# Quantitative Research Methods

## Chapter 13: Multiple Testing

Prof. Dr. Christoph Weisser

HSBI

Summer Semester 2026

# The course at a glance

| | Lecture | Topic |
|---|---|---|
| | 1 | Ch. 1 + 2 — Introduction; what is statistical learning |
| | 2 | Ch. 2 — Accuracy, bias–variance, Bayes classifier, KNN |
| | 3–4 | Ch. 3 — Linear regression (estimation, inference, diagnostics) |
| | 5–6 | Ch. 4 — Classification (logistic, LDA/QDA, naive Bayes, ROC) |
| | 7 | Ch. 5 — Resampling (validation, CV, bootstrap) |
| | 8 | Ch. 6 — Model selection and regularisation (ridge, lasso) |
| | 9 | Ch. 7 — Moving beyond linearity (splines, GAMs) |
| | 10 | Ch. 8 — Tree-based methods (trees, forests, boosting) |
| | 11 | Ch. 10 — Deep learning (MLPs, CNNs, training) |
| ▶ | **12** | **Ch. 13 — Multiple testing (FWER, FDR)** |

Twelve sessions of 180 minutes. ▶ marks today's chapter. Short exercises every ∼20 min; extended exercises every ∼45 min; each chapter has a companion lab notebook.

# Contents



**Optional and advanced material is collected in the appendix at the end of the deck.**

# Notation in this chapter

| Symbol | Meaning |
|---|---|
| $m$ | number of hypotheses tested simultaneously |
| $\alpha$ | per-test significance level; target FWER |
| $H_{0j}$ | $j$-th null hypothesis in the family |
| $p_j$, $p_{(j)}$ | $p$-value of test $j$; the $j$-th smallest (sorted) $p$-value |
| $V$ | number of false positives (true nulls rejected) |
| $R$ | total number of rejections, $R = V + S$ |
| $S$ | number of true positives (false nulls rejected) |
| $U$, $T$ | true nulls retained; false nulls retained (missed effects) |
| FWER | family-wise error rate $\Pr(V \geq 1)$ |
| FDR | false discovery rate $E\left[V / \max(R, 1)\right]$ |
| $q$ | target FDR level of Benjamini–Hochberg |
| Power | $E[S]/m_1$, with $m_0$ and $m_1$ the number of true and false nulls ($m_0 + m_1 = m$) |

# Sources and references

## Primary textbook

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning, with Applications in Python.* Springer.

# Learning objectives

By the end of this chapter you will be able to:

- **Explain** the multiple-testing problem and its consequences.
- **Apply** Bonferroni and Holm corrections to control family-wise error rate (FWER).
- **Apply** the Benjamini–Hochberg procedure to control false discovery rate (FDR).
- **Use** resampling-based $p$-values when the null distribution is hard to derive.
- **Avoid** the post-selection-inference and $p$-hacking pitfalls.

# Roadmap of this chapter

**Honest inference at scale**

1. The multiple-testing problem.
2. FWER controls: Bonferroni, Holm.
3. FDR control: Benjamini–Hochberg.
4. Resampling-based $p$-values.
5. Practical issues: $p$-hacking, post-selection inference.

# Where this chapter is used in industry

| Sector | Concrete application | Which decision it drives |
|---|---|---|
| Tech, e-commerce | Thousands of A/B tests a year, each judged on many metrics and segments | Ship or do not ship the change to all users |
| Pharma | Multiple endpoints, dose arms, interim looks and subgroups | Whether a confirmatory claim can go on the label |
| Genomics | Millions of variants in a genome-wide association study | Which variants are reported as hits |
| Asset management | Hundreds of backtested strategy variants on one price history | Which strategy gets capital |
| Marketing | Dozens of creatives, offers and CRM segments tested at once | Which offer is rolled out, not which looked best |
| Banking, insurance | Model-risk fairness and drift tests across many subgroups and metrics | Which gaps are real enough to act on |

**In industry — The common thread**

The bar for "significant" is set by the number of tests you *ran*, not the number you *report*. Genomics even has a standing constant for it: the genome-wide threshold $5 \times 10^{-8}$ is Bonferroni at roughly a million effectively independent tests.

# Industry case in depth: an online experimentation platform

## In industry — The family of tests inside one A/B test

One experiment carries **one primary metric** (say revenue per visitor), a panel of **secondary metrics** (conversion, basket size, retention), a set of **guardrail metrics** (latency, crash rate, unsubscribes) and automatic **segment** breakdowns by country, device and tenure. The statistic per metric is a difference in means with its $p$-value — so the family is easily $m > 100$.

## The rule the platform enforces and who signs off

The primary metric is *pre-registered* and decides ship / no-ship at the conventional level; the secondary panel gets BH-style FDR control; segment findings are labelled **exploratory** and become hypotheses for the next test. The experiment owner proposes, a launch review with product and data science approves. The number lands on the **product owner**, who must defend the call and explain why a segment win is only exploratory.

## Two honest caveats

Repeated *peeking* at a running test is the same problem in the time dimension — hence fixed horizons or sequential ("always-valid") procedures. And metrics within one experiment are strongly correlated, so BH's positive-dependence assumption is an approximation, not a guarantee.

# Outline

# Why naive testing fails at scale

Test $m$ hypotheses simultaneously, each at level $\alpha = 0.05$.

