Local FDR Simulation Example

Bradley Efron and Balasubramanian Narasimhan
Department of Statistics
Stanford University
Stanford, CA 94305

October 19, 2004

1 A Simulated Example

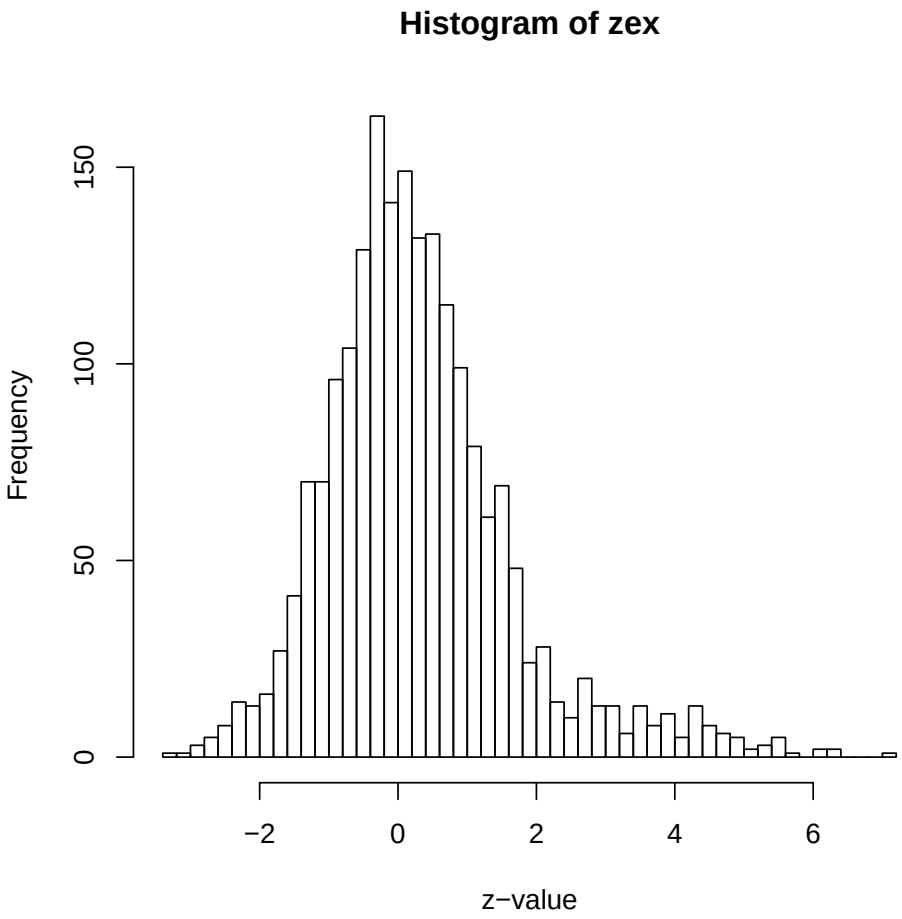
This simulation example involves 2000 "genes", each of which has yielded a test statistic z_i, with $z_i \approx N(mu_i, 1)$, independently for $i = 1, 2, \ldots, 2000$.

Here mu_i is the "true score" of gene i, which we observe only noisily. 1800 (90%) of the μ values are zero; the remaining 200 (10%) are from a $N(3, 1)$ distribution. The data are contained in the dataset `lfdrsim`, where the z_i are the column `zex`.

```
> library(locfdr)
> data(lfdrsim)
> zex <- lfdrsim[, 2]
```

A histogram shows that the z_i have a long tail to the right of zero, but with no obvious second mode near $z = 3$.

```
> hist(zex, breaks = seq(-3.4, 7.2, 0.2), xlab = "z-value")
```



The `locfdr` package allows us to compute the fdr values for the 2000 genes, the descriptive vector `f0.p0` and a 119×7 matrix of local fdr values.

```
> w <- locfdr(zex)

Loading required package: splines
```

Let's examine `f0.p0`.

