Local false discovery rates

Patrick Breheny

February 3

Introduction

- We concluded the previous lecture with a look at how false discovery rates can be viewed as either a frequentist methodology or an empirical Bayes estimate
- From a Bayesian standpoint, however, the false discovery rate is somewhat strange, in the sense that it involves conditioning on a rejection region $z_j \in \mathcal{Z}$
- A more natural thing to do, as least from a Bayesian perspective, is to condition on the actual value of z; in other words, to estimate

$$\text{fdr}(z_0) = \mathbb{P}(H_0 | Z = z_0);$$

the *local false discovery rate* for H_{0j} is therefore $\text{fdr}(z_j)$

FDR applies to the group, not a specific test

- One reason that the FDR is somewhat unsatisfying is that, by conditioning on $z_j \in \mathcal{Z}$, we calculate a probability/rate applying generally to all hypotheses in that region
- This, however, ignores the fact that some z-values are much more extreme than others, or to put it another way, that not all hypotheses are equally likely to be contributing the false discoveries
- For example, at an FDR of 1%, we can claim 734 discoveries; among them, $|z_j|$ ranges from 3.3 to 9.5
- FDR tells us to expect ≈ 7 false discoveries; those false discoveries are presumably much more likely to be coming from the tests with $z \approx 3$ than $z \approx 9$

The tale of the dishonest statistician

- To see why this might be a problem, let's take this line of reasoning to an extreme end: suppose we test $h = 1,000$ hypotheses, and the smallest p-value we get is $p = 0.001$
- If we want to control the FDR at 10%, this is well above the BH cutoff to reject the first gene (here, 0.0001)
- Suppose that the statistician, disappointed by the fact that we cannot reject any hypotheses, decides to add 10 additional tests for which they know in advance that the null hypothesis is false

The tale of the dishonest statistician (cont'd)

- As expected, the results for those 10 tests are highly significant
- Now, they go back to control the FDR for these 1,010 tests; the p-value cutoff for the 11th test is now $p = 0.0011$, so now we *can* reject the hypothesis that we couldn't on the previous slide
- This approach allows the statistician to publish a list of 11 "discoveries", of which 10 were known in advance, but hey, there's one interesting new discovery that we have "significant" statistical evidence for

Exchangeability

- This obviously flawed approach illustrates that false discovery rates come with a key assumption of exchangeability: if we're going to make significance statements about a *group* of tests, those tests should be as homogeneous as possible
- It isn't incorrect to say that the false discovery rate for those 11 discoveries is under 10%, but it's certainly misleading – it's pretty obvious which result is likely to be the false discovery
- This example is (hopefully) unrealistic, but the question of which hypotheses can be combined to form a relevant group arises quite often: for example, should we be combining the left and right tails?

Bayes rule again

- Following the same reasoning as at the end of the previous lecture, we can use Bayes rule to obtain an expression for the local false discovery rate:

$$\text{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)},$$

  where $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is the marginal density of z-values and $f_0(z)$ is the null density

- Note: Many authors (including me) use Fdr to refer to the false discovery rate and fdr to refer to the local FDR, reflecting the F/f convention for denoting distribution and density functions, respectively

