Lecture 10: Statistical Power & Multiple Hypothesis Correction

GENOME 560, Spring 2026

Saori Sakaue (sakaue@uw.edu)

Section 1

Statistical Power

The Null and Alternative Hypothesis

The null hypothesis, H_0:

- States the assumption (numerical) to be tested
- Begin with the assumption that the null hypothesis is TRUE
- Always contains the "$=$" sign

The alternative hypothesis, H_A:

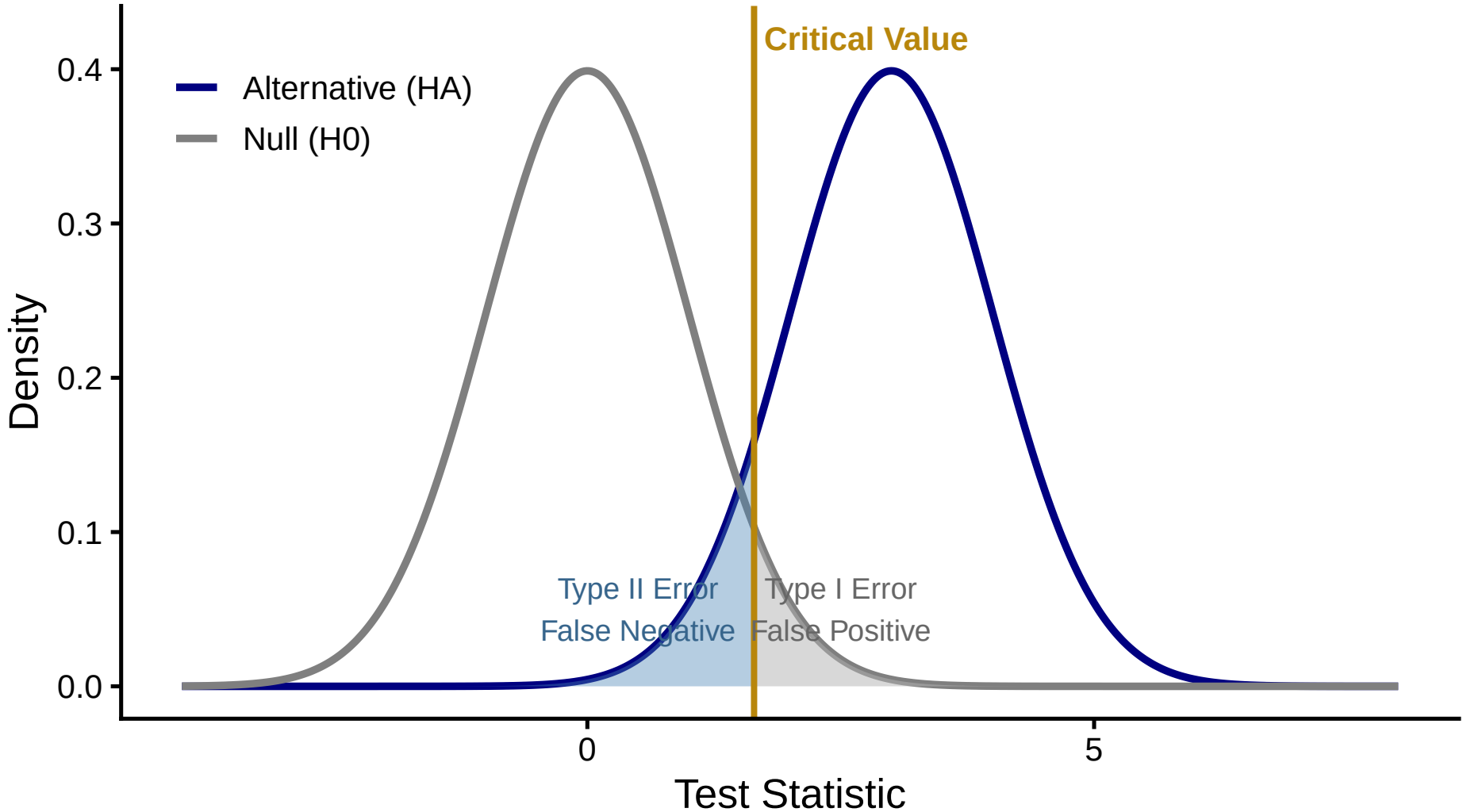
- Is the 'opposite' of the null hypothesis
- Challenges the status quo
- Never contains just the "$=$" sign
- Is generally the hypothesis believed to be true by the researcher

Errors in Hypothesis Testing

	H_0 True	H_0 False
Do Not Reject H_0	**Correct** $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	**Correct** $(1-\beta)$

- $\alpha = P(\text{Type I Error})$ — **False Positive**
- $\beta = P(\text{Type II Error})$ — **False Negative**

Errors in Hypothesis Testing — Visualized



Controlling α = Type I Error

We set $\alpha = 0.05$...

What does this mean?

- That 5% of the time we will reject a true null hypothesis

The α (controlling for false positive) and β (controlling for false negative) are **trade-offs**, depending on the 'critical value'

Statistical Power $= 1 - \beta$

Power is the probability of correctly rejecting the null when the alternative is true (i.e., "true positive decision")

$$\text{power} = 1 - P(\text{type II error}) = 1 - \beta$$

If we aim for a power $= 0.8$ or greater, that means an 80% chance you'll correctly reject the null hypothesis

Why Analyze Power?

Power analysis enables you to construct your experiment to ensure that you can detect biologically relevant/interesting changes

Nothing is worse than realizing you likely would never have succeeded in detecting the effect you were looking for in the first place!

