

Statistik für Data Scientists

Vorlesung 9: Gruppenvergleiche & Multiple Tests

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

Recap & Ziele heute

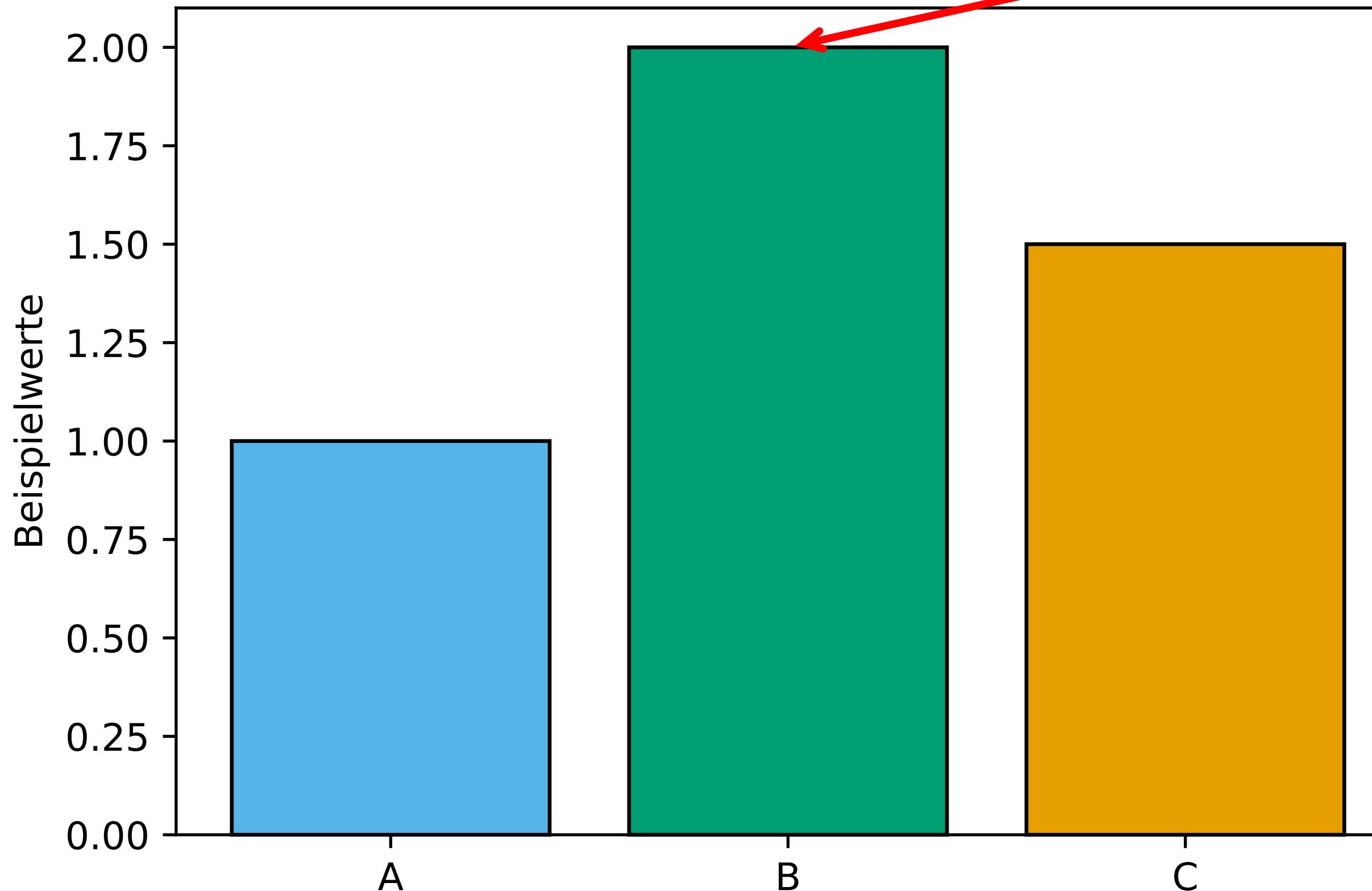
Wir erweitern die Vergleichslogik von zwei auf mehrere Gruppen.

- Recap V8: A/B, χ^2 , Permutation, Effektgrößen
- Heute: ANOVA, Kruskal, Post-hoc, Multiple Tests, FDR
- Fokus: Entscheidungslogik für 3+ Gruppen
- Ziel: passende Methode für Datentyp und Projekt finden

Mini-Check: Warum reicht ein t-Test nicht für drei Gruppen?

t-Test reicht nicht

Warum ein t-Test nicht für 3 Gruppen reicht

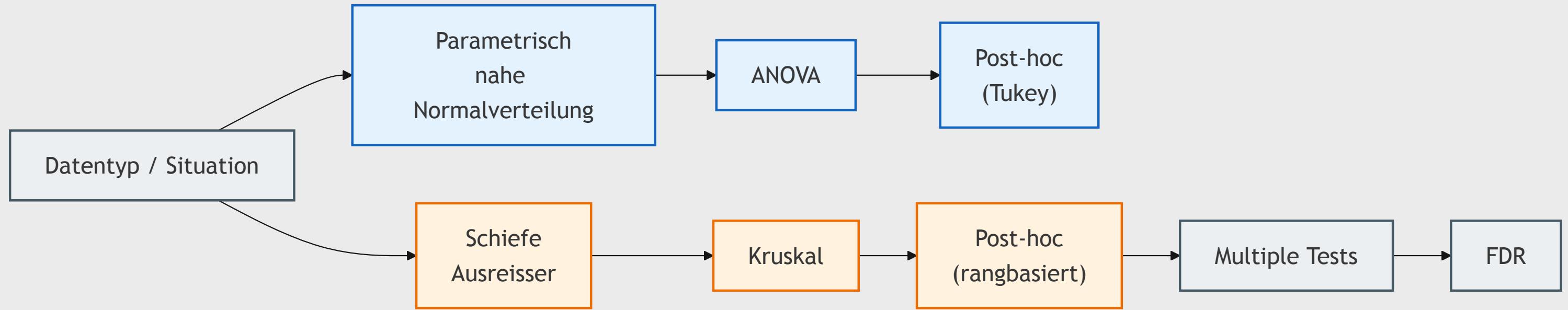


Lernziele heute

Du entscheidest am Ende sicher, welche Methode wann passt.

- ANOVA verstehen und korrekt anwenden
- Kruskal als robuste Alternative kennen
- Post-hoc Tests interpretieren können
- Multiple Tests und α -Inflation erkennen
- FDR nach Benjamini–Hochberg praktisch einsetzen
- Transfer zur Projektarbeit

Mini-Check: Welche Information liefert ANOVA **nicht**?



Motivation

Motivation: Gruppenvergleiche

Viele reale Datensätze haben mehr als zwei Gruppen: dafür reichen t-Tests nicht.

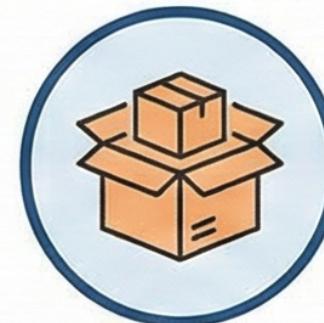
- Beispiele: Ländergruppen, Kategorien, Level, Kohorten
- t-Tests funktionieren nur für zwei Gruppen
- Viele Paarvergleiche → starke Fehlerinflation
- Wir brauchen ein **gemeinsames Modell** für 3+ Gruppen
- Ziel: prüfen, ob Gruppen sich insgesamt unterscheiden

Mini-Check: Was passiert bei 20 Tests mit $\alpha = 0.05$?

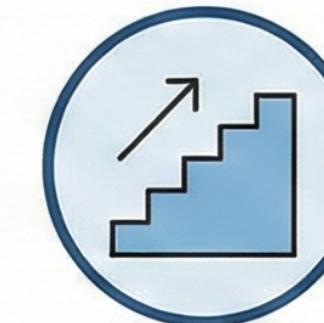
Das Problem: Mehr als zwei Gruppen



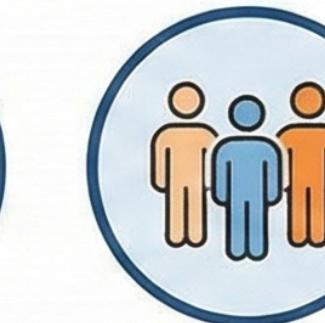
Ländergruppen Kategorien



Ländergruppen Kategorien



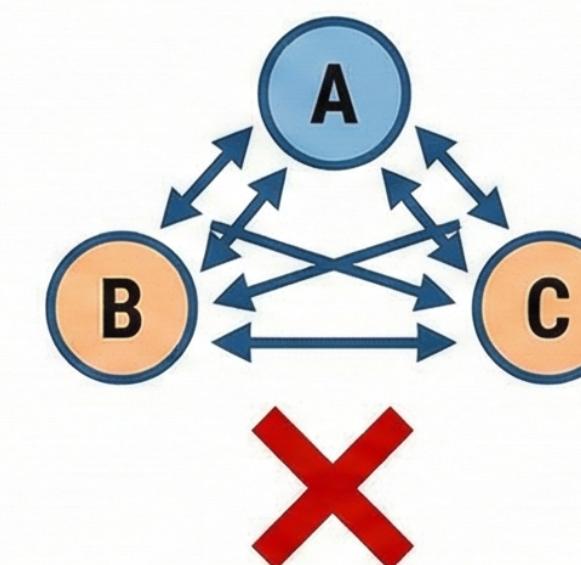
Level



Kohorten

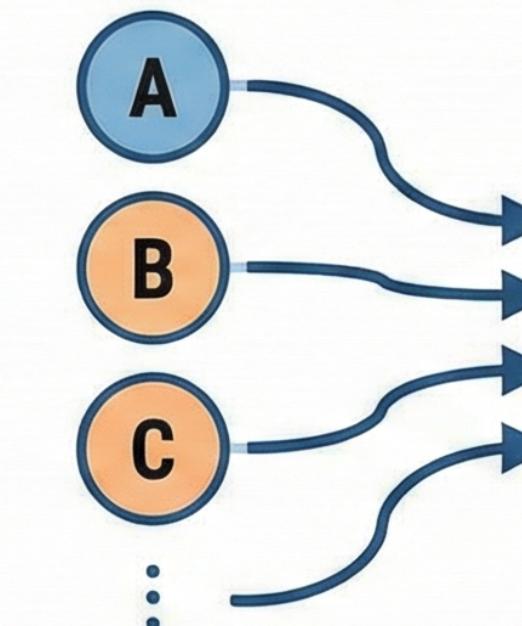


t-Tests: Nur für 2 Gruppen



Viele Paarvergleiche
→ Starke Fehlerinflatio

Die Lösung & Ein Beispiel



GEMEINSAMES MODELL (z.B. ANOVA)

Ziel: Prüfen, ob Gruppen sich insgesamt unterscheiden

MINI-CHECK



1 "False Positive" (bei $\alpha=0.05$)

Was passiert bei 20 Tests?

