False Discovery Rates, A New Deal

Matthew Stephens

2014/5/8

What you can expect from this talk

- No experimental data!

What you can expect from this talk

- No experimental data!
- No fancy heatmaps!

What you can expect from this talk

- No experimental data!
- No fancy heatmaps!
- 100% Money-back guarantee to half the False Discovery Rate (FDR) in your data!

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

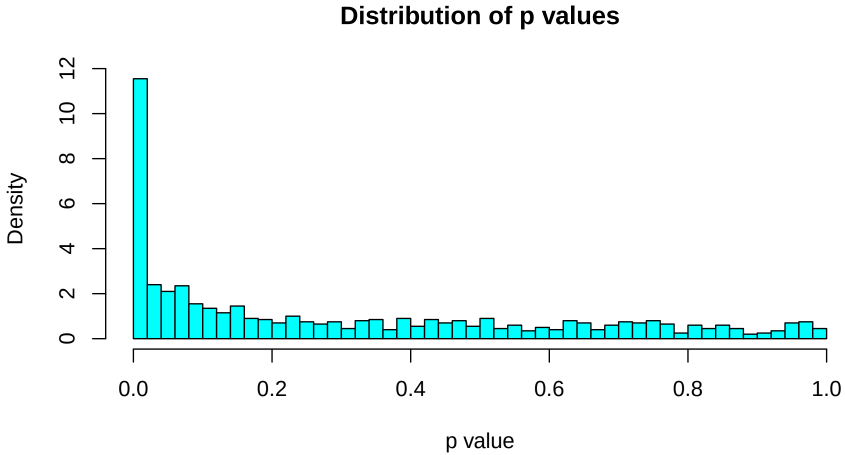
The Canonical Genomics Experiment

- 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.

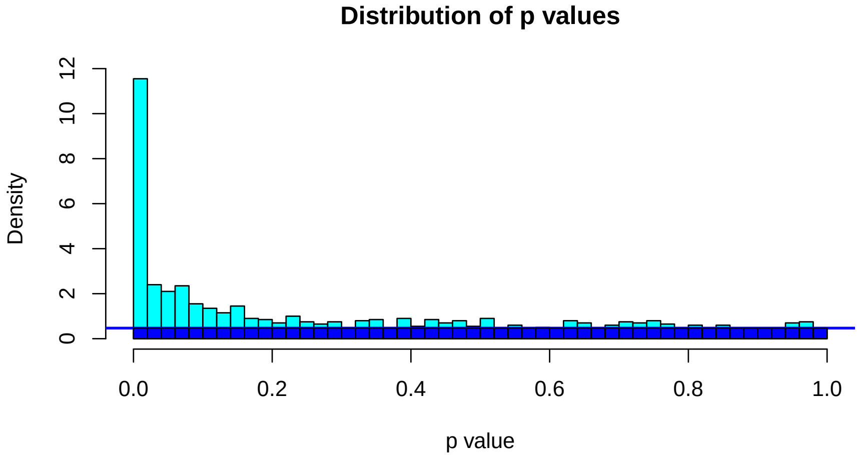
The Canonical Genomics Experiment

- 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.
- ...?

