Lecture 10: Multiple Testing

Goals

- Define the multiple testing problem and related concepts

- Methods for addressing multiple testing (FWER and FDR)

- Correcting for multiple testing in R

Type I and II Errors

Actual Situation “Truth”

Decision	H_0 True	H_0 False
Do Not Reject H_0	Correct Decision 1 - α	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision 1 - β

$$\alpha = P(Type\ I\ \ Error) \quad \beta = P(Type\ II\ \ Error)$$

power (검정력) = $P(\text{reject } H_0 \mid H_0 \text{ False})$

Why Multiple Testing Matters

Genomics = Lots of Data = Lots of Hypothesis Tests

A typical microarray experiment might result in performing 10000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect ***500*** genes to be deemed "significant" by chance.

Why Multiple Testing Matters

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

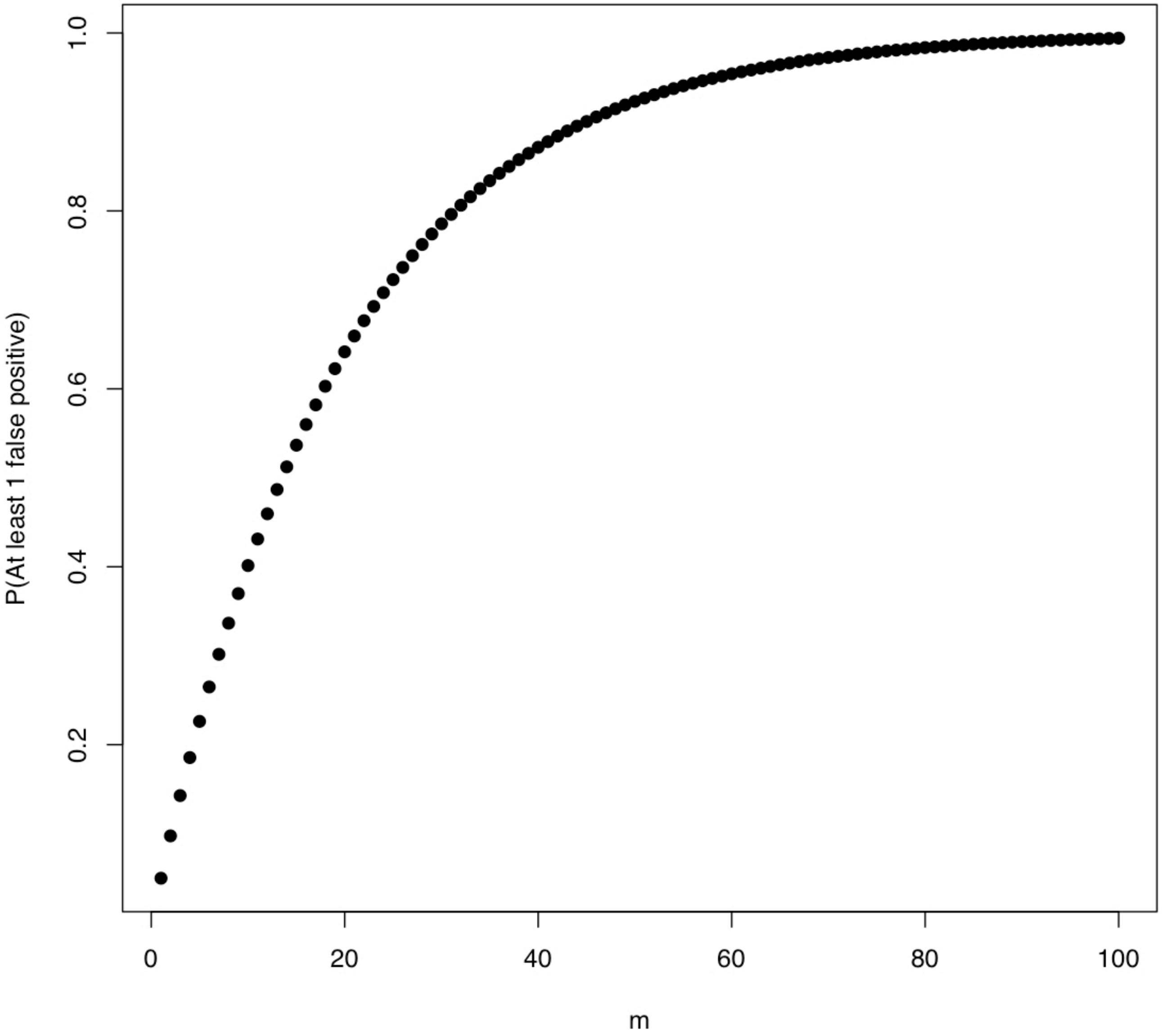
P(Making an error) = α 0.05

P(Not making an error) = 1 - α 0.95

P(Not making an error in m tests) = $(1 - \alpha)^m$ 0.95^{1000}

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

Probability of At Least 1 False Positive



Counting Errors

Assume we are testing H^1, H^2, …, H^m

m_0 = # of true hypotheses R = # of rejected hypotheses

	Null True	Alternative True	Total
Not Called Significant	U	T	$m - R$
Called Significant	V	S	R
	m_0	$m-m_0$	m

V = # Type I errors [false positives]

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
- Although these varied approaches have the same goal, they go about it in fundamentally different ways

Different Approaches To Control Type I Errors

- **Per comparison error rate** (PCER): the expected value of the number of Type I errors over the number of hypotheses,

  PCER = E(V)/m

- **Per-family error rate** (PFER): the expected number of Type I errors,

  PFE = E(V).

- **Family-wise error rate**: the probability of at least one type I error

  FEWR = P(V ≥ 1)

- **False discovery rate** (FDR) is the expected proportion of Type I errors among the rejected hypotheses

  FDR = E(V/R | R>0)P(R>0)

  유의하다고 판정된(귀무가설 기각) 유전자 중
  위양성 유전자의 분율

- **Positive false discovery** rate (pFDR): the rate that discoveries are false

