Local false discovery rates

Patrick Breheny

February 1

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

$$\mathbb{P}(H_{0j}|z_j = z)$$

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 504 discoveries; among them, $|z_j|$ ranges from 3.2 to 8.2
- FDR tells us to expect ≈ 5 false discoveries; those false discoveries are presumably much more likely to be coming from the tests with $z \approx 3$ than $z \approx 8$

The tale of the dishonest statistician

- Taking this line of reasoning to a more 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 he cannot reject any hypotheses, decides to add 10 additional tests for which he knows 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, he goes 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 he *can* reject the hypothesis that he couldn't on the previous slide
- This approach allows him 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

- The obviously dishonest approach laid out on the previous slide illustrates that false discovery rates come with a tacit 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 wrong to say that the false discovery rate for those 11 discoveries is under 10%, but it's certainly misleading, as it's pretty obvious where the false discovery would be coming from
- This example is extreme of course, 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

- Let us refer to $\mathbb{P}(H_{0j}|z_j = z)$ as the *local false discovery rate*
- As in our previous derivation, we can use Bayes rule to obtain an expression for the probability we are interested in:

$$\mathbb{P}(H_{0j}|z_j = z) = \frac{\pi_0 f_0(z_j)}{f(z_j)},$$

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

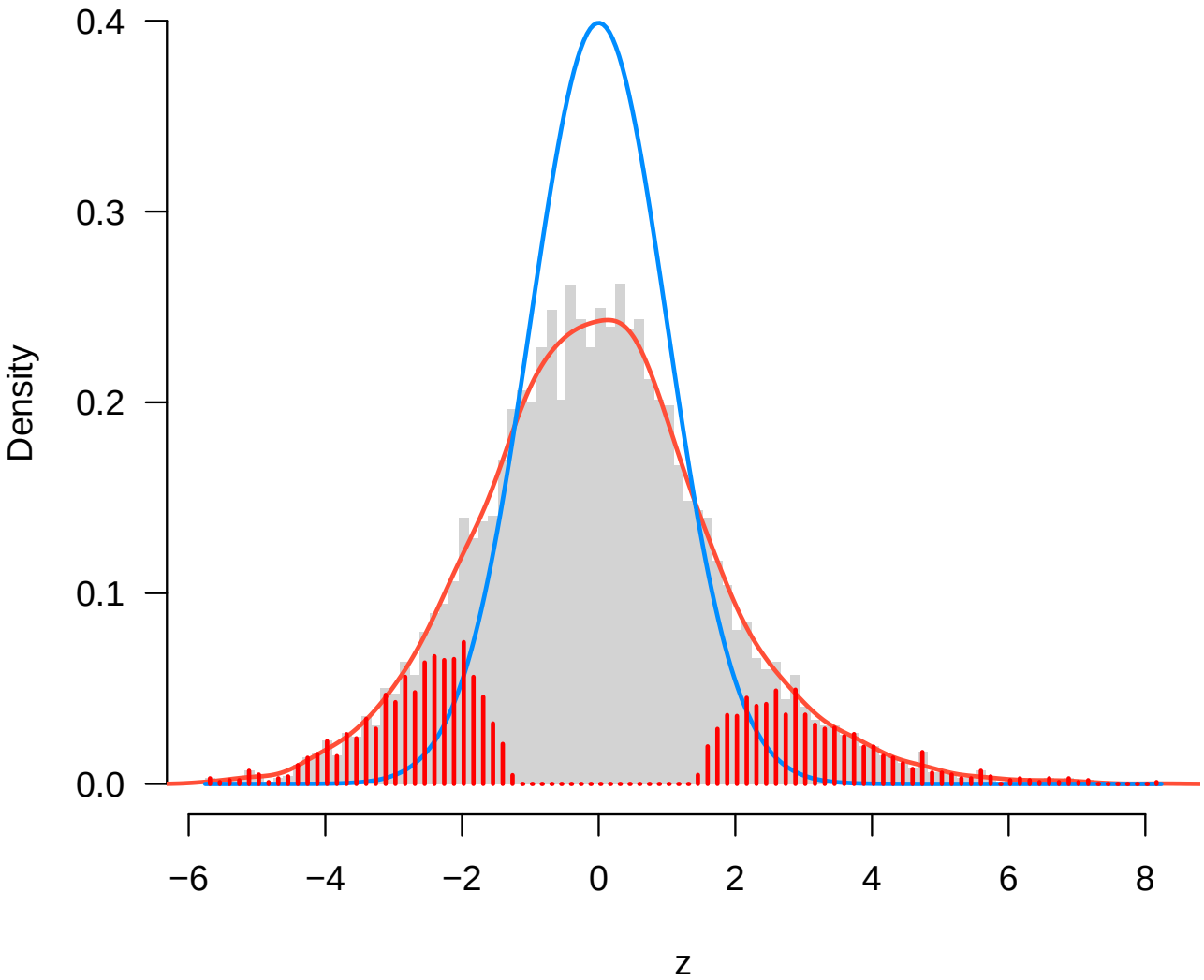
Remarks

- In principle, we could use binning, as in histograms, to estimate local FDRs
- For example, in the leukemia data
  - We observe 310 p-values between 0.01 and 0.02
  - We would expect 7,129/100=71.3 p-values between 0.01 and 0.02 by random chance (if $\pi_0 = 1$)
  - Thus, for a p-value between .01 and .02, the local FDR is 71.3/310=23%
- However, there are much better approaches to estimating densities, which we will discuss later

Remarks (cont'd)

- Two challenges faced by local FDRs for gaining widespread acceptance over the FDR from the previous lecture are:
  - Density estimation is much less straightforward than estimating a distribution function
  - No interpretation as a frequentist error rate control procedure is available
- From a Bayesian perspective, however, 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

