Statistik für Data Scientists

Vorlesung 8:Hypothesentests II

Prof. Dr. Siegfried Handschuh
DS-NLP
Universität St. Gallen

Recap & Ziele heute

Recap V7: Hypothesentests

Heute

- 👉 StatCoach
- χ^2-Tests
- A/B-Tests.
- Permutationstests
- Effektgrössen

Lernziele heute

Nach dieser Vorlesung kannst du:

- erklären, wie ein A/B-Test aufgebaut ist und warum Randomisierung zentral ist.
- den χ^2-Test bei Ja/Nein-Daten einordnen und seine Idee (O vs. E) verständlich erklären.
- die Grundidee eines Permutationstests beschreiben: Welt unter H_0 durch Mischen simulieren.
- den Unterschied zwischen Signifikanz und Effektstärke klar benennen.
- Cohen’s d, Risk Ratio (RR) und Odds Ratio (OR) interpretieren.
- entscheiden, welchen Test du bei zwei Gruppen und Ja/Nein-Daten sinnvoll einsetzt.

StatCoach

StatCoach

👉 StatCoach = Selbstlernaufgabe

- Selbstlernaufgabe mit StatCoach
- Prüfungsrelevant!
- Beispiel: 👉 StatCoach: Vergleich zweier Verteilungen mit dem χ^2-Test

StatCoach (ohne Symbol) = Optional

- Zusatzmaterial zur Vertiefung.
- Nicht prüfungsrelevant!
- Beispiel: StatCoach: Storytelling-Plots (Ridgeline, Raincloud)

A/B Test

Warum A/B-Tests? (Motivation)

Kleine Änderungen können grosse Effekte haben. Statistik zeigt, ob es echter Unterschied oder Zufall ist.

- Produktstory: Zwei App-Versionen, leichte Designänderung.
- Beobachtung: A = 5% Conversion, B = 7%.
- Frage: Ist B wirklich besser, oder ist das nur Zufall?
- Statistische Idee: Mittelwerte allein reichen nicht, wir brauchen einen Test.
- Prüfungsrelevanz: A/B-Tests sind der Einstieg in Hypothesentests.

Mini-Check: Warum kann man nicht einfach die beiden Mittelwerte vergleichen?

Ablauf eines A/B-Tests

Jeder A/B-Test besteht aus vier einfachen, wiederholbaren Schritten.

1. Randomisierung: Personen werden zufällig A oder B zugeordnet.
2. Messen: Wir erfassen eine Metrik (z. B. Conversion).
3. Testen: Vergleich der Gruppen (t-Test oder χ^2 bei Anteilen).
4. Effektgrösse: Wie stark ist der Unterschied praktisch?
5. Beispielzahlen: A = 5 von 100, B = 11 von 100 (5% vs. 11%).

Mini-Check: Welche Kennzahl würdest du testen: Mittelwertdifferenz oder Anteilsdifferenz?

Von A/B zum X²-Test (Warum Kategorien wichtig sind)

Bei Ja/Nein-Daten führt uns jeder A/B-Test automatisch zum χ^2-Test.

- Viele A/B-Tests arbeiten mit binären Outcomes (Conversion: Ja/Nein).
- Dafür nutzen wir nicht die Mittelwertlogik, sondern Kategorienlogik.
- χ^2 vergleicht beobachtete vs. erwartete Häufigkeiten.
- Wenn B wirklich besser ist, weicht die Kontingenztafel deutlich von der Zufallserwartung ab.

Kleine Tabelle:

Gruppe	Ja	Nein
A	5	95
B	11	89

Mini-Check: Wenn unter H_0 kein Unterschied besteht, wie würden die erwarteten Häufigkeiten aussehen?

Take-Away: A/B-Tests

- A/B-Test = strukturierter Entscheidungsprozess:
- Randomisieren $\rightarrow$ Messen $\rightarrow$ Testen $\rightarrow$ Bewerten.
- Bei binären Outcomes vergleicht man Anteile, nicht Mittelwerte.
- Praxisbezug: Produkt, Conversion, UX, Medizin


X2

X^2-Test

👉 StatCoach:

Vergleich zweier Verteilungen mit dem

χ^2-Test

X²-Test: Grundidee Schritt für Schritt

Der X^2 Test prüft, ob Beobachtungen stärker abweichen als durch Zufall erklärbar.