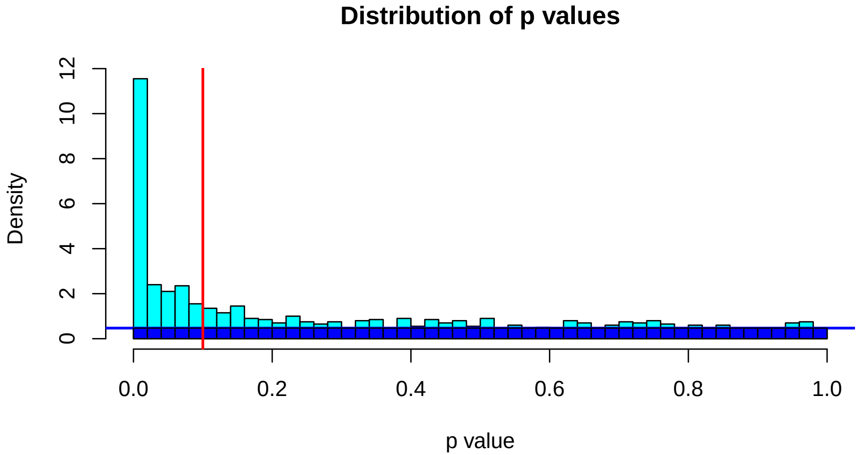
Example: FDR estimation



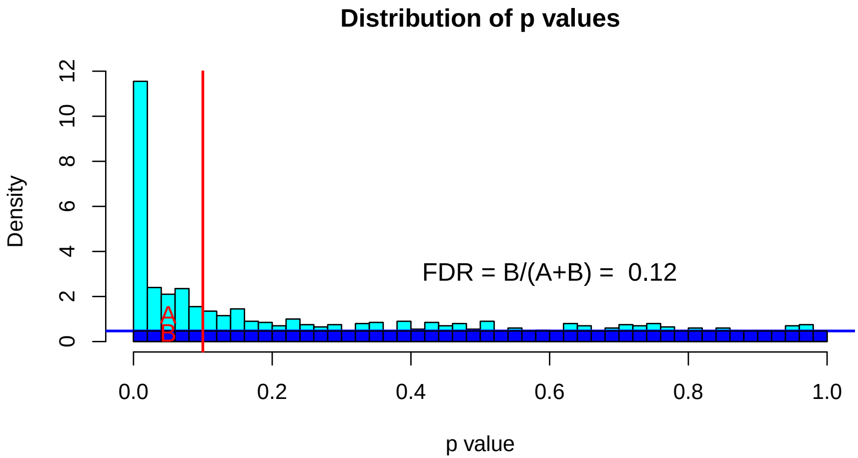
Example: FDR estimation



Example: FDR estimation



Example: FDR estimation



Problem 1: The Zero Assumption (ZA)

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

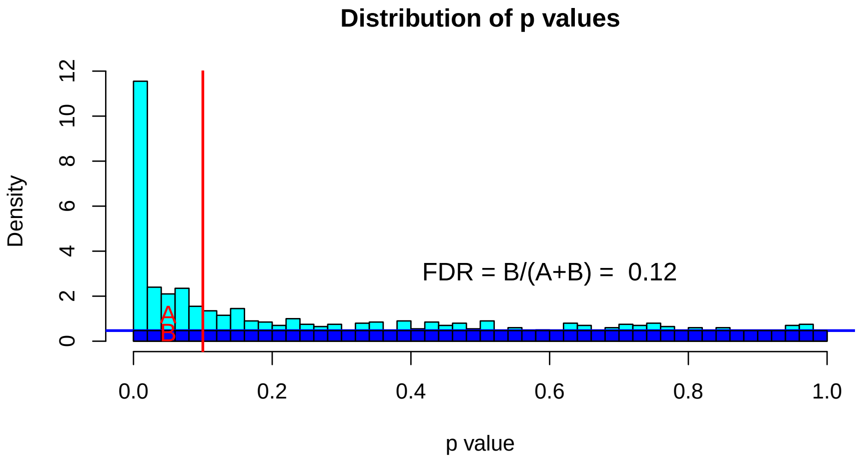
Problem 1: The Zero Assumption (ZA)

- 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".

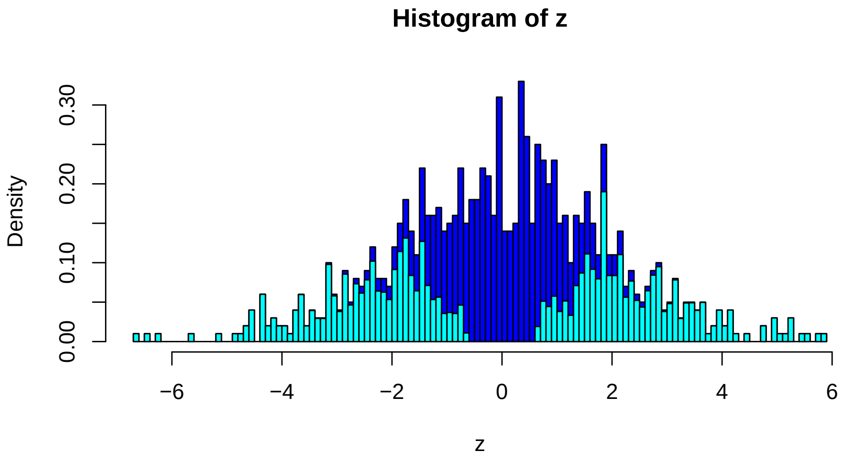
Problem 1: The Zero Assumption (ZA)