Components of Power Analysis

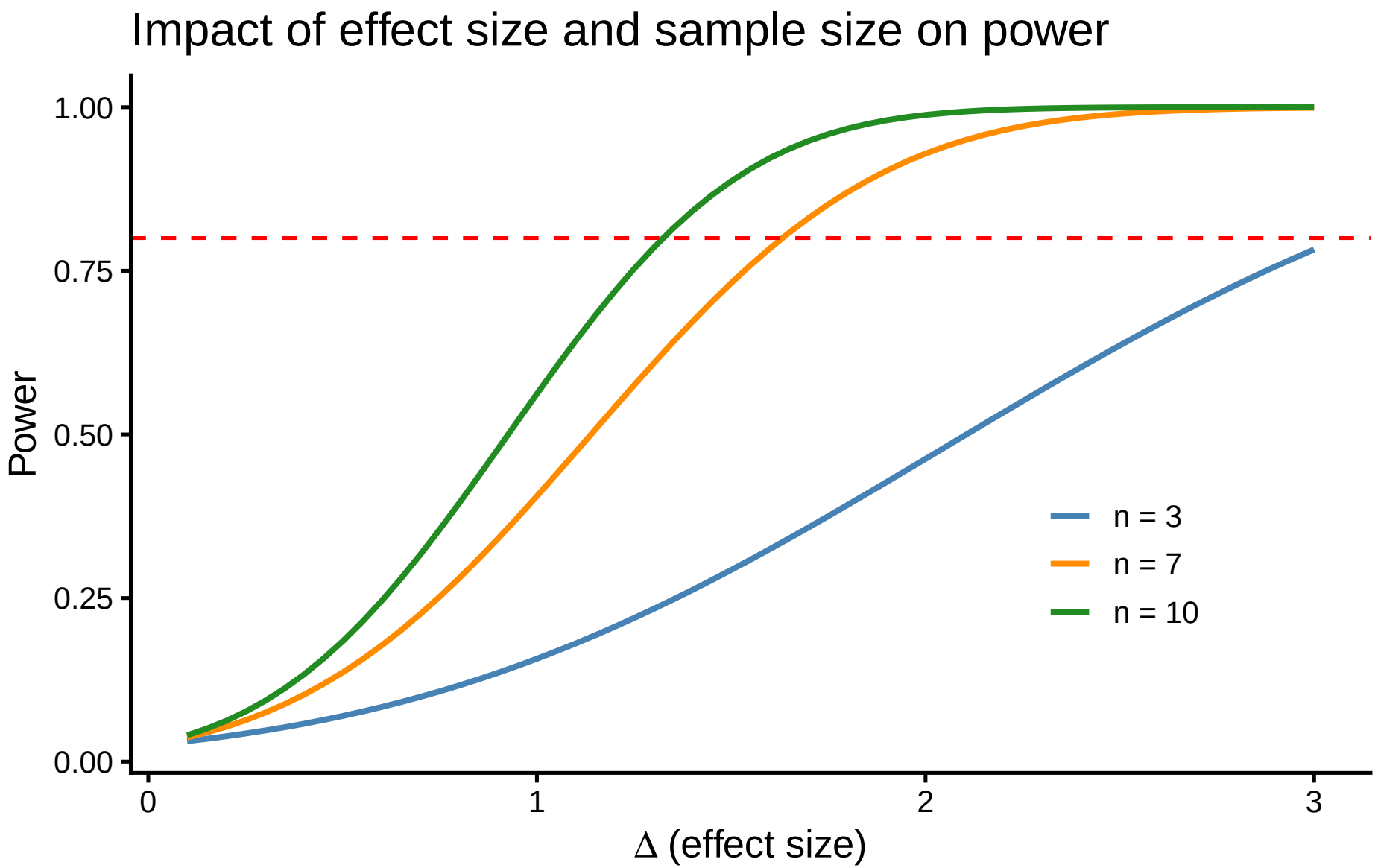
To do a power analysis, we need to know:

- σ^2 = variability/noise (as σ^2 decreases, power increases)
- Δ = distance between H_0 and H_1 (as Δ increases, power increases)
- n = sample size
- d = effect size = $\Delta/\sigma = (\mu_A - \mu_0)/\sigma$

And α?

- α = type I error rate (as α increases, power $[1 - \beta]$ increases)

Components of Power Analysis — Visualized



Power Analysis of Transcript Abundance

Let's say we wish to test whether a particular transcript is present at 4 copies per cell

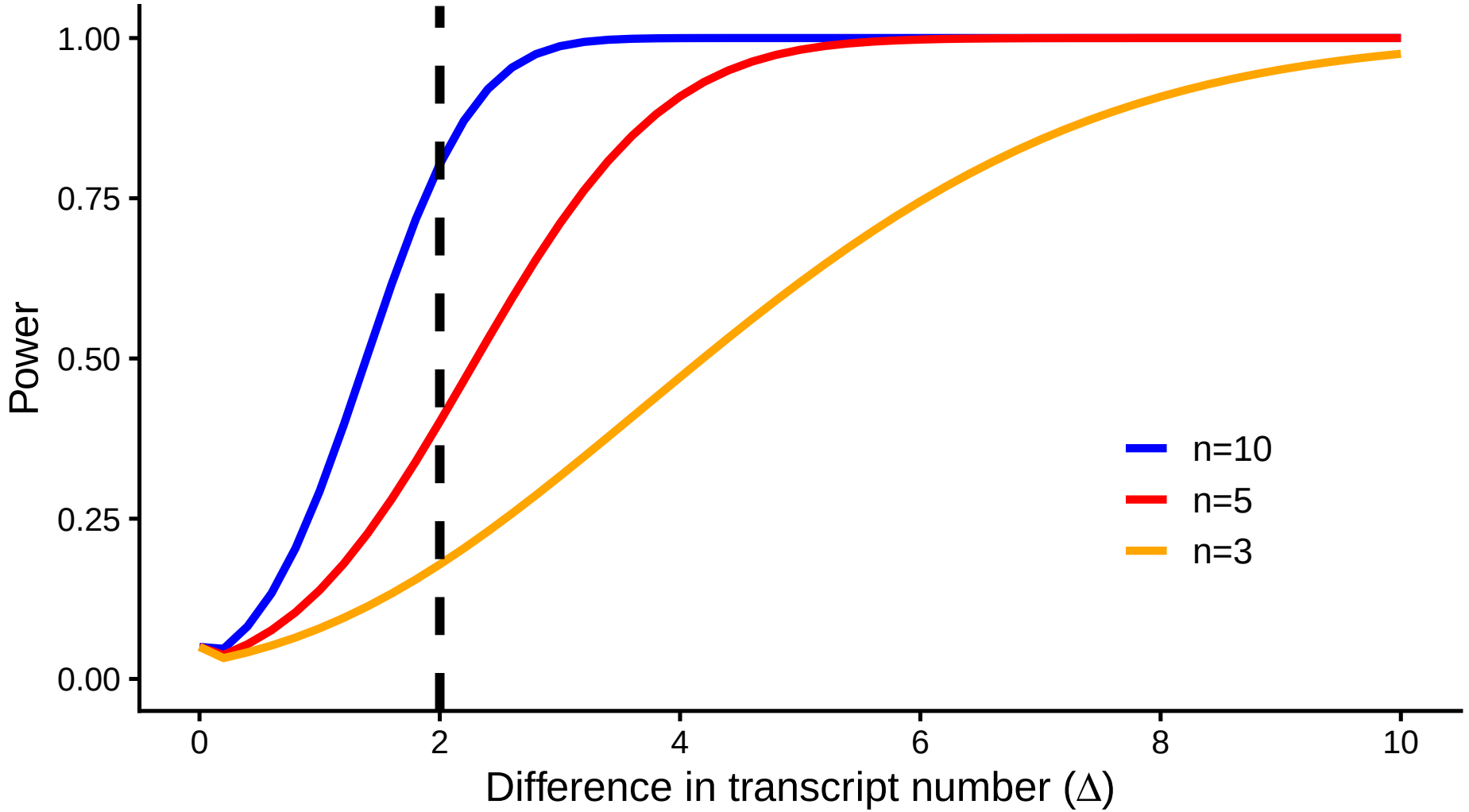
- We want to know the **power** to detect a difference if there are 6 copies ($\Delta = 2$)

$$H_0 : \mu = 4 \qquad H_1 : \mu = 6$$

- From past data, $\sigma = 2$
- Let's examine power for each sample size (n) with $\alpha = 0.05$

