False Discovery Rates, A New Deal

Matthew Stephens

2014/7/8

Before we get started: Let's get Organized!

- Over ~10 years of working with graduate students + postdocs, I've noticed something.

Before we get started: Let's get Organized!

- Over ~10 years of working with graduate students + postdocs, I've noticed something.
- Organized researchers get more done (and better!).

Before we get started: Let's get Organized!

- Over ~10 years of working with graduate students + postdocs, I've noticed something.
- Organized researchers get more done (and better!).
- Many of them are more organized than I am!

Before we get started: Let's get Organized!

- Over ~10 years of working with graduate students + postdocs, I've noticed something.
- Organized researchers get more done (and better!).
- Many of them are more organized than I am!
- Thought: I should get organized; I should help others get organized.

So what can you do?

- Buy a notebook; bring it to meetings; make notes!

So what can you do?

- Buy a notebook; bring it to meetings; make notes!
- Come to meetings with a written agenda.

So what can you do?

- Buy a notebook; bring it to meetings; make notes!
- Come to meetings with a written agenda.
- While doing research, record what you did and what the outcome was.

So what can you do?

- Buy a notebook; bring it to meetings; make notes!
- Come to meetings with a written agenda.
- While doing research, record what you did and what the outcome was.
- Use version control (git) and internet repositories (bitbucket, github) to organize notes, code, etc.

So what can you do?

- Buy a notebook; bring it to meetings; make notes!
- Come to meetings with a written agenda.
- While doing research, record what you did and what the outcome was.
- Use version control (git) and internet repositories (bitbucket, github) to organize notes, code, etc.
- Use *knitr* to help make your research reproducible.

So what can you do?

- Buy a notebook; bring it to meetings; make notes!
- Come to meetings with a written agenda.
- While doing research, record what you did and what the outcome was.
- Use version control (git) and internet repositories (bitbucket, github) to organize notes, code, etc.
- Use *knitr* to help make your research reproducible.
- Talk about the tools you find useful!

What are these repository things?

- A repository: a central place in which an aggregation of data is kept and maintained in an organized way (searcharticle.com)

What are these repository things?

- A repository: a central place in which an aggregation of data is kept and maintained in an organized way (searcharticle.com)
- Great for sharing material across multiple people (eg student and advisor!)

What are these repository things?

- A repository: a central place in which an aggregation of data is kept and maintained in an organized way (searcharticle.com)
- Great for sharing material across multiple people (eg student and advisor!)
- An amateur example: `http://github.com/stephens999/ash`

What is knitr?

- An R package

What is knitr?

- An R package
- A tool for literate programming

What is knitr?

- An R package
- A tool for literate programming
- Text, and R code are interleaved

What is knitr?

- An R package
- A tool for literate programming
- Text, and R code are interleaved
- When you compile the document, the code is run, and output inserted into the text.

What is knitr?

- An R package
- A tool for literate programming
- Text, and R code are interleaved
- When you compile the document, the code is run, and output inserted into the text.
- Great for writing reports, and keeping a track of what you did and what the result was!

What is knitr?

- An R package
- A tool for literate programming
- Text, and R code are interleaved
- When you compile the document, the code is run, and output inserted into the text.
- Great for writing reports, and keeping a track of what you did and what the result was!
- This talk was written with knitr (with RStudio)!

What is Reproducible Research?

- Principle: when publishing results of computational procedures, we should publish the code that produced the results.

What is Reproducible Research?

- Principle: when publishing results of computational procedures, we should publish the code that produced the results.
- "publishing figures or results without the complete software environment could be compared to a mathematician publishing an announcement of a mathematical theorem without giving the proof" (Buckheit and Donohoe)

What is Reproducible Research?

- Principle: when publishing results of computational procedures, we should publish the code that produced the results.
- "publishing figures or results without the complete software environment could be compared to a mathematician publishing an announcement of a mathematical theorem without giving the proof" (Buckheit and Donohoe)
- "an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." [Claerbout]

Why is reproducibility important?

- Not only because people are forgetful, error-prone, or dishonest!

Why is reproducibility important?

- Not only because people are forgetful, error-prone, or dishonest!
- Reproducing work is also the first step to extending it.

Why is reproducibility important?

- Not only because people are forgetful, error-prone, or dishonest!
- Reproducing work is also the first step to extending it.
- Helps communications among researchers (eg student + advisor).

Why is reproducibility important?

- Not only because people are forgetful, error-prone, or dishonest!
- Reproducing work is also the first step to extending it.
- Helps communications among researchers (eg student + advisor).
- If you do not publish code implementing your methods, your methods will likely go unused.

More on git, github, knitr, reproducibility

- Google "The git book", to get started on git.

More on git, github, knitr, reproducibility

- Google "The git book", to get started on git.
- Google "Karl Broman github tutorial" for statistics-oriented intro to github.

More on git, github, knitr, reproducibility

- Google "The git book", to get started on git.
- Google "Karl Broman github tutorial" for statistics-oriented intro to github.
- Google "donohoe buckheit" for "Wavelab and reproducible research"

The Canonical Genomics Experiment

- Measure lots of things, with error

The Canonical Genomics Experiment

- Measure lots of things, with error
- Get estimates of effects β_j ($\hat{\beta}_j$) and their standard errors s_j

The Canonical Genomics Experiment

- Measure lots of things, with error
- Get estimates of effects β_j ($\hat{\beta}_j$) and their standard errors s_j
- Turn these into Z-scores, $z_j = \hat{\beta}_j / s_j$

The Canonical Genomics Experiment

- Measure lots of things, with error
- Get estimates of effects β_j ($\hat{\beta}_j$) and their standard errors s_j
- Turn these into Z-scores, $z_j = \hat{\beta}_j / s_j$
- Turn these into p values, p_j

The Canonical Genomics Experiment

- Measure lots of things, with error
- Get estimates of effects β_j ($\hat{\beta}_j$) and their standard errors s_j
- Turn these into Z-scores, $z_j = \hat{\beta}_j / s_j$
- Turn these into p values, p_j
- Apply `qvalue` to identify findings "significant" at a given FDR.

The Canonical Genomics Experiment

- Measure lots of things, with error
- Get estimates of effects β_j ($\hat{\beta}_j$) and their standard errors s_j
- Turn these into Z-scores, $z_j = \hat{\beta}_j / s_j$
- Turn these into p values, p_j
- Apply `qvalue` to identify findings "significant" at a given FDR.
- . . . ?

FDR, local fdr, and q values

Although precise definitions vary depending on whether one takes a Bayesian or Frequentist approach to the problem, roughly

- The FDR at a threshold P is

$$\text{FDR}(P) = \Pr(\beta_j = 0 | p_j < P).$$

FDR, local fdr, and q values

Although precise definitions vary depending on whether one takes a Bayesian or Frequentist approach to the problem, roughly

- The FDR at a threshold P is

$$\text{FDR}(P) = \Pr(\beta_j = 0 | p_j < P).$$

- The q value for observation j is $q_j = \text{FDR}(p_j)$.

FDR, local fdr, and q values

Although precise definitions vary depending on whether one takes a Bayesian or Frequentist approach to the problem, roughly

- The FDR at a threshold P is

$$\text{FDR}(P) = \Pr(\beta_j = 0 | p_j < P).$$

- The q value for observation j is $q_j = \text{FDR}(p_j)$.
- The local false discovery rate, fdr, at threshold P is

$$\text{fdr}(P) = \Pr(\beta_j = 0 | p_j = P).$$

FDR, local fdr, and q values

Although precise definitions vary depending on whether one takes a Bayesian or Frequentist approach to the problem, roughly

- The FDR at a threshold P is

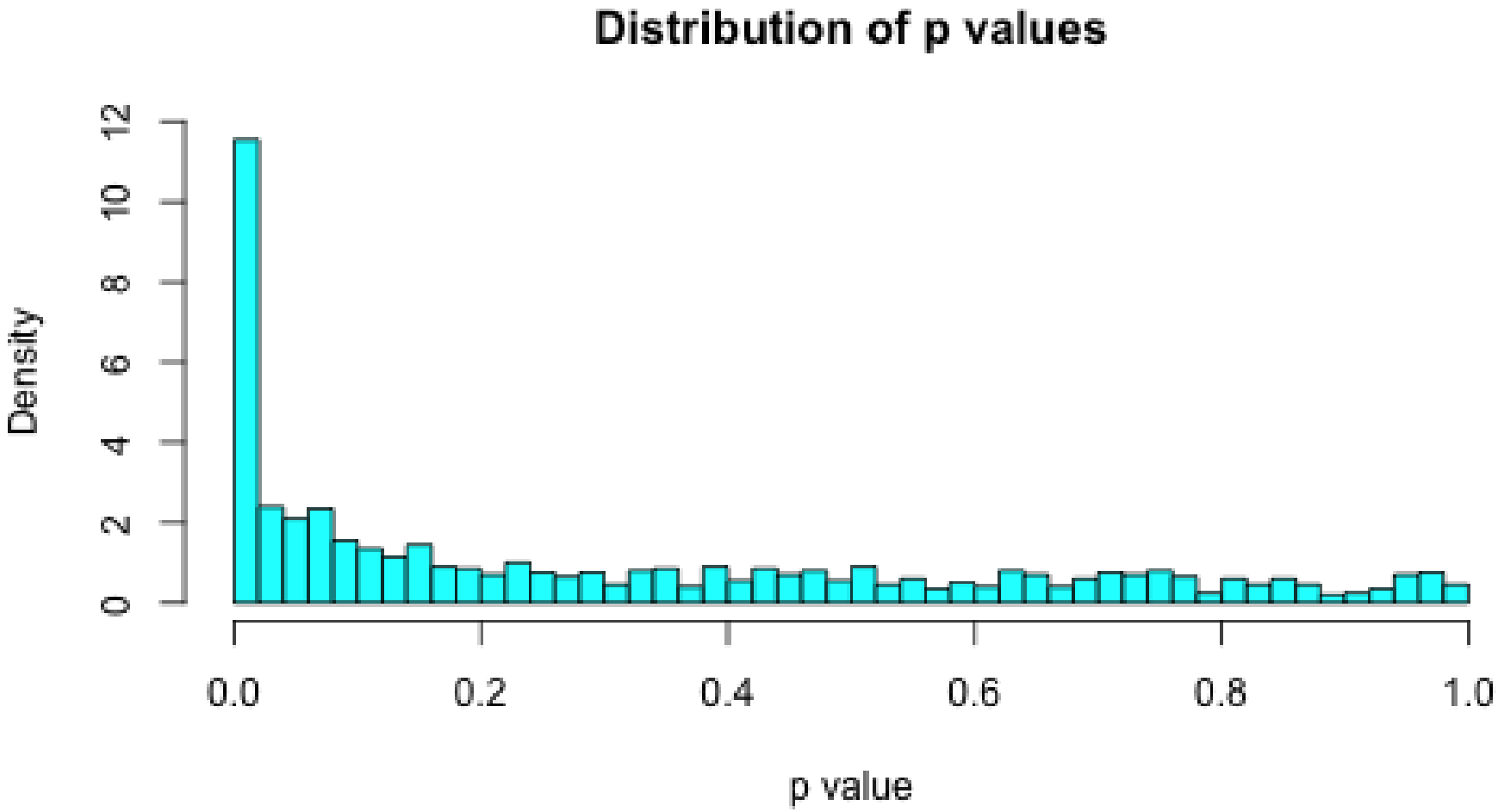
$$\text{FDR}(P) = \Pr(\beta_j = 0 | p_j < P).$$

- The q value for observation j is $q_j = \text{FDR}(p_j)$.
- The local false discovery rate, fdr, at threshold P is

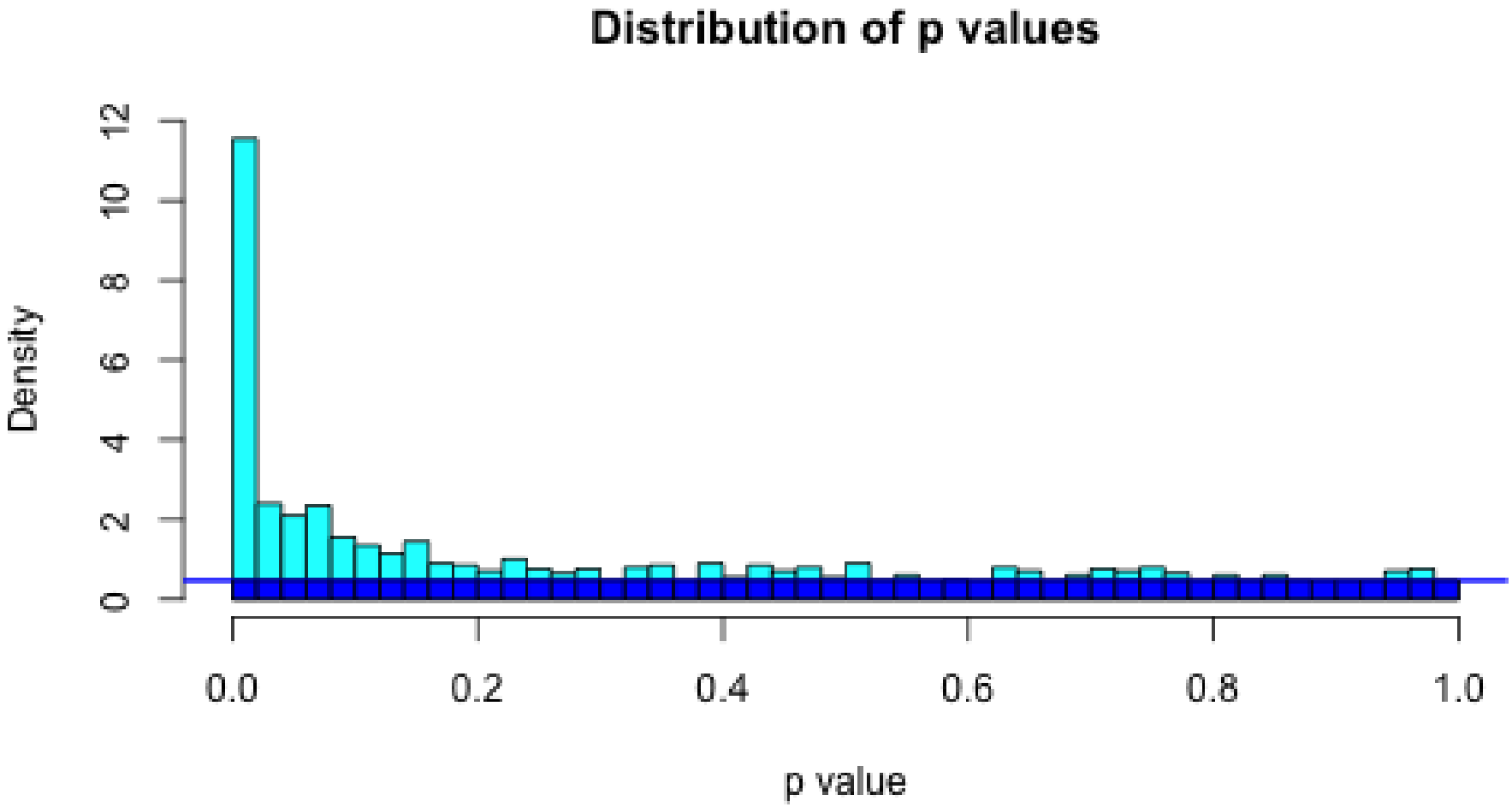
$$\text{fdr}(P) = \Pr(\beta_j = 0 | p_j = P).$$

- The fdr is more relevant, but slightly harder to estimate than FDR because it involves density estimation rather than tail-area estimation.

