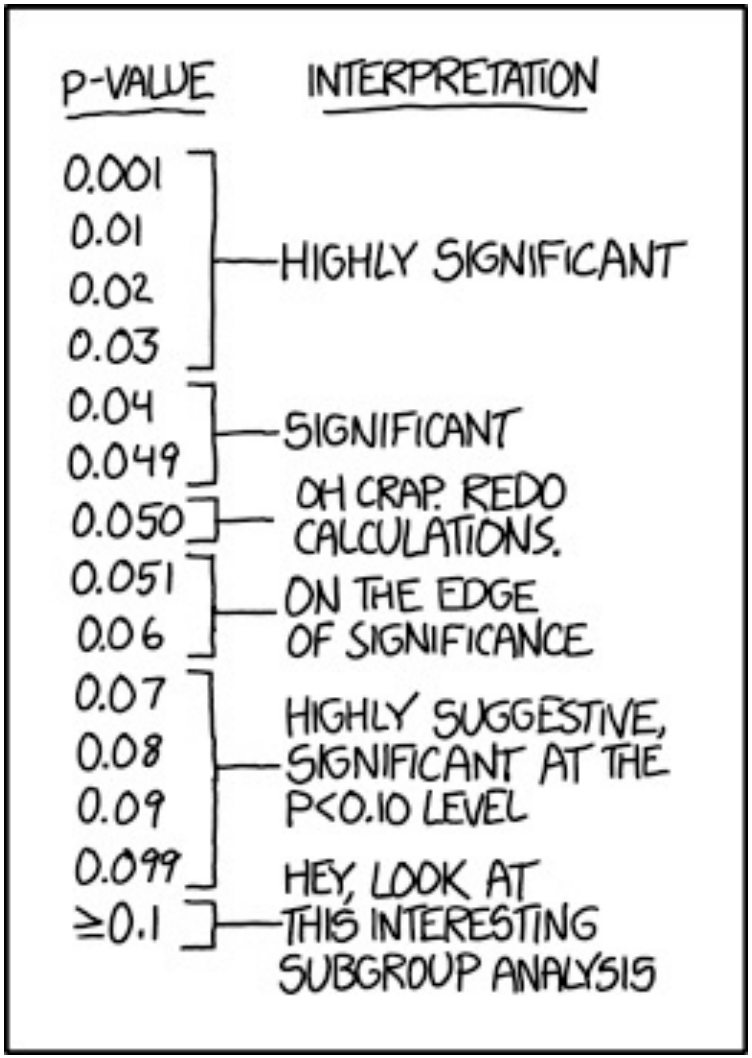
Multiple Hypothesis Testing

BMI 206



https://xkcd.com/1478/

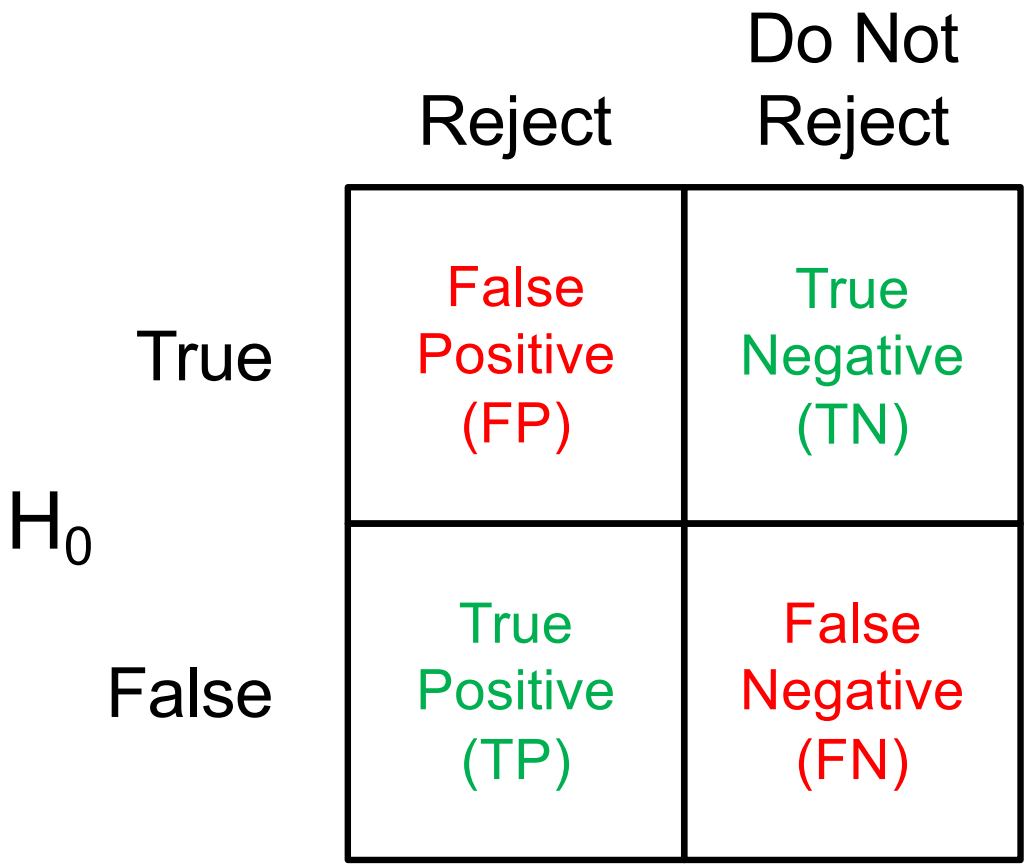
What is a p-value?

- **p-value:** the probability of obtaining a result at least as extreme as observed if H_0 is true.
  - Null hypothesis (H_0) is usually: chance/no effect

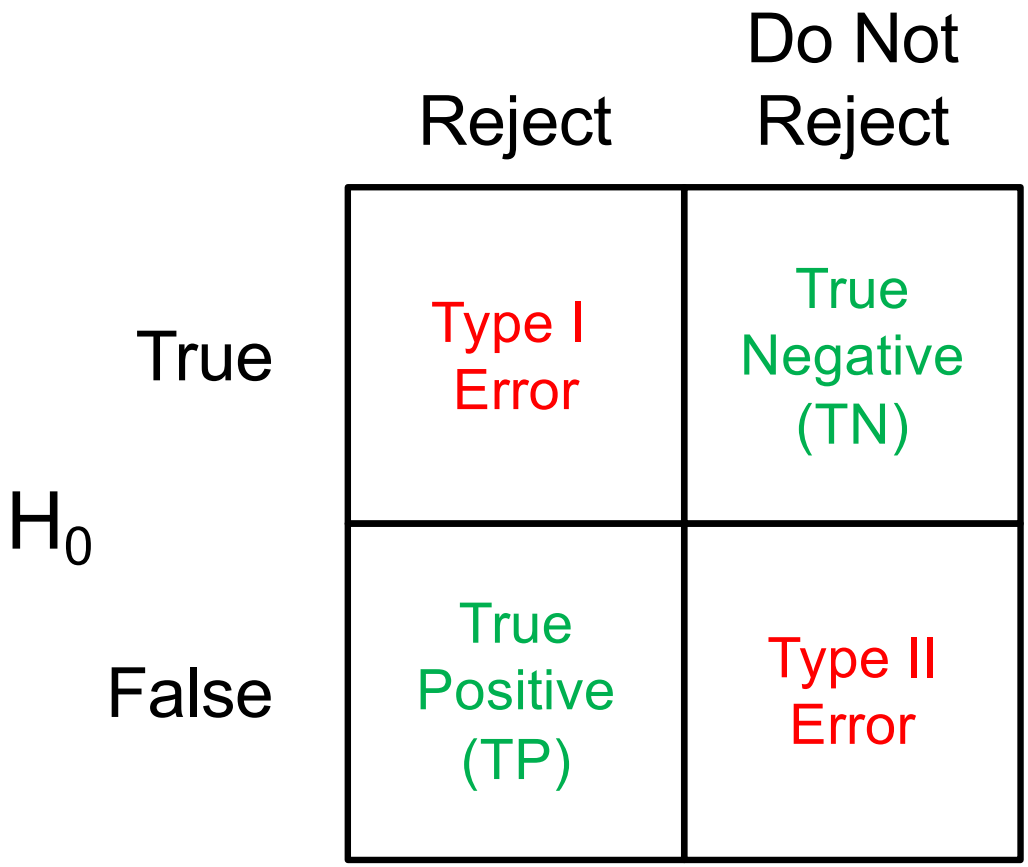
- $P < 0.05$ does not necessarily indicate a **meaningful** difference.