  pFDR = E(V/R | R > 0)

Digression: p-values

여담

- 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 calculate p-values

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 distributed of test statistics

Example: What to Permute?

- Gene expression matrix of m genes measured in 4 cases and 4 controls

Gene	Case 1	Case 2	Case 3	Case 4	Control 1	Control 2	Control 3	Control 4
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}
3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	X_{38}
4	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{46}	X_{47}	X_{48}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	X_{m1}	X_{m2}	X_{m3}	X_{m4}	X_{m5}	X_{m6}	X_{m7}	X_{m8}

Back To Multiple Testing: FWER

- Many procedures have been developed to control the Family Wise Error Rate (the probability of at least one type I error):

$$P(V \geq 1)$$

- Two general types of FWER corrections:

1. **Single step**: equivalent adjustments made to each p-value
2. **Sequential**: adaptive adjustment made to each p-value

① Single Step Approach: Bonferroni

- Very simple method for ensuring that the overall Type I error rate of α is maintained when performing m independent hypothesis tests

- Rejects any hypothesis with p-value $\leq \alpha/m$:

$$\tilde{p}_j = \min[mp_j,\ 1]$$

Bonferroni corrected p-value

- For example, if we want to have an experiment wide Type I error rate of 0.05 when we perform 10,000 hypothesis tests, we'd need a p-value of 0.05/10000 = 5×10^{-6} to declare significance

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
- simple, but conservative

② FWER: Sequential Adjustments

- Simplest sequential method is Holm's Method
  - Order the unadjusted *p*-values such that $p_1 \leq p_2 \leq \ldots \leq p_m$
  - For control of the FWER at level α, the step-down Holm adjusted p-values are

$$\tilde{p}_j = \min[(m - j + 1) \bullet p_j,\ 1]$$

  - The point here is that we don't multiply every p_i by the same factor m

- For example, when m = 10000:

$$\tilde{p}_1 = 10000 \bullet p_1,\ \tilde{p}_2 = 9999 \bullet p_2, \ldots, \tilde{p}_m = 1 \bullet p_m$$

