Block 2: Topic 5

Multiple Testing & p-Hacking

January 25, 2026

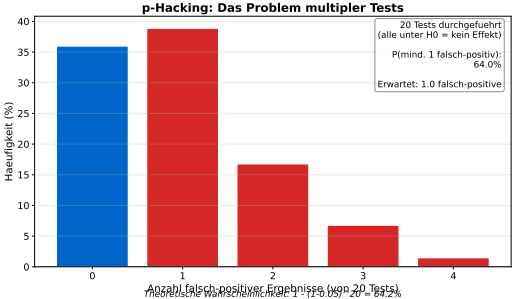
p-Hacking: Das Problem

Definition
Wiederholtes Testen und selektives Berichten bis $p < 0.05$ erreicht ist.

Typische Praktiken

- Viele Tests durchführen, nur signifikante berichten
- Ausreißer selektiv entfernen
- Variablen hinzufügen/entfernen bis es "passt"
- Stichprobe vergrößern bis $p < 0.05$

Resultat: Inflationsrate an falsch-positiven Ergebnissen!



p-Hacking zerstört die statistische Validität und führt zu Fehlentscheidungen.

Das Multiple Testing Problem

Grundproblem
Je mehr Tests wir durchführen, desto wahrscheinlicher finden wir "signifikante" Ergebnisse rein durch Zufall.

Beispiel: 20 Tests unter H_0
Bei $\alpha = 0.05$ erwarten wir:

$$0.05 \times 20 = 1 \text{ falsch-positives Ergebnis}$$

Das Problem verschärft sich

- Bei 100 Tests: erwarte 5 Zufallsbefunde
- Bei 1000 Tests: erwarte 50 Zufallsbefunde
- In Genomik/Neuroimaging: Millionen von Tests!

Lösung: Anpassung des Signifikanzniveaus für multiple Tests.

Ohne Korrektur führt multiples Testen zu massiven Fehlentscheidungen.

Bonferroni-Korrektur

Idee
Adjustiere das Signifikanzniveau um die Anzahl der Tests.

Formel

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

wobei m = Anzahl der Tests.

Beispiel: 20 Tests

$$\alpha_{\text{adj}} = \frac{0.05}{20} = 0.0025$$

Nur p-Werte < 0.0025 gelten als signifikant.

Nachteil

- Sehr konservativ: wenig Power, hohe Rate an falsch-negativen
- Bei vielen Tests wird es fast unmöglich, Signifikanz zu erreichen

Bonferroni kontrolliert familienweise Fehlerrate, aber um den Preis niedriger Power.

False Discovery Rate (FDR)

Moderne Alternative
Kontrolliert den erwarteten Anteil falscher Entdeckungen unter den als signifikant deklarierten Ergebnissen.

Benjamini-Hochberg Methode
In R:

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

Vorteile

- Weniger konservativ als Bonferroni
- Höhere Power bei vielen Tests
- Standard in Data Science, Genomik, Machine Learning
- Balanciert Fehler 1. Art und Power besser

Interpretation
Bei FDR = 0.05: maximal 5% der entdeckten Effekte sind falsch-positiv.

FDR ist der moderne Standard für multiple Testing in der Praxis.