- $P > 0.05$ does not necessarily indicate no **meaningful** difference.

Outcomes of One Test



Type I vs. Type II error



Life Hack: the boy who cried wolf

1. Caused a **type I error**:
    - townspeople thought there was a wolf when there was not (False Positive)
2. Then caused a **type II error**:
    - townspeople thought there was no wolf when there was (False Negative)

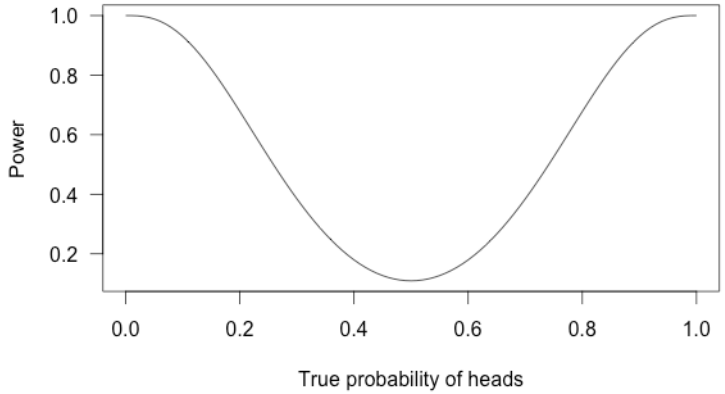
Statistical Power

- The **power** of a test is the probability of correctly rejecting a null hypothesis (1 – P(FN))

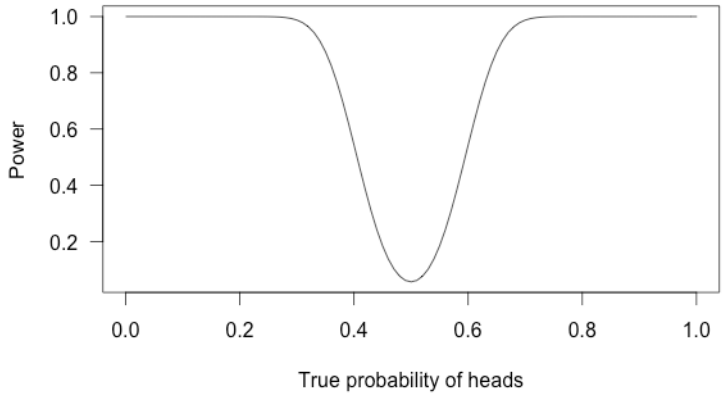
- Power varies based on the **effect size** and the **sample size**.

Statistical Power

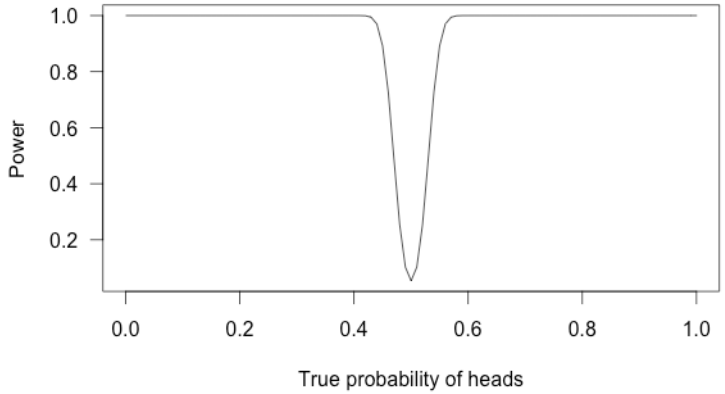
10 Flips



100 Flips

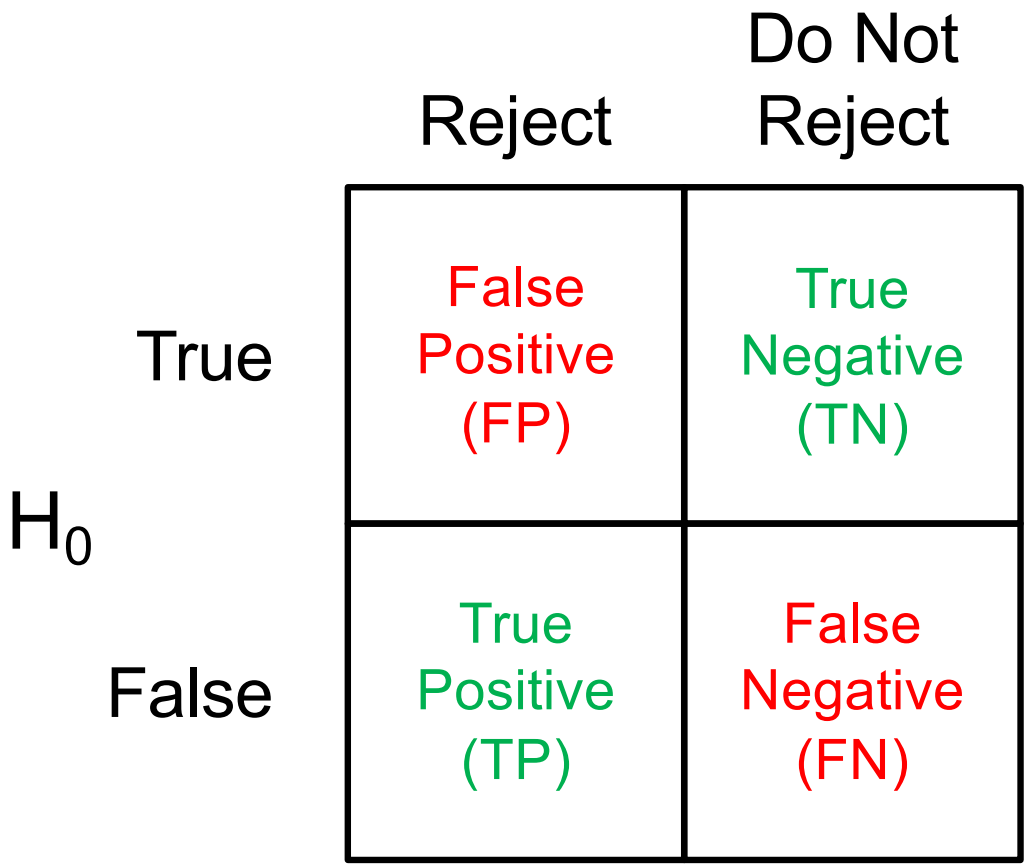


1000 Flips



- Power increases with sample size.
- Power increases with effect size.
- Many studies are underpowered.