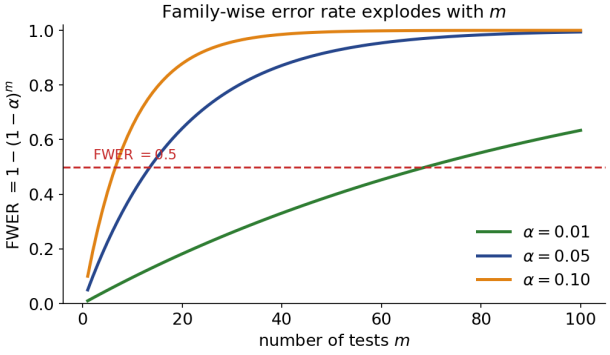
## The arithmetic

- Even if all $m$ nulls are true, we expect $0.05m$ false rejections.
- Test 10 000 genes $\Rightarrow\sim$ 500 false discoveries by chance.
- Naive thresholding fails dramatically.

## In industry — Technology: why "some metric always wins"

A firm running thousands of experiments a year, each scored on dozens of metrics and segments, faces exactly this arithmetic: without control, *something* looks significant in every test. That is how a change with no real effect gets shipped to every user.

# Why the family-wise error rate explodes




**Takeaway**

With $\alpha = 0.05$ the chance of $\geq 1$ false positive passes 50% by only $m = 14$ tests and approaches certainty as $m$ grows.

# The classical hypothesis-test refresher

## How to read this

- $H_0$: null hypothesis. $H_1$: alternative.
- Type I error: reject $H_0$ when true.
- Type II error: fail to reject when false.
- $p$-value: probability under $H_0$ of a test statistic at least as extreme as observed.
- "Reject $H_0$ if $p < \alpha$" controls the per-test Type I error rate at $\alpha$.

## Takeaway

A $p$-value is computed *assuming* $H_0$ is true; it is **not** the probability that $H_0$ is true, and a small $p$ is not proof of a large effect.

# The decision-table view

| | Not rejected | Rejected | Total |
|---|---|---|---|
| $H_0$ true | $U$ | $V$ | $m_0$ |
| $H_0$ false | $T$ | $S$ | $m_1$ |
| Total | $m - R$ | $R$ | $m$ |

## How to read this

- Family-wise error rate (FWER): $\Pr(V \geq 1)$.
- False discovery rate (FDR): $E[V/R \mid R > 0] \Pr(R > 0)$.
- Power: $E[S]/m_1$.

## In industry — Fraud and AML: the table is an alert queue

In a monitoring team $R$ is the alerts raised, $V$ those that waste an investigator's day, $V/R$ the false-positive rate operations argues about, and $T$ the missed fraud nobody sees.

# Exercise 13.1 — The multiple-testing problem [Concept]

## Exercise

1. Define a Type I error and a Type II error in terms of $H_0$.
2. Suppose $m$ hypotheses are tested independently, each at level $\alpha = 0.05$, and *all* nulls are in fact true. Write the probability of making at least one false rejection (the FWER) as a function of $m$.
3. Evaluate this probability for $m = 10$.
4. Find the smallest $m$ for which the FWER reaches 0.5.

# Solution — Exercise 13.1 (1/2)

## Solution

**Method.** The FWER is easiest to reach through its complement: rather than count "at least one false positive" directly, find the single event "no false positive at all" and subtract from 1.

**Part 1 — error types.**

- Type I (false positive): rejecting $H_0$ when it is true.
- Type II (false negative): failing to reject $H_0$ when it is false.

**Part 2 — FWER formula.** One test avoids a false positive with probability $1-\alpha$. Under independence the $m$ tests avoid it *jointly* with probability $(1-\alpha)^m$ (independent probabilities multiply), so the complement is

$$\mathrm{FWER} = \Pr(\text{at least one false positive}) = 1-(1-\alpha)^m.$$

## Common mistake

Adding the per-test rates to get $m\alpha$ — that only approximates the FWER (the first-order term) and can exceed 1.

# Solution — Exercise 13.1 (2/2)

## Solution

**Part 3 — $m = 10$.**

$$1 - (0.95)^{10} = 1 - 0.5987 = 0.4013 \approx 40\%.$$

A 5% per-test rate has ballooned to a 40% chance of at least one spurious "discovery". **Part 4 — FWER $= 0.5$.** Solve $1 - 0.95^m \geq 0.5$, i.e. $0.95^m \leq 0.5$:

$$m \geq \frac{\ln 0.5}{\ln 0.95} = \frac{-0.6931}{-0.05129} = 13.51,$$

so $m = 14$ tests already give a better-than-even chance of a false positive. This is exactly why corrections are needed.

## Common mistake

Dividing by $\ln 0.95$ without flipping the inequality. Since $\ln 0.95 < 0$, $0.95^m \leq 0.5 \iff m \ln 0.95 \leq \ln 0.5 \iff m \geq \ln 0.5 / \ln 0.95$ — the sign flips at the last step.

# Outline

# Bonferroni: split the error budget equally

## Rule

Reject $H_{0j}$ if $p_j \leq \alpha/m$.

## How to read this

- $m$: number of tests in the family; $\alpha$: target FWER (e.g. 0.05).
- Divide the budget evenly: each test is judged at the stricter level $\alpha/m$.
- Equivalently, adjust each $p$-value to $\tilde{p}_j = \min(1,\, m\,p_j)$ and compare to $\alpha$.

## Takeaway

Controls FWER at $\alpha$ under *any* dependence. Conservative: it throws away power, but is simple, popular, and a good starting point.

# Holm's step-down procedure (uniformly beats Bonferroni)

## Algorithm

1. Order $p$-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find the smallest $j$ with $p_{(j)} > \alpha/(m-j+1)$.
3. Reject all $H_{0(k)}$ for $k < j$.

