeglich)
Nach signifikanter ANOVA: Welche Gruppen unterscheiden sich?

```
TukeyHSD(model)
```

Achtung: ANOVA sagt nur "es gibt Unterschiede", nicht welche!

Bei mehr als 2 Gruppen: ANOVA statt mehrere t-Tests – viele Tests erhoehen die Chance auf Zufallstreffer (alpha-Inflation).

Regressionstabelle interpretieren

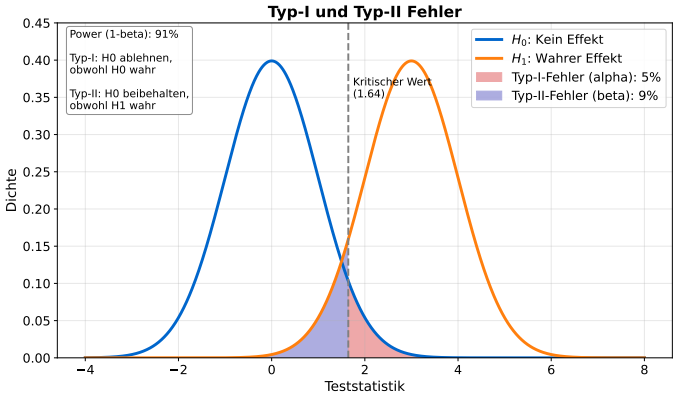
Typische Ausgabe von `summary(lm(...))`

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5000.00     500.00   10.00   < 0.001 ***
werbung          2.50       0.30    8.33   < 0.001 ***
preis         -100.00      15.00   -6.67   < 0.001 ***
qualitaet      500.00     100.00    5.00   < 0.001 ***
```

- **Estimate**: Geschätzter Koeffizient
- **Std. Error**: Unsicherheit der Schätzung
- **t value**: Estimate / Std. Error (je größer, desto signifikanter)
- **Pr(**$> |t|$**)**: p-Wert (Wahrscheinlichkeit unter H0)

Signifikanzsterne: *** p<0.001, ** p<0.01, * p<0.05

Fehler 1. und 2. Art



Beide Fehlertypen sind unvermeidlich – wir können nur den Tradeoff wählen.

Fehlertypen im Business-Kontext

Typ-I-Fehler (alpha)
H0 ablehnen, obwohl H0 wahr.
"Falsch Positiv"

Business-Beispiel:
Kampagne starten, die nicht wirkt

- Kosten: Kampagnenkosten
- Typisch: 5% Risiko (alpha = 0.05)

Kernfrage: Welcher Fehler ist teurer?

Typ-II-Fehler (beta)
H0 beibehalten, obwohl H1 wahr.
"Falsch Negativ"

Business-Beispiel:
Wirksame Kampagne nicht starten

- Kosten: Entgangener Gewinn
- Typisch: 20% Risiko (Power = 80%)

Im Business sollten alpha und beta basierend auf Kosten gewählt werden, nicht nur Konvention.

Power: Die Macht des Tests

Definition
Power $= 1 - \beta$ = Wahrscheinlichkeit, einen echten Effekt zu finden

Power hängt ab von:

1. Stichprobengröße n
2. Effektgröße d
3. Signifikanzniveau α
4. Varianz in den Daten

Regel: Power-Analyse VOR Datenerhebung durchführen!

Typische Power-Ziele

- 80% = Standard
- 90% = Konservativ
- <80% = Riskant

Problem:
Underpowered Studies finden echte Effekte nicht oder ueberschätzen sie.

Zu kleine Stichproben sind Ressourcenverschwendung – nichts wird gefunden.

Sample Size Berechnung

Ziel
Wie viele Beobachtungen brauche ich, um einen Effekt der Größe d mit Power $1 - \beta$ bei α zu finden?

Faustregel für t-Test:

$$n \approx \frac{16}{d^2}$$

pro Gruppe (für 80% Power)

Beispiel:
$d = 0.5 \Rightarrow n \approx 64$ pro Gruppe

In R:

```
library(pwr)

