STA 35C: Statistical Data Science III

Lecture 18: Multiple Hypotheses Testing

Dogyoon Song

Fall 2026, UC Davis

Announcements

Homework 5 is due tomorrow (Thu, Nov 5, 11:59 PM PT)

Extended office hours: Wednesdays, 1–3 PM (MSB 1143)

- The 2–3 PM hour is shared with STA 130A, whose students may receive priority

Agenda

Recap: Challenges & motivation for multiple hypotheses testing

- Issues arising with multiple tests
- Real-world concerns: p-hacking & data dredging

Today: How to control false positives

- **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 even when all nulls are true
- If all m nulls are true and each test is at level α, the expected number of false positives is at most $m \times \alpha$

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

Articles warning about misused statistical significance



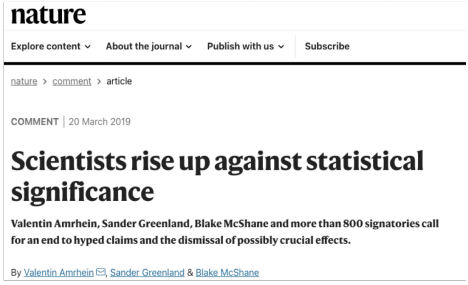




Figure: Many reproducibility crises trace back to undisclosed multiple testing or selective reporting. Proper adjustments can help mitigate these issues.

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*
- Example: trying many outcomes, subgroups, transformations, or model specifications, then reporting only the smallest p-value

Outcomes: Spurious "discoveries"

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

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

Recall: Hypothesis testing as classification

A single hypothesis test makes a decision about whether to reject H_0:

- **Goal:** Discover a "real phenomenon" (H_1) or conclude not enough evidence (H_0)
  - H_0 is true $\iff$ no real effect
  - H_0 is false $\iff$ there is a real effect (H_1)
  - We claim a "discovery" when H_0 is rejected
- **Analogy to classification:** Depending on the evidence in the data,
  - Reject $H_0 \iff$ classify $\hat{H} = 1$
  - Fail to reject $H_0 \iff$ classify $\hat{H} = 0$

	H_0 is true (**"H=0"**)	H_0 is not true (**"H=1"**)
Reject H_0 ($\widehat{\textbf{"H=1"}}$)	Type-I error (FP)	Correct (TP)
Not reject H_0 ($\widehat{\textbf{"H=0"}}$)	Correct (TN)	Type-II error (FN)

$\implies$ Pr(Type I error) = Pr(reject a true null) $\approx \frac{FP}{FP+TN}$ empirically among true nulls

- By setting a threshold α, we want to control Pr(Type I error) below α

Family-wise error rate (FWER): Definition

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

Multiple tests (m hypotheses): Familywise error rate (FWER) is defined as

$$\begin{aligned}\text{FWER} &= \Pr(\text{reject at least one true } H_0) \\ &= \Pr(\#\text{ Type-I error} \geq 1),\end{aligned}$$

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

If all m null hypotheses are true, the tests are independent, and each rejects with probability exactly α:

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

- When $m = 1$, $\text{FWER} = 1 - (1-\alpha)^m = 1 - (1-\alpha) = \alpha$
- Grows quickly with m
  - E.g. $m = 20$, $\alpha = 0.05 \implies \text{FWER} \approx 0.64 \gg 0.05$