- 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 effect sizes of rare SNPs have larger standard error than those of common SNPs

FDR problem 2: different measurement precision

- In some cases the measurement precisions differ among units
- Eg effect sizes of rare SNPs have larger standard error than those of common SNPs
- 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 effect sizes of rare SNPs have larger standard error than those of common SNPs
- 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

FDR problem 2: different measurement precision

- Simulation: effects $\beta_j \sim N(0, 1)$

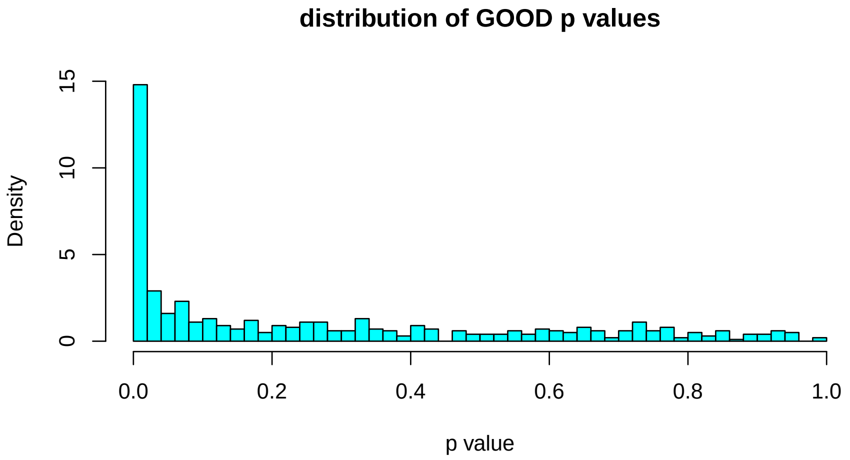
FDR problem 2: different measurement precision

- Simulation: effects $\beta_j \sim N(0, 1)$
- 500 "good" observations with low standard error ($s_j = 0.5$)

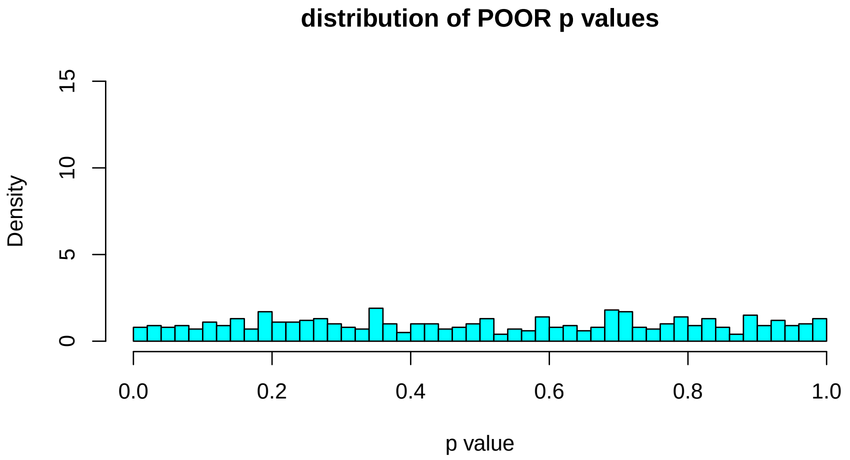
FDR problem 2: different measurement precision

- Simulation: effects $\beta_j \sim N(0, 1)$
- 500 "good" observations with low standard error ($s_j = 0.5$)
- 500 "poor" observations with very high standard error ($s_j = 10$)