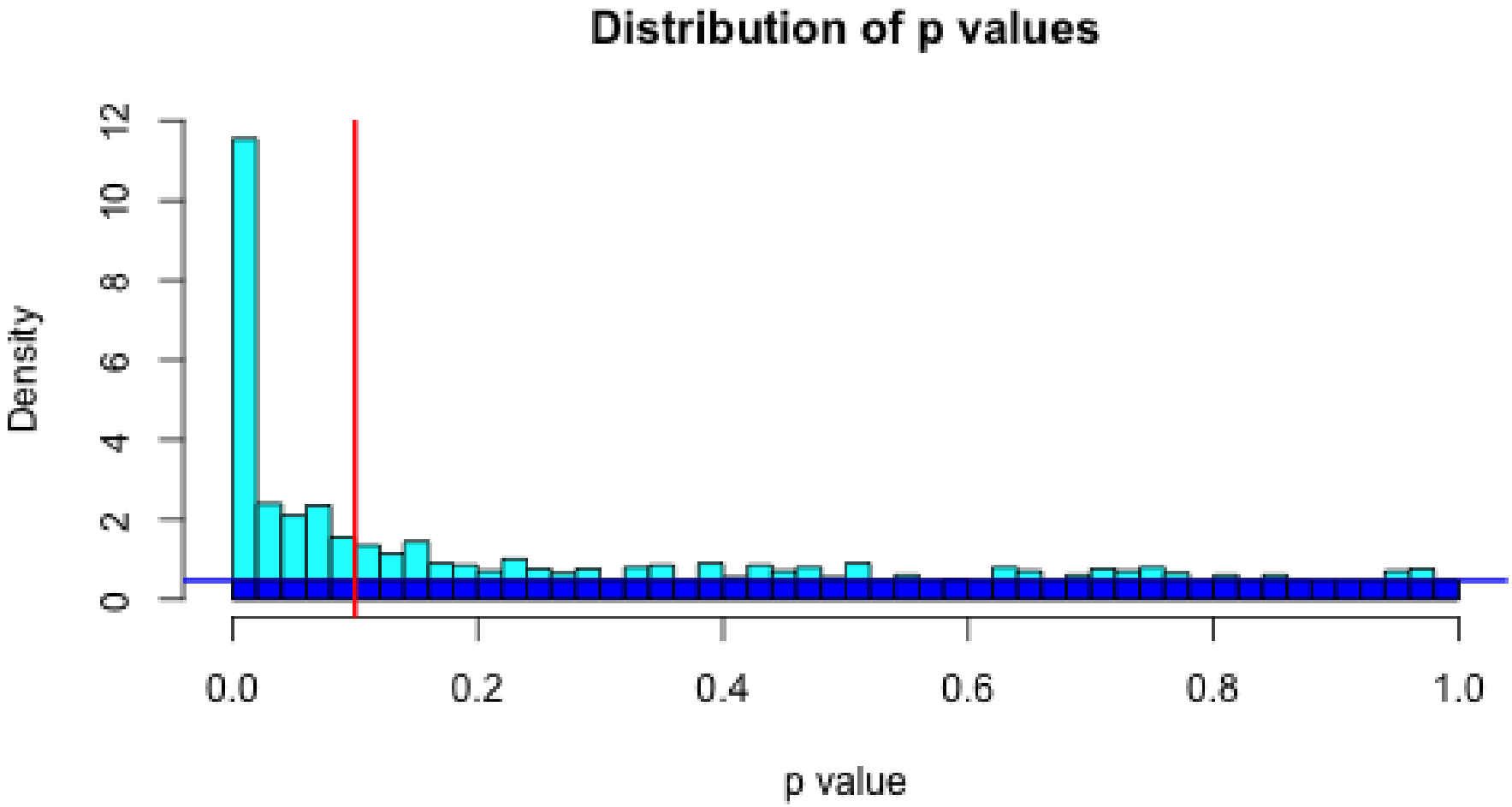
Example: FDR estimation



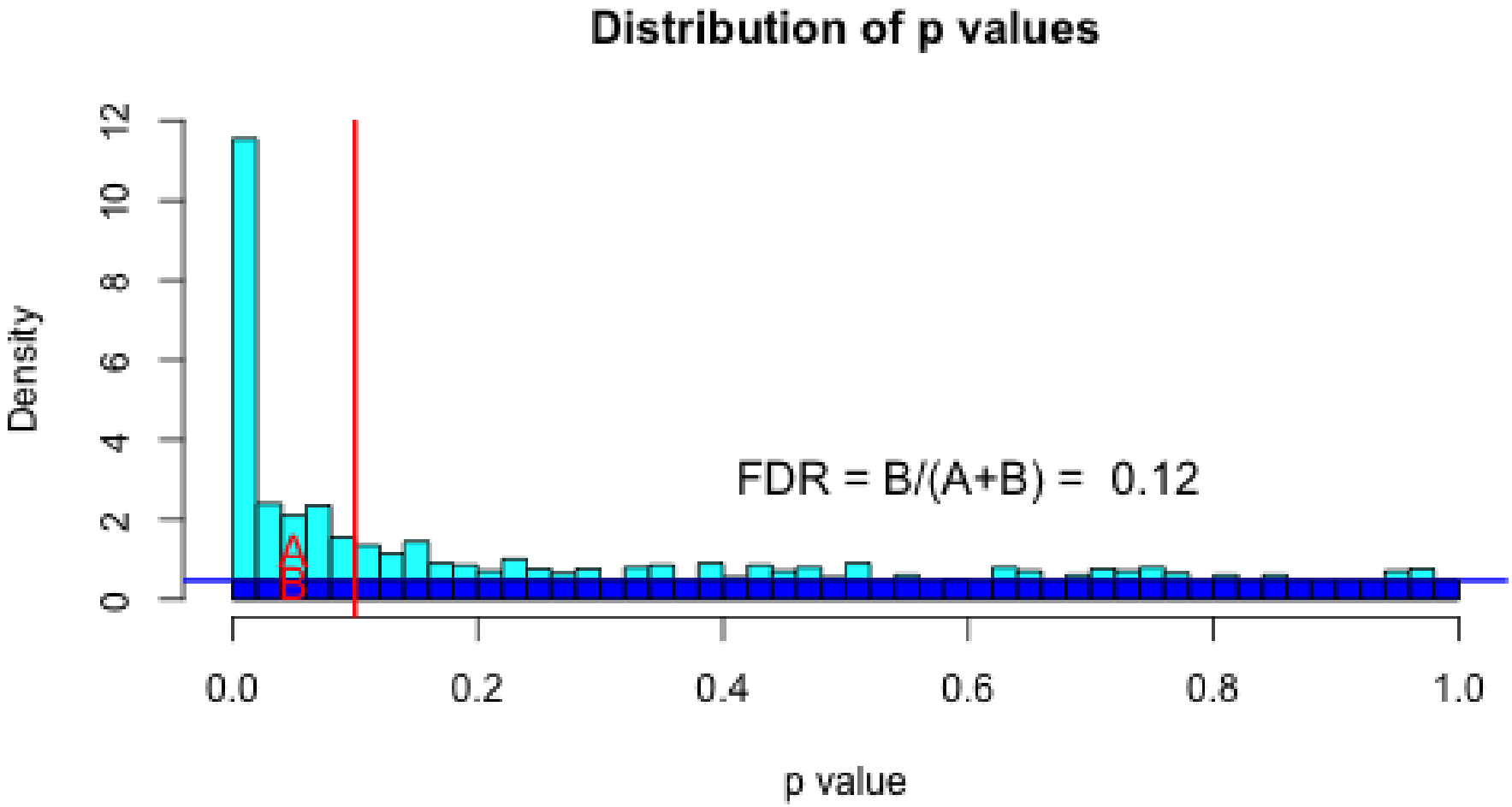
Example: FDR estimation



Example: FDR estimation



Example: FDR estimation



Is this an important problem?

- The original paper introducing FDR (Benjamini and Hochberg, 1995) has been cited 21,787 times (May 2014) according to Google Scholar.

Is this an important problem?

- The original paper introducing FDR (Benjamini and Hochberg, 1995) has been cited 21,787 times (May 2014) according to Google Scholar.
- That is three times a day for the last 19 years!

Problem 1: The Zero Assumption (ZA)

- The standard `qvalue` approach assumes that all the p values near 1 are null.

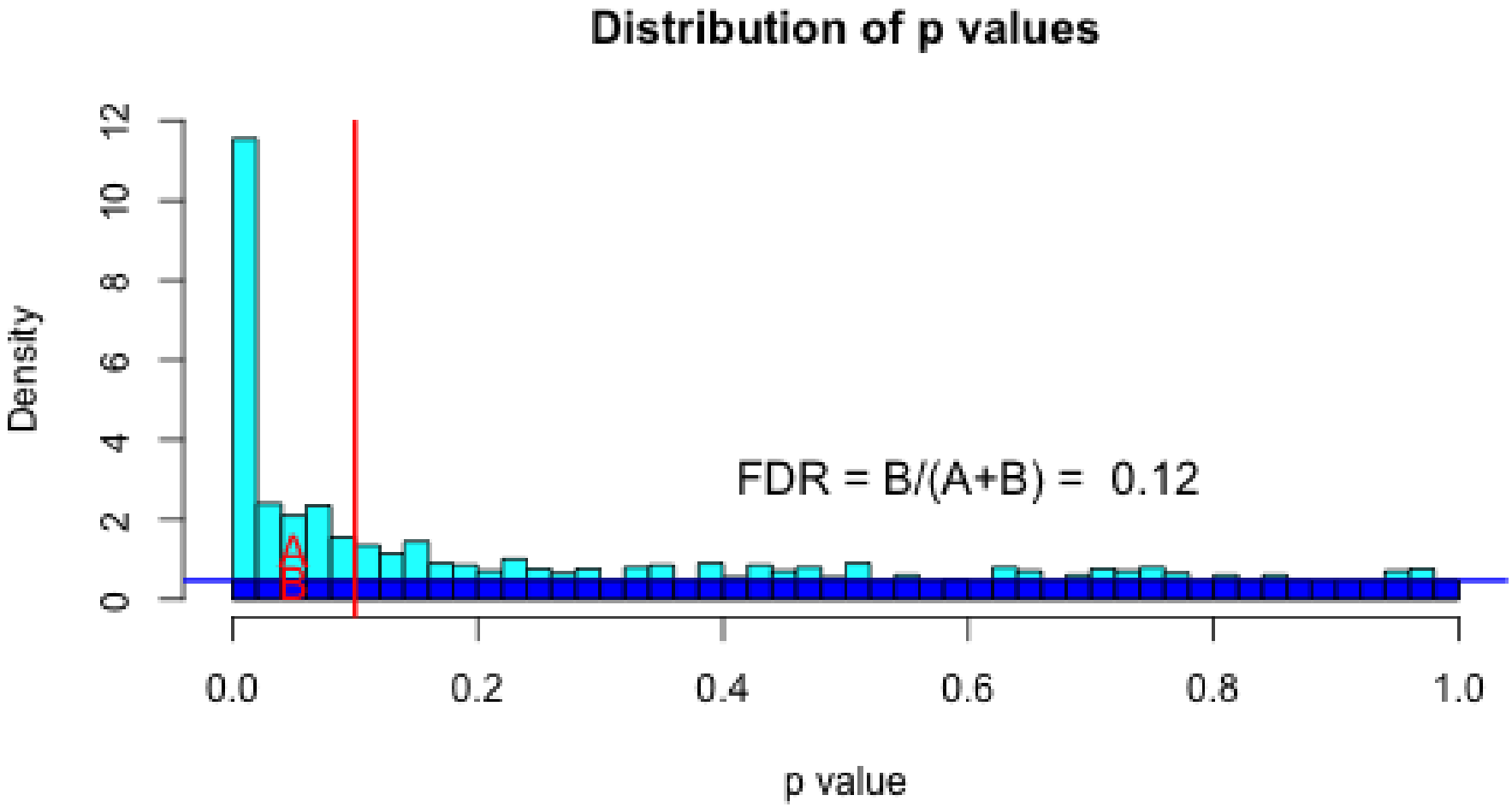
Problem 1: The Zero Assumption (ZA)

- The standard `qvalue` approach assumes that all the p values near 1 are null.
- Analogously, one can assume that all Z scores near 0 are null. Efron refers to this as the "Zero Assumption".

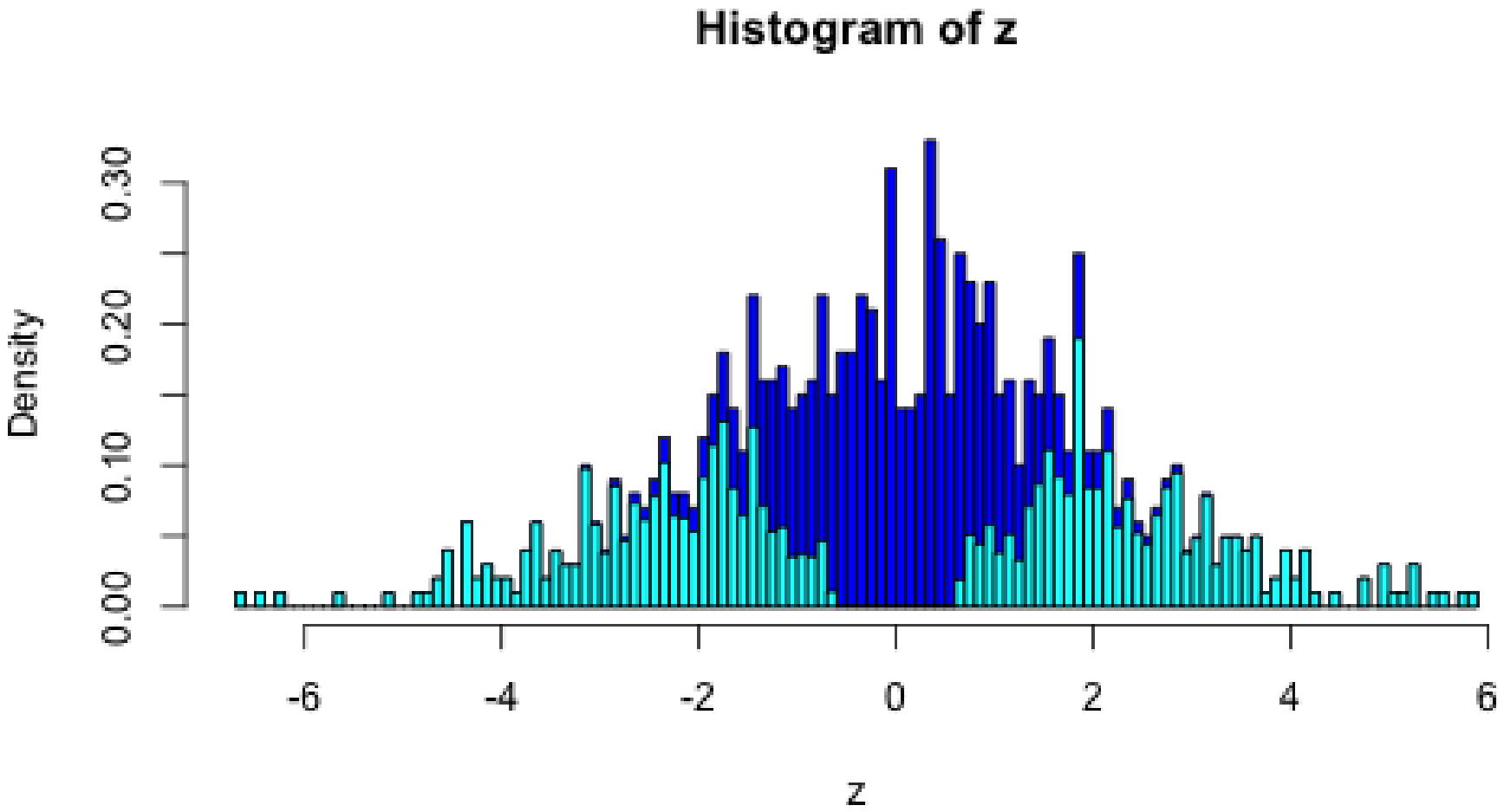
Problem 1: The Zero Assumption (ZA)

- The standard `qvalue` approach assumes that all the p values near 1 are null.
- Analogously, one can assume that all Z scores near 0 are null. Efron refers to this as the "Zero Assumption".
- Seems initially natural.

Implied distribution of p values under H_1



Implied distribution of Z scores under alternative



FDR problem 2: different measurement precision

- In some cases the measurement precisions differ among units

FDR problem 2: different measurement precision

- In some cases the measurement precisions differ among units
- Eg Expression levels of low-expressed genes have less precision than high-expressed genes

FDR problem 2: different measurement precision

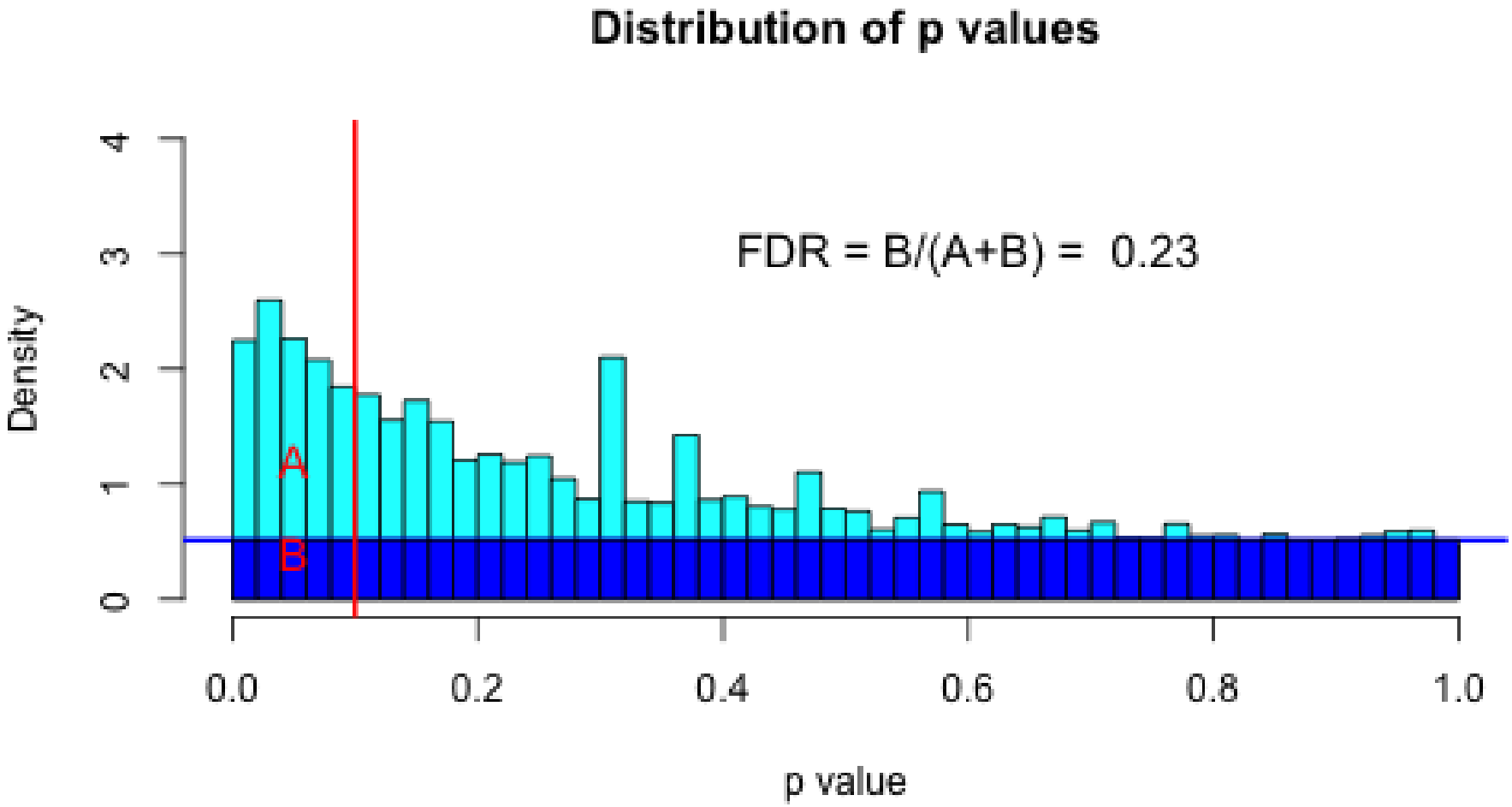
- In some cases the measurement precisions differ among units
- Eg Expression levels of low-expressed genes have less precision than high-expressed genes
- If some effects are measured less precisely than others, those tests "lack power" and dilute signal, increasing FDR

