STA 35C: Statistical Data Science III

Lecture 18: Multiple Hypotheses Testing

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 2 on Fri, May 16 (12:10 pm–1:00 pm in class)

- **Arrive early**: The exam starts at 12:10 pm and ends at 1:00 pm sharp
- **One hand-written cheat sheet:** Letter-size (8.5"×11"), double-sided, brief formulas/notes
- **Calculator**: A simple (non-graphing) scientific calculator is allowed
- **No other materials** beyond the single cheat sheet (no textbooks, etc.)
- **SDC accommodations:** Confirm scheduling with AES online ASAP

Preparation tips:

- Primary coverage: Lectures 12–19 (including Wed)
- A practice midterm and answer key are available on the course webpage
- Office hours this week:
  - Instructor: Wed, 4–6pm (extended); no OH Thu
  - TA: Mon/Thu 1–2pm

Today's topics

- **Multiple hypotheses testing**
  - Recap: Motivation & challenges
    - Issues arising with multiple tests
    - Real-world concerns: p-hacking & data dredging
  - **Family-wise error rate (FWER)**
    - Definition & intuition
    - Controlling FWER: Bonferroni correction & Holm's step-down
  - **False discovery rate (FDR)**
    - Definition & intuition
    - Controlling FDR: Benjamini-Hochberg procedure

Recap: Multiple testing

Single-hypothesis test:

- Typically set up H_0, and gather data to reject it if there is significant evidence
- Type I error $=$ false positive; Type II error $=$ false negative
- Each test has Type I error at most α (e.g. 0.05)

Modern data analysis: multiple tests simultaneously

- E.g. Testing thousands of predictors or biomarkers to discovery significant ones
- If m is large, false rejections can occur easily by chance
- On average, $\alpha \times m$ false positives if each test is at level α

Key challenge: Address the inflation of false positives as m grows

Related issues: p-hacking and data dredging

Real danger: Searching for “significant” results in many ways until something "works"

- Repeatedly testing different hypotheses/subgroups
- Eventually, some test may yield $p < 0.05$ *by chance*

Outcome: Spurious “discoveries”

- Published claims may fail to replicate
- True findings can be overshadowed by noise

Conclusion: Systematic multiple-testing corrections are crucial, especially for large m

Articles warning about misused statistical significance

PLOS Medicine   Browse   Publish   About

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

The Economist   Menu

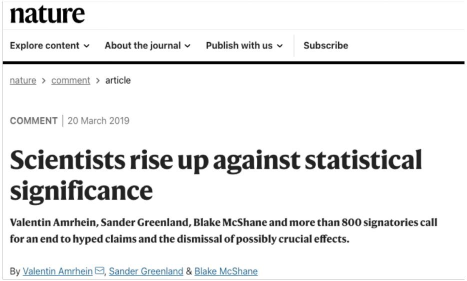
Weekly edition | The world in brief | War in the Middle East | War in Ukraine | United States | The world economy | Busines

Briefing | Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 18th 2013



nature

Explore content | About the journal | Publish with us | Subscribe

nature > comment > article

COMMENT | 20 March 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

By Valentin Amrhein, Sander Greenland & Blake McShane

Figure: Many reproducibility crises trace back to undisclosed multiple testing or selective reporting. Proper adjustments can help mitigate these issues.

Family-wise error rate (FWER): Definition

Single test: Pr(Type I error) = Pr(reject a true null)

Multiple tests (*m* hypotheses):

$$\mathrm{FWER} = \Pr(\text{reject at least one true } H_0),$$

i.e. the probability of *any* false positive among *m* tests

If tests are independent, and each are at level α:

$$\mathrm{FWER} = 1 - (1 - \alpha)^m,$$

which grows quickly with *m*

- E.g. $m = 20$, $\alpha = 0.05 \implies \mathrm{FWER} \approx 0.64 \gg 0.05$

Visualization: FWER grows as m increases

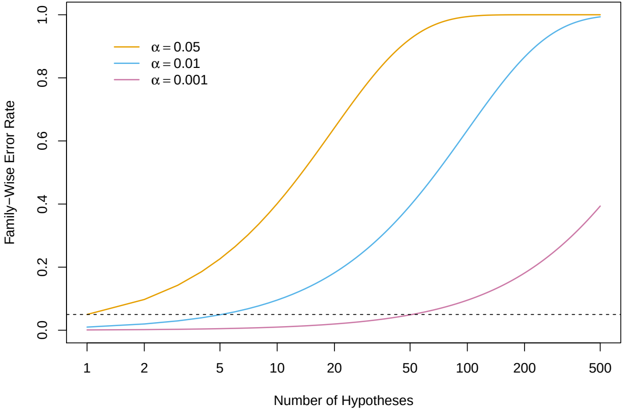


Figure: FWER vs. number of tests m (log scale) for $\alpha = 0.05$ (orange), 0.01 (blue), 0.001 (purple). The dashed line is 0.05. For $m = 50$ and target FWER=0.05, each test must be at $\alpha = 0.001$ [JWHT21, Figure 13.2].

The Bonferroni correction

Key idea:

- Observe that

$$\mathrm{FWER} = \Pr\left(\sum_{j=1}^{m}\{\text{Reject } H_j\}\right) \leq \sum_{j=1}^{m}\Pr\left(\{\text{Reject } H_j\}\right)$$

- Each test is done at level $\alpha/m \implies \Pr\left(\{\text{Reject } H_j\}\right) \leq \alpha/m \implies \mathrm{FWER} \leq \alpha$

The Bonferroni method (Bonferroni correction):

- For each hypothesis $H_1, \ldots, H_m$, reject H_j if only if $p_j < \frac{\alpha}{m}$

Pros & Cons:

- **Pros**: Simple & widely used
- **Cons**: Often *very conservative* $\implies$ few rejections (=discoveries) & lower power[1]

[1] Power = TPR = the fraction of false null hypotheses that are successfully rejected

Example: Bonferroni correction

Example

Let $m = 6$ hypotheses with p-values:

$$p_1 = 0.0018, \quad p_2 = 0.009, \quad p_3 = 0.021, \quad p_4 = 0.034, \quad p_5 = 0.045, \quad p_6 = 0.070.$$

At $\alpha = 0.05$, threshold $= \frac{\alpha}{m} = \frac{0.05}{6} \approx 0.00833$.

$$\text{Reject } H_j \text{ if } p_j < 0.00833.$$

Hence:

$$p_1 = 0.0018 < 0.00833 \implies \text{reject } H_1,$$

but $p_2 = 0.009 > 0.00833$ and the rest are larger. So Bonferroni rejects only H_1.

Conclusion: 1 rejection using Bonferroni, whereas naive $p < 0.05$ would reject 5 of them $(p_1, \ldots, p_5)$.

Holm's step-down procedure

Holm's method refines Bonferroni to be less conservative:

Holm's method

1. Specify α, the level at which to control the FWER
2. Compute the p-values for the m null hypotheses, $H_{01}, \dots, H_{0m}$
3. Sort p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
4. Define
$$L = \min \left\{ j : p_{(j)} > \frac{\alpha}{m+1-j} \right\}$$
5. Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$

Properties:

- Ensures $\mathrm{FWER} \leq \alpha$
- Rejects at least as many hypotheses as Bonferroni

Example: Holm's step-down procedure

Example

Step 1: Set $\alpha = 0.05$

Step 2: $p_1 = 0.0018$, $p_2 = 0.009$, $p_3 = 0.021$, $p_4 = 0.034$, $p_5 = 0.045$, $p_6 = 0.070$.

Step 3: Sort p-values $p_{(1)} = 0.0018$, $p_{(2)} = 0.009$, $p_{(3)} = 0.021$, $p_{(4)} = 0.034$, $p_{(5)} = 0.045$, $p_{(6)} = 0.070$.

Step 4: Find $L = 3$ because

$$p_{(1)} = 0.0018 \text{ ? } 0.0018 \leq \frac{0.05}{6+1-1} = \frac{0.05}{6} \approx 0.00833 \qquad \Longrightarrow \text{ reject } H_{(1)}\text{, continue}$$

$$p_{(2)} = 0.009 \text{ ? } 0.009 \leq \frac{0.05}{6+1-2} = \frac{0.05}{5} = 0.01 \qquad \Longrightarrow \text{ reject } H_{(2)}\text{, continue}$$

$$p_{(3)} = 0.021 \text{ ? } 0.021 \leq \frac{0.05}{6+1-3} = \frac{0.05}{4} = 0.0125\text{? No} \qquad \Longrightarrow \text{ stop; } L = 3$$

Step 5: We reject $H_{(1)}, H_{(2)}$ total 2 rejections. The rest are not rejected.