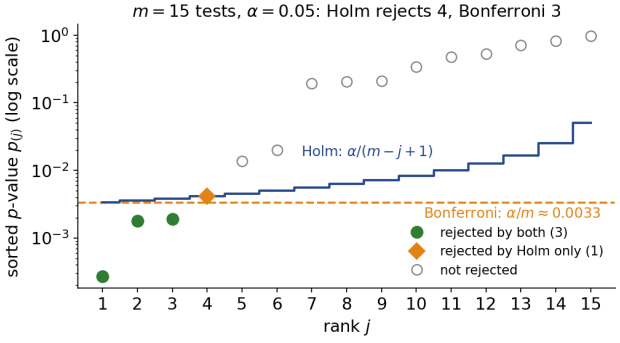
## Why it is more powerful

- The thresholds *grow* down the list: $\frac{\alpha}{m}, \frac{\alpha}{m-1}, \ldots, \alpha$ — the first equals Bonferroni, every later one is looser.
- "Step-down": start at the smallest $p$-value and walk down, stopping at the first test that fails; keep all larger $p$-values.

## Takeaway

Uniformly more powerful than Bonferroni, yet still controls FWER at $\alpha$.

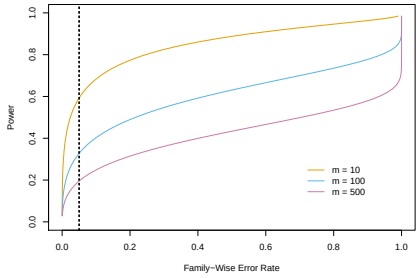
# Bonferroni vs. Holm on 15 simulated $p$-values




**Takeaway**

Bonferroni's flat bar $\alpha/m \approx 0.0033$ rejects 3 hypotheses; Holm's rising thresholds $\alpha/(m-j+1)$ additionally catch $p_{(4)} = 0.0041$, rejecting 4 — Holm always rejects at least as many as Bonferroni.

# Power loss with FWER control




**Takeaway**

Power collapses as $m$ grows: at FWER $= 0.05$ (dashed) it falls from about 0.6 at $m = 10$ to about 0.2 at $m = 500$ — FWER control costs real effects.

*Source: Figure 13.5 — ISLP, James et al. (2023).*

# Exercise 13.2 — Bonferroni correction [Math]

## Exercise

You run $m = 6$ tests and obtain the sorted $p$-values

$$p = (0.001,\ 0.008,\ 0.012,\ 0.030,\ 0.040,\ 0.600),$$

and wish to control the FWER at $\alpha = 0.05$.

1. State the Bonferroni rejection threshold.
2. Which hypotheses are rejected?
3. Compute the Bonferroni-adjusted $p$-values $\tilde{p}_j = \min(1,\ m\,p_j)$ and confirm they give the same decisions.

# Solution — Exercise 13.2

## Solution

**Step 1 — threshold.** $\alpha/m = 0.05/6 = 0.00833$.

**Step 2 — reject if $p_j \le 0.00833$.**

| $p_j$ | 0.001 | 0.008 | 0.012 | 0.030 | 0.040 | 0.600 |
|---|---|---|---|---|---|---|
| $\le 0.00833$? | ✓ | ✓ | × | × | × | × |

Reject $H_1$ and $H_2$ only ($0.008 < 0.00833$; $0.012 > 0.00833$).

**Step 3 — adjusted $p$-values $\tilde{p}_j = \min(1, 6p_j)$:**

$$(0.006,\ 0.048,\ 0.072,\ 0.180,\ 0.240,\ 1.000).$$

Rejecting when $\tilde{p}_j \le 0.05$ again flags exactly $H_1$ (0.006) and $H_2$ (0.048). Two rejections.

**Take-away.** Bonferroni is simple and valid under any dependence, but conservative — here it misses $H_3$ at $p = 0.012$.

*Common mistake:* correcting twice — compare raw $p_j$ with $\alpha/m$, *or* adjusted $\tilde{p}_j$ with $\alpha$. Comparing $\tilde{p}_j$ with $\alpha/m$ double-penalises.

# Exercise 13.3 — Holm's step-down procedure [Math]

## Exercise

Using the same data as Exercise 13.2,

$$p = (0.001,\ 0.008,\ 0.012,\ 0.030,\ 0.040,\ 0.600), \quad m = 6,\ \alpha = 0.05,$$

apply Holm's step-down procedure. For each rank $j$ compare $p_{(j)}$ with $\alpha/(m-j+1)$, stop at the first failure, and report the rejections. Compare with Bonferroni.

# Solution — Exercise 13.3

## Solution

The $p$-values are already sorted. Compare each with its Holm threshold $\alpha/(m-j+1)$:

| $j$ | $p_{(j)}$ | $\alpha/(m-j+1)$ | decision |
|---|---|---|---|
| 1 | 0.001 | $0.05/6 = 0.00833$ | $0.001 \le 0.00833$ ✓ |
| 2 | 0.008 | $0.05/5 = 0.01000$ | $0.008 \le 0.01000$ ✓ |
| 3 | 0.012 | $0.05/4 = 0.01250$ | $0.012 \le 0.01250$ ✓ |
| 4 | 0.030 | $0.05/3 = 0.01667$ | $0.030 > 0.01667$ × |

The first failure is at $j = 4$. Holm rejects all hypotheses with rank $< 4$, i.e. $H_{(1)}, H_{(2)}, H_{(3)}$ — **three** rejections.

**Comparison.** Holm also rejects $H_{(3)}$ at $p = 0.012$, which Bonferroni missed. Holm is uniformly more powerful than Bonferroni while still controlling the FWER at $\alpha$, because its first threshold matches Bonferroni's but every later one is larger.