Remarks

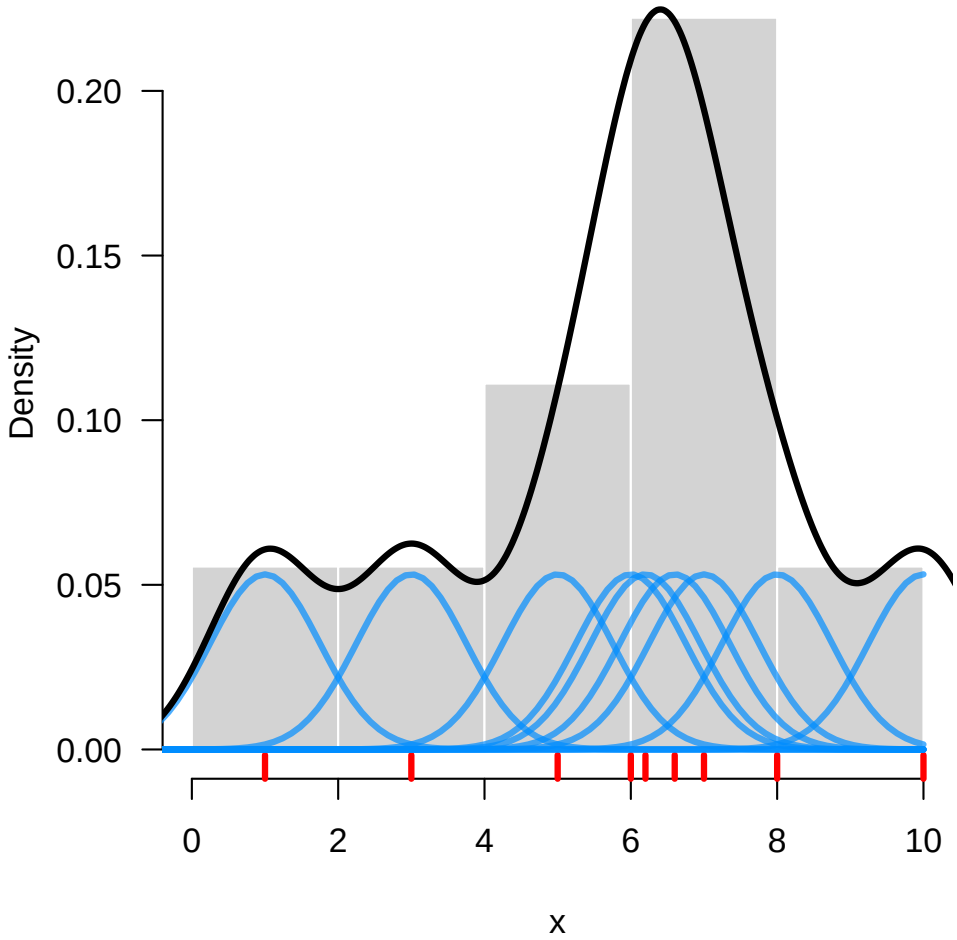
- Local FDRs offer a number of advantages over tail-area FDRs; for example, from a Bayesian perspective, conditioning on z is correct, not $z \in \mathcal{Z}$; in fact, the quantity $f_1(z)/f_0(z)$ is known as the *Bayes factor* for quantifying the level of empirical support for hypothesis 1 over hypothesis 0
- However, local FDR has faced two main challenges in terms of gaining widespread acceptance relative to tail-area FDR:
  - No interpretation as a frequentist error rate control procedure is available
  - Estimating a density (f) is far less straightforward than estimating a distribution (F), meaning that there are many variants of local FDR, unlike tail area FDR
- This may be changing (I've started to see local FDRs in prominent journals more often), but time will tell

Three ingredients

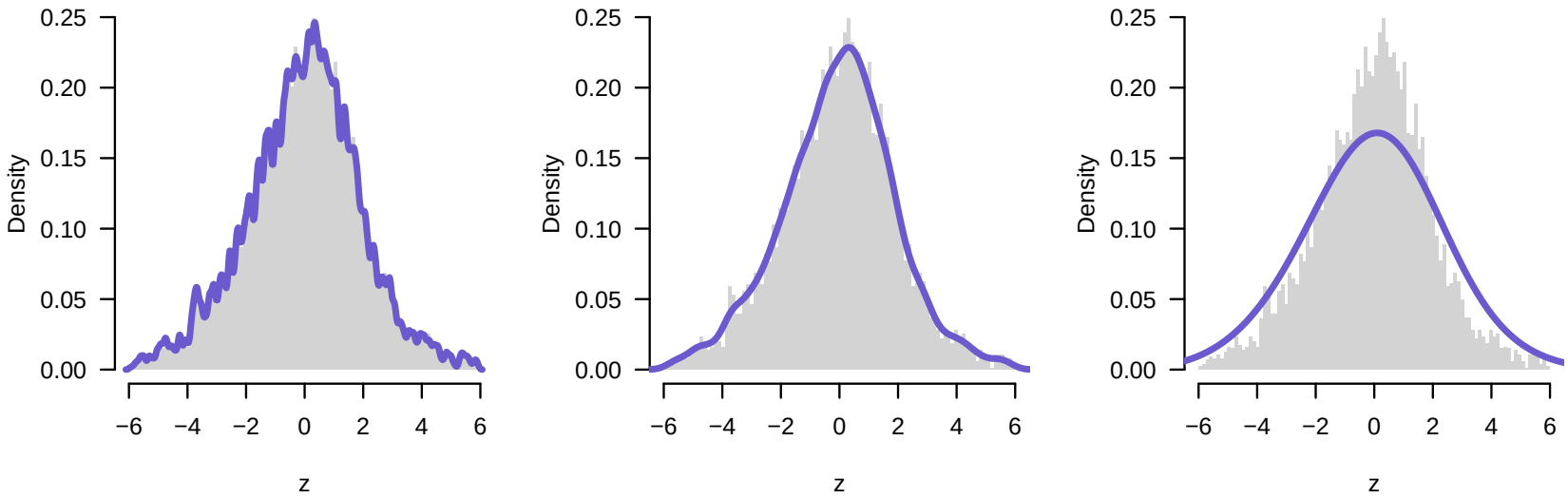
- The local false discovery rate has three components:
  - π_0
  - f
  - f_0
- Each of these can potentially be varied, producing different estimates of fdr
- Today, we will look at some relatively simple approaches for estimating these quantities, then look at one complex approach, although many alternatives exist