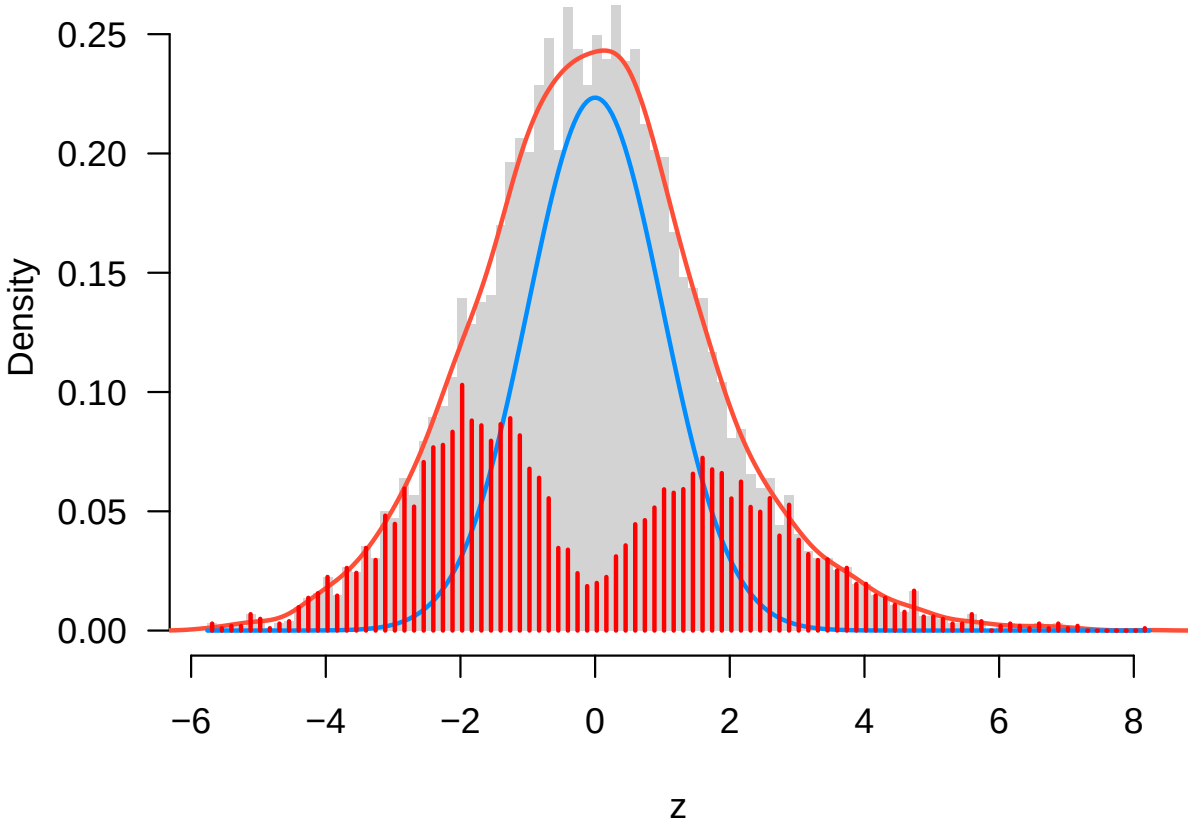
IFDR for leukemia data: Illustration



Remarks

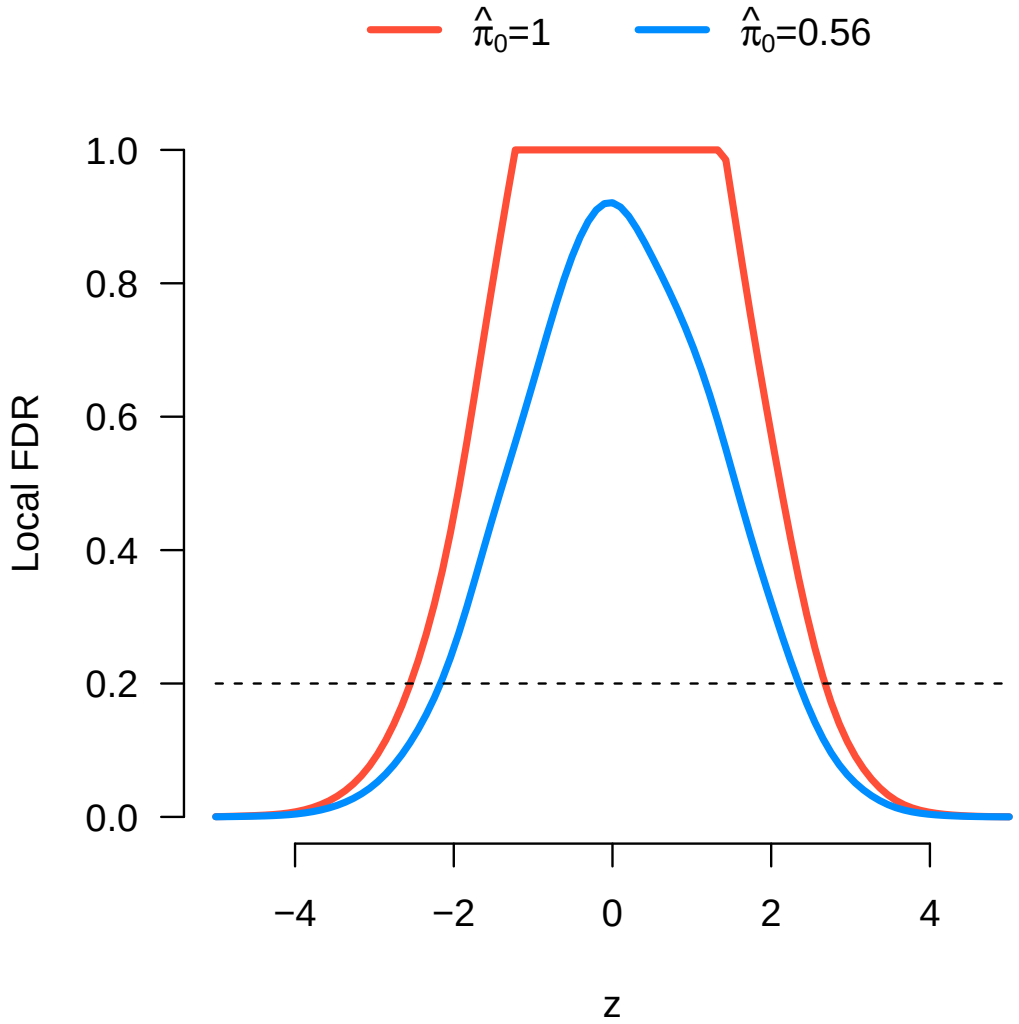
- For reasons we will discuss shortly, densities are much easier to estimate on the z-scale than the p-scale
- Subtracting the null density from the marginal density, it appears that the peak density of non-null z-values is around ± 2.5
- The preceding plot assumed $\pi_0 = 1$, which clearly results in some inconsistencies around $z = 0$, as the null density greatly exceeds the marginal density