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)

False Discovery Rate

	Null True	Alternative True	Total
Not Called Significant	U	T	$m - R$
Called Significant	V	S	R
	m_0	$m-m_0$	m

V = # Type I errors [false positives]

- False discovery rate (FDR) is designed to control the proportion of false positives ***among the set of rejected hypotheses*** (R)

FDR vs FPR

	Null True	Alternative True	Total
Not Called Significant	*U*	*T*	*m - R*
Called Significant	*V*	*S*	*R*
	m_0	$m-m_0$	*m*

$$FDR = \frac{V \text{ false positive}}{R \text{ total positive}} \qquad FPR = \frac{V}{m_0}$$

엄밀히 말하면, $FDR = E[Q] = E[\frac{V}{R}]$

where $\frac{V}{R}$ is defined to be 0 when $R=0$.

Benjamini and Hochberg FDR

- To control FDR at level δ:

1. Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$

2. Then find the test with the highest rank, j, for which the p value, p_j, is less than or equal to $(j/m) \times \delta$

3. Declare the tests of rank 1, 2, …, j as significant

$$p(j) \leq \delta \frac{j}{m}$$

B&H FDR Example

Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008 <	0.005	1
2	0.009 <	0.010	1
3	0.165 >	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Storey’s positive FDR (pFDR)

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

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

- Since P(R > 0) is ~ 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

What’s a q-value?

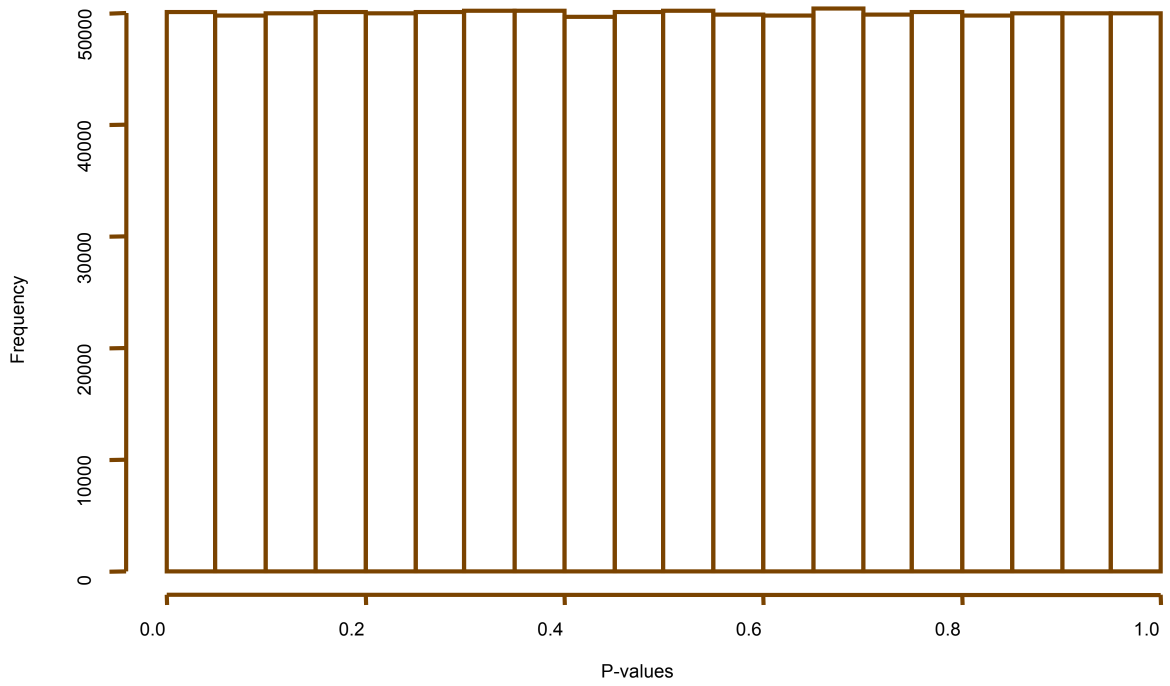
- q-value is defined as the minimum FDR that can be attained when calling that “feature” significant (i.e., expected proportion of false positives incurred when calling that feature significant)

- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)

- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

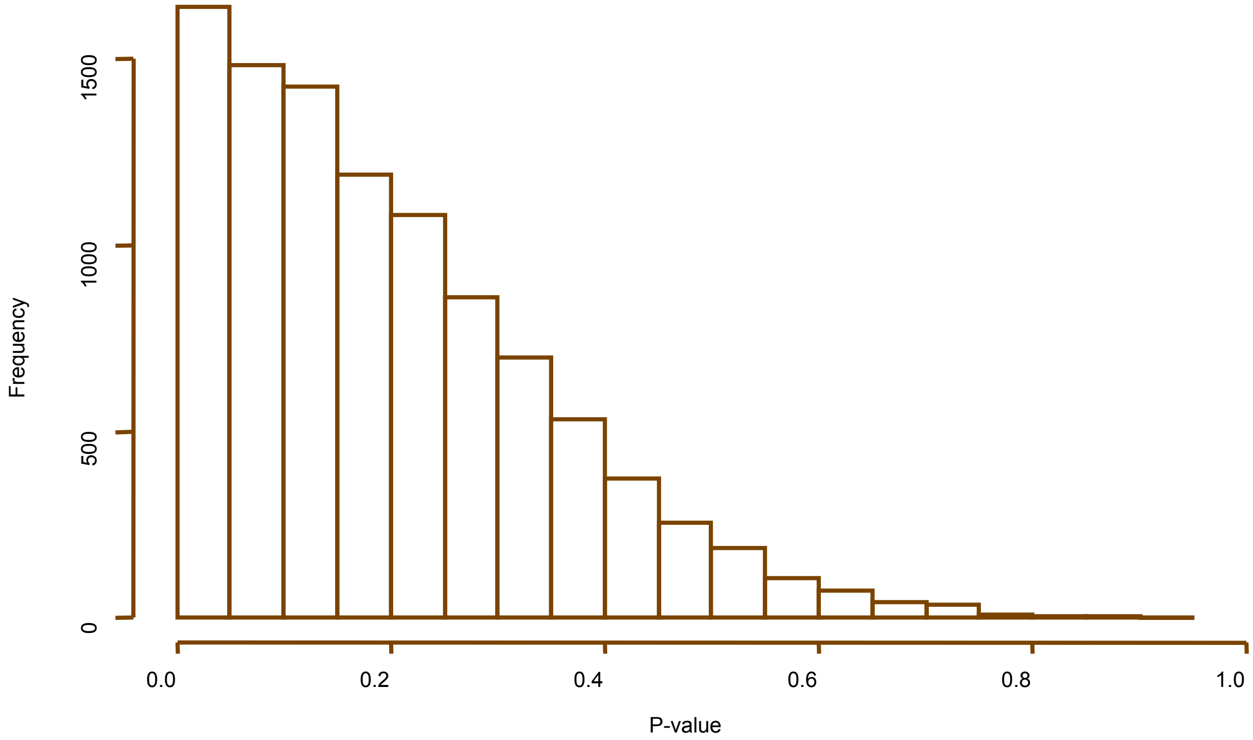
Estimating The Proportion of Truly Null Tests

- Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1



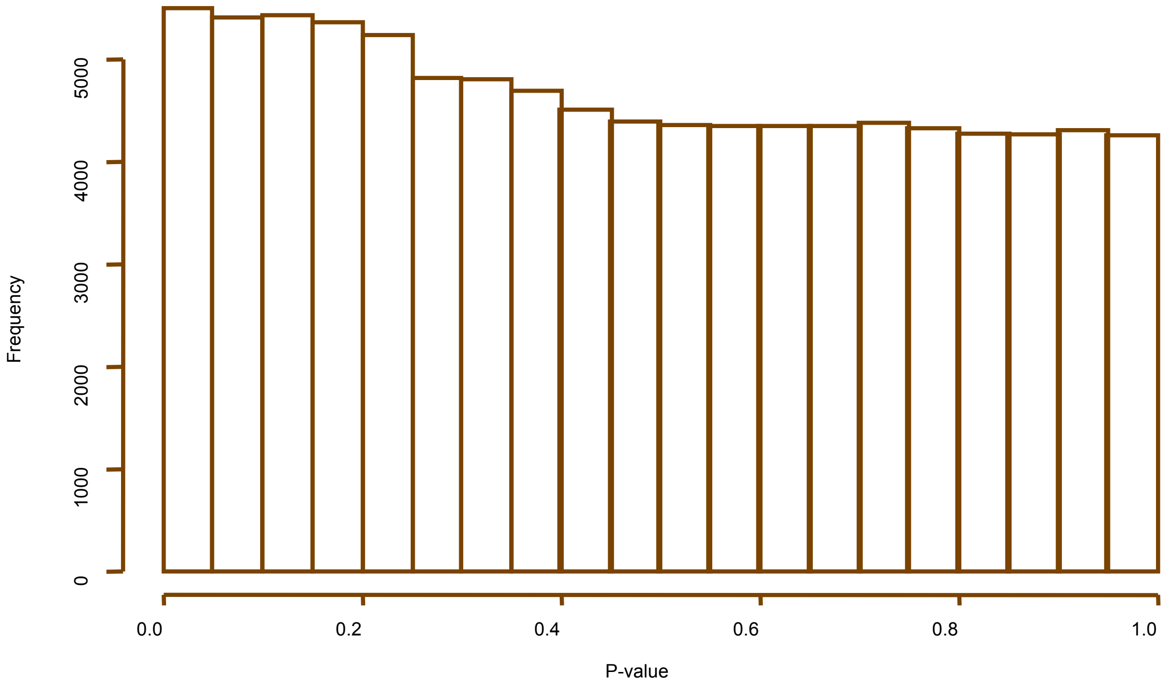
Estimating The Proportion of Truly Null Tests

- Under the alternative hypothesis p-values are skewed towards 0



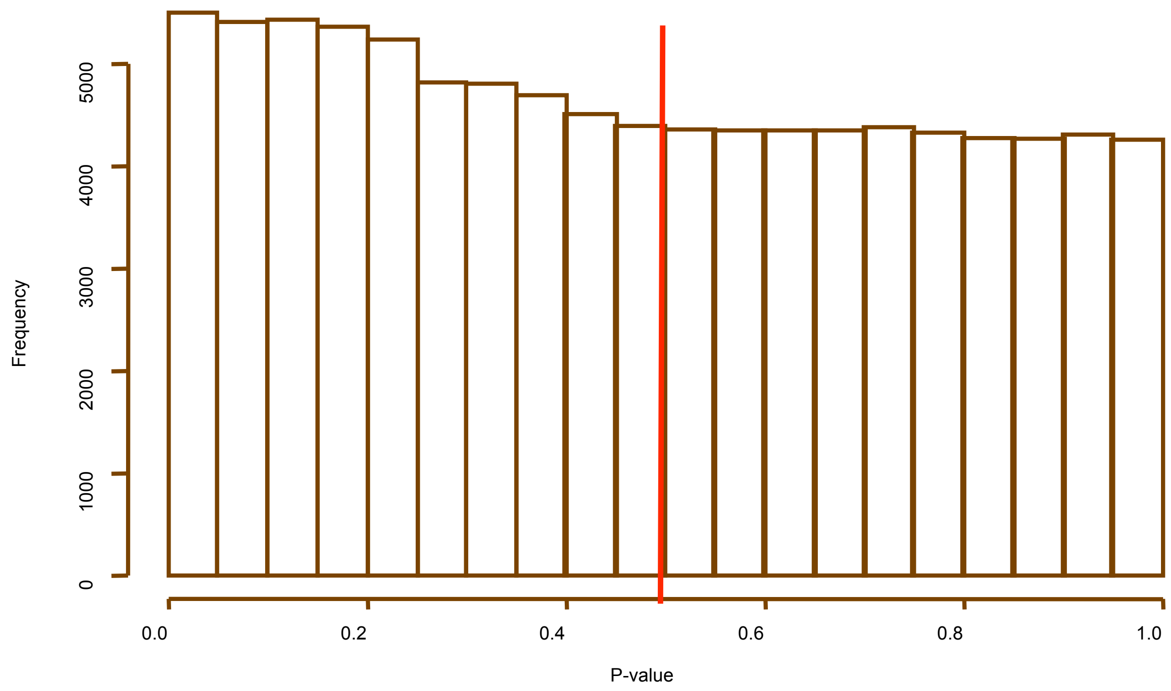
Estimating The Proportion of Truly Null Tests