Wahrscheinlichkeit für ≥ 1 Fehler $\approx 64\%$

Take-away: Einzeltests skalieren nicht für viele Gruppen

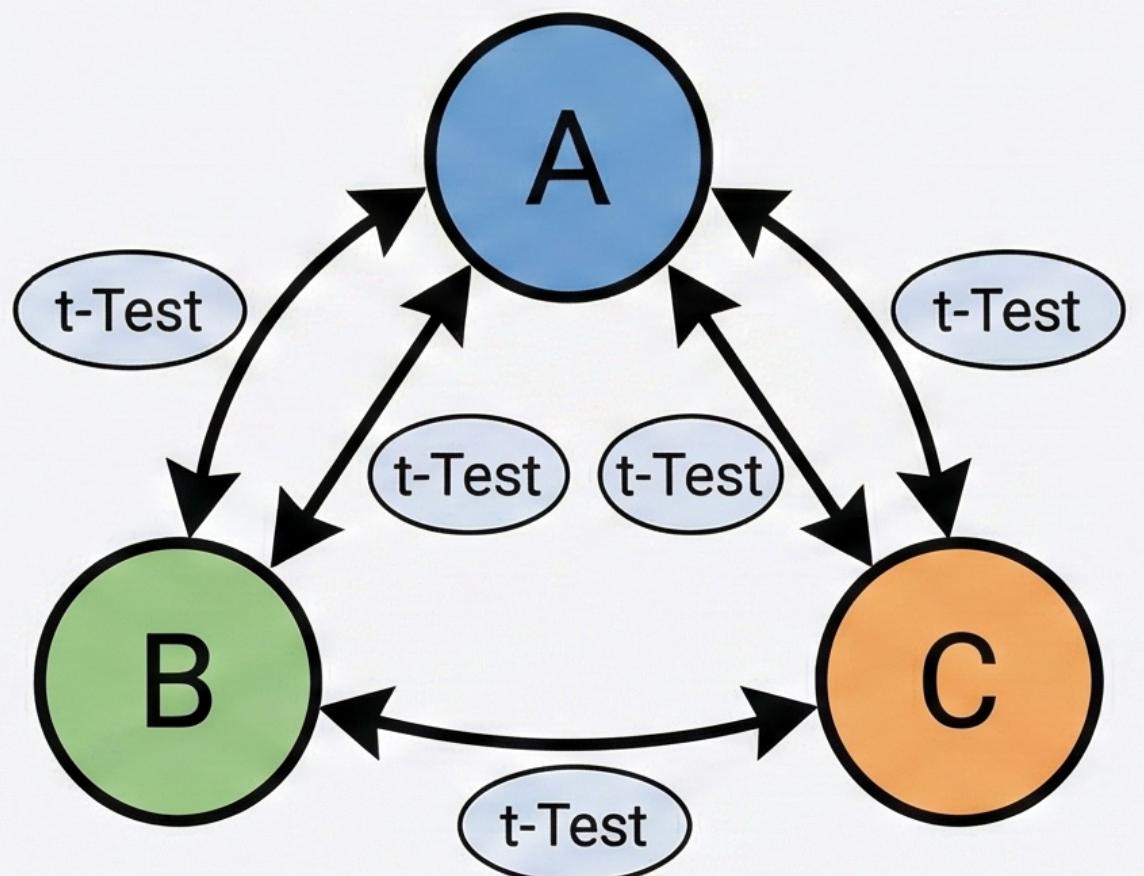
Warum Gruppenvergleiche?

Mehrere Gruppen → ein gemeinsamer Test statt vieler Einzelvergleiche.

- Viele Projekte haben 3+ Gruppen (PISA, WHO, Regionen)
- Paarweise Tests wären ineffizient und fehleranfällig
- Ein Gesamtmodell berücksichtigt alle Gruppen gleichzeitig
- Fokus: globale Unterschiede → danach lokale Paarvergleiche
- Ideale Werkzeuge: ANOVA oder Kruskal

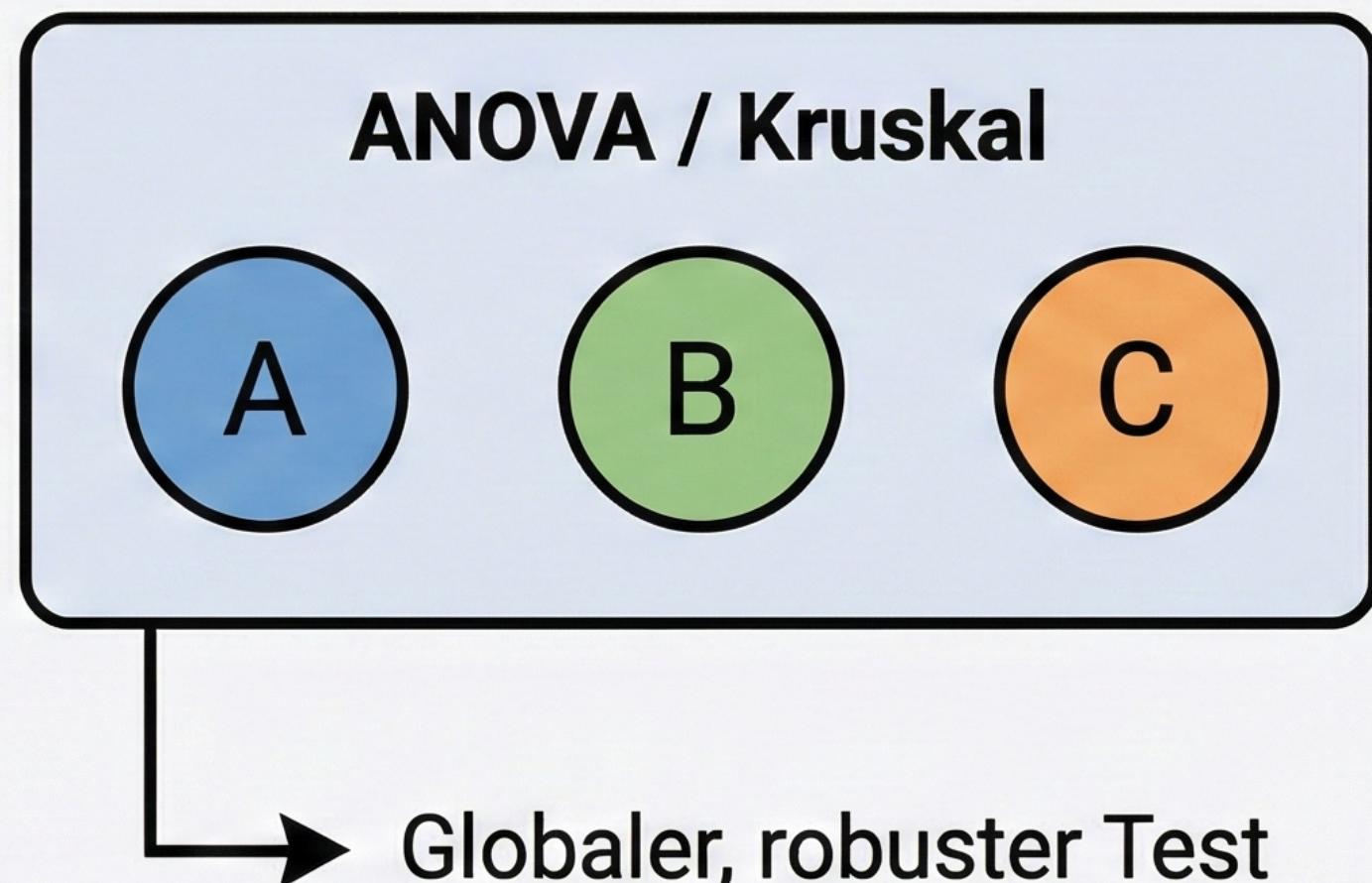
Mini-Check: Warum ist ein einzelner Gesamttest besser als viele paarweise t-Tests?

Viele Einzelvergleiche (Ineffizient)



Fehleranfällig & Aufwendig

Ein gemeinsamer Test (Effizient)



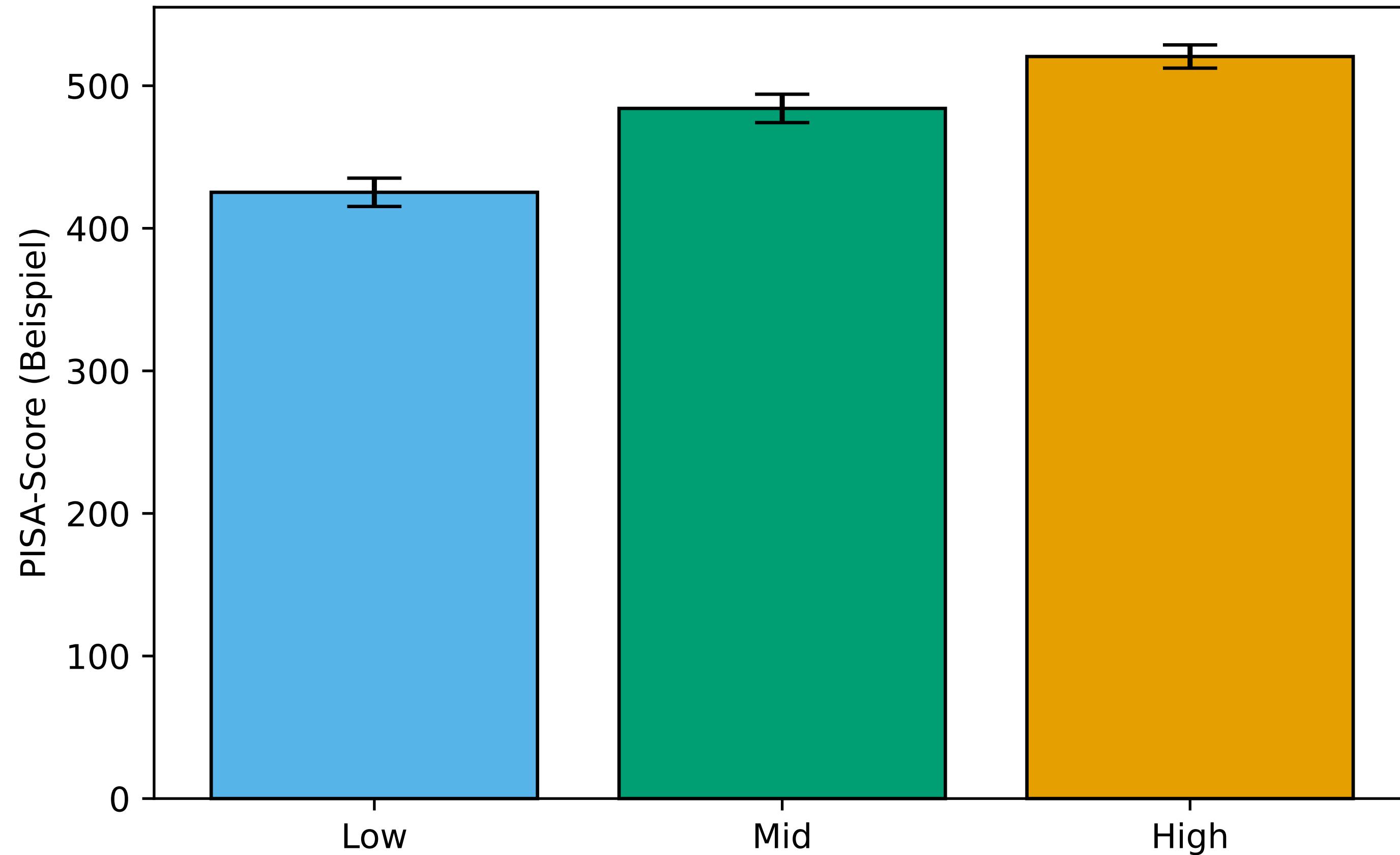
Beispiel: PISA SES-Level

Drei oder mehr Gruppen verlangen ein Modell, das alle gleichzeitig prüft.