Example: Mouse Heart Data

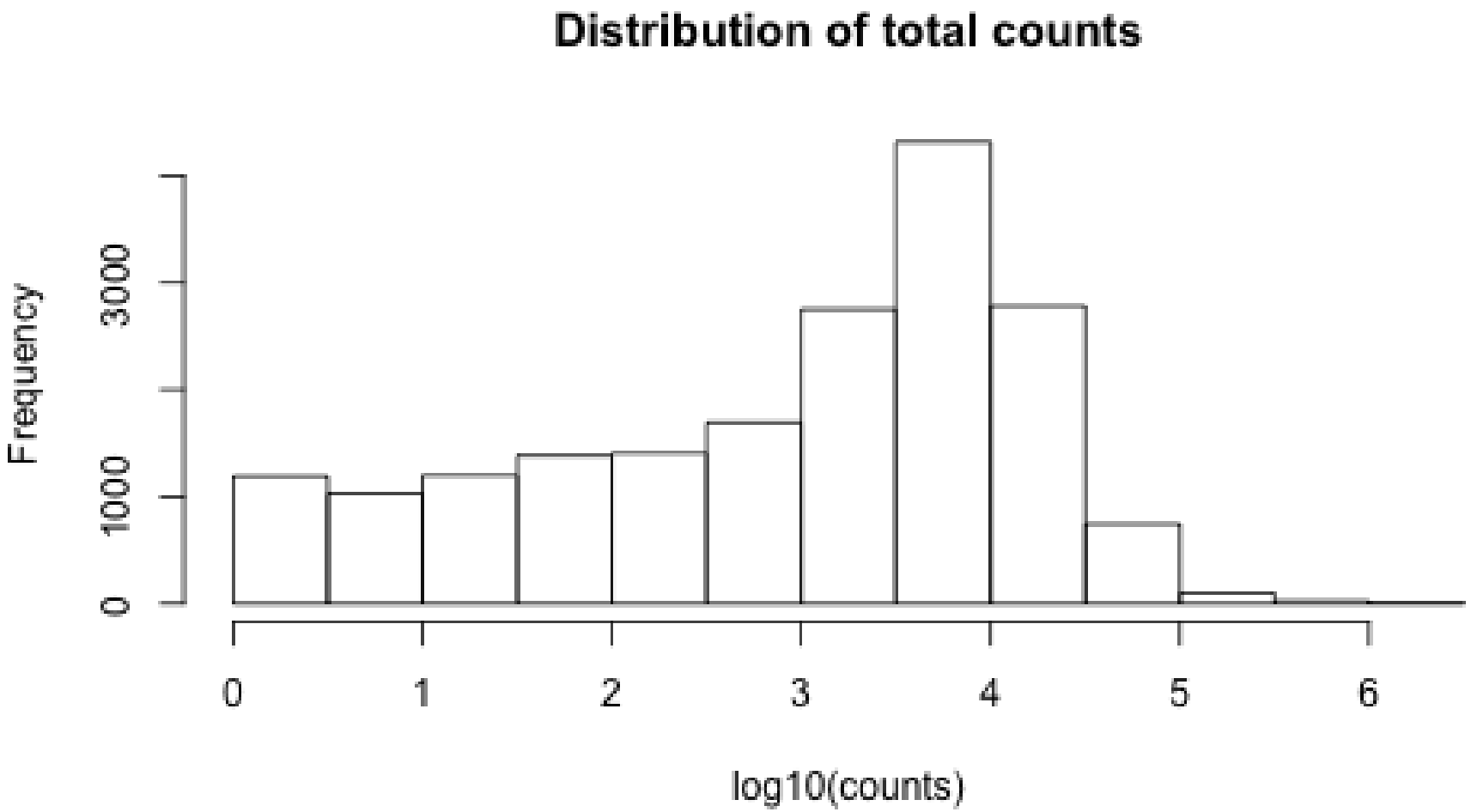
```
##      gene  lv1  lv2   rv1   rv2 genelength
## 1   Itm2a 2236 2174  9484 10883       1626
## 2  Sergef   97   90   341   408       1449
## 3 Fam109a  383  314  1864  2384       2331
## 4    Dhx9 2688 2631 18501 20879       4585
## 5   Ssu72  762  674  2806  3435       1446
## 8  Eif2b2  736  762  3081  3601       1565
```

- Data on 150 mouse hearts, dissected into left and right ventricle (courtesy Scott Schmemo, Marcelo Nobrega)

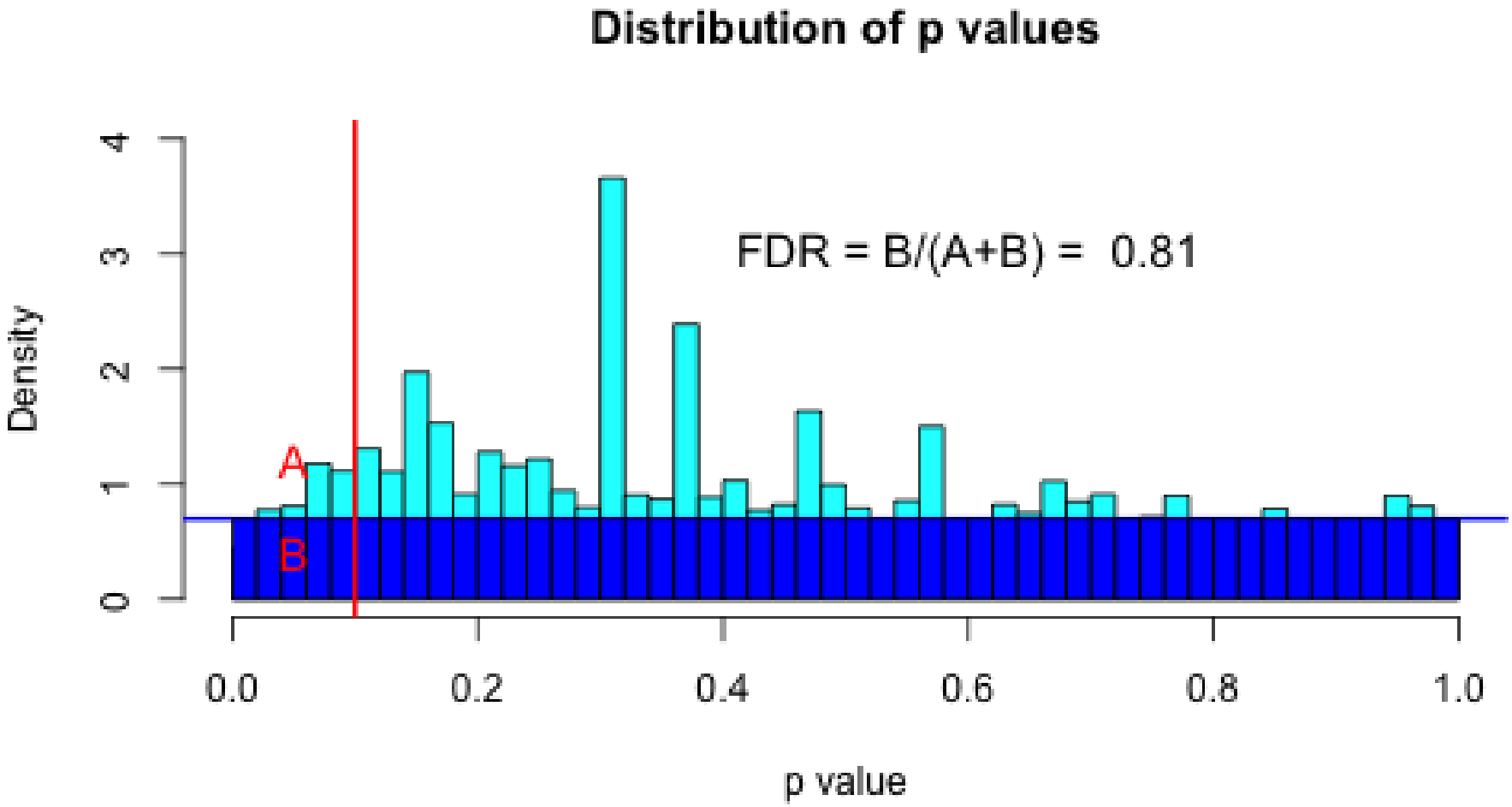
Example: Mouse Heart Data



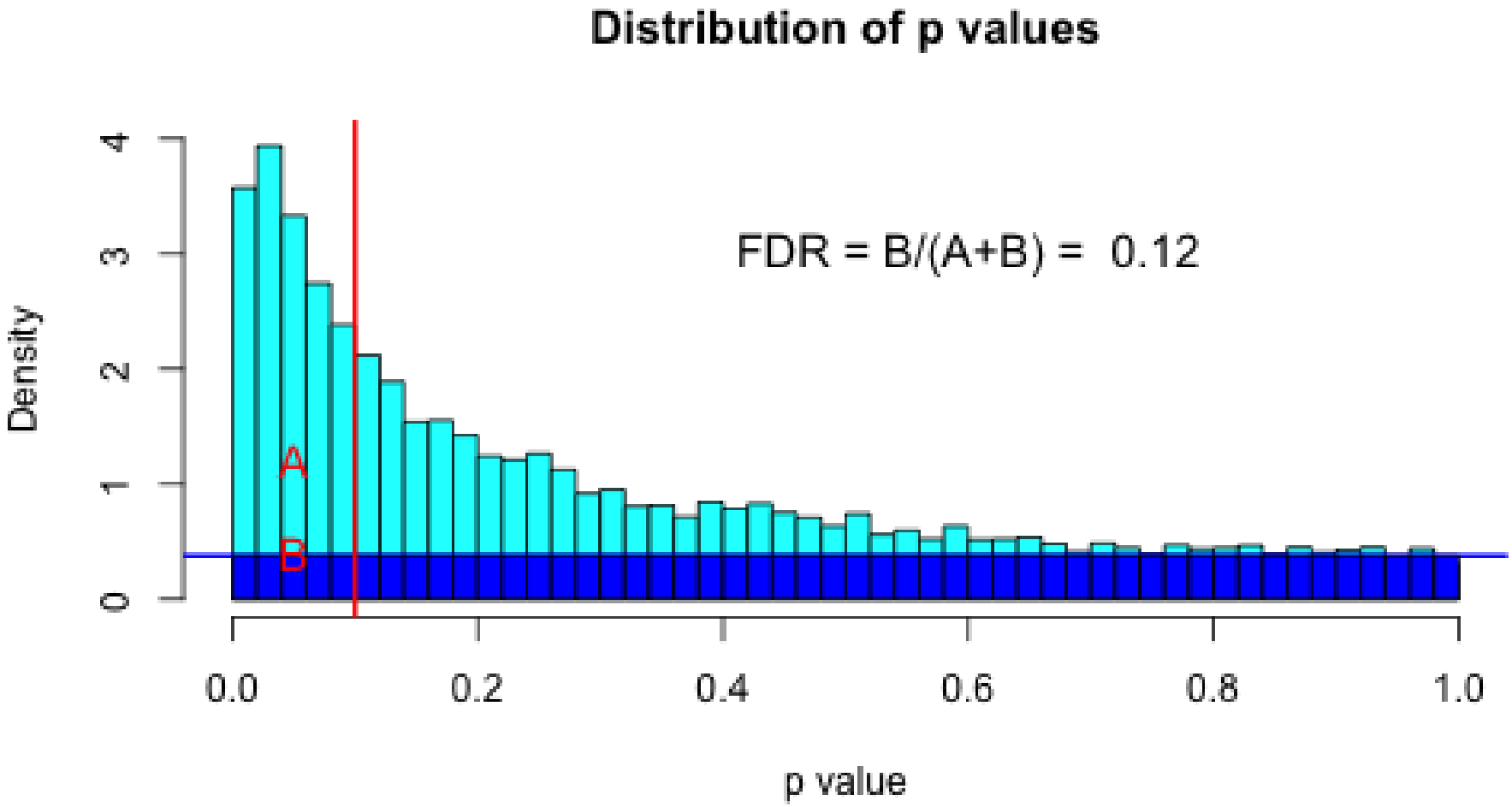
Mouse Data: Counts vary considerably across genes



