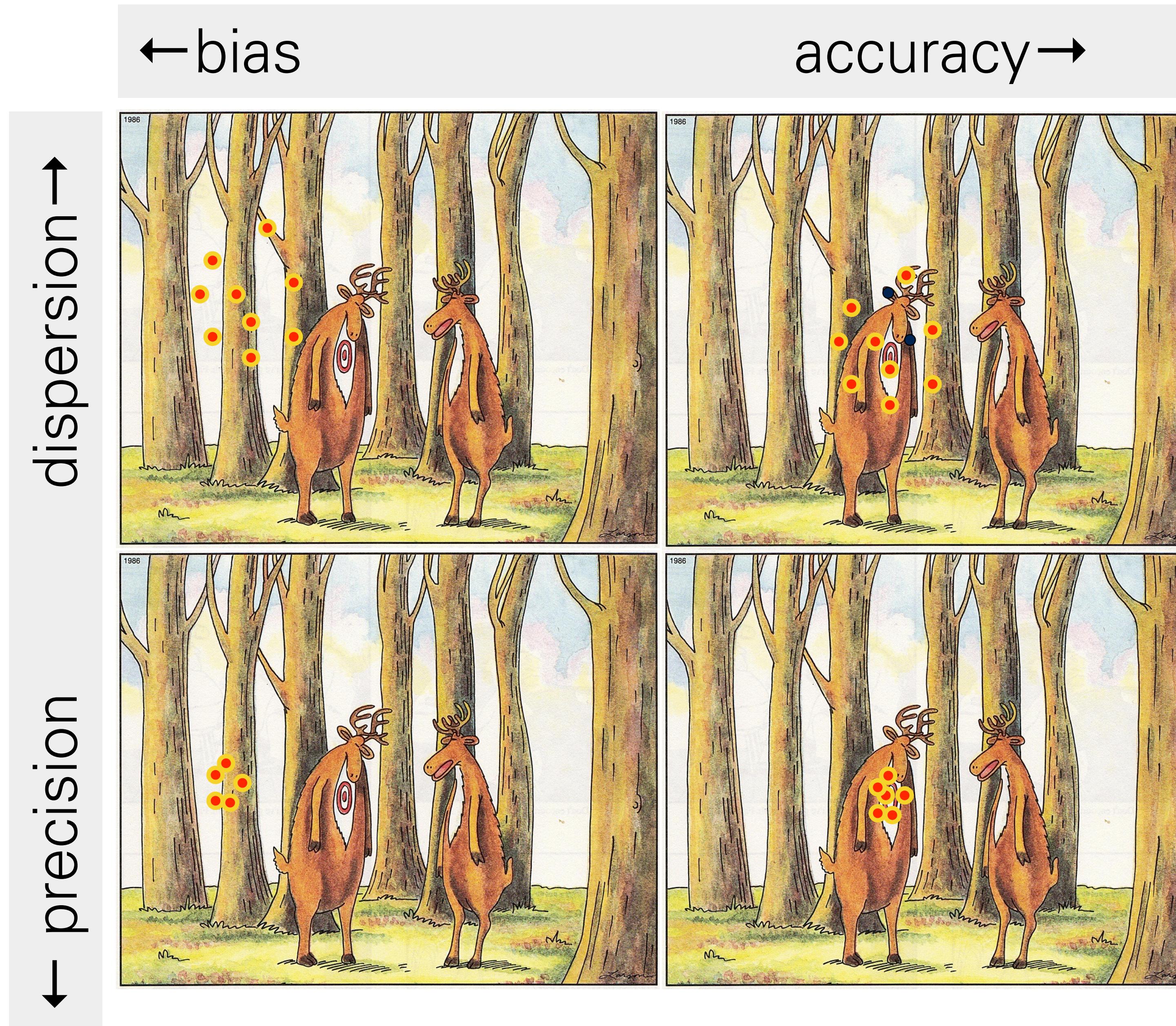


Hypothesis Testing

Wolfgang Huber



The fundamental tradeoff of statistical decision making



Comes in various guises

Accuracy vs Precision

Bias vs Variance

Model complexity vs overfitting

How to make rational decisions based on noisy, finite data?

Examples:

- Testing efficacy of a drug on people
 - lack of complete experimental control
 - finite sample size
- Prioritising the results from a biological high-throughput experiment (e.g. RNA-seq differential expression; CRISPR screen)
-

+: No understanding of mechanism involved / needed / desired

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

How to make rational decisions based on noisy, finite data?

Examples:

- Testing efficacy of a drug
 - lack of complete data
 - finite sample size
- Prioritising the results of RNA-seq differential expression analysis
-

- +**: No understanding of the data
- : Wouldn't we want to know more?

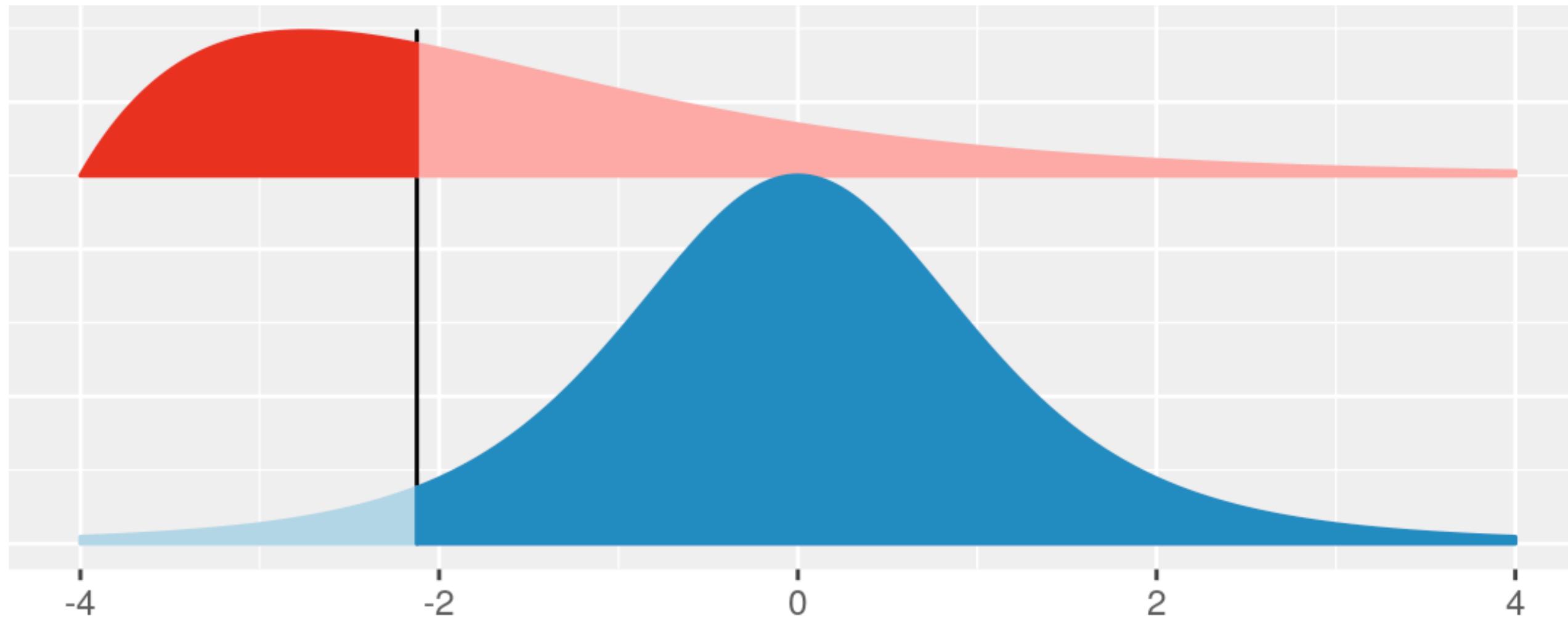


Any fool can work with infinite data.
Statisticians do finity.

Basic problem: binary decision



Drug works



some useful number x computed from the data

True Positive False Negative False Positive True Negative

False discovery rate

$$\text{FDR} = \frac{\text{area shaded in light blue}}{\text{sum of the areas left of the vertical bar (light blue + strong red)}}$$

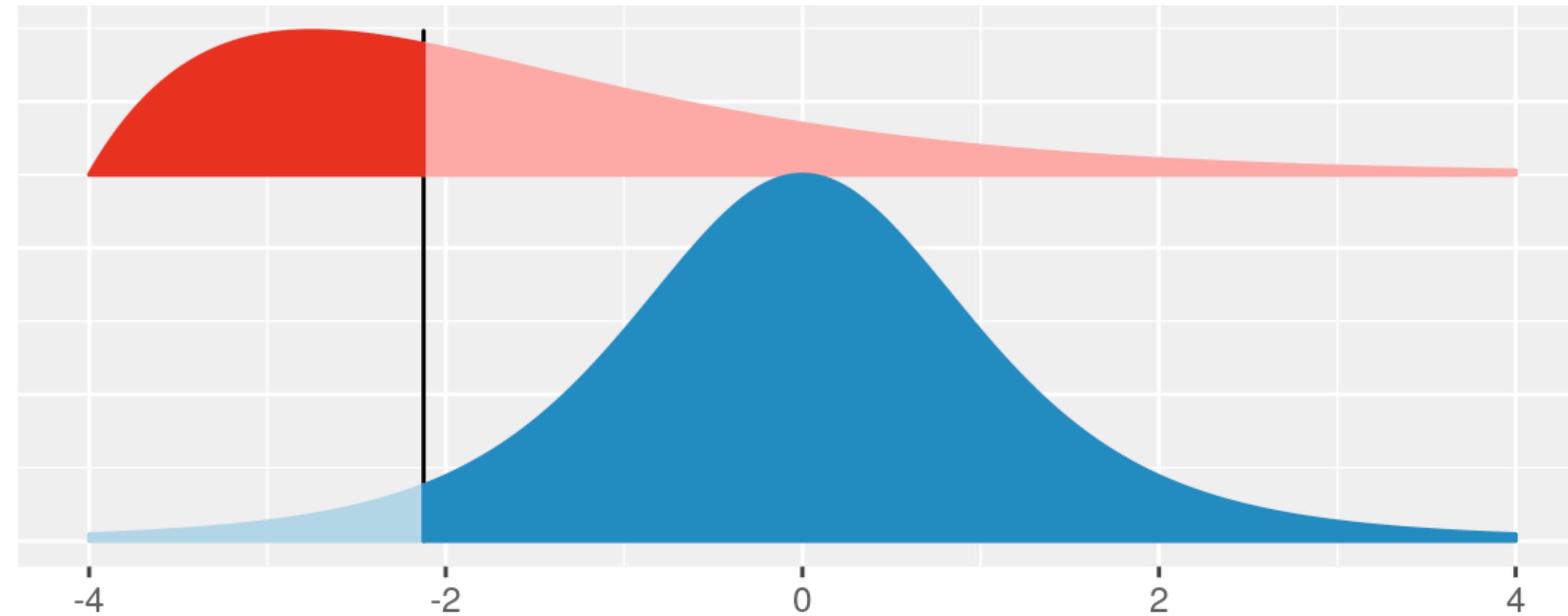


It's bad

Basic problem: binary decision



Drug works



True Positive False Negative False Positive True Negative

False discovery rate

$$\text{FDR} = \frac{\text{area shaded in light blue}}{\text{sum of the areas left of the vertical bar (light blue + strong red)}}$$



It's bad

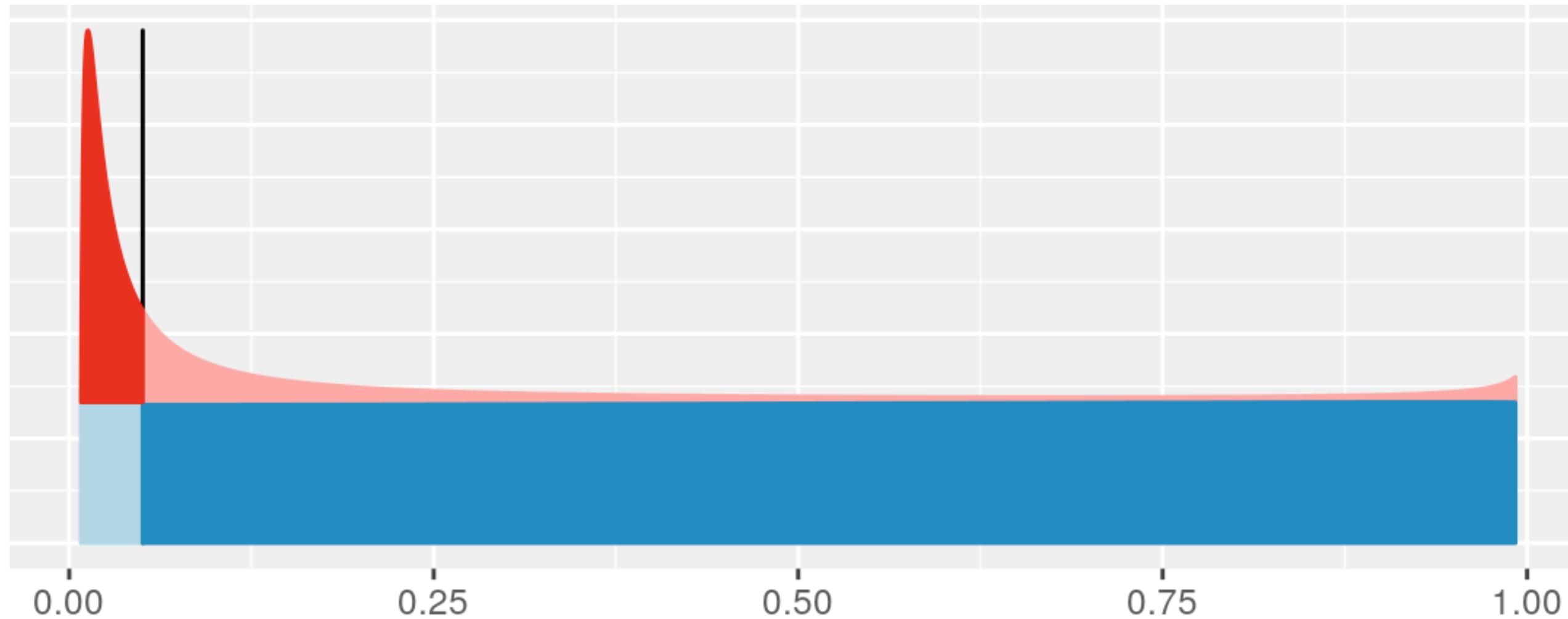
For this, we need to know:

1. the distribution of x in the blue class (the blue curve),
2. the distribution of x in the red class (the red curve),
3. the relative sizes of the blue and the red classes.