Controlling Errors in a Single Test

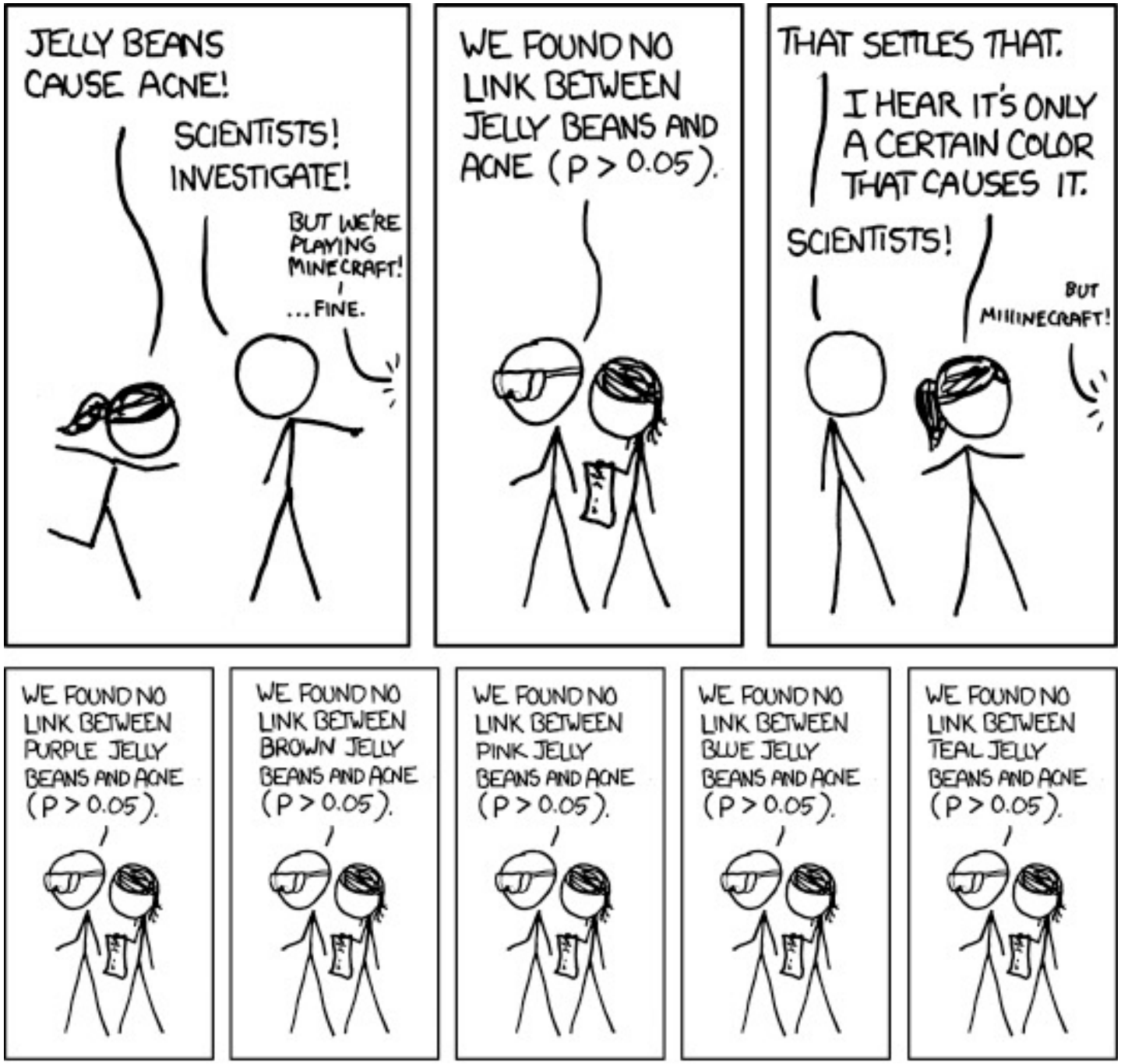


Significance Level (α) = P(FP)

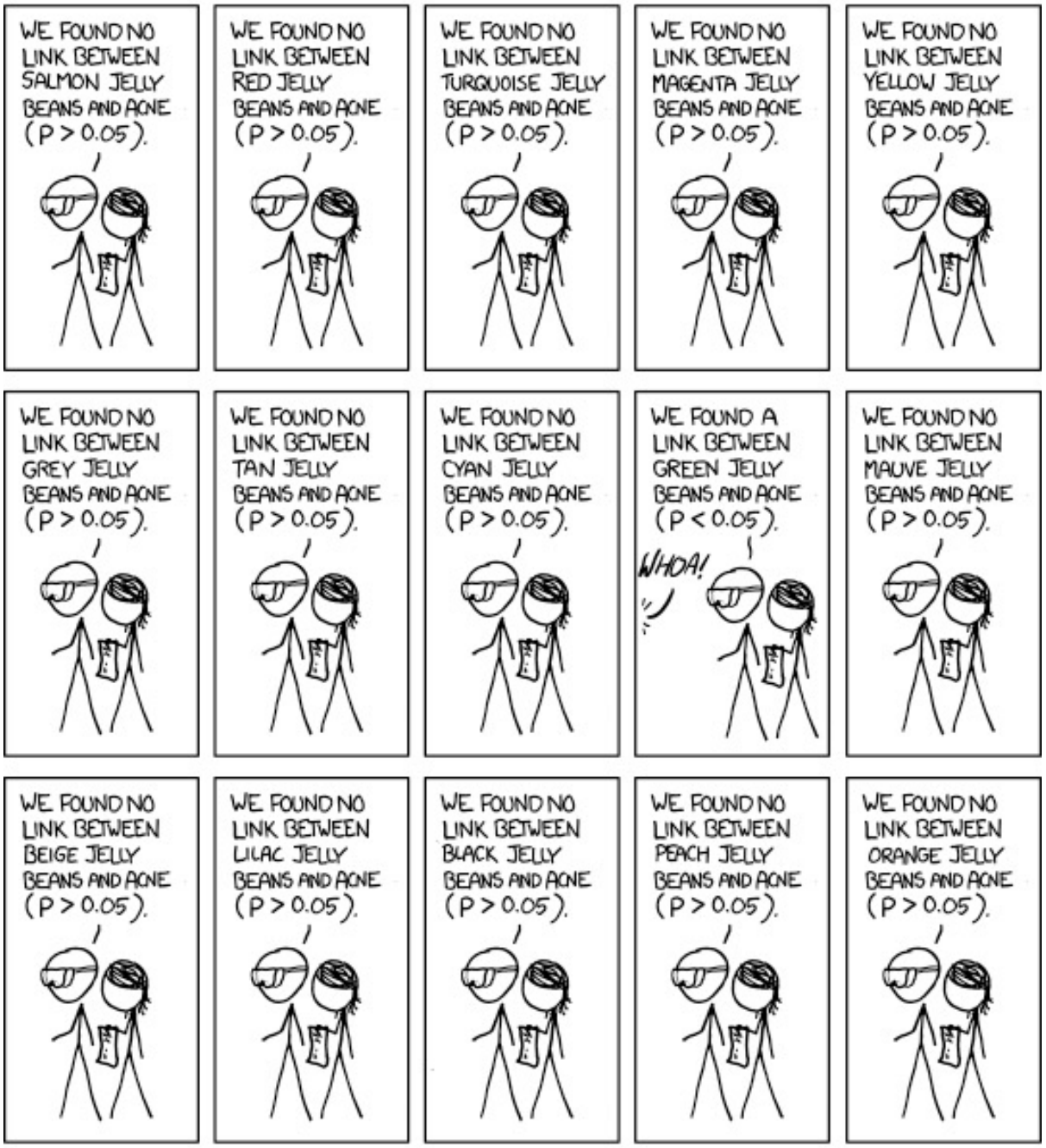
Power = 1 – P(FN) = 1 – β (H_0 false)

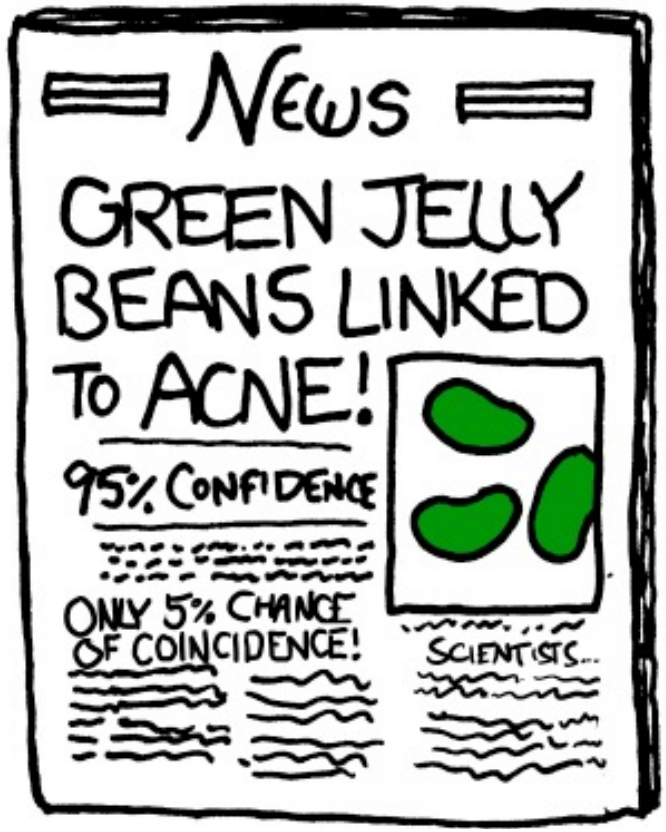
What happens when we test more than one hypothesis?