- Combined distribution is a mixture of p-values from the null and alternative hypotheses



Estimating The Proportion of Truly Null Tests

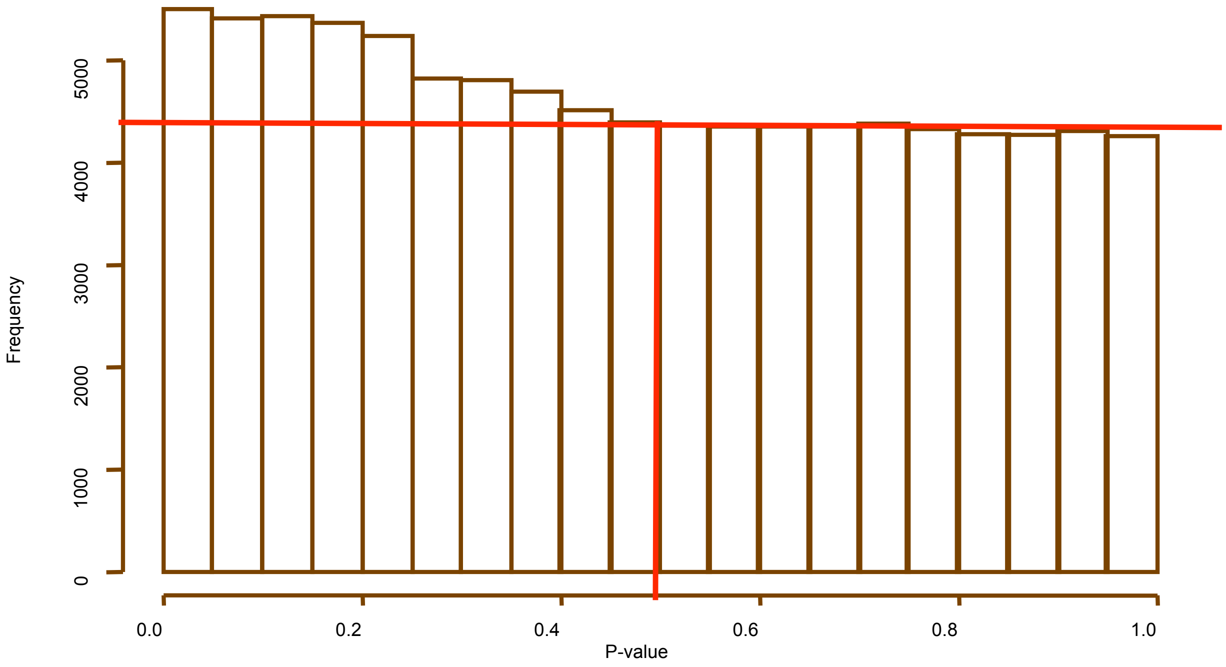
- For p-values greater than say 0.5, we can assume they mostly represent observations from the null hypothesis



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!)