Basic problem: binary decision



Drug works



another useful number computed from the data: p

True Positive False Negative False Positive True Negative



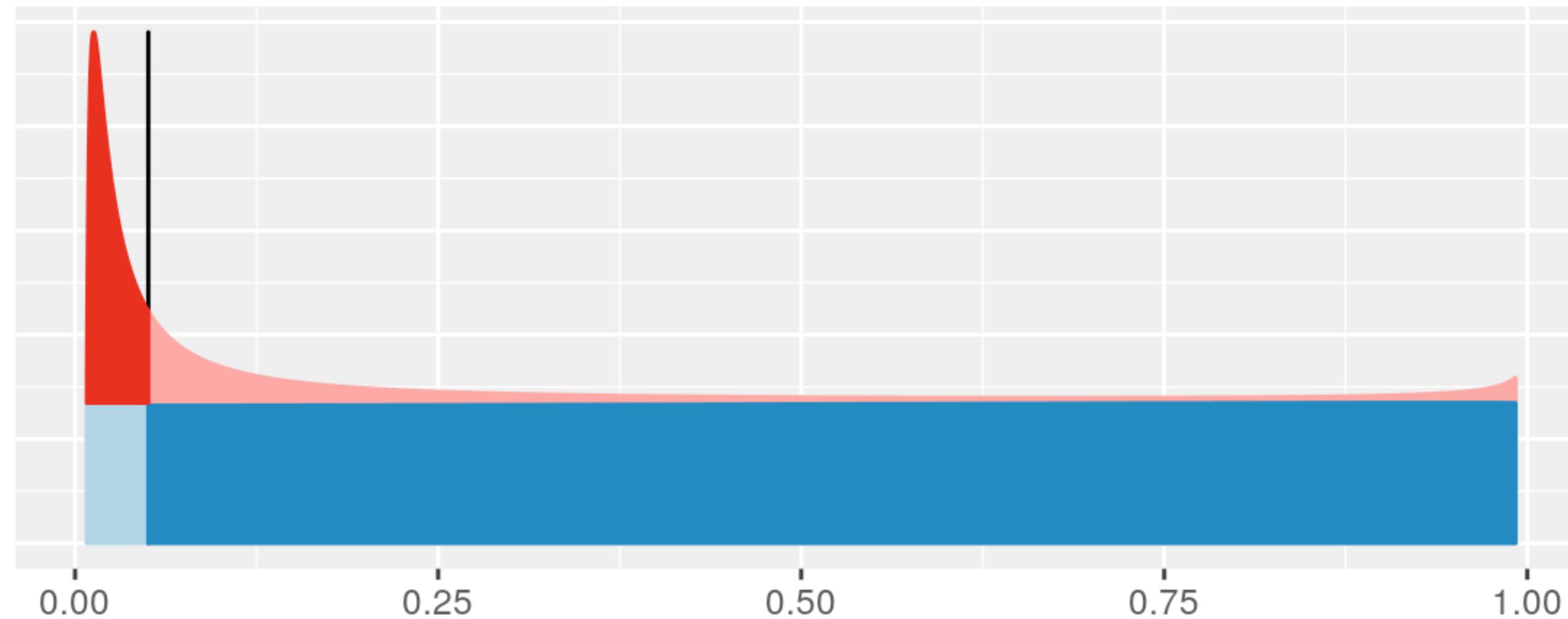
It's bad

$$p\text{-value} = \frac{\text{area shaded in light blue}}{\text{overall blue area}}$$

Basic problem: binary decision



Drug works



another useful number computed from the data: p

True Positive False Negative False Positive True Negative



It's bad

$$p\text{-value} = \frac{\text{area shaded in light blue}}{\text{overall blue area}}$$

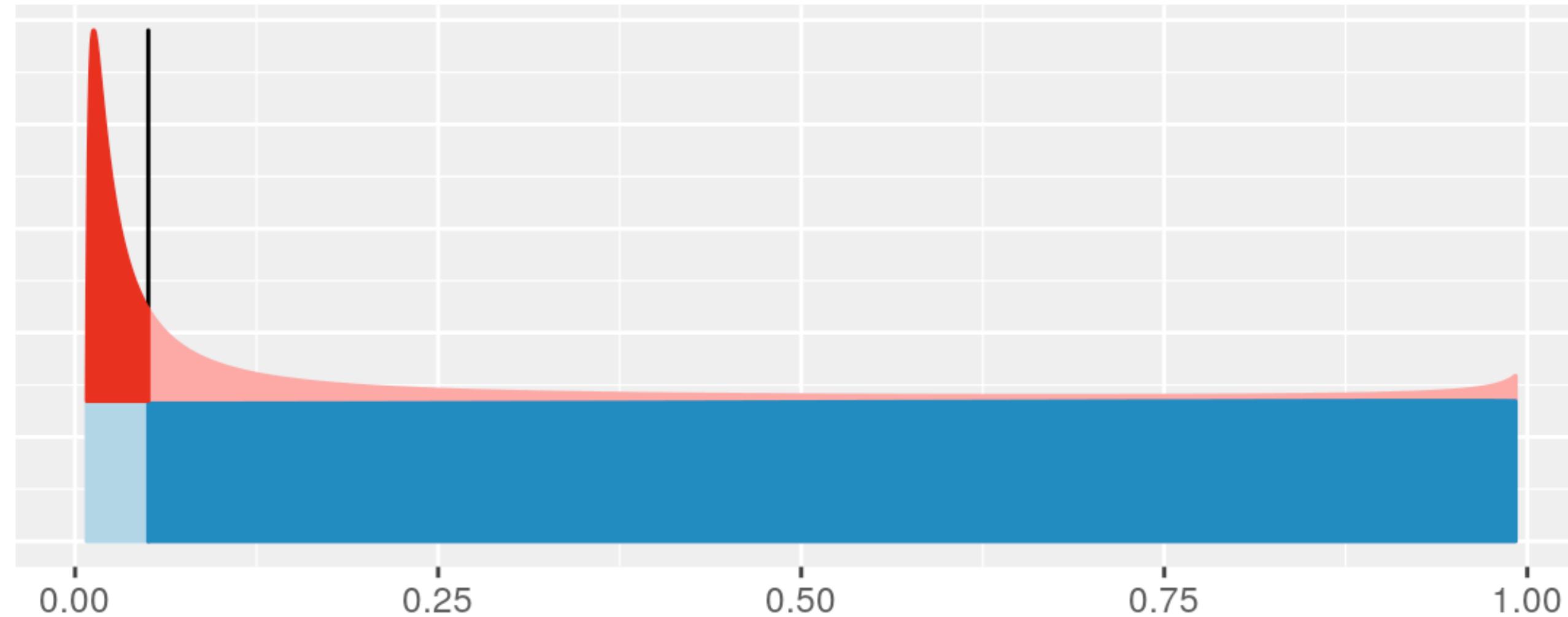
For this, we need to know:

1. the distribution of x in the blue class ("null hypothesis").
2. the distribution of x in the red class (the red curve),
3. the relative sizes of the blue and the red classes.

Basic problem: binary decision



Drug works



another useful number computed from the data: p

True Positive False Negative False Positive True Negative

$$p\text{-value} = \frac{\text{area shaded in light blue}}{\text{overall blue area}}$$

For this, we need to know:

1. the distribution of x in the blue class ("null hypothesis").
2. the distribution of x in the red class (the red curve),
3. the relative sizes of the blue and the red classes.



It's bad

What could possibly go wrong? What's the difference between p-value and FDR?

Machine Learning

Lots of free parameters

Lots of training data

Using multiple variables

... or objects that are not even
traditional variables (e.g. images)

Hypothesis testing

Some theory/model and no or few
parameters

No training data

More rigid/formulaic
Regulatory use



Machine Learning

Lots of free parameters

Lots of training data

Using multiple variables

... or objects that are not even
traditional variables (e.g. images)

Hypothesis testing

Some theory/model and no or few
parameters

No training data

More rigid/formulaic
Regulatory use



Count data: binomial test

Toss a coin a number of times ⇒

If the coin is fair, then heads $\sim 1/2$ the time.



But what is “roughly”?

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

We can use combinatorics / probability theory to quantify this.

Continuous data: the t-statistic

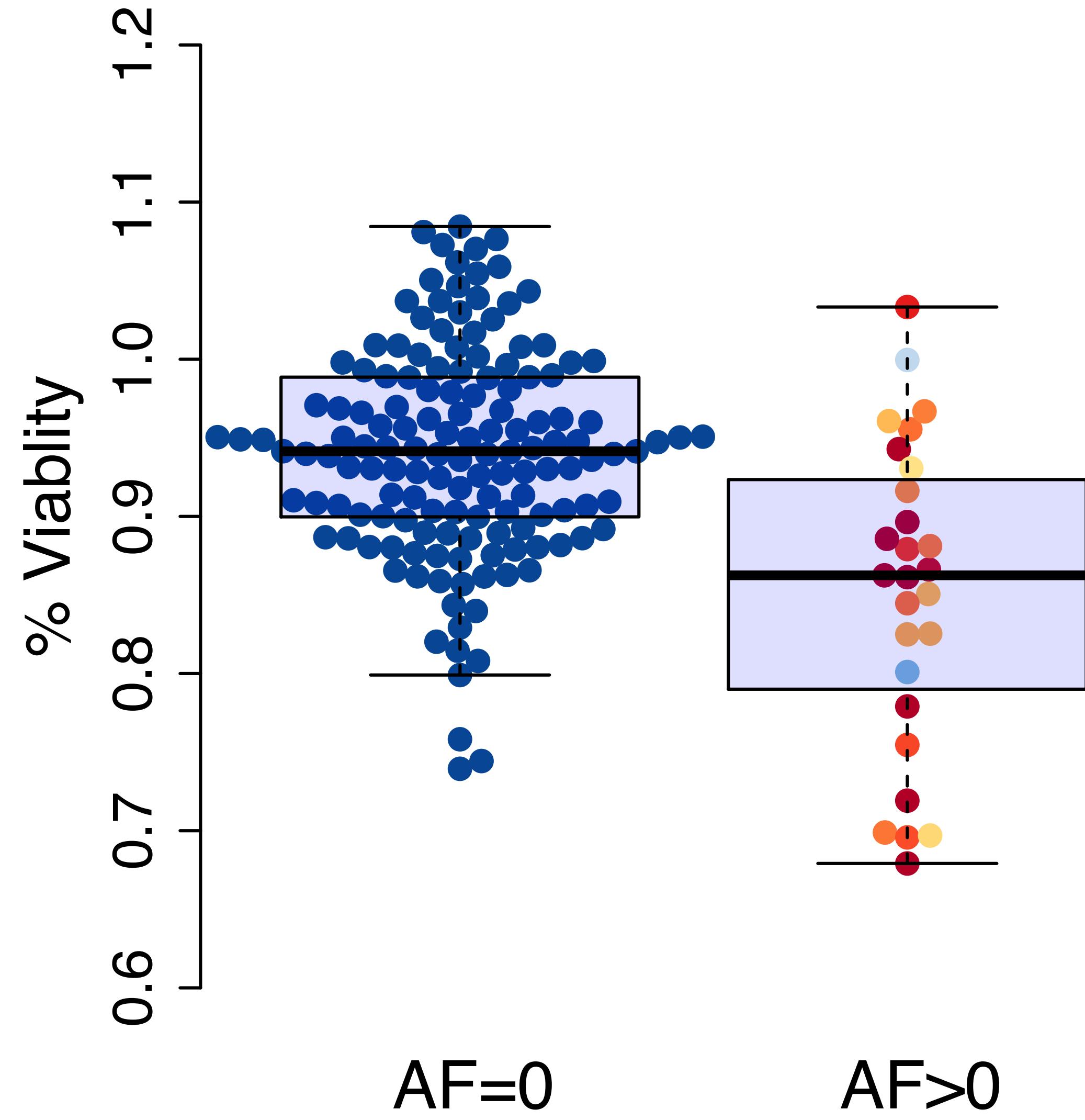
$$t = c \frac{m_1 - m_2}{s}$$

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

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

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

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



Binomial distribution

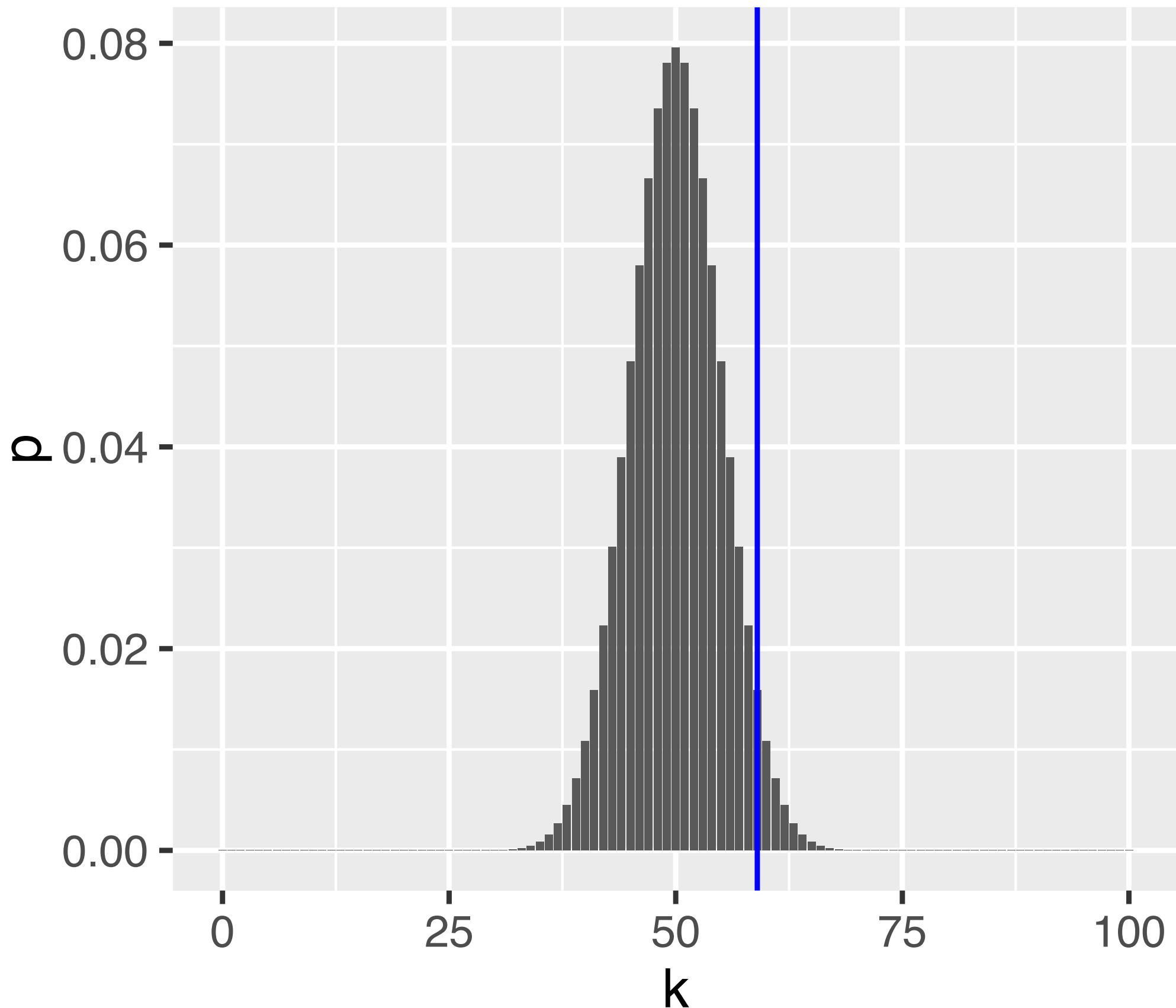


Figure 6.3: The binomial distribution for the parameters $n = 100$ and $p = 0.5$,

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

Rejection region

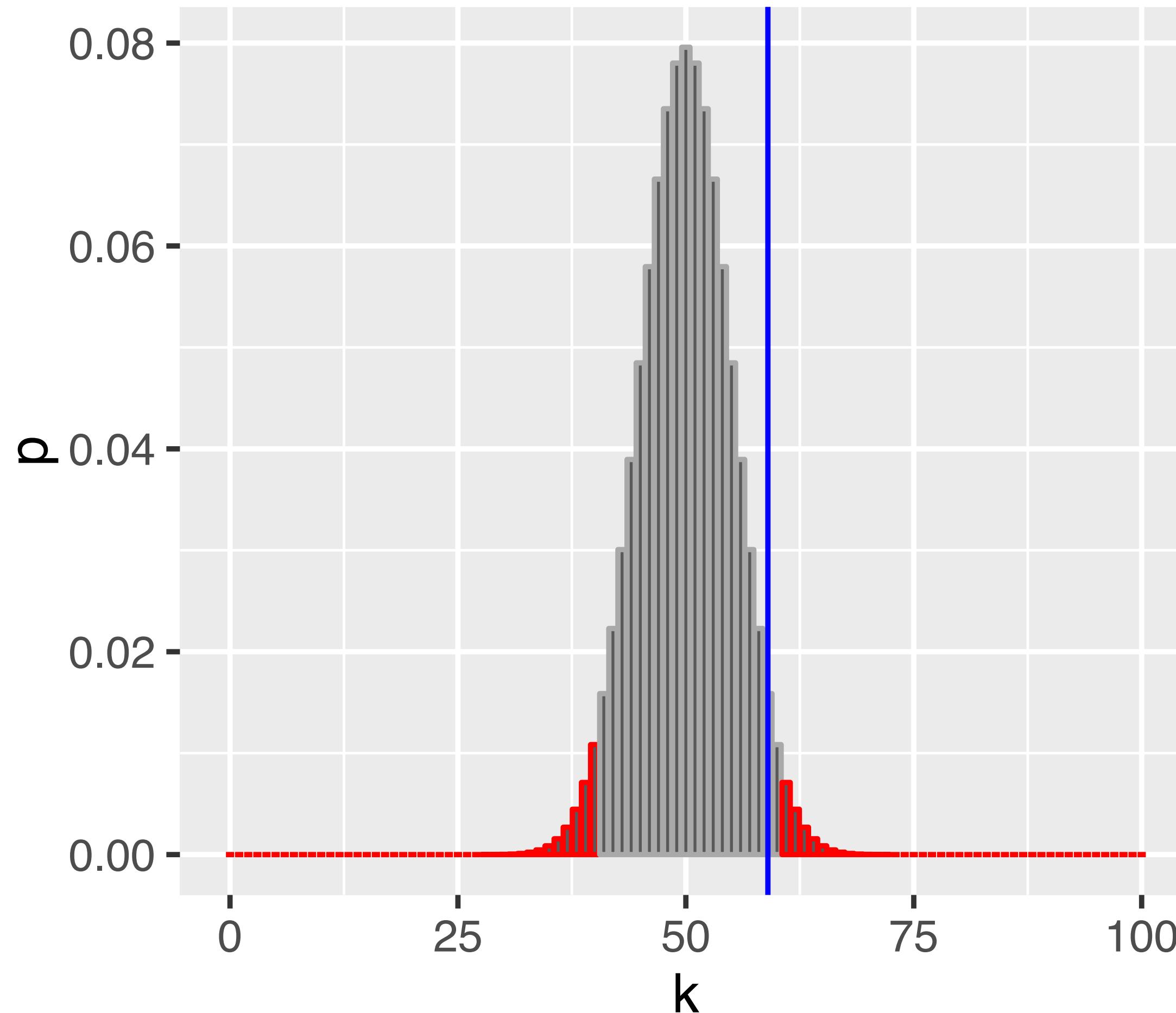


Figure 6.5: As Figure 6.3, with rejection region (red) whose total area is $\alpha = 0.05$.