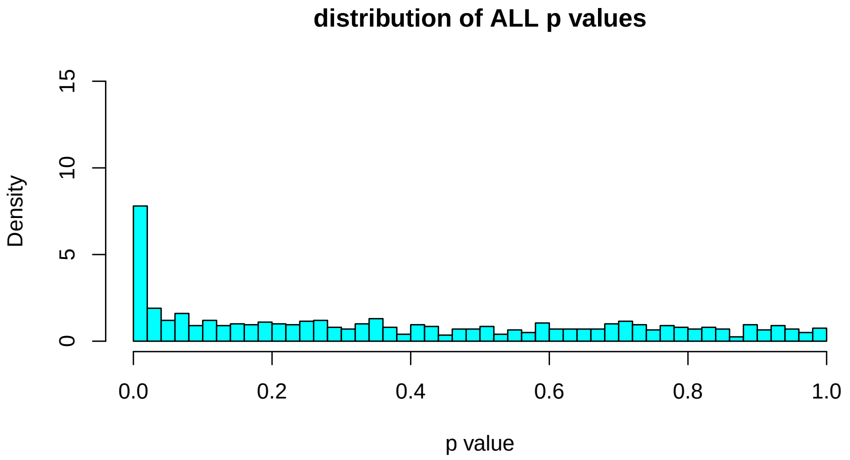
FDR problem 2: different measurement precision



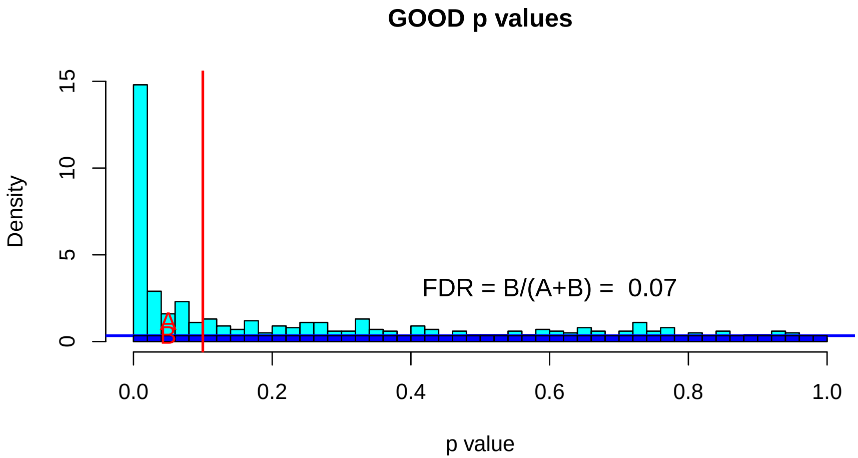
FDR problem 2: different measurement precision



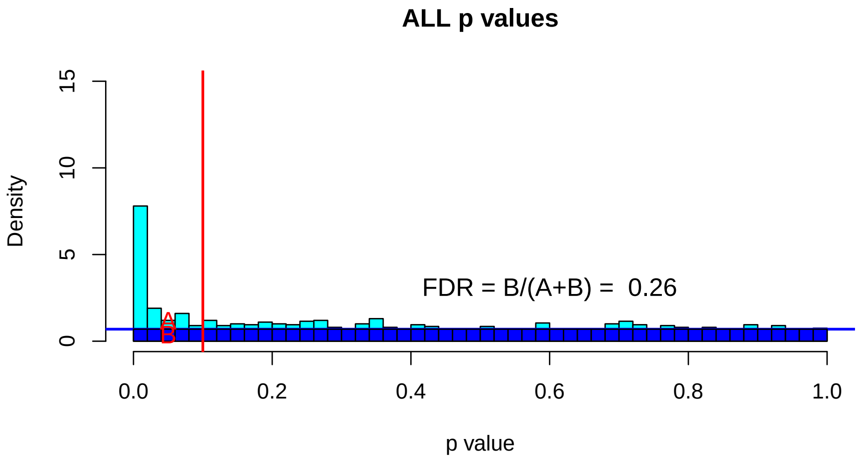
FDR problem 2: different measurement precision



FDR problem 2: different measurement precision



FDR problem 2: different measurement precision



Problems: Summary

Standard tools are highly 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 highly 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.
- 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 assuming

