

Hypothesis Testing

Wolfgang Huber, EMBL



Aims for this lecture

Understand the basic principles of hypothesis testing, its pitfalls, strengths, use cases and limitations

What changes when we go from single to multiple testing?

False discovery rates, p-value 'adjustments', filtering and weighting

See also

www.huber.embl.de/msmb Chapter 6

The screenshot shows a web browser window with three tabs: 'EMBL - Calendario - luglio 2019', 'CSAMA 2019 - Statistical Data', and 'Modern statistics modern biol...'. The address bar shows the URL 'https://www.cambridge.org/it/academic/subjects/statistics-probabili...'. The Cambridge University Press logo is in the top left, and navigation links for 'Academic', 'Cambridge English', 'Education', 'Bibles', 'Digital Products', 'About Us', and 'Careers' are in the top right. A search bar is located in the top right corner with the text 'Search for keyword, author, ISBN, etc.' and a search icon. Below the search bar are links for 'Include historic titles', 'Sign in', and 'Register'. A navigation menu includes 'Subjects', 'Blogs', 'News', 'Textbooks', 'Authors', 'Contact Us', 'Reference', and 'Conferences'. The main content area shows the book 'Modern Statistics for Modern Biology' by Susan Holmes and Wolfgang Huber. The book cover is on the left, and the title and authors are on the right. The price is £ 49.99 Paperback. There are buttons for 'Add to cart' and 'Add to wishlist'. A 'Request inspection copy' button is also present. Below the book information are links for 'Description', 'Contents', 'Resources', 'Courses', and 'About the Authors'. At the bottom, there is a footer with the text 'If you are a biologist and want to get the best out of the powerful methods of modern computational' and a 'RELATED BOOKS' section.

EMBL - Calendario - luglio 2019 X CSAMA 2019 - Statistical Data X Modern statistics modern biol... X

https://www.cambridge.org/it/academic/subjects/statistics-probabili...

CAMBRIDGE UNIVERSITY PRESS Academic Cambridge English Education Bibles Digital Products About Us Careers Italy

Cart (0)

Academic

Unlocking potential with the best learning and research solutions

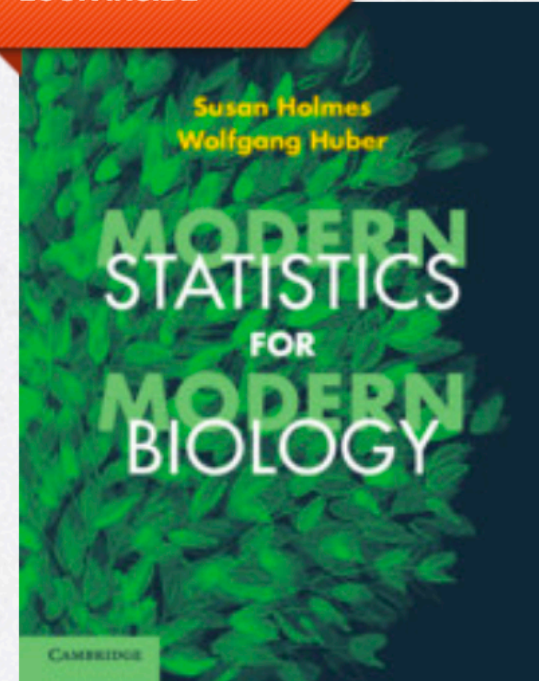
Search for keyword, author, ISBN, etc.

Include historic titles Sign in Register

Subjects Blogs News Textbooks Authors Contact Us Reference Conferences

Home Academic Statistics and probability Statistics for life sciences, medicine and health

LOOK INSIDE



Modern Statistics for Modern Biology

TEXTBOOK

AUTHORS:
Susan Holmes, Stanford University, California
Wolfgang Huber, European Molecular Biology Laboratory

DATE PUBLISHED: February 2019

AVAILABILITY: In stock

FORMAT: Paperback

ISBN: 9781108705295

Rate & review

£ 49.99
Paperback

Add to cart Add to wishlist

Request inspection copy
Lecturers may request a copy of this title for inspection
Request

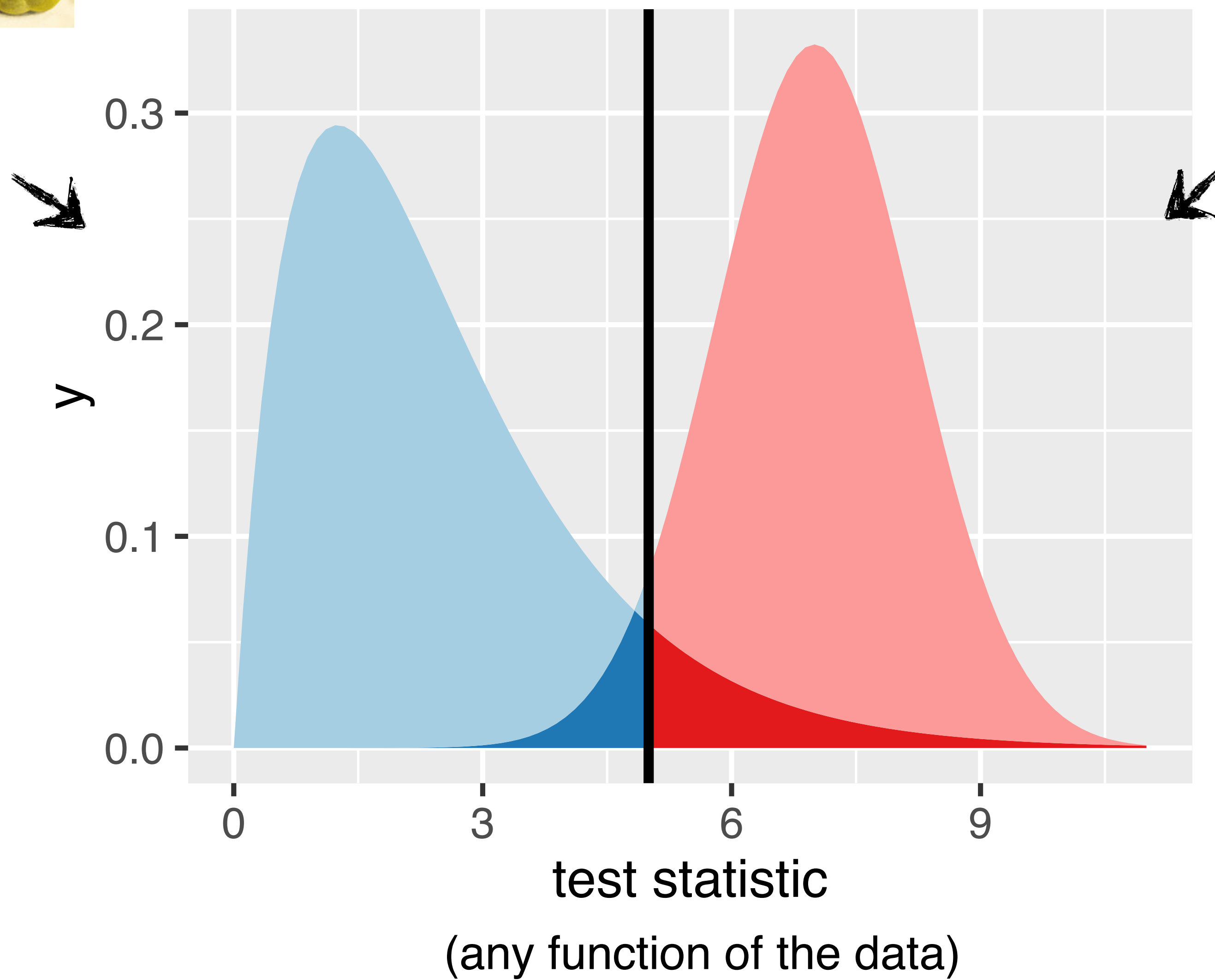
I want this title to be available as an eBook

Description Contents Resources Courses About the Authors

If you are a biologist and want to get the best out of the powerful methods of modern computational

RELATED BOOKS

Testing vs Classification



Accuracy vs Precision - Bias vs Variance

← bias

accuracy →

dispersion →



← precision



How to make rational decisions based on noisy, finite data

Prototypical examples:

- Testing efficacy of a drug on people
 - lack of complete experimental control
 - finite sample size
- Effect of fertilizer, genetic variants, ... on phenotype of plants in an outdoors field trial
- Lady testing tea, clairvoyant, telepath, ...
- Toxicology

+: No understanding of mechanism involved / needed / desired

-: Wouldn't we want to use any available understanding or 'priors'?

Example



Toss a coin a number of times \Rightarrow

If the coin is fair, then heads should appear half of the time (roughly).

But what is “roughly”? We use combinatorics / probability theory to quantify this.

Suppose we flipped the coin 100 times and got 59 heads. Is this ‘significant’?

Binomial distribution

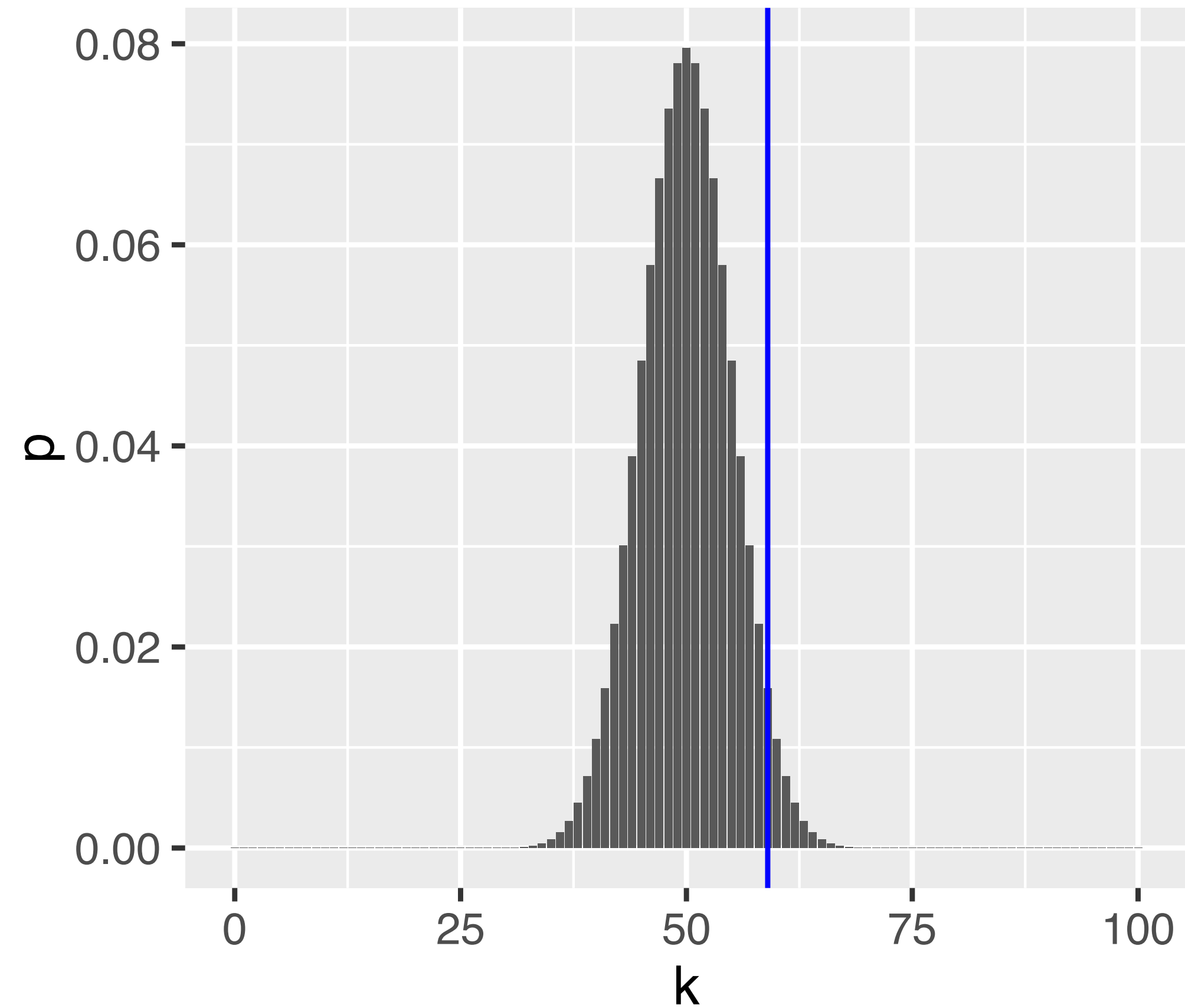


Figure 6.3: The binomial distribution for the parameters $n = 100$ and $p = 0.5$,

$$P(K = k | n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Rejection region

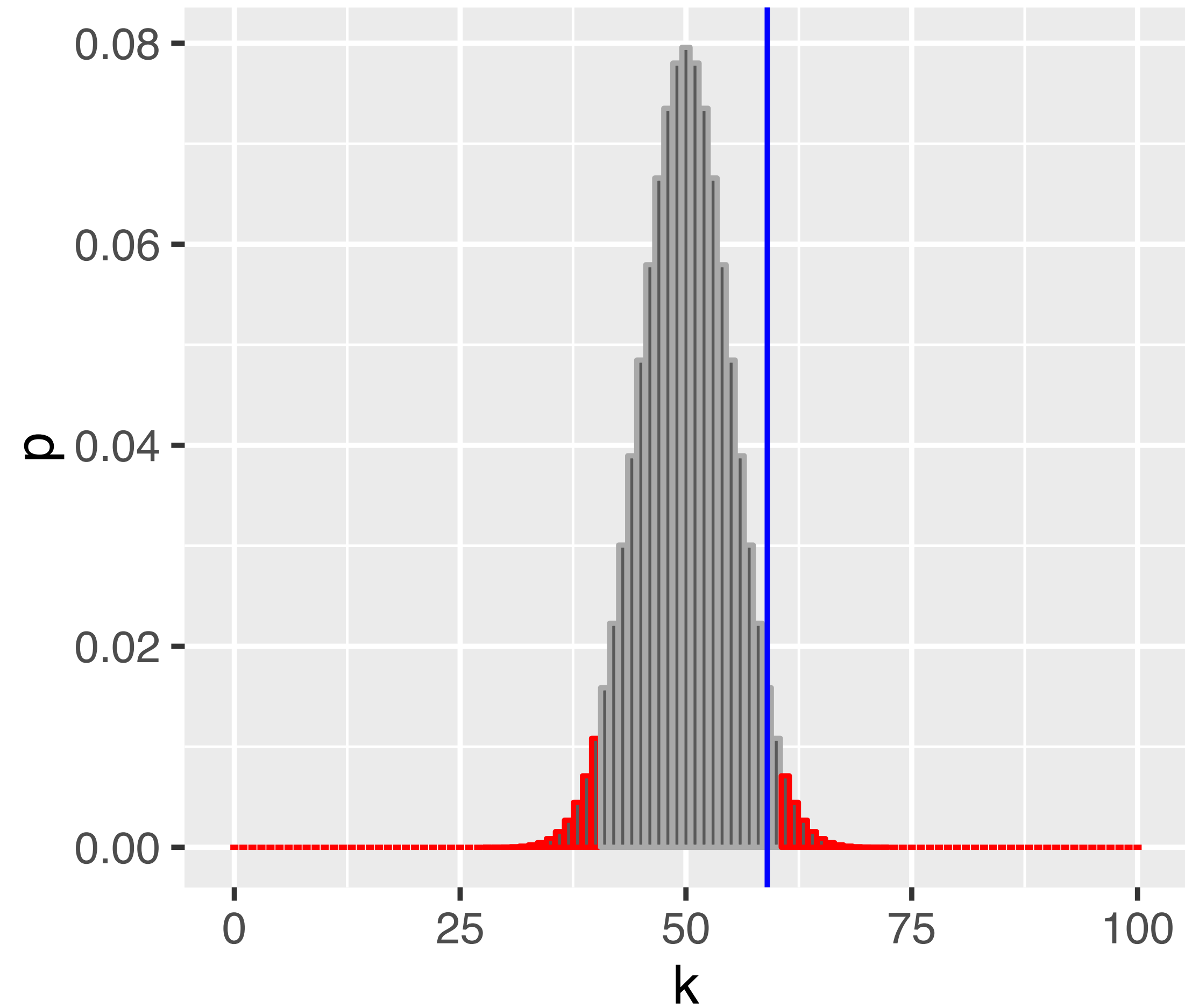


Figure 6.5: As Figure 6.3, with rejection region (red) whose total area is $\alpha = 0.05$.

Questions

- Does the fact that we don't reject the null hypothesis mean that the coin is fair?
- Would we have a better chance of detecting an unfair coin if we did more coin tosses? How many?
- If we repeated the whole procedure and again tossed the coin 100 times, might we then reject the null hypothesis?
- Our rejection region is asymmetric - its left part ends with 40, while its right part starts with 61. Why is that? Which other ways of defining the rejection region might be useful?