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 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

Parametric Theory vs Simulation

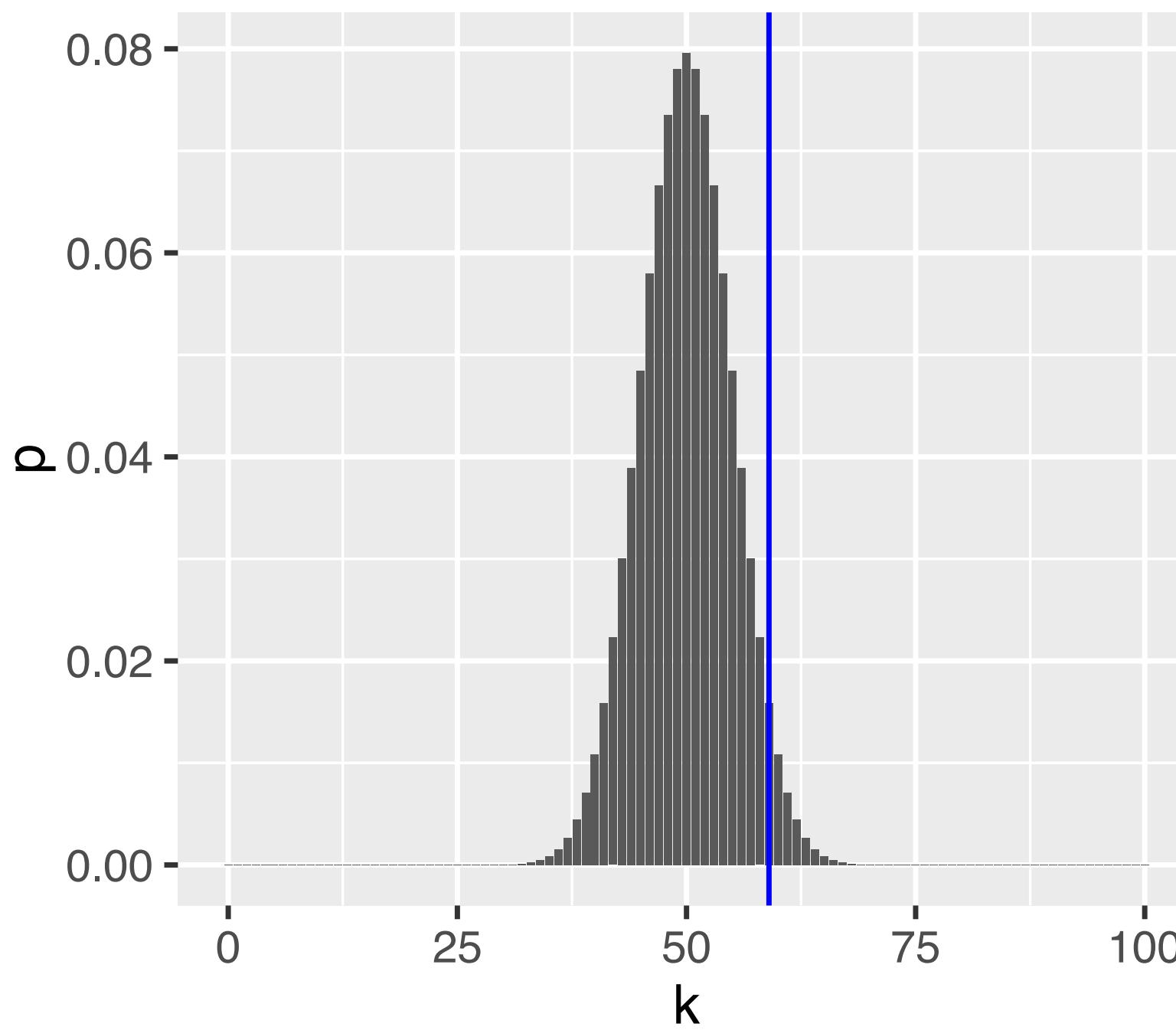


Figure 6.3: The binomial distribution for the parameters $n = 100$ and $p = 0.5$, according to Equation (6.1).

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

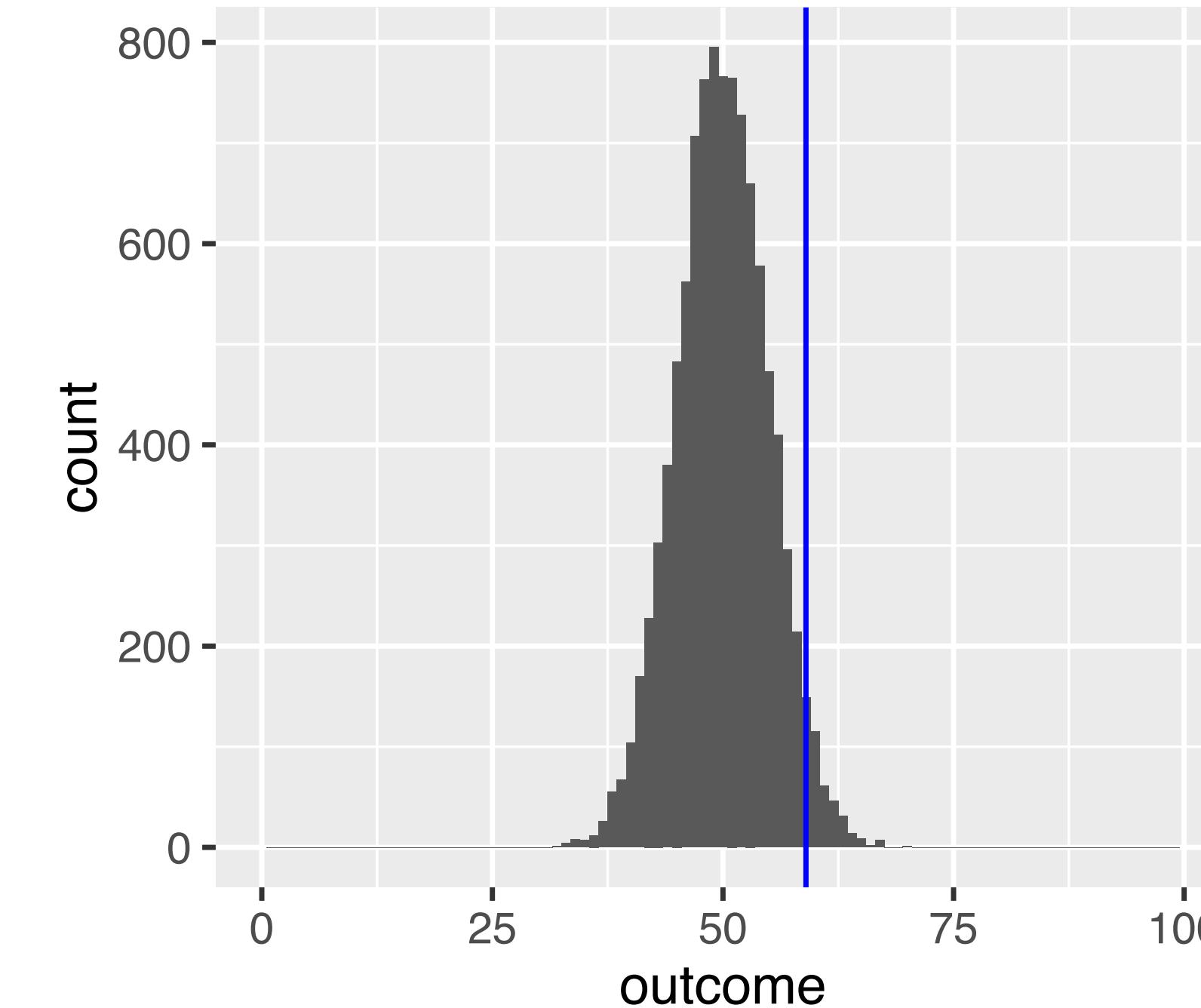
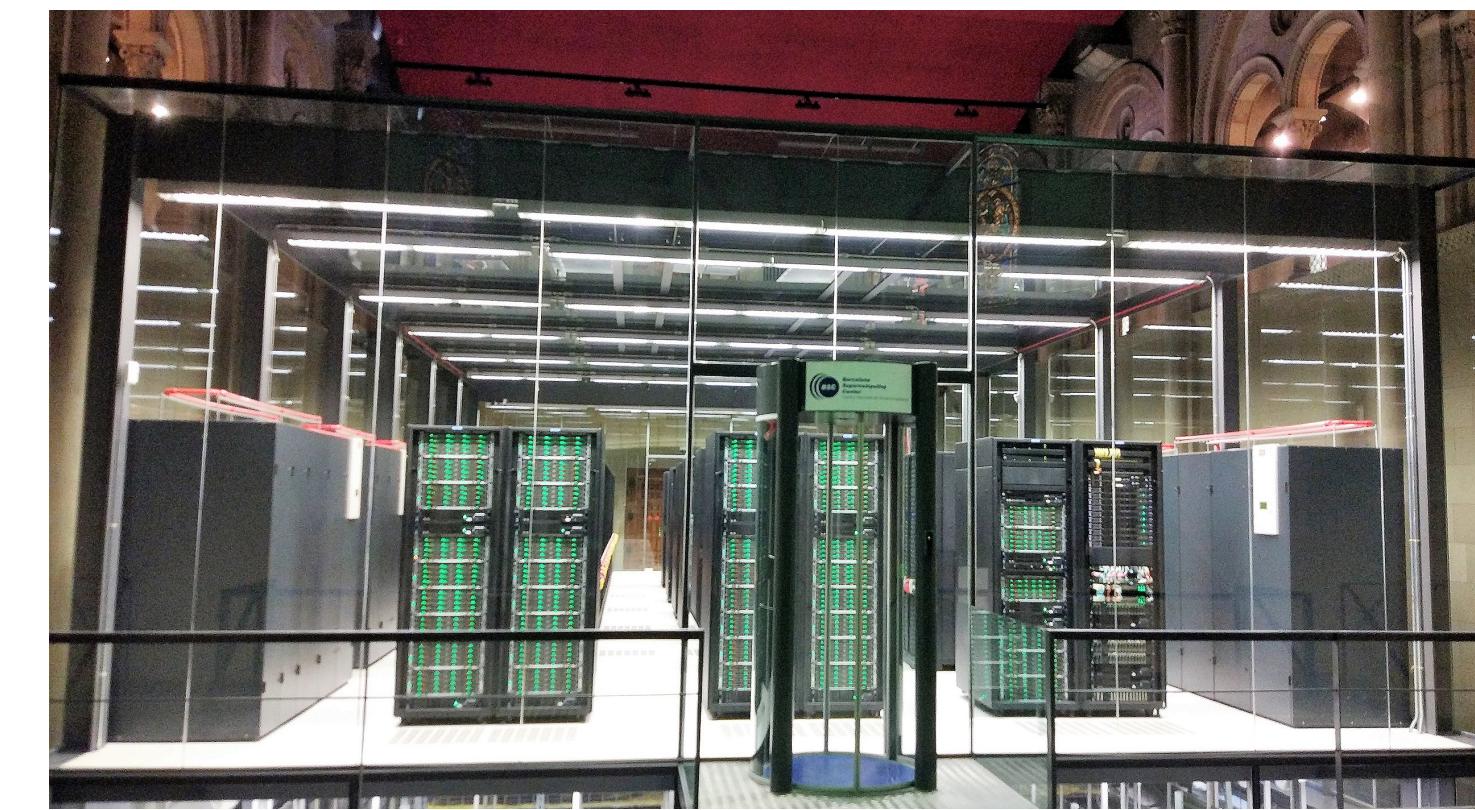


Figure 6.4: An approximation of the binomial distribution from 10^4 simulations (same parameters as Figure 6.3).