Conclusion: Holm's method rejects 2, whereas Bonferroni rejected only 1.

Visualization: Bonferroni vs. Holm

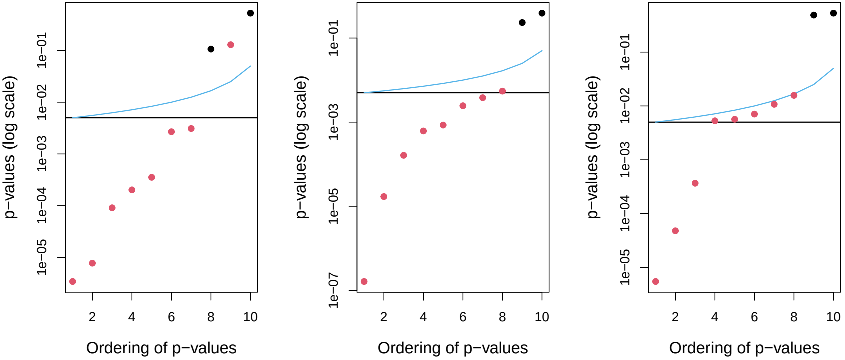


Figure: Each panel shows sorted p-values from a separate simulation of $m = 10$ null hypotheses, with the two true nulls in black and the others in red. Controlling the FWER at 0.05, Bonferroni rejects all points below the **black** line, while Holm rejects all below the **blue** line. The gap between these lines indicates the additional hypotheses Holm rejects but Bonferroni does not. In the middle panel, Holm rejects one more null than Bonferroni; in the right panel, it rejects five more [JWHT21, Figure 13.3].

FWER vs. power trade-off

FWER demands *no* false rejections with probability at least $1 - \alpha$:

- Very stringent if m is large
- Tends to reduce power (fewer true positives found)

In modern "exploratory" studies:

- We may tolerate a small fraction of false positives to discover more true ones
- This leads to the *false discovery rate (FDR)* approach

Illustration: power vs. FWER

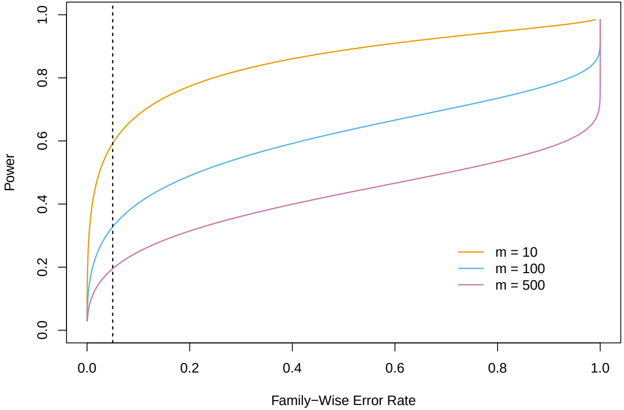


Figure: In a simulation with 90% of m nulls true, the power is displayed against FWER. Colors of the curves: $m = 10$ (orange), $m = 100$ (blue), $m = 500$ (purple). Larger m reduces power. The vertical dashed line marks FWER=0.05 [JWHT21, Figure 13.5].

False discovery rate (FDR): Definition and motivation

Motivation: Controlling FWER can be too conservative for large m

Instead: control the fraction of rejected hypotheses that are *false positives*

$$\text{FDP} = \frac{\#\text{ false positives}}{\#\text{ total rejections}} \quad (= 0 \text{ if no rejections}).$$

False discovery rate (FDR) $= \mathbb{E}[\text{FDP}]$

- Among the “discoveries," we want to ensure a fraction up to q may be false positives *on average*

Properties:

- Accept a small fraction of false positives, in exchange for more total discoveries
- Typically yields more rejections ("discoveries") than FWER-based methods