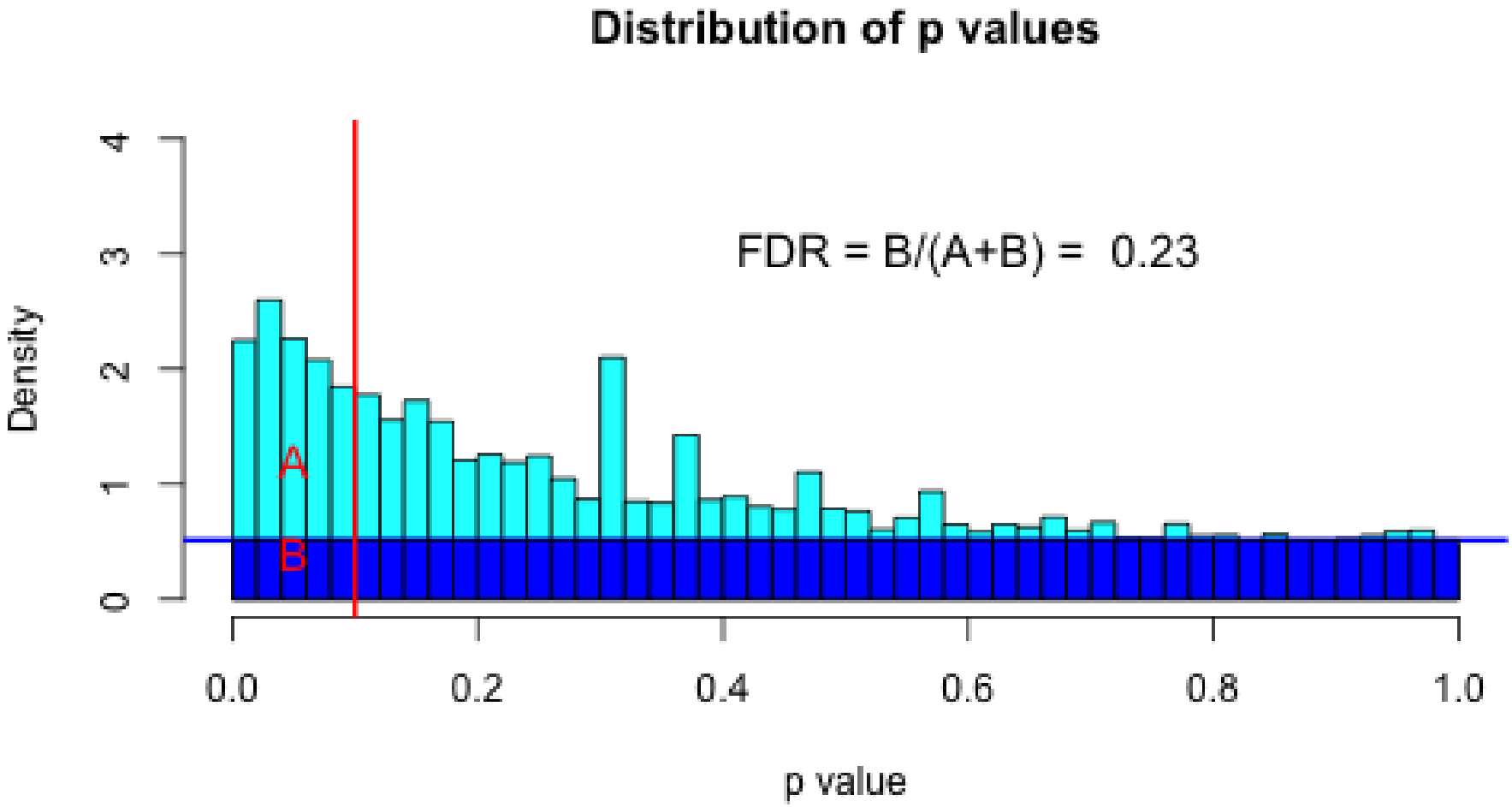
Lower count genes, less power



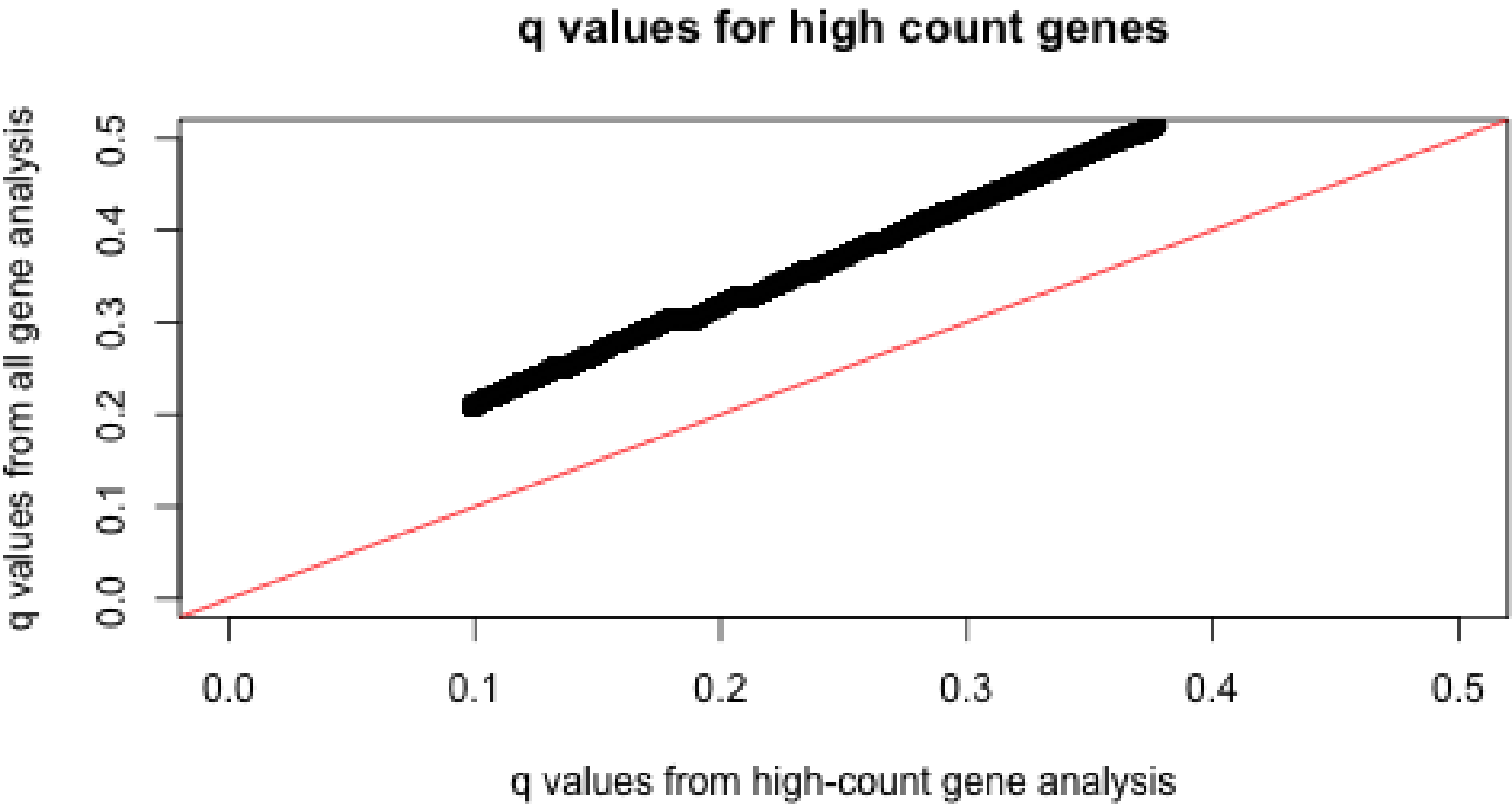
Higher count genes, more power



Low-count genes dilute signal at high-count genes



FDR problem 2: low count genes add noise, increase q values



Problems: Summary

Standard tools are unduly conservative.

- The ZA, which implies actual effects have a (probably unrealistic) bimodal distribution; causes overestimate of π_0, losing power.

Problems: Summary

Standard tools are unduly conservative.

- The ZA, which implies actual effects have a (probably unrealistic) bimodal distribution; causes overestimate of π_0, losing power.
- By focussing on p values, low-precision measurements can dilute high-precision measurements.

FDR via Empirical Bayes

- Following previous work (e.g. Newton, Efron, Muralidharan) we take an empirical Bayes approach to FDR.

FDR via Empirical Bayes

- Following previous work (e.g. Newton, Efron, Muralidharan) we take an empirical Bayes approach to FDR.
- Eg Efron assumes that the Z scores come from a mixture of null, and alternative:

$$Z_j \sim f_Z(.) = \pi_0 N(.; 0, 1) + (1 - \pi_0) f_1(.)$$

where f_1, π_0 are to be estimated from the data.

FDR via Empirical Bayes

- Following previous work (e.g. Newton, Efron, Muralidharan) we take an empirical Bayes approach to FDR.
- Eg Efron assumes that the Z scores come from a mixture of null, and alternative:

$$Z_j \sim f_Z(.) = \pi_0 N(.; 0, 1) + (1 - \pi_0) f_1(.)$$

where f_1, π_0 are to be estimated from the data.

- Various semi-parametric approaches taken to estimating f_1. For example, Efron uses Poisson regression; Muralidharan uses mixture of normal distributions.

FDR via Empirical Bayes

- Following previous work (e.g. Newton, Efron, Muralidharan) we take an empirical Bayes approach to FDR.
- Eg Efron assumes that the Z scores come from a mixture of null, and alternative:

$$Z_j \sim f_Z(.) = \pi_0 N(.; 0, 1) + (1 - \pi_0) f_1(.)$$

  where f_1, π_0 are to be estimated from the data.
- Various semi-parametric approaches taken to estimating f_1. For example, Efron uses Poisson regression; Muralidharan uses mixture of normal distributions.
- Once f_1 and π_0 estimated, FDR calculations are straightforward.

FDR: A New Deal

- Instead of modelling Z scores, model the effects β,

$$\beta_j \sim \pi_0 \delta_0(.) + (1 - \pi_0) g(.)$$

FDR: A New Deal

- Instead of modelling Z scores, model the effects β,

$$\beta_j \sim \pi_0 \delta_0(.) + (1 - \pi_0) g(.)$$

- Constrain g to be unimodal about 0; estimate g from data.

FDR: A New Deal

- Instead of modelling Z scores, model the effects β,

$$\beta_j \sim \pi_0 \delta_0(.) + (1 - \pi_0) g(.)$$

- Constrain g to be unimodal about 0; estimate g from data.
- *Incorporate precision* of each observation $\hat{\beta}$ into the likelihood. Specifically, approximate likelihood for β_j by a normal:

$$L(\beta_j) \propto \exp(-0.5(\beta_j - \hat{\beta}_j)^2 / s_j^2).$$

[From $\hat{\beta}_j \sim N(\beta_j, s_j)$] Or, better, use a t likelihood if s_j estimated using few observations.

FDR - A New Deal

- A convenient way to model g: mixture of 0-centered normal distributions:

$$g(\beta; \pi) = \sum_{k=1}^{K} \pi_k N(\beta; 0, \sigma_k^2)$$

FDR - A New Deal

- A convenient way to model g: mixture of 0-centered normal distributions:

$$g(\beta; \pi) = \sum_{k=1}^{K} \pi_k N(\beta; 0, \sigma_k^2)$$

- Estimating g comes down to estimating π. Joint estimation of π_0, π easy by maximum likelihood (EM algorithm).

FDR - A New Deal

- A convenient way to model g: mixture of 0-centered normal distributions:

$$g(\beta;\pi) = \sum_{k=1}^{K} \pi_k N(\beta; 0, \sigma_k^2)$$

- Estimating g comes down to estimating π. Joint estimation of π_0, π easy by maximum likelihood (EM algorithm).
- By allowing K large, and σ_k to span a dense grid of values, we get a flexible unimodal symmetric distribution.

FDR - A New Deal

- A convenient way to model g: mixture of 0-centered normal distributions:

$$g(\beta;\pi) = \sum_{k=1}^{K} \pi_k N(\beta; 0, \sigma_k^2)$$

- Estimating g comes down to estimating π. Joint estimation of π_0, π easy by maximum likelihood (EM algorithm).
- By allowing K large, and σ_k to span a dense grid of values, we get a flexible unimodal symmetric distribution.
- Can approximate, arbitrarily closely, any scale mixture of normals. Includes almost all priors used for sparse regression problems (spike-and-slab, double exponential/Laplace/Bayesian Lasso, horseshoe).

FDR - A New Deal

- Alternatively, a mixture of uniforms, with 0 as one end-point of the range, provides still more flexibility, and in particular allows for asymmetry.

FDR - A New Deal

- Alternatively, a mixture of uniforms, with 0 as one end-point of the range, provides still more flexibility, and in particular allows for asymmetry.
- If allow a very large number of uniforms this provides the non-parametric mle for g; cf Grenander 1953; Cordy + Thomas 1997.