Parametric Theory vs Simulation

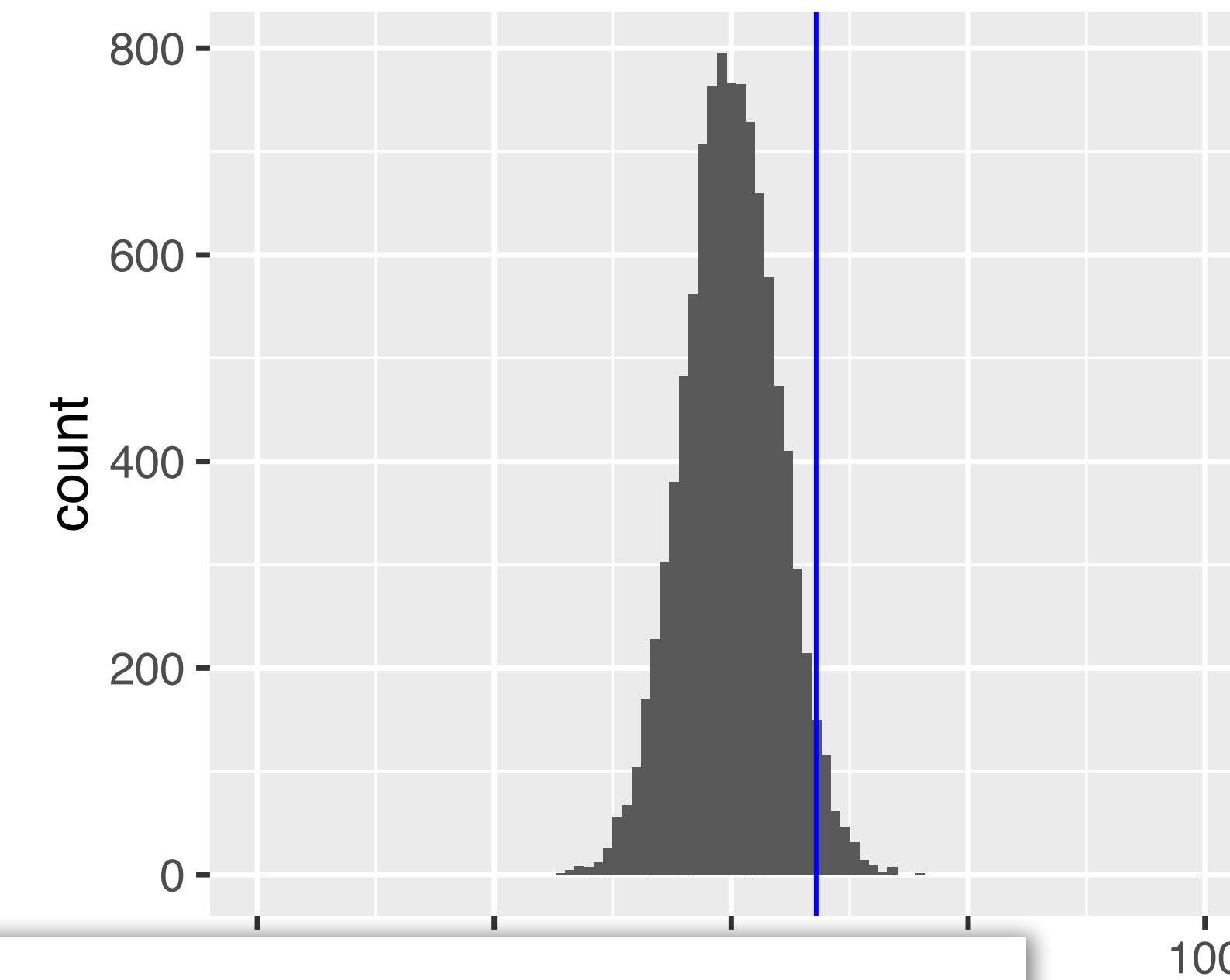
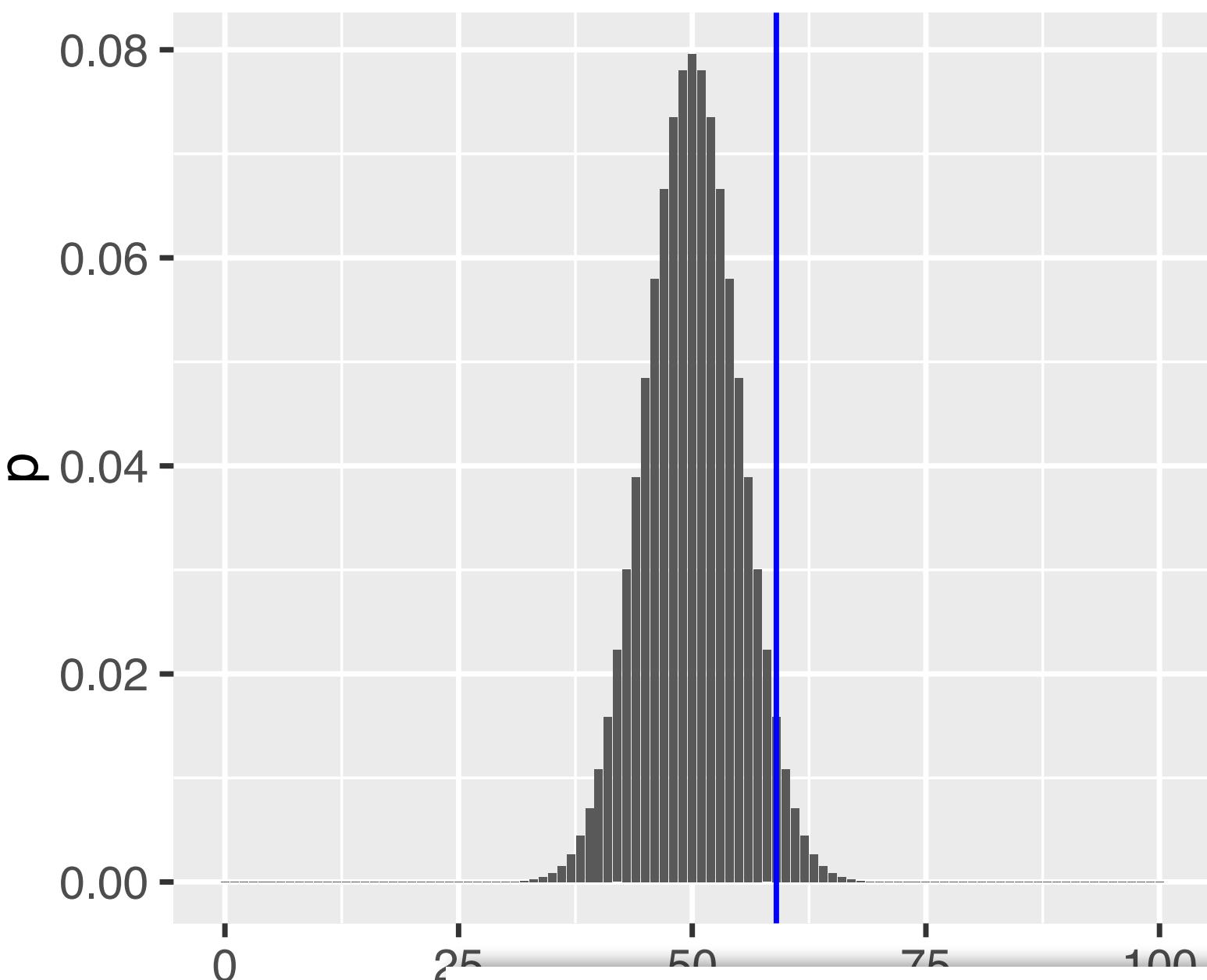
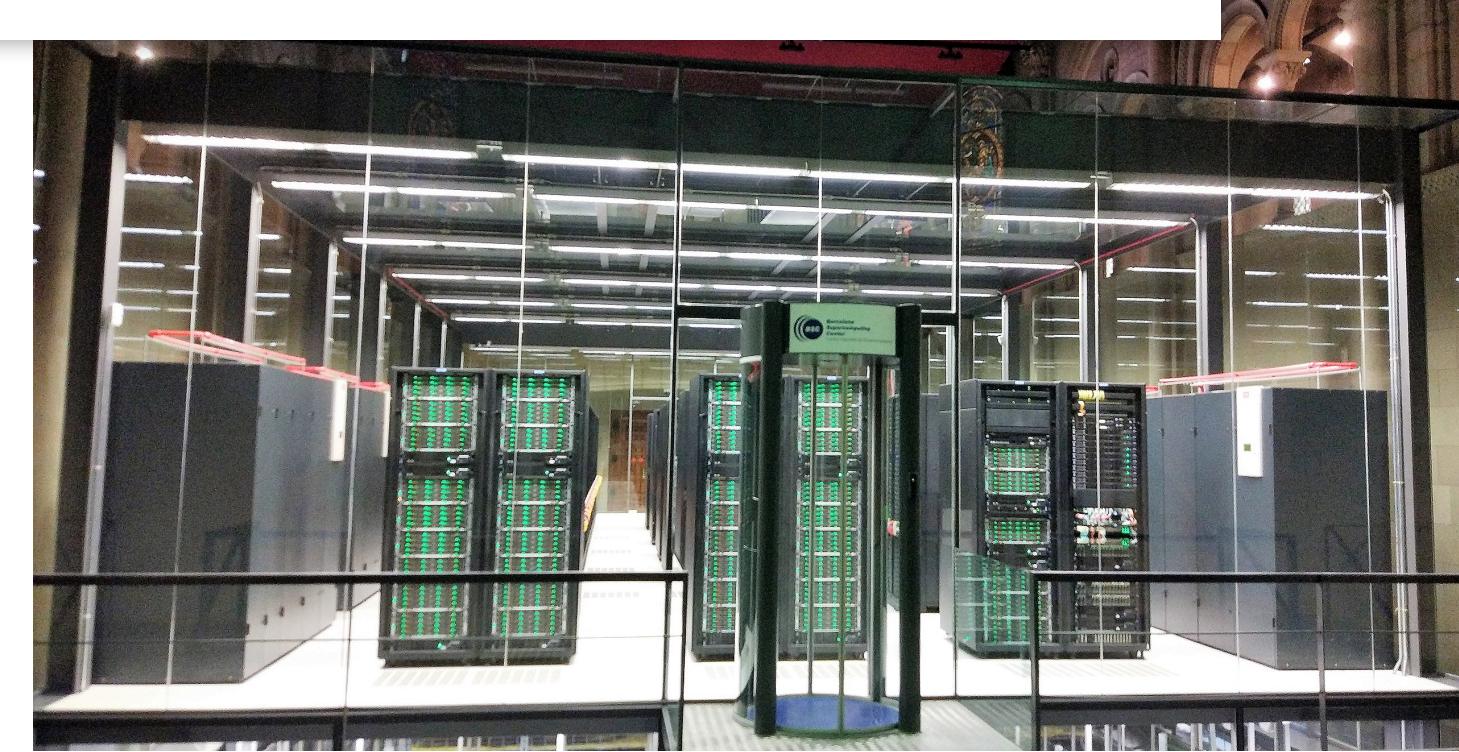


Figure 6.3: The k-th bin of the histogram of simulated data for the parameters $n = 100$ and $p = 0.5$ according to Equations 6.1 and 6.2. The two histograms show the same distribution, which is the binomial distribution with parameters $n = 100$ and $p = 0.5$.

Q:
Discuss pros and contras for each

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



t-distribution

If the data are identically normal distributed and independent, with sample sizes n_1, n_2 , then under H_0 , the *t*-statistic follows a *t*-distribution

$$f(t, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

$$\nu = n_1 + n_2$$

Comments and Pitfalls

“identical normal and independence” are sufficient, not necessary, conditions

Comments and Pitfalls

“identical normal and independence” are sufficient, not necessary, conditions

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

Comments and Pitfalls

“identical normal and independence” are sufficient, not necessary, conditions

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

Options: transform data, use permutations, simulations

Comments and Pitfalls

“identical normal and independence” are sufficient, not necessary, conditions

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

Options: transform data, use permutations, simulations

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

Comments and Pitfalls

“identical normal and independence” are sufficient, not necessary, conditions

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

Options: transform data, use permutations, simulations

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”

Avoid Fallacy

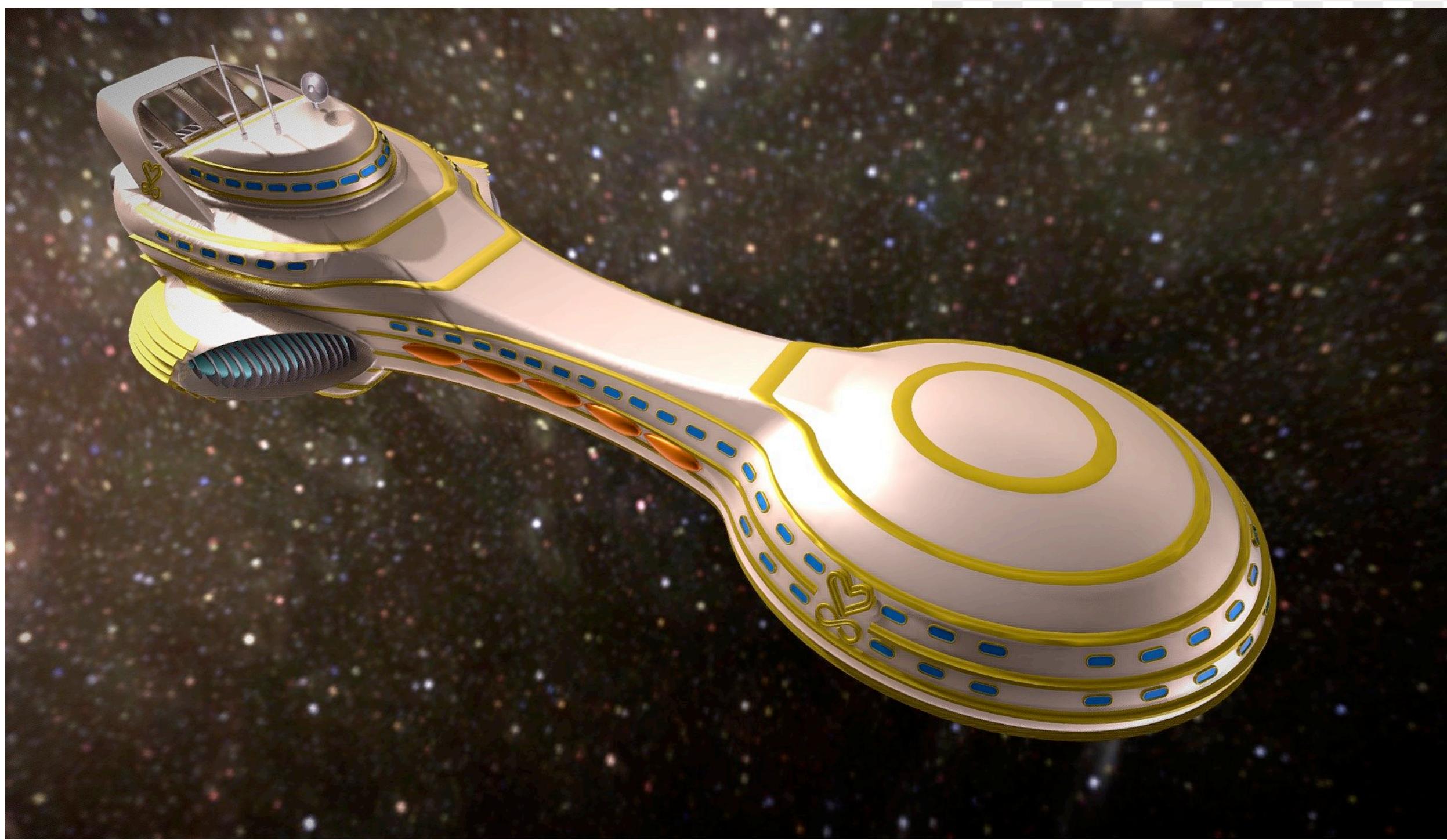
The p-value is the probability that the data could happen, under the condition that the null hypothesis is true.

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

Absence of evidence ≠ evidence of absence



Don't report absurdly small p-values



Reporting p values, W. Huber, Cell Systems, DOI:
10.1016/j.cels.2019.03.001

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

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. Lazarus

DOI: 10.1080/00031305.2016.1154108

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. Lazarus

DOI: 10.1080/00031305.2016.1154108

What can we do about this?

The p-value is 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, ...)? (a.k.a. “false discovery probability”)

The p value is:

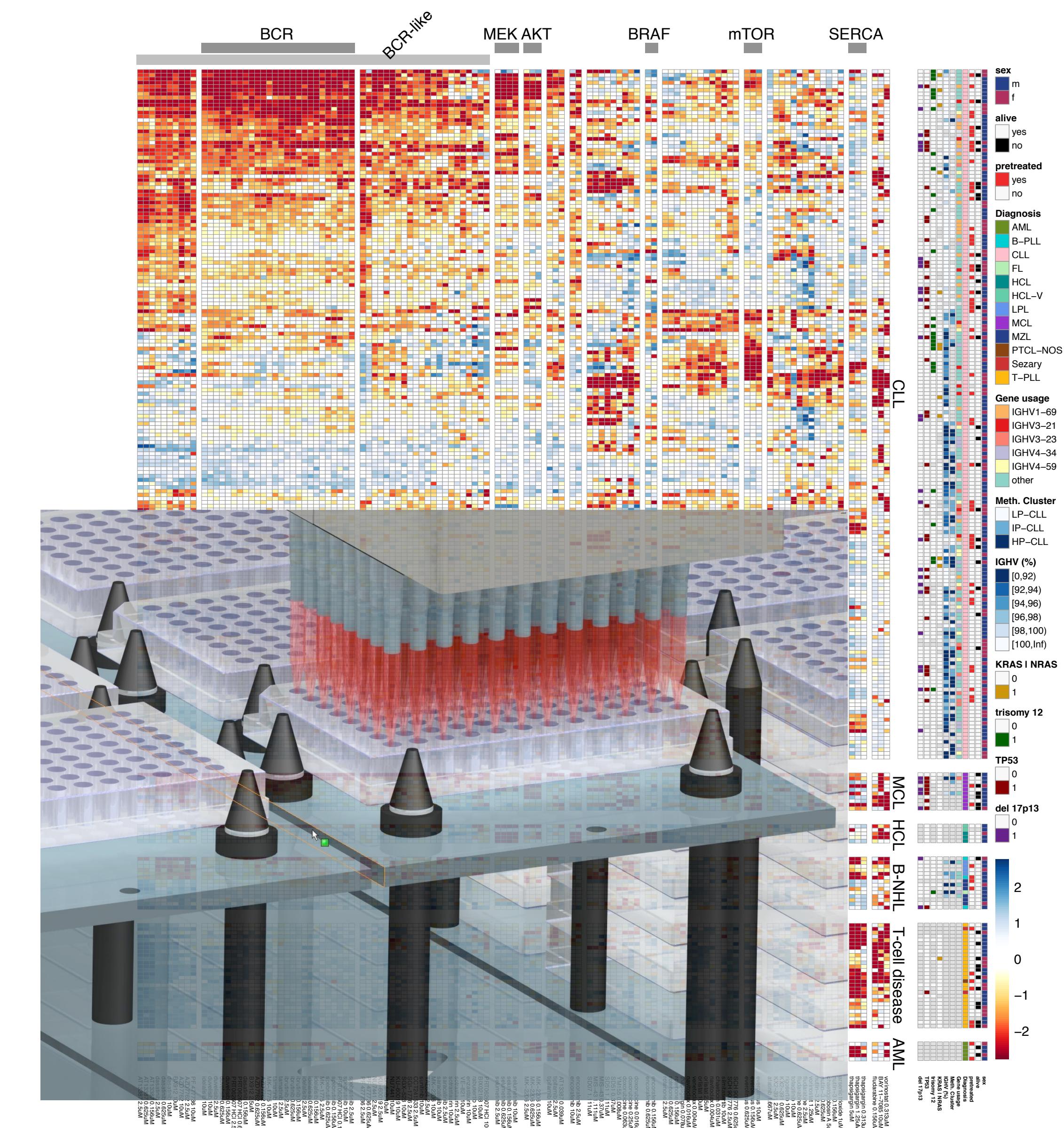
If the finding is wrong (null hypothesis is true), what is the probability of seeing the data.