The Five Steps of Hypothesis Testing

Choose an experimental design and a data summary function for the effect that you are interested in: the test statistic

Set up a null hypothesis that lets you compute the possible outcomes and each

This is the idealised scenario, “orthodoxy”.

of reality that possible

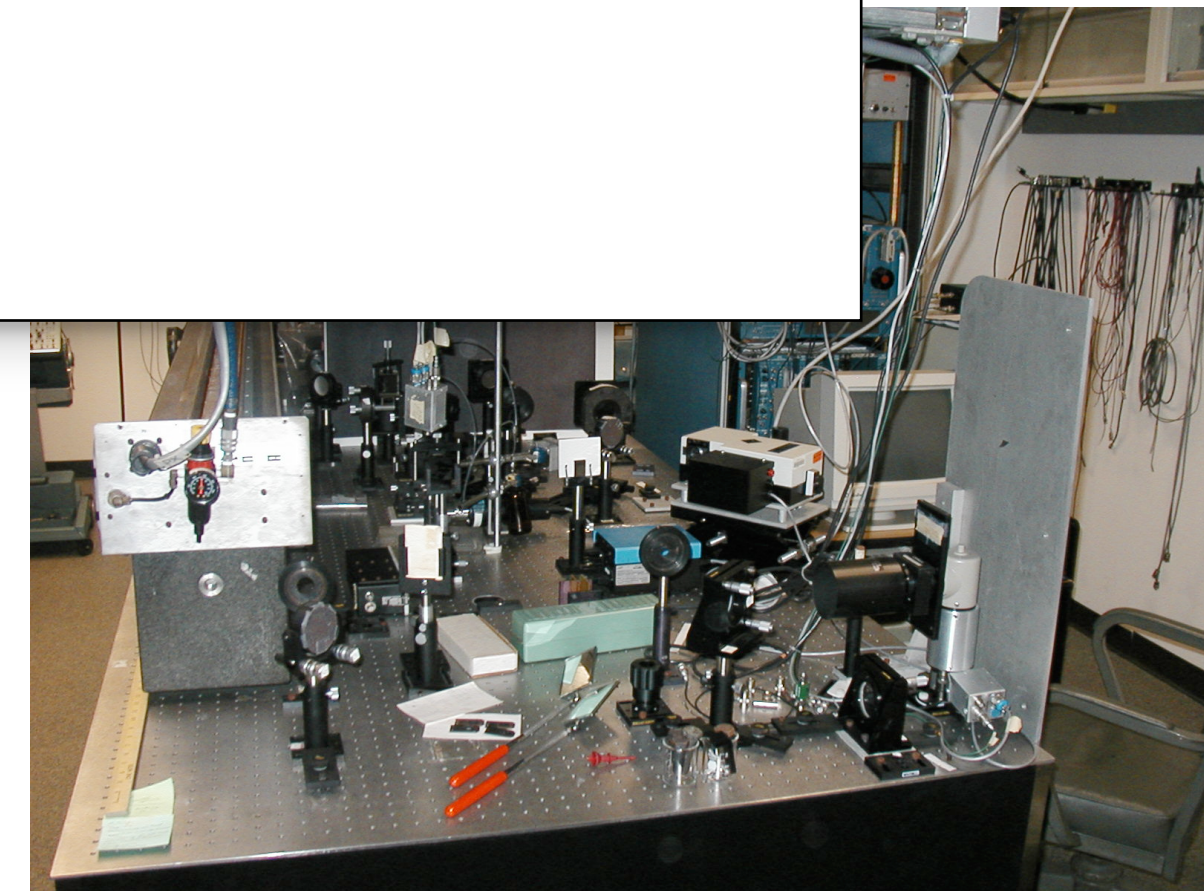
Decide on the rejection probability is small (significance level)

Reality, esp. in retrospective ‘data-mining’ can be quite different.

whose total

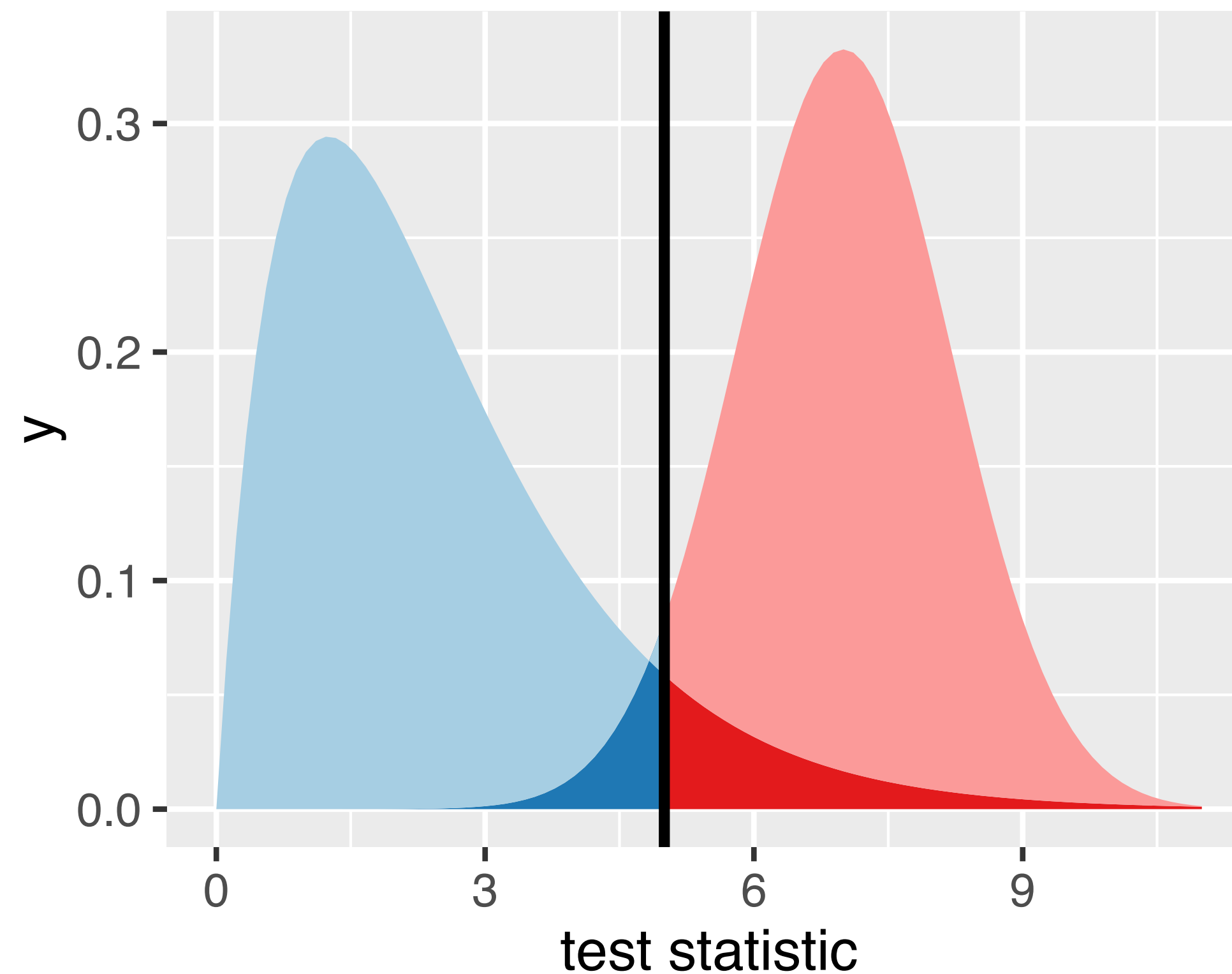
Do the experiment and compute the test statistic.

Make a decision: reject null hypothesis if the test statistic is in the rejection region.



Types of Error in Testing

Test vs reality	Null hypothesis is true	...is false
Reject null hypothesis	Type I error (false positive)	True positive
Do not reject	True negative	Type II error (false negative)



Parametric Theory vs Simulation

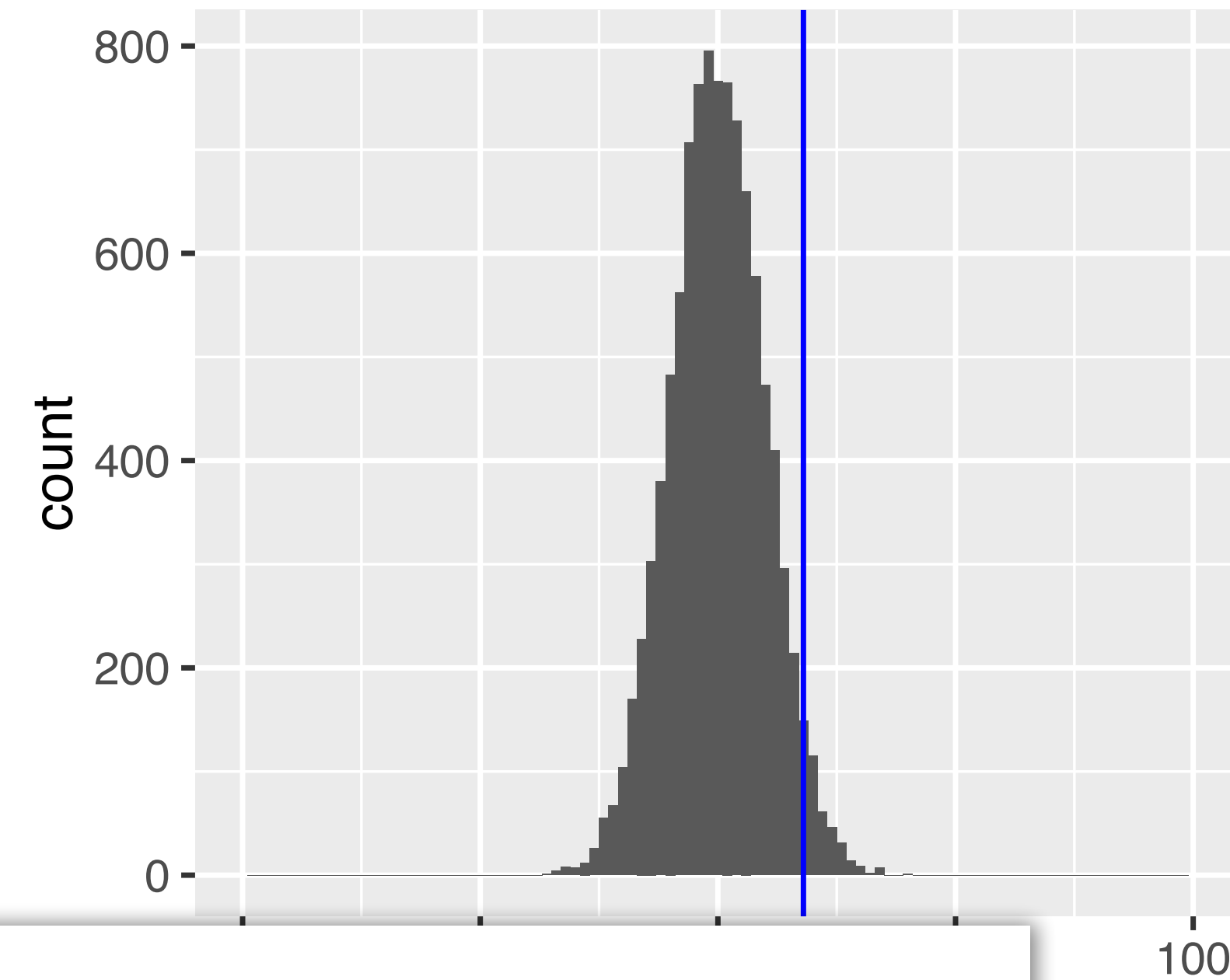
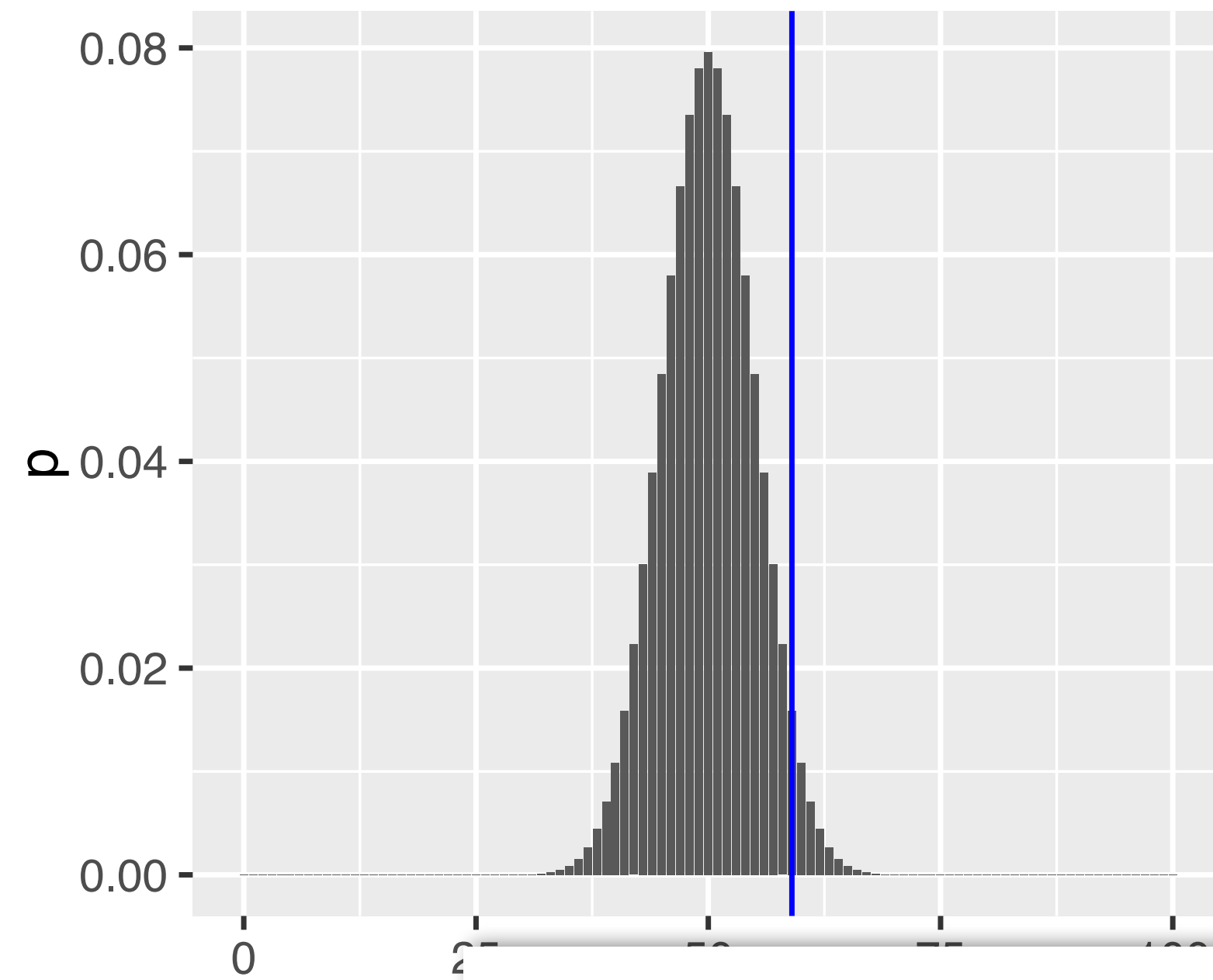


Figure 6.3: The histograms show the distribution of the number of successes K in n trials, with the parameters n and p according to Equation 6.1.

Q:

Discuss pros and cons for each

the
simulations

$$P(K = k | n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$



The choice of the test statistic

Suppose we observed 50 tails in a row, and then 50 heads in a row. Is this a perfectly fair coin?

We could use a different test statistic: number of times we see two tails in a row

Is this statistic generally and always preferable?