Illustration: g a mixture of 0-centered normals

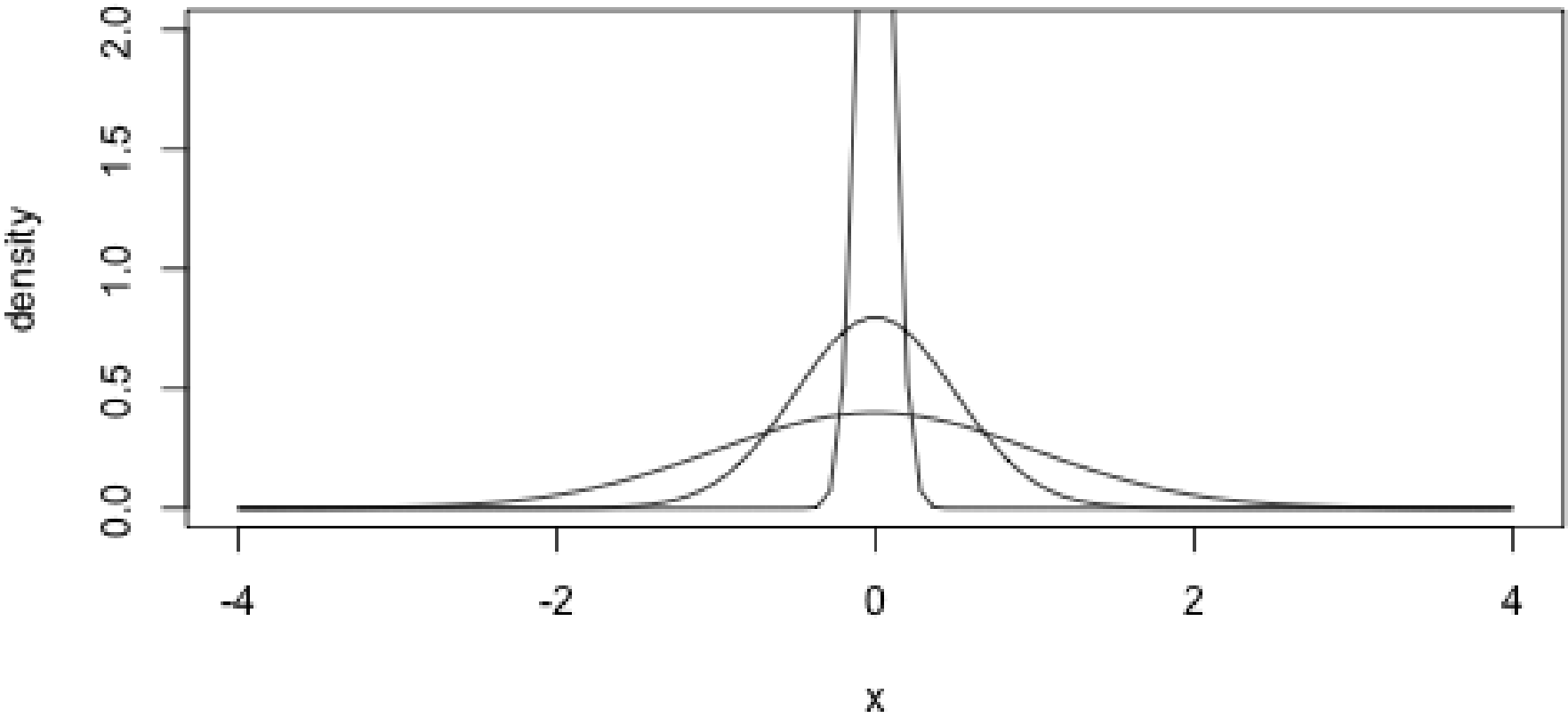


Illustration: g a mixture of 0-centered normals

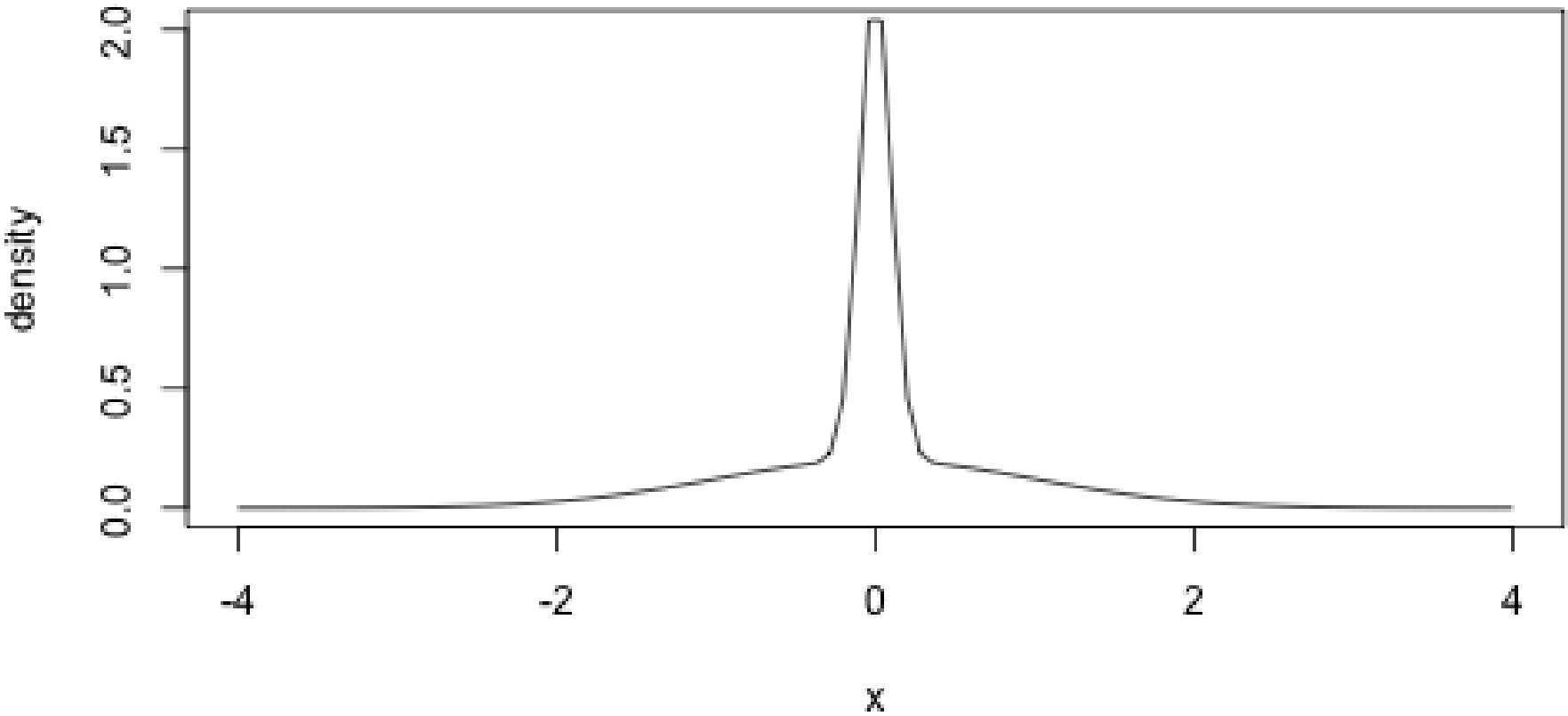


Illustration: g a mixture of 0-anchored uniforms

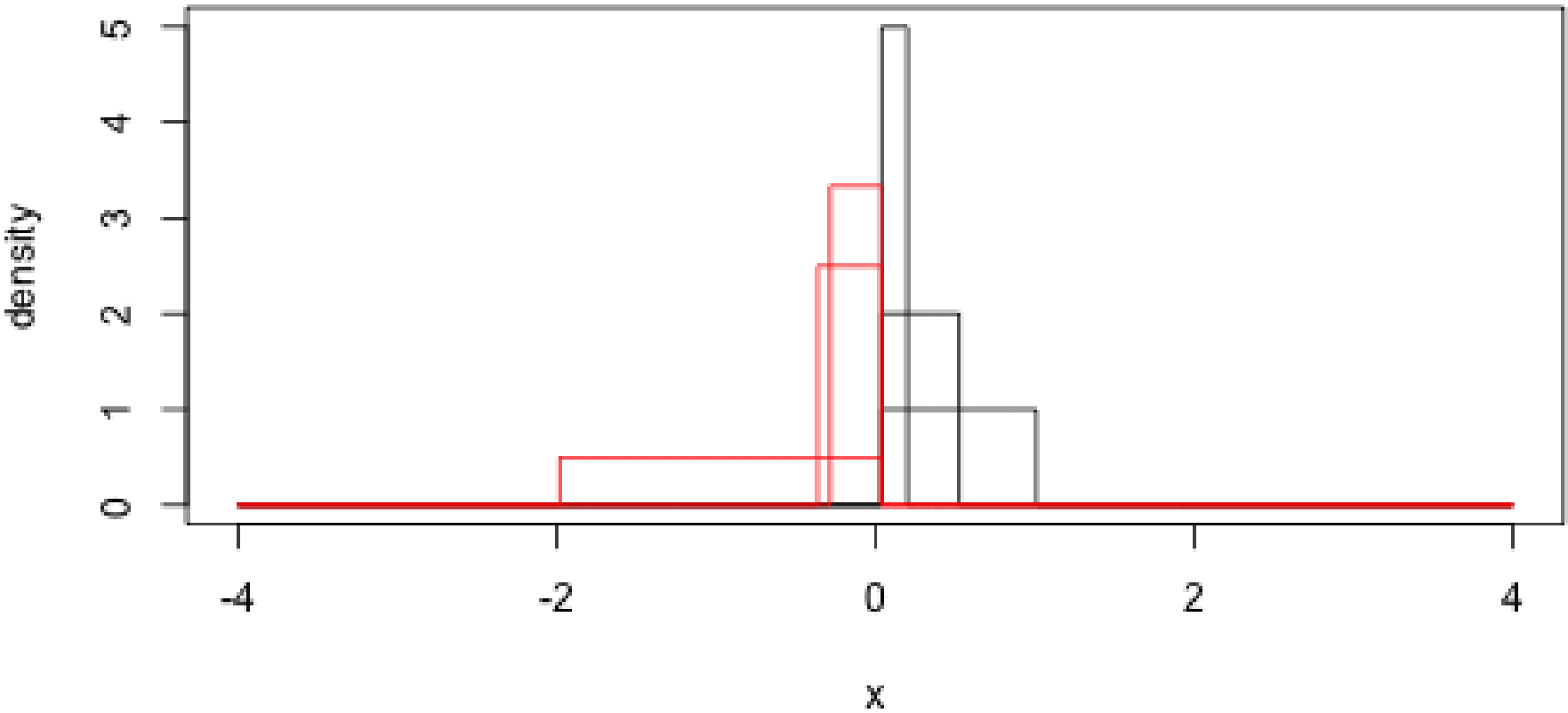
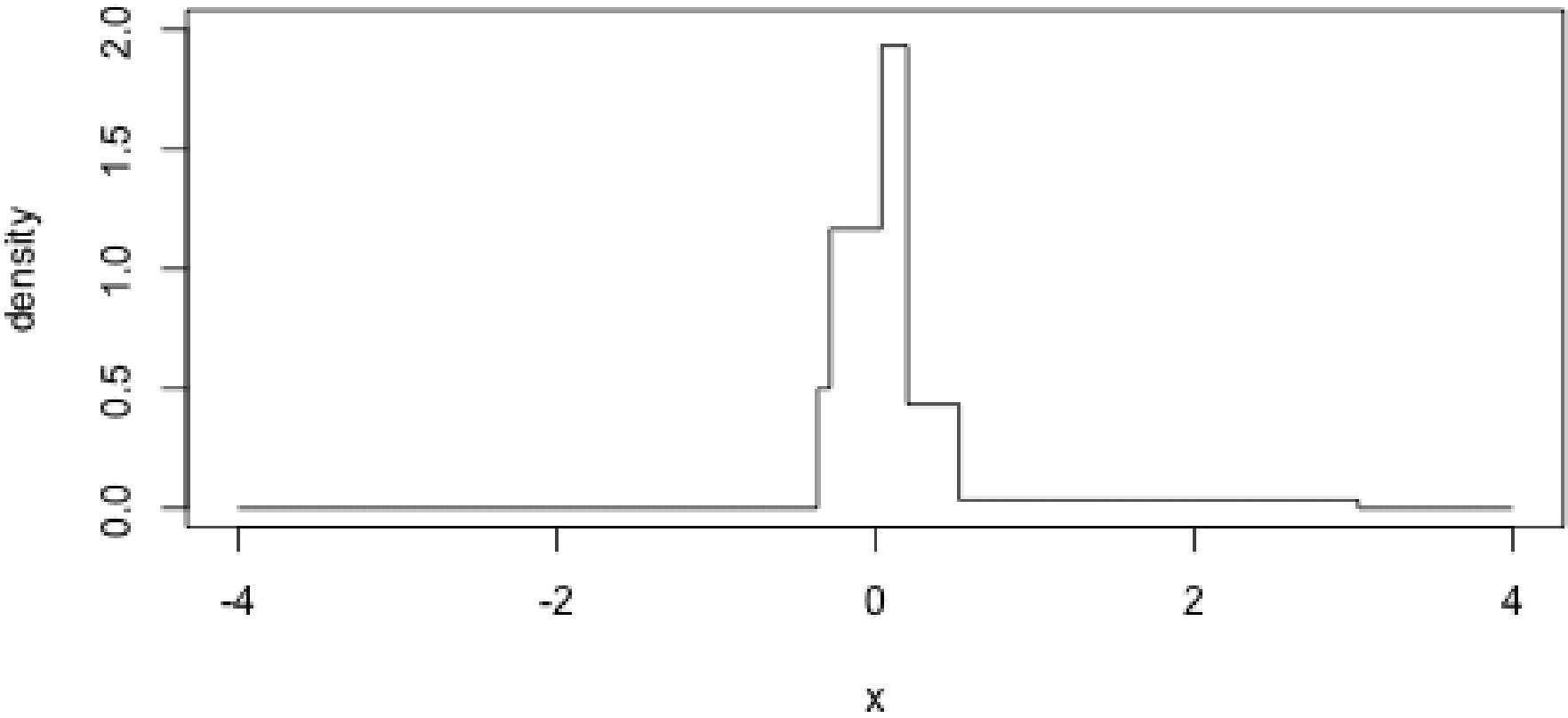


Illustration: g a mixture of 0-anchored uniforms



Adaptive Shrinkage

- This approach actually provides a full posterior distribution for each β_j.

Adaptive Shrinkage

- This approach actually provides a full posterior distribution for each β_j.
- So easy to obtain point estimates and credible intervals.

Adaptive Shrinkage

- This approach actually provides a full posterior distribution for each β_j.
- So easy to obtain point estimates and credible intervals.
- Because $g(\beta)$ is unimodal, the point estimates (and CIs) will tend to be "shrunk" towards the overall mean (0).

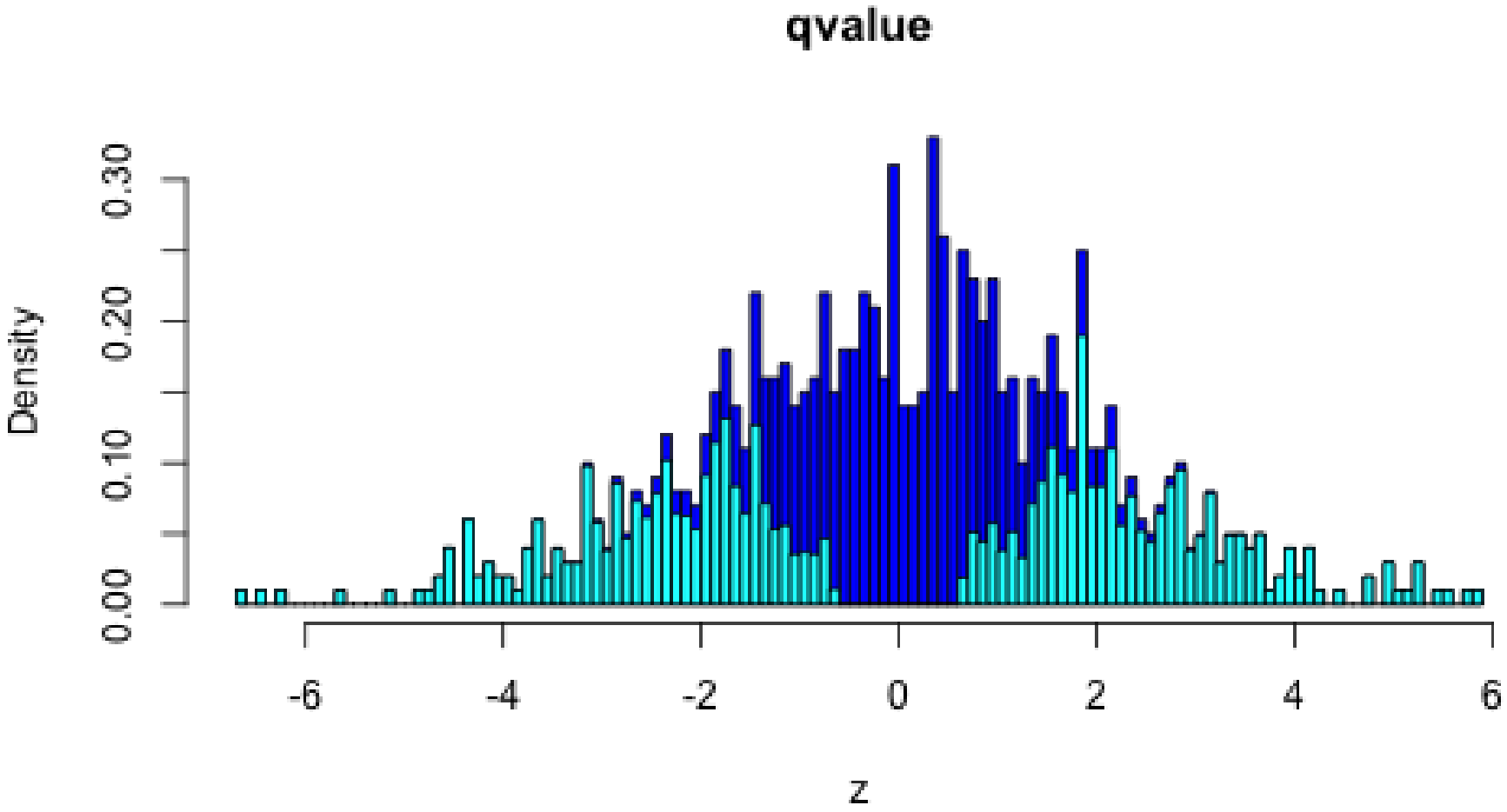
Adaptive Shrinkage

- This approach actually provides a full posterior distribution for each β_j.
- So easy to obtain point estimates and credible intervals.
- Because $g(\beta)$ is unimodal, the point estimates (and CIs) will tend to be "shrunk" towards the overall mean (0).
- Because $g(\beta)$ is estimated from the data, the amount of shrinkage is adaptive to signal in the data. And because of the role of s_j, the amount of shrinkage adapts to the information on each observation.

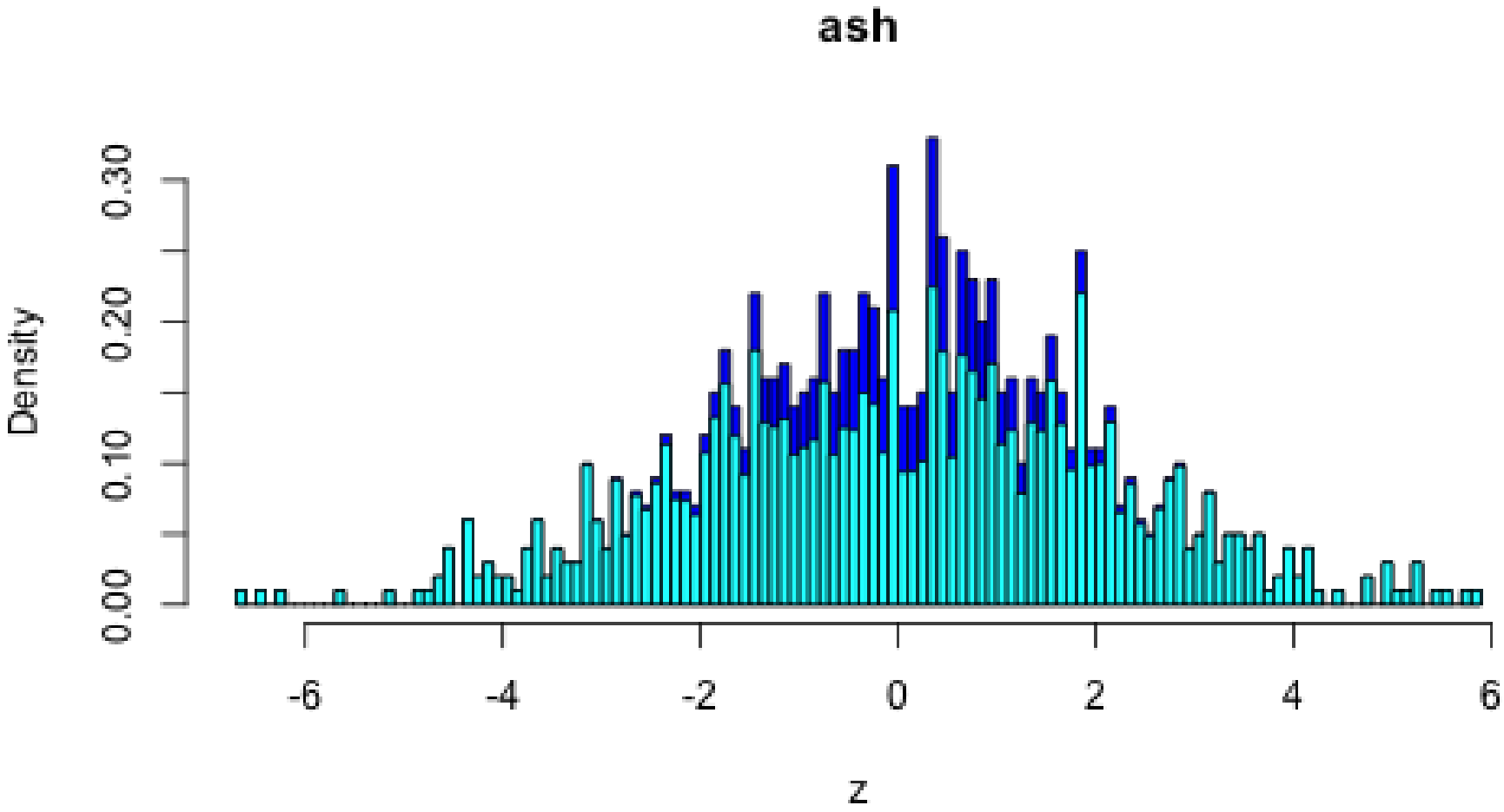
Adaptive Shrinkage

- This approach actually provides a full posterior distribution for each β_j.
- So easy to obtain point estimates and credible intervals.
- Because $g(\beta)$ is unimodal, the point estimates (and CIs) will tend to be "shrunk" towards the overall mean (0).
- Because $g(\beta)$ is estimated from the data, the amount of shrinkage is adaptive to signal in the data. And because of the role of s_j, the amount of shrinkage adapts to the information on each observation.
- So we call the approach "Adaptive Shrinkage" (ASH).

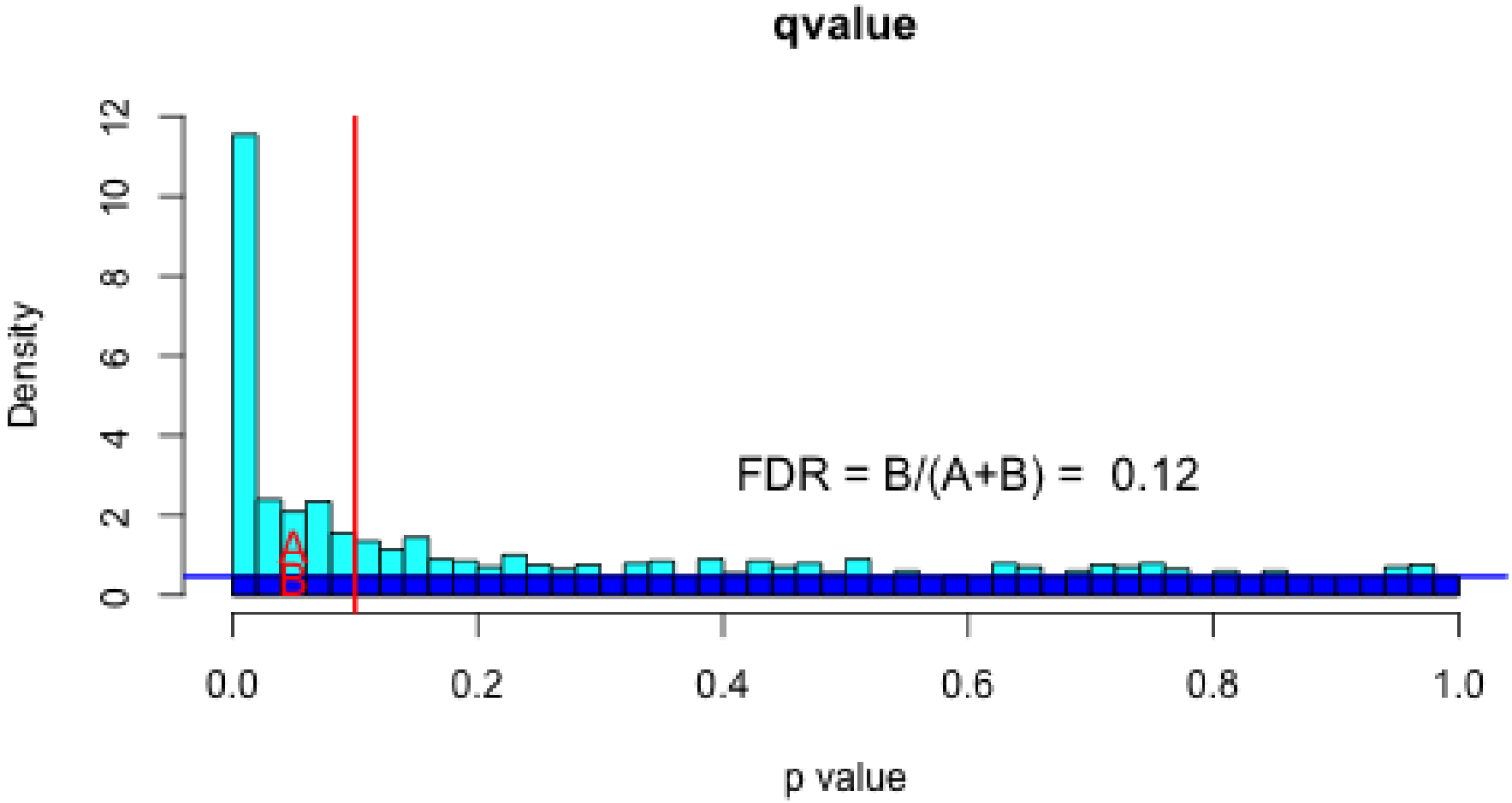
Recall Problem 1: distribution of alternative Z values multimodal