- Wir vergleichen O_{ij} (observed) mit E_{ij} (expected).
- Unter H_0 gilt: «Kein Unterschied zwischen A und B» $\rightarrow$ gleiche Grundrate.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Interpretation: Je grösser χ^2 , desto unwahrscheinlicher ist H_0
- Freiheitsgrade bei 2×2: (2–1)(2–1) = 1.

Mini-Check: Wie verändert sich X^2, wenn O und E weit auseinanderliegen?

Freiheitsgrad im χ^2-Test (df)

- In einer 2×2-Tabelle sind die Randwerte (Summen) bekannt.
- Dadurch ist nur eine einzige Zelle frei wählbar → alle anderen ergeben sich automatisch aus den Summen.
- Deshalb gilt im 2×2-Fall immer: $df = (2–1)(2–1) = 1$
- Der Freiheitsgrad bestimmt, welche χ^2-Kurve wir zur p-Wert-Bestimmung verwenden.

Beispiel:

	Ja	Nein	Summe
A	a	b	40
B	c	d	40
Summe	50	30	80

Nur eine Entscheidung ist frei: z. B. a = 30 → b, c und d ergeben sich automatisch.

X² Beispiel mit konkreten Zahlen (Kaffee & Geschlecht)

Ein reales 2×2-Beispiel zeigt, wie O und E berechnet werden.

Beobachtet (O):

	Ja	Nein	Summe
Männer	30	10	40
Frauen	20	20	40

Gesamt: 50 Ja, 30 Nein, 80 Personen.

Erwartete Werte (unter H_0):

- $E_{11} = \frac{40 \cdot 50}{80} = 25, \quad E_{12} = 15$
- $E_{21} = 25, \quad E_{22} = 15$

χ²-Berechnung (Skizze):

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(10-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(20-15)^2}{15} \approx 5.33$$

p ≈ 0.02 → Zusammenhang plausibel (signifikant auf 5%-Niveau).

Mini-Check: Warum ist die Abweichung «30 statt 25 Ja bei Männern» statistisch relevant?

Mini-Check: Erwartete Häufigkeiten berechnen

E_{ij}-Berechnung ist simpel:
man braucht nur Zeilen-, Spaltensummen und N.

$$E_{ij} = \frac{(\text{Zeilensumme}_i)(\text{Spaltensumme}_j)}{N}$$

Beispiel:

	Ja	Nein	Summe
Gruppe A	?	?	45
Gruppe B	?	?	55
Summe	72	28	100

- Aufgabe: Berechne $E_{A,Ja}$ und $E_{B,Nein}$.
- Hinweis: Rundung erst am Ende.
- Mini-Check: $E_{A,Ja}$ = ?

Take-Away: X²-Test

- Der X²-Test prüft, ob beobachtete Häufigkeiten stärker abweichen als durch Zufall erklärbar.
- Grosse Abweichung zwischen O und E → hoher χ^2-Wert → kleiner p-Wert.
- Ideal für 2×2-Tabellen und alle Fragen: «Hängen diese Kategorien zusammen?»

Permutationstest

Permutationstest: Intuition

Ein Permutationstest simuliert die Welt, in der es keinen Unterschied gibt und prüft, ob unsere Daten ungewöhnlich wären.

- Unter H_0 gilt: «A und B sind gleich» $\rightarrow$ Labels A/B sind austauschbar.
- Idee: Wir mischen die Gruppenzuweisung tausendfach neu.
- Jede Neuzuweisung erzeugt eine neue Differenz Δ^*
- Vergleich: Ist die beobachtete Differenz Δ extrem im Vergleich zu den Δ^*?
- Vorteil: Funktioniert ohne Normalverteilungsannahmen.

Mini-Check: Warum ist das Durchmischen der Labels eine Simulation der Nullhypothese?