We Examine a Range of Alternate Hypotheses...

...and generate a power curve:



Section 2

Multiple Hypothesis Correction

If You Were a Prophet...

- You have 27 clients
- You will predict the weather — sunny, cloudy, rainy — for three consecutive days
- There are $3^3 = 27$ combinations of the weather

- If you tell each client one of all possible weather combinations, you are 100% likely to get one client who will get the correct weather prediction from you
- **Without any ability to predict the weather!**

Why Multiple Testing Matters

In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

$$P(\text{Making an error}) = \alpha$$

$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

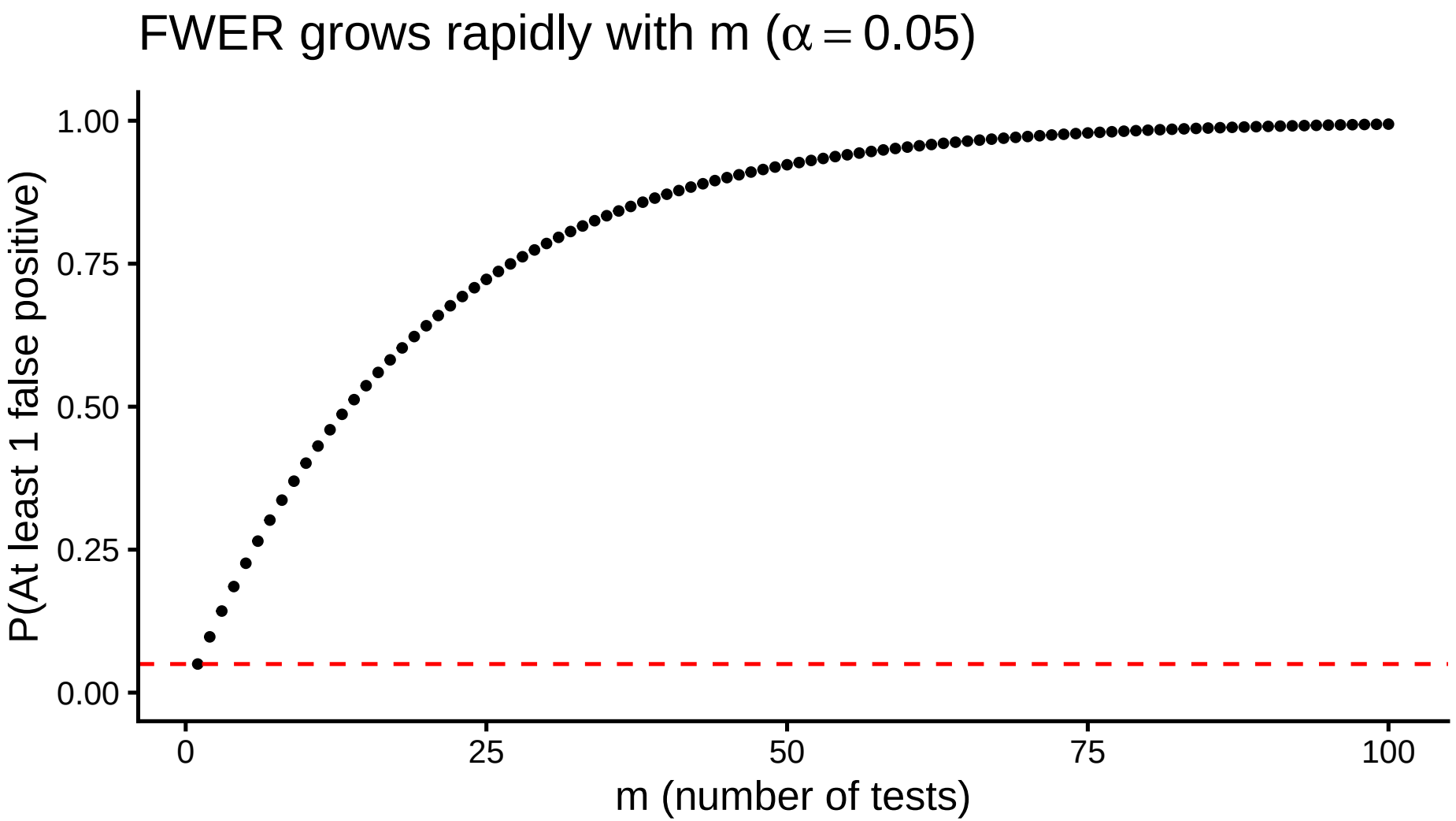
Why Multiple Testing Matters — Example

If we toss 10 fair coins 1024 times, in one occasion we are expected to have all 10 'head' coins

But if we ignore the fact that we tossed the coins 1024 times, simply looking at the all 10 'head' coins event, the p-value for having an unbiased coin is $p = 2 \times (1/2)^{10} = 2/1024$. Quite small — therefore we may reject H_0...

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

Probability of At Least 1 False Positive



The p-value

What is it?

- The chance of getting a test statistic value as extreme or more extreme than the one you got under the null hypothesis (H_0)
- ...for **one** test...

What do we need to do when we have many tests?

- Adjust our expectations accordingly!
- OK, but why?
- And more importantly...how?

What Does Correcting for Multiple Testing Mean?