*Common mistake:* continuing down the list after the first failure. Holm is *step-down*: once $p_{(4)}$ fails, ranks 4–6 are all retained — even if a later $p_{(j)}$ were to sit below its own threshold.

# Outline

# The Benjamini–Hochberg (BH) procedure

## Algorithm (control FDR at level $q$)

1. Order the $p$-values: $p_{(1)} \leq \cdots \leq p_{(m)}$.
2. Find the largest $L$ such that $p_{(L)} \leq Lq/m$.
3. Reject the $L$ smallest hypotheses.

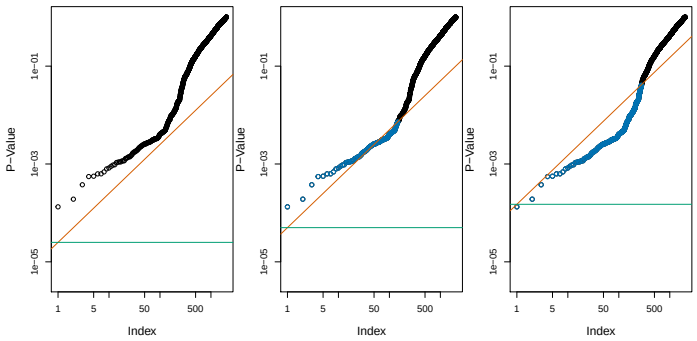
## Reading the threshold $Lq/m$

- $q$: target FDR (e.g. 0.10); $m$: number of tests; $L$: rank in the sorted list.
- The bar $Lq/m$ *rises* with rank, unlike Bonferroni's flat $\alpha/m$.
- "Step-up": scan from the *largest* $p$-value down; reject everything up to the first one that clears its bar.

## Takeaway

Under independence (or positive dependence), BH controls FDR at $q$.

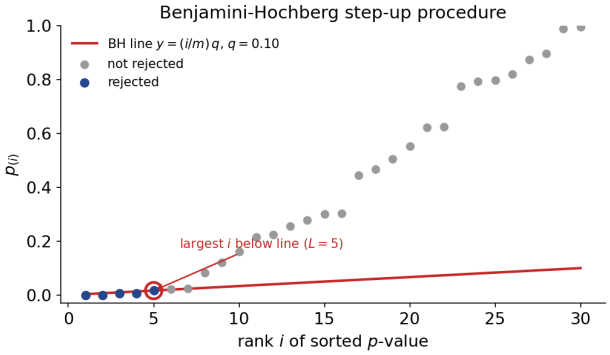
# Visualising BH




Sorted $p$-values vs. the line $Lq/m$. Reject everything below the highest crossing.

*Source: Figure 13.6 — ISLP, James et al. (2023).*

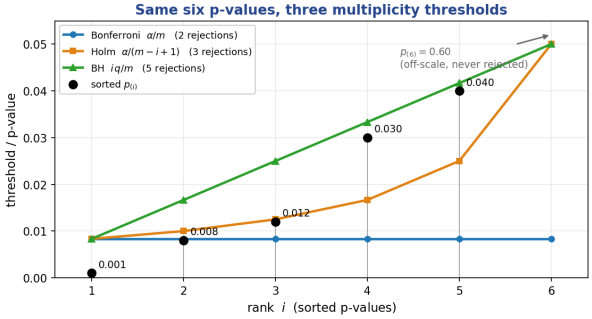
# Benjamini–Hochberg as a staircase




## Takeaway

Compare each sorted $p_{(i)}$ with the line $(i/m)q$; take the *largest* rank below it and reject everything up to there (5 of 30 here at $q = 0.10$).

# Three thresholds on the same six p-values




## Takeaway

A $p$-value is a candidate for rejection when it sits *below* its method's curve. All three meet at rank 1; then Bonferroni stays flat while Holm and BH rise — so here they reject 2, 3 and 5 hypotheses (Exercises 13.2/13.3/13.5).

# FDR vs. FWER: which to use?

## How to read this

- Use **FWER** when even one false discovery is unacceptable (drug-approval trials).
- Use **FDR** when an exploratory list of candidates is acceptable and false positives are tolerable in moderation (genomics, A/B testing).
- BH is much more powerful than Bonferroni when $m$ is large and a non-trivial fraction of nulls are false.

## In industry — Governance: FWER or FDR?

Use **FWER** when a single false claim is expensive and public: a drug label, a regulatory filing, a change shipped to every user. Use **FDR** when the output is a *ranked shortlist for further investigation* — candidate biomarkers, leads for a fraud team, hypotheses for the next experiment — and a known fraction of false leads is an acceptable cost of doing business. Regulators expect the first: FDA guidance on multiple endpoints requires a *pre-specified* multiplicity strategy for any confirmatory claim, and an unadjusted subgroup finding does not support a label claim.

# Exercise 13.4 — FDR vs. FWER [Concept]

## Exercise

1. Give the definitions of the FWER and the FDR in terms of the decision-table quantities $V$ (false rejections) and $R$ (total rejections).
2. Explain why controlling the FDR is generally less conservative (more powerful) than controlling the FWER.
3. For each scenario, state whether FWER or FDR control is more appropriate and why: (i) a confirmatory phase-III drug trial; (ii) screening 20 000 genes for follow-up experiments.

# Solution — Exercise 13.4 (1/2)

## Solution

**Part 1 — definitions.**

$$\text{FWER} = \Pr(V \geq 1), \qquad \text{FDR} = E\left[\frac{V}{R} \,\middle|\, R > 0\right] \Pr(R > 0).$$

FWER is the probability of *any* false positive; FDR is the *expected proportion* of false positives among the rejections.