IFDR for leukemia data: $\hat{\pi}_0 = 0.56$



Using $\hat{\pi}_0 = 0.56$, our estimate from the previous lecture, we seem to obtain much more realistic estimations of the null and alternative distributions

z vs local FDR



For a 20% local FDR cutoff:

- Using $\hat{\pi}_0 = 1$, critical value of $z = 2.67$; 937 significant results
- Using $\hat{\pi}_0 = 0.56$, critical value of $z = 2.35$; 1,328 significant results

R code

- There are a number of R packages for calculating local FDRs, all of which take different approaches to the estimation of the marginal density and the estimation of a null density, both of which we will discuss shortly
- The one that I am most familiar with is `locfdr`, developed by Efron and colleagues:

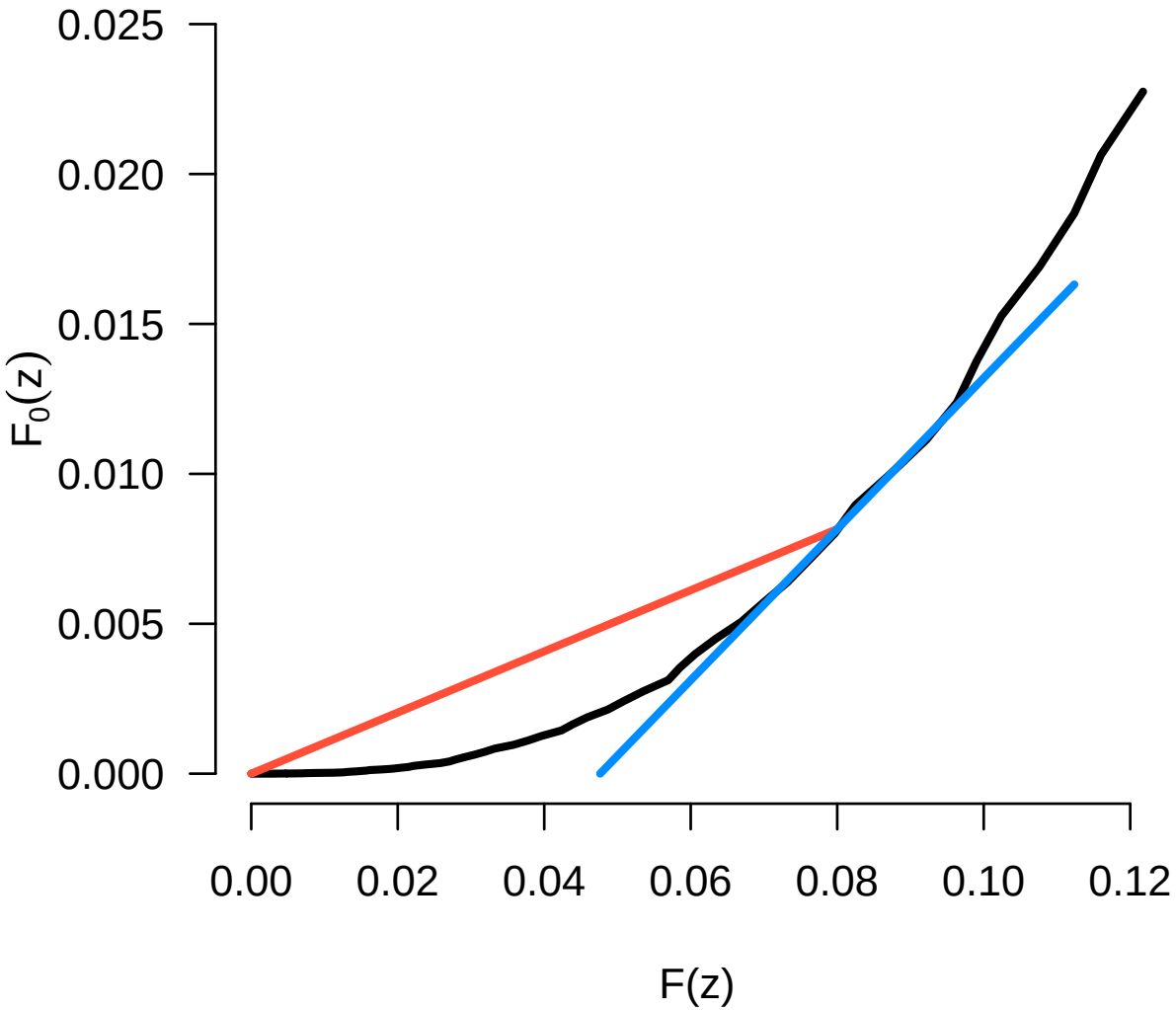
```
res <- locfdr(z)
```

  The function accepts z-values and returns a bunch of stuff, most importantly `res$fdr`, the estimated local FDR
- The function also produces plots similar to those on slides 10 and 12
- Another popular package for calculating local FDRs is `fdrtool`

Cutoff comparison

- It is worth spending a few slides on a deeper examination of FDR versus local FDR in terms of results and interpretation
- Using $\pi_0 = 1$, and a 10% cutoff,
  - FDR: Critical $z = 2.02$; 1,537 significant findings
  - Local FDR: Critical $z = 2.94$; 640 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% local FDR

FDR vs. local FDR: Geometry



Remarks

- Geometrically, the FDR is the slope of the secant line connecting that point to the origin
- Meanwhile, the local FDR is the slope of the tangent line
- The tangent line will have a higher slope provided that we are in the tail of the distribution (and that the marginal distribution has thicker tails than the null distribution)