Can we compute the answer to the interesting question instead?

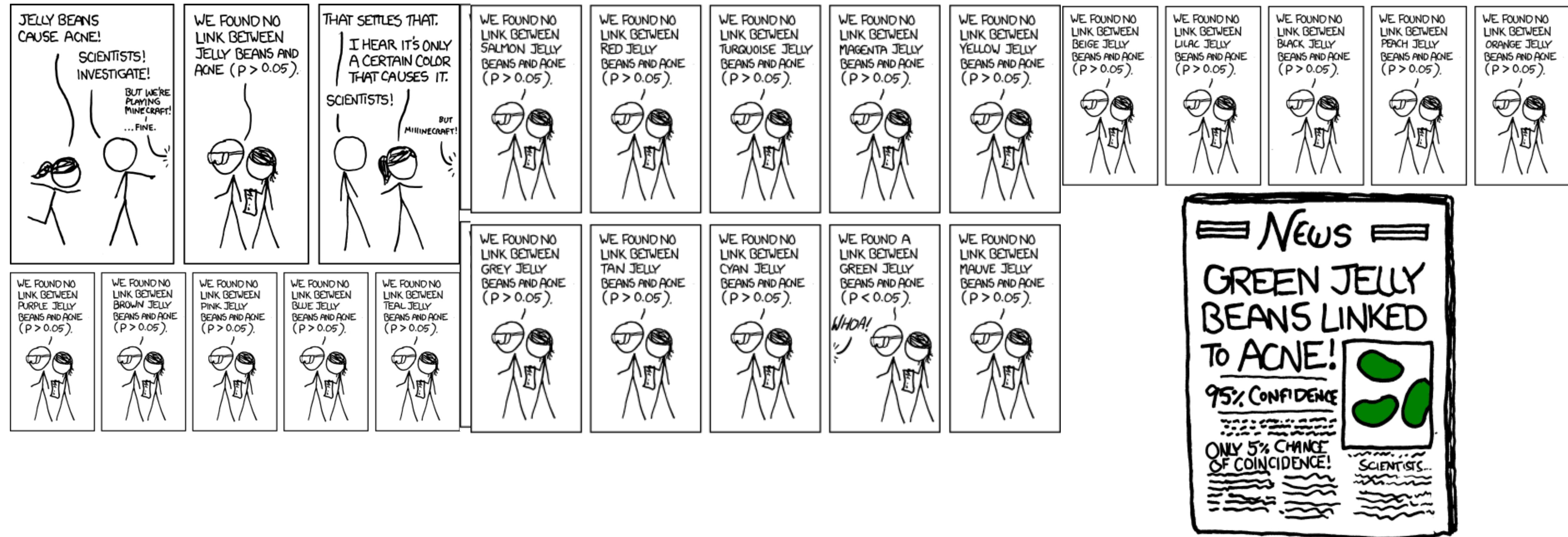
Multiple Testing

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
 - ...



Multiple Testing



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?"

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 a .

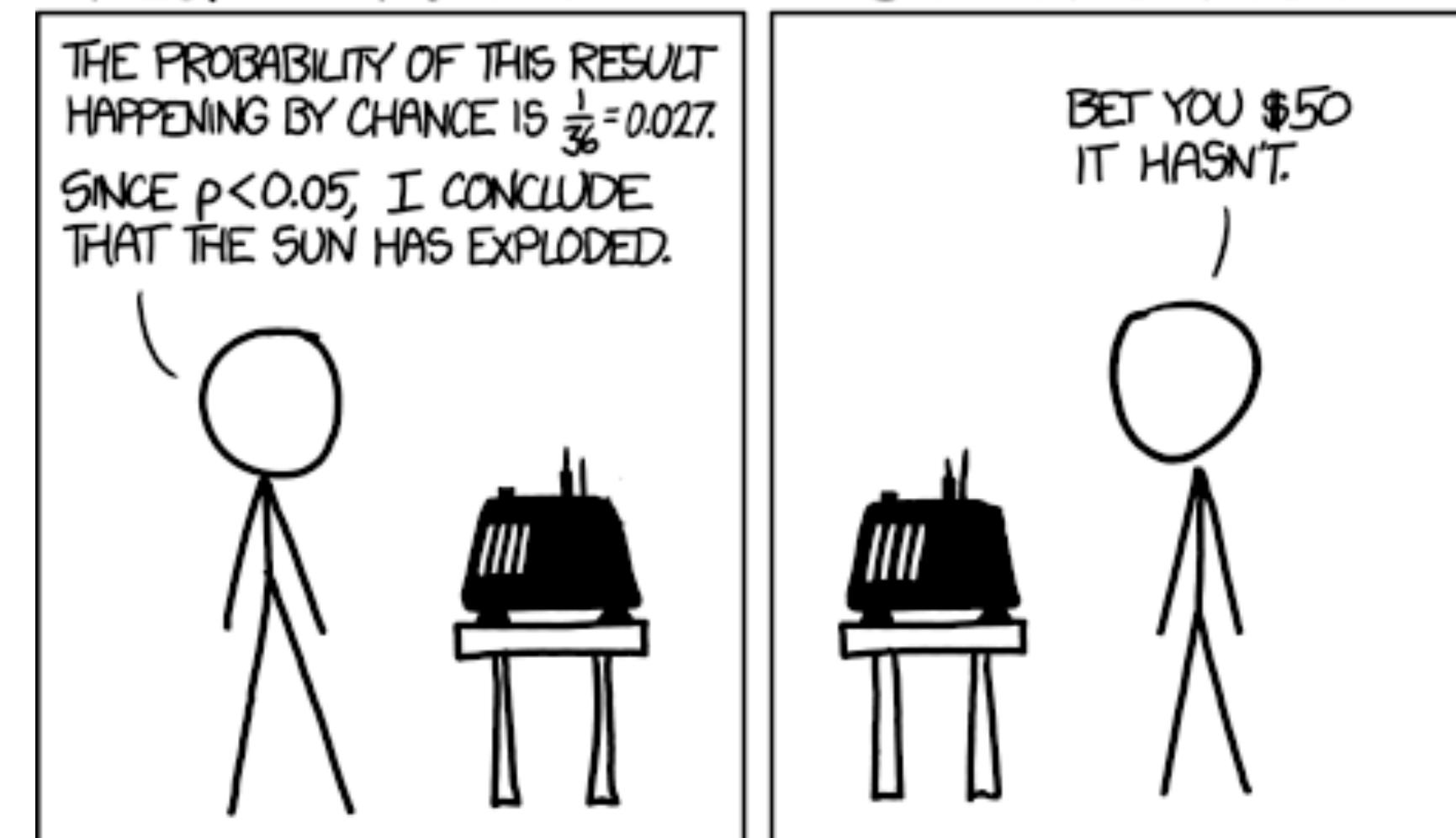
The Multiple Testing Opportunity

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

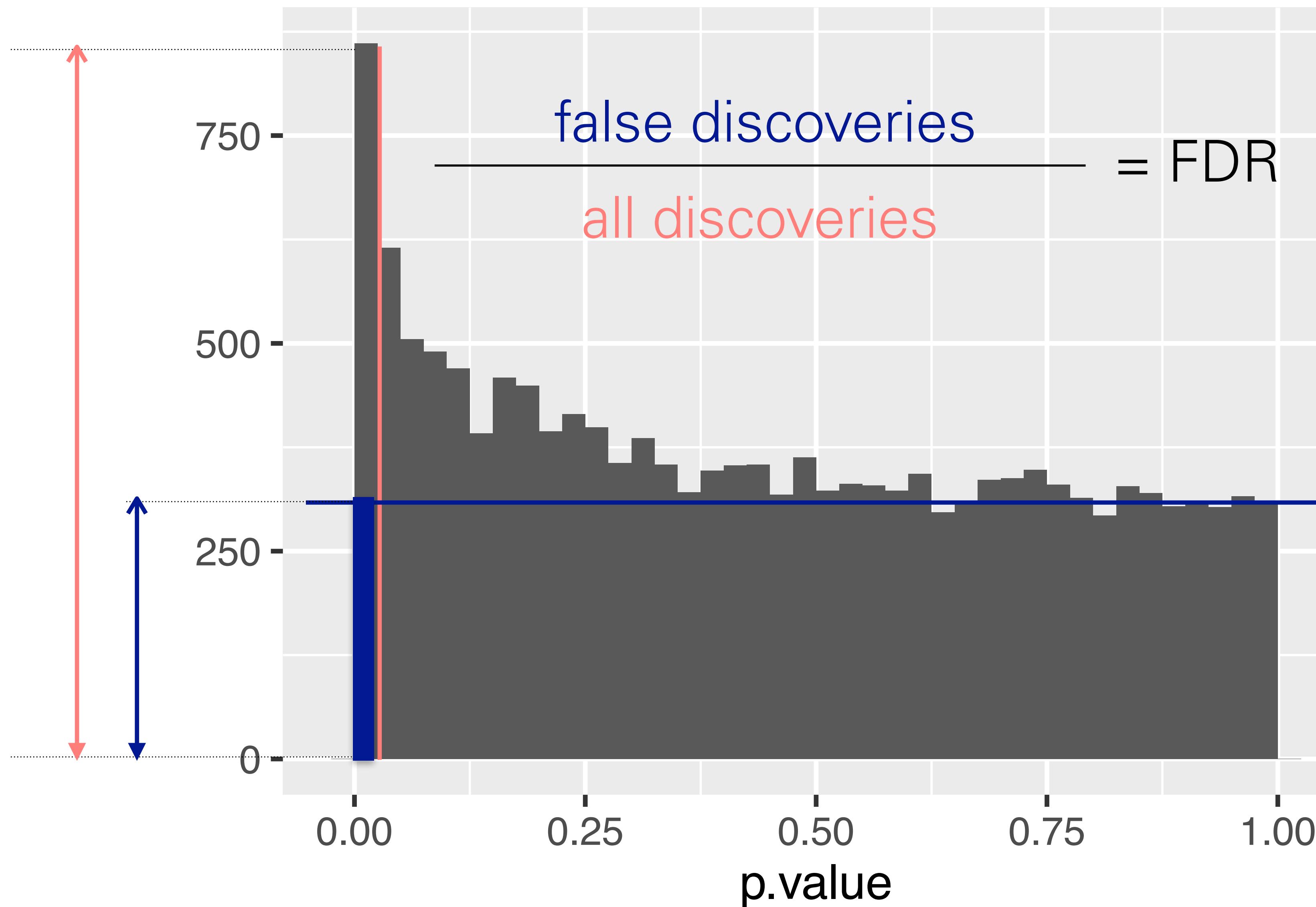


FREQUENTIST STATISTICIAN: BAYESIAN STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$. SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.

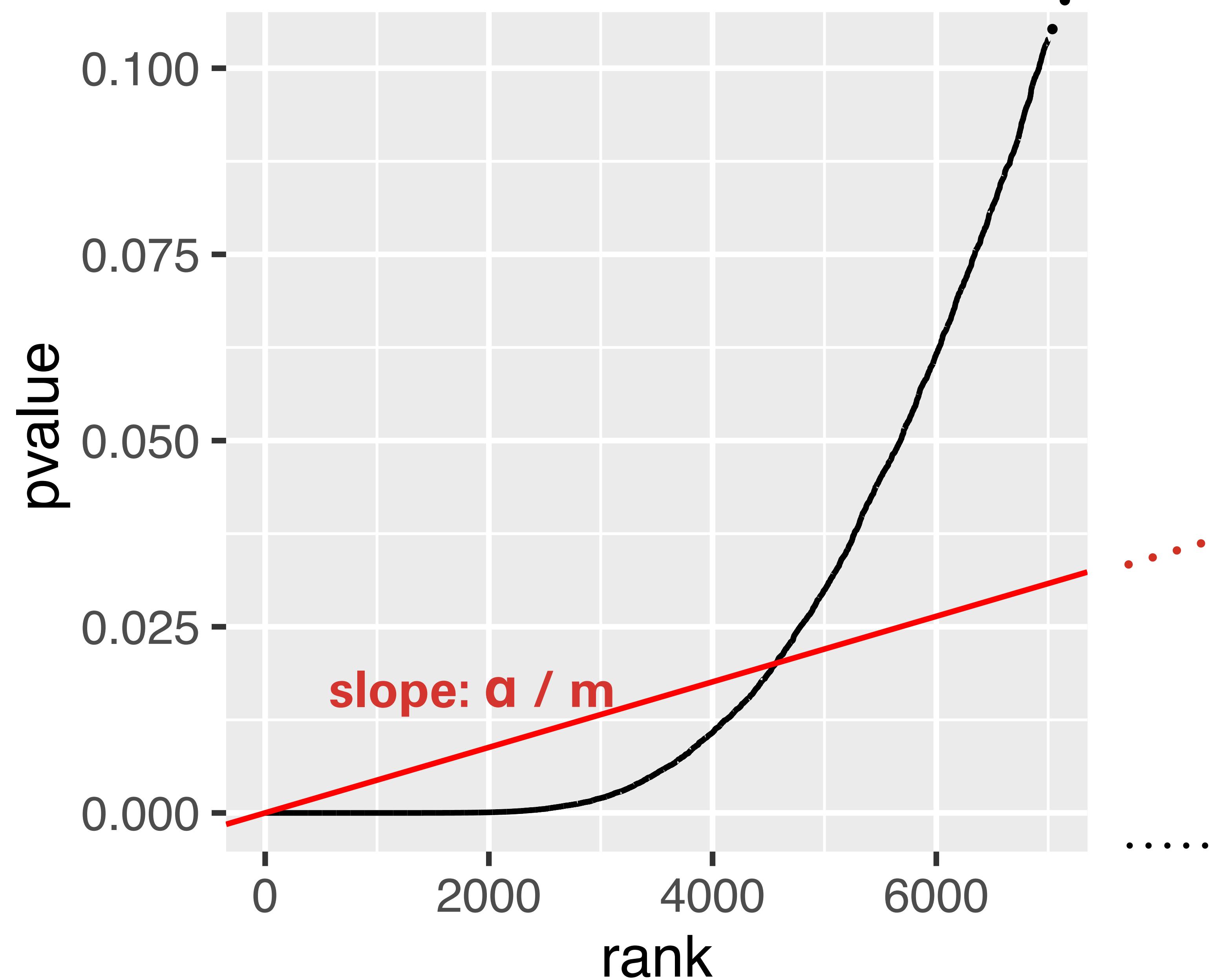


False Discovery Rate



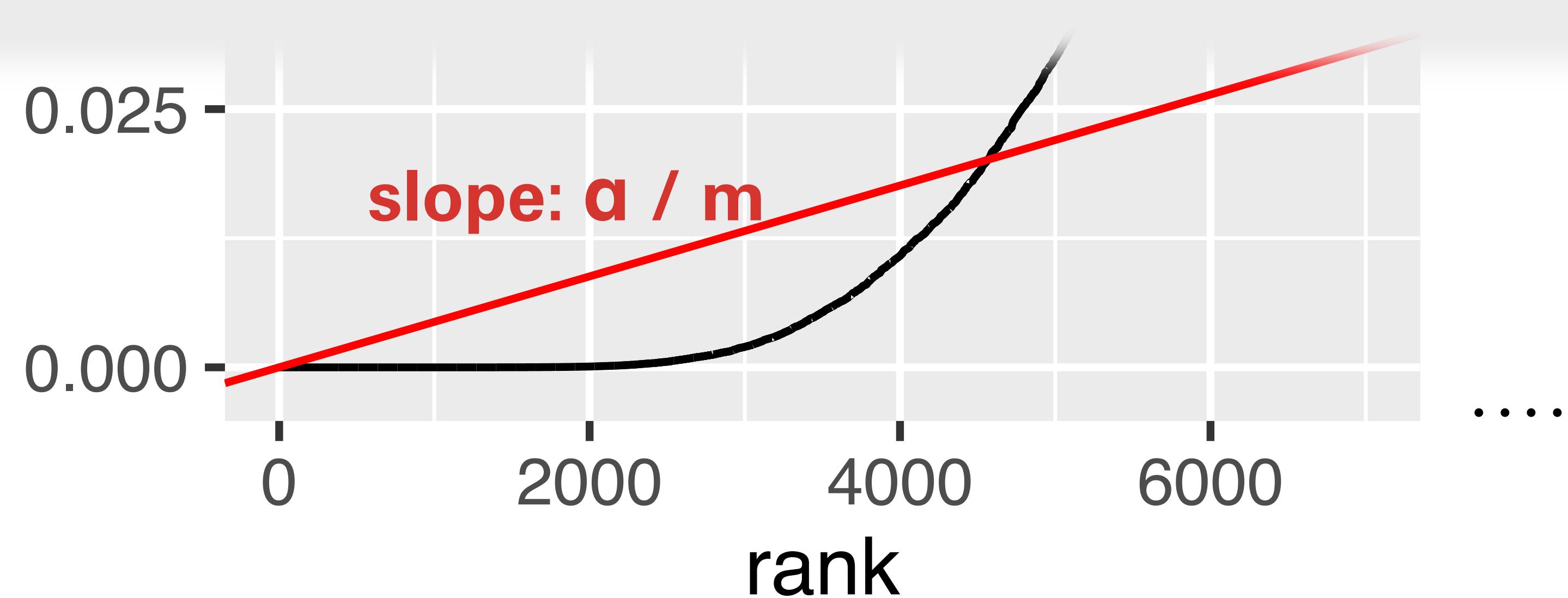
Method of Benjamini & Hochberg (1995)