**Part 2 — why FDR is more powerful.** FWER guards against a single false positive anywhere, so its threshold must shrink with $m$ (roughly $\alpha/m$). FDR only requires the false positives to be a small *fraction* of discoveries, so it tolerates a few errors when there are many true discoveries. This lets it use a larger, adaptive threshold and reject far more when many nulls are false.

# Solution — Exercise 13.4 (2/2)

## Solution

**Part 3 — scenarios.**

(i) **FWER** (Bonferroni/Holm): a single false approval is unacceptable; the cost of any Type I error is very high.

(ii) **FDR** (Benjamini–Hochberg): an exploratory candidate list is the goal; a modest fraction of false leads is acceptable and will be filtered by follow-up experiments. FWER here would be far too conservative and kill power.

## Common mistake

Defaulting to FWER "to be safe". Safety is not free — the stricter criterion buys its guarantee with missed true effects, so the choice is a design decision, not a formality.

# Exercise 13.5 — Benjamini–Hochberg procedure [Math]

## Exercise

Using the same six $p$-values,

$$p = (0.001,\ 0.008,\ 0.012,\ 0.030,\ 0.040,\ 0.600), \quad m = 6,$$

apply the Benjamini–Hochberg (BH) procedure at FDR level $q = 0.05$.

1. For each rank $L$ compute the BH threshold $Lq/m$.
2. Find the largest $L$ with $p_{(L)} \leq Lq/m$.
3. State the rejections and compare with Bonferroni and Holm.

# Solution — Exercise 13.5 (1/2)

## Solution

**Steps 1–2 — compare each $p_{(L)}$ with $Lq/m$,** where $q/m = 0.05/6 = 0.00833$:

| $L$ | $p_{(L)}$ | $Lq/m$ | $p_{(L)} \leq Lq/m$? |
|---|---|---|---|
| 1 | 0.001 | 0.00833 | ✓ |
| 2 | 0.008 | 0.01667 | ✓ |
| 3 | 0.012 | 0.02500 | ✓ |
| 4 | 0.030 | 0.03333 | ✓ |
| 5 | 0.040 | 0.04167 | ✓ |
| 6 | 0.600 | 0.05000 | × |

The largest $L$ satisfying the condition is $L = 5$.

# Solution — Exercise 13.5 (2/2)

## Solution

**Step 3 — reject.** BH rejects the 5 smallest hypotheses $H_{(1)}, \ldots, H_{(5)}$ (everything except $p = 0.600$).

**Comparison on identical data.**

$$\text{Bonferroni: } 2 \;<\; \text{Holm: } 3 \;<\; \text{BH: } 5 \text{ rejections.}$$

Controlling the FDR yields many more discoveries than controlling the FWER — at the price of allowing a small expected fraction ($\leq q$) of them to be false. Note BH is a *step-up* rule: even though $p_{(4)} = 0.030$ and $p_{(5)} = 0.040$ individually clear their thresholds here, BH would reject them via the largest passing $L$ regardless of any earlier borderline case.

## Common mistake

Treating $q$ like $\alpha$. They bound different things: $\alpha$ caps the chance of *any* false rejection; $q$ caps only the expected *fraction* of the rejected list that is false.

# Extended Exercise 13.1 — Full workflow on ten p-values [Math]

## Extended exercise (15 min)

Ten hypotheses give the sorted $p$-values

$$p = (0.0008,\ 0.0041,\ 0.0060,\ 0.0100,\ 0.0180,\ 0.0290,\ 0.0850,\ 0.150,\ 0.240,\ 0.790),$$

with $m = 10$. Control the FWER at $\alpha = 0.05$ (Bonferroni, Holm) and the FDR at $q = 0.10$ (Benjamini–Hochberg).

1. List the rejections under **Bonferroni**, **Holm** and **BH**.
2. Verify the ordering $R_{\text{Bonf}} \leq R_{\text{Holm}} \leq R_{\text{BH}}$.
3. State the guaranteed FWER bound, and comment on what the FDR guarantee buys.

# Solution — Extended Exercise 13.1 (1/3)

## Solution — Bonferroni and Holm (FWER $\leq 0.05$)

**Bonferroni:** reject if $p_j \leq \alpha/m = 0.005$. Only $p_{(1)} = 0.0008$ and $p_{(2)} = 0.0041$ qualify $\Rightarrow$ **2 rejections**.
**Holm:** compare $p_{(j)}$ with $\alpha/(m-j+1)$, stopping at the first failure:

| $j$ | $p_{(j)}$ | $\alpha/(m-j+1)$ | |
|---|---|---|---|
| 1 | 0.0008 | $0.05/10 = 0.00500$ | ✓ |
| 2 | 0.0041 | $0.05/9 = 0.00556$ | ✓ |
| 3 | 0.0060 | $0.05/8 = 0.00625$ | ✓ |
| 4 | 0.0100 | $0.05/7 = 0.00714$ | × stop |

Holm rejects ranks 1–3 $\Rightarrow$ **3 rejections**.

# Solution — Extended Exercise 13.1 (2/3)

## Solution — Benjamini–Hochberg (FDR $\leq 0.10$)

Compare each $p_{(L)}$ with $Lq/m = 0.01L$ and take the *largest* passing $L$:

| $L$ | $p_{(L)}$ | $Lq/m$ | |
|---|---|---|---|
| 5 | 0.0180 | 0.050 | ✓ |
| 6 | 0.0290 | 0.060 | ✓ ← largest |
| 7 | 0.0850 | 0.070 | × |
| 8 | 0.150 | 0.080 | × |

(Ranks 1–4 pass too.) The largest $L$ with $p_{(L)} \leq Lq/m$ is $L = 6$, so BH rejects the six smallest $\Rightarrow$ **6 rejections** — including $p_{(6)} = 0.029$ and ranks $4, 5$ that the FWER rules missed.