$$\hat{\beta}_j \sim N(\beta_j, s_j)$$

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

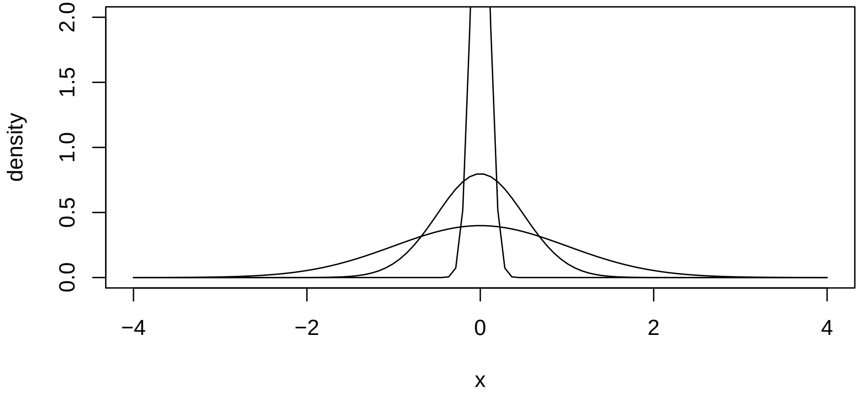


Illustration: g a mixture of 0-centered normals

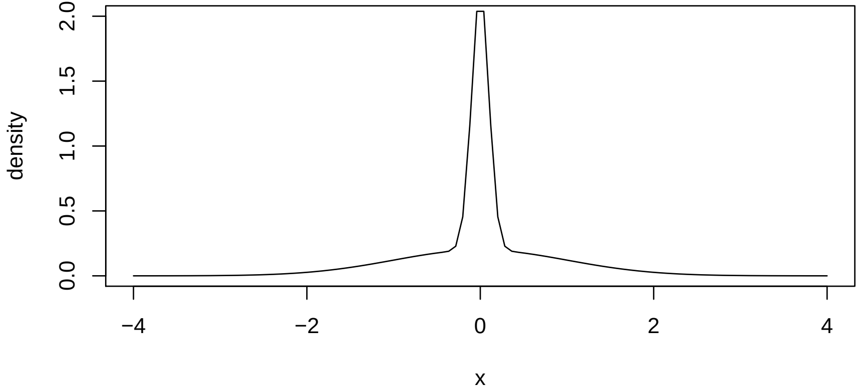


Illustration: g a mixture of 0-anchored uniforms

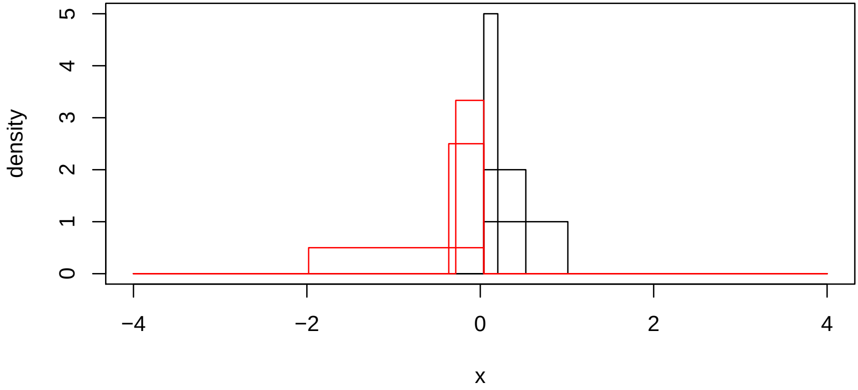
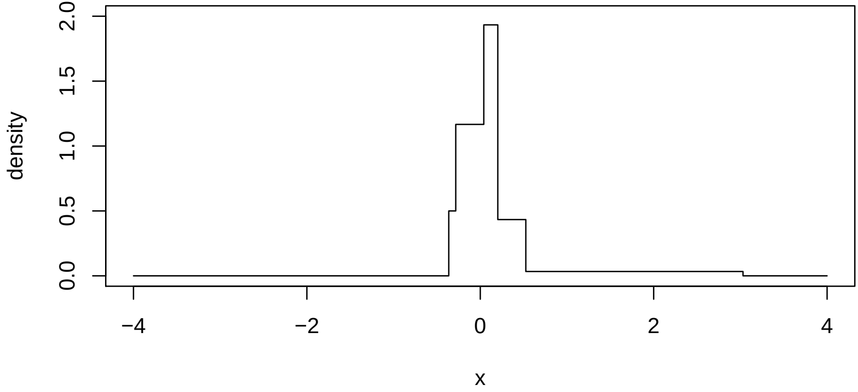
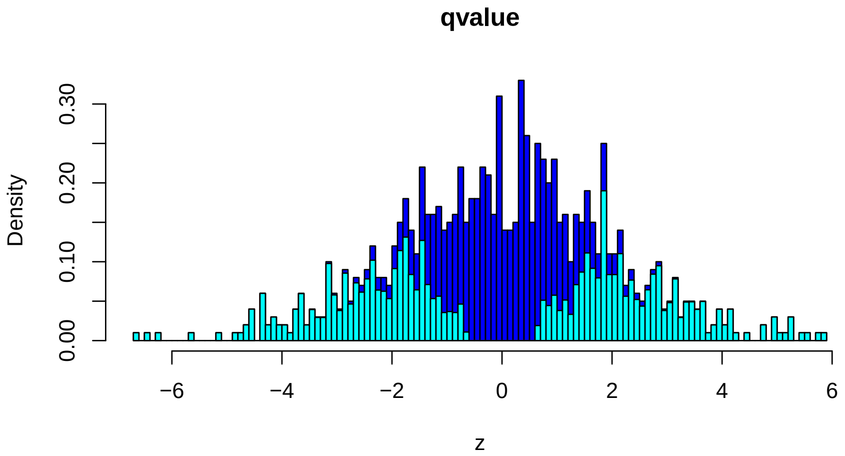