Visualization: FWER grows as m increases

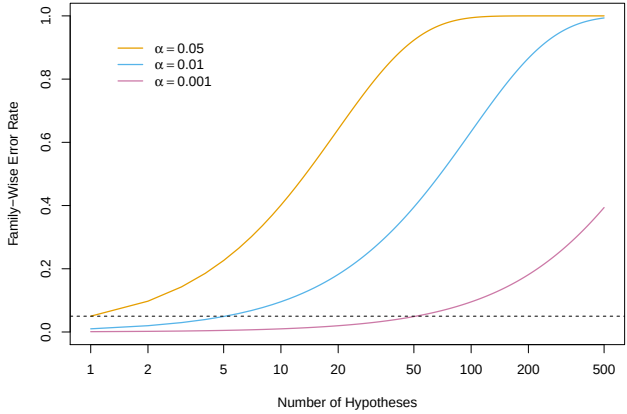


Figure: FWER vs. number of tests m (log scale) for $\alpha = 0.05$ (orange), 0.01 (blue), 0.001 (purple). The dashed line is 0.05. For $m = 50$, a per-test level of 0.001 suffices for $\mathrm{FWER} \leq 0.05$ [JWHT21, Figure 13.2].

Testing multiple hypotheses at level α

Suppose we test m hypotheses $H_{0,1}, \ldots, H_{0,m}$, all at level α, obtaining confusion matrix:

	H_0 is true	H_0 is not true
Reject H_0	N_{FP}	N_{TP}
Not reject H_0	N_{TN}	N_{FN}

- $N_{\mathsf{FP}}, N_{\mathsf{TP}}, N_{\mathsf{TN}}, N_{\mathsf{FN}}$ are random variables that sum to m
- Roughly, we expect $N_{\mathsf{FP}} \approx m \cdot p_{\mathsf{FP}}$; when all m nulls are true, $N_{\mathsf{FP}} \approx m \cdot \alpha$

If all m nulls are true and independent tests each reject with probability exactly α,

- Probability of *at least one* false positive $= 1 - (1 - \alpha)^m$
- For $m = 20, \alpha = 0.05$, that probability is $\approx 64\%$

Goal of FWER control: Keep the probability of making even one false positive small:

$$\mathrm{FWER} = \Pr(N_{\mathsf{FP}} \geq 1) \leq \alpha$$

The Bonferroni correction for FWER control

Key idea: Use the union bound over the true null hypotheses:

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

- 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 $p_j \leq \frac{\alpha}{m}$ (instead of $\leq \alpha$)

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

$$\text{Reject } H_j \quad \text{if } p_j \leq \frac{0.05}{6}.$$

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 for FWER control

Holm's method scans from smallest to largest and stops at the first failure.

Holm's method

1. Specify α, the level at which to control the FWER
2. Compute the p-values for the m null hypotheses, $H_{01}, \ldots, H_{0m}$
3. Sort p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
4. Define the first failing rank (set $L = m + 1$ if none fails):

$$L = \min \left\{ j : p_{(j)} > \frac{\alpha}{m + 1 - j} \right\}$$

5. Reject the first $L - 1$ ordered hypotheses: none if $L = 1$, all if $L = m + 1$

Properties:

- For valid p-values, ensures $\mathrm{FWER} \leq \alpha$ without requiring independence
- 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: Stop at the first failure, $L = 3$:

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

$$p_{(2)} = 0.009 \leq \frac{0.05}{5} = 0.01 \quad \Longrightarrow \text{ reject } H_{(2)}, \text{ continue}$$

$$p_{(3)} = 0.021 > \frac{0.05}{4} = 0.0125 \quad \Longrightarrow \text{ stop; } L = 3$$

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

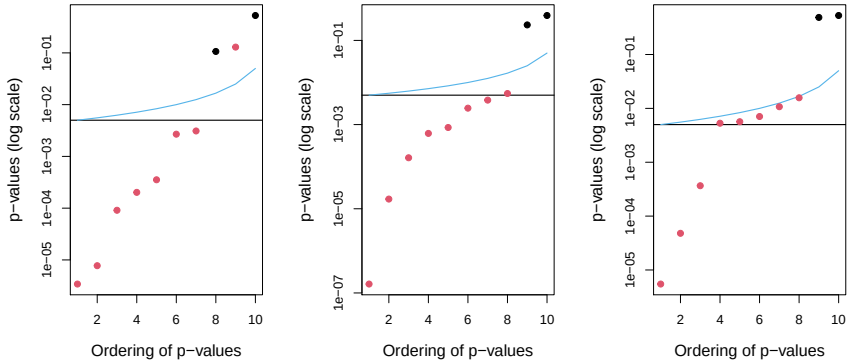


Figure: Each panel shows sorted p-values from a separate simulation of $m = 10$ null hypotheses, with the two true nulls in black and the others in red. Controlling the FWER at 0.05, Bonferroni rejects all points below the **black** line, while Holm scans from the smallest p-value and stops at the first point above the **blue** threshold, rejecting only the preceding points. In the middle panel, Holm rejects one more null than Bonferroni; in the right panel, it rejects five more [JWHT21, Figure 13.3].