- When people say "adjusting p-values for the number of hypothesis tests performed" what they mean is **controlling the Type I error rate**
- Very active area of statistics — many different methods have been described

Different Approaches to Control Type I Errors

- **Per comparison error rate** (PCER): $E(V)/m$
- **Per-family error rate** (PFER): $E(V)$
- **Family-wise error rate** (FWER): $P(V \geq 1)$
- **False discovery rate** (FDR): $E(V/R \mid R > 0) \cdot P(R > 0)$
- **Positive false discovery rate** (pFDR): $E(V/R \mid R > 0)$

Where V = number of false positives, R = total rejections, m = total tests

Family-wise Error Rate (FWER): Bonferroni Correction

$$\begin{aligned}\text{FWER} &= P(\text{falsely reject } \textbf{at least one} \text{ null hypothesis}) \\ &= 1 - (1 - \alpha_{\text{one-test}})^m \quad \text{for } m \text{ tests}\end{aligned}$$

If we want FWER $\leq \alpha$, then:

$$p_i \leq \frac{\alpha}{m} \qquad p_{\text{adj.}} = \min(p \times m, 1)$$

Simple, problem solved... but...

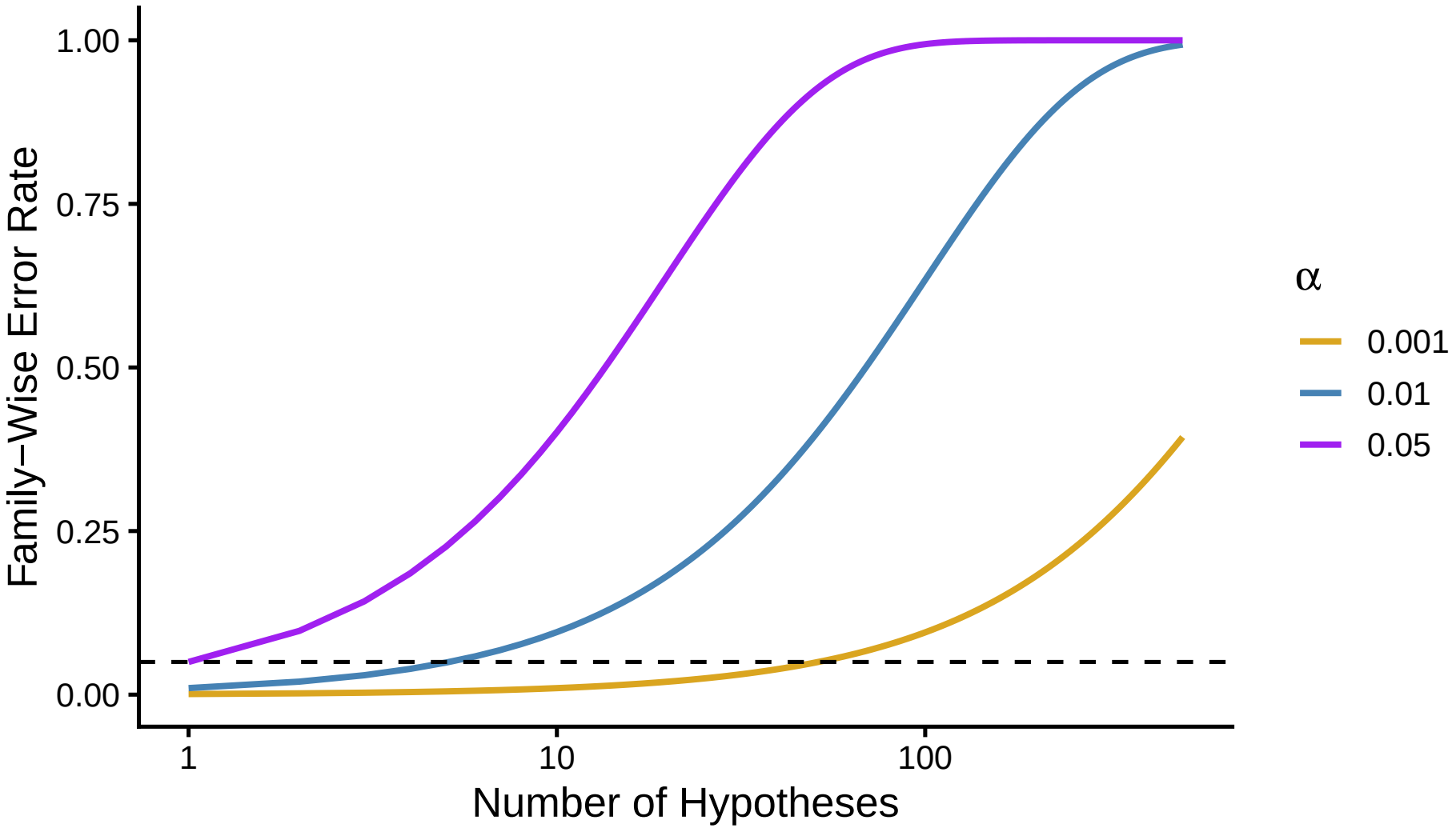
Bonferroni Correction — Example

$p_{\text{adj.}} = \min(p \times m, 1)$

- For instance, to control the FWER at level 0.05 while testing $m = 100$ null hypotheses,
- the Bonferroni procedure requires us to control the Type I error for each null hypothesis at level $0.05/100 = 0.0005$
- i.e., to reject all null hypotheses for which the p-value is below 0.0005

The Bonferroni correction is very conservative = false negatives (type II errors)...

Bonferroni Correction by m



Philosophical Objections to Bonferroni Corrections

"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" — Perneger (1998)

- Counter-intuitive: interpretation of finding depends on the number of other tests performed
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist

Who Cares About Not Making ANY Type I Errors?

- FWER is appropriate when you want to guard against **ANY** false positives
- However, in many cases (particularly in genomics) we can live with a certain number of false positives
- In these cases, the more relevant quantity to control is the **false discovery rate (FDR)**

Estimating Error Rates

	Declared non-significant	Declared significant	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
	$m - R$	R	m

FWER $= P(V \geq 1)$ — probability of at least one Type I Error

FDR $= E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)$

Finding Another Way — False Discovery Rate

- When m is large, trying to prevent **any** false positives (as in **FWER control**) is simply **too stringent**
- Instead, we make sure that the ratio of false positives **to total positives** is sufficiently low
  - FDR $= E\left(\frac{V}{R}\right) = \frac{FP}{FP+TP}$
  - Instead of keeping the chance of one False Positive low, we will now accept that **X% of our total positive values** are false