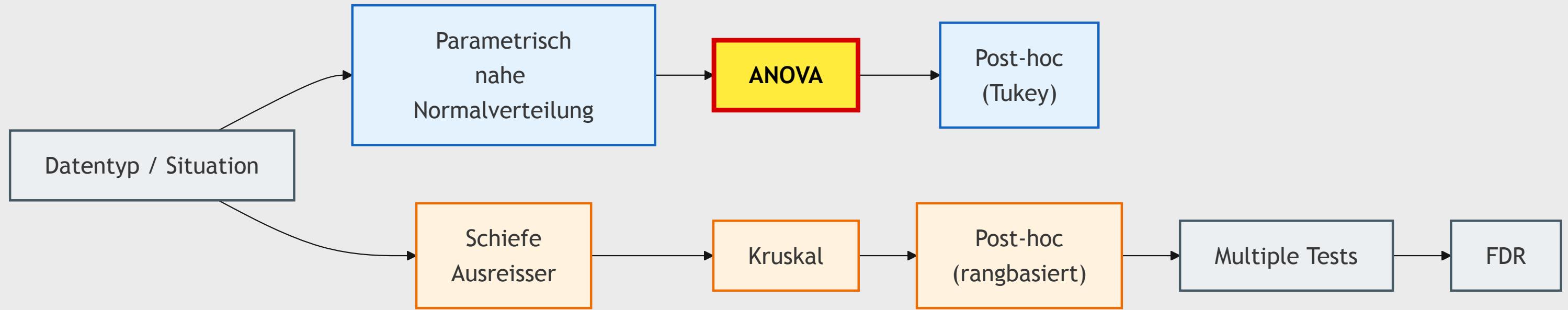
- **SES** (Socioeconomic Status)-Level: Low, Mid, High
- Frage: Unterscheiden sich die Scores zwischen den drei Gruppen?
- t-Tests ungeeignet → nur 2 Gruppen möglich
- **Lösung**: ANOVA (parametrisch) oder Kruskal (robust)
- Datenstruktur entscheidet, welche Methode passt

Mini-Check: Welche Art von Variable ist SES?

PISA-Beispiel: Scores nach SES-Level (95%-Konfidenzintervalle)



ANOVA



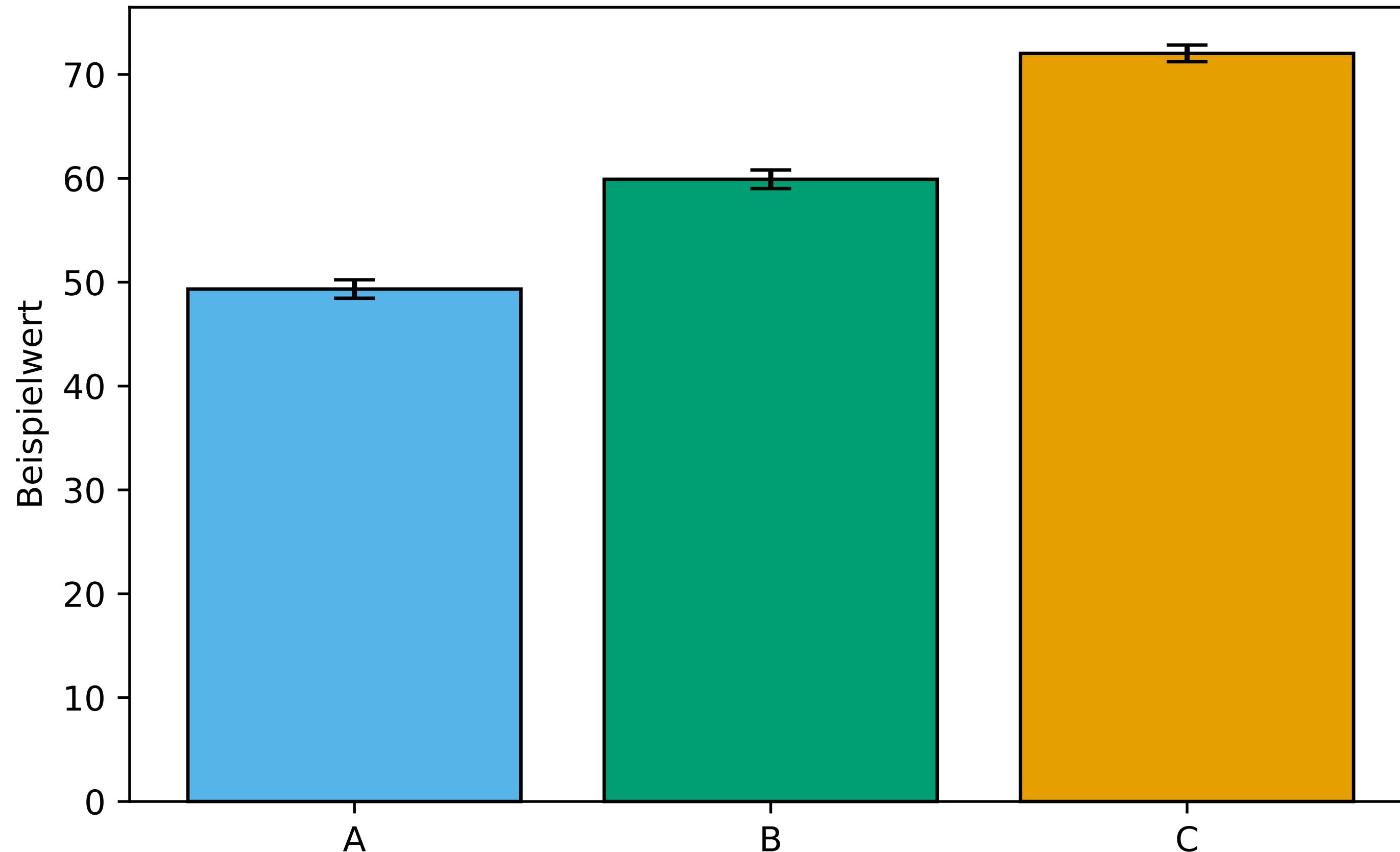
Motivation: ANOVA

ANOVA prüft globale Mittelwertunterschiede für 3+ Gruppen.

- ANOVA = **AN**alysis **O**f **V**Ariance (Varianzanalyse)
- Frage: Gibt es **mindestens eine** Gruppe mit anderem Mittelwert?
- Kernidee: Verhältnis von Gruppen- zu Innenvarianz
- Ergebnis: eine gemeinsame Teststatistik F
- Basis für Post-hoc Tests
- Effizienter als viele t-Tests

Mini-Check: Wann ist ANOVA sinnvoll?

ANOVA-Motivation: Drei Gruppen, unterschiedliche Mittelwerte



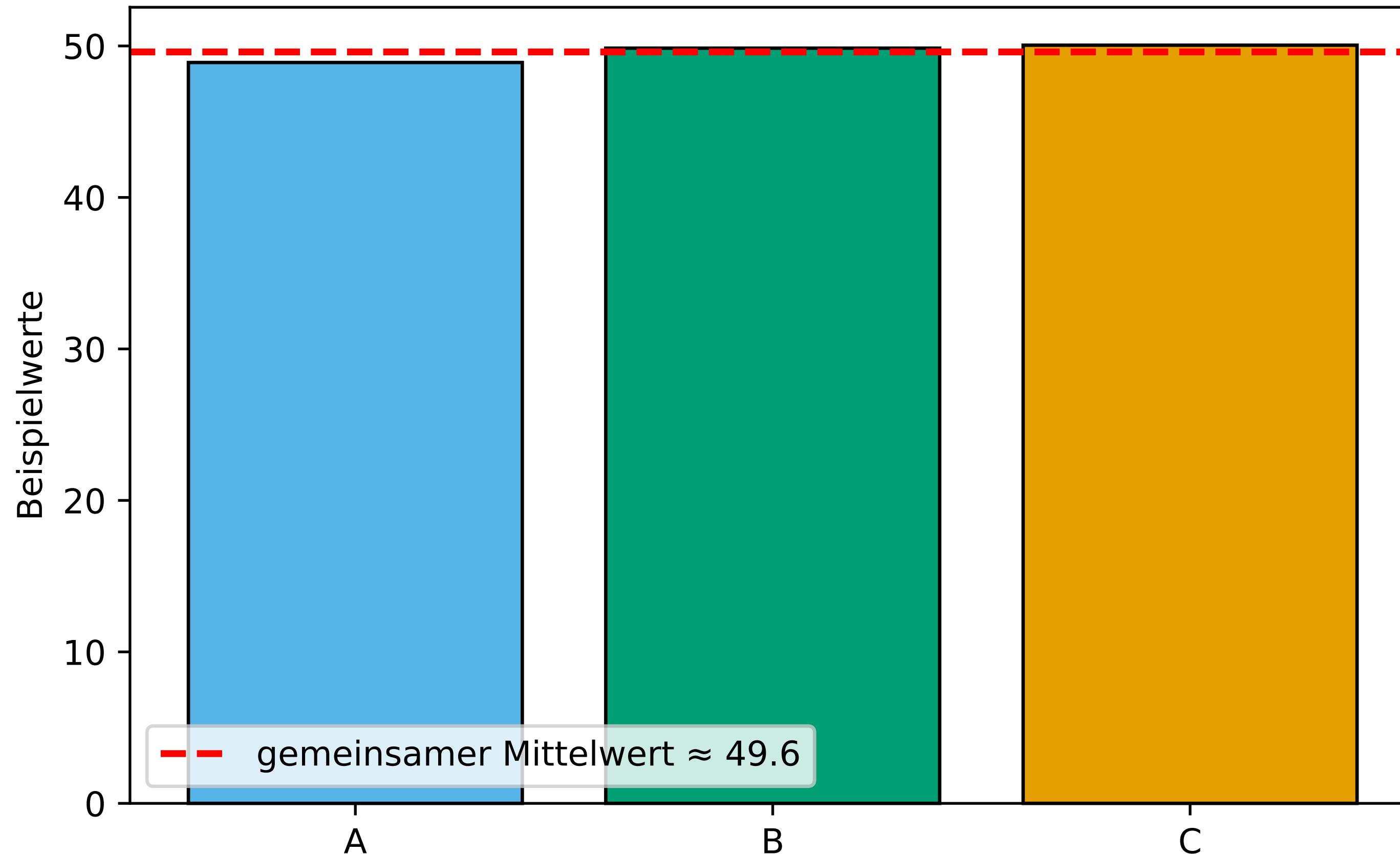
ANOVA – Grundidee

ANOVA vergleicht Varianz zwischen Gruppen mit der Varianz innerhalb der Gruppen.

- Between-Varianz: Unterschiede der Gruppenmittelwerte
- Within-Varianz: Streuung innerhalb jeder Gruppe
- Teststatistik (MS=Mean Square): $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$
- Hoher $F \rightarrow$ Gruppen unterscheiden sich stärker als zufällig erwartet

Mini-Check: Was passiert mit F , wenn alle Gruppenmittel gleich sind?