# Solution — Extended Exercise 13.1 (3/3)

## Solution — comparison and error guarantees

$$\underbrace{2}_{\text{Bonferroni}} \leq \underbrace{3}_{\text{Holm}} \leq \underbrace{6}_{\text{BH}}.$$

**FWER bound.** Bonferroni and Holm both guarantee $\mathrm{FWER} = \Pr(V \geq 1) \leq \alpha = 0.05$ under *any* dependence: the chance of *even one* false rejection is at most 5%. **What FDR buys.** BH instead guarantees $\mathrm{FDR} = E[V/R] \leq q\, m_0/m \leq 0.10$. Among its 6 discoveries, at most an *expected* 10% ($\lesssim 1$ here) are false. Tolerating a small false *fraction* rather than *any* false positive is exactly why BH rejects more.

## Common mistake

Reading "FDR $\leq 0.10$" as "each discovery is 90% likely true". The FDR is an average over the whole rejected list (and over repetitions), not a per-discovery probability.

# Outline

# $p$-hacking and selection bias

## Don't

- Report only the most-significant subset of tests.
- Test thousands of hypotheses and call the smallest $p$-value "significant".

## Takeaway

Multiple-testing correction must include *every* test you ran, not only the ones you report. Pre-register hypotheses where possible; out-of-sample replication is the gold standard.

## In industry — Quantitative finance: backtest overfitting

Backtesting hundreds of strategy variants on the same price history guarantees an impressive Sharpe ratio by chance. Practitioners call this data snooping, deflate the reported performance for the number of trials, and insist on out-of-sample and out-of-time evidence before allocating capital — the Chapter 5 lesson, restated as a multiplicity problem.

# Extended Exercise 13.2 — FWER or FDR? A design decision [Integrative]

## Extended exercise (15 min)

For each study, choose an error criterion and justify it, using the naive expected number of false discoveries $E[V] = m_0\,\alpha$ (all-null worst case $m_0 = m$) and the power consequences.

1. **Scenario A — confirmatory phase-III trial:** $m = 4$ co-primary endpoints; one false "significant" endpoint could license an ineffective drug.
2. **Scenario B — exploratory screen:** $m = 20\,000$ genes tested for follow-up; a shortlist with a few false leads is acceptable.
3. Contrast the statistical power the two criteria leave on the table.

# Solution — Extended Exercise 13.2 (1/3)

## Solution — scenario A: control FWER

With $m = 4$ tests at $\alpha = 0.05$, naive testing expects

$$E[V] = m_0\alpha = 4(0.05) = 0.20 \text{ false positives}, \qquad \mathrm{FWER} = 1 - 0.95^4 = 0.185.$$

A $\sim$19% chance of *any* false endpoint is unacceptable when the decision is regulatory and irreversible. **Choose FWER control** (Bonferroni threshold $\alpha/m = 0.0125$, or Holm). One false positive is the outcome we must not allow, so we pay the power cost gladly — with only $m = 4$ tests that cost is tiny.

# Solution — Extended Exercise 13.2 (2/3)

## Solution — scenario B: control FDR

With $m = 20\,000$ and (worst case) $m_0 \approx 20\,000$, naive testing expects

$$E[V] = m_0\alpha = 20\,000(0.05) = 1000 \text{ false discoveries!}$$

Bonferroni would shrink the threshold to $\alpha/m = 2.5 \times 10^{-6}$, cutting $E[V]$ to $\approx 0.05$ but **destroying power** for realistic effect sizes. **Choose FDR control** (BH at $q = 0.10$): accept that an expected $\leq 10\%$ of the shortlist is false in exchange for detecting far more true signals — the false leads are removed by follow-up experiments.

# Solution — Extended Exercise 13.2 (3/3)

## Solution — the power trade-off

| | Confirmatory ($m = 4$) | Exploratory ($m = 20\,000$) |
|---|---|---|
| Criterion | FWER (Bonferroni/Holm) | FDR (BH) |
| Threshold shape | $\alpha/m$ (fixed, tiny) | adaptive $Lq/m$ |
| Naive $E[V]$ | 0.20 | 1000 |
| Priority | no false positive | many true findings |

**Trade-off.** The FWER threshold shrinks like $\alpha/m$, so power collapses as $m$ grows — fine for a handful of endpoints, ruinous for 20 000 genes. FDR asks only that false positives be a small *fraction* of discoveries, so its adaptive threshold keeps power high at scale. **Rule of thumb:** match the criterion to the cost of a single false positive.

## Common mistake

Quoting $E[V] = m\alpha$ as a fact about your study — it is the all-null worst case ($m_0 = m$). With real effects present the count is smaller, though still ruinous at scale.

# Outline

# Python lab — run it live in the notebook

## Companion notebook: `chapter_13_lab.ipynb`

The code in this section is meant to be **demonstrated live**. Open the companion Jupyter notebook and run it cell by cell:

- §1, *Simulation* — $m = 1000$ tests with a fraction of true alternatives;
- §2, *Compare procedures* — Bonferroni, Holm and Benjamini–Hochberg side by side;
- §3, *Visualise the Benjamini–Hochberg line*, then §4, *Permutation test*.

Data loads via the `ISLP` package or the bundled CSV files; §5, *Exercises* ends the notebook with hands-on tasks.