Pop-up quiz: Bonferroni vs. Holm

Suppose $m = 4$, $\alpha = 0.05$, and the sorted p-values are

$$p_{(1)} = 0.010, \qquad p_{(2)} = 0.015, \qquad p_{(3)} = 0.030, \qquad p_{(4)} = 0.200.$$

Question: How many hypotheses are rejected by Bonferroni and by Holm's method?

A) Bonferroni rejects 1; Holm rejects 2.

B) Bonferroni rejects 2; Holm rejects 1.

C) Both reject 1.

D) Both reject 2.

Answer: A. Bonferroni uses $0.05/4 = 0.0125$, so it rejects only $p_{(1)}$. Holm compares $0.010 \leq 0.0125$, then $0.015 \leq 0.05/3$, then stops at $0.030 > 0.05/2$, so it rejects two.

Illustration: Power (TPR) vs. FWER trade-off

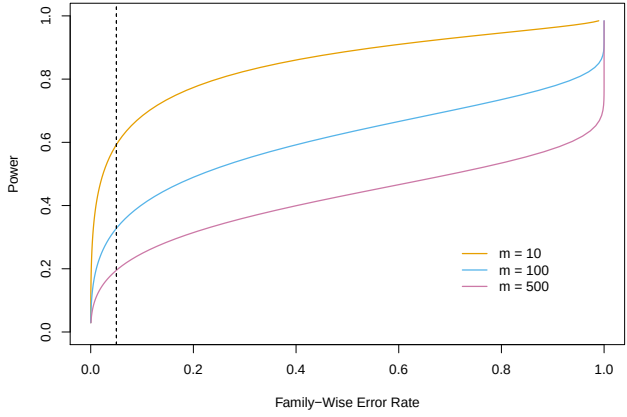


Figure: In a simulation with 90% of m nulls true, the power is displayed against FWER. Colors of the curves: $m = 10$ (orange), $m = 100$ (blue), $m = 500$ (purple). Larger m reduces power. The vertical dashed line marks FWER=0.05 [JWHT21, Figure 13.5].

FWER control may not suffice

FWER demands *no* false rejections (at all) 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

	H_0 is true	H_0 is not true
Reject H_0	Type-I error (FP)	Correct (TP)
Not reject H_0	Correct (TN)	Type-II error (FN)

New goal: Find more true discoveries, controlling the expected proportion of false discoveries.

False discovery rate (FDR): Definition and motivation

Motivation: Controlling FWER can be too conservative for large m

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

$$\text{FDP} = \frac{\#\text{ false positives}}{\#\text{ total rejections}} = \frac{\#\,\text{FP}}{\#\,\text{FP} + \#\,\text{TP}}$$

- Directly controlling FDP is impossible because we never know which H_{0j} are true/false
- Thus, we **control FDP in expectation**, not necessarily in every dataset

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

- Allow up to fraction q of false positives *on average* among the "claimed discoveries"
- The choice of q is context- and dataset-dependent (no gold standard like $\alpha = 0.05$)

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

Benjamini–Hochberg 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. Check **all ranks**, even after a failed comparison, and define

$$L = \max\left(\left\{j : p_{(j)} \leq \frac{qj}{m}\right\} \cup \{0\}\right)$$

5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$ (if $L = 0$, reject none)

Result:

- Under standard conditions, ensures $\mathrm{FDR} \leq q$, but not necessarily small FWER
- Reject every rank through L, including earlier ranks that failed their own comparisons