ANOVA - Grundidee: Alle Gruppen haben denselben Mittelwert

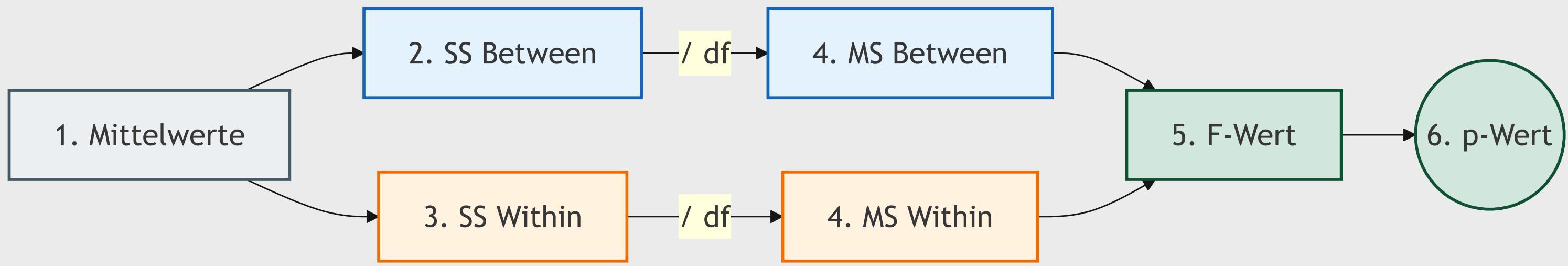


ANOVA – Ablauf

ANOVA folgt einer klaren Schrittfolge, die auf Varianzzerlegung basiert.

- Schritt 1: Gruppenmittelwerte berechnen
- Schritt 2: Between-Varianz ($SS = \text{Sum of Squares}$) SS_{between} bestimmen
- Schritt 3: Within-Varianz SS_{within} bestimmen
- Schritt 4: Mittelquadrate berechnen: $MS = SS/df$
- Schritt 5: $F = MS_{\text{between}}/MS_{\text{within}}$
- Schritt 6: p -Wert berechnen

Mini-Check: Warum brauchen wir Freiheitsgrade df ?



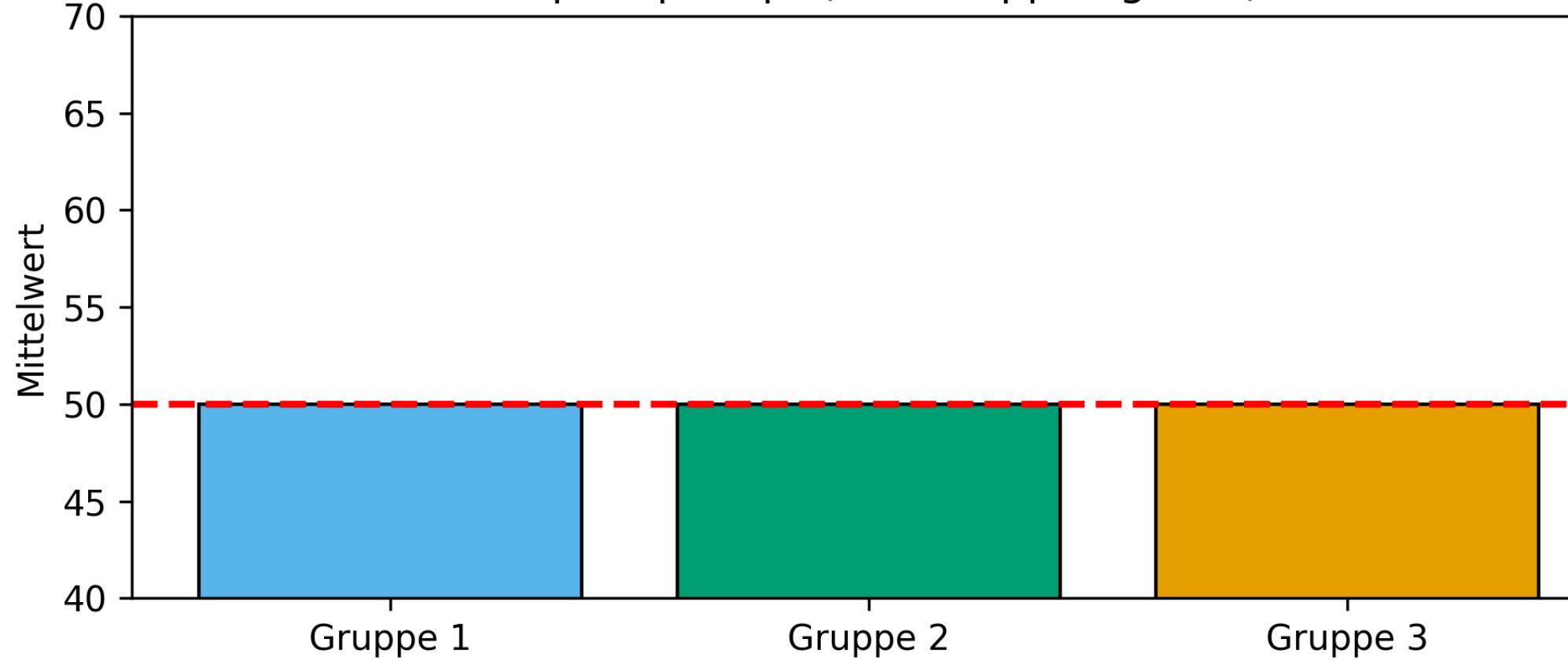
H_0 und H_1 bei ANOVA

ANOVA testet eine globale Nullhypothese über alle Gruppenmittelwerte.

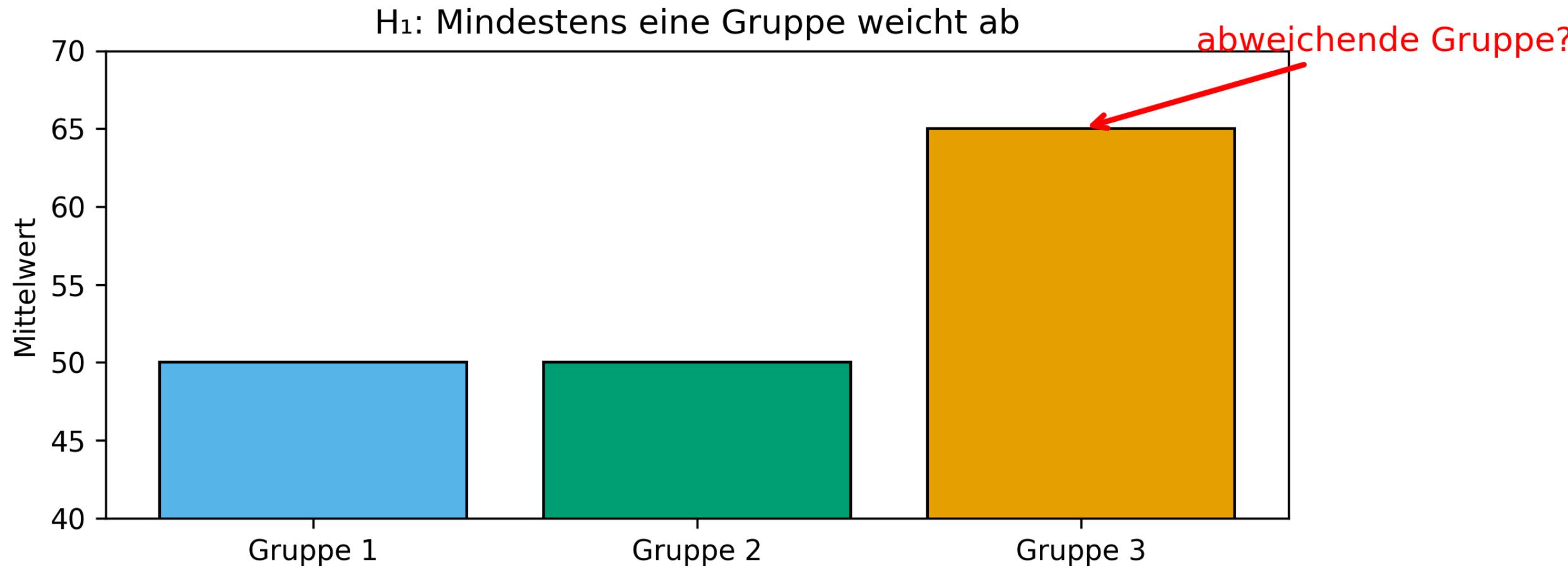
- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
- H_1 : Mindestens eine Gruppe weicht ab
- ANOVA sagt **nicht**, welche Gruppe anders ist
- Dafür braucht es Post-hoc Tests
- Ergebnis: ein globaler F -Test

Mini-Check: Warum brauchen wir trotz signifikantem F Post-hoc Tests?

$H_0: \mu_1 = \mu_2 = \mu_3$ (alle Gruppen gleich)



$H_1:$ Mindestens eine Gruppe weicht ab



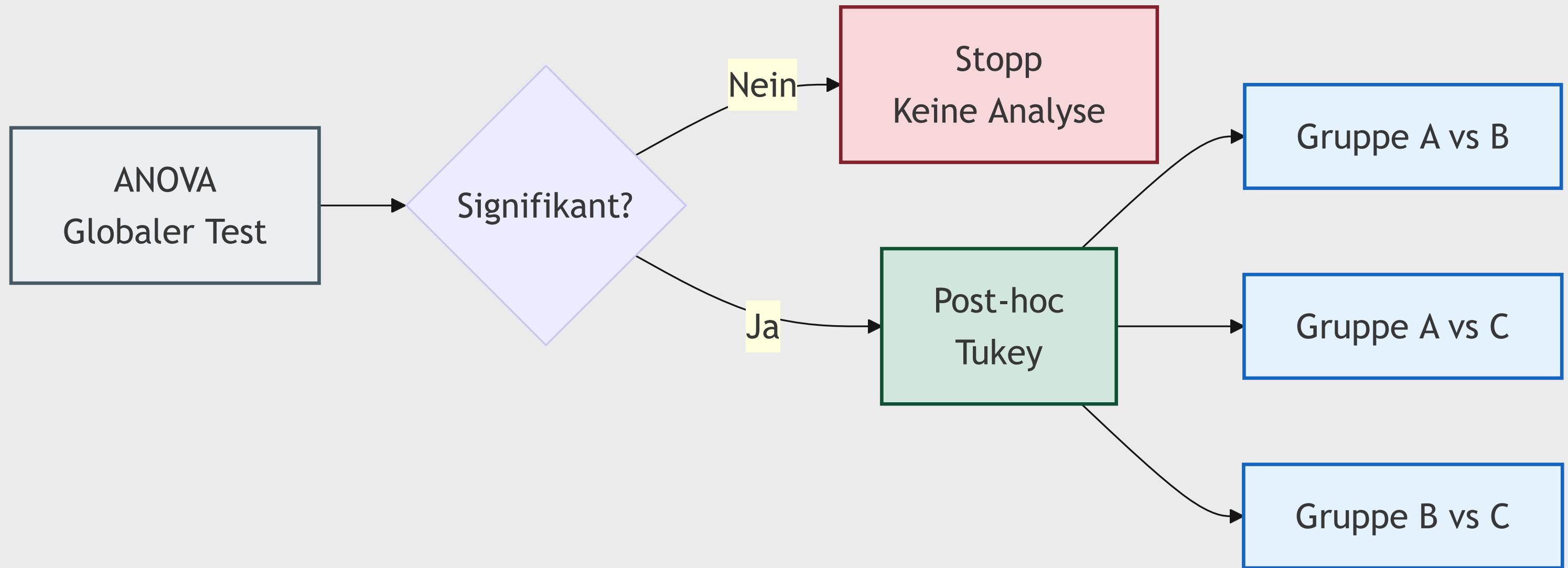
Post-hoc Test

Motivation: Post-hoc Tests

Post-hoc Tests beantworten **wo** genau Unterschiede bestehen.