# Bonferroni, Holm, BH in statsmodels

```
from statsmodels.stats.multitest import multipletests # corrections
import numpy as np # arrays

p = np.array([1e-6, 0.01, 0.04, 0.05, 0.30, 0.45, 0.80]) # raw p-values

for method in ["bonferroni", "holm", "fdr_bh"]: # FWER, FWER, FDR control
    rej, p_corr, *_ = multipletests(p, alpha=0.05, method=method) # adjust
    print(f"{method:12s}", rej) # rejection flags (bool)
    print(f"{'p_adj':12s}", np.round(p_corr, 4)) # adjusted p-values
```

# Extended Exercise 13.3 — Simulating FDR and power [Python]

## Extended exercise (15 min)

Simulate $m = 1000$ one-sided $z$-tests of which $m_1 = 100$ have a true effect (mean shift $\mu = 3$); the other 900 are null. Over many repeats, apply Benjamini–Hochberg at $q = 0.10$ and Bonferroni at $\alpha = 0.05$, and estimate the *realised* FDR ($V/R$) and power ($S/m_1$) of each.

1. Write the simulation and run $\sim 2000$ repeats.
2. Report the realised FDR and power for BH and for Bonferroni.
3. Check BH against its theoretical bound $q\, m_0/m$, and comment on the power gap.

## Run this live

The worked solution sits in `chapter_13_lab.ipynb` under *Lecture exercises — worked Python solutions*: the cell headed *Extended Exercise 13.3 — Simulating FDR and power [Python]*.

# Solution — Extended Exercise 13.3 (1/3)

## Solution — code: setup and the BH rule

```
import numpy as np # arrays and random numbers
from scipy import stats # normal cdf for p-values
rng = np.random.default_rng(0) # seeded generator: reproducible
m, m1, mu = 1000, 100, 3.0 # 100 true effects, mean shift mu
q, alpha = 0.10, 0.05 # FDR target q, FWER target alpha
truth = np.arange(m) < m1 # first m1 tests are non-null

def bh_reject(p, q): # Benjamini--Hochberg step-up
    order = np.argsort(p); ps = p[order] # sort p-values ascending
    thr = (np.arange(1, m + 1) / m) * q # BH line: (j/m) * q
    hit = np.where(ps <= thr)[0] # ranks below the line
    rej = np.zeros(m, bool) # start: reject nothing
    if len(hit): rej[order[:hit.max() + 1]] = True # up to largest hit
    return rej # boolean rejection mask
```

# Solution — Extended Exercise 13.3 (2/3)

## Solution — code: simulate FDR and power

```
def simulate(method, reps=2000): # average FDR / power over reps
    fdr = pw = 0.0 # running totals
    for _ in range(reps):
        z = rng.normal(0, 1, m); z[:m1] = rng.normal(mu, 1, m1) # effects: mean mu
        p = 1 - stats.norm.cdf(z) # one-sided p-values
        rej = bh_reject(p, q) if method == "bh" else (p <= alpha / m) # BH or Bonf.
        V = (rej & ~truth).sum(); S = (rej & truth).sum(); R = rej.sum() # V, S, R
        fdr += V / max(R, 1); pw += S / m1 # realised FDP and power
    return fdr / reps, pw / reps # averages over the repeats

for meth in ["bh", "bonferroni"]: # compare both procedures
    f, pw = simulate(meth) # run the experiment
    print("%-11s FDR=%.3f power=%.3f" % (meth, f, pw)) # summary line
```

# Solution — Extended Exercise 13.3 (3/3)

## Solution — expected numbers and interpretation

**Typical printout.**

```
bh  FDR=0.090  power=0.723
bonferroni  FDR=0.002  power=0.186
```

- **BH** holds the realised FDR at 0.090 — essentially its theoretical ceiling $q\,m_0/m = 0.10 \times 900/1000 = 0.09$ — while recovering about **72%** of the true effects.
- **Bonferroni** is far stricter: realised FDR $\approx 0.002$, but power collapses to about **19%** — it misses four-fifths of the signals.
- **Lesson:** at $m = 1000$, FDR control detects roughly *four times* as many true effects as FWER control, for a small, bounded false fraction.

*Common mistake:* expecting the realised FDR to hit $q = 0.10$ exactly. The guarantee is $q\,m_0/m = 0.09$ here — BH adapts to the unknown share of true nulls, and any single run fluctuates around it.

# Outline

# Chapter 13 in one slide

## What we accomplished

Honest inference when many hypotheses are tested at once.

- Set up the multiple-testing problem.
- Family-wise error rate (FWER): Bonferroni and Holm.
- False discovery rate (FDR): Benjamini–Hochberg.
- Resampling-based $p$-values and Storey's $q$-values.
- Pitfalls of post-selection inference and $p$-hacking.

# Key formulas at a glance

**FWER** $\mathrm{FWER} = \Pr(V \geq 1) = 1 - (1-\alpha)^m$ (independent tests).

**Bonferroni** reject $H_{0j}$ if $p_j \leq \alpha/m$ (controls FWER).

**Holm** reject while $p_{(j)} \leq \dfrac{\alpha}{m-j+1}$, stepping down from the smallest.

**FDR** $\mathrm{FDR} = E\left[\dfrac{V}{\max(R,1)}\right]$ (expected false-discovery proportion).

**Benjamini–Hochberg** reject the $L$ smallest, where $L = \max\{j : p_{(j)} \leq \frac{j}{m} q\}$.

**BH guarantee** $\mathrm{FDR} = E[V/R] \leq q\, m_0/m \leq q$ ($m_0$ true nulls; independence or positive dependence).

**Power** $\mathrm{Power} = E[S]/m_1$.

# Vocabulary checklist

| Term | Meaning |
|---|---|
| Type I error | Reject $H_0$ when true |
| Type II error | Fail to reject when false |
| FWER $= \Pr(V \geq 1)$ | Family-wise error rate |
| FDR $= E[V/R]$ | False discovery rate |
| Bonferroni / Holm | FWER controls |
| BH procedure | Step-up FDR control |
| Storey's $q$-value | FDR-calibrated $p$-value |
| Permutation test | Null distribution by relabelling |
| Knockoffs | Modern FDR control with covariate engineering |

# Decision rules of thumb

- Always specify the family of tests *before* looking at the data.
- Confirmatory / regulated $\Rightarrow$ FWER (Bonferroni or Holm).
- Exploratory / discovery $\Rightarrow$ BH on FDR.
- Report both raw and adjusted $p$-values.
- Replicate findings on an independent sample whenever possible.
- For valid post-selection inference: sample-splitting, selective inference, knockoffs.