Method of Benjamini & Hochberg



Method of Benjamini & Hochberg

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



Not all Hypothesis Tests are Created Equal

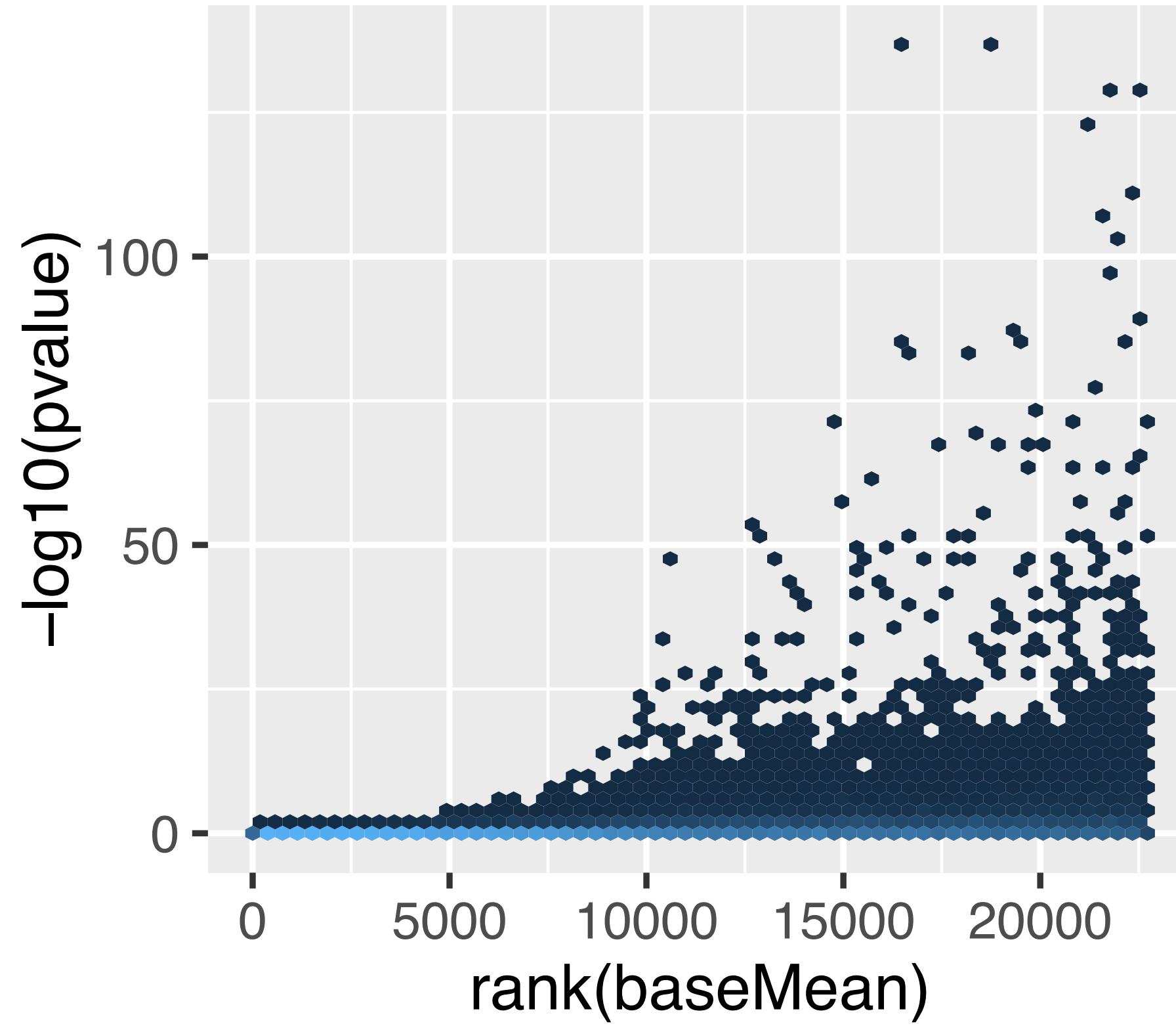
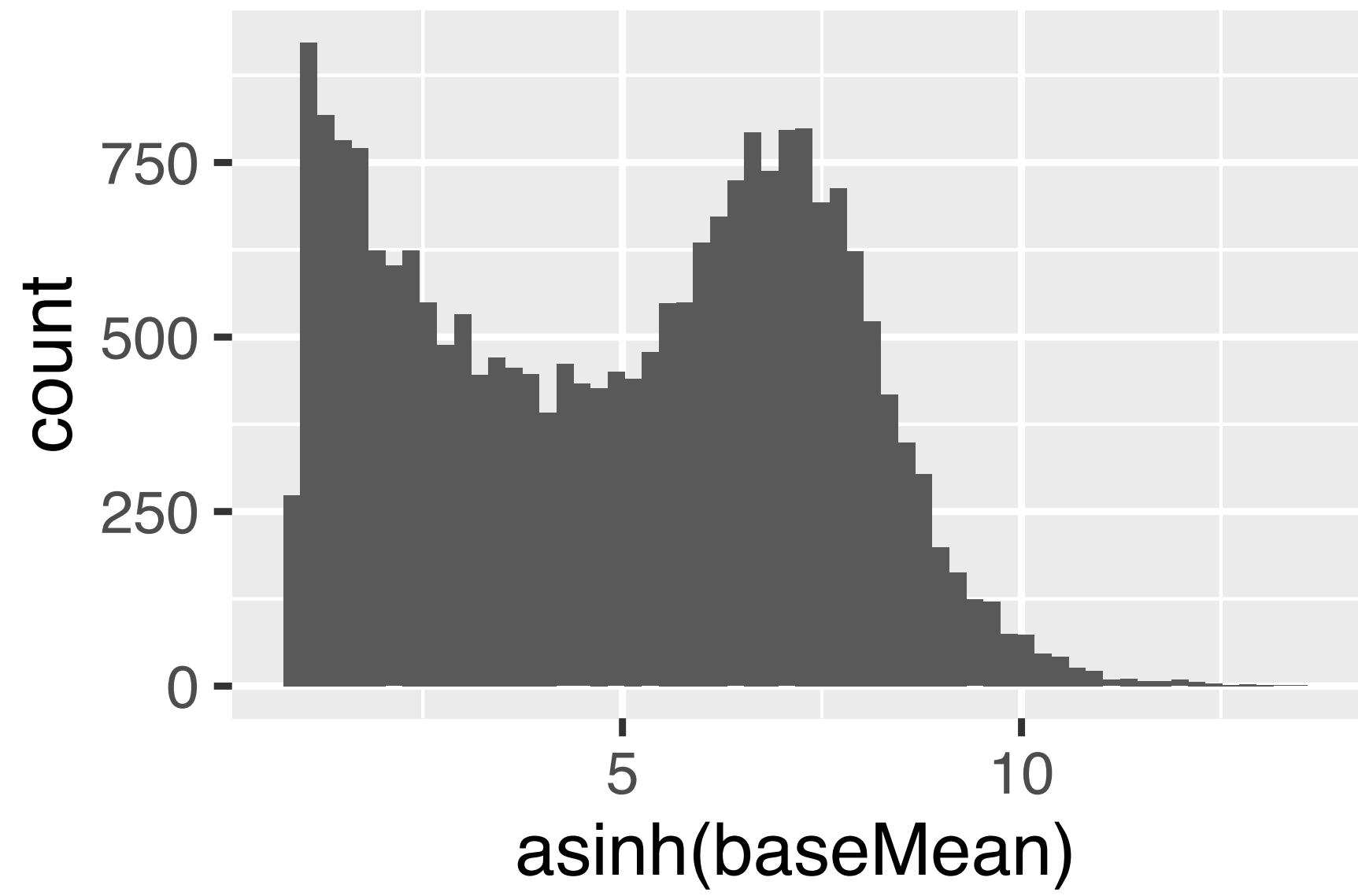
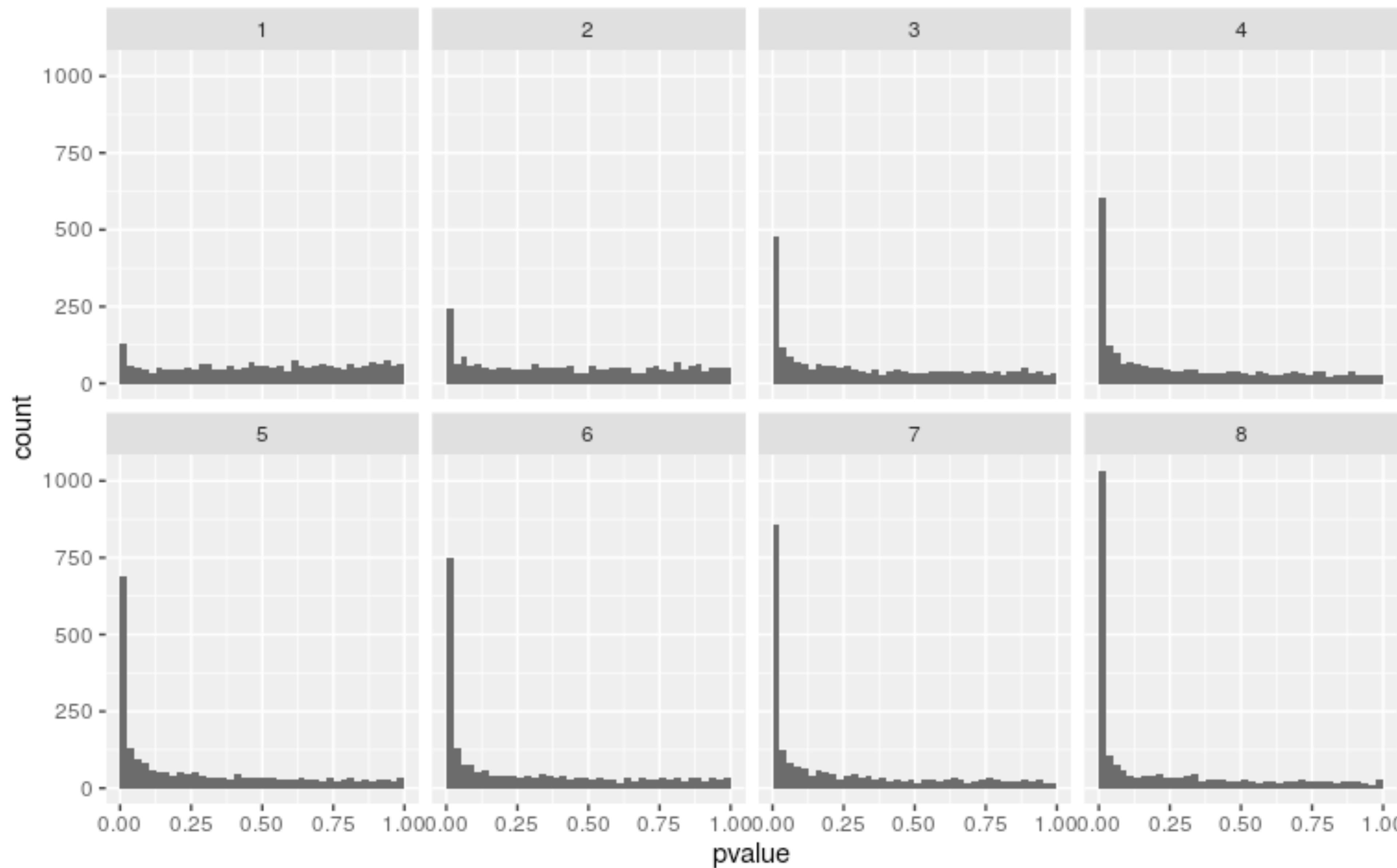


Figure 6.15: Histogram of baseMean . We see that it covers a large dynamic range, from close to 0 to around 3.3×10^5 .

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

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



Weighted Benjamini-Hochberg method

- Let $w_i \geq 0$ and $\frac{1}{m} \sum_{i=1}^m w_i = 1$ (“weight budget”).
- Define $Q_i = P_i/w_i$.
- Apply BH to Q_i instead of P_i .
- Proven Type-I error (FDR) control (Genovese, Roeder, Wasserman *Biometrika* 2006).
- If $w_i > 1$, then H_i is easier to reject.
- $Q_i \leq t \Leftrightarrow P_i \leq w_i t =: t_i$

Weighted Benjamini-Hochberg method

- Let $w_i \geq 0$ and $\frac{1}{m} \sum_{i=1}^m w_i = 1$
- Define $Q_i = P_i/w_i$.
- Apply BH to Q_i instead of P_i to control Type-I error rate.
- Proven Type-I error control (Genovese, Roeder, Wasserman *Biometrika*, 2006).
- If $w_i > 1$, then H_i is easier to reject.
- $Q_i \leq t \Leftrightarrow P_i \leq w_i t =: t_i$

Problem: how to know the weights?



Independent hypothesis weighting (IHW)

- 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

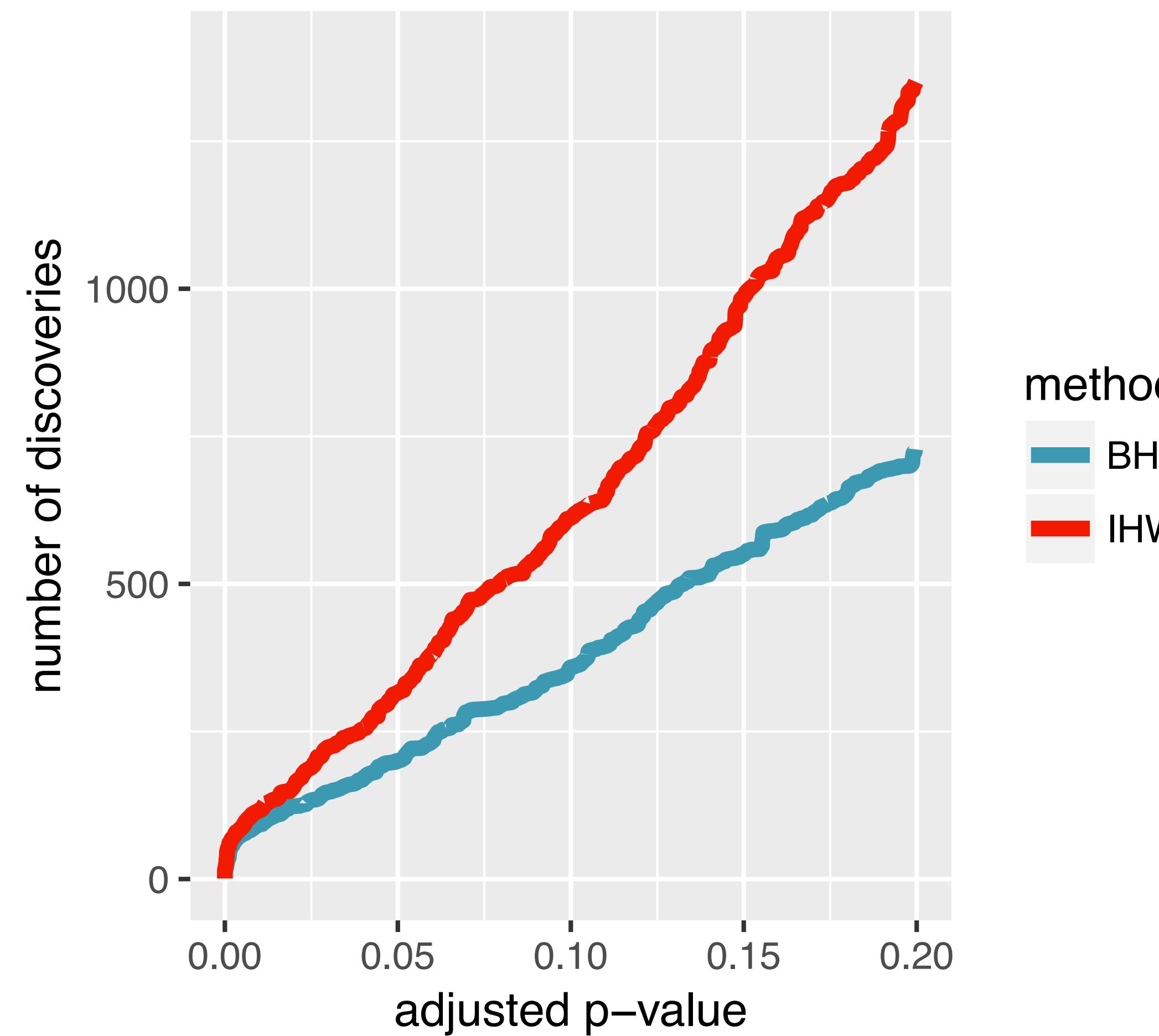
Bioconductor package **IHW**

Ignatiadis et al.,

- Nature Methods 2016, DOI 10.1038/nmeth.3885
- JRSSB 2021, DOI 10.1111/rssb.12411