Density estimation using Gaussian kernels

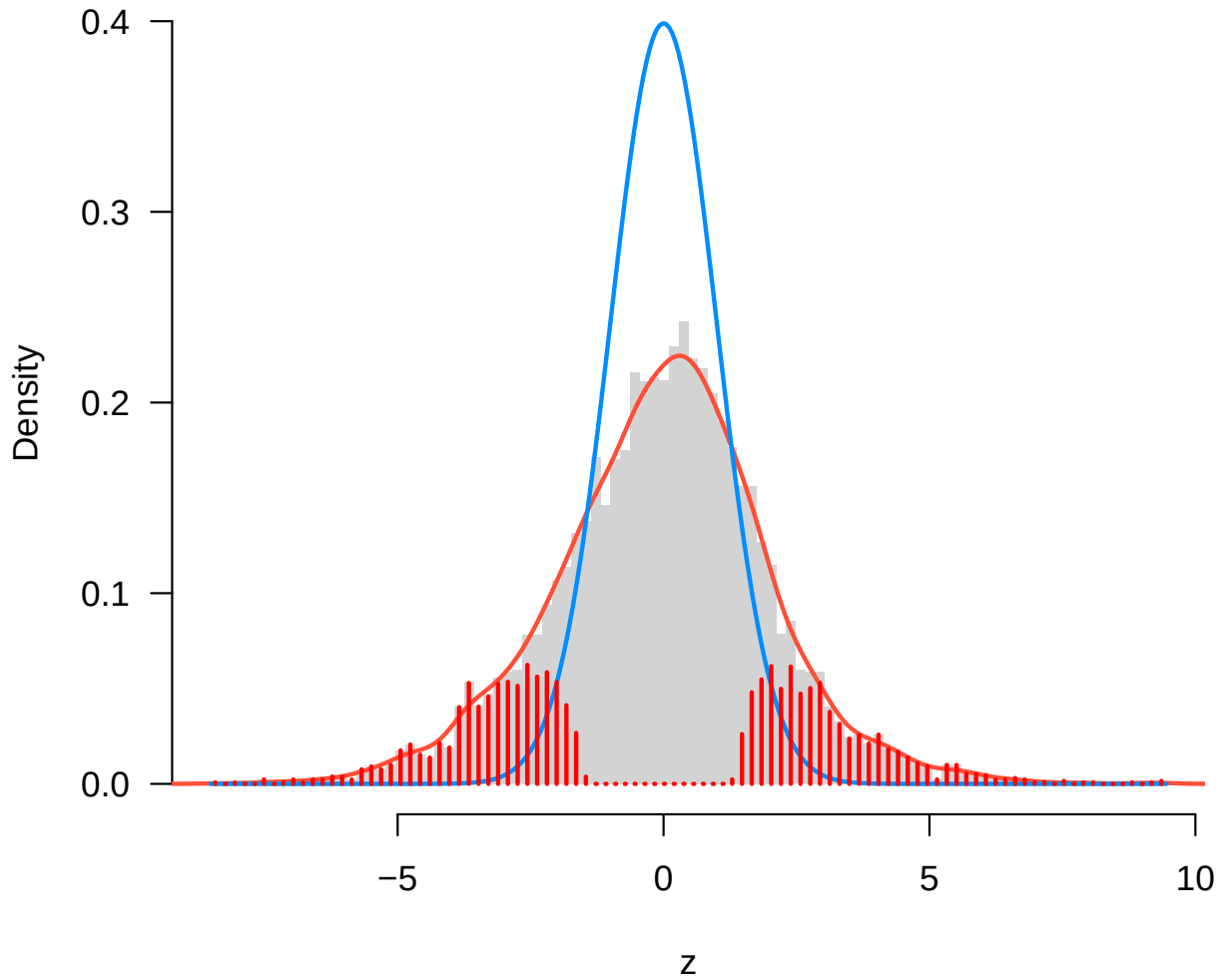
One common approach is *kernel density estimation*:



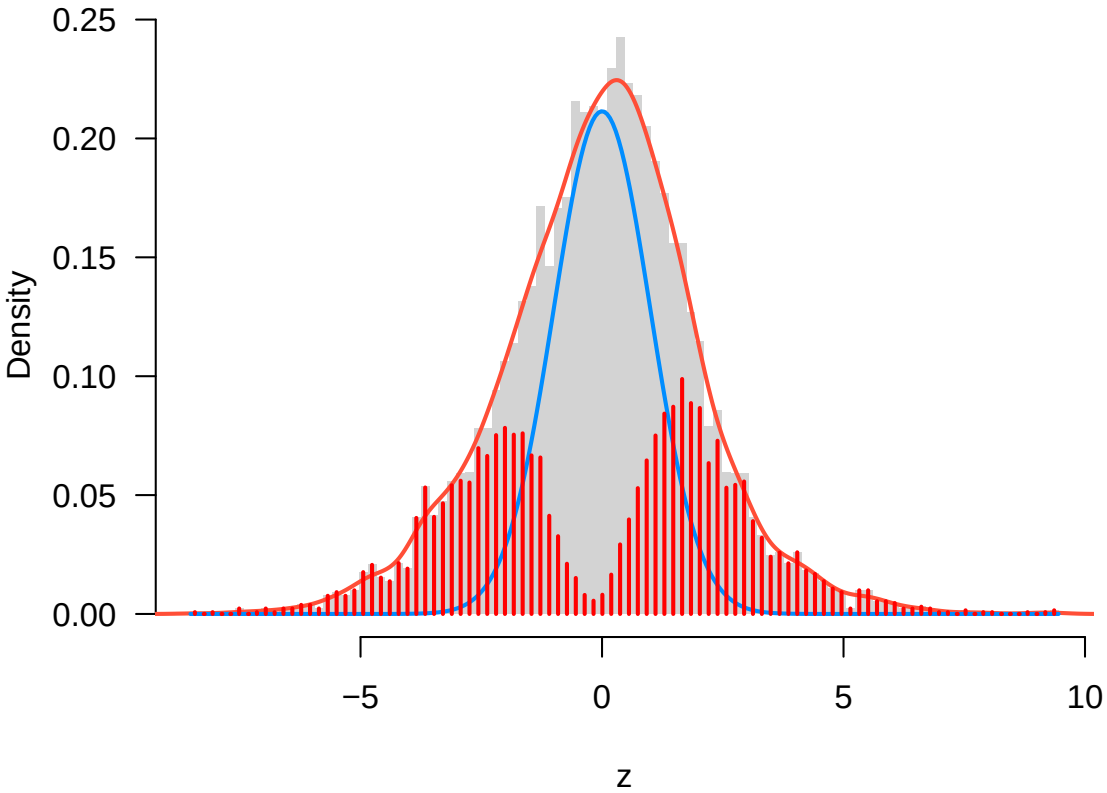
Choice of bandwidth



Local fdr for leukemia data: Illustration

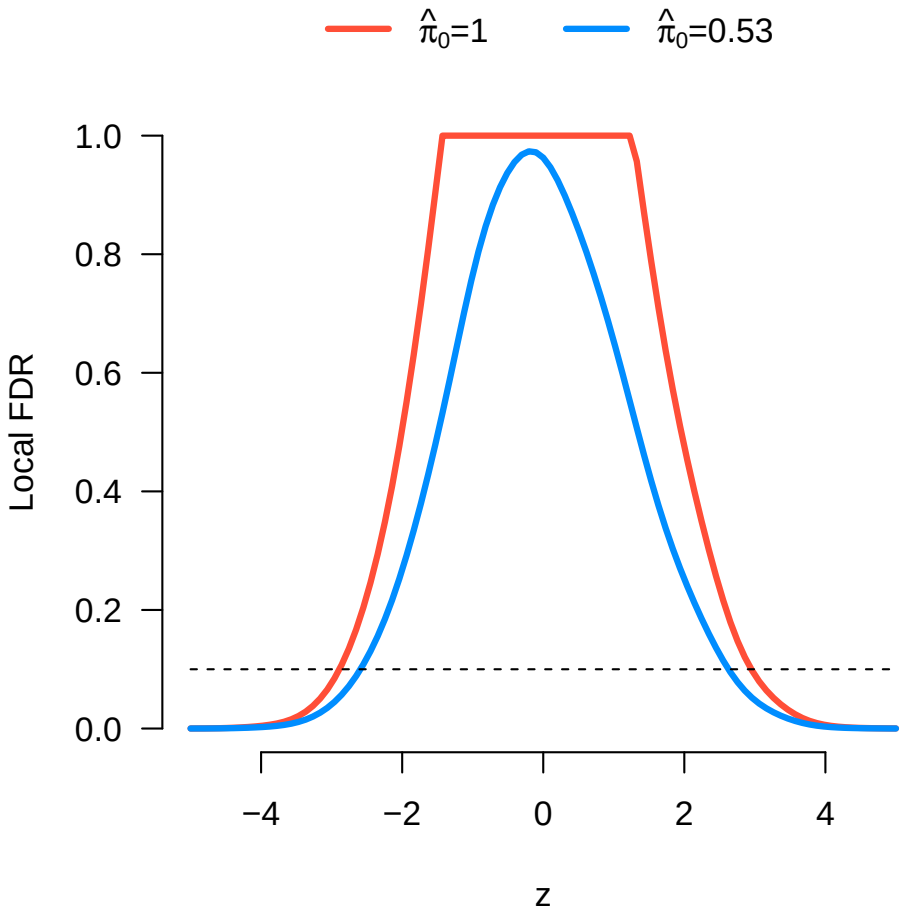


Local fdr for leukemia data: $\hat{\pi}_0 = 0.53$



Using $\hat{\pi}_0 = 0.53$, our estimate from the previous lecture, we seem to obtain more realistic estimations of the null and alternative distributions

z vs local FDR



For a 10% local FDR cutoff:

- Using $\hat{\pi}_0 = 1$, critical value of $z = 2.95$; 986 significant results
- Using $\hat{\pi}_0 = 0.53$, critical value of $z = 2.63$; 1,266 significant results

Estimating a null distribution?

- Lastly, one could consider estimating f_0 as well
- This is admittedly a somewhat weird idea – using the data to estimate the null – however, it has been proposed in the literature and studied by many authors
- The basic idea is to assume that $Z \sim \mathrm{N}(\delta_0, \sigma_0^2)$ and use the "central" part of the data to estimate δ_0 and σ_0
- It is certainly possible, for a variety of reasons, for the theoretical null $N(0, 1)$ not to hold; whether we can fix these problems by estimating a null is not always clear
- It's an interesting idea, but I'm not going to say much more about it in this lecture

Cutoff comparison

- It is worth spending a few slides on a deeper examination of Fdr versus fdr in terms of results and interpretation
- Using $\pi_0 = 1$, and a 10% cutoff,
  - Fdr: Critical $z = 2.27$; 1,635 significant findings
  - fdr: Critical $z = 2.95$; 986 significant findings
- For any given percentage cutoff, local FDR is considerably more conservative than FDR about declaring a result significant
- To put it another way, a 10% Fdr does not mean the same thing as a 10% fdr