Conditional expectation relationship

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

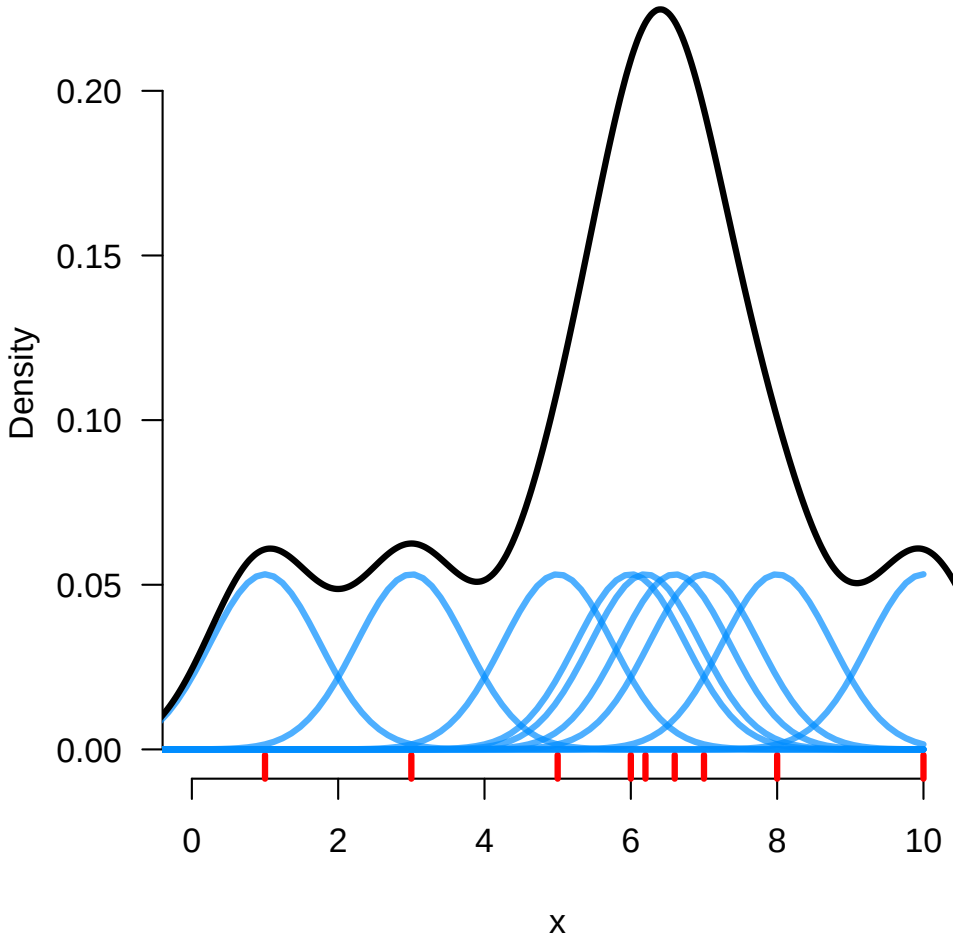
$$\mathbb{E}\{\phi(z)|z \in \mathcal{Z}\} = \phi(\mathcal{Z}),$$

  where $\phi(z)$ is the local FDR at point z and $\phi(\mathcal{Z})$ is the FDR over the set $\mathcal{Z}$
- Roughly, then, we should expect the average local FDR among the significant features to equal the FDR:
  - Left tail: Average lFDR for features with $\mathrm{FDR} < 0.1$ is 0.097
  - Right tail: Average lFDR for features with $\mathrm{FDR} < 0.1$ is 0.093
- This relationship does not exactly work out for two-sided tests unless we specifically estimate a combined tail density $f(|z|)$