Power

There can be several test statistics, with different power, for different types of alternative

Continuous data: the t-statistic

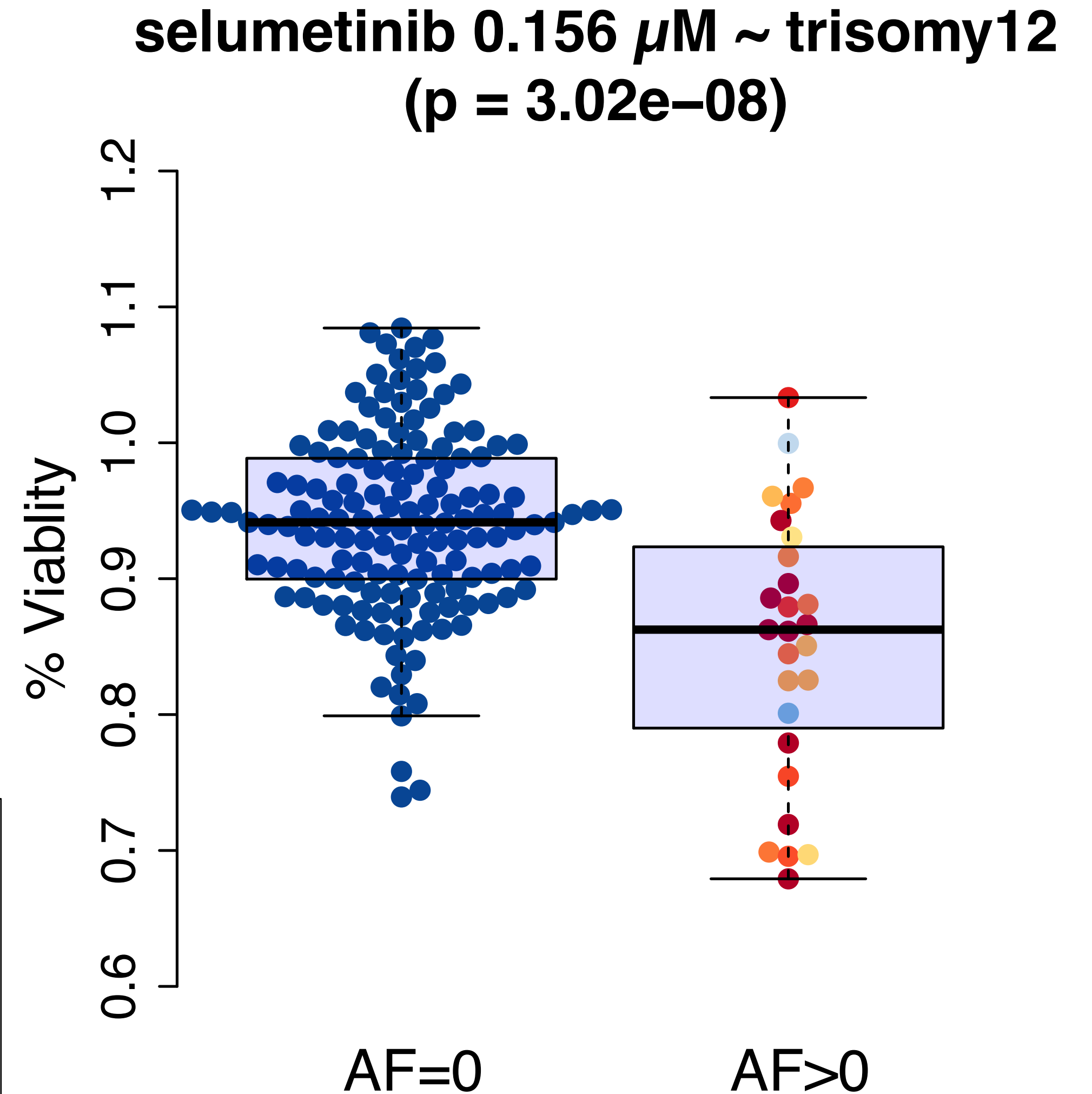
$$t = c \frac{m_1 - m_2}{s}$$

- Can also be adapted to one group only
- Relation to z-score

$$m_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{g,i} \quad g = 1, 2$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_{1,i} - m_1)^2 + \sum_{j=1}^{n_2} (x_{2,j} - m_2)^2 \right)$$

$$c = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$



t-distribution

If the data are identically normal distributed and independent, then under H_0 , t follows a 't-distribution' with parameter n_1+n_2 (a.k.a. degrees of freedom)

Q:

How does the distribution of $|t|$ look?

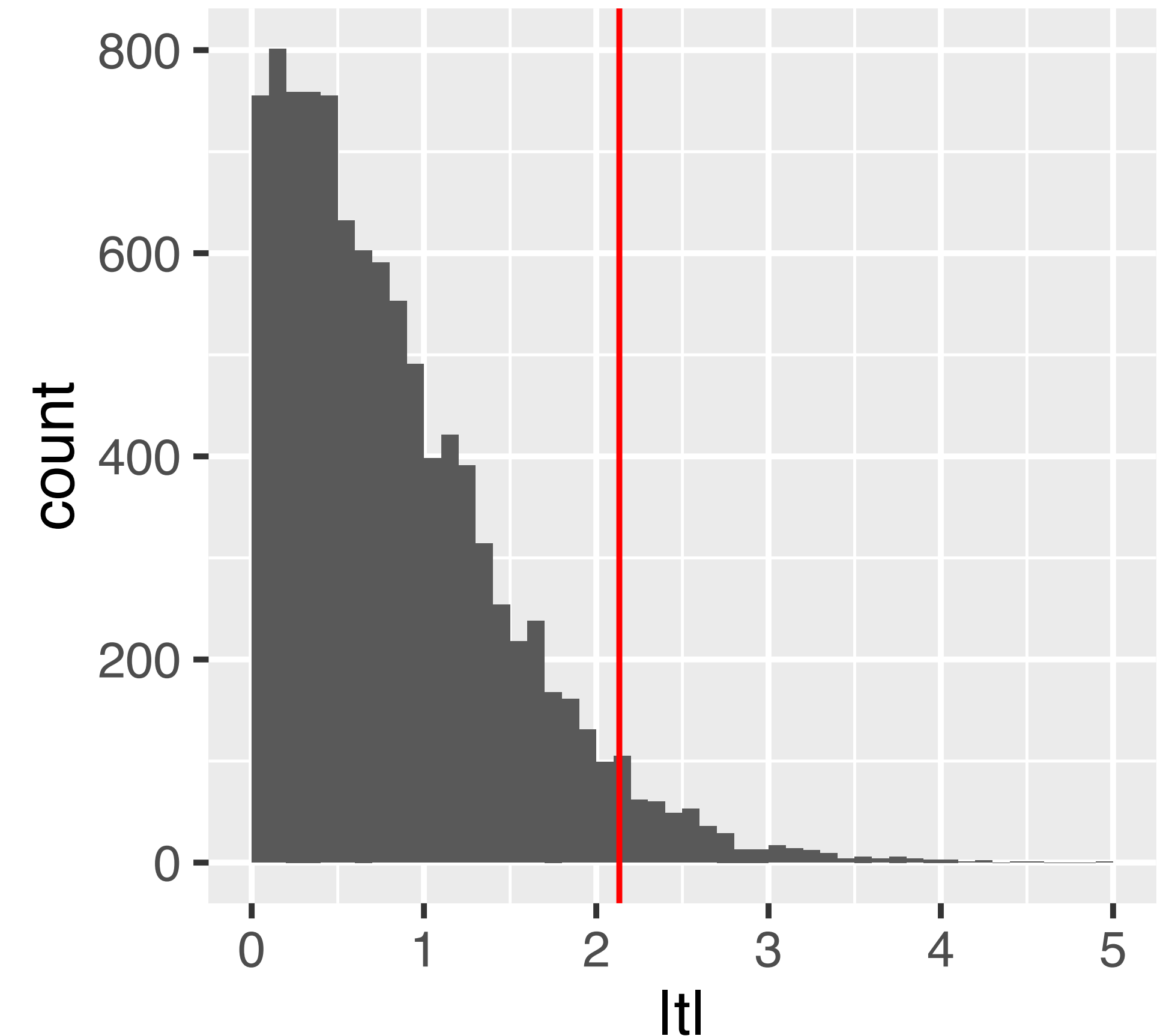


Figure 6.8: The null distribution of the (absolute) t -statistic determined by simulations – namely, by random permutations of the group labels.

Comments and Pitfalls

The proof that the t-statistic follows a t-distribution assumes that observations are independent and follow a normal distribution: this is a sufficient, but not necessary, condition

Deviation from normality (heavier tails): test typically maintains type-I error control, but no longer has optimal power.

Options: use permutations;
use a different test (e.g., Wilcoxon)

Deviation from independence: type-I error control is lost, p-values will likely be totally wrong (e.g., for positive correlation, too optimistic).

No easy options:

... try to model the dependence / remove it ...

... empirical null (Efron et al.) ...

Avoid Fallacy

The p-value is the probability that the data could happen, under the condition that the null hypothesis is true.

It is not the probability that the null hypothesis is true.

Absence of evidence \neq
evidence of absence



Recap: Single Hypothesis Testing

p-values are random variables: uniformly distributed if the null hypothesis is true - and should be close to zero if the alternative holds.

Note: We only observe one draw.

We prove something by disproving ('rejecting') the opposite (the null hypothesis). Reject = Discover.

Not rejecting does not prove the null hypothesis

Repeating the experiment (under the null): Around 5% of the times the p-value will be less than 0.05 by chance

All this reasoning is probabilistic. Testing & p-values are for rational decision making in uncertain contexts.

Limitations of p-value based hypothesis testing

Too much power: often, the 'null' is small (point-like), alternative is large (region-like)

Summarizing the data into one single number mashes together effect size and sample size

No place to take into account plausibility or 'prior' knowledge

What is p-Value Hacking ?

On the same data, try different tests until one is significant

On the same data, try different hypotheses until one is significant (HARKing - hypothesizing after results are known)

Moreover....:

retrospective data picking

'outlier' removal

the 5% threshold and publication bias

The ASA's Statement on p-Values:
Context, Process, and Purpose
Ronald L. Wasserstein & Nicole A.
Lazara DOI:
10.1080/00031305.2016.1154108

What can we do about this?

The right answer to the wrong question

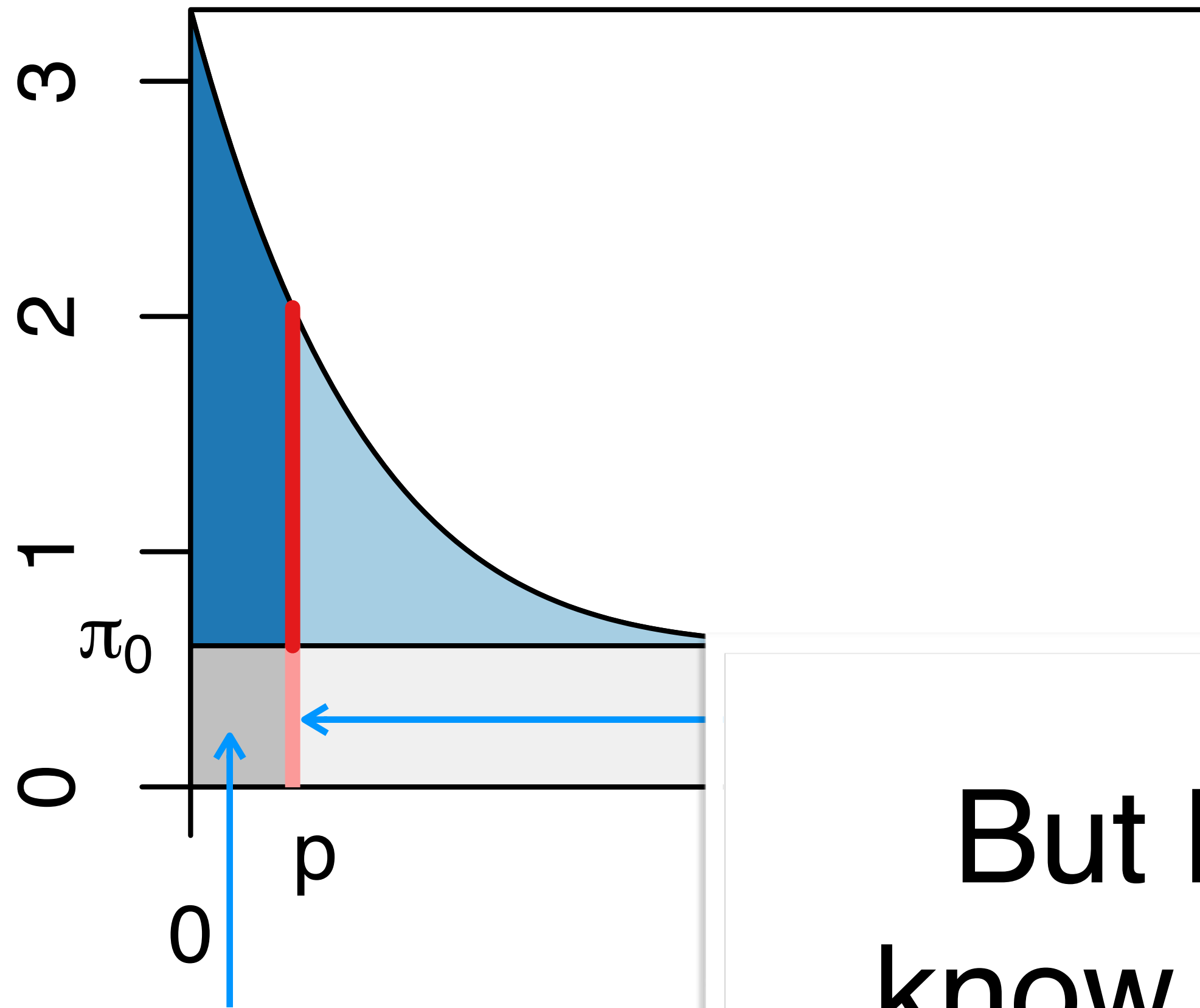
Researchers (regulators, investors, etc.) usually want to know:

If I publish this finding (allow this drug, invest in this product, ...), what is the probability that I'll later be proven wrong (cause harm, lose my money, ...)?

The p-value is the probability of seeing the data if the null hypothesis is true. It has little to do with the probability that my subsequent decision is wrong (a.k.a. "false discovery").

Can we compute a False Discovery Probability instead?

The two-groups model and the (local) false discovery rate



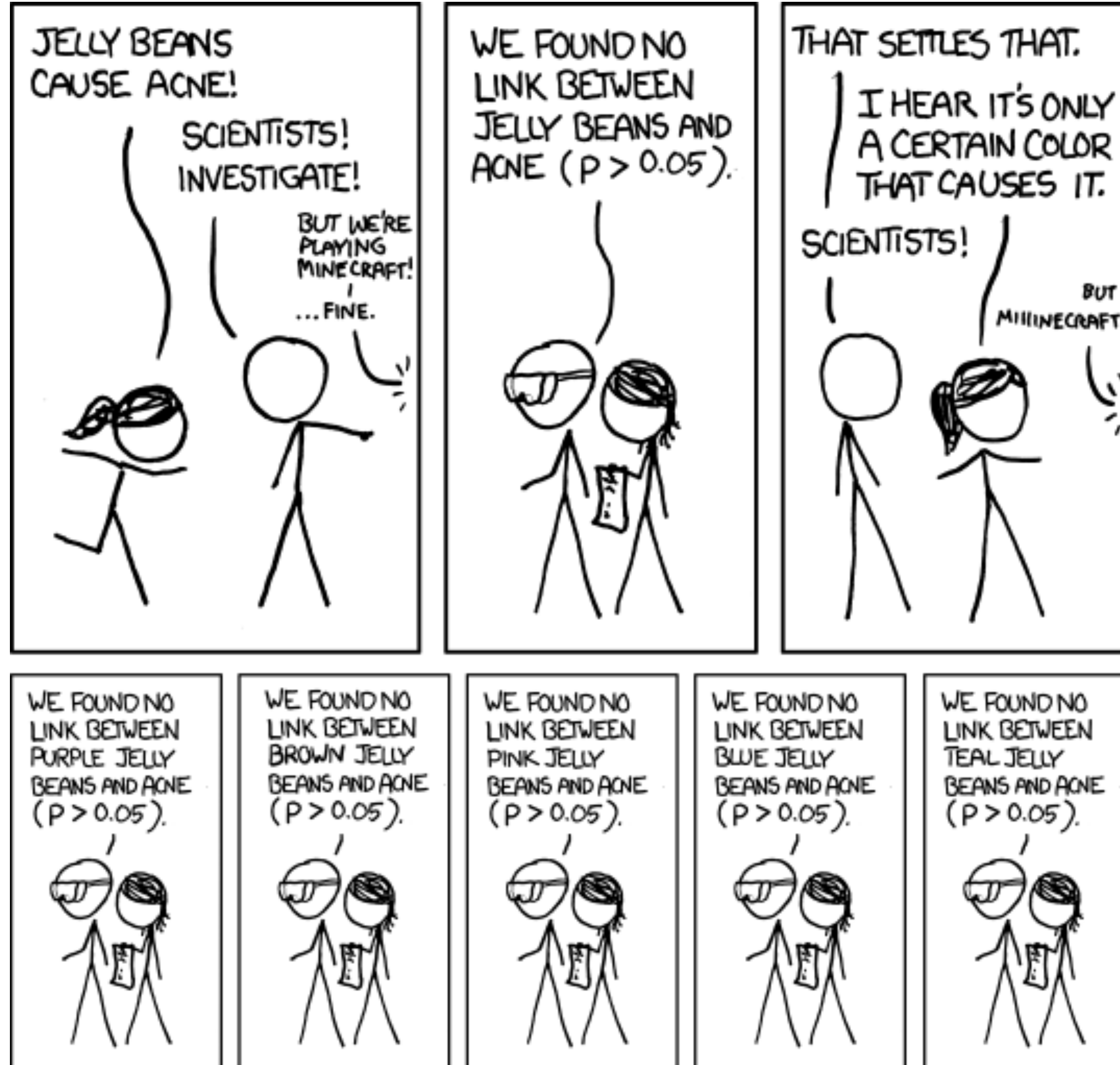
$$f(p) = \pi_0 + (1 - \pi_0)f_{\text{alt}}(p),$$