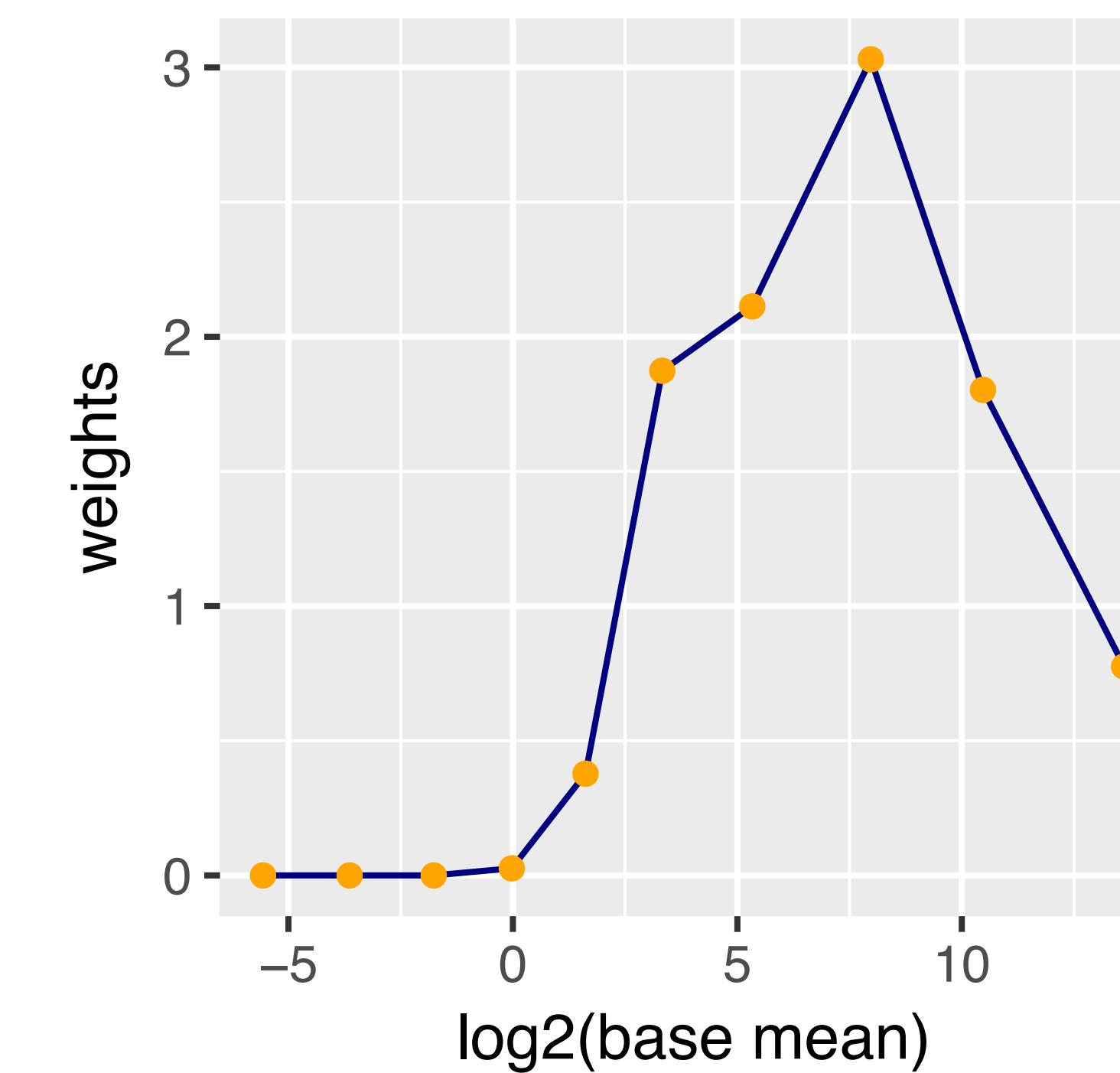
RNA-Seq example (DESeq2)

power

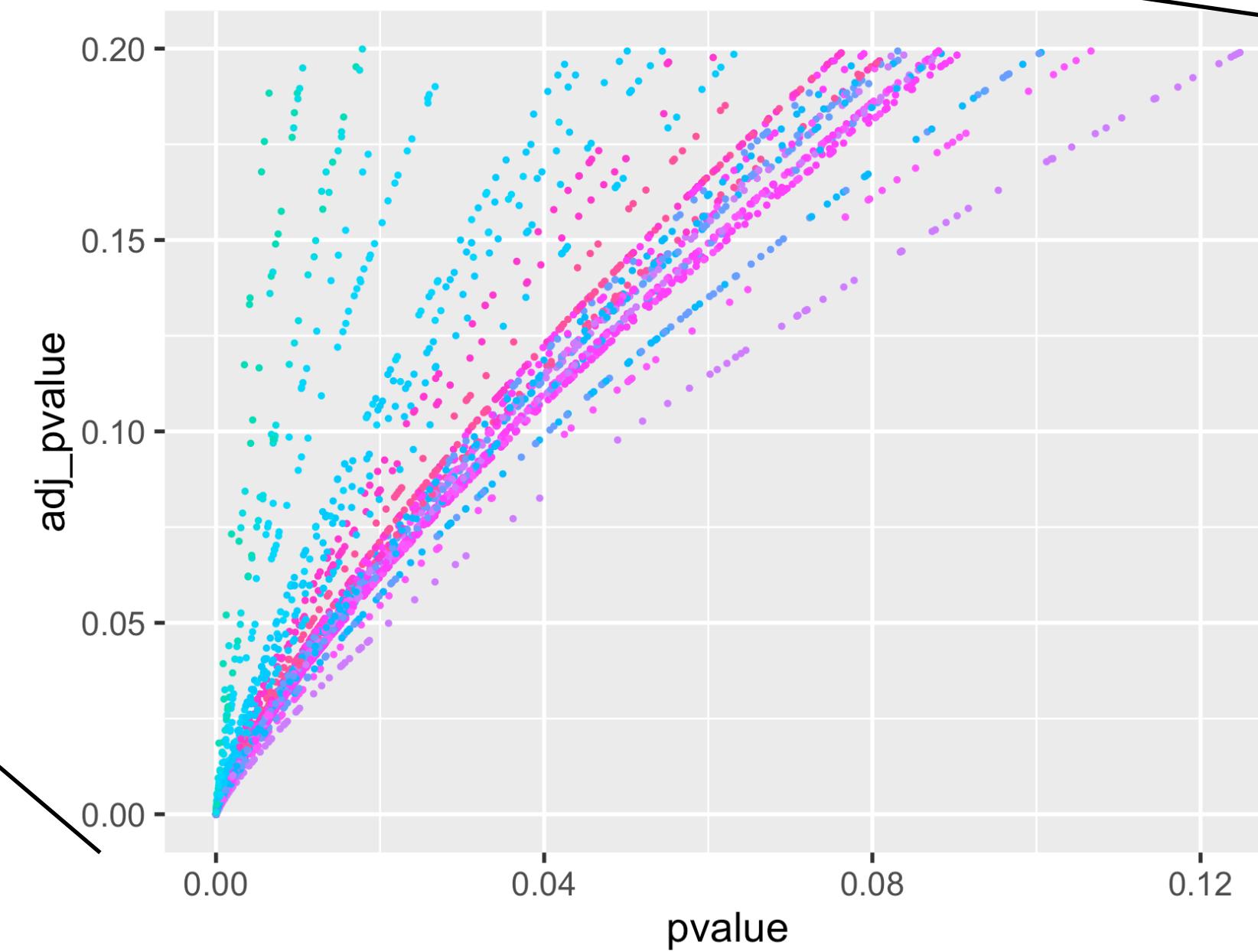
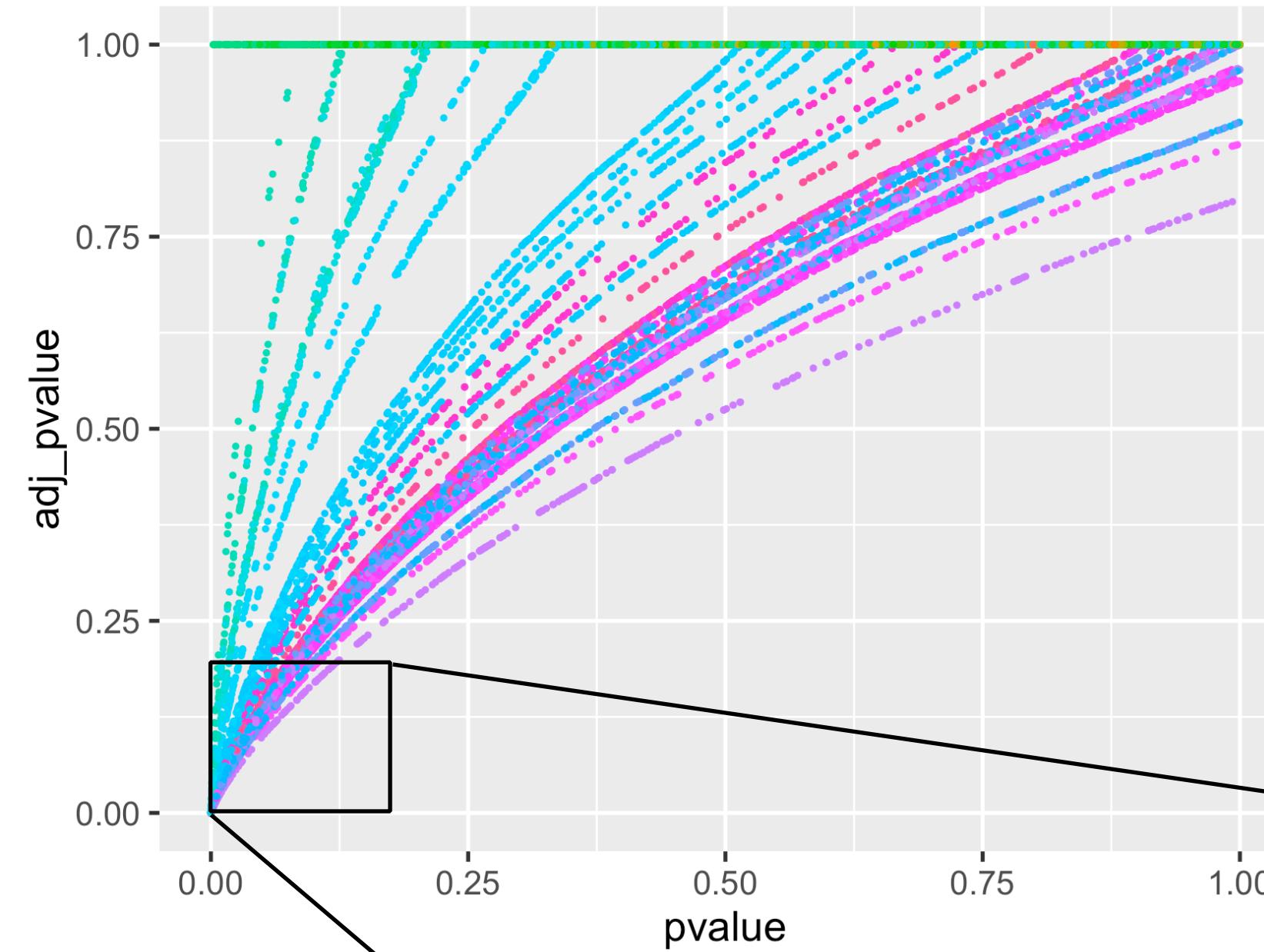