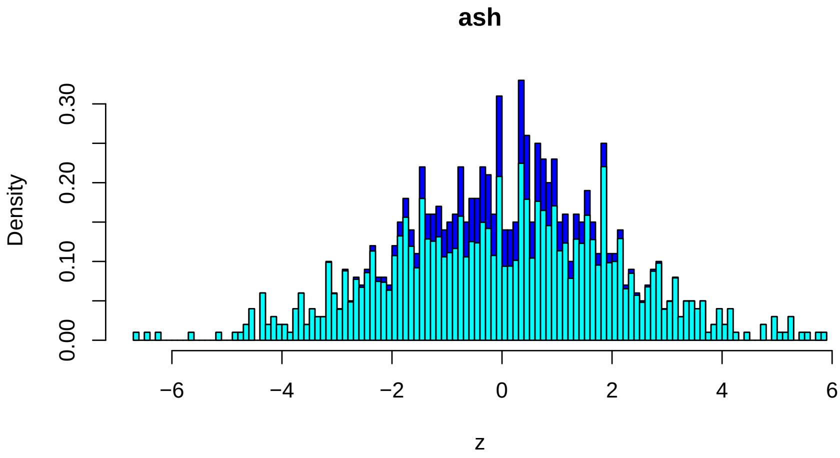
Illustration: g a mixture of 0-anchored uniforms



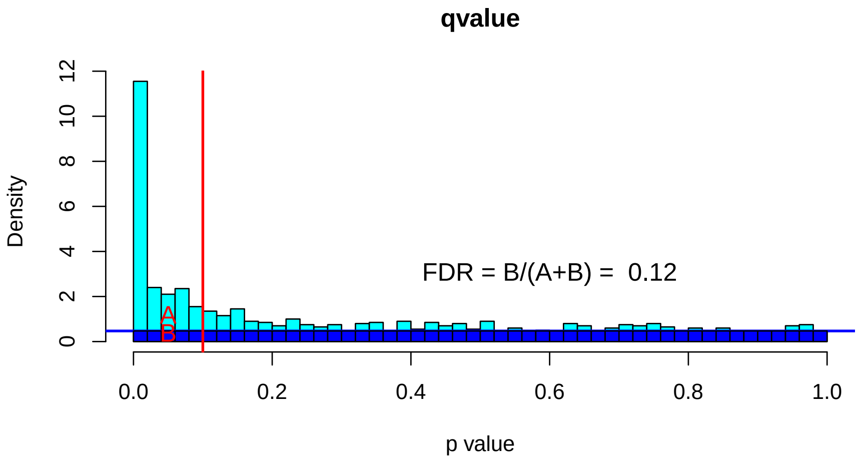
Recall Problem 1: distribution of alternative Z values multimodal