$$\text{fdr}(p) = \frac{\pi_0}{f(p)}.$$

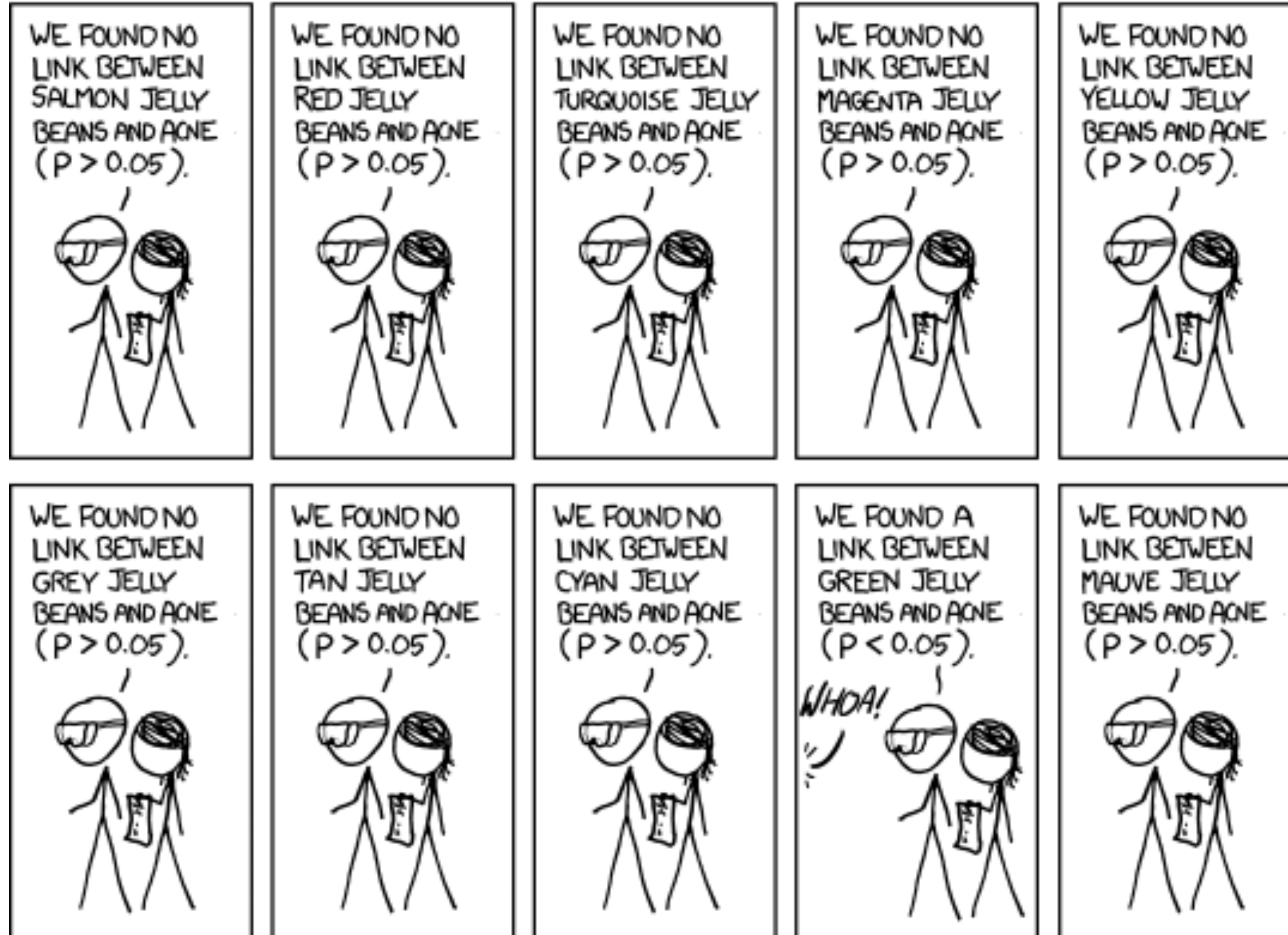
But how do we know π_0 and f_{alt} ?

FDR: Ratio between set property. It applies to a set of hypotheses (discoveries).

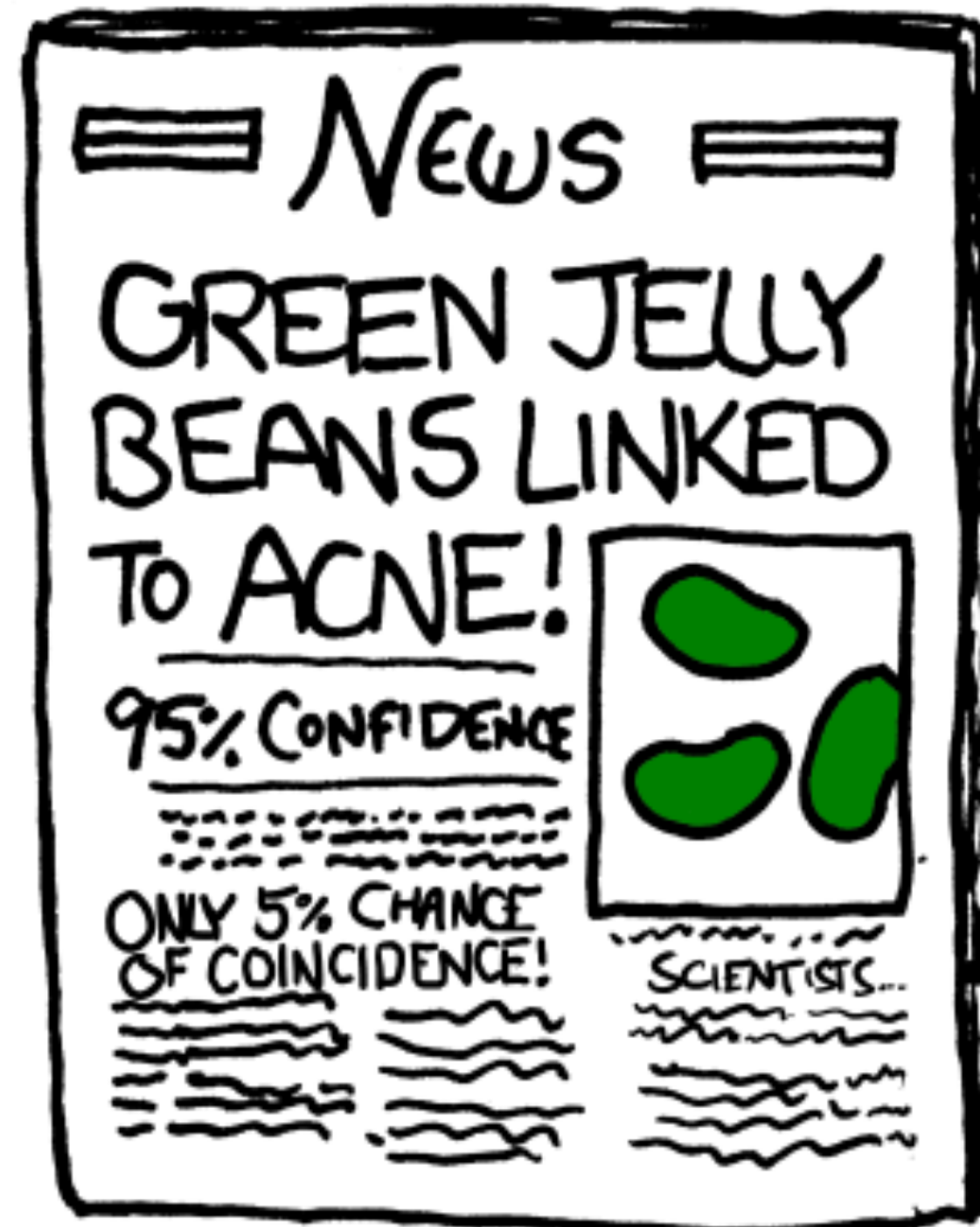
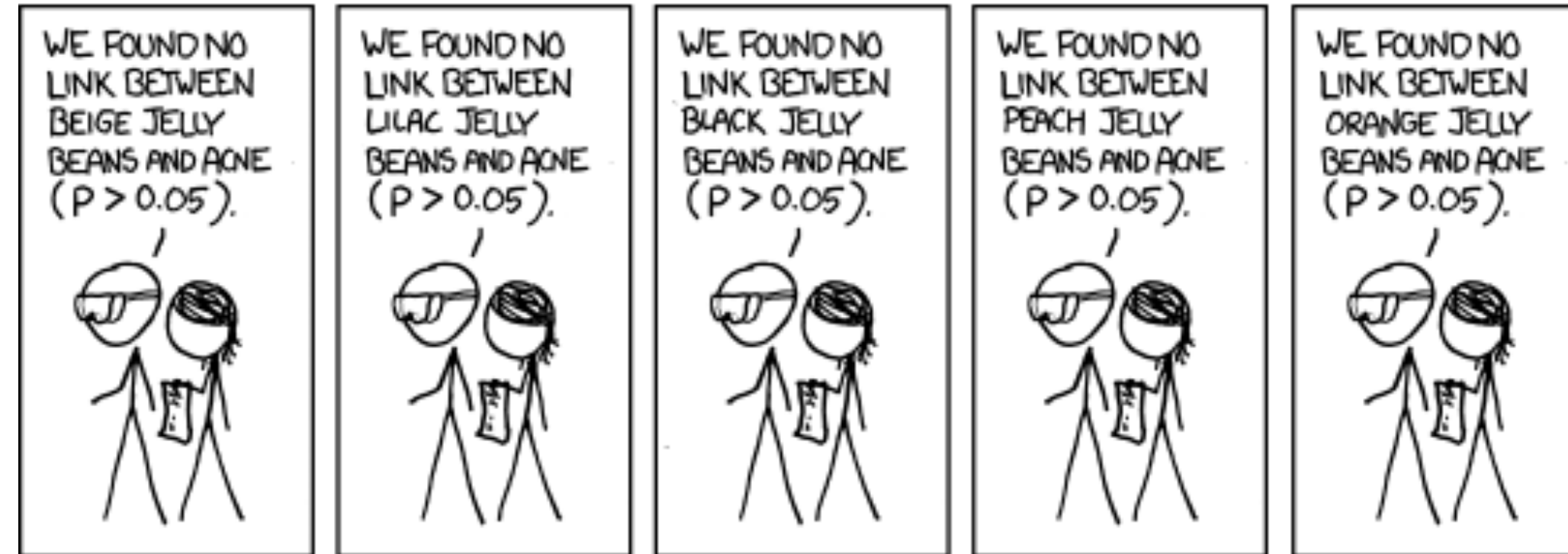
Multiple Testing



Multiple Testing



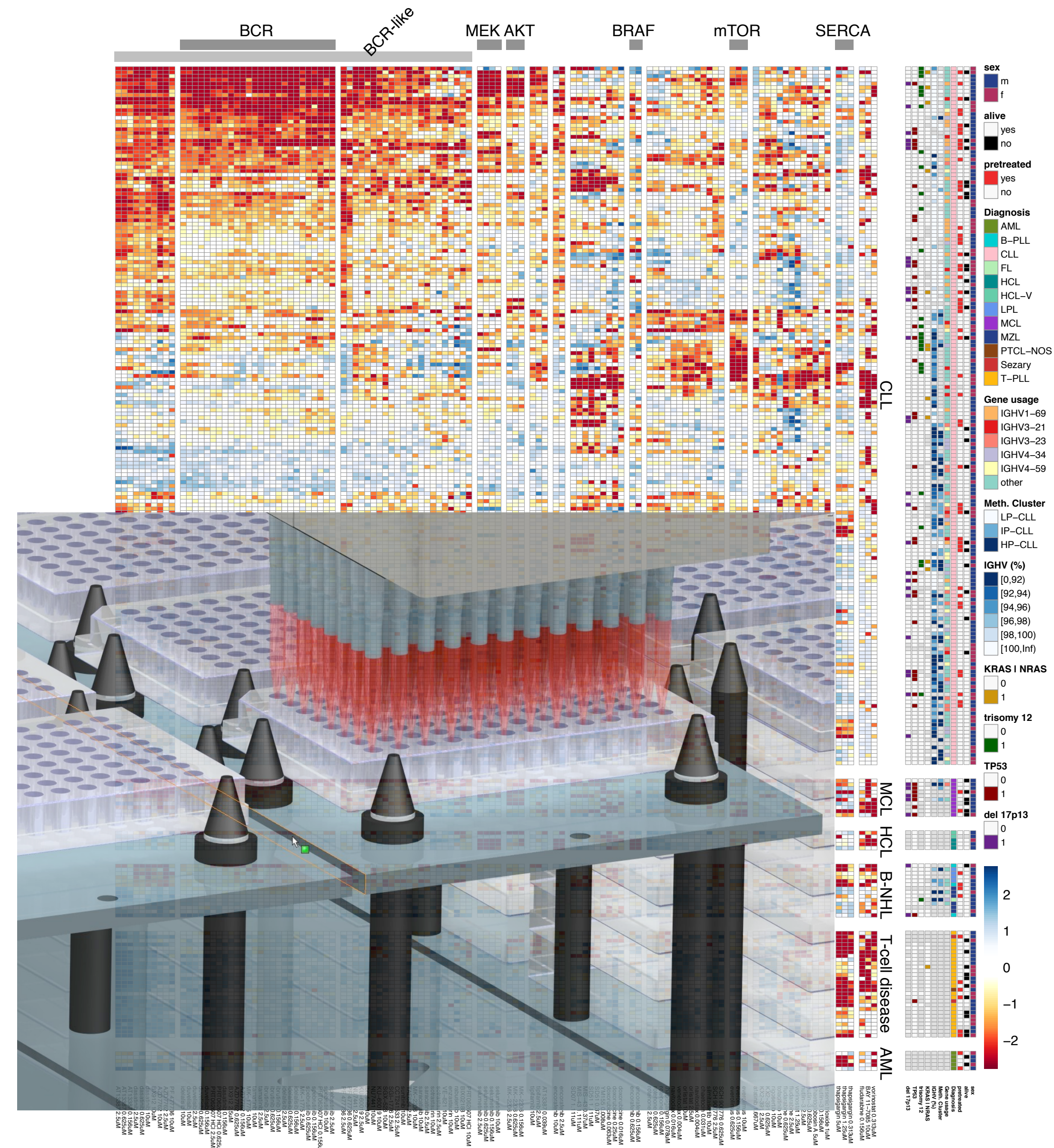
Multiple Testing



Multiple Testing

Many data analysis approaches in genomics employ item-by-item testing:

- Expression profiling
- Differential microbiome analysis
- Genetic or chemical compound screens
- Genome-wide association studies
- Proteomics
- Variant calling
- ...



False Positive Rate and False Discovery Rate

FPR: fraction of FP among all true negatives

FDR: fraction of FP among hits called

Example:

20,000 genes, 500 are d.e.,
100 hits called, 10 of them wrong.

FPR: $10/19,500 \approx 0.05\%$

FDR: $10/100 = 10\%$



"Wait a minute! Isn't anyone here a real sheep?"

Experiment-Wide Type I Error Rates

Test vs Reality	Null Hypothesis is true	...is false	Total
Rejected	V	S	R
Not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m : total number of hypotheses
- m_0 : number of null hypotheses
- V : number of false positives (a measure of type I error)

Family-wise error rate (FWER): The probability of one or more false positives, $P(V > 0)$. For large m_0 , this is difficult to keep small.

False discovery rate (FDR): The expected fraction of false positives among all discoveries, $E[V / \max \{R, 1\}]$.