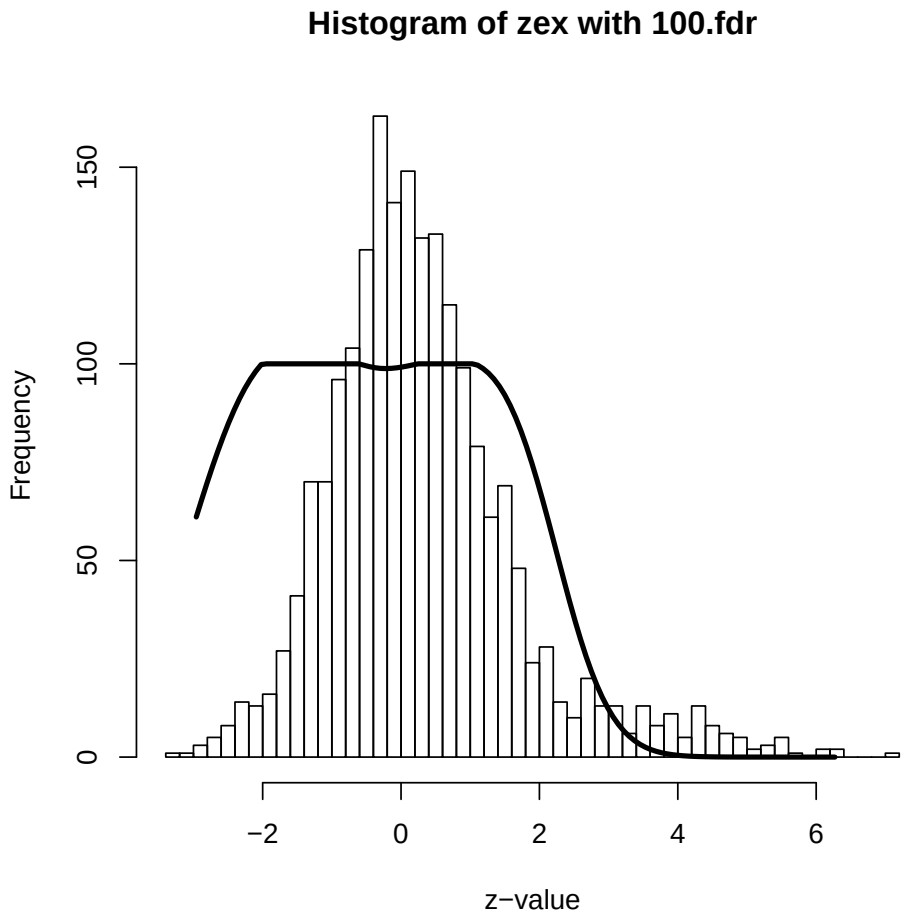
```
> print(w$f0.p0)

      zmax        sig         p0
0.02138590 0.97936902 0.91644463
```

The above indicates that the empirical null $f_0(z)$ was estimated to be $N(.0214, .979^2)$, and that the estimated proportion of null cases is 0.916. (The fitting method is conservative in the sense of tending to overestimate the true null proportion.) In this case the empirical null has done a good job of estimating what we happen to know is the true $N(0, 1)$ null.

We now add the fdr plot to the histogram (scaled up by a factor of 100).

```
> hist(zex, breaks = seq(-3.4, 7.2, 0.2), xlab = "z-value", main = "Histogram of zex with 100.f
> lines(w$mat[, 1], 100 * w$mat[, 2], lwd = 3)
```



This shows that the only small $fdr(z)$ values are on the right side, as they should be, with $fdr(z)$ declining to zero as z goes from 2 to 4.

We can now compute z such that $fdr(z) = 0.2$.

```
> zp <- approx(w$mat[, 2], w$mat[, 1], 0.2)
> print(zp)

$x
[1] 0.2

$y
[1] 2.791499
```

So, $fdr(2.79) = 0.2$. We now compute how many genes have fdr less than 0.2.

```
> sum(zex > zp$y)

[1] 117
```

Thus, we have 117 genes with fdr less than .2. Of these 117, only 4 are actually Nulls, i.e. have $\mu_i = 0$. This is in rough agreement with the tail area Fdr, `Fdrleft`, which equals .041 at $z = 2.79$.

The fact that the local fdr is nearly five times greater, .2 compared to .041, shows that genes near the boundary point 2.79 are much more likely to be false discoveries than the average gene having $z > 2.79$.