Controlling FDR: Benjamini–Hochberg procedure

Benamini-Hockberg procedure

1 Specify q, the level at which to control the FDR

2 Compute the p-values for the m null hypotheses, $H_{01}, \ldots, H_{0m}$

3 Sort p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$

4 Define

$$L = \max \left\{ j : p_{(j)} < \frac{qj}{m} \right\}$$

5 Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$

Result:

- Ensures FDR $\leq q$, but but not necessarily small FWER
- Typically more powerful, yielding more rejections, than Bonferroni/Holm if m is large

Example: Benjamini-Hochberg procedure

Example

Step 1: Set $q = 0.05$

Step 2: $p_1 = 0.0018, p_2 = 0.009, p_3 = 0.021, p_4 = 0.034, p_5 = 0.045, p_6 = 0.070.$

Step 3: Sort p-values $p_{(1)} = 0.0018, p_{(2)} = 0.009, p_{(3)} = 0.021, p_{(4)} = 0.034, p_{(5)} = 0.045, p_{(6)} = 0.070.$

Step 4: Find $L = 3$ because

$$
\begin{aligned}
k = 1: &\quad 0.0018 \leq 0.05 \times \frac{1}{6} \approx 0.0083?\ \checkmark \\
k = 2: &\quad 0.009 \leq 0.05 \times \frac{2}{6} \approx 0.0167?\ \checkmark \\
k = 3: &\quad 0.021 \leq 0.05 \times \frac{3}{6} = 0.025?\ \checkmark \\
k = 4: &\quad 0.034 \leq 0.05 \times \frac{4}{6} \approx 0.0333?\ \text{No}\ (\,0.034 > 0.0333\,)
\end{aligned}
$$

Step 5: Reject $H_{(1)}, H_{(2)}, H_{(3)}$.

Conclusion: BH rejects 3, while Holm rejects 2, Bonferroni rejects 1.

Visual comparison: Bonferroni vs. Benjamini-Hochberg

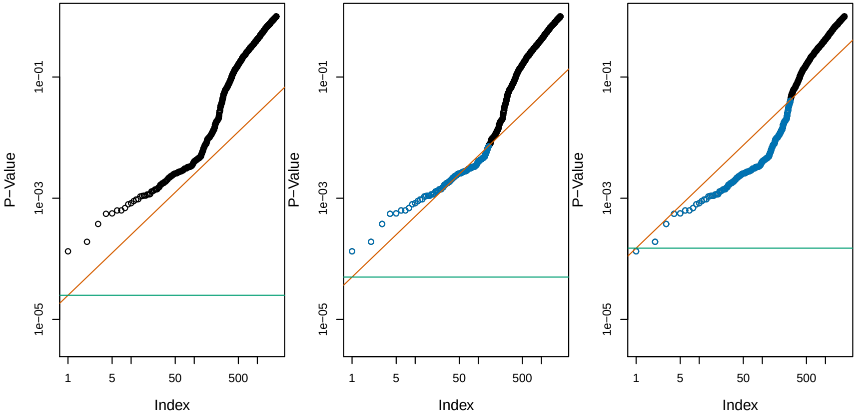


Figure: Panels: same set of $m = 2000$ sorted p-values for the Fund dataset. Green lines: thresholds for FWER control (Bonferroni) at $\alpha = 0.05$, 0.1, 0.3 (left to right). Orange lines: thresholds for FDR control (Benjamini-Hochberg) at $q = 0.05, 0.1, 0.3$ (left to right). E.g., When the FDR is controlled at $q = 0.1$, 146 nulls are rejected (center, blue points). At $q = 0.3$, 279 nulls are rejected (right, blue points) [JWHT21, Figure 13.6].

Wrap-up

- **FWER** (Bonferroni/Holm):
  - Strictly ensures Pr(any false positive) $\leq \alpha$
  - Conservative for large m, leading to fewer rejections & reduced power
- **FDR** (Benjamini–Hochberg):
  - Controls the expected fraction of false positives among rejections
  - Typically yields more rejections than FWER, especially for large m
- **Practical consideration:**
  - Use FWER for strict confirmatory analyses needing minimal Type I error
  - Use FDR for exploratory, large-scale studies, tolerating some false positives to gain more discoveries

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.