- Nur sinnvoll nach signifikanter ANOVA
- Vergleichen alle relevanten Gruppenpaare
- Kontrollieren Fehlerwahrscheinlichkeit
- Zeigen Richtung und Stärke der Unterschiede
- Typische Methoden: Tukey HSD

Mini-Check: Warum dürfen Post-hoc Tests nicht ohne ANOVA gestartet werden?



Post-hoc: Tukey HSD

Tukey vergleicht alle Gruppenpaare stabil und kontrolliert Fehler 1. Art.

- HSD = Honest Significant Difference
- Nutzt gemeinsame Varianzschätzung aus der ANOVA
- Liefert für jedes Paar: Differenz, Konfidenzintervall, Signifikanz
- Robust bei vielen Paarvergleichen
- Standard für Post-hoc nach ANOVA

Mini-Check: Warum ist eine gemeinsame Varianzschätzung sinnvoll?

Beispiel: ANOVA + Tukey HSD

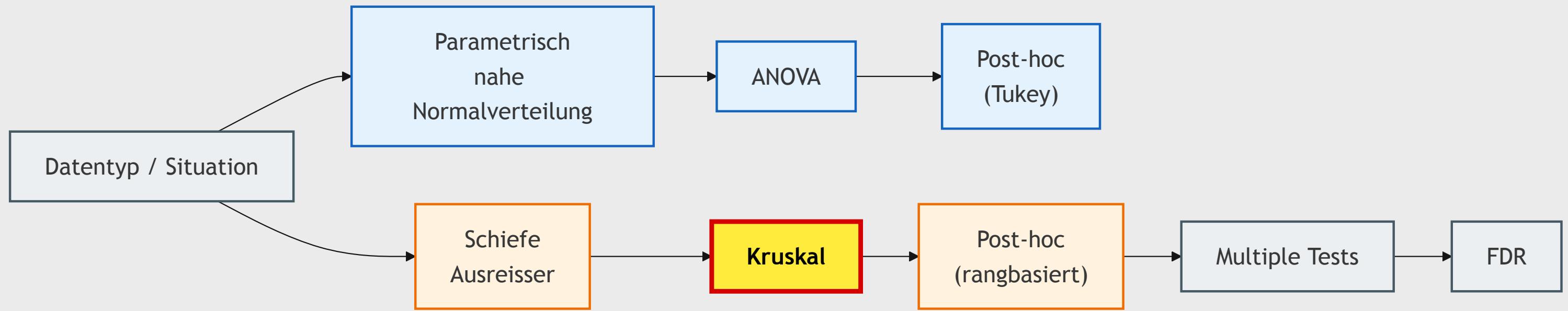
ANOVA entdeckt den globalen Unterschied, Tukey zeigt die konkreten Paarunterschiede.

- Daten: Drei Gruppen (A, B, C) mit unterschiedlichen Mittelwerten
- ANOVA-Ergebnis:
 - $F = 249.203$
 - $p = 6.65 \times 10^{-43}$
 - → globaler Unterschied hochsignifikant
- Tukey HSD (paarweise Vergleiche):
 - A–B: **10.95 Punkte**, signifikant
 - A–C: **23.14 Punkte**, signifikant
 - B–C: **12.20 Punkte**, signifikant
- Interpretation: Alle drei Gruppen unterscheiden sich voneinander.

Mini-Check: Was liefert Tukey, was ANOVA nicht liefert?

ANOVA zeigt ob,
Post-hoc zeigt wo Unterschiede liegen.

Kruskal-Wallis



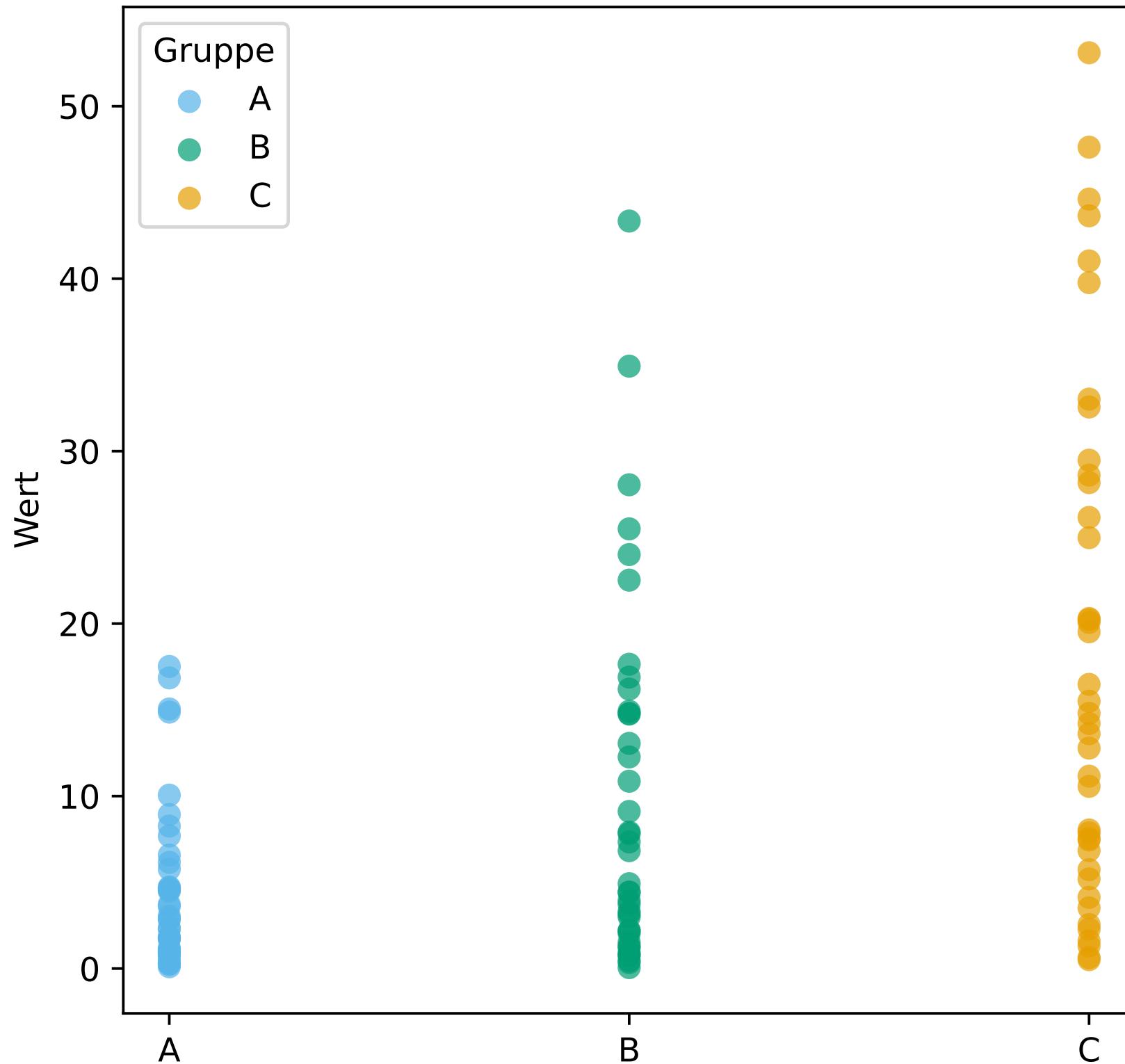
Motivation: Kruskal-Wallis

Kruskal ist die robuste Alternative zur ANOVA, wenn Verteilungsannahmen verletzt sind.

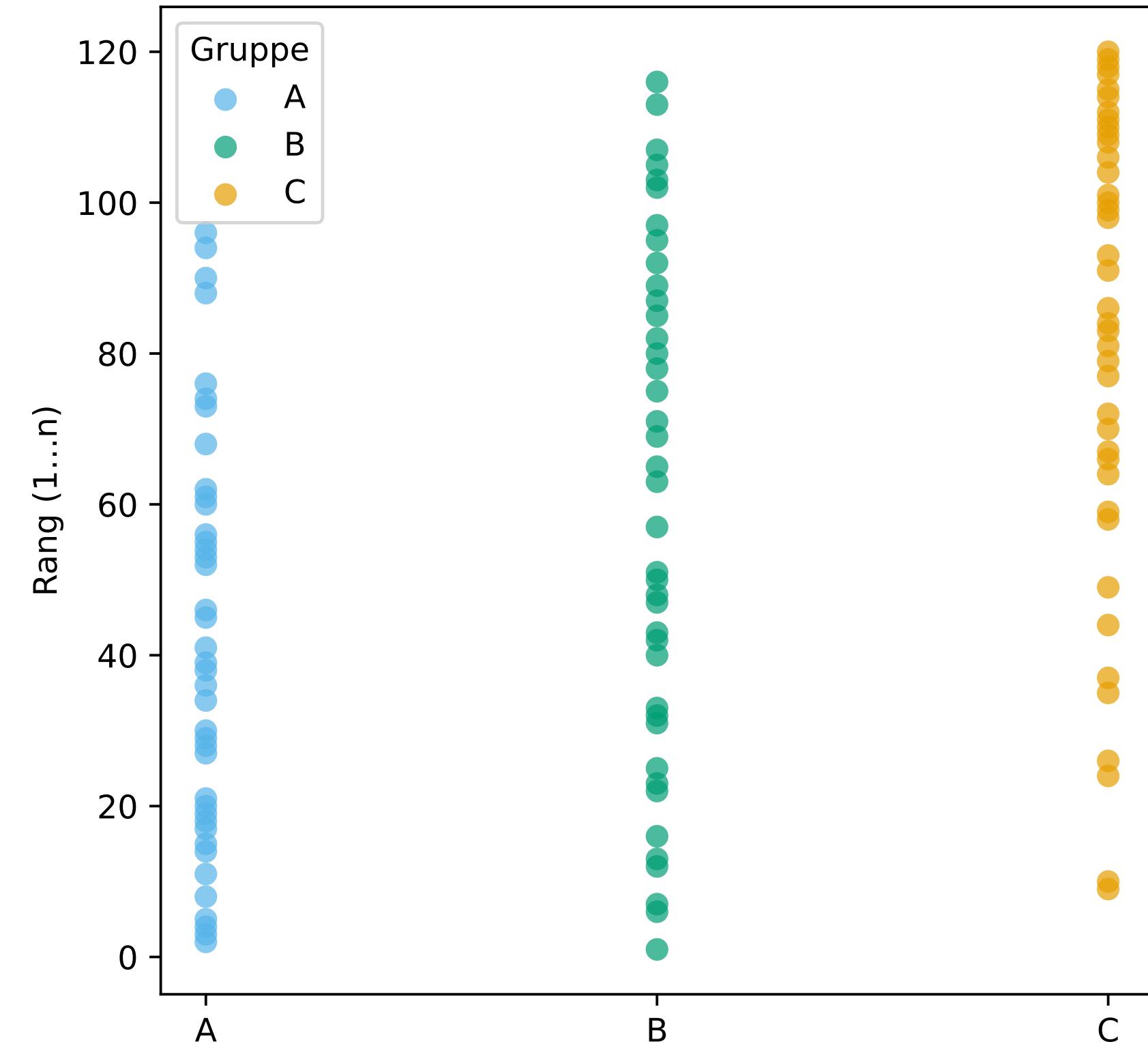
- ANOVA verlangt Normalität und vergleichbare Varianzen
- Realwelt-Daten oft: Schiefe, Heavy-Tails, Ausreisser
- Kruskal nutzt **Ränge** statt Rohwerte
- Funktioniert für 3+ Gruppen ohne Normalverteilung
- Testet Lage-/Medianunterschiede

Mini-Check: Warum machen Ränge den Test robust?