Benjamini-Hochberg Method

1. Pick q which is the desired **False Discovery Rate (FDR)**
2. Rank all p-values from smallest to largest: $p_1, p_2, \ldots, p_m$
3. Find **the largest** k and p-value p_k such that:

$$p_k \leq \frac{i}{m} \times q$$

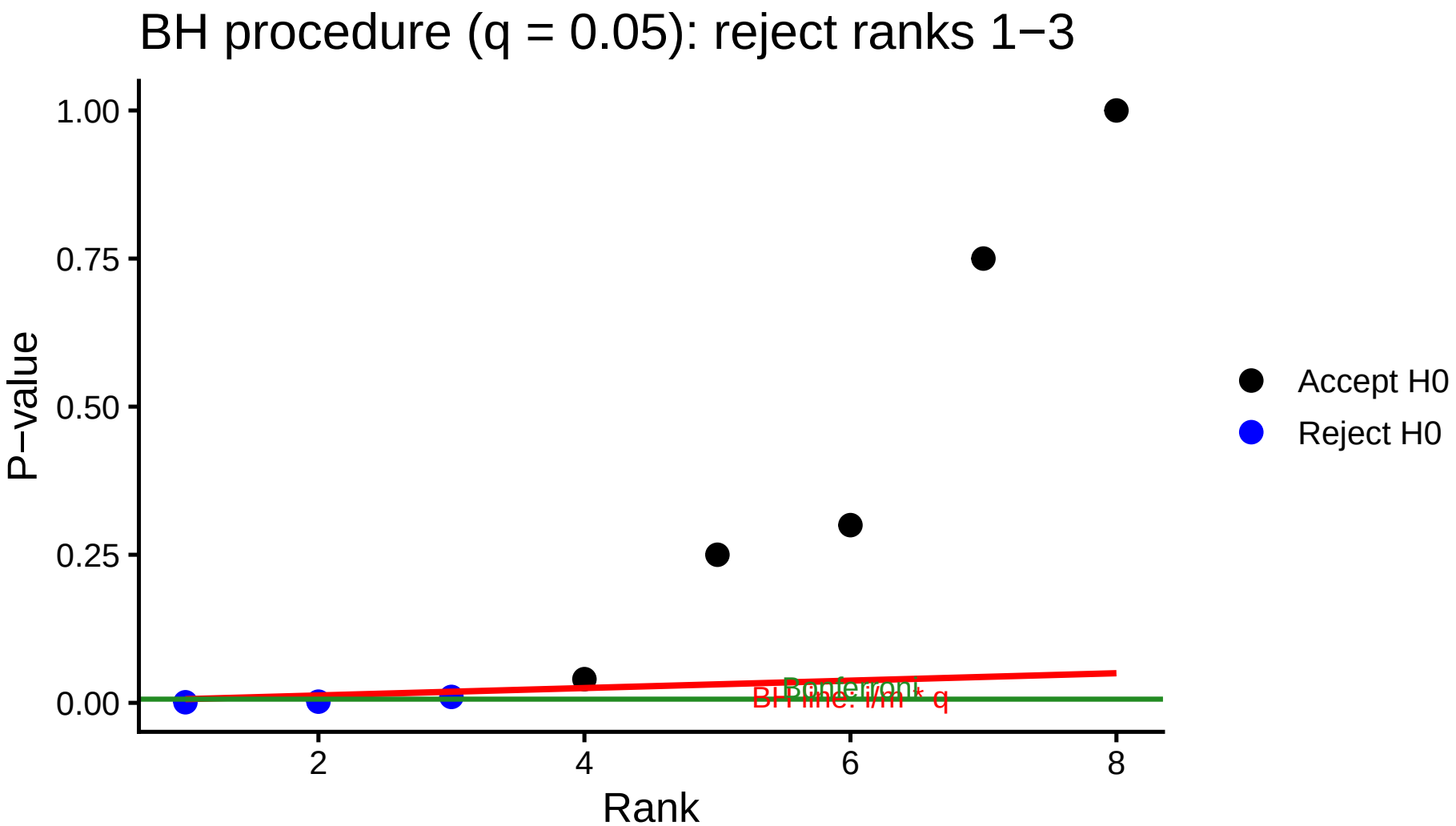
where $i =$ rank of p-value, $m =$ total number of p-values