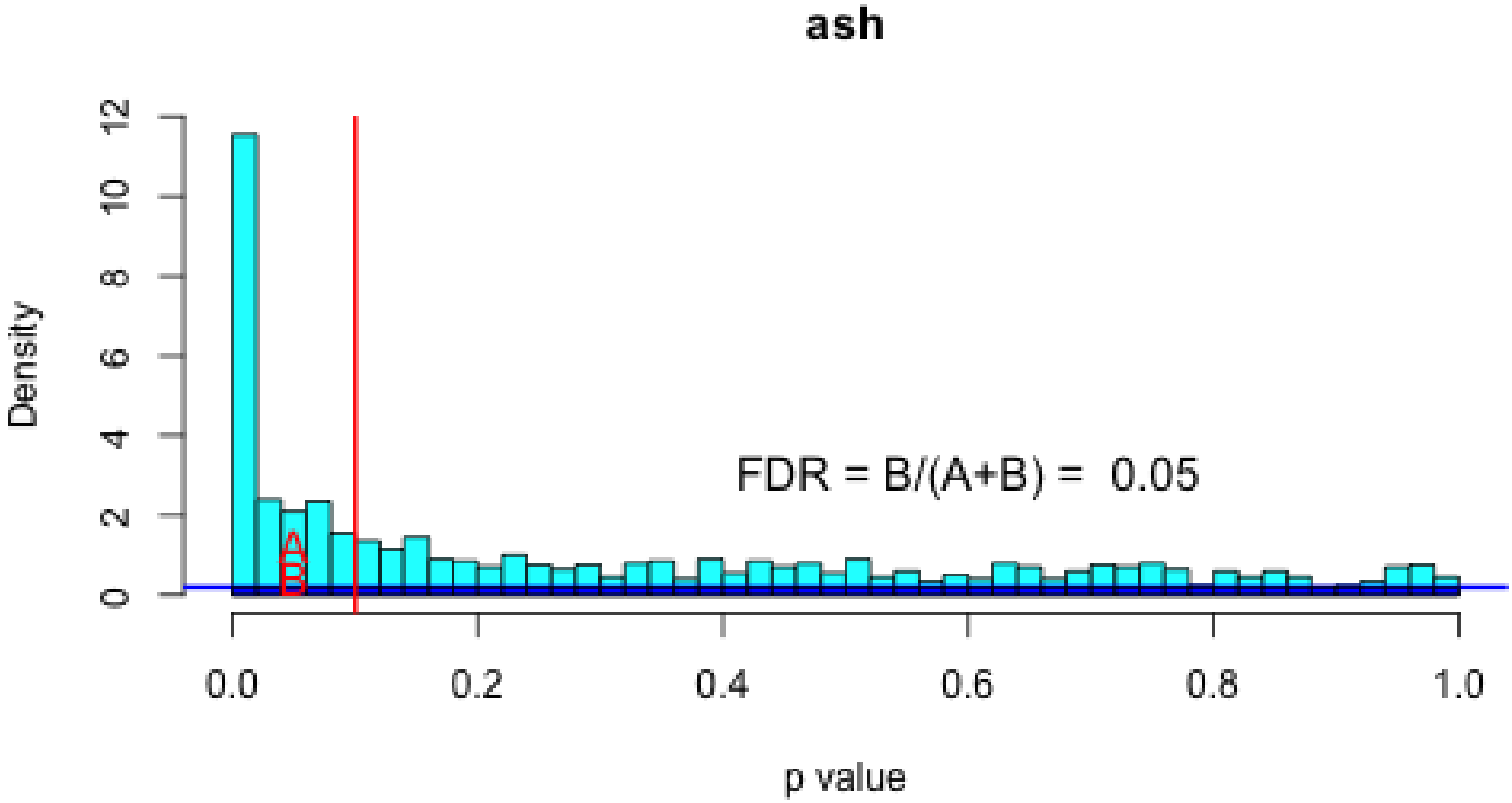
Problem Fixed: distribution of alternative Z values unimodal



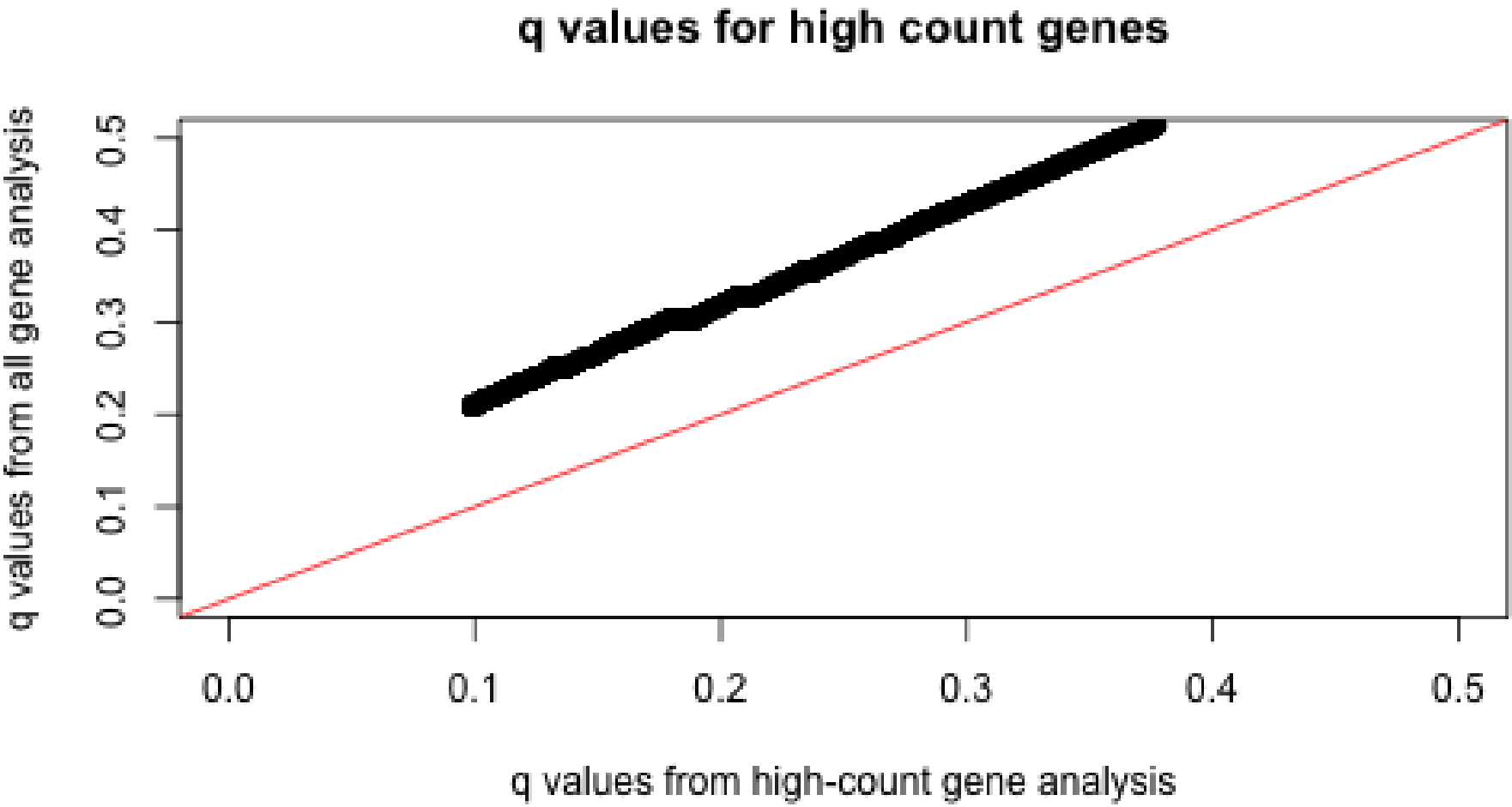
Example: FDR estimation



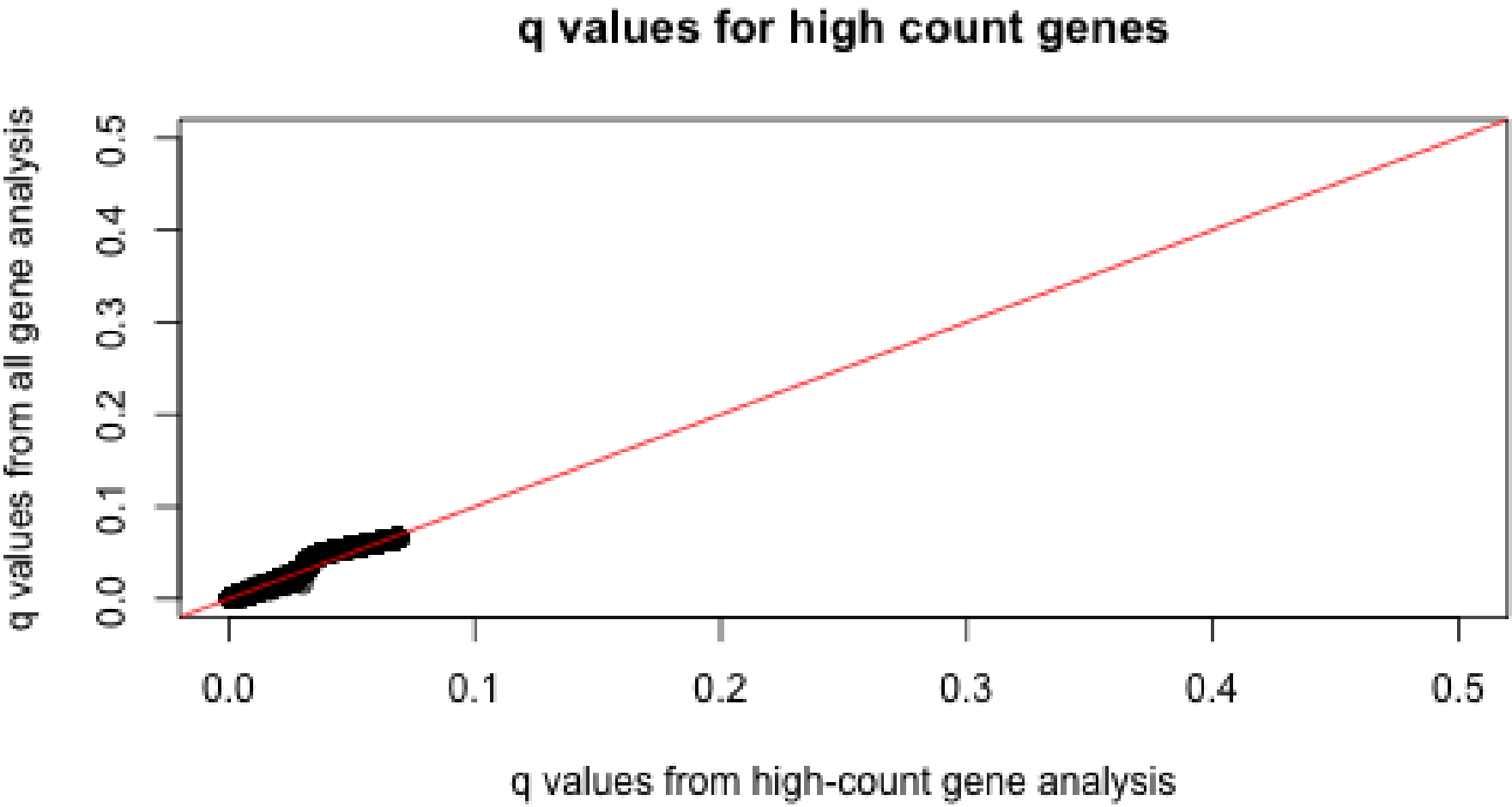
Example: FDR estimation



Recall Problem 2: low count genes add noise, increase q values



Problem Fixed: incorporating precision reduces influence of low-count genes



A new problem: an embarrassment of riches

- If the null is mostly false, the new approach can provide unsettling results

A new problem: an embarrassment of riches

- If the null is mostly false, the new approach can provide unsettling results
- The FDR can be small for all observations, even those with $p \approx 1$!

A new problem: an embarrassment of riches

- If the null is mostly false, the new approach can provide unsettling results
- The FDR can be small for all observations, even those with $p \approx 1$!
- In the illustrative example, the maximum q value is 0.18

Perhaps we didn't really understand the question?

- Problem arises only if we insist on asking question "is $\beta_j = 0$?"

Perhaps we didn't really understand the question?

- Problem arises only if we insist on asking question "is $\beta_j = 0$?"
- Given enough signal, we become convinced that very few of the $\beta_j = 0$

Perhaps we didn't really understand the question?

- Problem arises only if we insist on asking question "is $\beta_j = 0$?"
- Given enough signal, we become convinced that very few of the $\beta_j = 0$
- But for some β_j we still may have little information about actual value

Perhaps we didn't really understand the question?

- Problem arises only if we insist on asking question "is $\beta_j = 0$?"
- Given enough signal, we become convinced that very few of the $\beta_j = 0$
- But for some β_j we still may have little information about actual value
- Suggests a change of focus: ask for which β_j are we confident about the sign (cf Gelman et al, 2012).

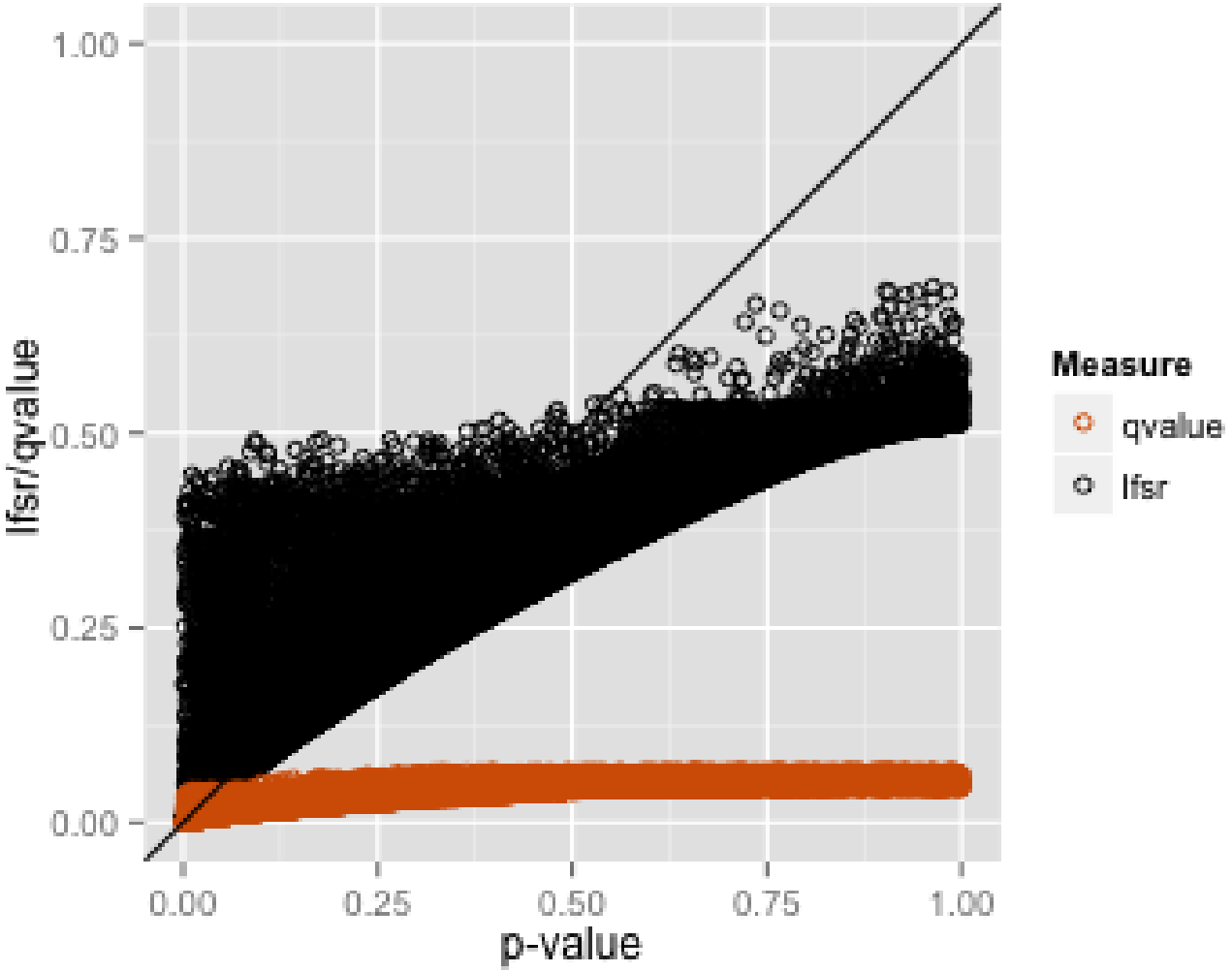
The False Sign Rate

- Suggestion: replace FDR with local false sign rate (lfsr), the probability that if we say an effect is positive (negative), it is not.