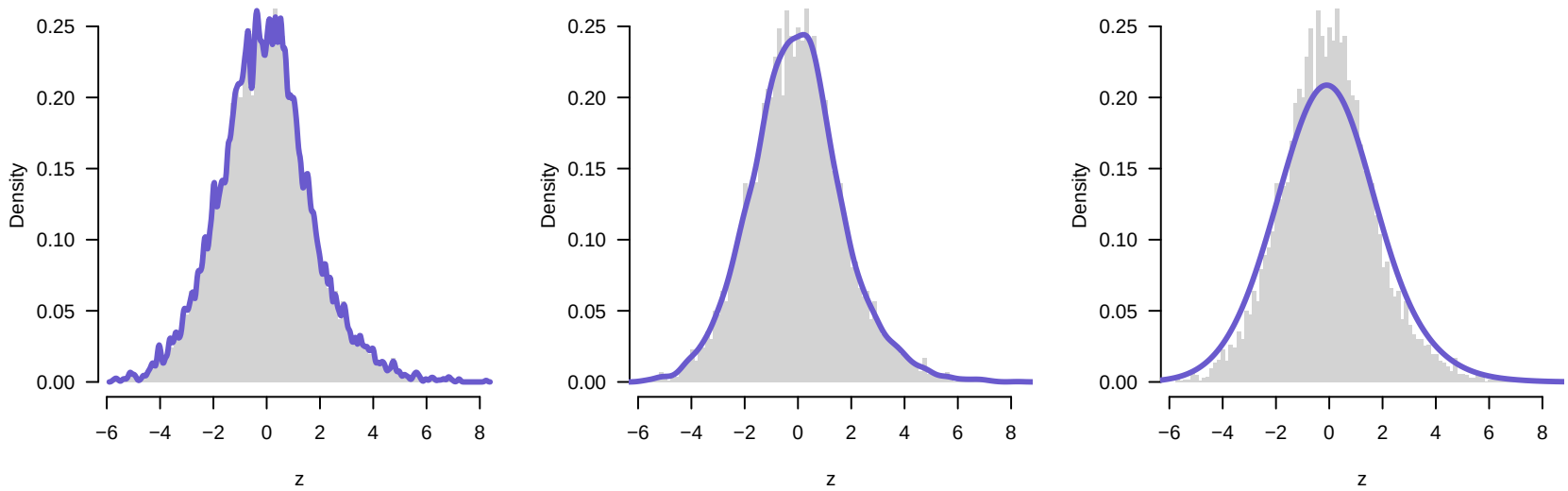
Introduction

- Accurate estimation of local FDRs obviously depends on obtaining accurate estimates of the marginal density $f(z)$
- There are many ways to do this – in part, this is why there are several packages for estimating local FDRs
- A comprehensive treatment of density estimation is beyond the scope of this course, but we can see the basic idea of a very common approach, *kernel density estimation*, in just a few slides

Density estimation using Gaussian kernels



Choice of bandwidth



Remarks

- The standard `R` function for estimating densities is `density()`, which also provides several automatic approaches to selecting an optimal bandwidth
- The `density()` function uses the kernel approach; the `locfdr` package uses a different approach; however, for the leukemia data the agreement between the resulting local FDRs is essentially perfect ($\hat{\rho} > 0.999$)
- Regardless of the method you use, I would urge you to always look at a resulting plot of histogram and density estimate to make sure that the method is producing reasonable results

Estimating a null distribution?

- Our last concept for today is the idea of estimating the null distribution
- This is somewhat controversial and not a necessary part of using local false discovery rates, but many software packages for local FDRs, including `locfdr`, do this, so it is important to be aware of the issues
- Estimating a null distribution is, for the most part, not something one does in conventional hypothesis testing
- Large-scale testing, however, opens up the possibility of empirically estimating the null distribution