NB: if $m_0=m$, then $FDR=FWER$

The Multiple Testing Burden

When performing several tests, type I error goes up: for $\alpha = 0.05$ and n indep. tests, probability of no false positive result is

$$\underbrace{0.95 \cdot 0.95 \cdot \dots \cdot 0.95}_{n\text{-times}} \lll 0.95$$



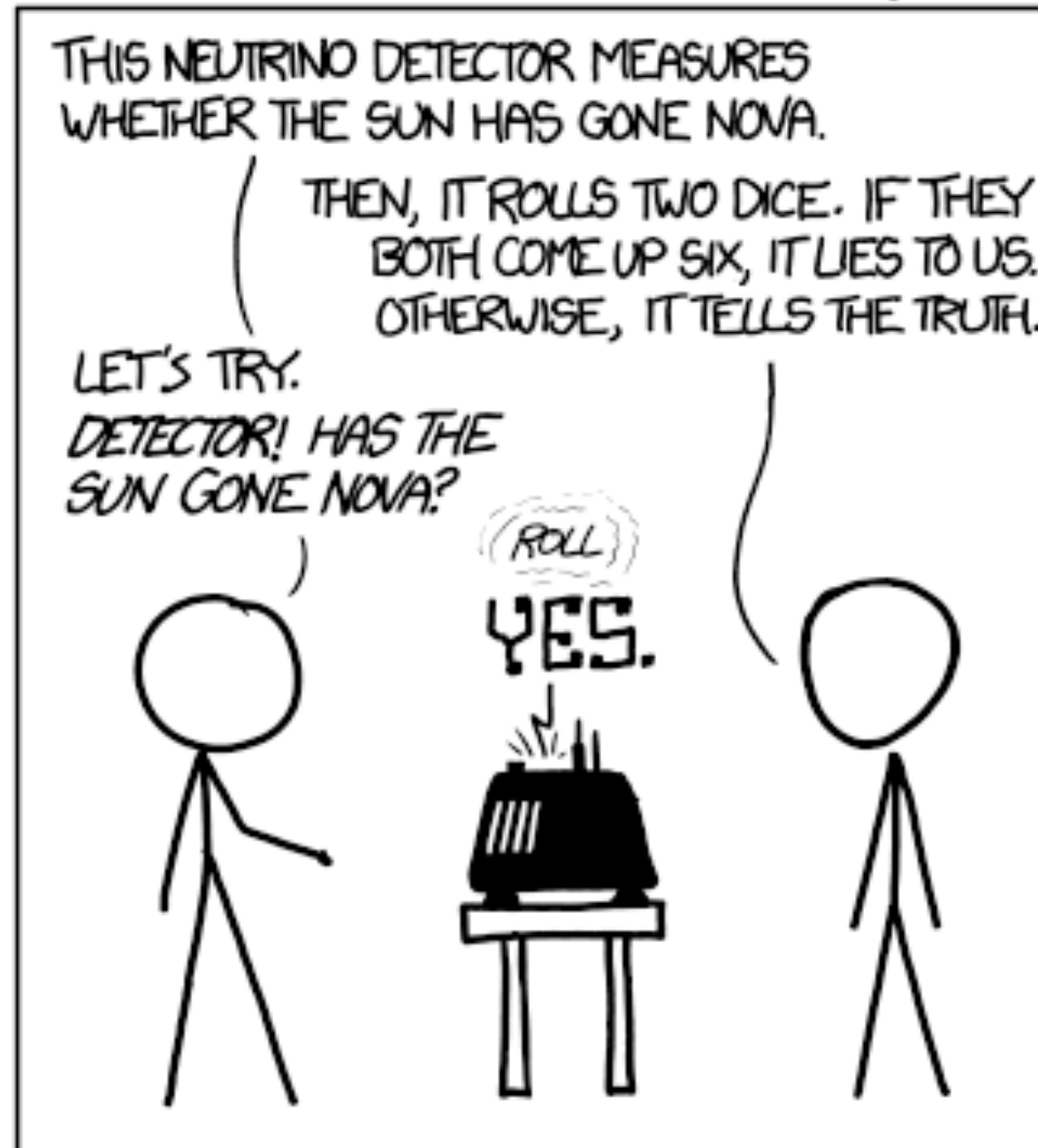
Bonferroni Correction



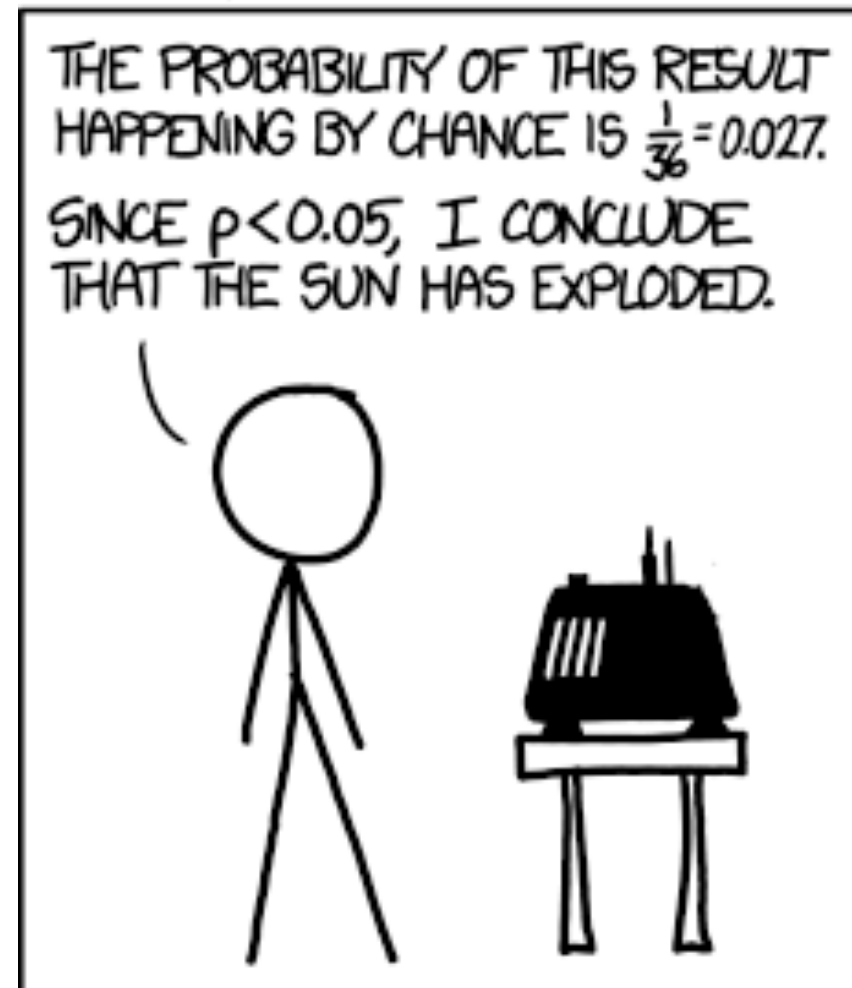
For m tests, multiply each p-value with m .
Then see if anyone still remains below α .

The Multiple Testing Opportunity

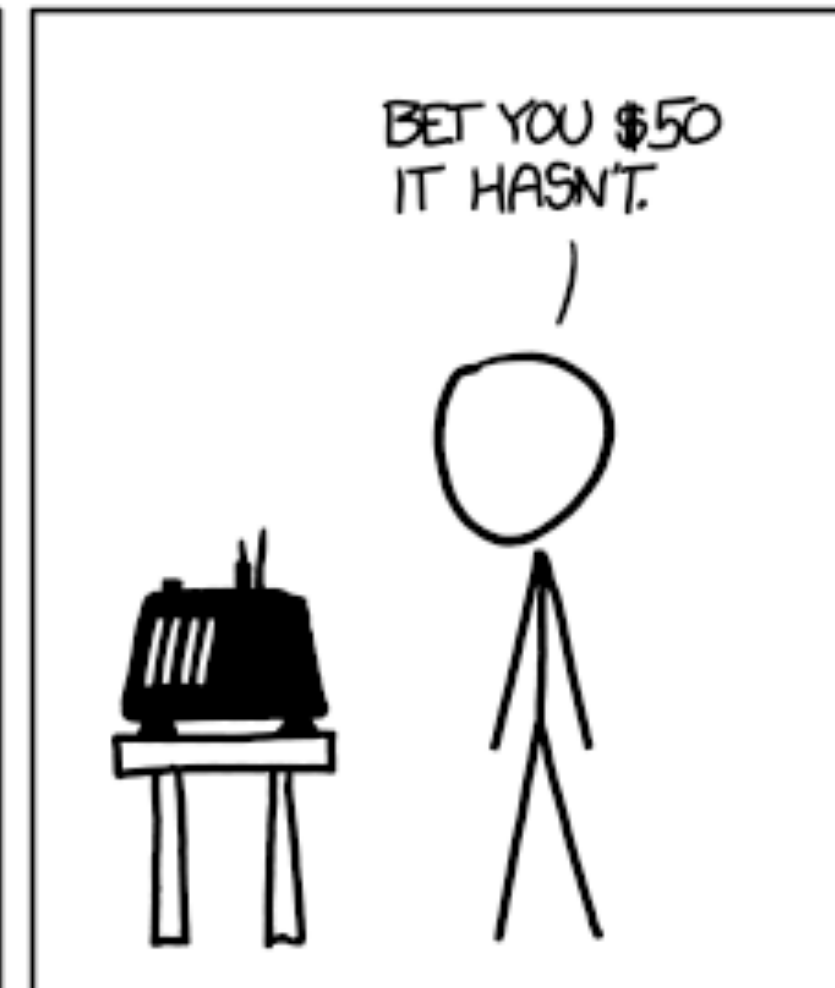
DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:



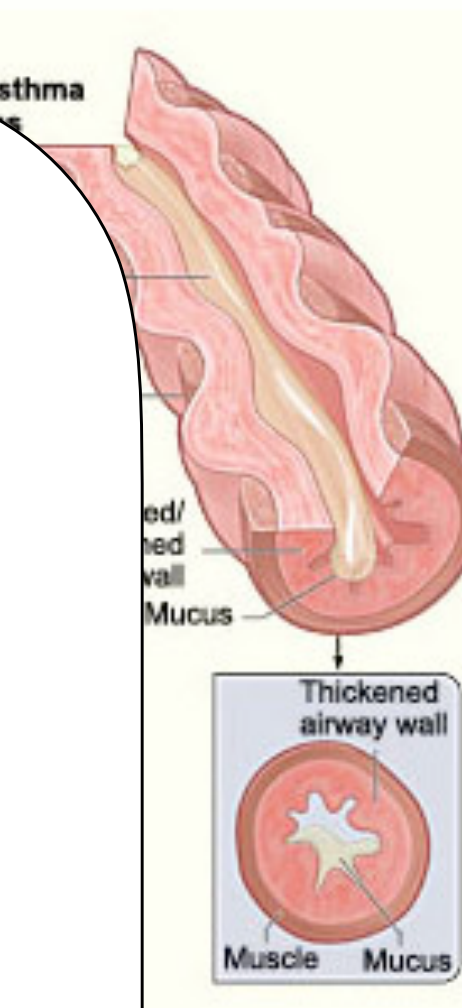
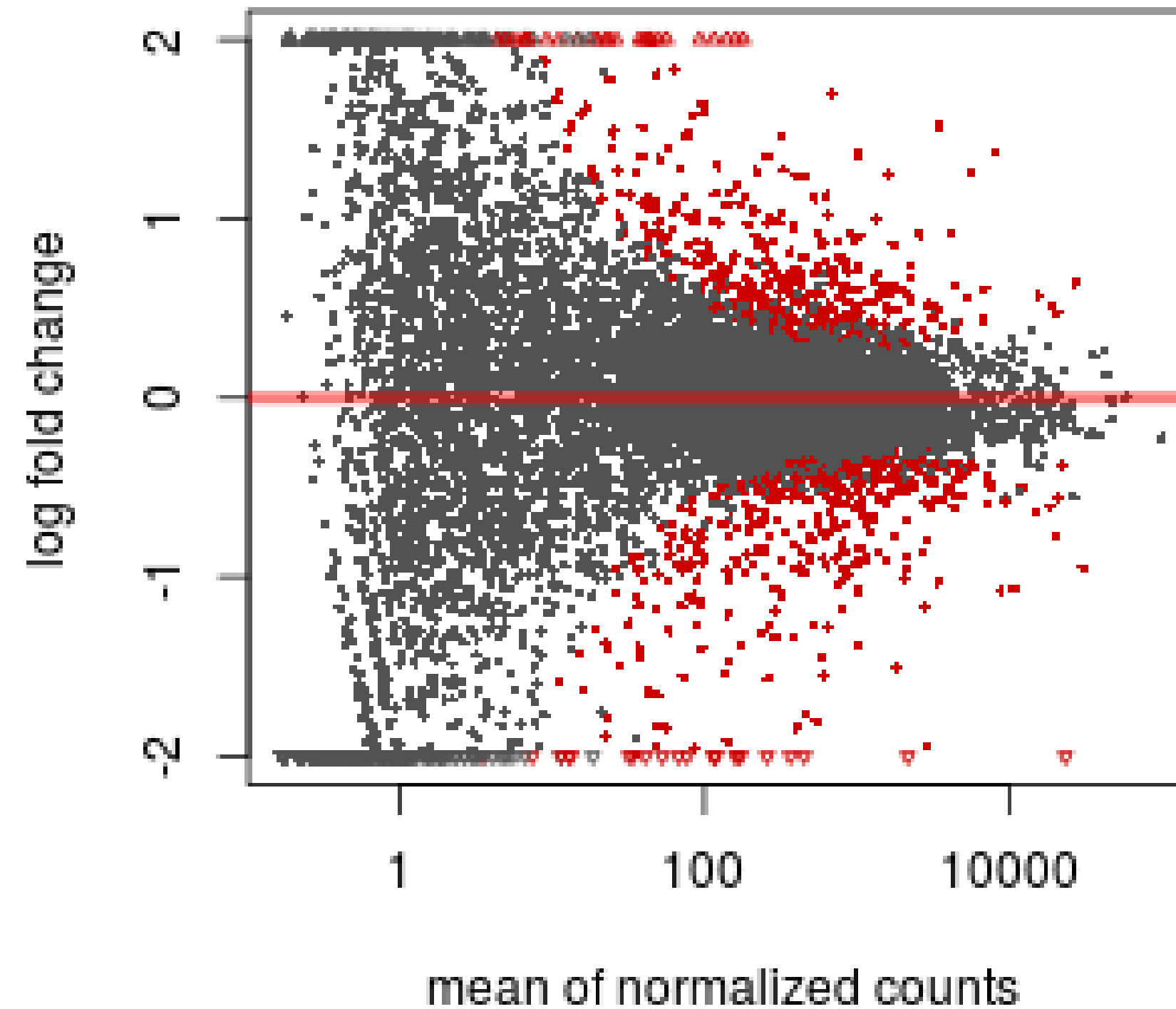
BAYESIAN STATISTICIAN:



Data set 1: RNA-Seq

Transcriptome
 samples
 smooth muscle
 dexamethasone
 glucocorticoid

cellline
 N61
 N61
 N05
 N05
 N08
 N080
 N061011
 N061011 trt



analysis:

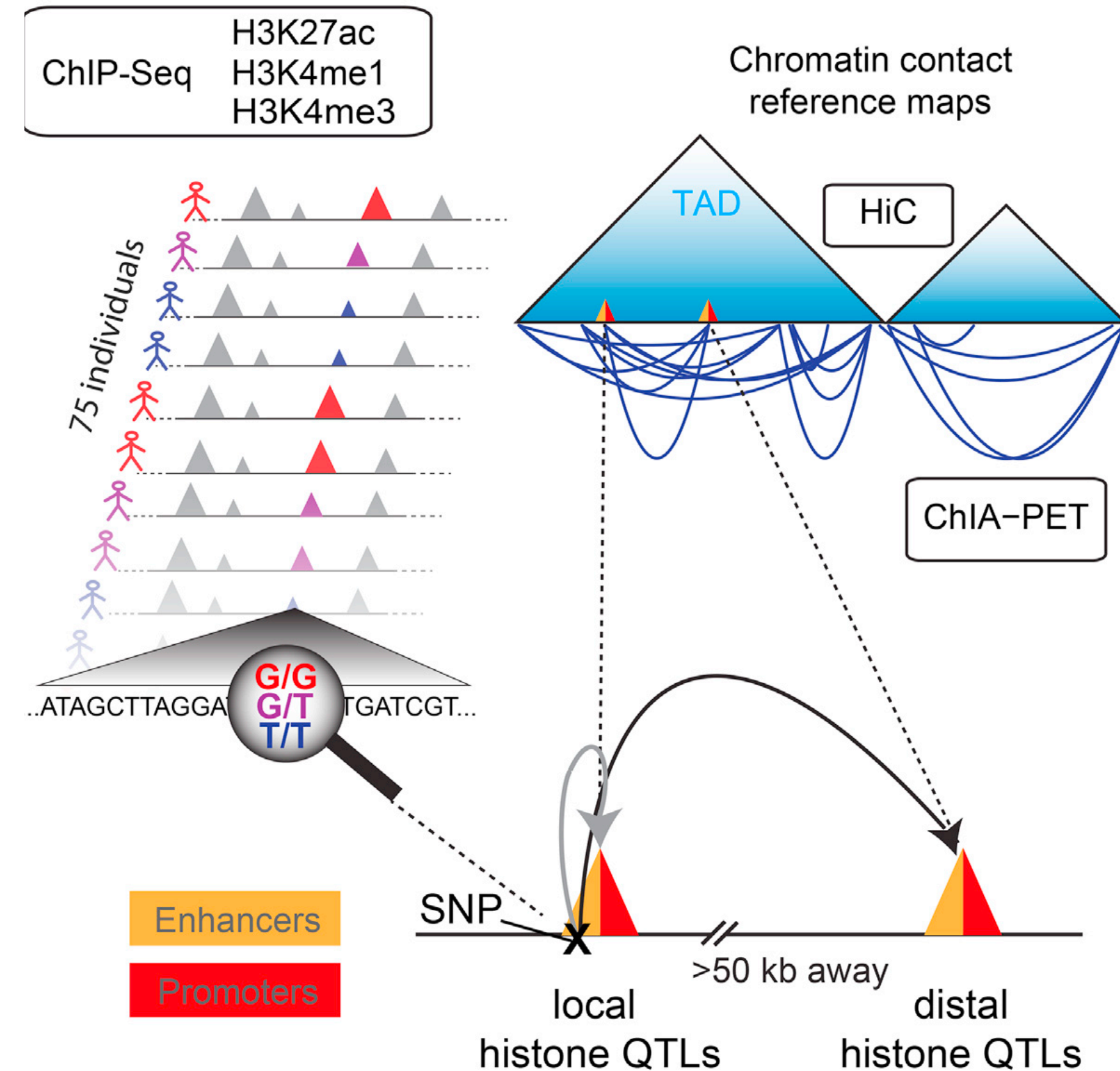
expression = α_j

design <- ~ cellline + dexamethasone

Himes et al. "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells." PLoS One. 2014 GEO: [GSE52778](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52778).