A motivational cartoon…



https://xkcd.com/882/

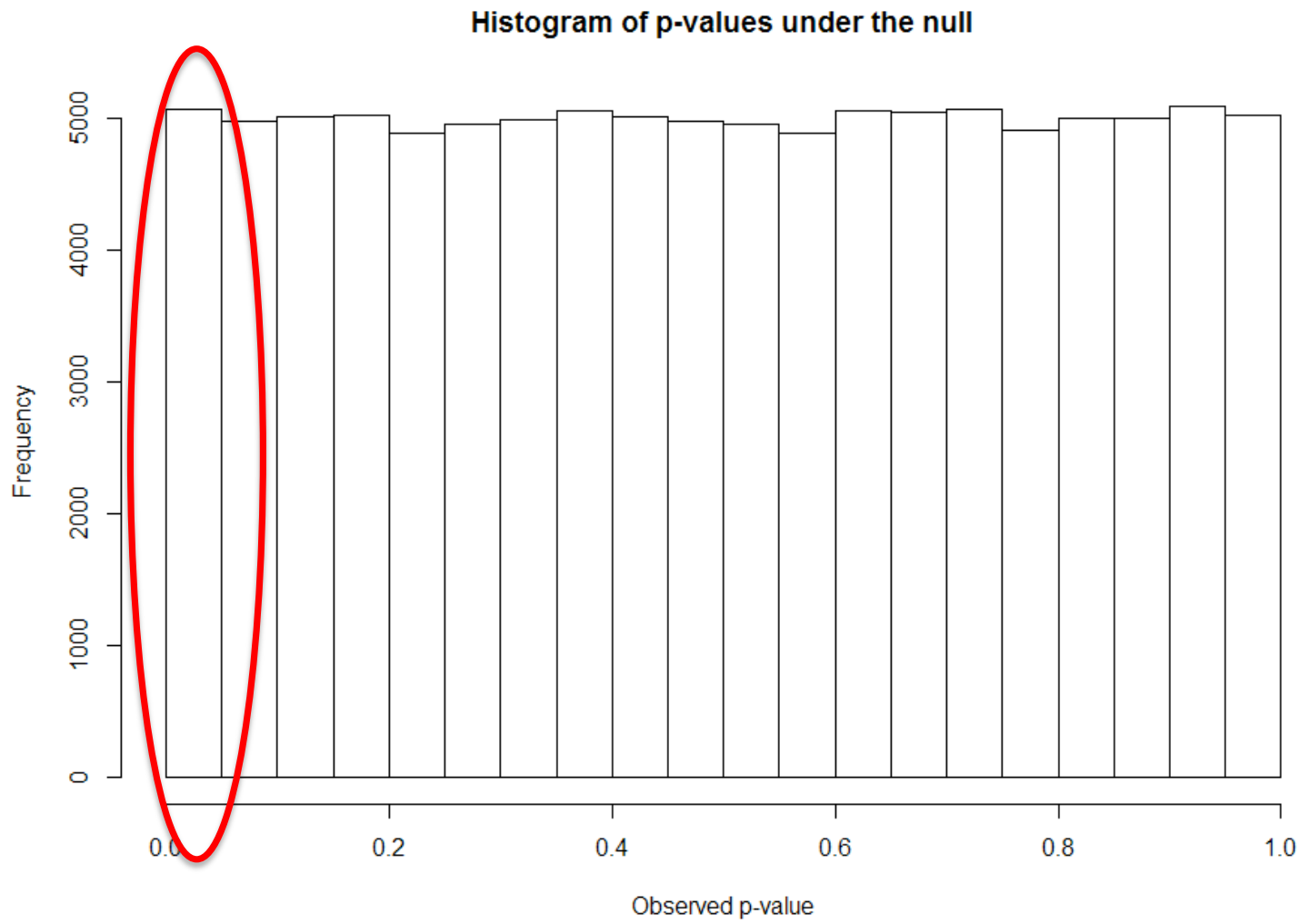




https://xkcd.com/882/

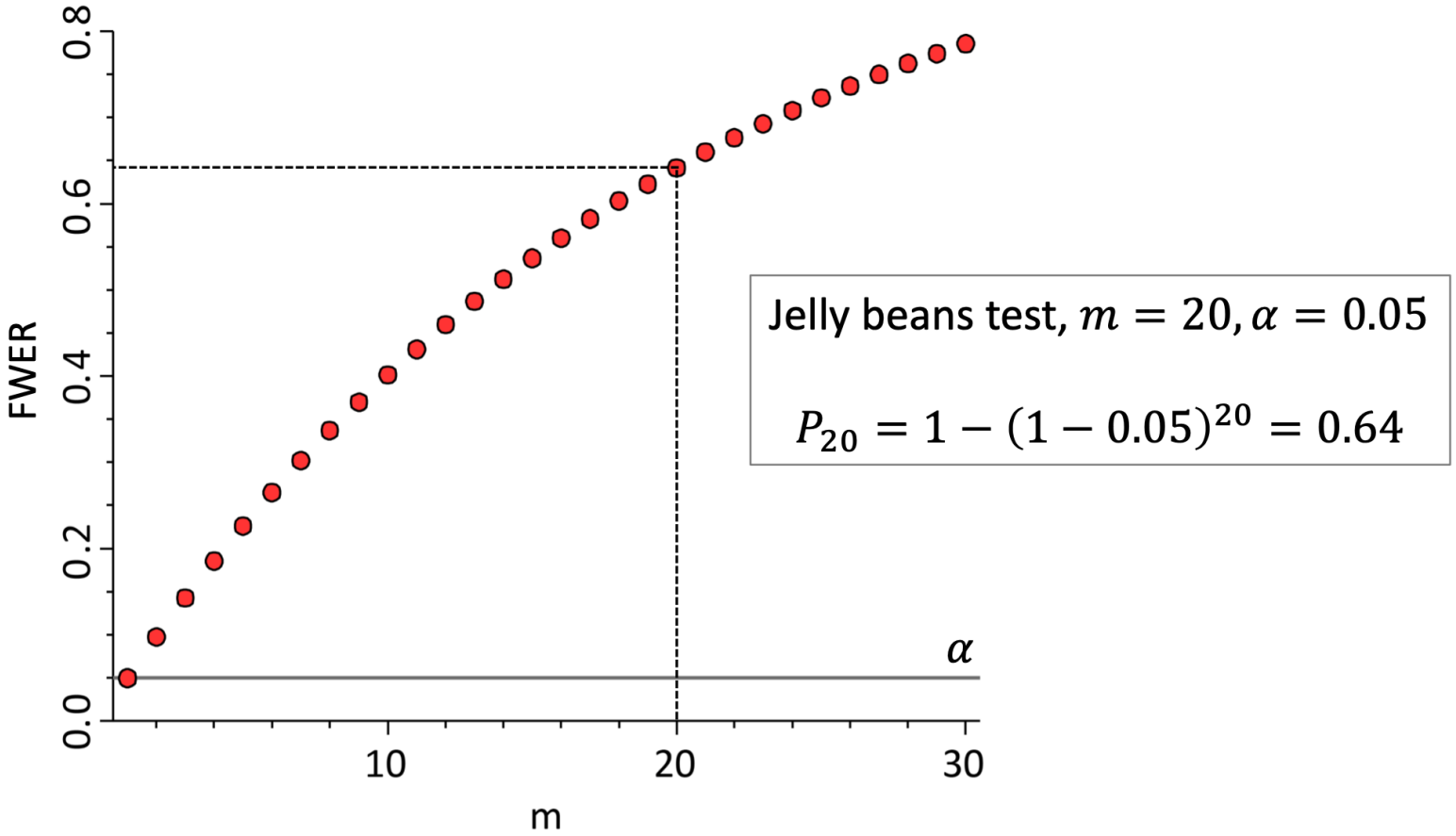
Why is testing multiple hypotheses a problem?

What is the distribution of p-values under the null?