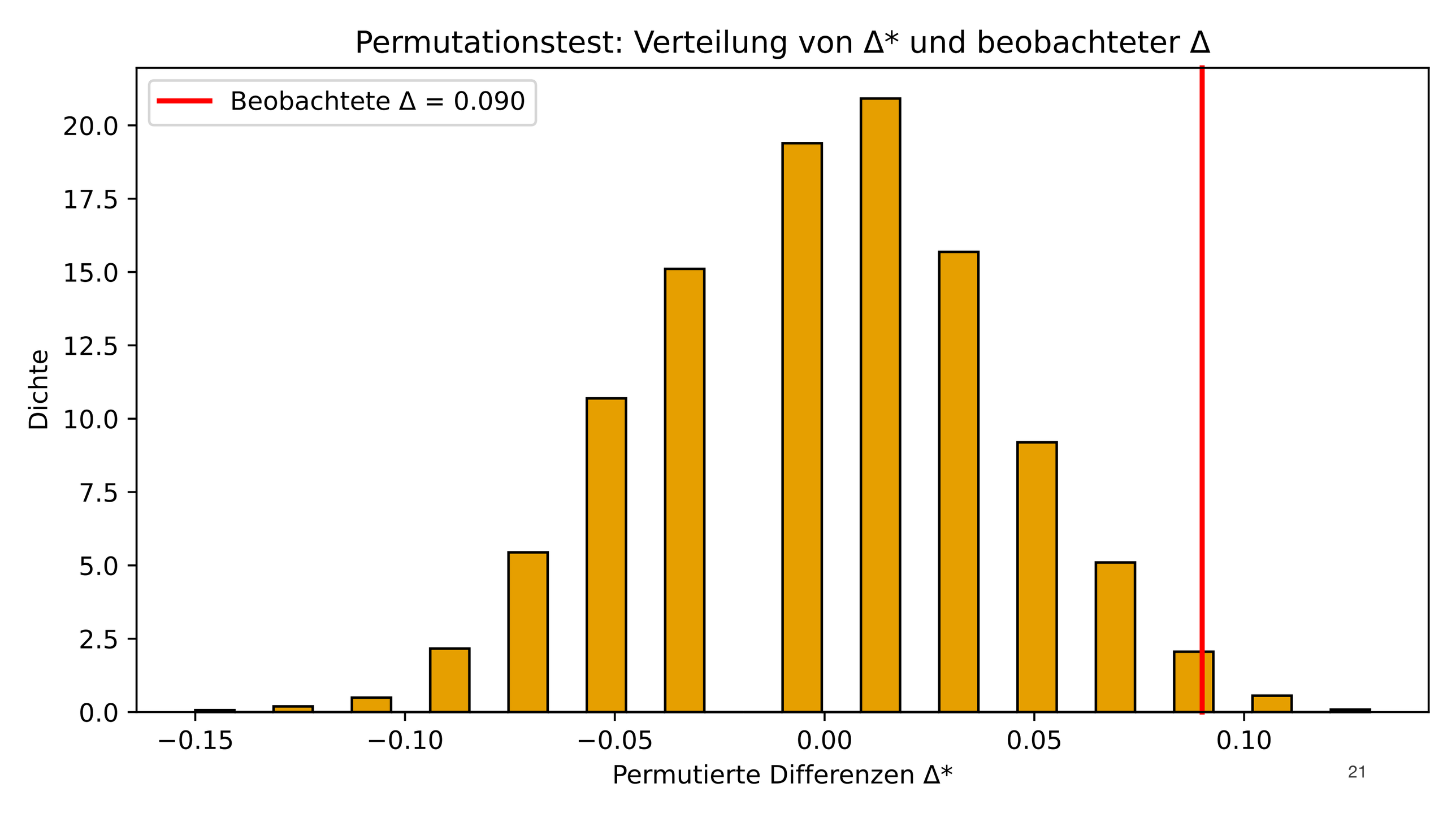
Permutationstest: Schrittfolge (4 einfache Schritte)

Permutationstests folgen einem klaren, allgemeinen Algorithmus.

1. Beobachtete Differenz Δ berechnen, z. B. Mittelwert(B) – Mittelwert(A).
2. Labels mischen: Daten werden zufällig auf Gruppen verteilt.
3. Neue Differenz Δ^* berechnen, Schritte 2 und 3 werden sehr häufig wiederholt (z.B. 1000-5000 Mal)
4. p-Wert: Anteil der Δ^* mit $|\Delta^*| \geq |\Delta|$.

Interpretation: Kleiner p, die beobachtete Differenz wäre selten unter H_0.