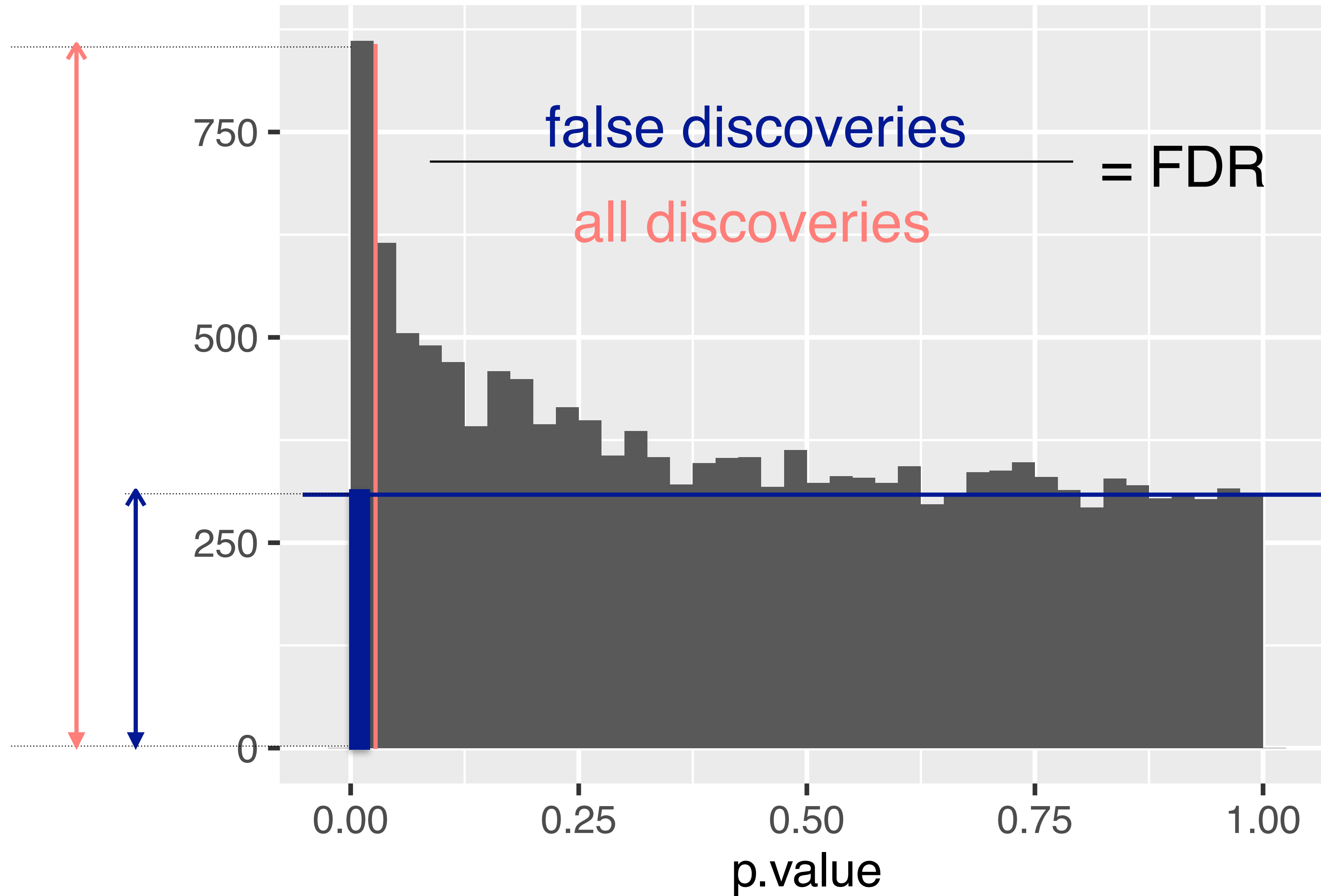
Data set 2: hQTL

ChIP-seq for histone marks in lymphoblastoid cell lines from 75 sequenced individuals. Local QTLs: find best-correlated SNP within 2kb of peak boundaries: 14,142 hQTLs, involving ~10% of all H3K27ac peaks (FDR=0.1, permutations)
Distal: distance cutoffs from 50 to 300 kb; also HiC



Grubert, Zaugg, Kasowski, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell (2015).

False Discovery Rate



Method of Benjamini & Hochberg (1995)

Method of Benjamini & Hochberg

0.100 -

```
BH = {  
  i <- length(p) : 1  
  o <- order(p, decreasing = TRUE)  
  ro <- order(o)  
  pmin(1, cummin(n/i * p[o]))[ro]  
}
```

takes a list of p-values as input and returns a matched list of 'adjusted' p-values.

0.000 -

0

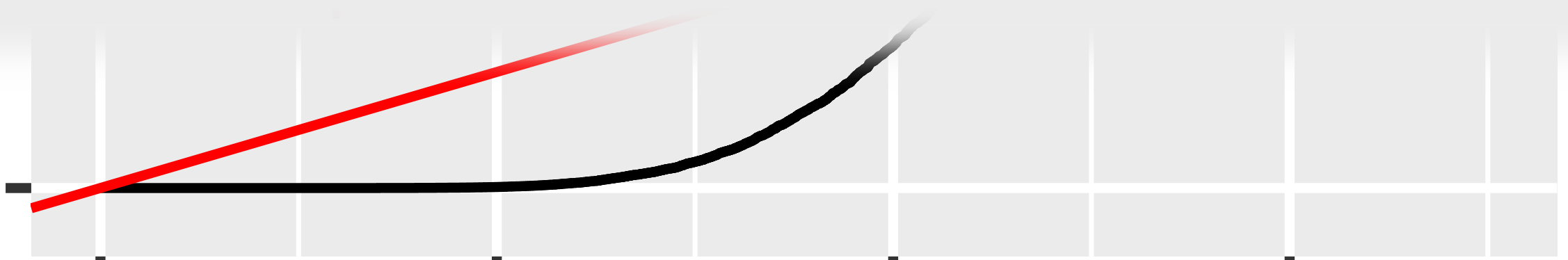
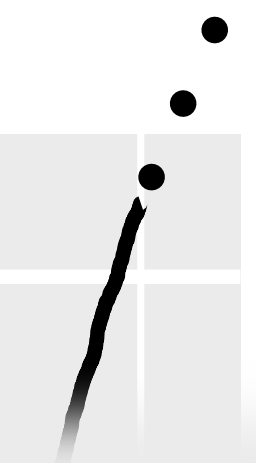
2000

4000

6000

rank

.....



Not all Hypothesis Tests are Created Equal

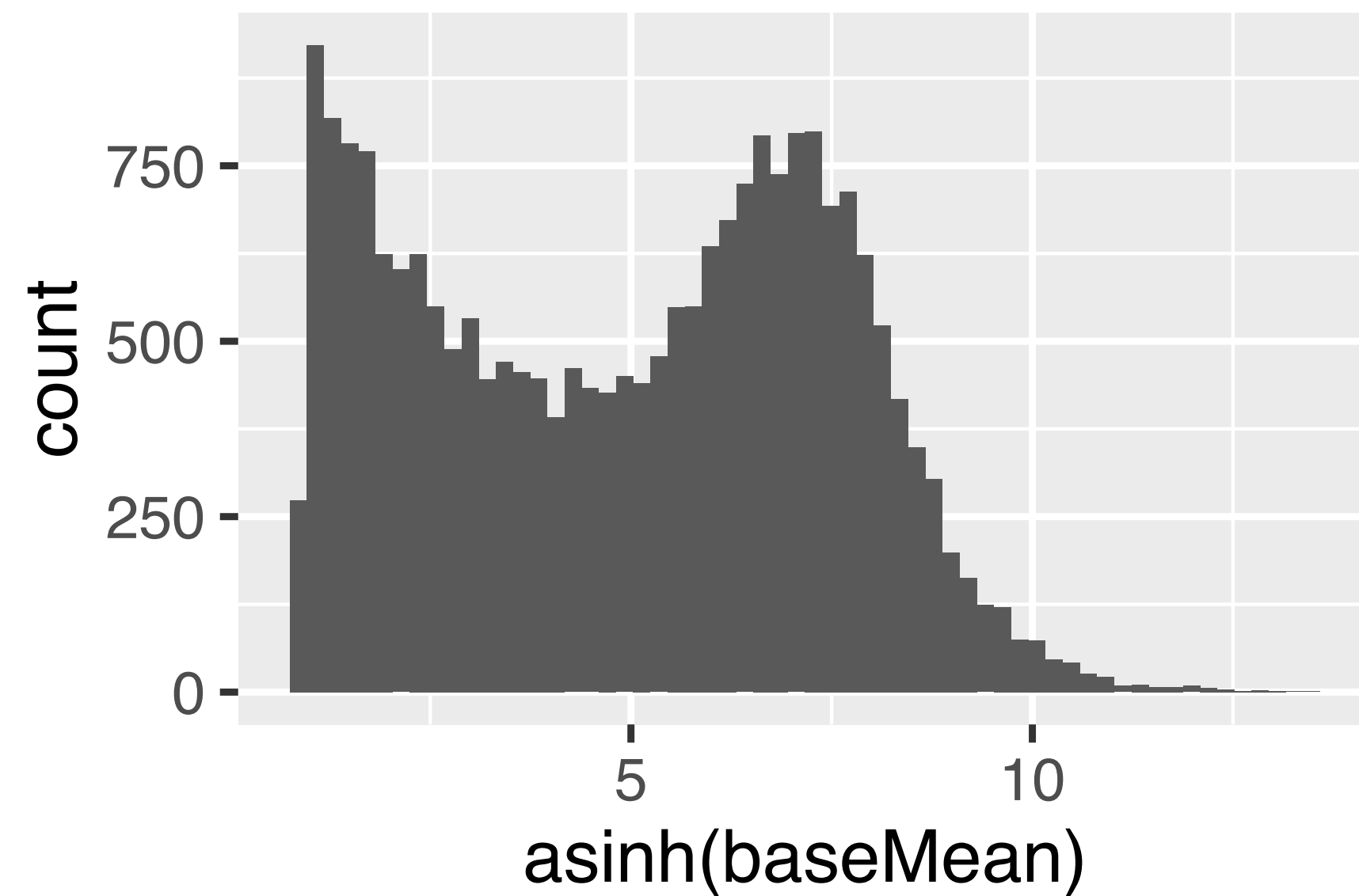
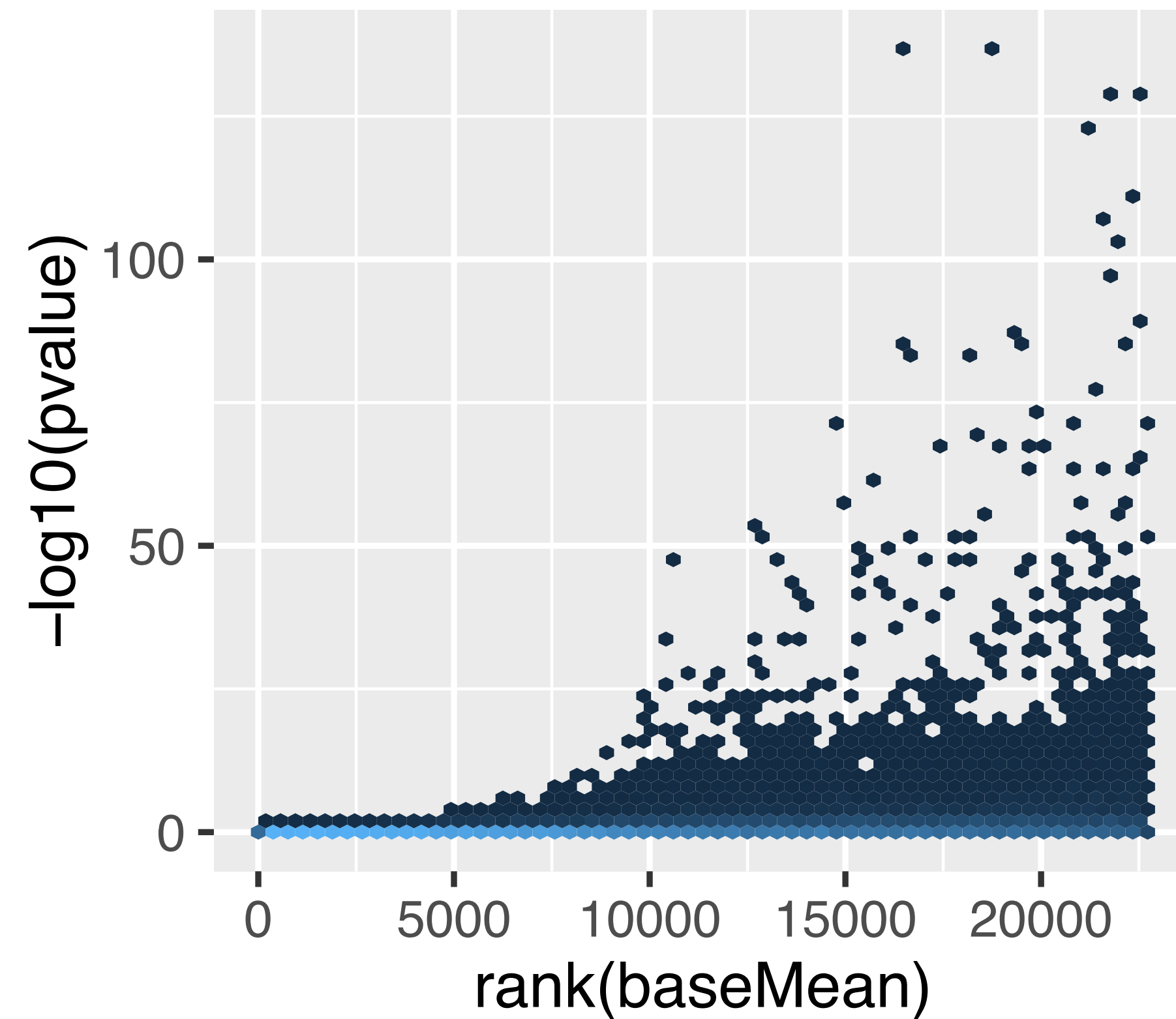


Figure 6.15: Histogram of baseMean. We see that it covers a large dynamic range, from close to 0 to around 3.3×10^5 .



Covariates - examples

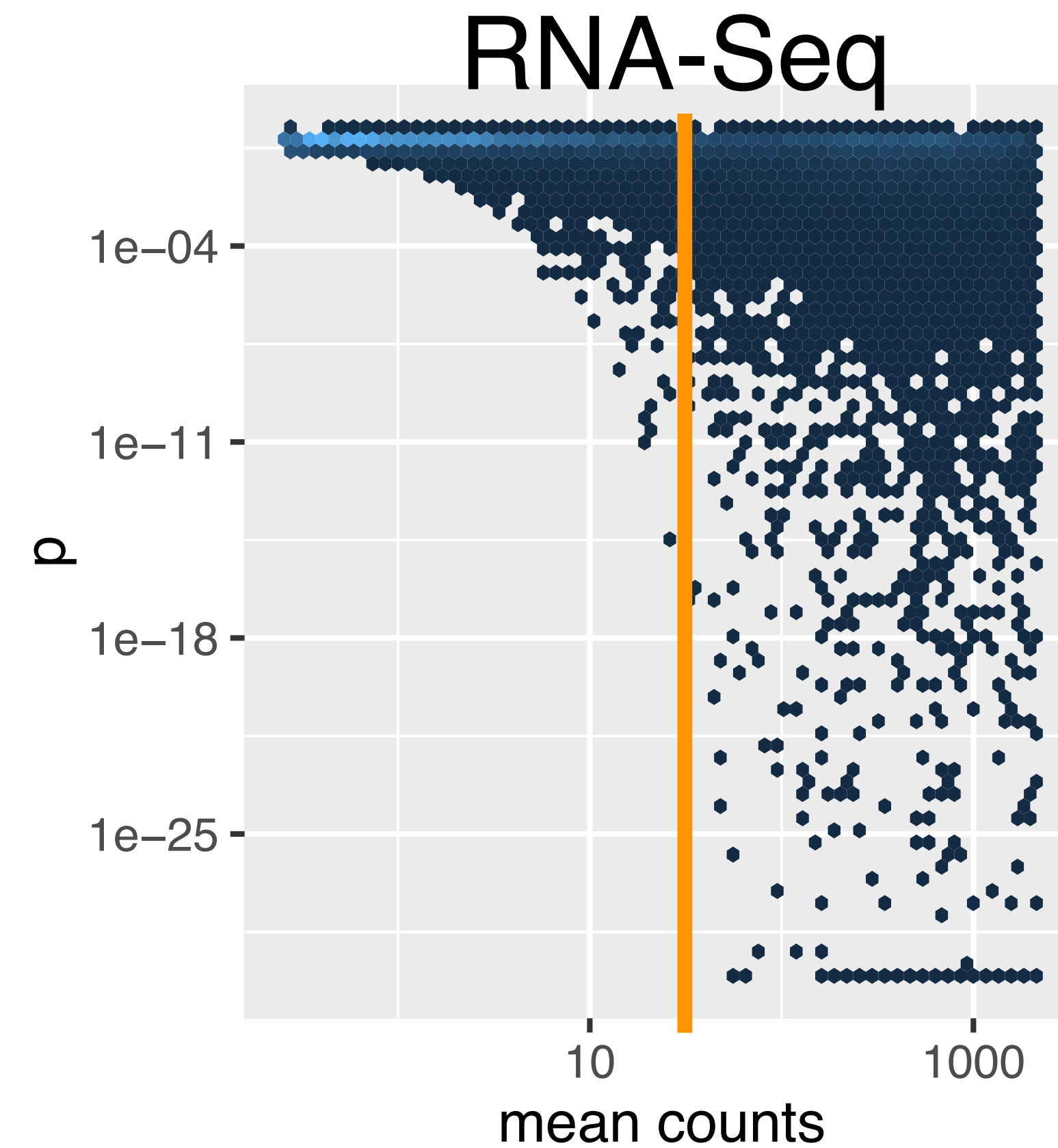
Application	Covariate
Differential RNA-Seq, ChIP-Seq, CLIP-seq, ...	(Normalized) mean of counts for each gene
eQTL analysis	SNP – gene distance
GWAS	Minor allele frequency
<i>t</i> -tests	Overall variance
Two-sided tests	Sign
All applications	Sample size; measures of signal-to-noise ratio