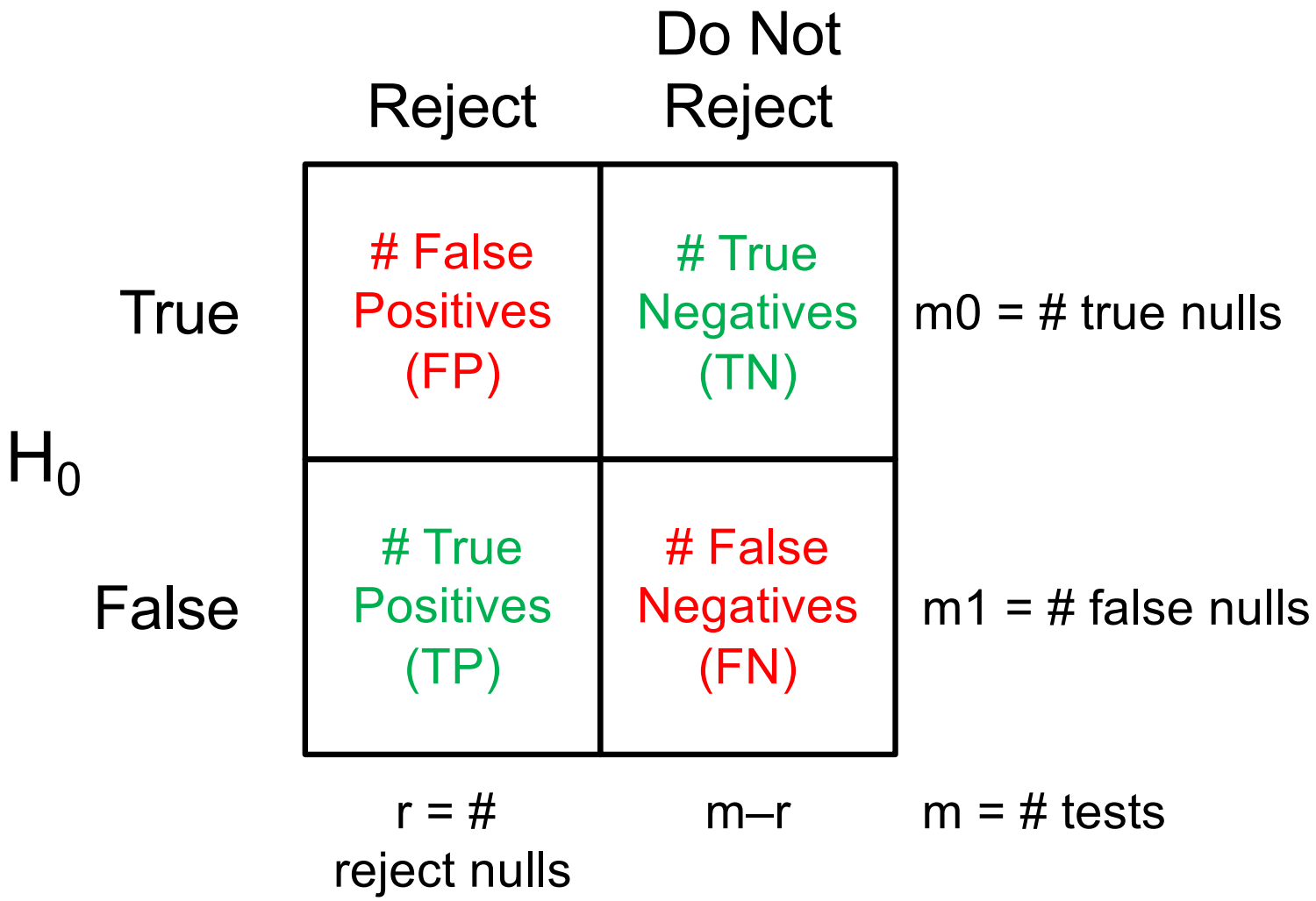
What is the chance under the null of at least one p-value < α in m ind. tests?

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



http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/p-values8.pdf

Outcomes of Many Tests



What can we do?

- In one test, α controls the family-wise error rate (**FWER**), the probability of at least one false positive :

$$\mathrm{P}(\mathrm{FP} > 0) \leq \alpha$$

- Over all m tests, this is:

$$\mathrm{P}(\#\mathrm{FP} > 0) \leq \alpha$$

Bonferroni Correction

To control FWER over **m** tests, adjust the p-value threshold (α) we use:

$$\alpha_{Bonferroni} = \alpha / m$$

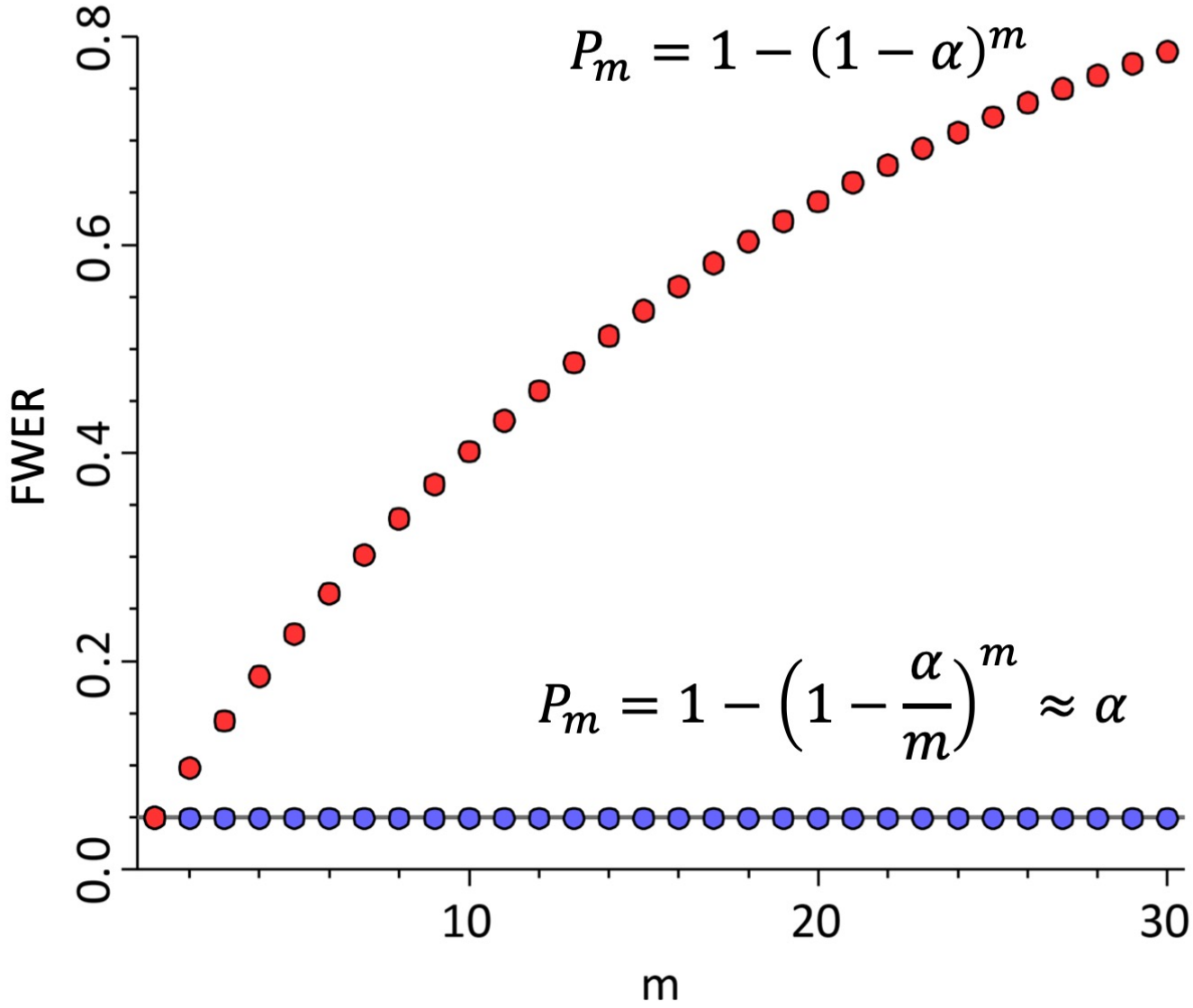
If α=.05 and 20 tests:

$$\alpha_{Bonferroni} = 0.05 / 20 = 0.0025$$

Or, equivalently, correct the p-values:

$$p_{Bonferroni} = p * 20$$

Bonferroni Correction Graph



http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/p-values8.pdf

Proof

- Let $p_1,\ldots,p_m$ be the p-values for all tests
- Let I_0 be the indices of all m_0 true null hypoth.
- We are interested in:

$$P(p_i \leq \frac{\alpha}{m}) \text{ for at least one i in } \mathrm{I}_0$$

- By Boole's inequality, this is $\leq$:

$$\sum_{i \in I_0} P(p_i \leq \frac{\alpha}{m}) = \sum_{i \in I_0} \frac{\alpha}{m} = \frac{m_0 \alpha}{m} \leq \alpha$$