method
BH
IHW

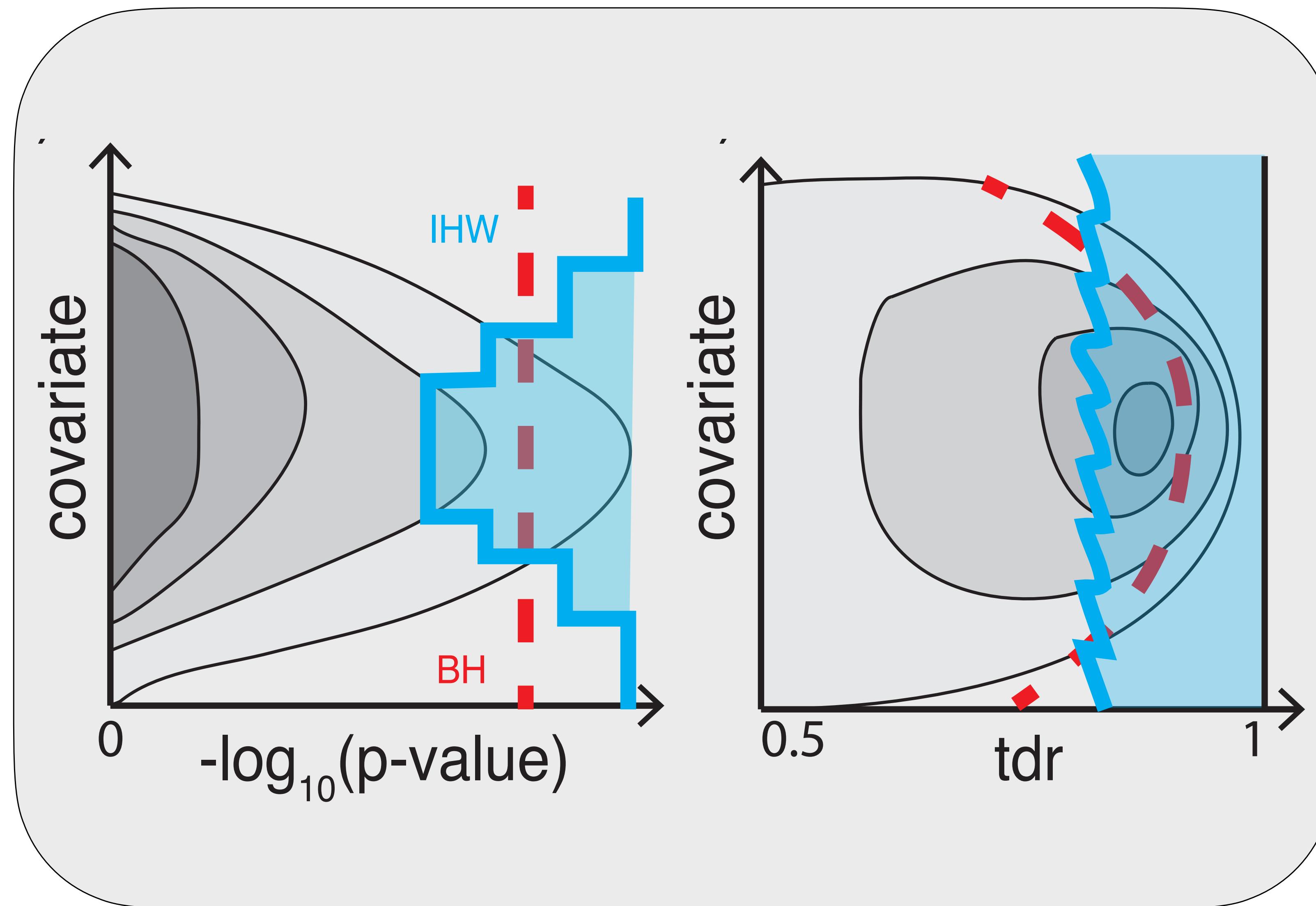
weights



Ranking is not monotonous in raw p-values



The decision boundaries is in two dimensions



Summary

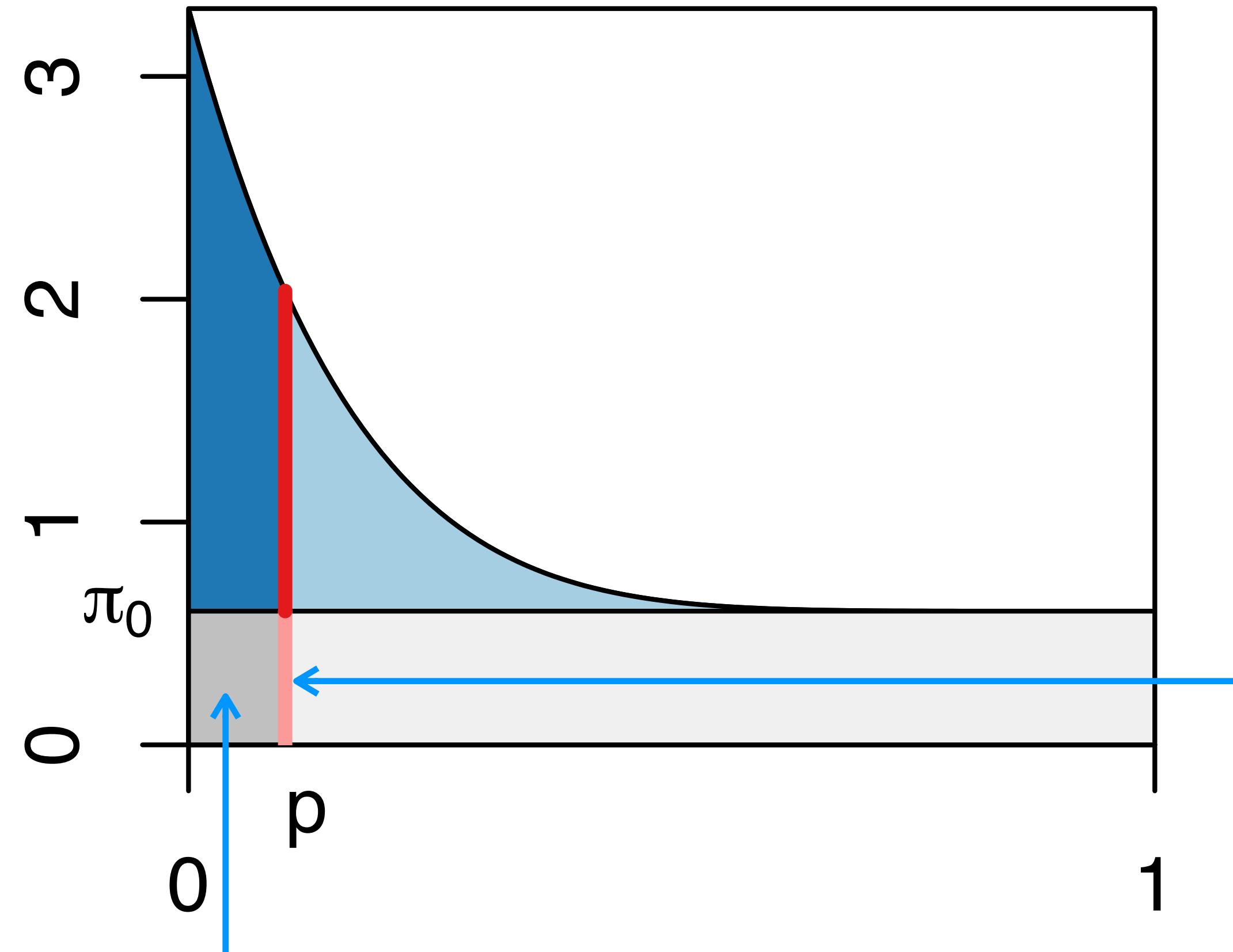
- Multiple testing is not a problem but an opportunity
- Heterogeneity across tests
- Informative covariates are often apparent to domain scientists
 - independent of test statistic under the null
 - informative on π_1 , F_{alt}
- Can do data-driven weighting (“IHW”)
 - Scales well to millions of hypotheses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

The p value is not enough



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

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



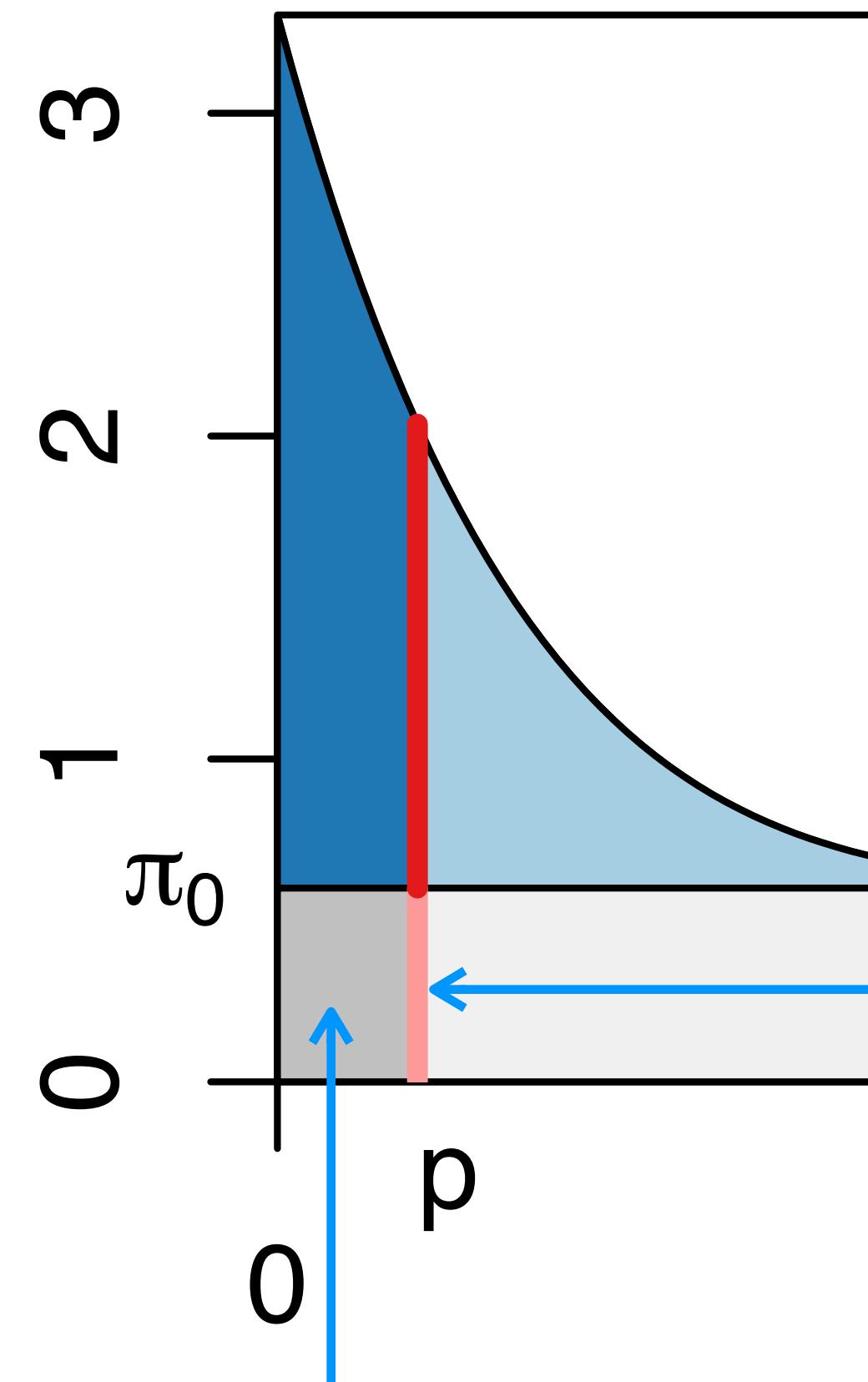
FDR: Ratio between the areas. An average property of all tests rejected below the threshold.

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

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

fdr: Ratio between the line segment lengths. Applies to tests rejected just at this particular threshold.

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



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

But how do we
know π_0 and f_{alt} ?

the line
Applies
to tests rejected just at this
particular threshold.

FDR: Ratio between the areas. An average property of all tests rejected below the threshold.

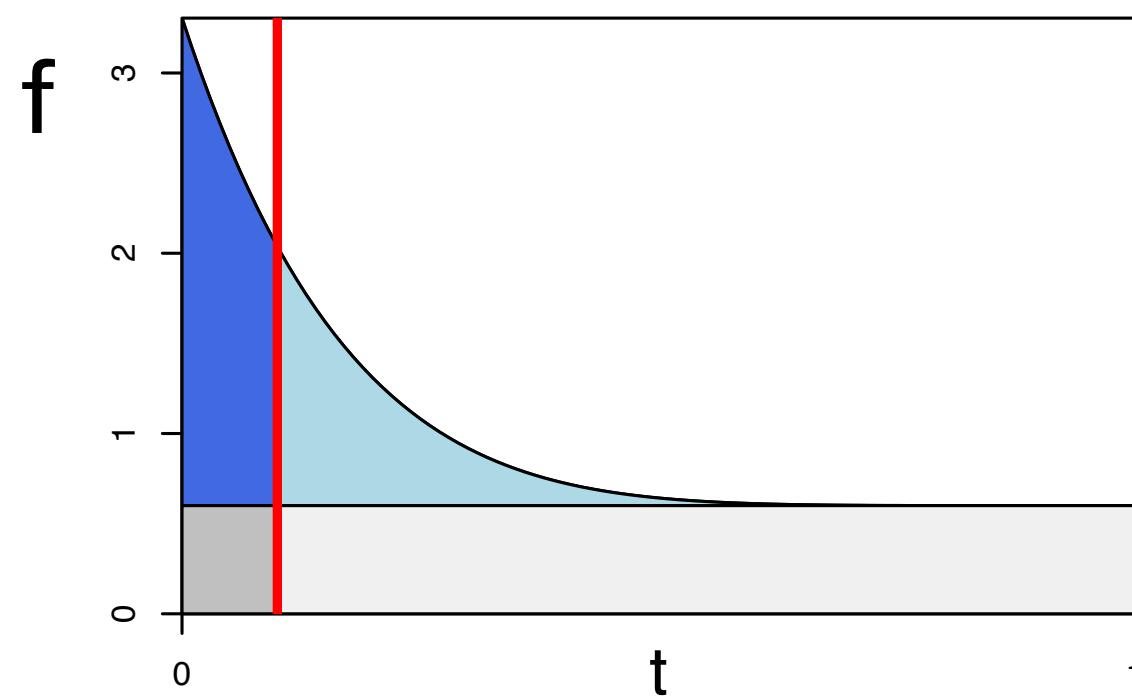
Same p-value, different FDR / fdr

$$X_i \sim \mathbb{P}^X$$

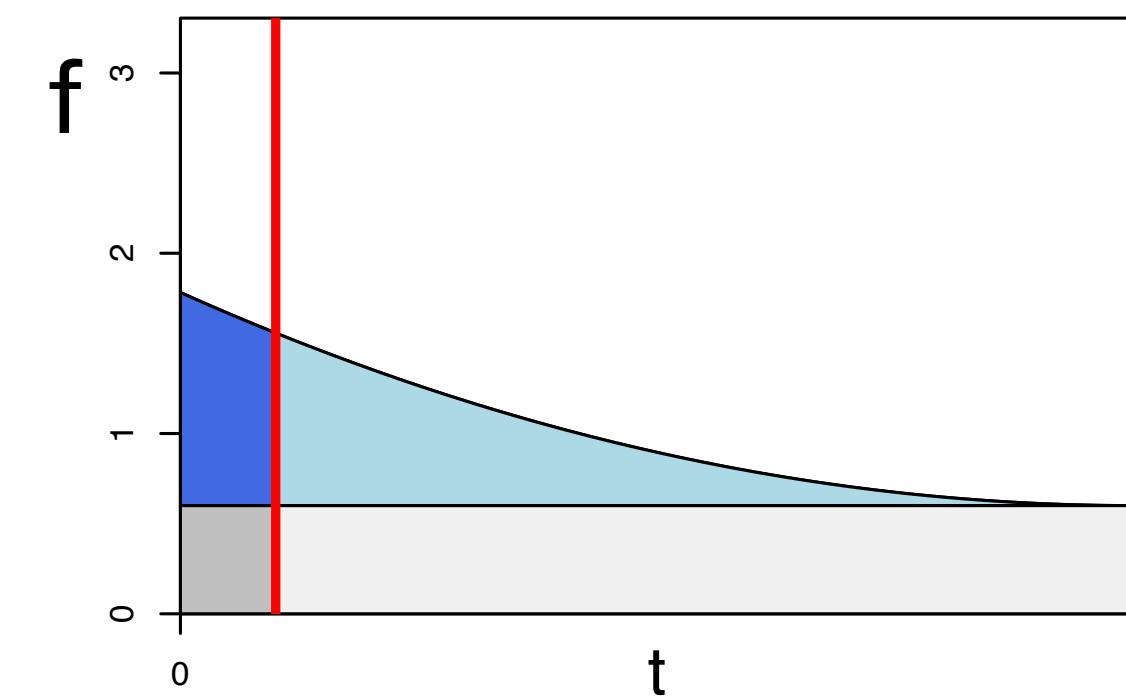
$$H_i \mid X_i \sim \text{Bernoulli}(1 - \pi_0(X_i))$$

$$P_i \mid (H_i = 0, X_i) \sim U[0, 1]$$

$$P_i \mid (H_i = 1, X_i) \sim F_{\text{alt}|X_i}$$

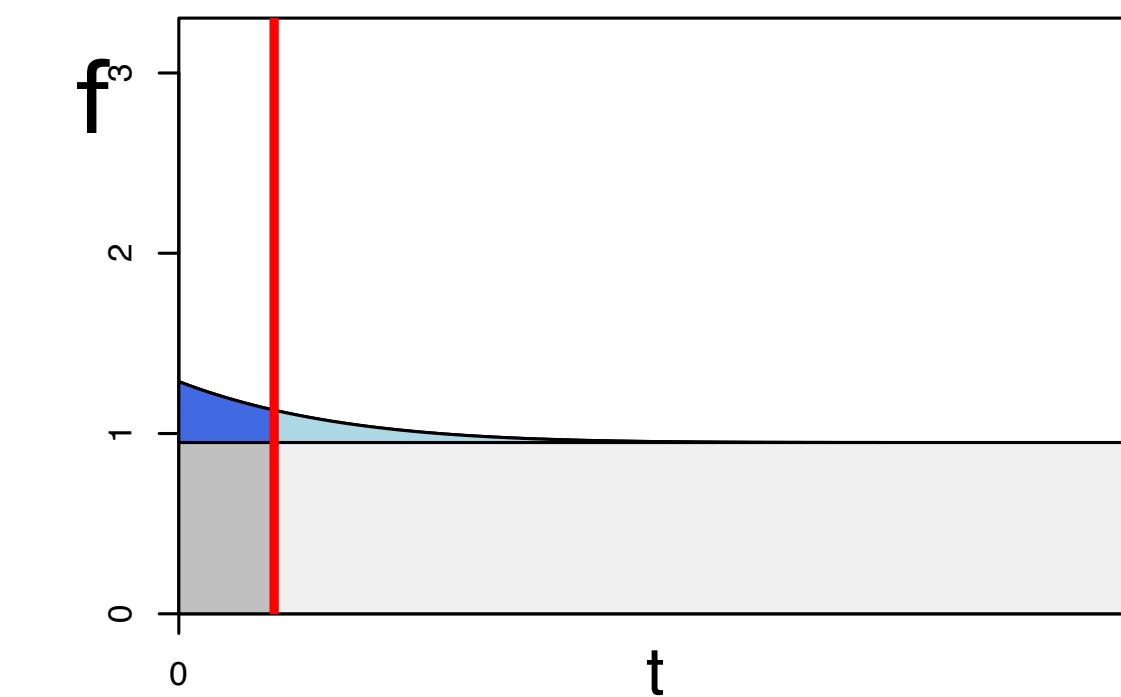


$\pi_0 = 0.6$



$\pi_0 = 0.6$

different F_{alt}



$\pi_0 = 0.95$

same F_{alt}