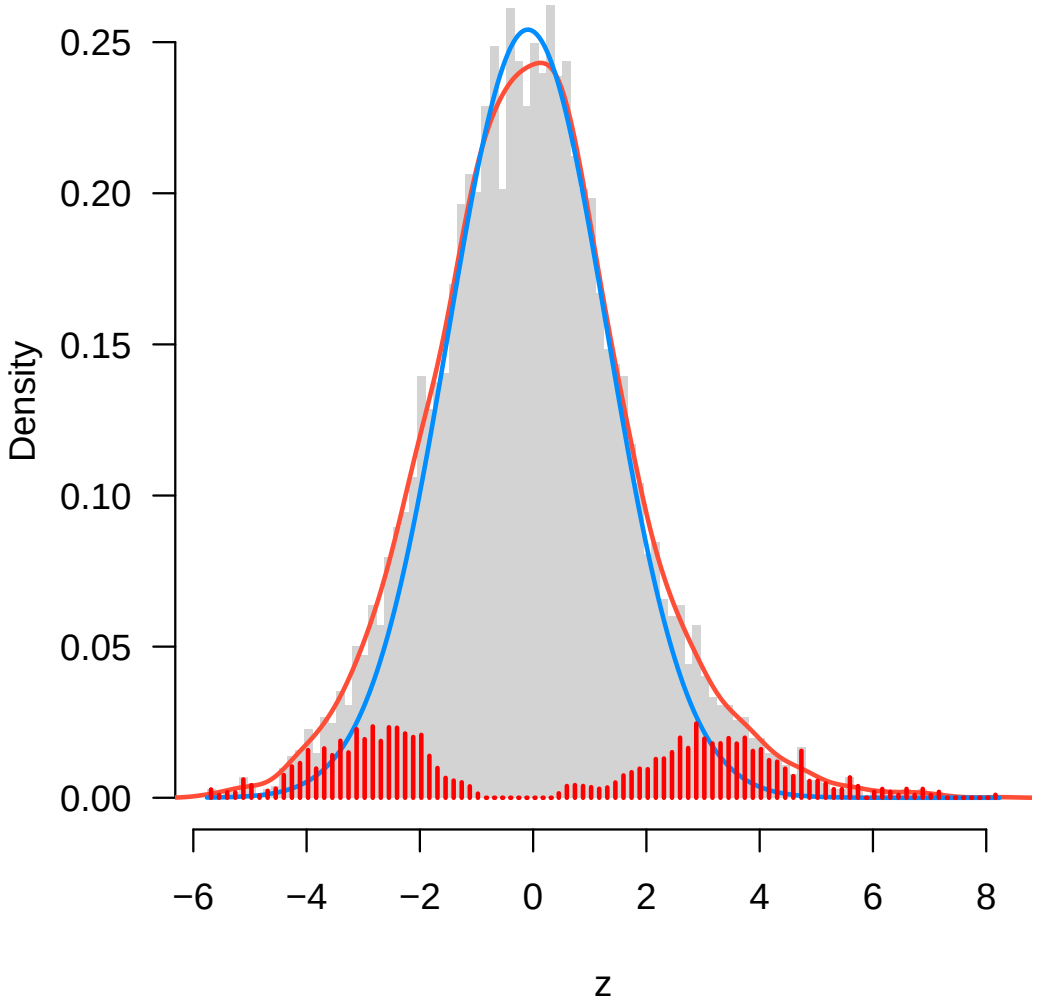
Methods for estimating the null distribution

- Estimating a null distribution typically relies on two assumptions:
  - The non-null density is close to zero near $z = 0$: $f_1(0) \approx 0$
  - The null distribution is still normal, albeit with a possibly different mean and variance: $Z \overset{H_0}{\sim} \mathrm{N}(\delta_0, \sigma_0^2)$
- *Central matching* estimates the null distribution by assuming that $\log f(z)$ is quadratic near zero: estimating the three parameters of the quadratic function provide estimates for π_0, δ_0, and σ_0
- *Maximum likelihood* defines a central region, $\mathcal{A}_0$, and uses maximum likelihood methods for truncated normal distributions to estimate the parameters of the null distribution

Leukemia data: Results

	$\hat{\delta}_0$	$\hat{\sigma}_0$	$\hat{\pi}_0$
Theoretical	0.00	1.00	0.69
Central matching	-0.09	1.40	0.89
Maximum likelihood	-0.13	1.52	0.95

Leukemia results w/ estimation of null



Remarks

- It is certainly possible, for a variety of reasons, for the theoretical null $N(0, 1)$ not to hold
- Whether this is a more convincing explanation, in the case of the leukemia data, than the explanation that a large number of hypotheses are legitimately false is difficult for me to say
- Furthermore, if the theoretical null is badly violated, it may indicate fundamental problems with the experiment that simply estimating an empirical null distribution cannot fix
- Regardless, it's a very interesting idea and one worth being aware of, although it's important to think critically about whether it is the right thing to do for the problem at hand