Boole's Inequality

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

The probability that **at least one** of the events happens is no greater than the sum of the probabilities of the individual events

Problems with Bonferroni

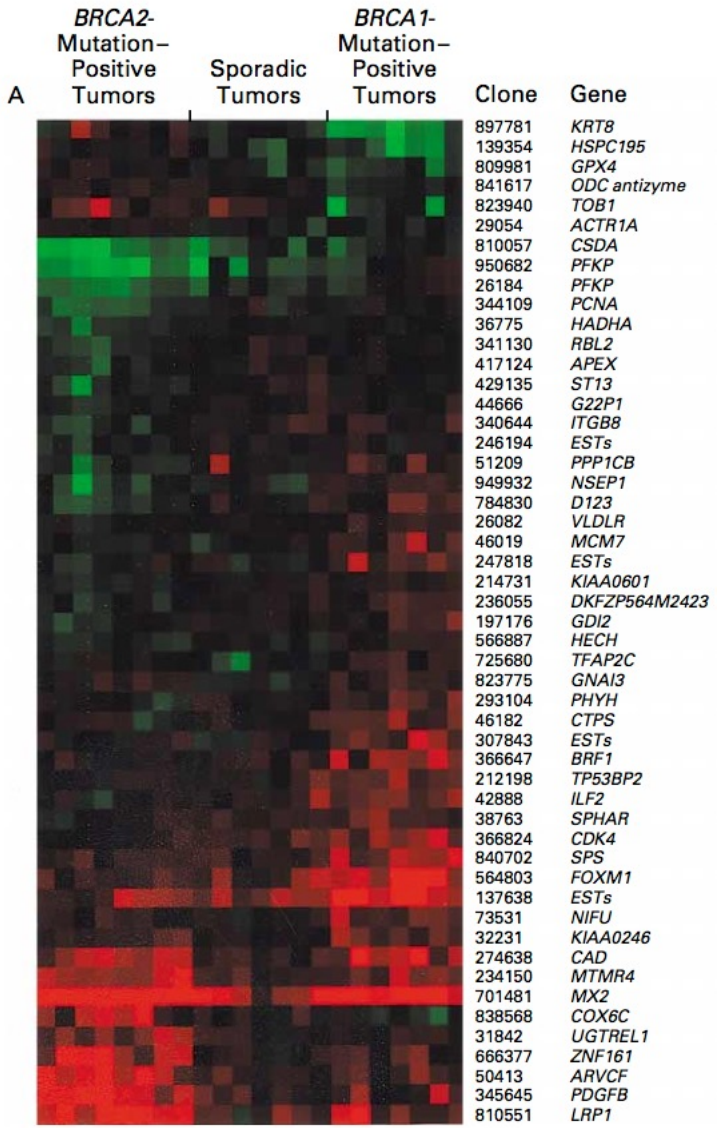
- Bonferroni correction is conservative
  - Can use Holm-Bonferroni instead:

$$P_k < \frac{\alpha}{m + 1 - k}$$

- Bonferroni says little about the mix of TPs and FPs in the set of hypotheses called significant.

- If we expect that many tests should reject H_0, we may be fine with more than one FP.

Genome-wide Analyses



Many genes are likely to be differentially expressed between conditions.

Why not control # FPs over tests that reject the null?

False Discovery Rate (FDR)

FP / (FP+TP)

vs.

False Positive Rate (FPR)

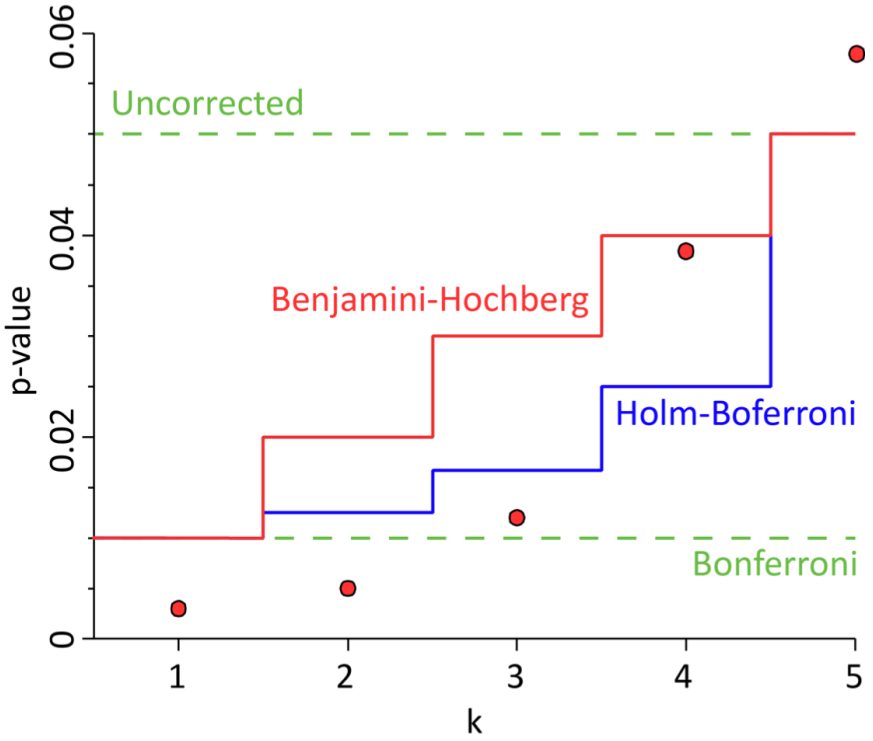
FP / (FP+TN)

H_0	Reject	Do Not Reject
True	# False Positives (FP)	# True Negatives (TN)
False	# True Positives (TP)	# False Negatives (FN)

q-value: the FDR analog of the p-value

Benjamini-Hochberg Procedure

1. Rank p-values in ascending order: $P_{(1)} \ldots P_{(m)}$.
2. For a given α, find largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
3. Reject the null for all $H_{(i)}$ for $i = 1, \ldots, k$.