Mini-Check: Warum steigt die Aussagekraft, wenn wir viele Permutationen machen, z. B. 5000?


Permutationstest: Verteilung von Δ* und beobachteter Δ
Beobachtete Δ = 0.090
Dichte
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−0.15
−0.10
−0.05
0.00
0.05
0.10
Permutierte Differenzen Δ*

Permutationstest: kompakter Python-Code

```
import numpy as np

np.random.seed(42)

A = np.random.binomial(1,0.05,100)
B = np.random.binomial(1,0.11,100)


data = np.concatenate([A,B])
nA = len(A)

obs = A.mean() - B.mean()

diffs = []
for _ in range(5000):
  np.random.shuffle(data)
  diffs.append(data[:nA].mean() - data[nA:].mean())

p = np.mean(np.abs(diffs) ⩾ abs(obs))
print(p)
```

Permutationstest: kompakter Python-Code

Der Permutationstest ist programmatisch einfach

```python
import numpy as np

np.random.seed(42)

A = np.random.binomial(1,0.05,100)
B = np.random.binomial(1,0.11,100)


data = np.concatenate([A,B])
nA = len(A)

obs = A.mean() - B.mean()

diffs = []
for _ in range(5000):
  np.random.shuffle(data)
  diffs.append(data[:nA].mean() - data[nA:].mean())

p = np.mean(np.abs(diffs) ⩾ abs(obs))
print(p)
```

- Hinweis: Für Conversions arbeitet man mit Anteilen.
- Wichtig: Interpretation des p-Werts bleibt identisch wie im t- oder χ^2 -Test.

Mini-Check: Was sagt ein p-Wert von z. B. 0.03 hier aus?

Take-Away: Permutationstests

- Permutationstests simulieren direkt die Nullhypothese durch zufälliges Mischen.
- Keine Modellannahmen: robust bei Schiefe, Ausreissern oder kleinen Stichproben.
- Der p-Wert ist empirisch: Wie ungewöhnlich ist unser Effekt in einer zufälligen Welt?

Effektgrössen

Warum Effektgrössen? (Signifikanz ≠ Relevanz)

Ein signifikanter Unterschied kann völlig irrelevant sein.
Effektgrössen zeigen, wie gross ein Unterschied wirklich ist.

- p-Werte sagen nur: «Unterschied auffällig?», nicht: «Unterschied wichtig?»
- Bei grossen Stichproben wird fast alles signifikant.
- Effektgrössen messen Stärke eines Effekts, unabhängig von n.
- Beispiel: A=5%, B=6% bei n=10’000 → hoch signifikant, winziger Effekt.

Cohen’s d – Effektgrösse für Mittelwerte

Cohen’s d misst die Distanz zweier Mittelwerte in Einheiten der gemeinsamen Streuung.

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

Pooled SD:

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

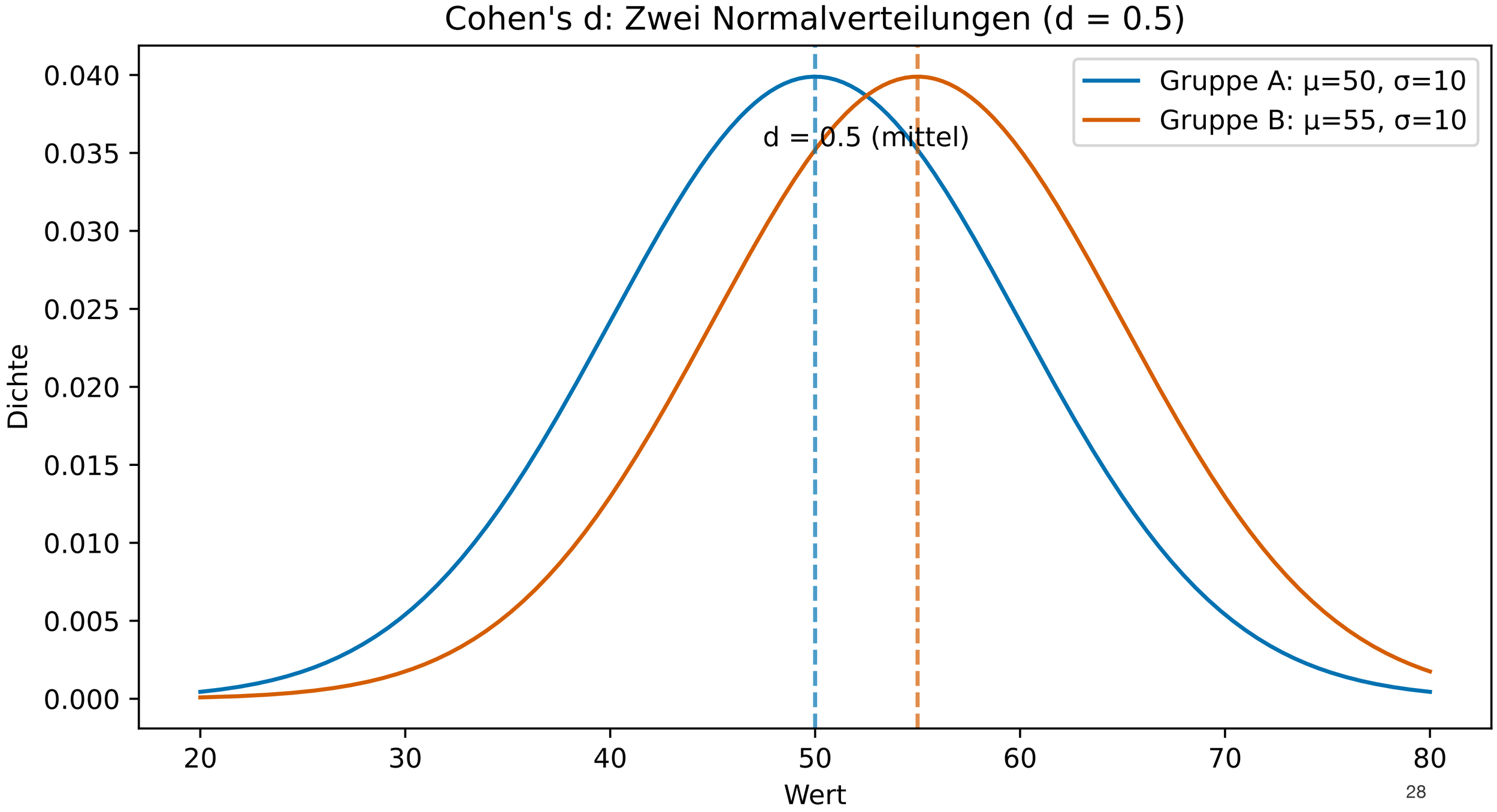
Richtwerte (Cohen): 0.2 klein, 0.5 mittel, 0.8 gross.

Gruppe A: M=50, SD=10;
Gruppe B: M=55, SD=10

$$d = \frac{(55 - 50)}{10} = 0.5 \Rightarrow \text{mittlerer Effekt}$$

d misst Stärke, nicht Signifikanz.

Mini-Check: Was bedeutet d=0.8 in Worten?


Cohen's d: Zwei Normalverteilungen (d = 0.5)
Gruppe A: μ=50, σ=10
Gruppe B: μ=55, σ=10
d = 0.5 (mittel)
Dichte
Wert
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
80

Odds & Odds Ratio, bspw. Fussballwetten ⚽

Odds sind Chancenverhältnisse:

$$\text{Odds} = \frac{P(\text{Ereignis})}{P(\text{nicht-Ereignis})}$$

Statistische Odds: 1:4 → sehr geringe Chance (1 Erfolg, 4 Misserfolge). Für 1 CHF Einsatz bekommst du 4 CHF Gewinn

(Fussballwetten nutzen genau dieses Prinzip, aber schreiben es umgekeht als Payout System auf, also 4:1 in diesem Fall)

Mathematisch entspricht z. B. 1:4:

$$\text{Odds} = \frac{1}{4} = 0.25$$

- Odds Ratio (OR) vergleicht die Chancen zweier Teams/ Gruppen:
  $$OR = \frac{\text{Odds}_1}{\text{Odds}_2}$$
- Beispiel:
  Team A: 1:4 → $\text{Odds}_A = 0.25$
  Team B: 1:2 → $\text{Odds}_B = 0.50$
  $$OR = \frac{0.25}{0.50} = 0.5$$
  → Team A hat halb so hohe Chance wie Team B.
- Merksatz:
  Odds sind wie Sportwetten-Verhältnisse.
  Die Odds Ratio vergleicht diese Chancen zwischen zwei Gruppen.