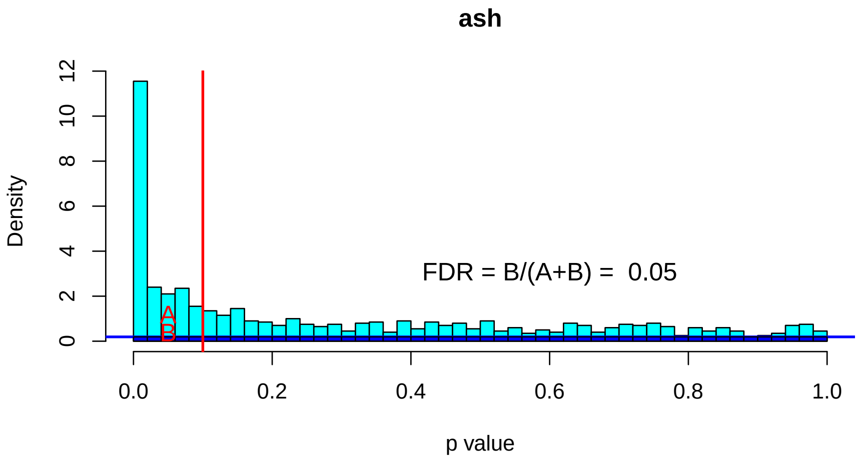
Problem Fixed: distribution of alternative Z values unimodal



Example: FDR estimation



Example: FDR estimation



Recall Problem 2: poor measurements dilute good measurments

Number of findings at (estimated) FDR=0.05:

```
## id          qvalue    ash
## ------   ---------   -----
## ALL            114    211
## GOOD           193    209
```

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: assume *none* of the β_j are zero ("one group approach"), and ask for which β_j are we confident about the sign (Gelman et al, 2012).

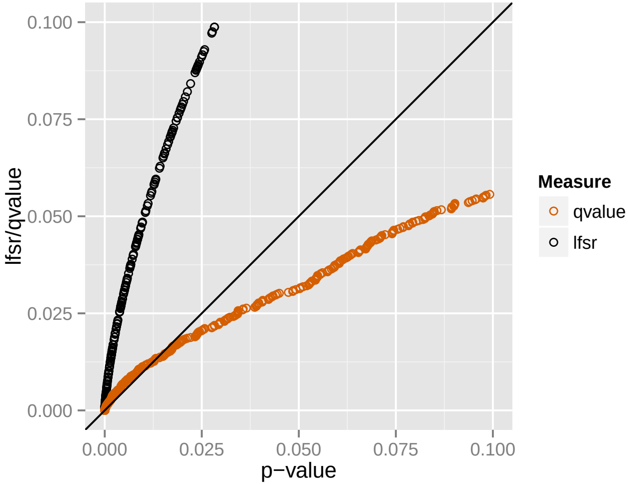
The False Sign Rate

- 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 $\Pr(\beta_j < 0) = 0.975$ and $\Pr(\beta_j > 0) = 0.025$. Then we report β_j as a “(negative) discovery“, and estimate its fsr as 0.025.

Even with many signals, large *p* values have high lfsr



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.
- Unimodal assumption for effects reduces conservatism

Summary

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

Summary

- ASH provides a generic approach to shrinkage estimation, as well as false discovery (sign) rates.
- Unimodal assumption for effects reduces conservatism
- But by using two numbers $(\hat{\beta}, s)$ instead of one (p values or z scores) precision of different measurement is better accounted for.
- In high-signal contexts, False Sign Rate is preferable to False Discovery Rate.

Reproducible Research?

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

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)

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]

This talk is reproducible!

- http://www.github.com/stephens999/ash

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).