http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/p-values8.pdf

Benjamini-Hochberg Procedure

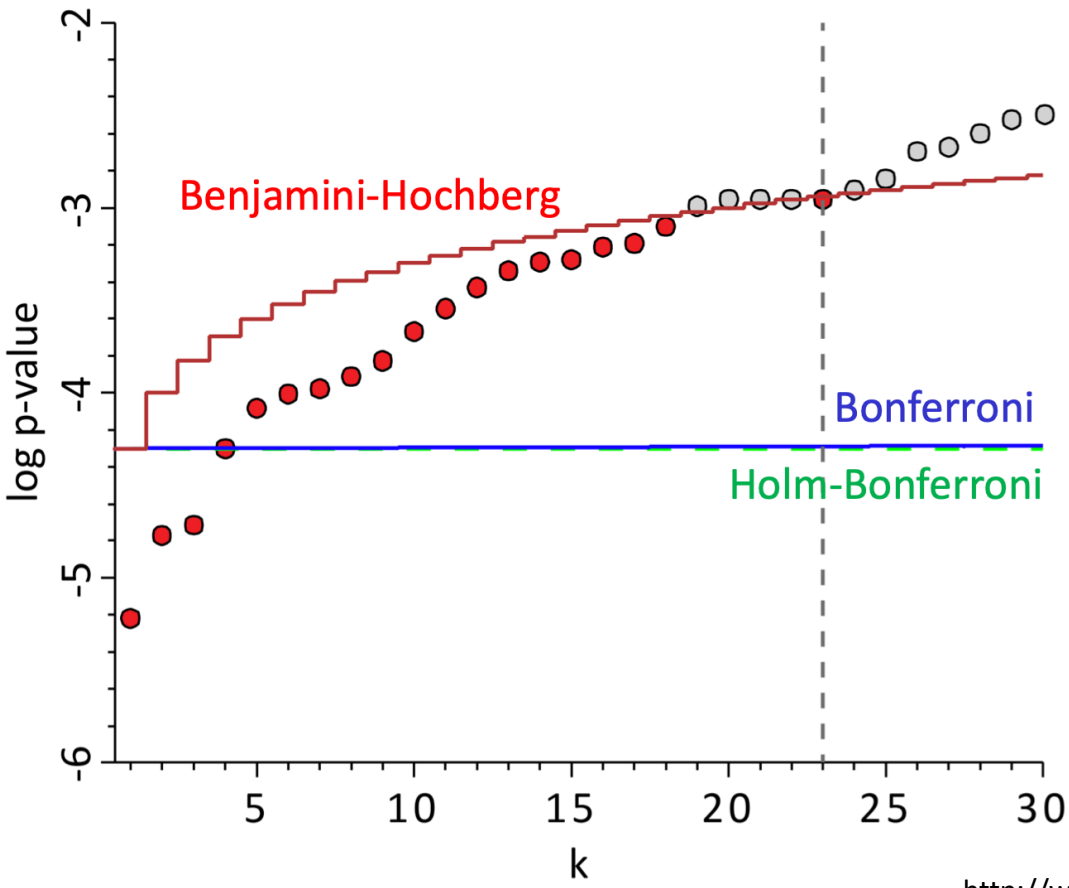
1. Rank p-values in ascending order: $P_{(1)} \ldots P_{(m)}$.
2. For a given α, find largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
3. Reject the null for all $H_{(i)}$for $i = 1, \ldots, k$.

- BH procedure is less conservative than Bonferroni correction.

- In genomics, we often expect many rejections of the null and can tolerate a few false positives.

BH Graphical Example

Imagine performing 1000 tests, 30 with H_0 false:



Bonferroni			
	No effect	Effect	Total
Significant	0	4	4
Not significant	970	26	996
Total	970	30	1000

Benjamini-Hochberg			
	H_0 true	H_0 false	Total
Significant	2	21	23
Not significant	968	9	977
Total	970	30	1000

http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/p-values8.pdf

Other useful metrics

Sensitivity, Recall, True Positive Rate

TP / (TP+FN)

Specificity, True Negative Rate

TN / (TN+FP)

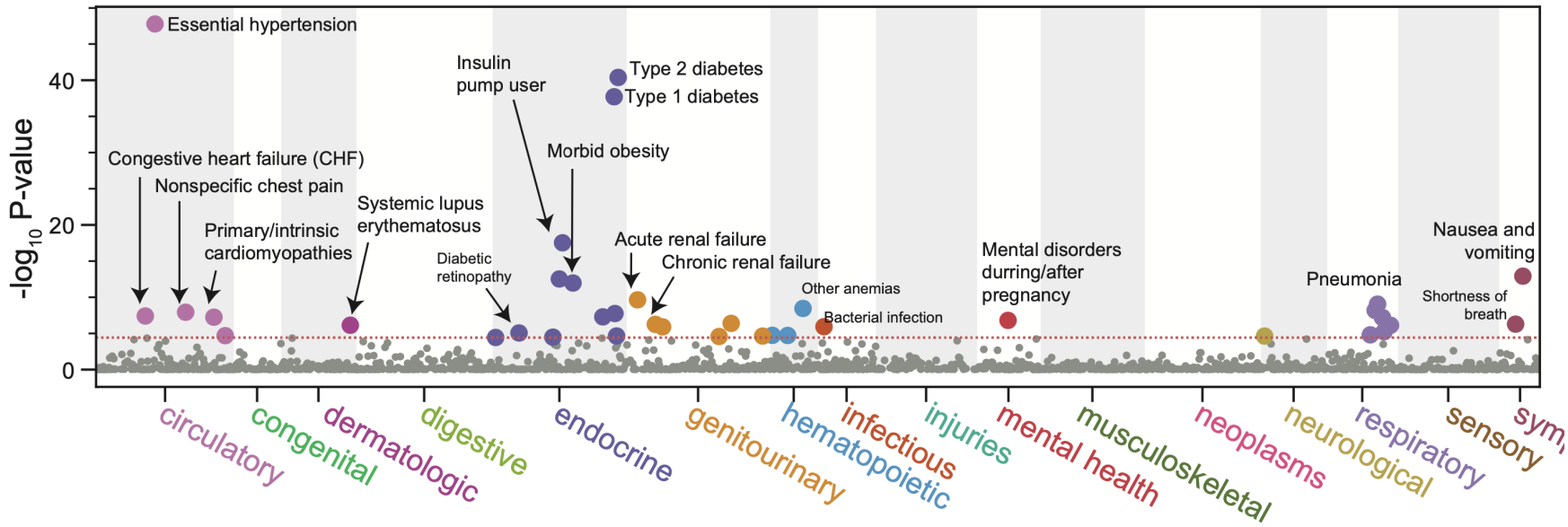
Precision, Positive Predictive Value

TP / (TP+FP)

H_0	Reject	Do Not Reject
True	# False Positives (FP)	# True Negatives (TN)
False	# True Positives (TP)	# False Negatives (FN)

Discussion

1. How can we account for correlation structure among the results of our multiple tests?



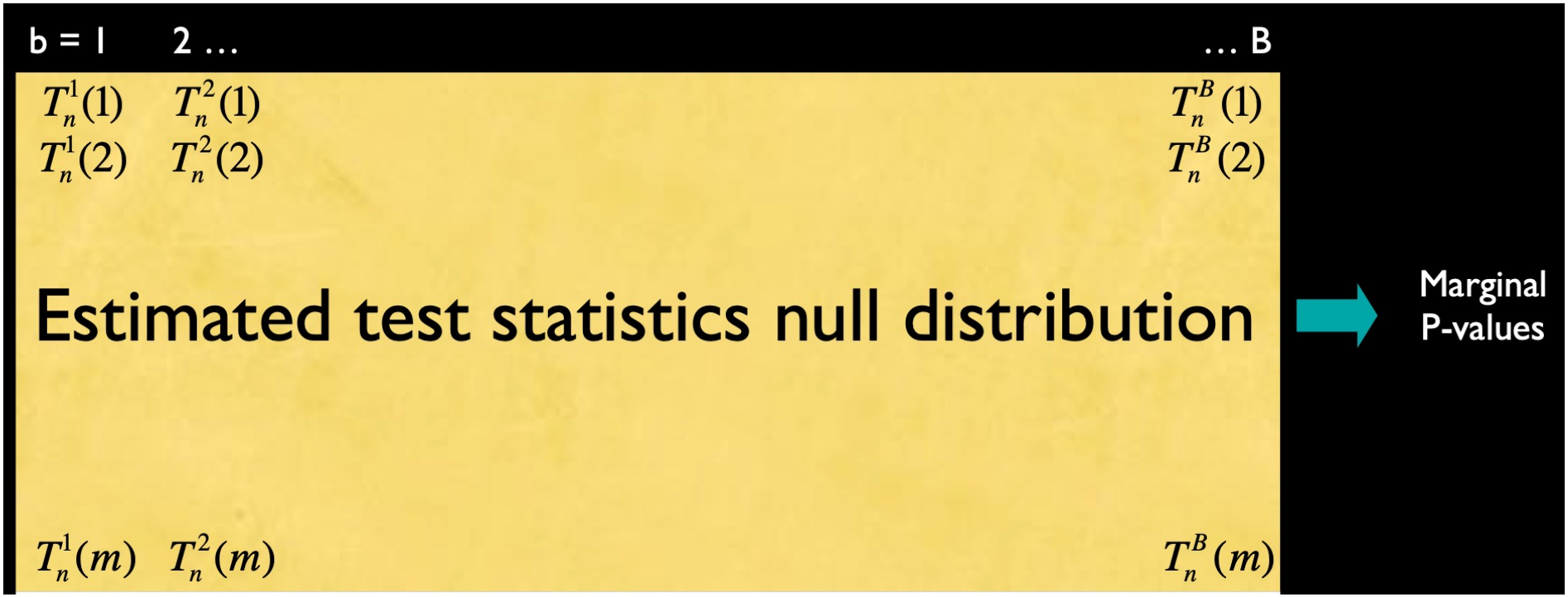
2. Should you perform multiple testing correction for all the hypotheses you test in your life?