Rohwerte (schief, heavy-tail)



Ränge statt Rohwerte



Warum Kruskal?

Kruskal prüft Gruppenunterschiede ohne param. Annahmen und eignet sich für schiefe Daten.

- Schiefe Verteilungen → ANOVA wird unzuverlässig
- Kruskal prüft Lageunterschiede über Rangsummen
- Robust gegenüber Ausreisern und Varianzheterogenität
- Sinnvoll z. B. bei Fraud-, Income- oder Web-Traffic-Daten
- Ergebnis: globaler H -Test ähnlich zu ANOVA-Logik

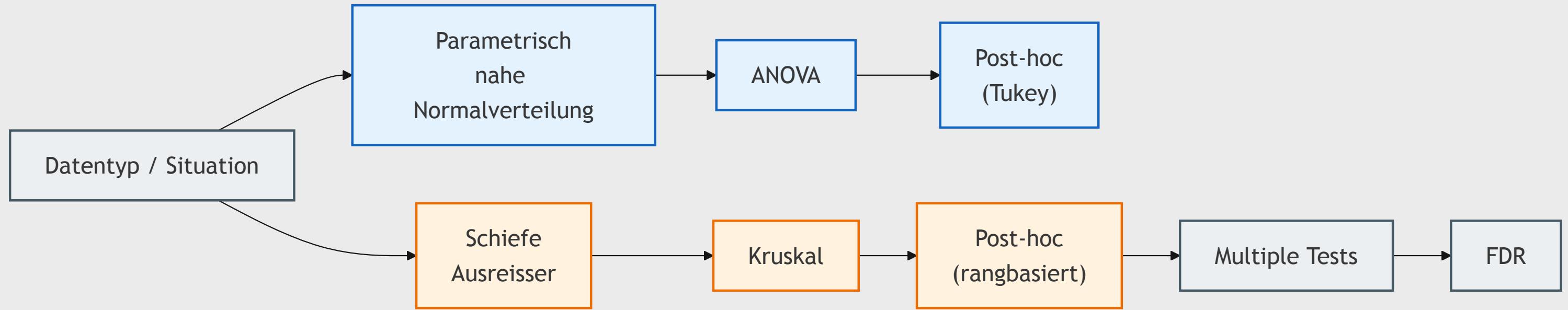
Mini-Check: Wann würdest du Kruskal statt ANOVA wählen?



StatCoach

Kruskal-Wallis

Multiple Tests



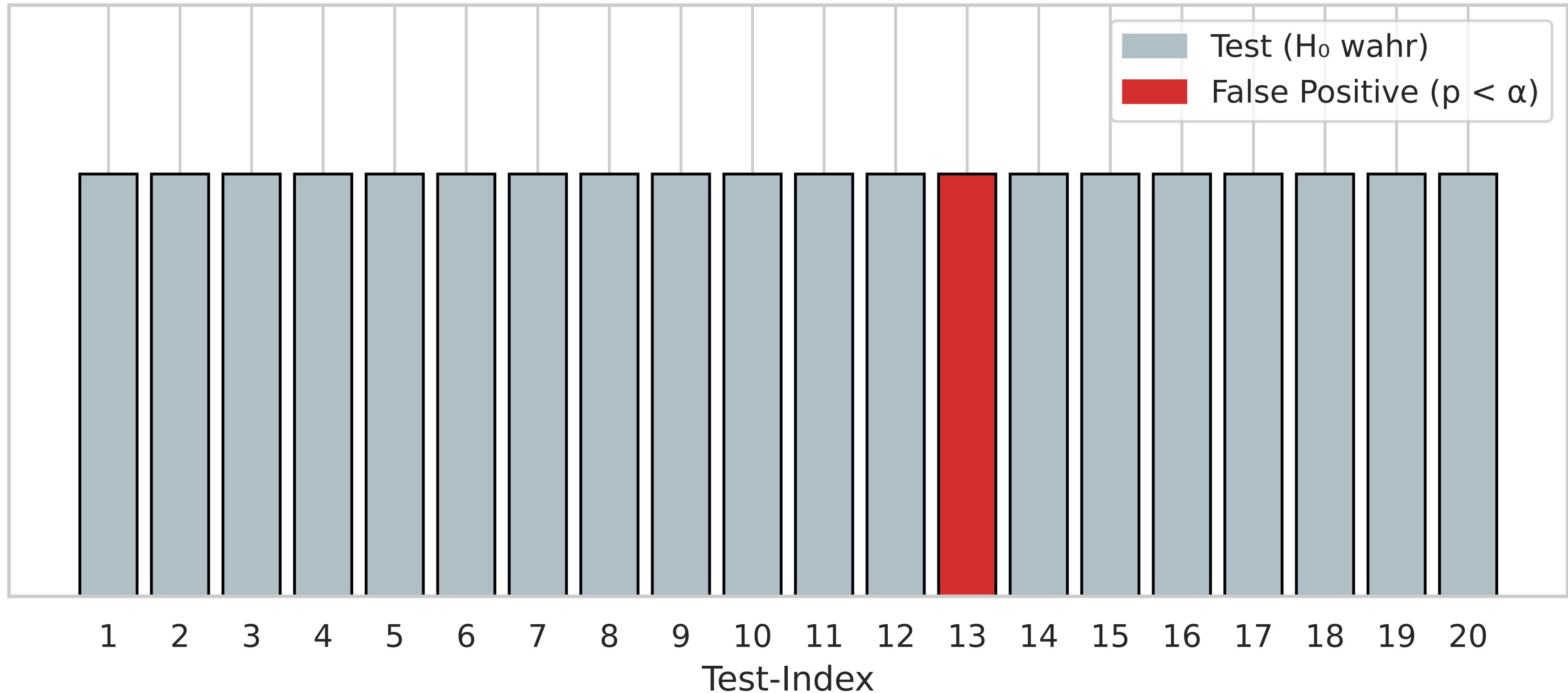
Motivation: Multiple Tests

Viele parallele Tests erhöhen die Wahrscheinlichkeit für Zufallstreffer massiv.

- Jeder Test trägt eigenes Fehlerrisiko α
- Viele Tests \rightarrow Gesamtfehler steigt schnell an
- Beispiel: 20 Tests bei $\alpha = 0.05 \rightarrow \approx 64\%$ Fehlerchance
- 100 Tests \rightarrow Fehler fast garantiert
- Lösung: Verfahren zur Fehlerkontrolle

Mini-Check: Warum steigt die Fehlerwahrscheinlichkeit bei vielen Tests so stark?

20 parallele Tests (ein Balken als 'False Positive' markiert)



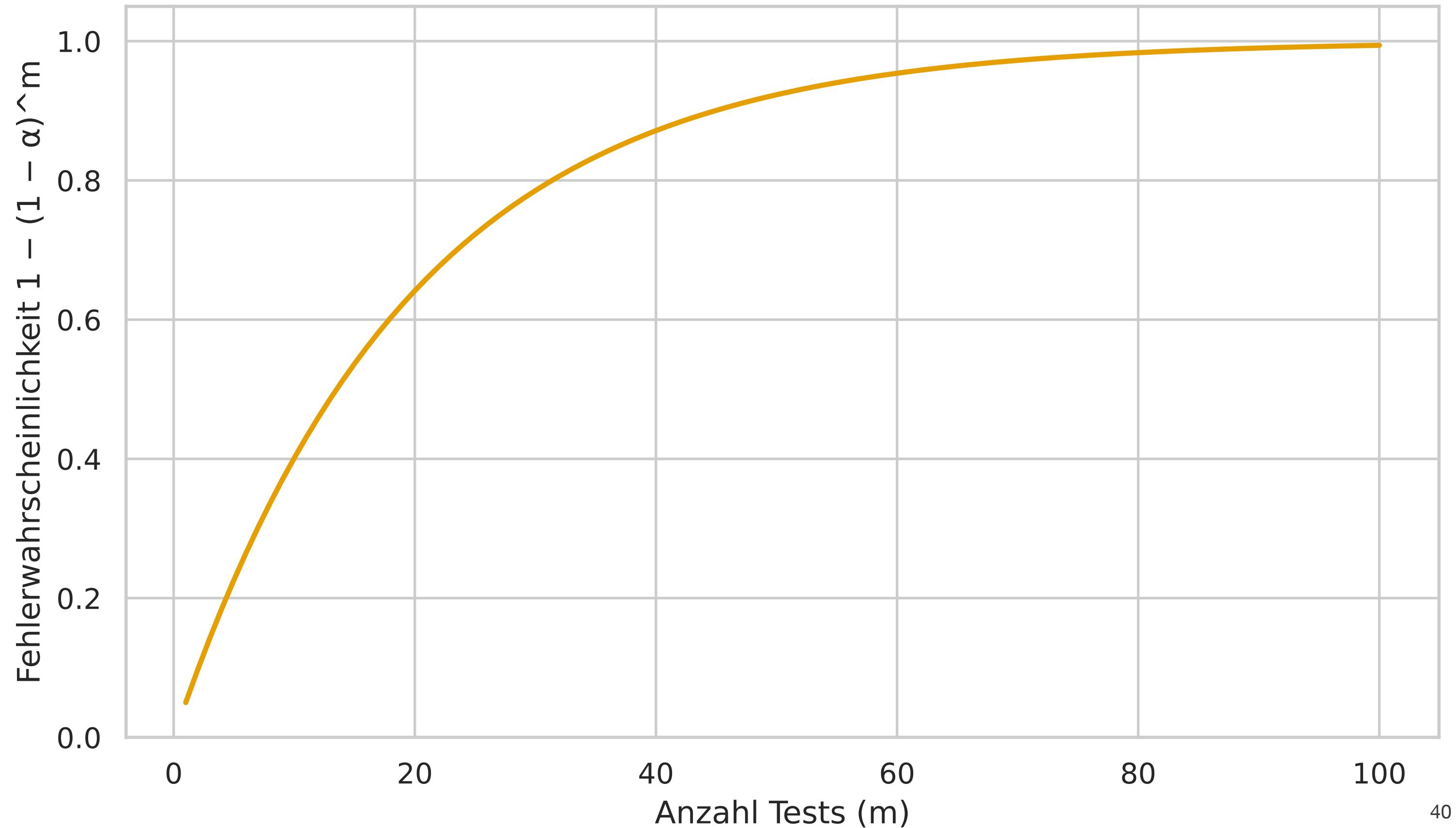
Das Problem vieler Tests

Ohne Korrektur entstehen bei vielen Kategorien massenhaft Zufallstreffer.