4. **Reject** null for all values: $H_i = 1, 2, \ldots, k$

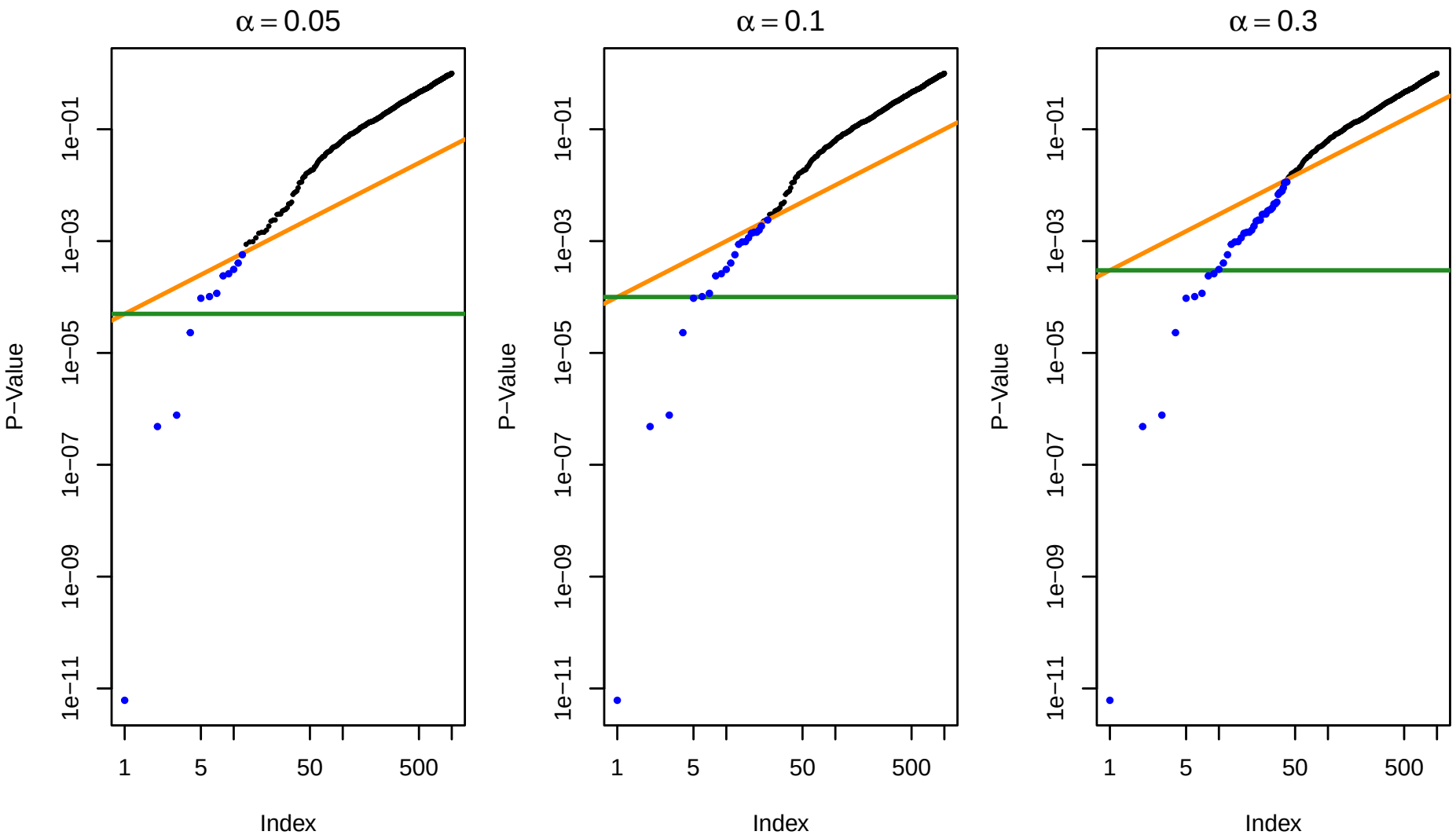
Benjamini-Hochberg — Choosing q

- How to pick appropriate q?
- It will be very arbitrary, or context dependent
- If follow-up experiments are costly and labor intensive, you may want to control FDR small, like 5%
- Whereas if your follow-up is easy and you don't want to miss interesting things, you may be happy with the FDR of 30%

Benjamini-Hochberg — Example



Benjamini-Hochberg — Graphically



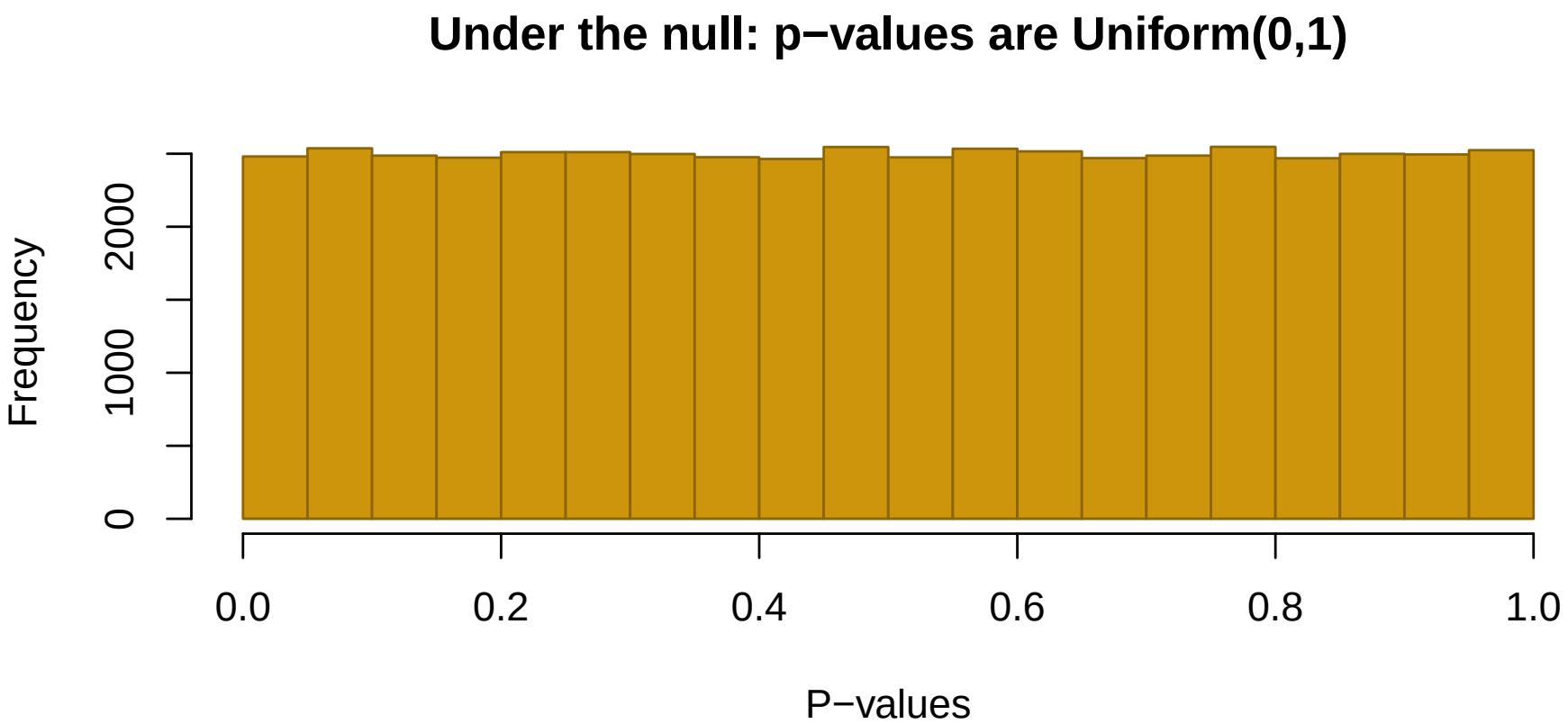
Storey's Positive FDR (pFDR)