Conditional expectation relationship

- Further insight into the relationship between FDR and local FDR is given by this result:

$$\mathbb{E}\{\text{fdr}(z)|z \in \mathcal{Z}\} = \text{Fdr}(\mathcal{Z})$$

- Roughly, then, we should expect the average local FDR among the significant features to equal the FDR:
  - Left tail: Average fdr for features with $\text{Fdr} < 0.1$ is 0.102
  - Right tail: Average fdr for features with $\text{Fdr} < 0.1$ is 0.097
- This relationship does not exactly work out for two-sided tests unless we specifically estimate a combined tail density $f(|z|)$

R code

- There are a number of R packages for calculating local FDRs, all of which take different approaches to the estimation of π_0, f, and potentially f_0
- I will discuss one package in some detail today called `ashr`: "False discovery rates: a new deal", by Stephens (2017), *Biostatistics*
- Other popular packages include `locfdr` and `fdrtool`

Gaussian mixture model

- Let $\{\theta_j\}$ denote the effects of interest, with corresponding standard errors $\{s_j\}$
- Consider the following empirical Bayes mixture model:

$$\hat{\theta}_j | \theta_j, s_j \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(\theta_j, s_j^2)$$

$$\theta_j \overset{\perp\!\!\!\perp}{\sim} \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k \mathrm{N}(0, \sigma_k^2),$$

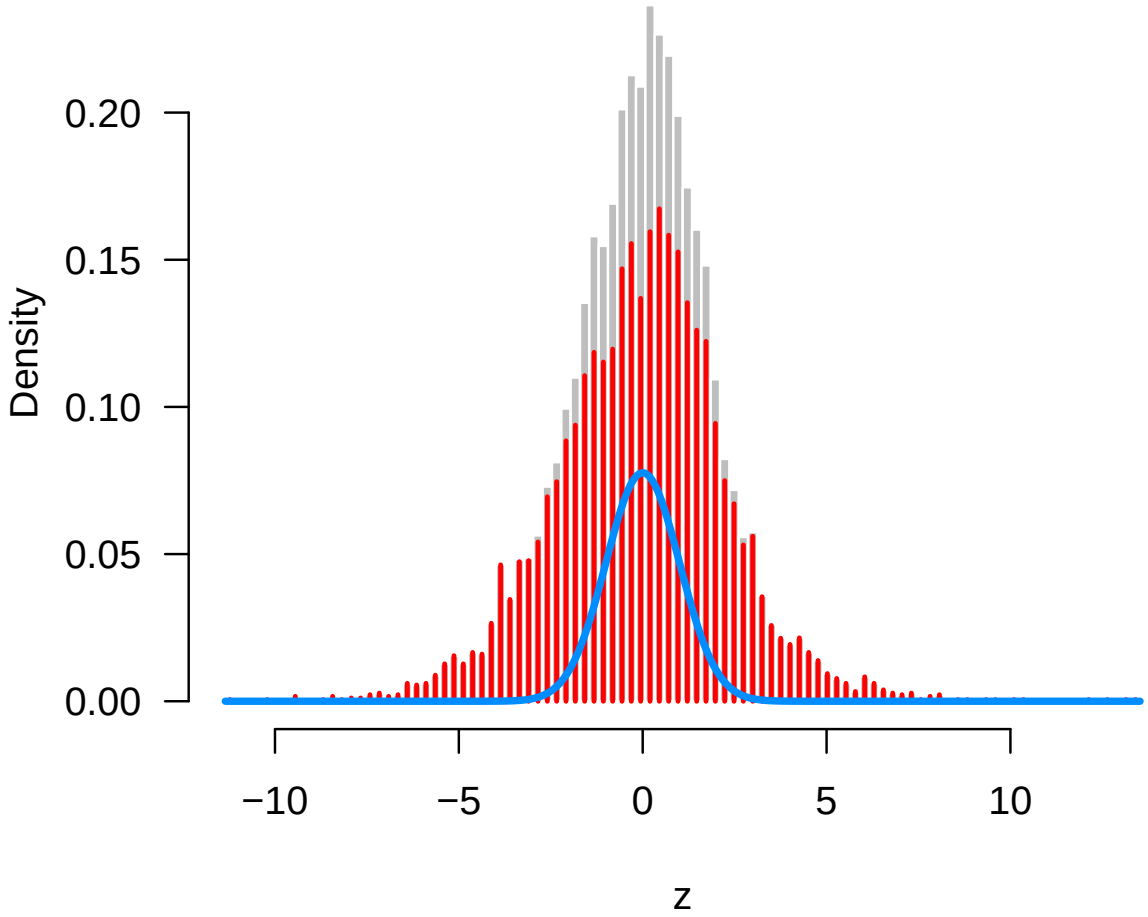
  where $\delta(\cdot)$ denotes a point mass at zero
- By Bayes' rule,