- Formel für Gesamtfehler: $1 - (1 - \alpha)^m$
- Unkorrigiert: viele scheinbare Treffer
- Viele Effekte verschwinden bei echter Fehlerkontrolle
- Notwendig: Bonferroni oder FDR

Mini-Check: Warum ist ein einzelner kleiner p -Wert bei vielen Tests nicht aussagekräftig?

Akkumulation der Fehlerwahrscheinlichkeit bei multiplen Tests



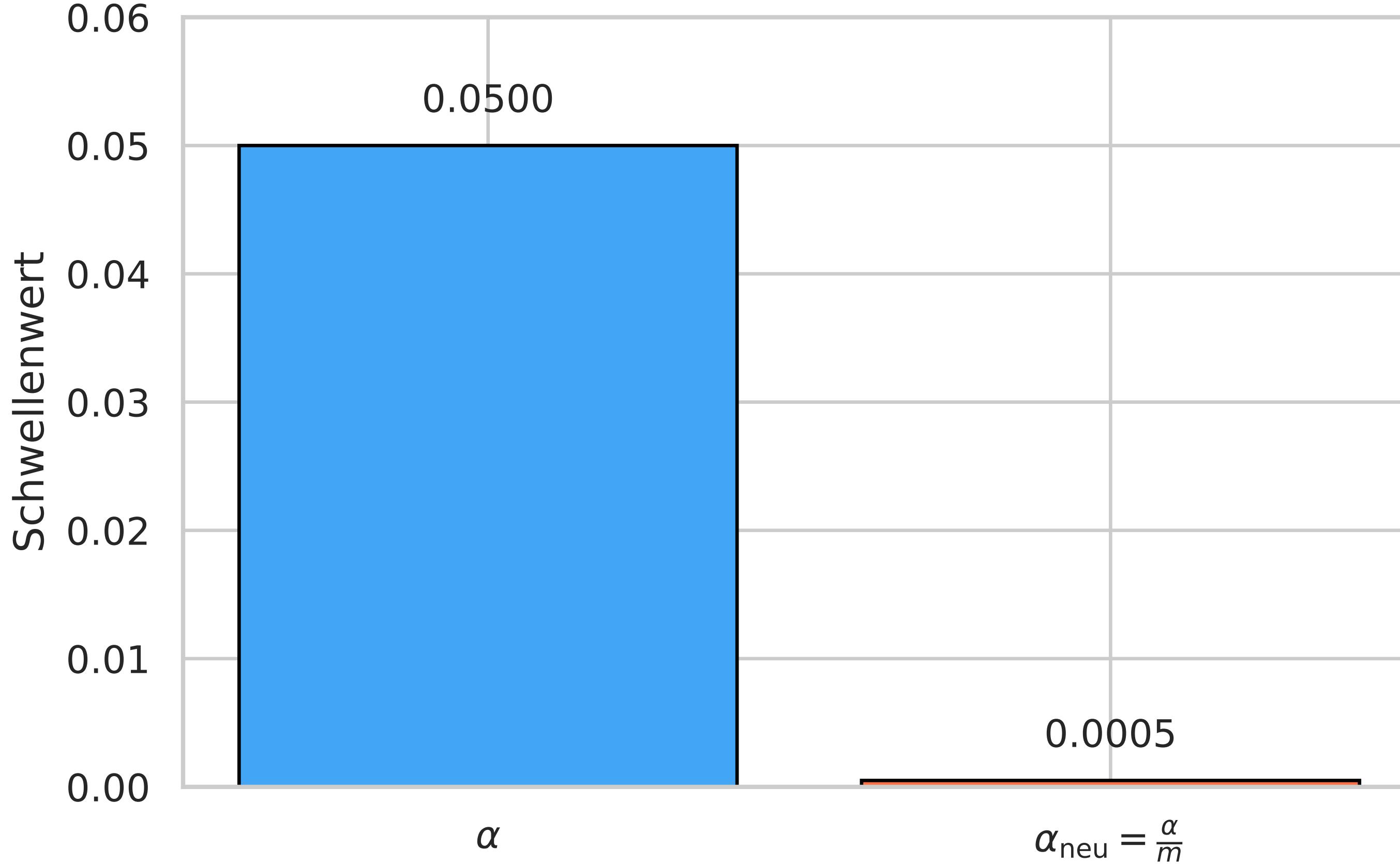
Bonferroni (zu streng)

Bonferroni schützt zuverlässig, wird aber bei vielen Tests extrem konservativ.

- Idee: Gesamtfehler unter Kontrolle halten
- Neue Schwelle: $\alpha_{\text{neu}} = \frac{\alpha}{m}$
- Beispiel: 100 Tests $\rightarrow \alpha_{\text{neu}} = 0.0005$
- Vorteil: Sehr stark gegen Fehlalarme
- Nachteil: Kaum noch Power, viele echte Effekte verschwinden

Mini-Check: Warum verliert Bonferroni bei vielen Tests echte Effekte?

Bonferroni-Korrektur bei $m = 100$ Tests



Viele parallele Tests machen zufällige «Signifikanzen» zur Regel
nicht zur Ausnahme.
Bonferroni-Korrektur möglich, aber zu streng

FDR (False Discovery Rate)

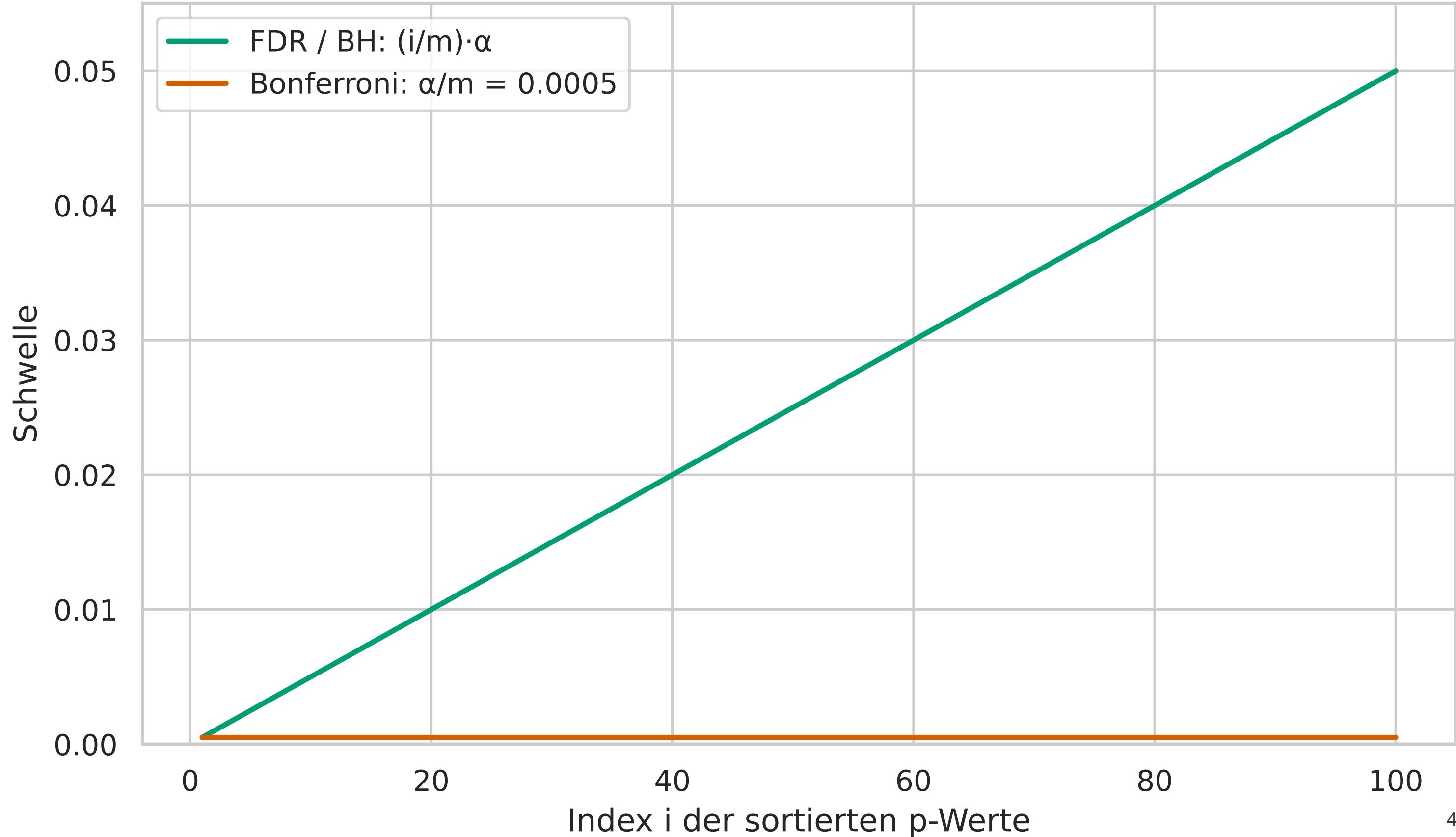
Motivation: False Discovery Rate (FDR)

FDR erkennt echte Effekte bei vielen Tests, ohne zu streng zu sein.

- Bonferroni schützt sehr stark → verliert aber Power
- FDR kontrolliert **Anteil** falscher Entdeckungen
- Ideal für grosse Hypothesensets (10+, 50+, 100+)
- Standard in Bio, Medizin und Data Science
- Basis: Benjamini–Hochberg-Verfahren

Mini-Check: Warum ist FDR oft besser als Bonferroni?

Vergleich: Bonferroni vs. FDR (Benjamini-Hochberg)