$$\text{BH: FDR} = E\left[\frac{V}{R} \mid R > 0\right] P(R > 0)$$

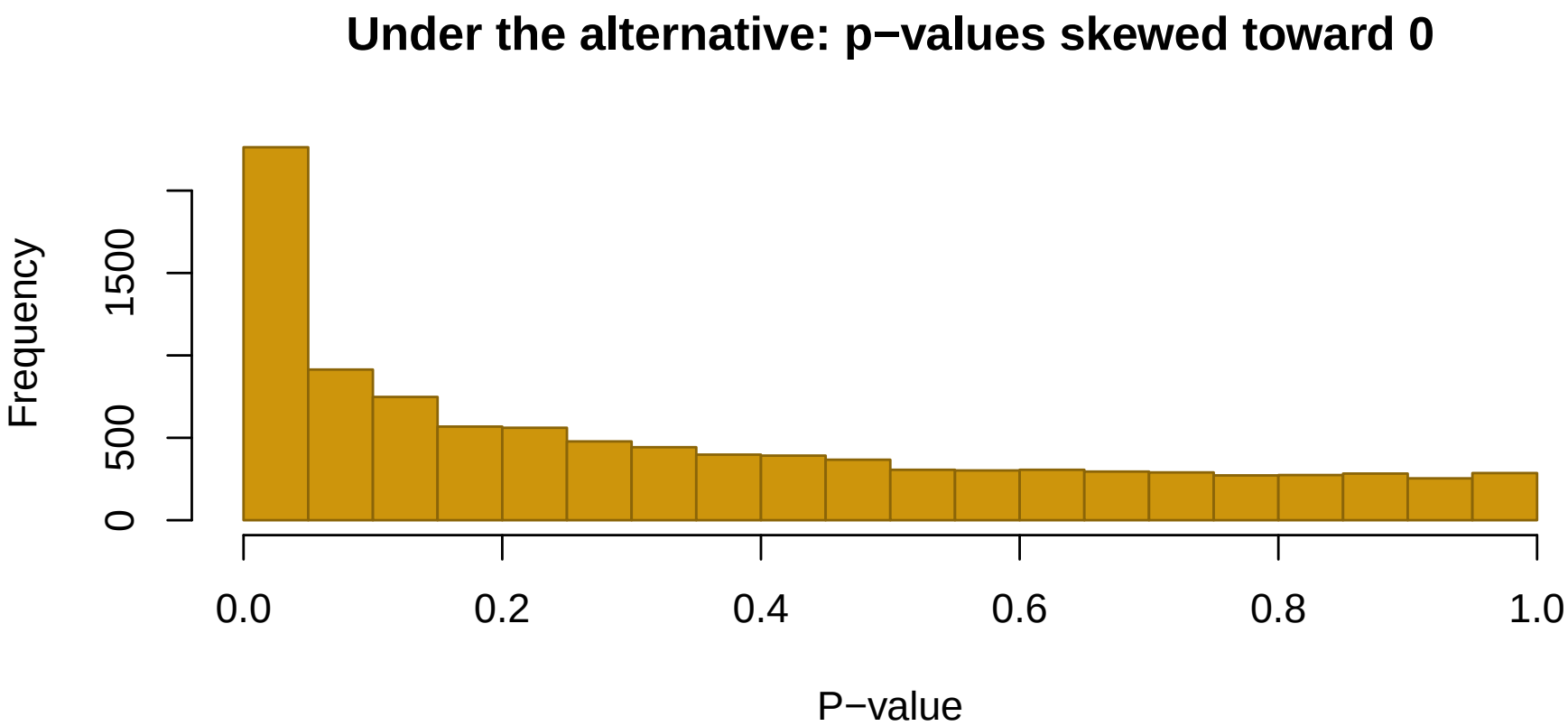
$$\text{Storey: pFDR} = E\left[\frac{V}{R} \mid R > 0\right]$$

- Since $P(R > 0) \approx 1$ in most genomics experiments, FDR and pFDR are very similar
- Omitting $P(R > 0)$ facilitated development of a measure of significance in terms of the FDR for each hypothesis

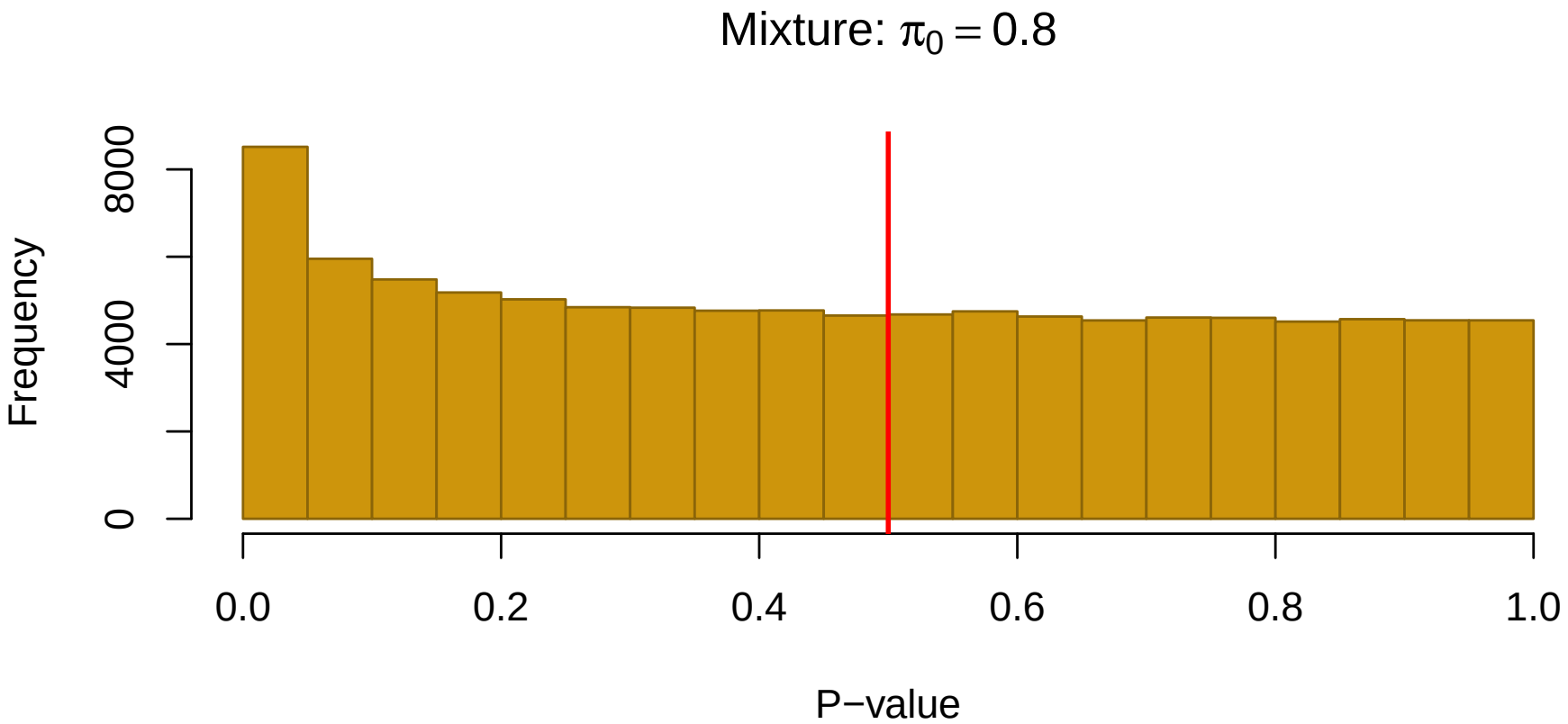
Estimating The Proportion of Truly Null Tests



Estimating The Proportion of Truly Null Tests



Estimating The Proportion of Truly Null Tests



Definition of π_0

$\hat{\pi}_0$ is the proportion of truly null tests:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, 2, \ldots, m\}}{m(1-\lambda)}$$

$1 - \hat{\pi}_0$ is the proportion of truly alternative tests (very useful!)