Independent Filtering

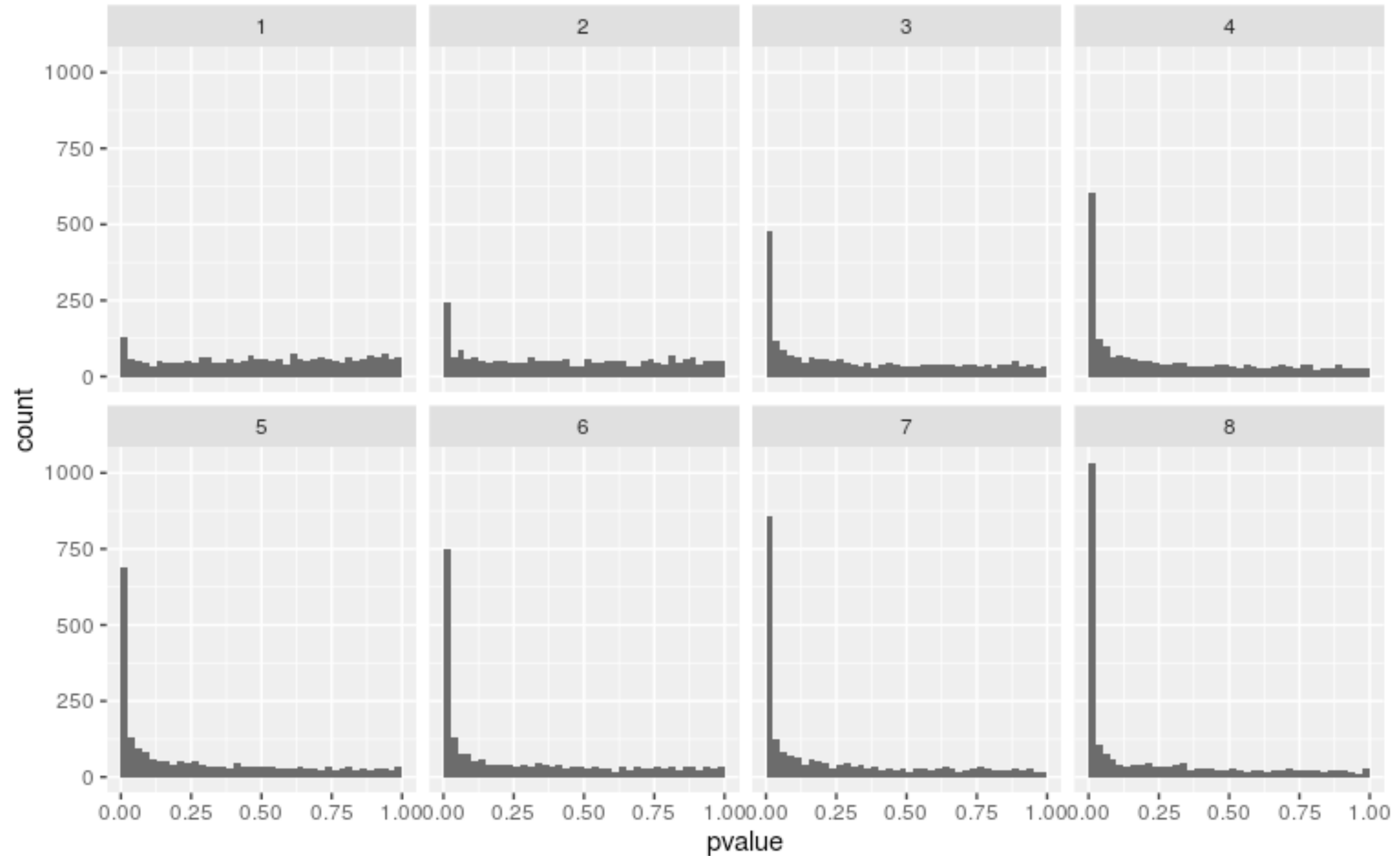
Two steps:

- All hypotheses H_i with $X_i < x$ get filtered.
- Apply BH to remaining hypotheses.

(Bourgon, Gentleman, Huber
PNAS 2010)



RNA-Seq p-value histogram stratified by average read count



Weighted Benjamini-Hochberg method

- Let $w_i \geq 0$ and $\frac{1}{m} \sum w_i = 1$ (budget”).
- Define $Q_i = \frac{p_i}{w_i}$
- Apply BH
- Proven
- Wasserman
- If $w_i > 1$,
- $Q_i \leq t \Leftrightarrow$

Problem: how to know the weights?



der,

Independent hypothesis weighting (IHW): basic idea

hypothesis weighting

- Stratify the tests into G bins, by covariate X
- Choose α
- For each possible weight vector $\mathbf{w} = (w_1, \dots, w_G)$ apply weighted BH procedure. Choose \mathbf{w} that maximizes the number of rejections at level α .
- Report the result with the optimal weight vector \mathbf{w}^* .



Nikos Ignatiadis

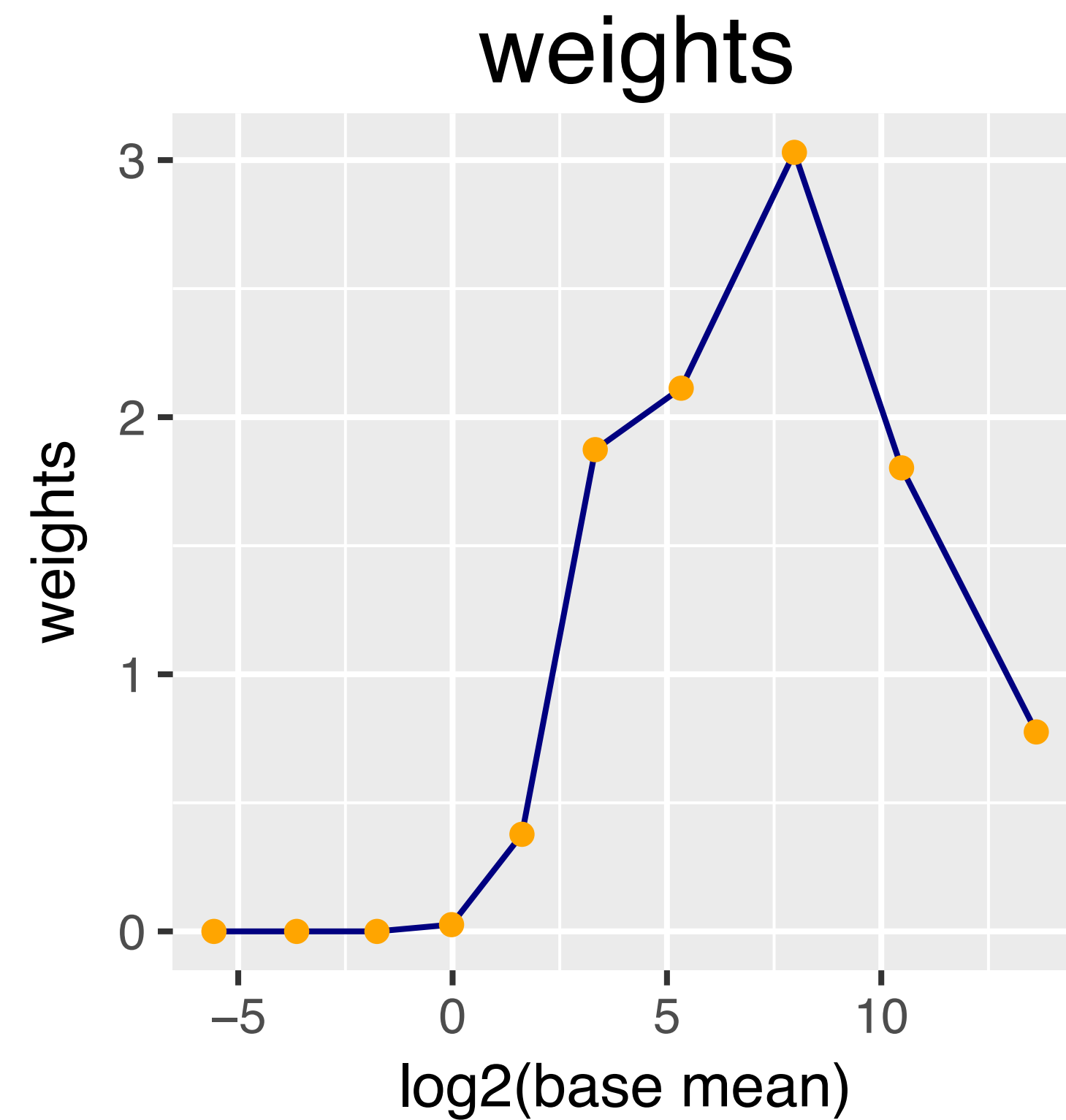
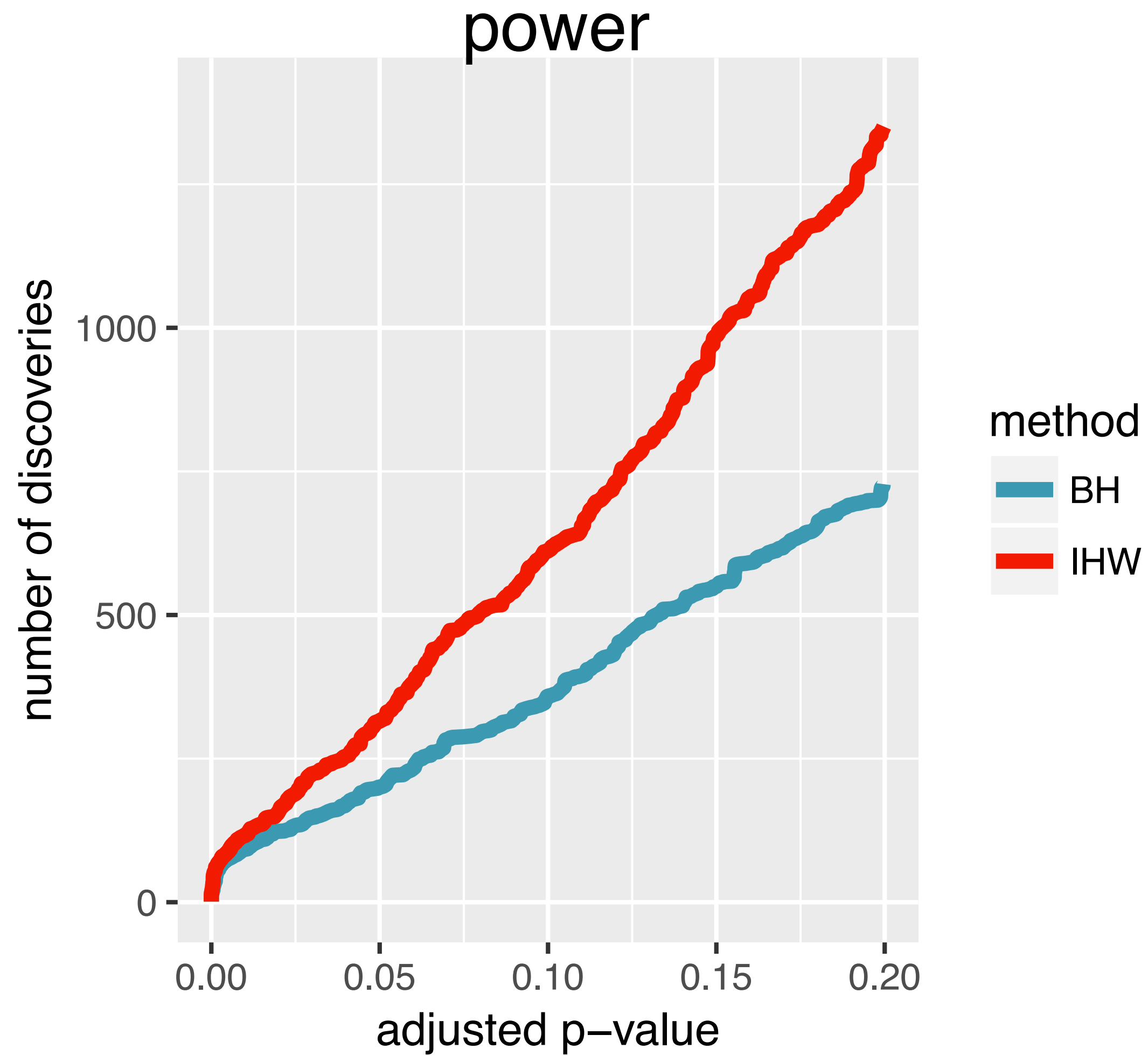
Ignatiadis et al.,