$$f(\theta_j | \hat{\theta}_j) \propto f(\theta_j) f(\hat{\theta}_j | \theta_j)$$

Remarks

- What makes this an "empirical" Bayes model is that we will be estimating $\{\pi_k, \sigma_k\}$ from the data, rather than specifying priors on them and fitting a fully Bayesian model
- This could be done a variety of ways, for example using the EM algorithm, although we will skip the details
- One key difference from the earlier approach is that here, f_1 is unimodal by construction; recall that it was bimodal with peaks around ± 2 earlier

ashr: Leukemia data



π_0

- According to the mixture model, even the z values near 0 are likely to be non-null; $\hat{\pi}_0 = 0.19$
- This varies somewhat depending on what mixture you assume, but is always much lower than the Storey approach for this data set:
  - Uniform: $\hat{\pi}_0 = 0.25$
  - Half-uniform: $\hat{\pi}_0 = 0.22$

False sign rate

- One advantage of this approach is that we obtain a posterior, and can carry out some interesting calculations unavailable to us in the frequentist framework
- Of particular interest is the idea of a *local false sign rate*:

$$\text{fsr} = \mathbb{P}(\theta \leq 0 | \hat{\theta} > 0);$$

  the definition for $\hat{\theta} < 0$ is similar
- John Tukey: "The effects of A and B are always different – in some decimal place – for any A and B. Thus asking 'Are the effects different?' is foolish . . . the more meaningful question [is]: 'is the evidence strong enough to support a belief that the observed difference has the correct sign?'"

Example: False sign rate

- This is straightforward to calculate with the mixture model, since the posterior has a simple, closed form
- To illustrate, let's consider the gene TERF1
  - $z = 0.2$
  - $p = 0.84$
  - $\text{Fdr} = 0.49 (\hat{\pi}_0 = 0.53)$
  - $\text{fdr} = 0.92 (\text{kernel}, \hat{\pi}_0 = 0.53)$
  - $\text{fdr} = 0.25 (\texttt{ashr})$
  - $\text{fsr} = 0.59 (\texttt{ashr})$

Prioritizing discoveries

- When the number of discoveries is large, one typically wishes to prioritize the most promising or significant findings
- Prioritizing based on p-value/Fdr/fdr/fsr is sometimes unsatisfactory, as a feature can be highly significant without a large effect size if the variance is small
- However, prioritizing on the basis of mean difference/fold change is often worse, as it gives too much emphasis to noisy features with inconsistent effects

Posterior means: Leukemia data

- Again, with a posterior distribution, there is another attractive option available to us: the posterior mean
- The posterior mean reflects the effect size, but is shrunken towards zero by the prior; how much shrinkage depends on the feature's noise level (s_j):

Gene	$\hat{\theta}$	s	fsr	fdr	PM
MCL1	1.26	0.28	0.0007	0.0005	1.03
PTX3	1.11	0.17	0.0000	0.0000	1.03
CSF1R	1.32	0.32	0.0034	0.0022	1.03
FAH	1.08	0.13	0.0000	0.0000	1.02
M63438_s_at	2.22	0.68	0.1058	0.0537	1.02
PLCB2	1.11	0.17	0.0000	0.0000	1.02

PM: Posterior mean

ashr: Usage

- Basic usage of the ashr package:

```
fit <- ash(theta, se)
```

  the mixcompdist option changes the type of mixture distribution (normal/uniform/etc)
- Main results are included in fit$result, but some other functions of interest:
  - get_pm(fit): Posterior means
  - get_lfdr(fit): Local false discovery rate
  - get_lfsr(fit): Local false sign rate
  - get_pi0(fit): $\hat{\pi}_0$