# Common pitfalls

## Don't

1. Test thousands of hypotheses and report only the "significant" ones.
2. Use BH when tests are strongly negatively dependent.
3. Confuse FWER and FDR — they answer different questions.
4. Apply post-hoc multiple-testing fixes after looking at the data.
5. Use uncorrected $p$-values from stepwise / lasso.
6. Treat a $p$-value as a probability that $H_0$ is true.

# Course-wide summary

## The conceptual arc

1. Vocabulary and bias-variance (Ch. 1–2).
2. Linear baselines and how to evaluate them (Ch. 3–5).
3. Regularisation and dimension reduction (Ch. 6).
4. Adding flexibility: splines, GAMs, trees, SVMs, deep nets (Ch. 7–10).
5. Special responses: censored times, no response (Ch. 11–12).
6. Honest inference at scale (Ch. 13).

# Closing thoughts

**The four habits of an effective statistical learner**

1. Always plot the data.
2. Always validate on data you didn't train on.
3. Always interpret what the model says about the world.
4. Always check assumptions and limitations before claiming results.

Thank you for taking this course. Good luck applying statistical learning to your own data!

# Appendix — what is in here, and why

Nothing in the appendix is needed to follow the chapter: each slide is a formal derivation, a longer worked exercise, or a side topic. Read it when you want the full story, or when a later chapter sends you back here.

## Contents of the appendix

- **The four outcomes, drawn** — a diagram of the same $2 \times 2$ table the main deck already gives as a table.
- **Resampling-based inference** — permutation $p$-values, the resampling estimate of the FDR, and what the histogram of many $p$-values looks like. Use it when the null distribution is not available in closed form.
- **Post-selection inference** — why a $p$-value computed after looking at the data is not a $p$-value. Important in research practice, beyond the exam.

# The four outcomes of a test decision

| | **Not rejected** | **Rejected** | |
|---|---|---|---|
| $H_0$ **true** | $U$ true negatives | $V$ Type I (false pos.) | $m_0$ |
| $H_0$ **false** | $T$ Type II (false neg.) | $S$ true positives | $m_1$ |
| | $m - R$ | $R = V + S$ | |

**Takeaway**

**Type I errors** are cell $V$ — rejected true nulls. The FWER $= \Pr(V \geq 1)$ guards against *any* false positive; the FDR $= E[V/R]$ controls their *expected fraction* among the $R$ rejections.

# Resampling-based $p$-values

When the test statistic's null distribution is hard to derive:

## How to read this

- **Permutation**: shuffle the response many times to get a null distribution.
- **Bootstrap (under $H_0$)**: resample under a constraint that enforces the null.
- Empirical $p$-values are then thresholded by Bonferroni / Holm / BH as before.

## Takeaway

Empirical $p$-value $= \dfrac{1 + \#\{\text{resampled statistics} \geq \text{observed}\}}{1 + B}$ over $B$ resamples — no parametric null needed. The "$+1$" keeps it strictly positive.

# Resampling-based FDR

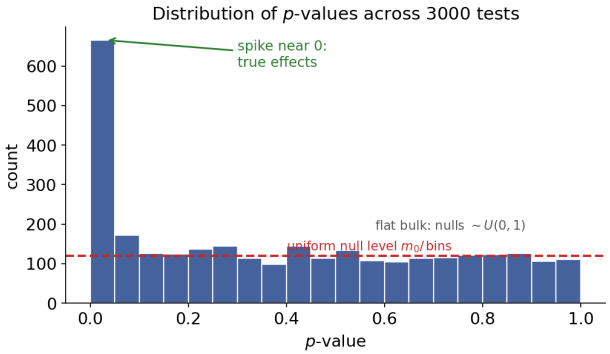
## Storey's $q$-value

Estimate the proportion of true nulls $\hat{\pi}_0$ from the $p$-value distribution. Calibrate FDR more accurately than BH. Especially relevant when many nulls are false.

## How $\hat{\pi}_0$ is read off

True nulls give $p$-values that are $\sim U(0,1)$, so the histogram is *flat* away from 0; the height of that flat part estimates $\hat{\pi}_0$. BH implicitly assumes $\hat{\pi}_0 = 1$; plugging in the smaller true value makes FDR control less conservative and recovers more discoveries.

# The distribution of $p$-values under many tests




**Takeaway**

True nulls give $p$-values that are $\sim U(0,1)$ (the flat bulk), while true effects pile up near 0. Storey's method reads $\hat{\pi}_0$ off the height of that flat part to calibrate the FDR.

# Post-selection inference

After variable selection (lasso, stepwise) standard $p$-values are invalid.

**Takeaway**

The bias: the *same data* both *chose* the hypothesis and *tested* it, so the reported $p$-values are optimistically small. Valid inference must account for the selection step.

**Modern methods**

- Sample splitting / data-carving.
- Selective inference (Lee, Sun, Sun, Taylor).
- Knockoffs (Barber, Candès).