• Nature Methods 2016, DOI10.1038/nmeth.3885

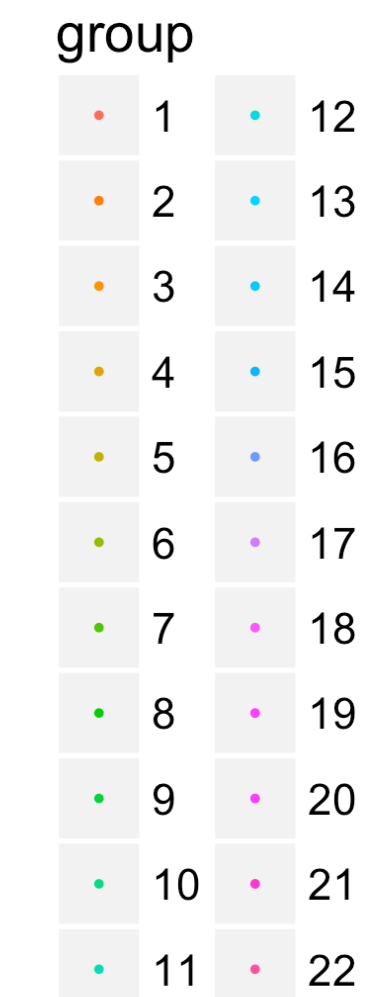
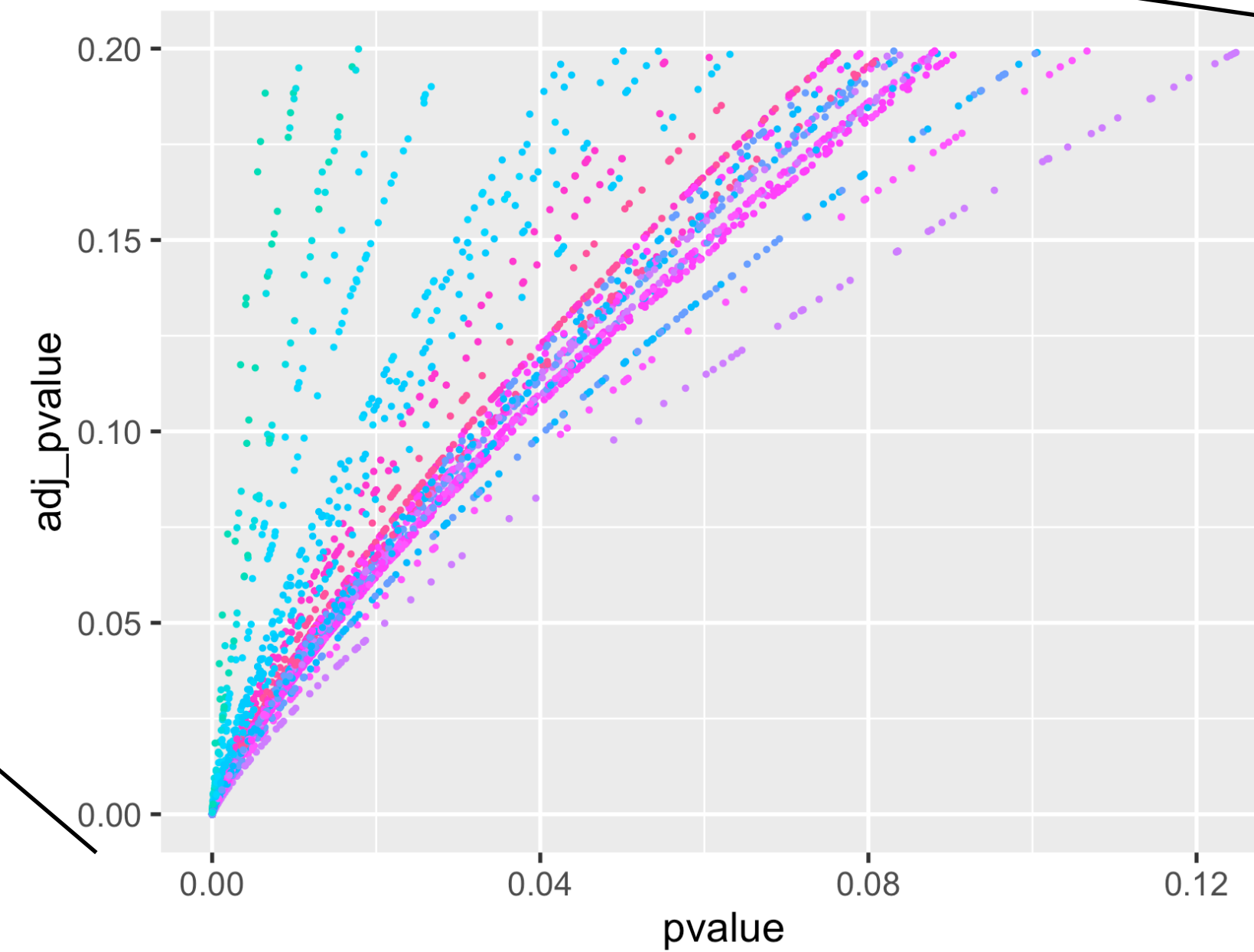
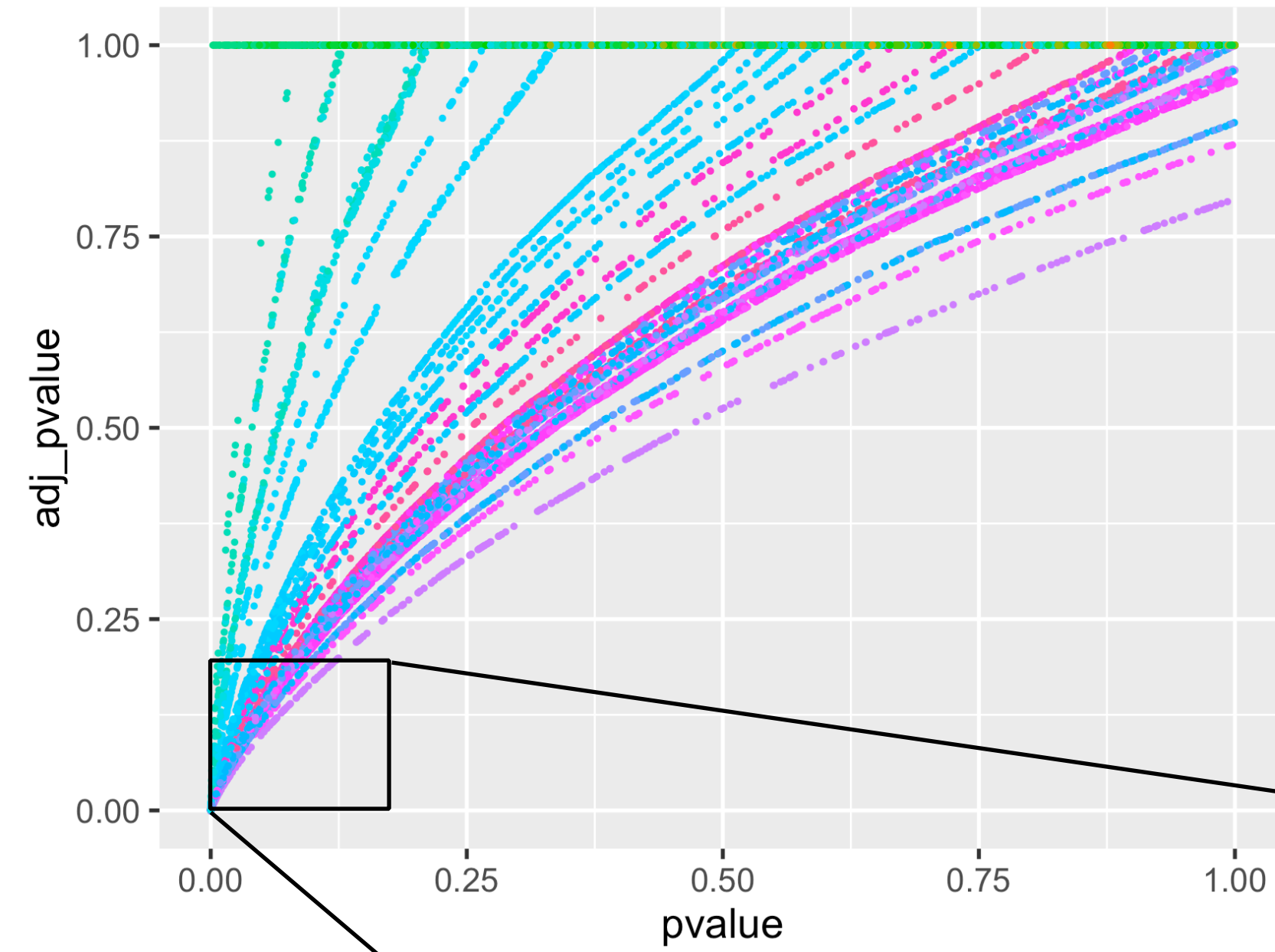
• arXiv:1701.05179

Bioconductor package IHW

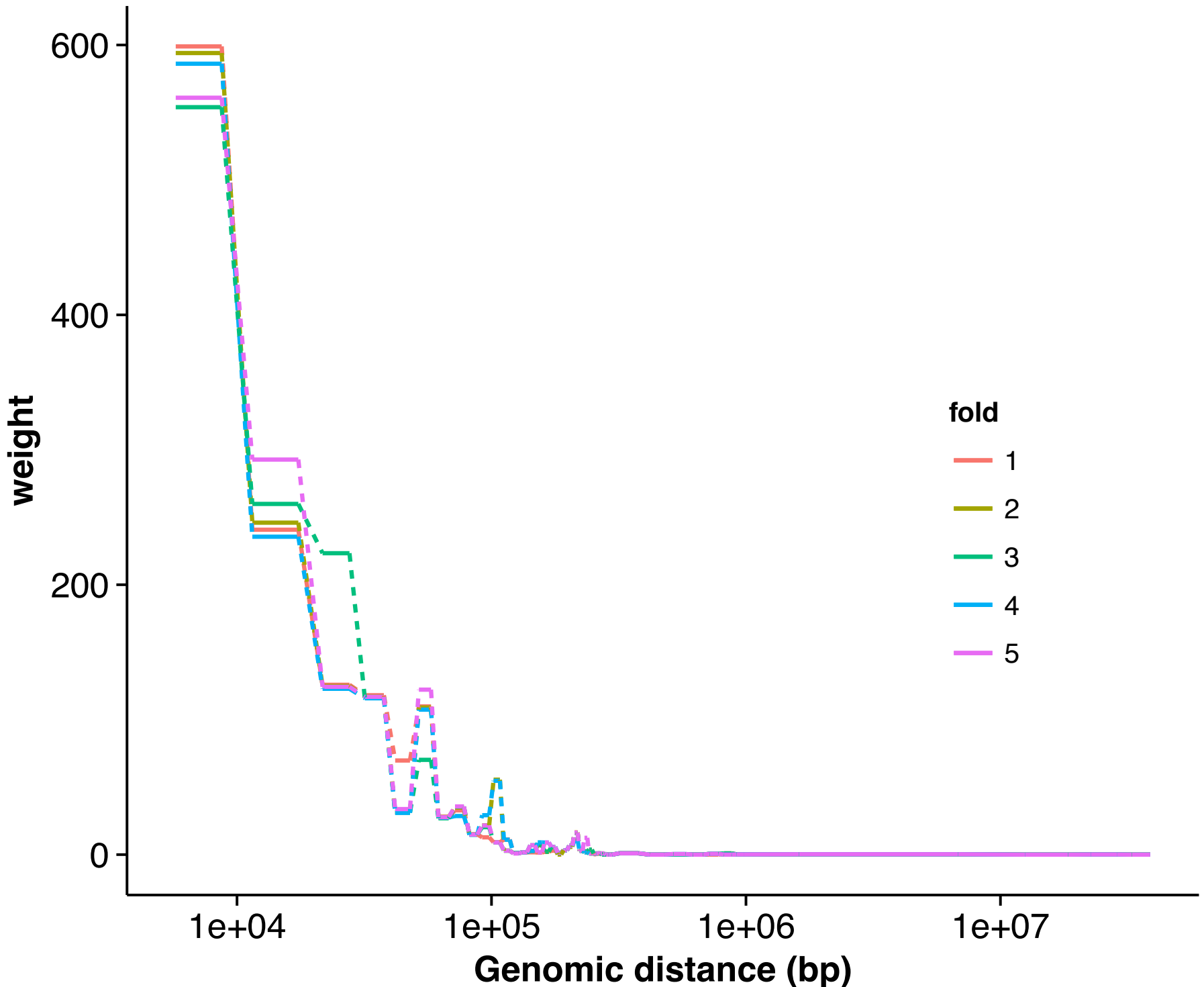
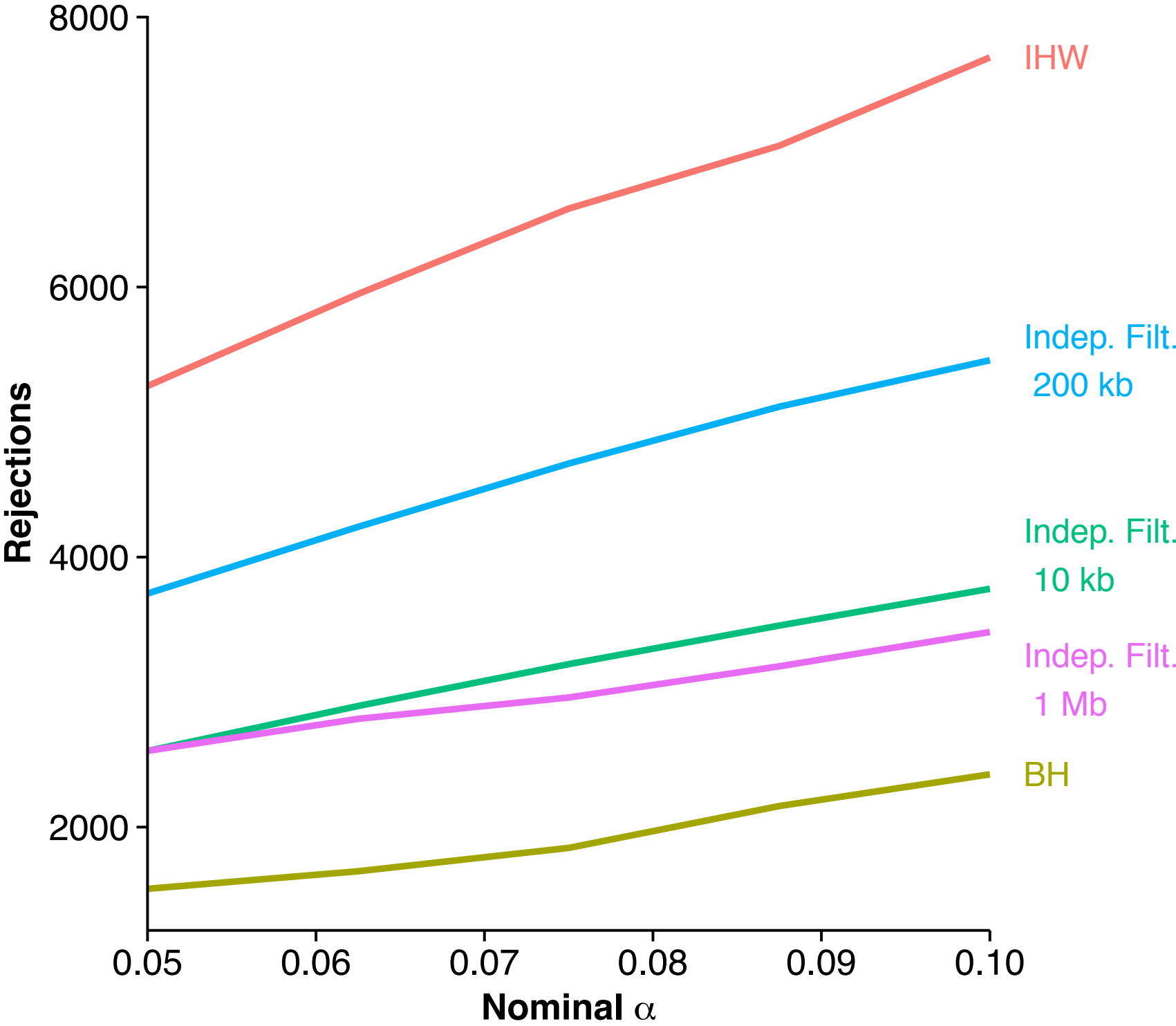
RNA-Seq example (DESeq2)



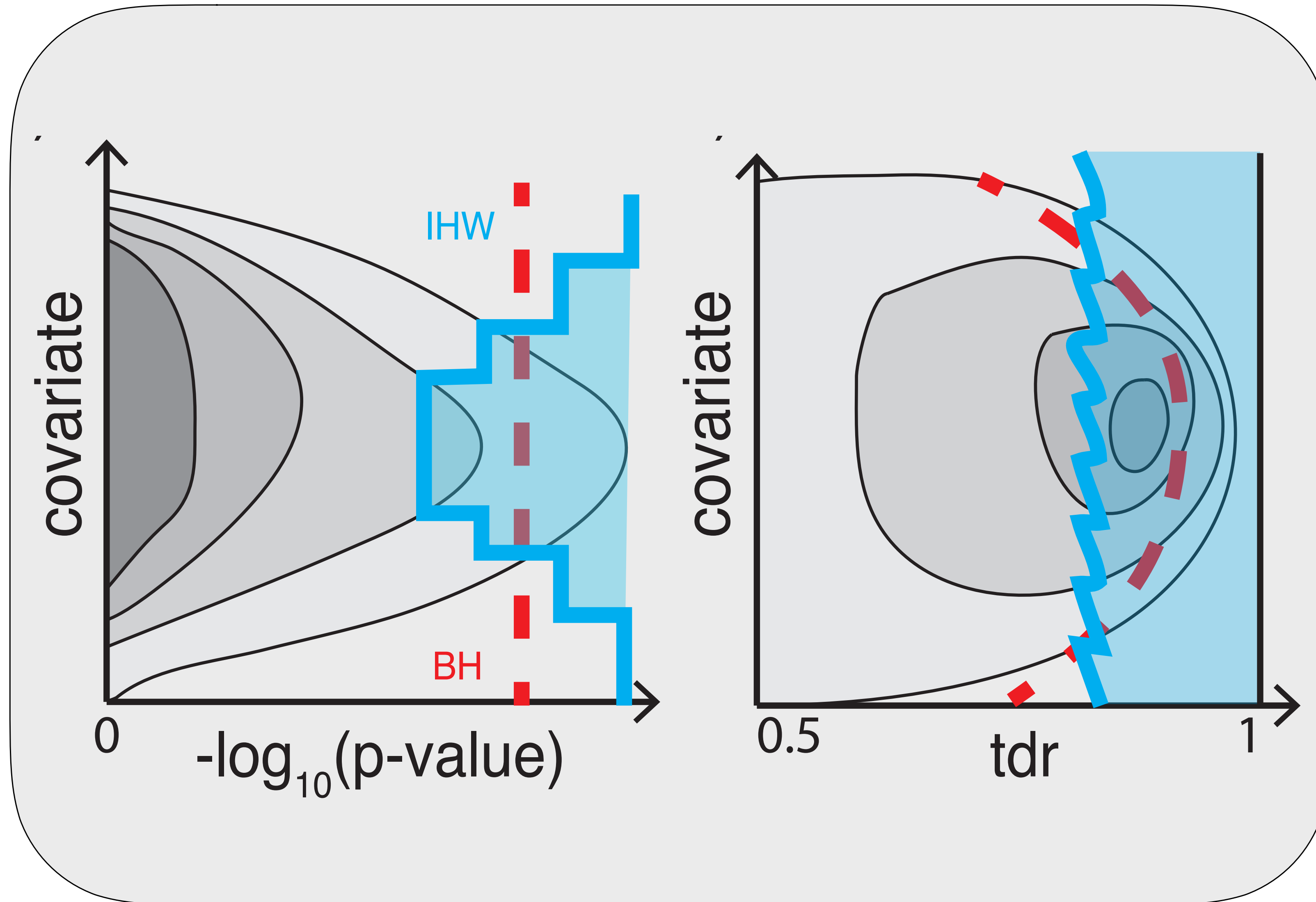
Ranking is not monotonous in raw p-values



Histone-QTL example (H3K27ac)



2D decision boundaries



Summary

- Multiple testing is not a problem but an opportunity
- Heterogeneity across tests
- Informative covariates are often apparent to domain scientists
 - independent of test statistic under the null
 - informative on π_1 , F_{alt}
- Data-driven weighting
- Scales well to millions of hypotheses
- Controlling ‘overoptimism’

A promotional image for the James Bond film 'Casino Royale'. It features Daniel Craig as James Bond, wearing a grey three-piece suit, a white shirt, and a patterned tie. He is looking upwards and to the right with a serious expression. The background is a mix of blue and orange, suggesting a sky and fire. The text 'The p-value Is Not Enough 007' is overlaid on the image.

The p-value Is Not Enough
007

P-VALUE

INTERPRETATION

0.001]— HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04]— SIGNIFICANT
0.049	
0.050]— OH CRAP. REDO CALCULATIONS.
0.051]— ON THE EDGE OF SIGNIFICANCE
0.06	
0.07]— HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099]— HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	