Marginal correction methods assume independence of tests

- Bonferroni
- Holm-Bonferroni
- Benjamini-Hochberg
- etc.

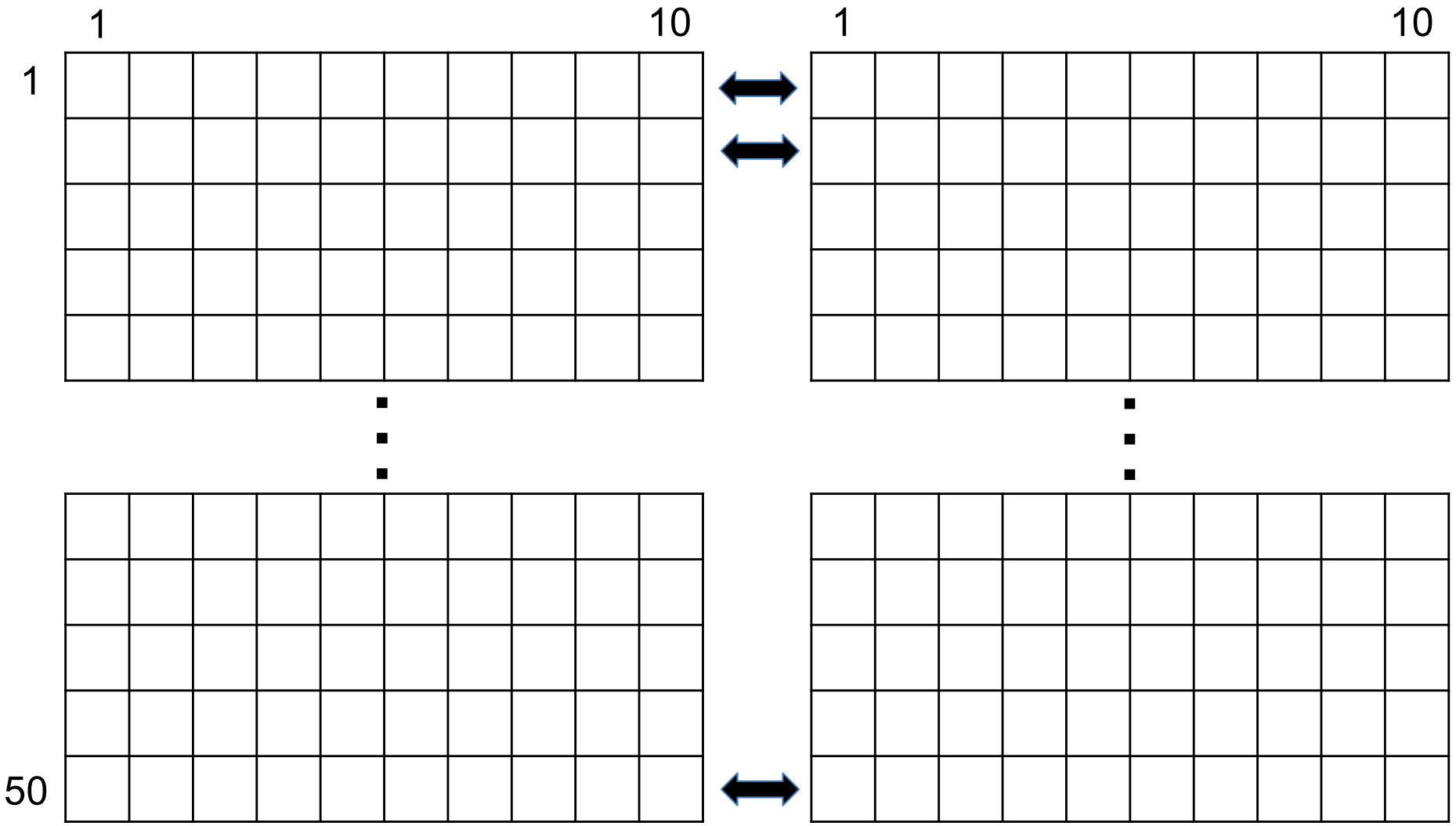
Bootstrapping can help…

Joint methods for adjusted p-values



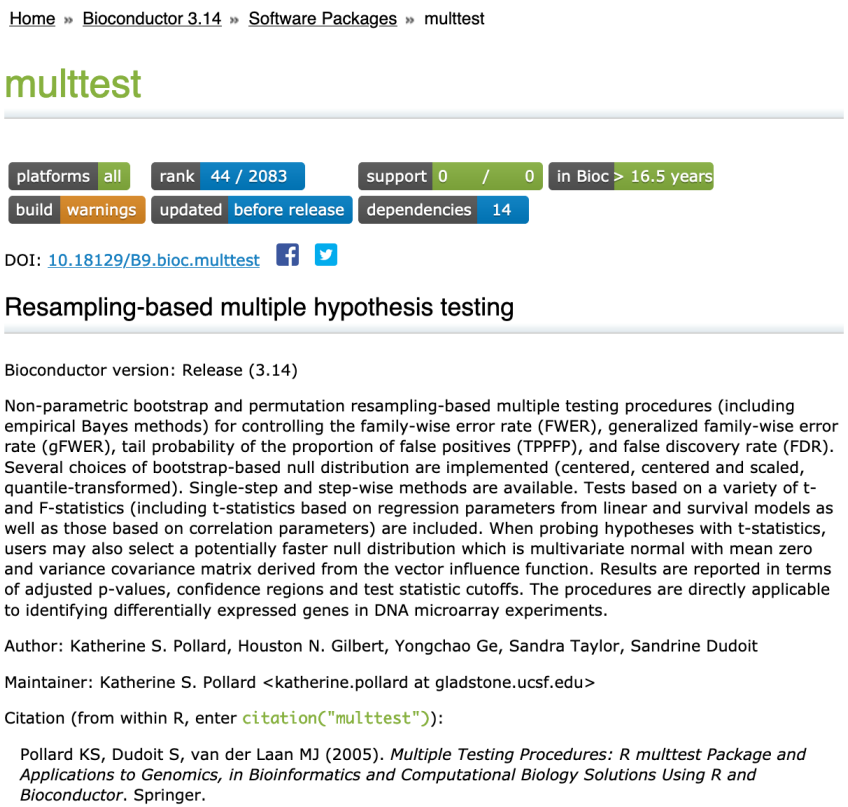
Comparing the observed stats to the joint null distribution of the test statistics enables direct control of Type I error rates.

Simulation Example



Tomorrow

- Examples of permutation and bootstrap methods for jointly adjusting for multiple testing



Home » Bioconductor 3.14 » Software Packages » multtest

multtest

platforms all | rank 44 / 2083 | support 0 / 0 | in Bioc > 16.5 years

build warnings | updated before release | dependencies 14

DOI: 10.18129/B9.bioc.multtest

Resampling-based multiple hypothesis testing

Bioconductor version: Release (3.14)

Non-parametric bootstrap and permutation resampling-based multiple testing procedures (including empirical Bayes methods) for controlling the family-wise error rate (FWER), generalized family-wise error rate (gFWER), tail probability of the proportion of false positives (TPPFP), and false discovery rate (FDR). Several choices of bootstrap-based null distribution are implemented (centered, centered and scaled, quantile-transformed). Single-step and step-wise methods are available. Tests based on a variety of t- and F-statistics (including t-statistics based on regression parameters from linear and survival models as well as those based on correlation parameters) are included. When probing hypotheses with t-statistics, users may also select a potentially faster null distribution which is multivariate normal with mean zero and variance covariance matrix derived from the vector influence function. Results are reported in terms of adjusted p-values, confidence regions and test statistic cutoffs. The procedures are directly applicable to identifying differentially expressed genes in DNA microarray experiments.

Author: Katherine S. Pollard, Houston N. Gilbert, Yongchao Ge, Sandra Taylor, Sandrine Dudoit

Maintainer: Katherine S. Pollard <katherine.pollard at gladstone.ucsf.edu>

Citation (from within R, enter `citation("multtest")`):

Pollard KS, Dudoit S, van der Laan MJ (2005). *Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.