The False Sign Rate

- Suggestion: replace FDR with local false sign rate (lfsr), the probability that if we say an effect is positive (negative), it is not.
- Example: suppose we estimate that $\Pr(\beta_j < 0) = 0.95, \Pr(\beta_j = 0) = 0.025$ and $\Pr(\beta_j > 0) = 0.025$. Then we report β_j as a "(negative) discovery", and estimate its lfsr as 0.05.

Even with many signals, large p values have high lfsr



Advantages of False Sign Rate

- The FDR is typically not actually identifiable from data.

Advantages of False Sign Rate

- The FDR is typically not actually identifiable from data.
- This is because data cannot distinguish between $\beta_j = 0$ and β_j "very small". So π_0 is not identifiable, and FDR is very sensitive to π_0.

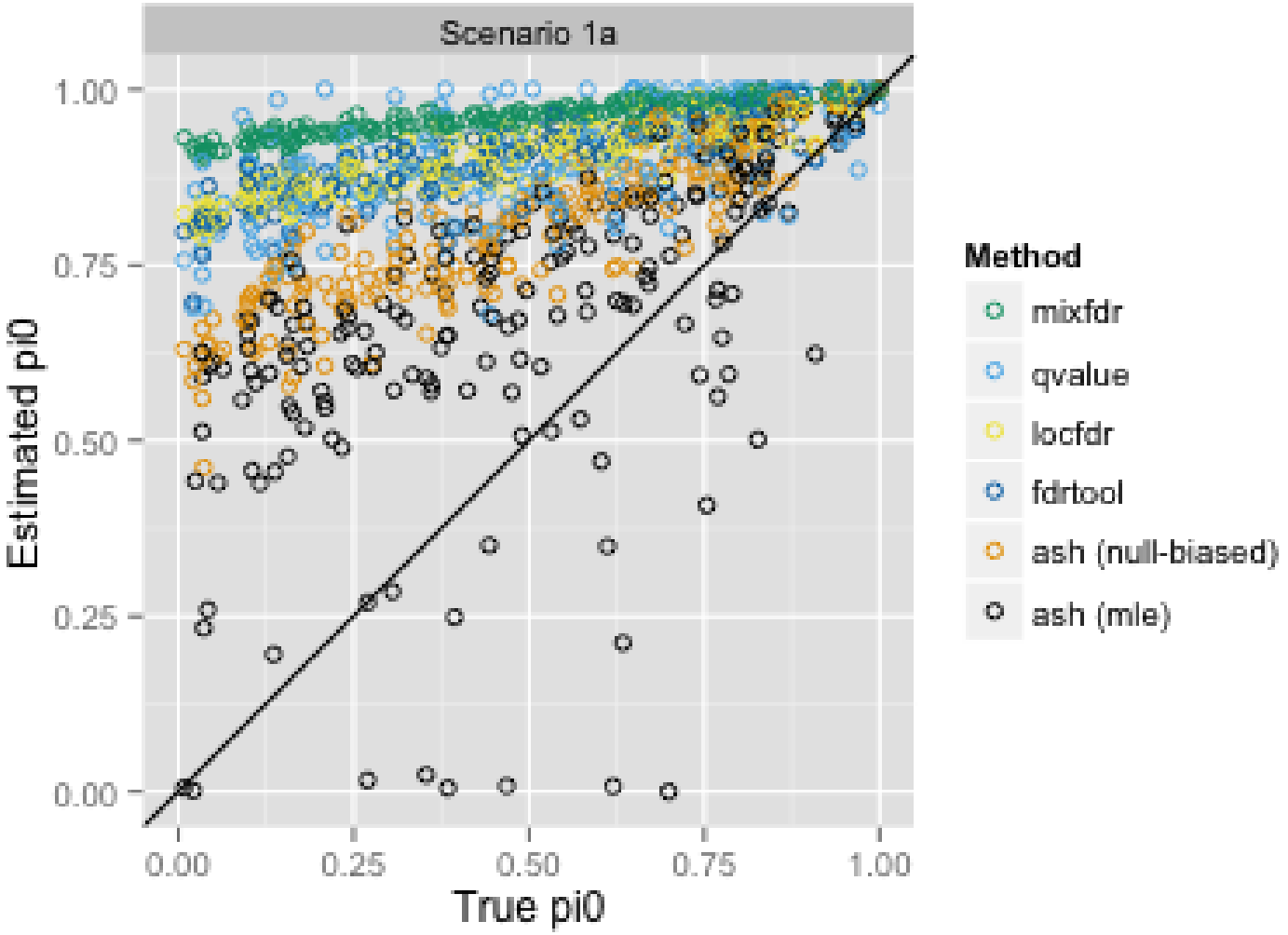
Advantages of False Sign Rate

- The FDR is typically not actually identifiable from data.
- This is because data cannot distinguish between $\beta_j = 0$ and β_j "very small". So π_0 is not identifiable, and FDR is very sensitive to π_0.
- So methods for estimating π_0 and FDR, including those presented here, are designed to be "conservative" (i.e. overestimate the FDR).

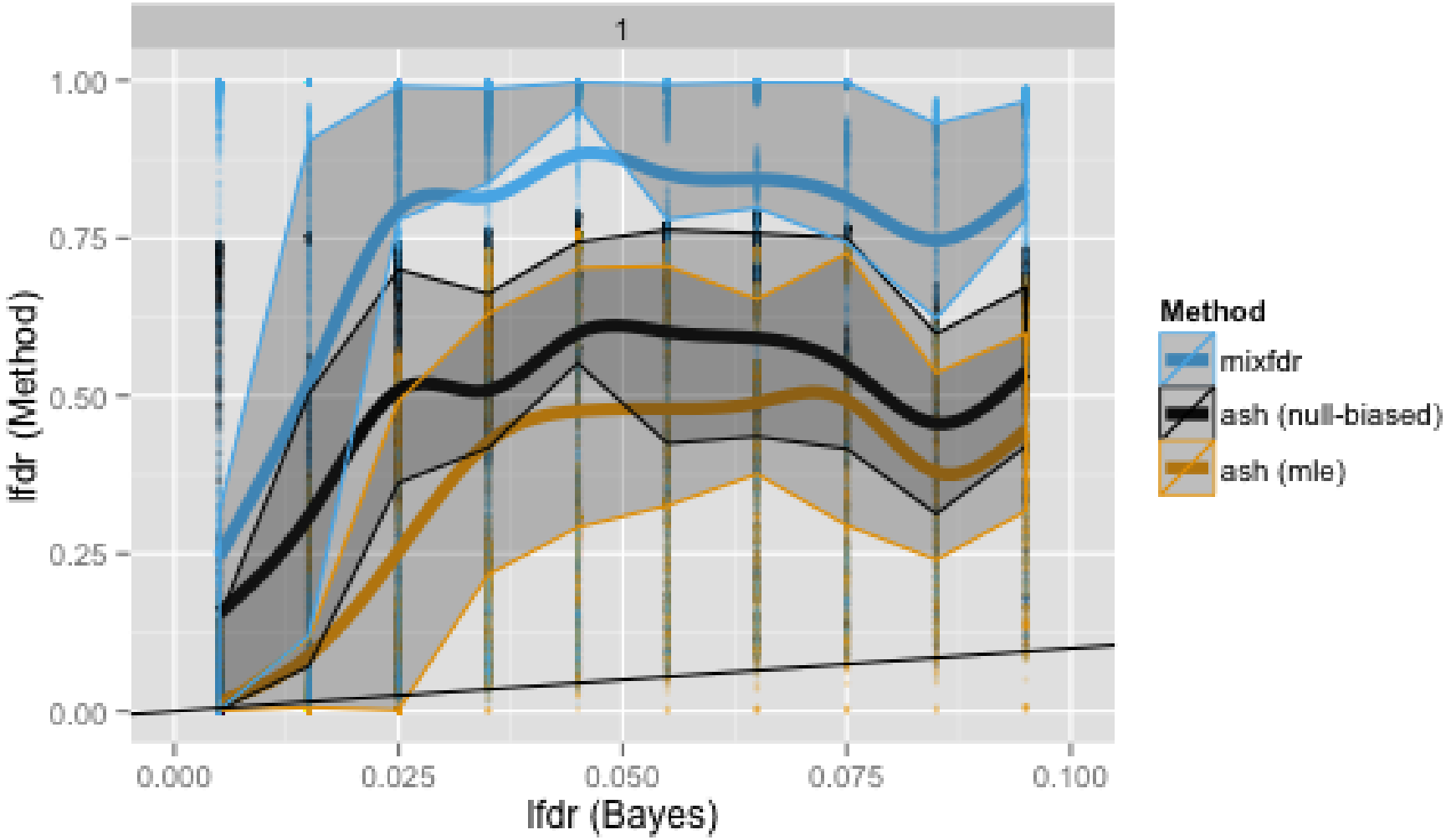
Advantages of False Sign Rate

- The FDR is typically not actually identifiable from data.
- This is because data cannot distinguish between $\beta_j = 0$ and β_j "very small". So π_0 is not identifiable, and FDR is very sensitive to π_0.
- So methods for estimating π_0 and FDR, including those presented here, are designed to be "conservative" (i.e. overestimate the FDR).
- The False Sign Rate is much less senstive to π_0, and hence more identifiable from data!

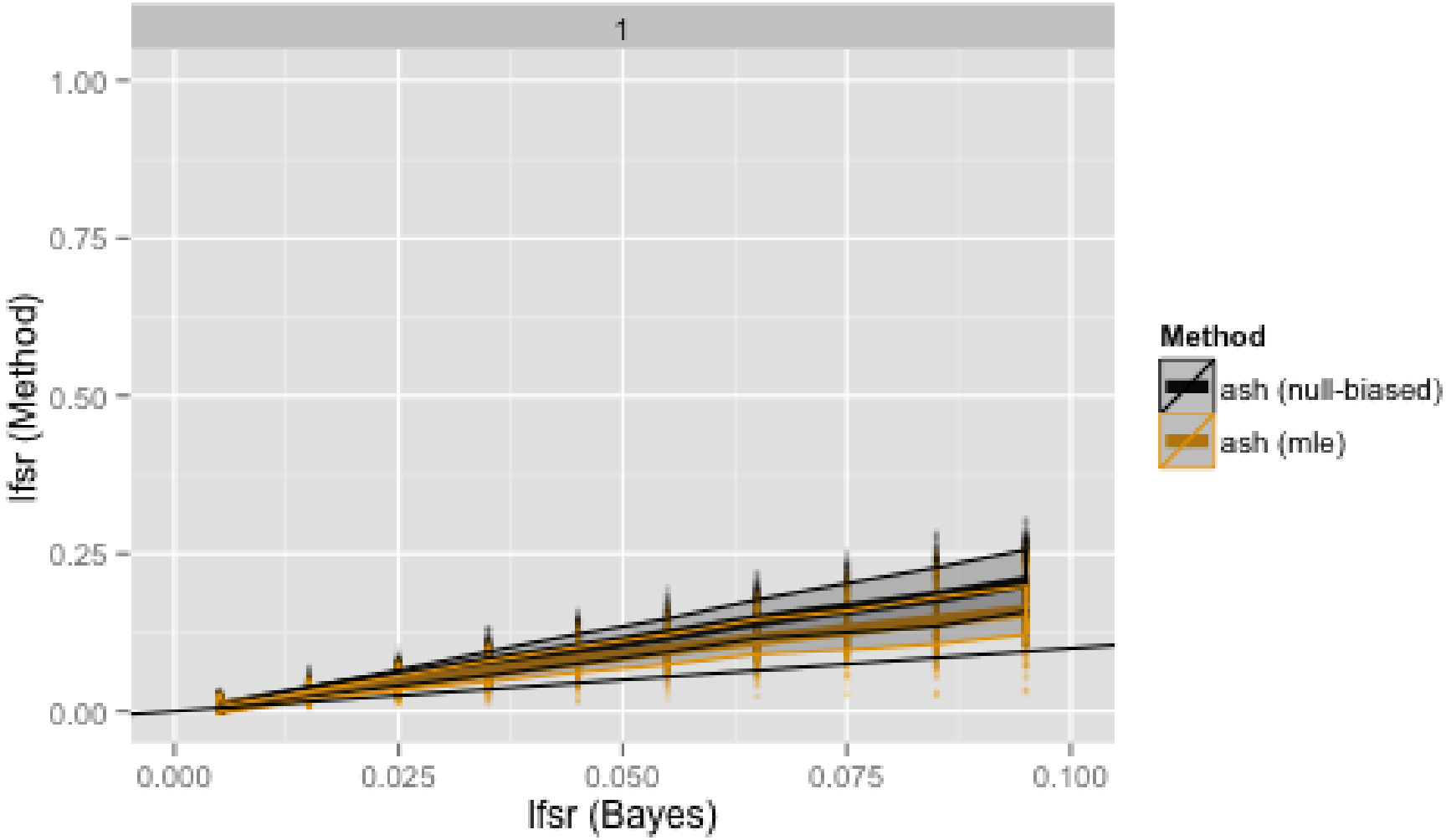
Simulated example: π_0 not identifiable.



Simulated Example: so fdr not identifiable



Simulated Example: fsr much more identifiable



Summary

- ASH provides a generic approach to shrinkage estimation, as well as false discovery (sign) rates.

Summary

- ASH provides a generic approach to shrinkage estimation, as well as false discovery (sign) rates.
- But by using two numbers $(\hat{\beta}, s)$ instead of one (p values or z scores) varying precision of measurements is better accounted for.

Summary

- ASH provides a generic approach to shrinkage estimation, as well as false discovery (sign) rates.
- But by using two numbers $(\hat{\beta}, s)$ instead of one (p values or z scores) varying precision of measurements is better accounted for.
- Unimodal assumption for effects reduces conservatism

Summary

- ASH provides a generic approach to shrinkage estimation, as well as false discovery (sign) rates.
- But by using two numbers $(\hat{\beta}, s)$ instead of one (p values or z scores) varying precision of measurements is better accounted for.
- Unimodal assumption for effects reduces conservatism
- False Sign Rate preferable to False Discovery Rate: more identifiable, and better representation of information in data for "high-signal" situations.

Other Applications

- Widely applicable: requiring only an estimated effect size and standard error for each object.

Other Applications

- Widely applicable: requiring only an estimated effect size and standard error for each object.
- E.g. Currently applying it to wavelet shrinkage applications.

Next steps?

- Incorporate shrinkage of variances and not just means. (e.g. "moderated t test", Smyth et al)

Next steps?

- Incorporate shrinkage of variances and not just means. (e.g. "moderated t test", Smyth et al)
- Allow $g(\cdot; \pi)$ to depend on covariates X.

Next steps?

- Incorporate shrinkage of variances and not just means. (e.g. "moderated t test", Smyth et al)
- Allow $g(\cdot; \pi)$ to depend on covariates X.
- Allow for correlations in the measured $\hat{\beta}_j$.

Thanks

- to the developers of **R**, **knitr**, **Rstudio** and **Pandoc**.

Thanks

- to the developers of **R**, **knitr**, **Rstudio** and **Pandoc**.
- to the several postdoctoral researchers and students who have worked with me on related topics.

Thanks

- to the developers of **R**, **knitr**, **Rstudio** and **Pandoc**.
- to the several postdoctoral researchers and students who have worked with me on related topics.
- Including Ester Pantaleo, Scott Powers, Mengyin Lu, Sen Tian, Wei Wang, Zhengrong Xing.

Thanks

- to the developers of **R**, **knitr**, **Rstudio** and **Pandoc**.
- to the several postdoctoral researchers and students who have worked with me on related topics.
- Including Ester Pantaleo, Scott Powers, Mengyin Lu, Sen Tian, Wei Wang, Zhengrong Xing.
- NHGRI for funding.

Thanks

- to the developers of **R**, **knitr**, **Rstudio** and **Pandoc**.
- to the several postdoctoral researchers and students who have worked with me on related topics.
- Including Ester Pantaleo, Scott Powers, Mengyin Lu, Sen Tian, Wei Wang, Zhengrong Xing.
- NHGRI for funding.
- `ashr` package: `http://www.github.com/stephens999/ash`

Pandoc Command used

```
pandoc -s -S -i --template=my.beamer -t beamer -V
theme:CambridgeUS -V colortheme:beaver slides.md -o
slides.pdf
```

(alternative to produce html slides; but figures would need reworking)

```
pandoc -s -S -i -t dzslides --mathjax slides.md -o
slides.html
```

Here is my session info:

```
print(sessionInfo(), locale=FALSE)
```

```
## R version 3.0.2 (2013-09-25)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## attached base packages:
## [1] splines   stats     graphics  grDevices utils     datas
## [8] base
```

Some odd things in the data

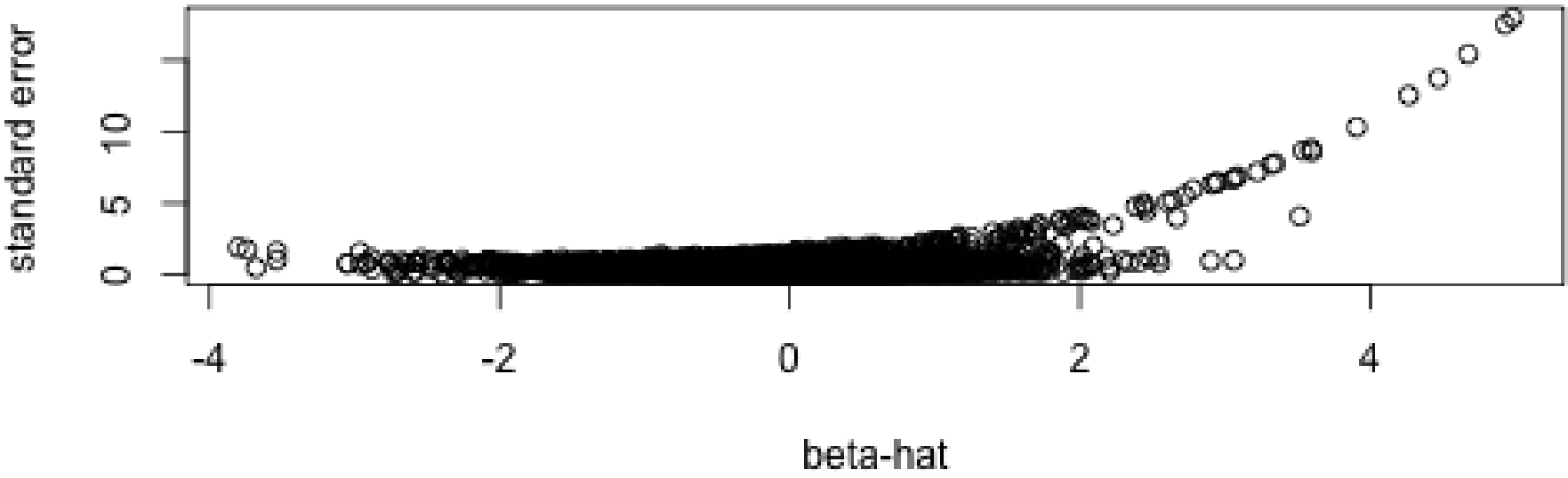


Figure : plot of chunk unnamed-chunk-35

```
## Error: object 'dd' not found
```

A technicality

- Suppose you estimate $\Pr(\beta_j < 0) = 0.98$, $\Pr(\beta_j > 0) = 0.01$, $\Pr(\beta_j = 0) = 0.01$.