The p-value... Again

- Implicit in all multiple testing procedures is the assumption that the distribution of p-values is "correct"
- This assumption often is not valid for genomics data where p-values are obtained by asymptotic theory
- Thus, **resampling methods** are often used to calculate p-values

Resampling Approaches for p-values

- When theoretical distribution of test statistics does not exist...
- When assumptions for statistical tests do not hold...

- You can empirically generate the distribution of test statistic by resampling or repeated sampling of datasets
- If you want to know the distribution of test statistic under null, **generating null with permutation** is very useful

Permutations

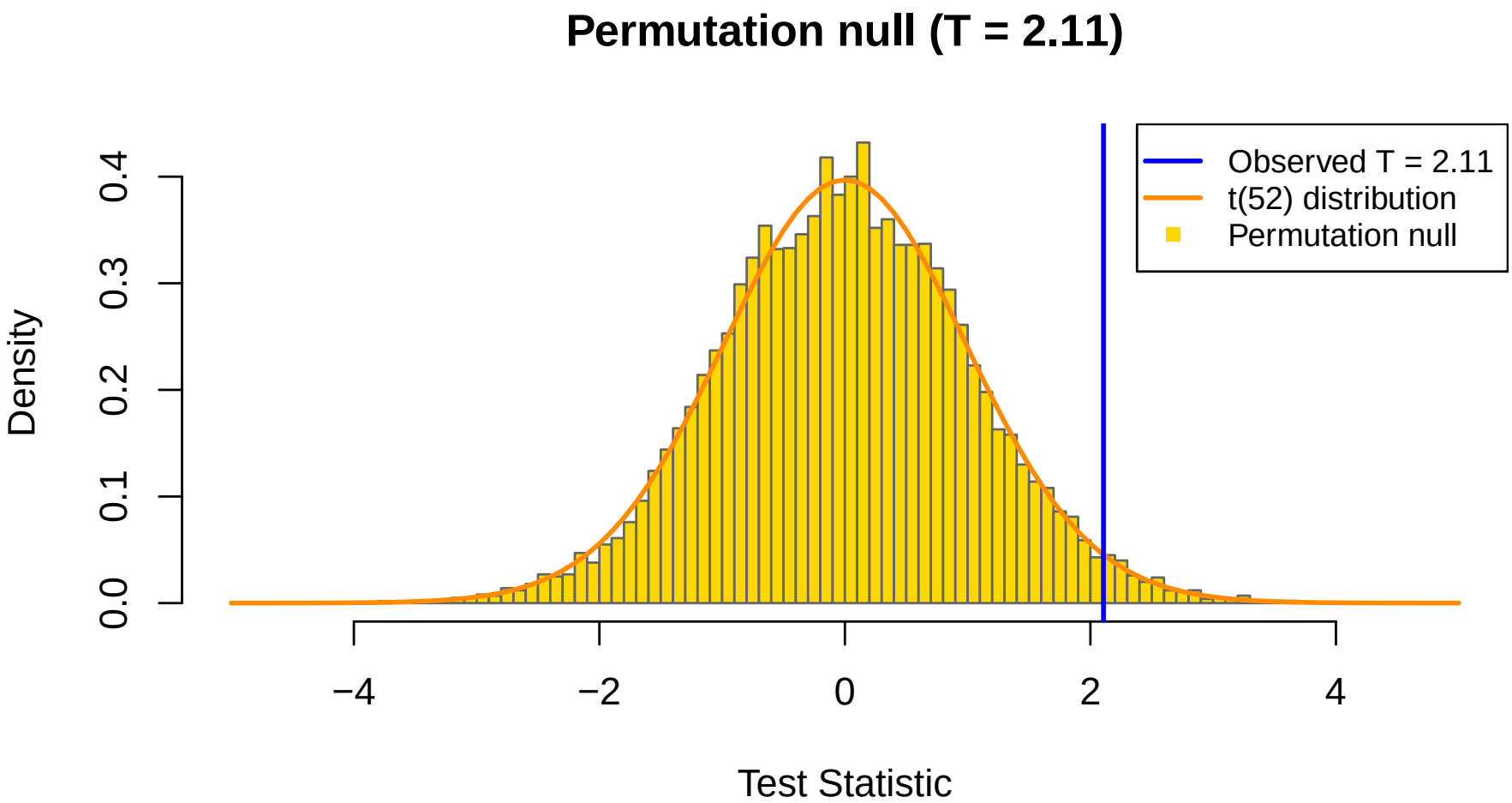
1. Analyze the problem: think carefully about the null and alternative hypotheses
2. Choose a test statistic
3. Calculate the test statistic for the original labeling of the observations
4. **Permute the labels and recalculate the test statistic**
   - Do all permutations: Exact Test
   - Randomly selected subset: Monte Carlo Test
5. Calculate p-value by comparing where the observed test statistic value lies in the permuted distribution of test statistics

Permutation Approach for Two-sample t-test

Algorithm (Re-Sampling p-Value for a Two-Sample t-Test):

1. Compute T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$
2. For $b = 1, \ldots, B$ (e.g., $B = 10{,}000$):
   a. Permute the $n_X + n_Y$ observations at random
   b. Compute T^{*b} on the permuted data
3. The p-value is: $\frac{\sum_{b=1}^{B} \mathbf{1}_{(|T^{*b}| \geq |T|)}}{B}$

Empirical Distribution of t-score Under Null



Practical Points in Permutations

How to permute?

- You need to create **null (no association)**, by breaking the relationship between X and Y, **while keeping all the structure the same** as in the real dataset

How many permutations?

- With 1000 permutations the smallest possible p-value is 0.001
- A useful strategy is to start with 1000 permutations and continue to larger numbers only if p is small enough to be interesting, e.g., $p < 0.1$

Section 3

Summary

Summary: Power Analysis

Power Analysis

- Depends on several factors: effect size, sample variance, α/β, the test you perform
- Will tell us what effect size we can detect given the test performed

To increase power:

- Increase n
- Reduce variance
- Set a more liberal threshold (?)

Summary: Multiple Hypothesis Testing

Multiple hypothesis testing

- Multiple tests $\rightarrow$ FWER $= \alpha \times m$
- **FWER** control (e.g., Bonferroni) controls the chance of getting a single FP
- **FDR** (e.g., BH) guarantees an upper bound on FDR across all tests
  - $FDR = \frac{FP}{FP+TP}$