Example: Benjamini–Hochberg procedure

Example

Step 1: Set $q = 0.05$

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

Steps 3–4: Sort the p-values and check **all six ranks** to find $L = \max\left\{j : p_{(j)} < \frac{qj}{m}\right\} = 3$ because

j	$p_{(j)}$	qj/m (rounded)	$p_{(j)} \le qj/m$?
1	0.0018	0.00833	Yes
2	0.0090	0.01667	Yes
3	0.0210	0.02500	Yes
4	0.0340	0.03333	No: continue
5	0.0450	0.04167	No: continue
6	0.0700	0.05000	No: scan complete

Step 5: The largest passing rank is $L = 3$; reject $H_{(1)}, H_{(2)}, H_{(3)}$.

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

Visual comparison: Bonferroni vs. Benjamini–Hochberg

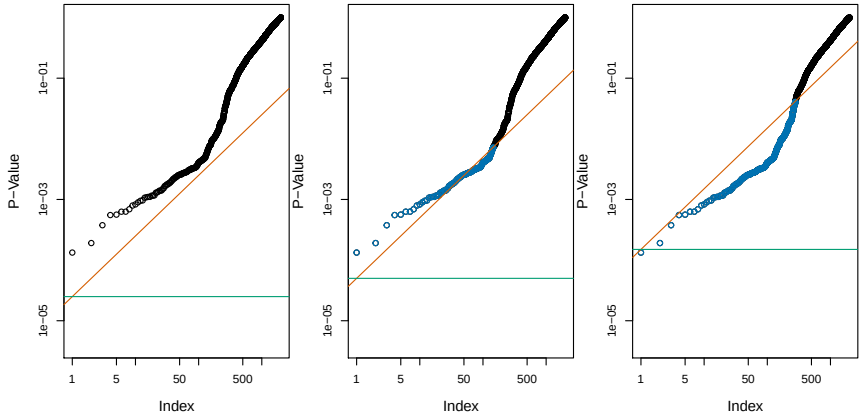


Figure: Panels: same set of $m = 2000$ sorted p-values for the `Fund` dataset. Green lines: thresholds for FWER control (Bonferroni) at $\alpha = 0.05, 0.1, 0.3$ (left to right). Orange lines: thresholds for FDR control (Benjamini–Hochberg) at $q = 0.05, 0.1, 0.3$ (left to right). E.g., When the FDR 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].

Pop-up quiz: Comparing FDR vs. FWER

Suppose there are m hypotheses to test simultaneously. **Which statement best captures the difference between FDR and FWER**?

A) FDR forces the probability of *zero* false positives to stay below α, whereas FWER allows a small fraction q.
B) FDR aims to keep $\mathbb{E}$[fraction of false positives among rejections] $\leq q$, while FWER demands Pr(at least one false positive) $\leq \alpha$.
C) Under FDR control, no false positives are allowed once you discover enough true positives.
D) FDR only works for independent tests, but FWER can handle correlated tests without adjustments.

Answer: B.
FDR control (e.g., Benjamini–Hochberg) allows a certain fraction of false positives *on average*, whereas FWER control (e.g., Bonferroni/Holm) requires the chance of *any* false positive be controlled below α.

Wrap-up

Multiple testing increases the chance of false positives; ordinary $p < 0.05$ rules are not enough when many hypotheses are tested.

FWERcontrols the probability of making false positives:

- Strictly ensures $\Pr(N_{\mathrm{FP}} \geq 1) \leq \alpha$, e.g., via Bonferroni/Holm
- Conservative for large m, leading to fewer rejections & reduced power

FDR controls the expected fraction of false positives among rejections:

- Controls $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] \leq q$, e.g., via Benjamini–Hochberg
- Typically yields more rejections than FWER, especially when m is large

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

Suggested reading: [JWHT21, Ch. 13.3–13.4]

Next lecture: Review for Midterm 2

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.