A technicality

- Suppose you estimate $\Pr(\beta_j < 0) = 0.98$, $\Pr(\beta_j > 0) = 0.01$, $\Pr(\beta_j = 0) = 0.01$.
- Should you declare an fdr of 0.01 or 0.02?

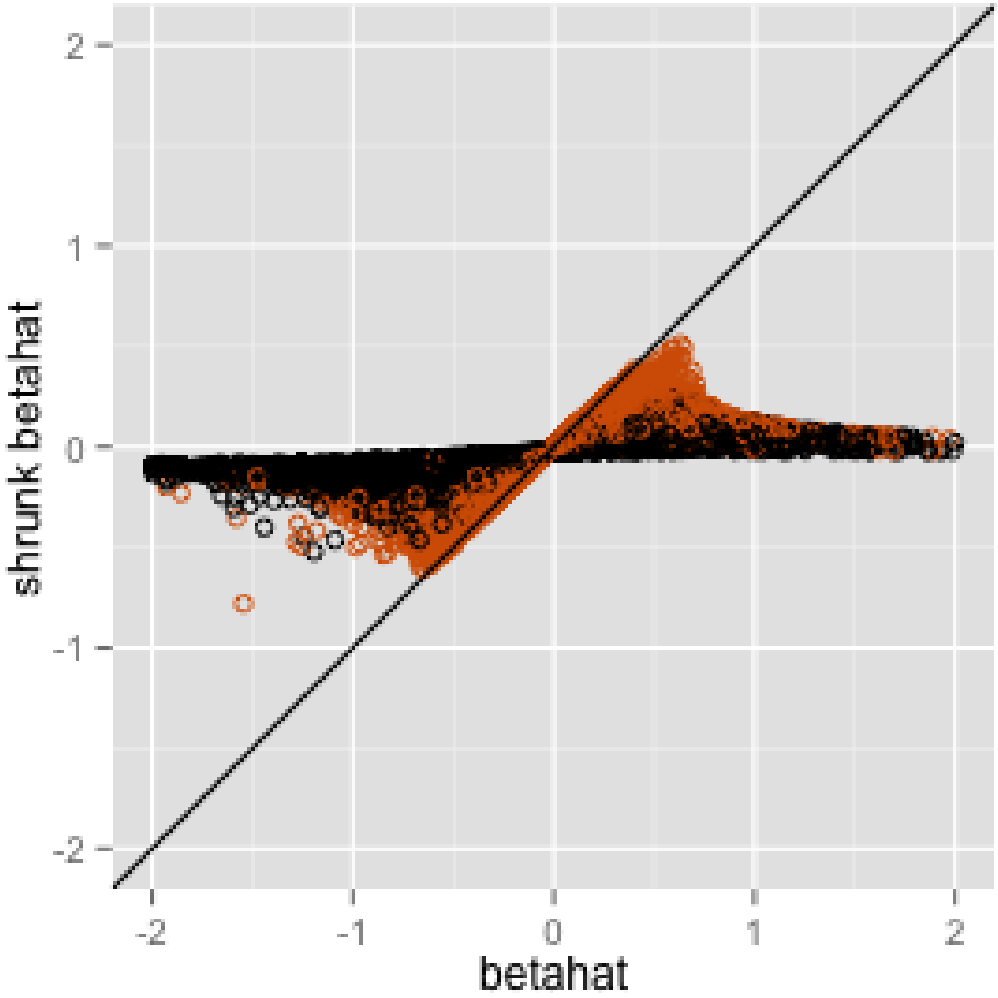
A technicality

- Suppose you estimate $\Pr(\beta_j < 0) = 0.98$, $\Pr(\beta_j > 0) = 0.01$, $\Pr(\beta_j = 0) = 0.01$.
- Should you declare an fdr of 0.01 or 0.02?
- Maybe fsr makes more sense anyway?

Adaptive Shrinkage of point estimates

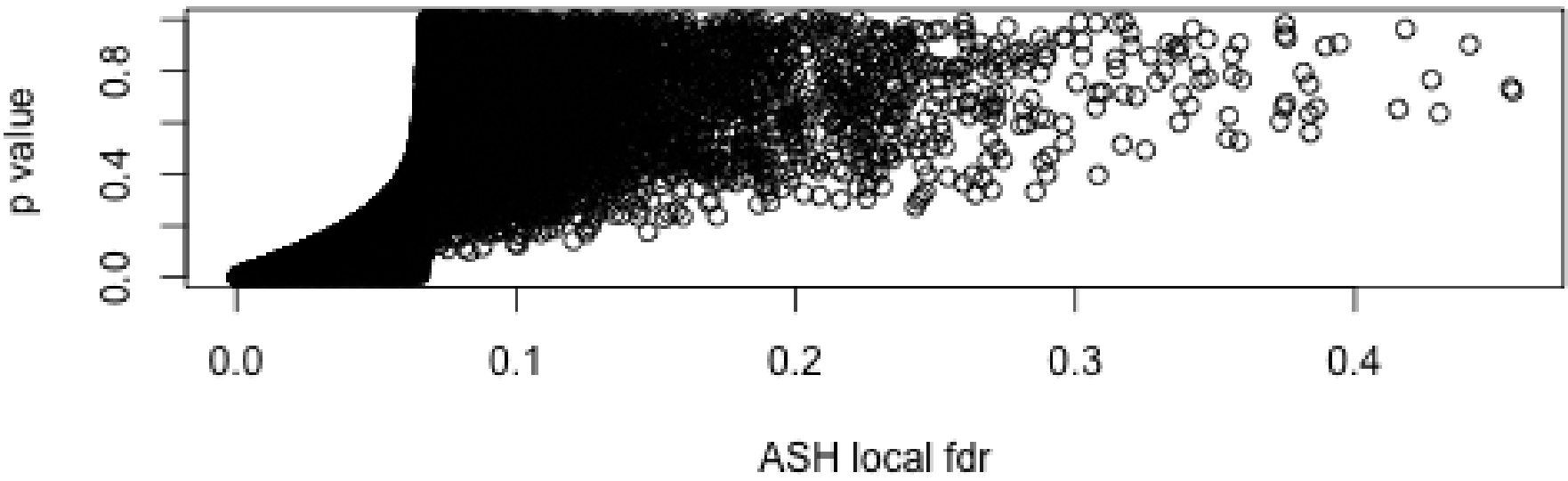
- Recall idea: amount of shrinkage depends on measurement precision, s_j.

Adaptive Shrinkage of point estimates

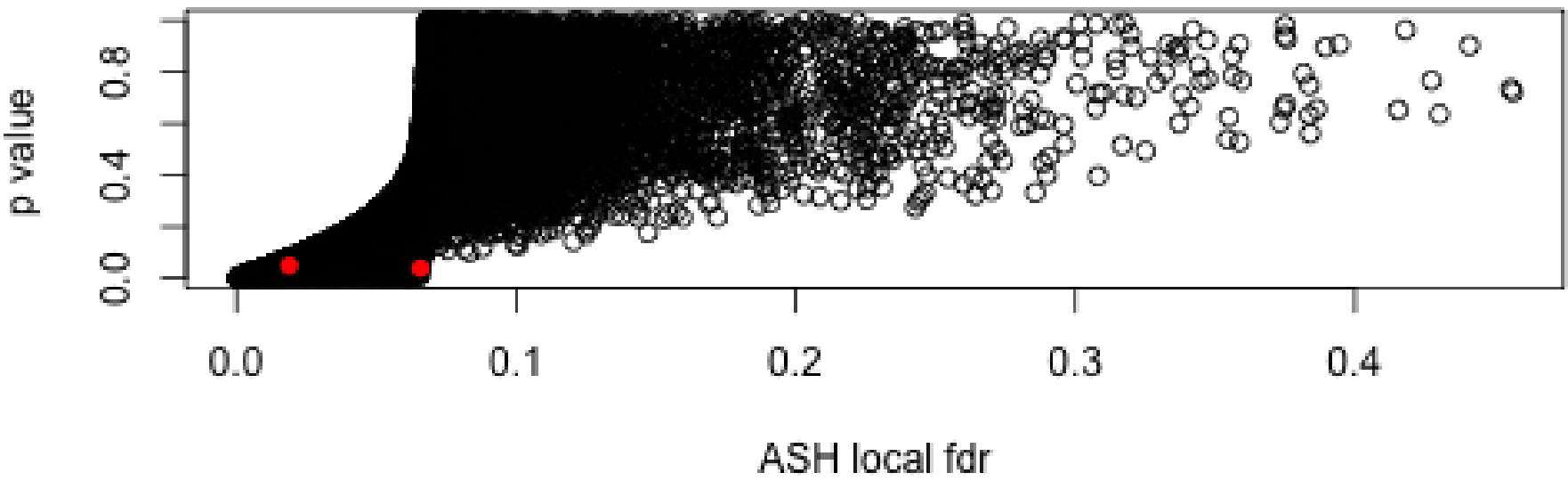


Shrinkage is adaptive to information

Need to fix counts.associate to use fdr method in ash



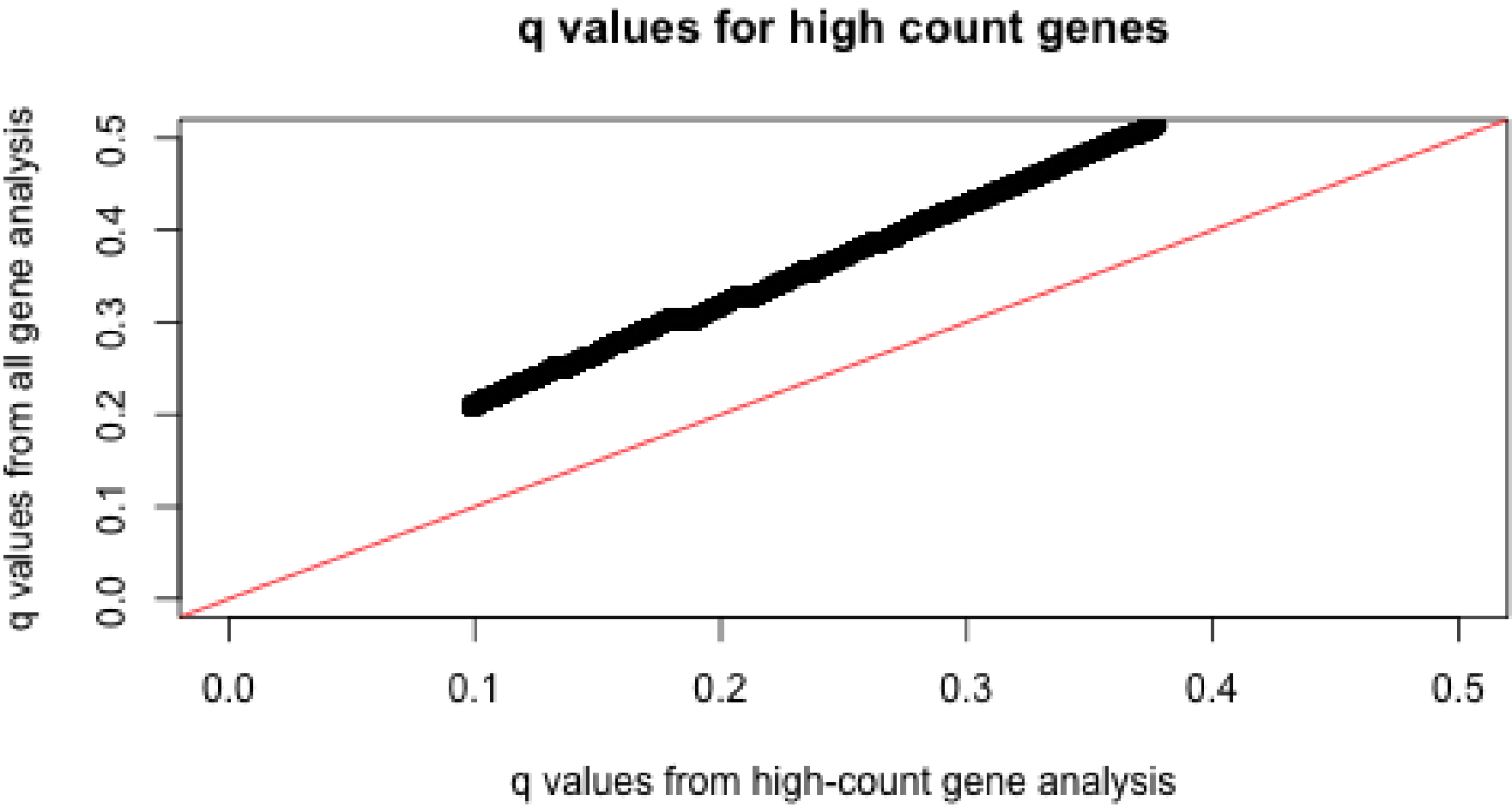
Shrinkage is adaptive to information



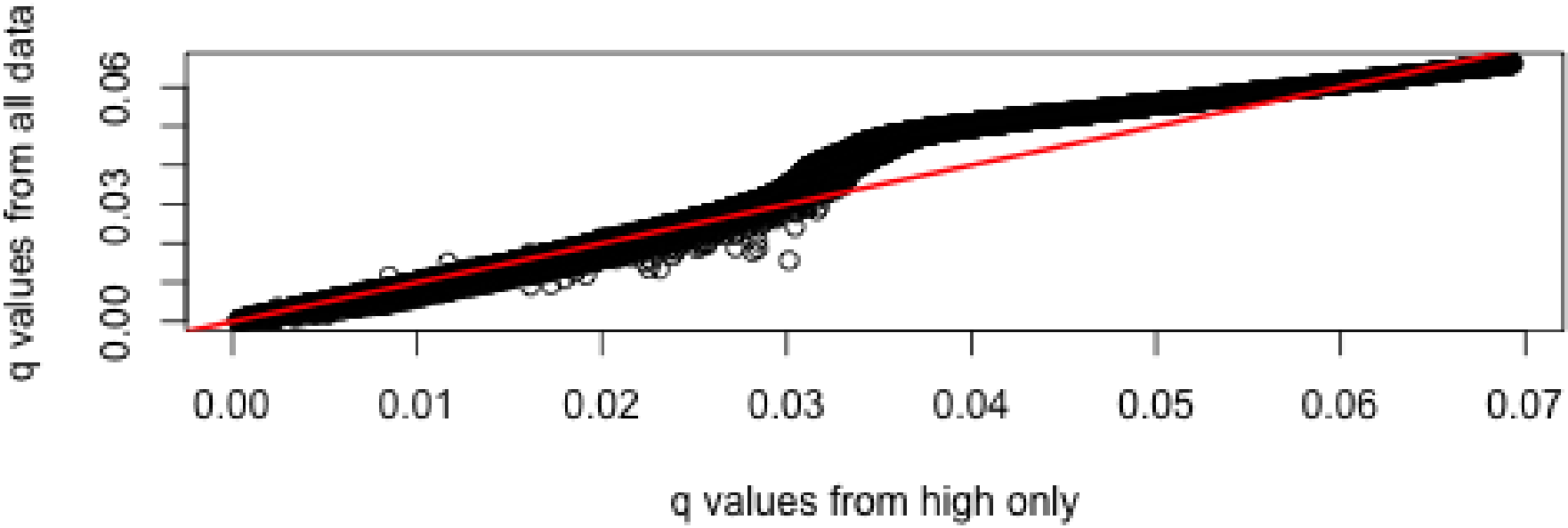
Shrinkage is adaptive to information

```
##          gene  lv1  lv2  rv1  rv2    pval zdat.ash$lfdr
## 19422 Mgat5b    7   10  320  452 0.03795        0.06575
## 20432  Sec63 1042 1034 5496 6649 0.04908        0.01895
```

Recall FDR problem 1: q values increased by low count genes



ASH q values more robust to inclusion of low count genes



Compare fitted $f(\beta)$, both estimating π_0 and fixing $\pi_0 = 0$.

```
## Error: object 'hh.ash.fdr' not found
```

```
## Error: object 'hh.ash.shrink' not found
```