# t-Test
pwr.t.test(
  d = 0.5,      # Effektgroesse
  sig.level = 0.05,
  power = 0.80,
  type = "two.sample"
)

# ANOVA
pwr.anova.test(
  k = 3,        # Gruppen
  f = 0.25,     # Effektgroesse
  sig.level = 0.05,
  power = 0.80
)
```

Power-Analyse ist essentiell für Budgetplanung von Studien und Experimenten.

Expected Value: Entscheidung unter Unsicherheit

P(Erfolg) = 60% P(Misserfolg) = 40%

	Kampagne erfolgreich	**Kampagne nicht erfolgreich**
Kampagne starten	+100k CHF	-30k CHF
Kampagne nicht starten	0k CHF	0k CHF

E[Starten] = 0.6*100 + 0.4*(-30) = 48k CHF
E[Nicht starten] = 0k CHF

Empfehlung: Kampagne starten (EV = +48k CHF)

Entscheidungen unter Unsicherheit erfordern Kosten-Nutzen-Abwaegung, nicht nur p-Werte.

Entscheidungsmatrix: Business-Anwendung

Beispiel: Neue Kampagne testen

	Kampagne wirkt	Kampagne wirkt nicht
Starten	+100k CHF	-50k CHF
Nicht starten	0 CHF	0 CHF

Expected Value:
$EV(Start) = P(wirkt) \times 100k + P(nicht) \times (-50k)$
$EV(Nicht) = 0$

Entscheidungsregel
Starten wenn $EV(Start) > 0$
$\Rightarrow P(wirkt) > \frac{50k}{150k} = 33\%$

Mit Test-Ergebnis:
$p < 0.05$ ist oft zu konservativ!
Wenn Upside » Downside, akzeptiere mehr Risiko.

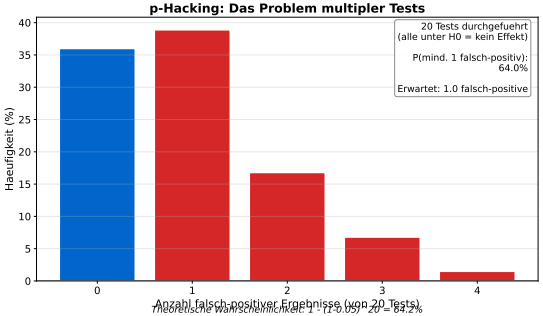
Optimale Entscheidungen beruecksichtigen Kosten und Wahrscheinlichkeiten, nicht nur p-Werte.

p-Hacking: Das Problem

Definition
Manipulieren der Analyse, bis $p < 0.05$ erreicht wird.

Typische Praktiken

- Viele Tests durchführen, nur signifikante berichten
- Ausreißer selektiv entfernen
- Variablen hinzufuegen/entfernen
- Subgruppen bilden bis signifikant
- Daten sammeln bis signifikant



Bei 20 Tests unter H0 sind im Schnitt 1 falsch-positiv (5% von 20).

Multiple Testing: Lösungen

Bonferroni-Korrektur

$$\alpha_{adj} = \frac{\alpha}{m}$$

Bei 20 Tests: $\alpha = 0.05/20 = 0.0025$
Nachteil: Sehr konservativ

False Discovery Rate (FDR)
Kontrolliert den erwarteten Anteil falscher Entdeckungen unter den signifikanten.

```
p.adjust(p_values, method = "BH")
```

Vorteil: Weniger konservativ, mehr Power

Best Practice: Bei vielen Tests immer FDR oder Bonferroni verwenden!

Ohne Korrektur sind viele "signifikante" Ergebnisse in grossen Studien Zufallsbefunde.

FDR-Korrektur in R

Beispiel: 20 Hypothesen testen
BH-Verfahren: p-Werte sortieren, jeden mit $\frac{i}{m} \times \alpha$ vergleichen.

```
# Simulierte p-Werte (18 unter H0, 2 echte Effekte)
p_values <- c(runif(18), 0.001, 0.003)

# Ohne Korrektur:
sum(p_values < 0.05)  # Viele "signifikante" Ergebnisse

# Bonferroni-Korrektur
p_bonf <- p.adjust(p_values, method = "bonferroni")
sum(p_bonf < 0.05)    # Konservativ

# Benjamini-Hochberg FDR
p_fdr <- p.adjust(p_values, method = "BH")
sum(p_fdr < 0.05)     # Besserer Kompromiss

# Vergleich
data.frame(original = p_values, bonferroni = p_bonf, fdr = p_fdr)
```

FDR ist der Standard in Data Science bei vielen simultanen Tests.

Signifikanz vs. Relevanz

Statistische Signifikanz

- $p < 0.05$
- "Wahrscheinlich kein Zufall"
- Abhängig von Stichprobengröße!

Problem mit Big Data:
Bei $n = 1'000'000$ wird fast alles signifikant, auch triviale Effekte.

Praktische Relevanz

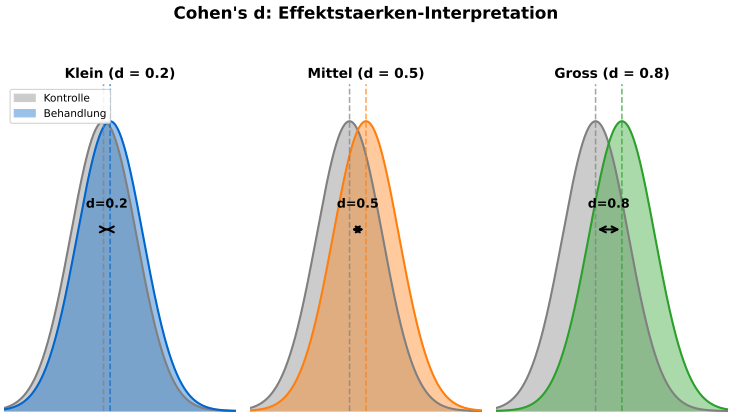
- Effektstaerke (z.B. Cohen's d)
- "Wie gross ist der Effekt?"
- Unabhängig von n

Effektstaerken:

- $d = 0.2$: klein
- $d = 0.5$: mittel
- $d = 0.8$: gross

Immer fragen: Ist der Effekt gross genug, um praktisch relevant zu sein?

Cohen's d: Effektstaerke



Cohen's d = Differenz in Standardabweichungen – unabhängig von Stichprobengröße.

Cohen's d berechnen

Formel

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

wobei

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Interpretation

- $|d| < 0.2$: vernachlässigbar
- $|d| \approx 0.2$: klein
- $|d| \approx 0.5$: mittel
- $|d| \geq 0.8$: gross

In R:

```
# Mit effsize-Paket
library(effsize)
cohen.d(gruppe_a, gruppe_b)

# Manuell
cohens_d <- function(x, y) {
  nx <- length(x)
  ny <- length(y)
  s_pooled <- sqrt(
    ((nx-1)*sd(x)^2 +
     (ny-1)*sd(y)^2) /
    (nx + ny - 2)
  )
  (mean(x) - mean(y)) / s_pooled
}
```

Cohen's d immer zusammen mit p-Wert berichten!

Fallstudie: E-Commerce Conversion

Situation
Online-Shop testet neuen Checkout-Prozess.

Daten

- Gruppe A (alt): n=5000, 2.1% Conversion
- Gruppe B (neu): n=5000, 2.3% Conversion

Statistische Analyse

- Chi-Quadrat: $p = 0.04$
- Cohen's $h = 0.03$ (Effektmass fuer Proportionen, sehr klein)

Business-Kontext

- 100k Besucher/Monat
- Durchschnittlicher Warenkorb: 80 CHF
- 0.2% mehr = 160 zusaetzliche Kaeufe
- = 12'800 CHF/Monat Mehrertrag

Entscheidung:
Trotz kleiner Effektgröße: Implementieren!
Absoluter Gewinn rechtfertigt es.

Bei hohem Volumen können auch kleine Effekte wirtschaftlich relevant sein.

Fallstudie: Medikamenten-Studie

Situation
Pharma-Firma testet neues Blutdruck-Medikament.

Daten

- Placebo: n=200, BD = 145 mmHg
- Medikament: n=200, BD = 143 mmHg

Statistische Analyse

- t-Test: p = 0.02
- Cohen's d = 0.15 (sehr klein)
- 95%-KI: [0.5, 3.5] mmHg

Klinische Bewertung

- 2 mmHg Senkung ist klinisch irrelevant
- Nebenwirkungen nicht gerechtfertigt
- Kosten pro Patient hoch

Entscheidung:
Trotz Signifikanz: Nicht zulassen. Effekt zu klein für Nutzen.

Statistische Signifikanz allein rechtfertigt keine klinische Entscheidung.

Fallstudie: HR Analytics

Situation
Unternehmen untersucht Zusammenhang zwischen Weiterbildung und Kündigung.

Daten (n = 500)

- Mit Weiterbildung: 8% Kündigung
- Ohne Weiterbildung: 15% Kündigung

Statistische Analyse

- Chi-Quadrat: $p < 0.001$
- Odds Ratio: 0.49 (= halbe Kuendigungs-Odds)
- Cohen's h = 0.22 (Proportionen-Effekt, klein-mittel)

Kritische Fragen

- Kausalität? Oder Selektion?
- Wer bekommt Weiterbildung?
- Confounders (Stoervariablen): Motivation, Potenzial?

Entscheidung:
Signifikant ja, aber keine Kausalität ohne Experiment. Nur Beobachtungsdaten (kein randomisierter Versuch)!

Signifikante Korrelation beweist keine Kausalität – immer kritisch hinterfragen.

Block 2: Zusammenfassung

Konzepte

- Inferenz: Stichprobe $\rightarrow$ Population
- Bootstrap für Konfidenzintervalle
- Permutationstest für Signifikanz
- Typ-I/II Fehler als Business-Risiken
- p-Hacking und FDR
- Effektstaerken

Praktische Skills

- infer-Paket für Bootstrap
- Permutationstests durchführen
- Regressionstabellen interpretieren
- Expected Value berechnen
- p.adjust() für FDR
- Cohen's d berechnen

Kernbotschaft: Signifikanz allein reicht nicht – Effektstaerke und Business-Kontext sind entscheidend!

Block 3: Komplexität, Kausalität & Generalisierung

Vielen Dank für Ihre Aufmerksamkeit!

Fragen?

joerg.osterrieder@fhgr.ch