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
```

Issue: you can't actually estimate π_0!

- Data cannot distinguish between $\beta_j = 0$ and β_j "very small"

Issue: you can't actually estimate π_0!

- Data cannot distinguish between $\beta_j = 0$ and β_j "very small"
- As a result π_0 cannot be estimated: the data can never rule out $\pi_0 = 0$.

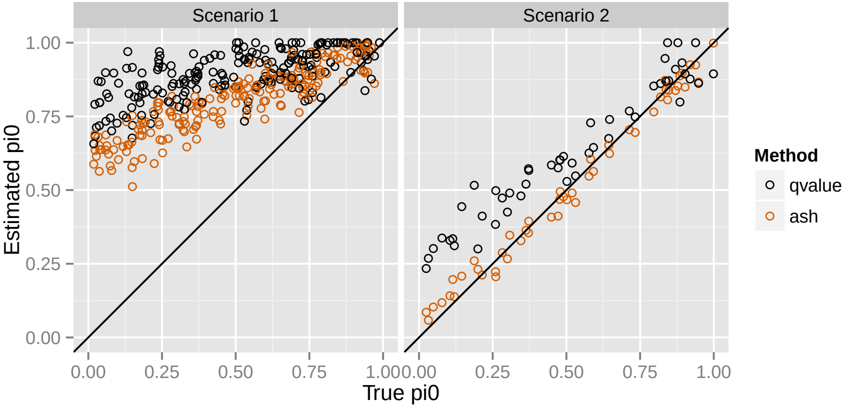
Issue: you can't actually estimate π_0!

- Data cannot distinguish between $\beta_j = 0$ and β_j "very small"
- As a result π_0 cannot be estimated: the data can never rule out $\pi_0 = 0$.
- But the unimodal constraint bounds how big π_0 can be.

Issue: you can't actually estimate π_0!

- Data cannot distinguish between $\beta_j = 0$ and β_j "very small"
- As a result π_0 cannot be estimated: the data can never rule out $\pi_0 = 0$.
- But the unimodal constraint bounds how big π_0 can be.
- Use penalized likelihood to make π_0 "as big as possible", subject to the unimodal constraint.

Approach remains conservative (if unimodal assumption holds)



Next steps?

- Incorporate t likelihood as well as normal.

Next steps?

- Incorporate t likelihood as well as normal.
- Incorporate shrinkage of variances and not just means.

Next steps?

- Incorporate t likelihood as well as normal.
- Incorporate shrinkage of variances and not just means.
- Extend to allow $g(\cdot; \pi)$ to depend on covariates X.

Next steps?

- Incorporate t likelihood as well as normal.
- Incorporate shrinkage of variances and not just means.
- Extend to allow $g(\cdot;\pi)$ to depend on covariates X.
- Extend to allow for correlations in the measured $\hat{\beta}_j$.

Is this an important problem?

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

Is this an important problem?

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

FDR and q values

Although precise definitions vary, roughly

- The FDR at a threshold P is

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

FDR and q values

Although precise definitions vary, 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 $q_j = \text{FDR}(p_j)$.

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.
- So for example we can easily compute fdrs for discoveries other than "non-zero" (eg compute $\Pr(\beta_j > 2|\hat{\beta}_j)$).

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.
- So for example we can easily compute fdrs for discoveries other than "non-zero" (eg compute $\Pr(\beta_j > 2|\hat{\beta}_j)$).
- And use it to obtain point estimates and credible intervals for each β_j, taking account of information from all the other β_j.

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.
- So for example we can easily compute fdrs for discoveries other than "non-zero" (eg compute $\Pr(\beta_j > 2|\hat{\beta}_j)$).
- And use it to obtain point estimates and credible intervals for each β_j, taking account of information from all the other β_j.
- Because $f(\beta)$ is unimodal, the point estimates will tend to be "shrunk" towards the overall mean (0).

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.
- So for example we can easily compute fdrs for discoveries other than "non-zero" (eg compute $\Pr(\beta_j > 2|\hat{\beta}_j)$).
- And use it to obtain point estimates and credible intervals for each β_j, taking account of information from all the other β_j.
- Because $f(\beta)$ is unimodal, the point estimates will tend to be "shrunk" towards the overall mean (0).
- Because $f(\beta)$ is estimated from the data, the amount of shrinkage is adaptive to the data. And because of the role of s_j, the amount of shrinkage adapts to the information on each gene.

Estimation and Shrinkage

- Besides allowing one to estimate fdr and fsr, this approach also provides a posterior distribution for each β_j.
- So for example we can easily compute fdrs for discoveries other than "non-zero" (eg compute $\Pr(\beta_j > 2|\hat{\beta}_j)$).
- And use it to obtain point estimates and credible intervals for each β_j, taking account of information from all the other β_j.
- Because $f(\beta)$ is unimodal, the point estimates will tend to be "shrunk" towards the overall mean (0).
- Because $f(\beta)$ is estimated from the data, the amount of shrinkage is adaptive to the data. And because of the role of s_j, the amount of shrinkage adapts to the information on each gene.
- So we call the approach "Adaptive Shrinkage" (ASH).