Risk Ratio & Odds Ratio: Effektgrössen für Anteile

Bei Ja/Nein-Daten sind Risk Ratio und Odds Ratio die passenden Effektgrössen.

Risk Ratio (RR): $RR = \frac{p_1}{p_2}$

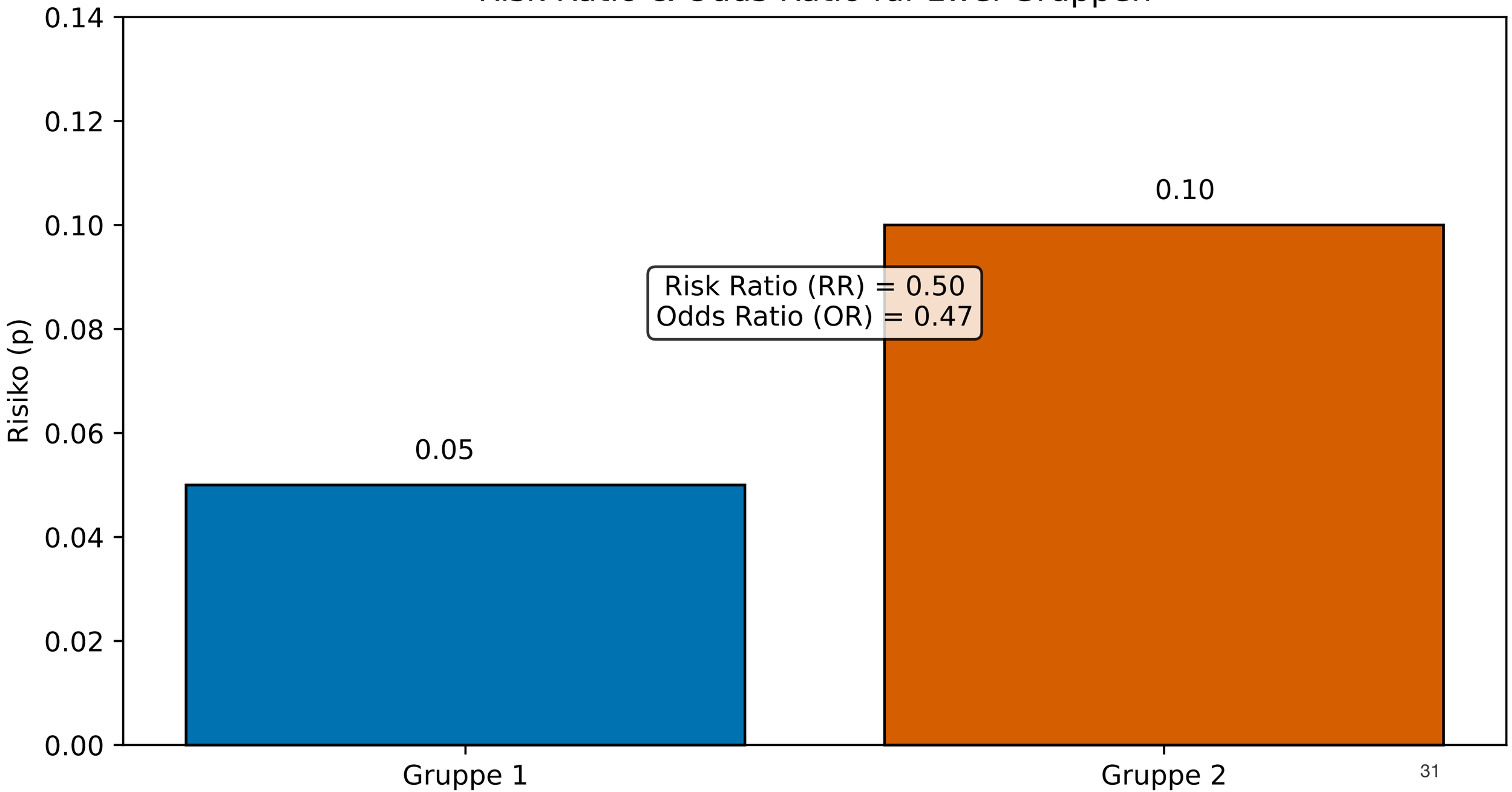
Beispiel: Impfgruppe 1% Risiko,
Kontrollgruppe 3% Risiko $\rightarrow$ RR = 0.33

Odds Ratio (OR): $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

Interpretation:
- RR < 1 → Gruppe 1 hat geringeres Risiko.
- OR nützlich bei seltenen Ereignissen.