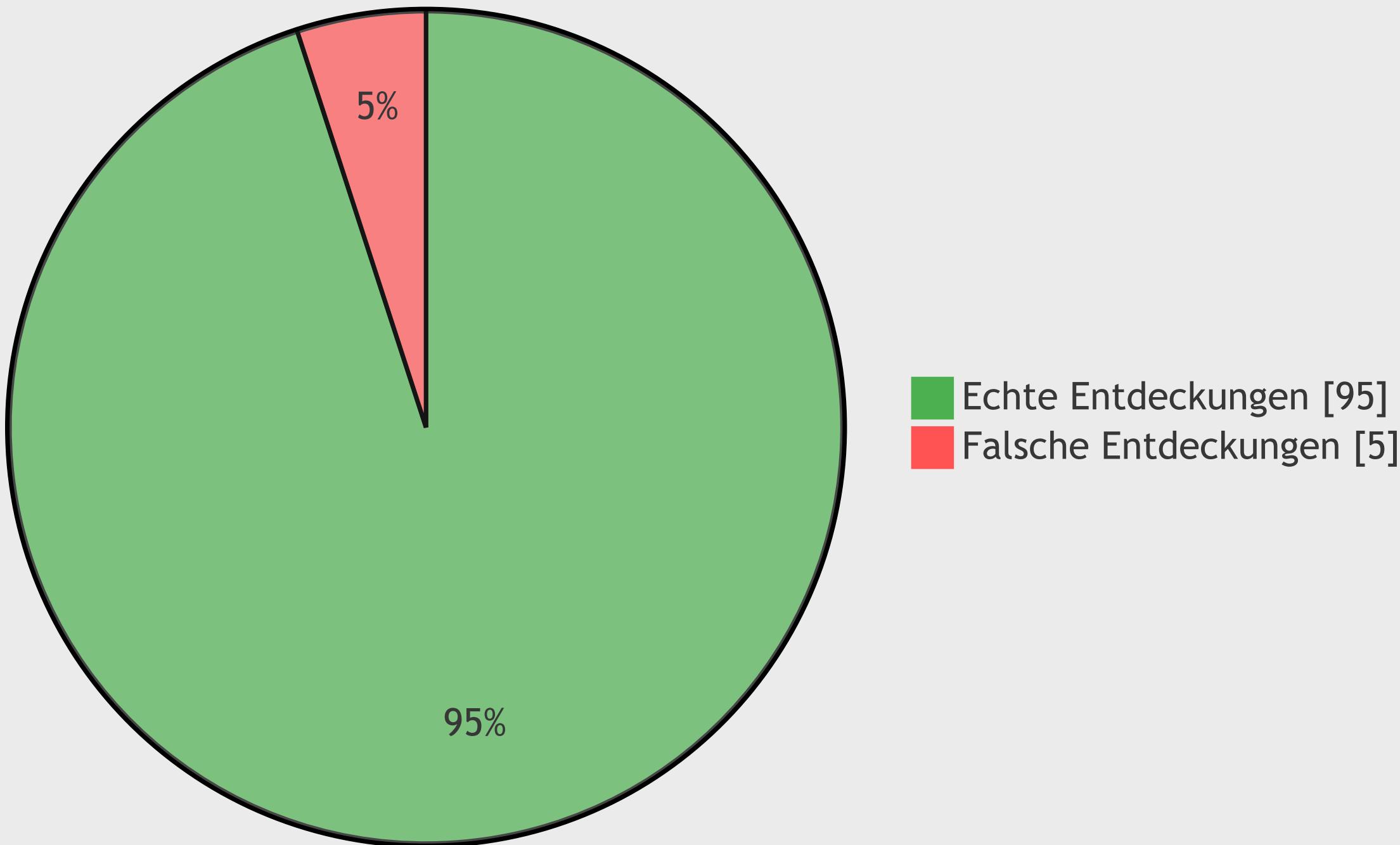
Was ist FDR?

FDR kontrolliert den erwarteten Anteil falscher Entdeckungen unter allen signifikanten Ergebnissen.

- Fokus nicht auf Fehler **pro Test**, sondern Fehler **unter Entdeckungen**
- Einige Fehler sind bei vielen Tests unvermeidlich
- Wichtig: Anteil der falschen Treffer klein halten
- Praktisch: deutlich höhere Power als Bonferroni
- Grundlage: sortierte p -Werte + adaptive Schwelle

Mini-Check: Warum ist der Anteil falscher Treffer informativer als die absolute Zahl?

falscher Entdeckungen bei FDR (alpha = 0.05)



Benjamini–Hochberg: Schritt-für-Schritt

BH berechnet eine adaptive Schwelle, die mit i ansteigt.

- Schritt 1: Sortiere p -Werte
 $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
- Schritt 2: Berechne für jedes i :
$$T_i = \frac{i}{m} \alpha$$
- Schritt 3: Finde das grösste i , für das
 $p_{(i)} \leq T_i$
- Schritt 4: Alle $p_{(1)}, \dots, p_{(i)}$ sind signifikant

Mini-Check: Warum steigt T_i mit i ?

Benjamini-Hochberg: sortierte p-Werte und Schwellen ($\alpha = 0.05$, $m = 10$)

i	p_(i)	T_i	Signifikant?
1	0.006	0.005	
2	0.016	0.010	
3	0.016	0.015	
4	0.037	0.020	
5	0.060	0.025	
6	0.060	0.030	
7	0.071	0.035	
8	0.073	0.040	
9	0.087	0.045	
10	0.095	0.050	

```
from statsmodels.stats.multitest import multipletests

pvals = [0.001, 0.02, 0.03, 0.20, 0.04, 0.07, 0.15]

reject, pvals_corrected, _, _ = multipletests(
    pvals,
    alpha=0.05,
    method='fdr_bh'
)

print("Signifikant?:", reject)
print("FDR-korrigierte p-Werte:", pvals_corrected)
```

```
Signifikant?: [ True False False False False False False]
FDR-korrigierte p-Werte: [0.007 0.07  0.07  0.2   0.07  0.098 0.175]
```

StatCoach: FDR-korrigierte p-Werte

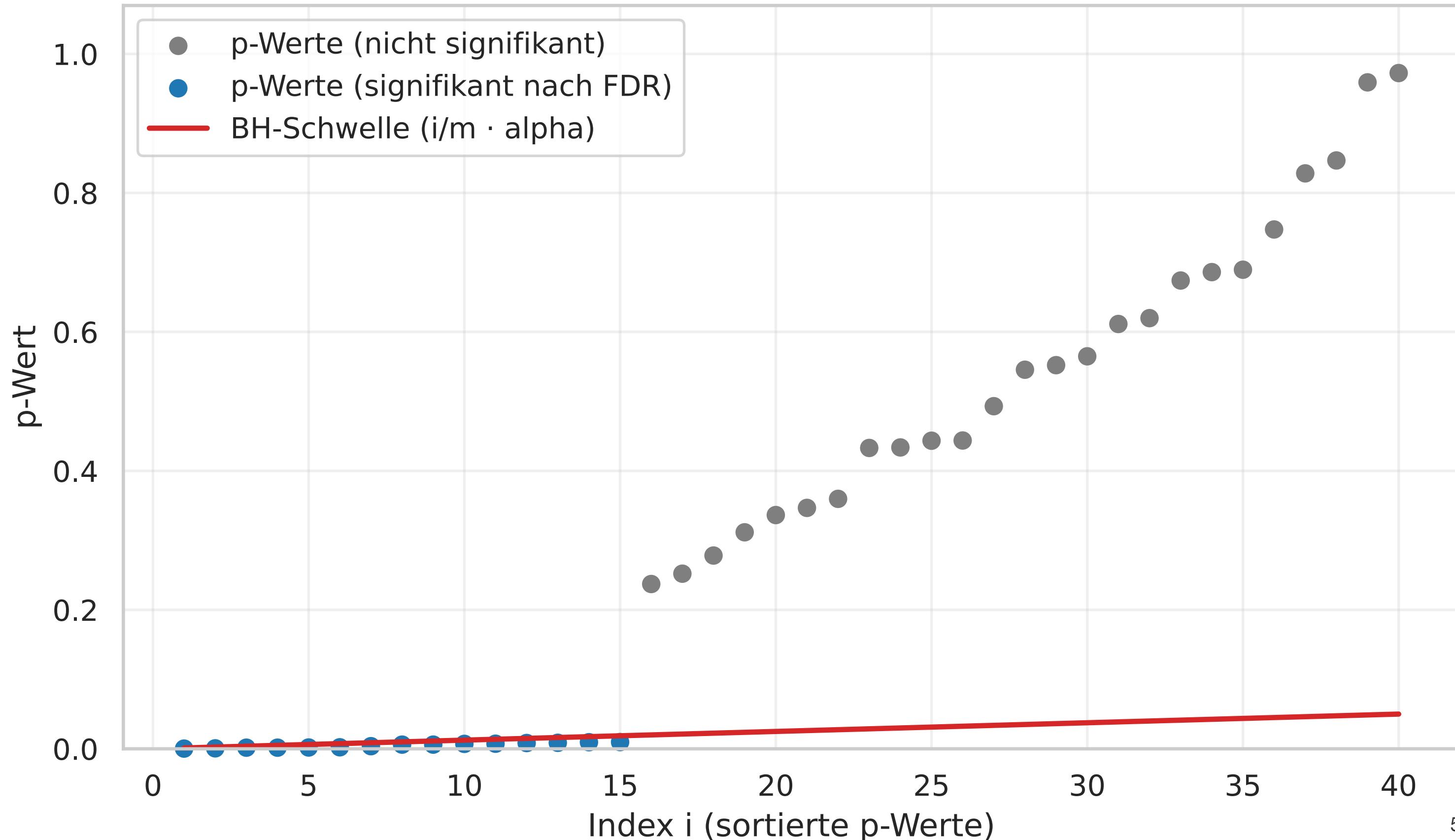
Visualisierung: FDR-Linie

Die BH-Linie ist eine steigende Grenze, die sortierte p -Werte schneidet.

- Sortierte p -Werte als Punkte
- Steigende BH-Linie $\frac{i}{m} \alpha$
- Punkte unter der Linie → signifikant
- Punkte über der Linie → nicht signifikant
- Intuitive Methode zum Verständnis von FDR

Mini-Check: Warum erkennt FDR mehr echte Effekte als Bonferroni?

Benjamini-Hochberg: Sortierte p-Werte und FDR-Schwelle



Was sind «Tests»?

Tests sind **einzelne Hypothesenprüfungen**, nicht Gruppen.

- Ein Test = ein p-Wert
- Mehrere Gruppen → mehrere Tests
- FDR arbeitet **immer** auf der Anzahl der Tests, nicht auf der Anzahl der Gruppen

Mini-Check: Wenn du 6 Gruppen hast, wie viele Tests entstehen bei allen Paarvergleichen?

Beispiele:

- ANOVA → 1 Test
- Tukey Post-hoc mit 5 Gruppen → 10 Tests
- 107 Kategorien (z. B. Chicago 311) → 107 Tests
- Feature-Screening: jedes Feature = 1 Test

Take-Away: FDR

FDR ist die praktikable Fehlerkontrolle für mehrere parallele Tests.

- Kontrolliert den Anteil falscher Entdeckungen
- Funktioniert schon bei wenigen Tests (z. B. 5–10)
- Wird besonders wertvoll, wenn die Testzahl steigt (20+, 50+, 100+)
- Weniger streng als Bonferroni, deshalb höhere Power
- Robust, intuitiv und Standard in moderner Data Science

Mini-Check: Bei welchen Projekten würdest du FDR besonders dringend einsetzen?

Zusammenfassung

Zusammenfassung heute

ANOVA, Kruskal und FDR bilden das Fundament für Mehrgruppenanalysen.

- ANOVA: prüft globale Mittelwertunterschiede
- Post-hoc (Tukey): zeigt **welche** Gruppen sich unterscheiden
- Kruskal: robuste Alternative bei Schiefe oder Ausreissern
- Multiple Tests: viele Vergleiche → Fehlerinflation
- FDR: kontrolliert Anteil falscher Entdeckungen bei vielen Tests

Mini-Check: Welche Methode würdest du bei schiefen Verteilungen bevorzugen?

Quiz: Aktive Wiederholung

Kahoot Quiz VL9

- Fokus: ANOVA, Post-hoc, Multiple Tests, FDR
- Ziel: eigene Wissenslücken erkennen

Ausblick: Regression I & II

Regression verallgemeinert Gruppenvergleiche zu vollständigen Erklärungsmodellen.

- Nächste Woche: Regression als lineares Modell
- Schwerpunkt: Effekte, Trends, Confounding
- ANOVA als Spezialfall der Regression
- Anwendung: WHO, IMF, Refugees, eigene Projekte
- Transfer: von Gruppenunterschieden zu multiplen Einflussfaktoren

Mini-Check: Warum ist Regression die natürliche Fortsetzung nach ANOVA?