Beispiel:
p_1=0.05, p_2=0.10 $\rightarrow$ $RR = 0.5, \quad OR \approx 0.47$

Mini-Check: Wenn RR = 0.5, was bedeutet das praktisch?


Risk Ratio & Odds Ratio für zwei Gruppen
Risiko (p)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.05
0.10
Risk Ratio (RR) = 0.50
Odds Ratio (OR) = 0.47
Gruppe 1
Gruppe 2

Mini-Check: Effektgrössen in der Praxis

Effektgrössen helfen bei der Bewertung, ob ein Unterschied praktisch relevant ist.

- Szenario: A=5% Conversion, B=7% Conversion.
- Statistisch: Unterschied signifikant bei n=2000.
- Praktisch: Wie gross ist der Effekt wirklich?

Effektgrösse Anteil:

$RR = \frac{0.07}{0.05} = 1.4 \rightarrow$ Gruppe B hat 40% höhere Conversion.

Entscheidungslogik: «Lohnt sich die Änderung?» hängt von RR, nicht vom p-Wert ab.

Mini-Check: Wenn RR=1.05, ist das wahrscheinlich praktisch relevant?

Take-Away: Effektgrössen

- Signifikanz $\neq$ Relevanz: grosse $n \rightarrow$ fast alles wird signifikant.
- Effektgrössen messen die Grösse eines Unterschieds, nicht nur seine Existenz.
- d, RR und OR zeigen, ob ein Effekt praktisch bedeutsam ist.

Zusammenfassung

Zusammenfassung

- A/B-Tests: strukturierter Weg zu datenbasierten Entscheidungen.
- χ^2-Test: vergleicht beobachtete vs. erwartete Häufigkeiten bei kategorialen Daten.
- Permutationstest: robuste Alternative ohne Modellannahmen.
- Effektgrössen (d, RR, OR): messen praktische Relevanz.
- Prüfen $\neq$ Verstehen: Signifikanz allein reicht nicht.

Mini-Check: Welchen Test würdest du bei zwei Gruppen und Ja/Nein-Daten nutzen?

Quiz: Aktive Wiederholung

Kahoot Quiz VL8: Hypothesentests II

Projekte

Kürzel	Projekt / Datensatz
TAX	Yellow Taxi Trip Data (NYC Taxis)
FRD	Credit Card Transactions & Fraud Detection
UNH	UNHCR Refugee Data (Persons of Concern, Asylum etc.)
EUR	Eurostat Labour Market / Employment Statistics
PIS	PISA: Socioeconomic Status & Academic Performance
IMF	IMF World Economic Outlook (Macro Indicators)
NYC	NYC Parking Violations
WHO	WHO Mortality Database
CHI	Chicago 311 Service Requests
FSH	Global Fish Consumption (FAO Food Balance Sheets)

Methoden für alle Projekte (VL9–VL11)

Methode / Test	TAX	FRD	UNH	EUR	PIS	IMF	NYC	WHO	CHI	FSH
ANOVA / Kruskal (VL9)		X		X	X			X	X	
FDR (VL9)		X		X				X	X	
Bootstrap CI (VL10)	X	X	X			X				X
Permutation (VL10)		X	X		X	X		X	X	
Regression (VL10)			X	X	X	X		X	X	
Robuste Tests (VL11)		X	X			X	X	X	X	X
Outlier / robuste Lage (VL11)	X		X			X	X	X		X

Ausblick

Nächste Woche: VL 9: Methoden für Projekte mit vielen Gruppen und vielen Tests

ANOVA: Analysis of Variance

- Relevant für: Eurostat Labour Market, WHO Mortality Database, PISA (Sozioökonomischer Status), Credit Card Fraud, Chicago 311 Service Requests

FDR: False Discovery Rate

- Relevant für: Credit Card Fraud, Chicago 311 Service Requests, WHO Mortality Database, Eurostat Labour Market