Multiple Testing par II

Alexander McLain

November 15th, 2021

Example of leukemia data

We'll illustrate the multiple testing methods with gene expression data from the leukemia ALL/AML study of Golub et al. (1999). Load the leukemia dataset:

```
library(multtest) #Useful package for multiple testing and data
```

```
## Warning: package 'BiocGenerics' was built under R version 4.0.5
```

```
library(genefilter)
library(locfdr)
library(qvalue)
# Load some functions from Patrick Breheny's website
source('http://myweb.uiowa.edu/pbreheny/7600/s16/notes/fun.R')
data(golub)
dim(golub)
```

```
## [1] 3051   38
```

Note that each column is a sample (subject), and $golub_{j,i}$ is the expression level for gene j in tumor mRNA sample i. All of the genes have identifiers and tumor class labelss (0 for ALL, 1 for AML).

```
dim(golub.gnames)
```

```
## [1] 3051    3
```

```
golub.gnames[1:4, ]
```

36	AFFX-HUMISGF3A/M97935_MA_at (endogenous control)	AFFX-HUMISGF3A/M97935_MA_at
37	AFFX-HUMISGF3A/M97935_MB_at (endogenous control)	AFFX-HUMISGF3A/M97935_MB_at
38	AFFX-HUMISGF3A/M97935_3_at (endogenous control)	AFFX-HUMISGF3A/M97935_3_at
39	AFFX-HUMRGE/M10098_5_at (endogenous control)	AFFX-HUMRGE/M10098_5_at

```
golub.cl
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
```

We'll use the *rowttests* to compute the test statistics and p-values. We'll also calculate the adjusted p-values for the methods from last class.

```
t.tests <- rowttests( golub, factor(golub.cl) )
str(t.tests)
```

```
'data.frame':   3051 obs. of  3 variables:
 $ statistic: num  -2.502 -1.156 0.11 0.273 1.187 ...
 $ dm       : num  -0.4923 -0.2179 0.0199 0.1695 0.7266 ...
 $ p.value  : num  0.017 0.255 0.913 0.787 0.243 ...
 - attr(*, "df")= num [1:3051] 36 36 36 36 36 36 36 36 36 36 ...
```

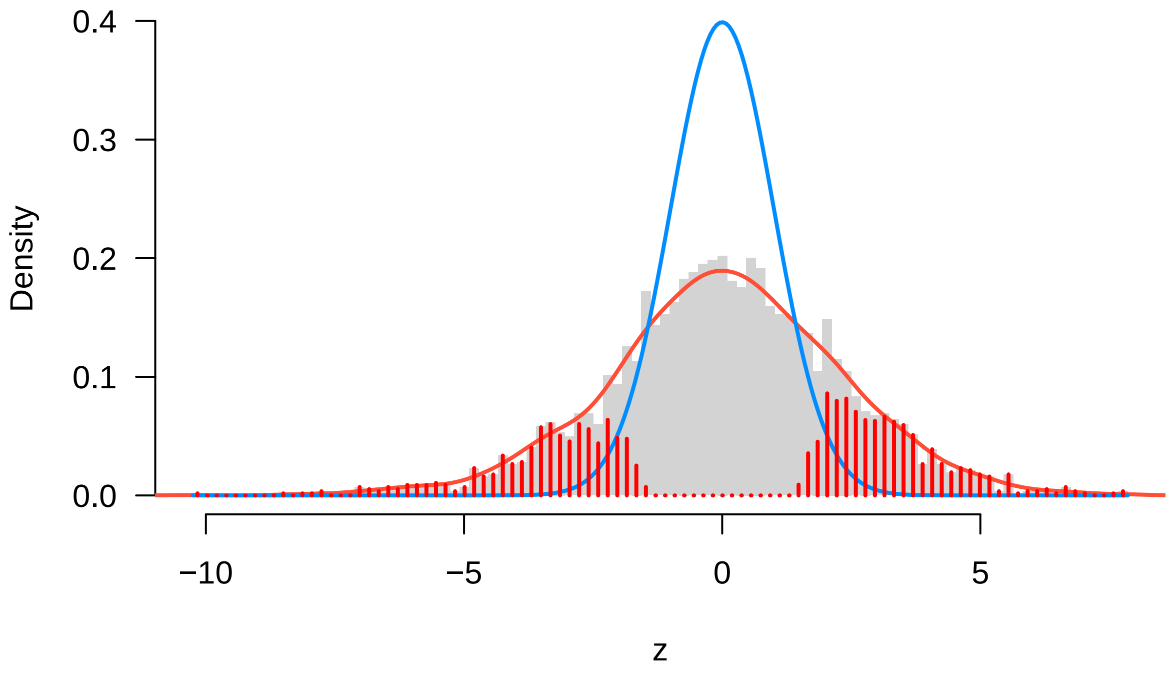
```
test.stat <- t.tests$statistic
p_vals <- t.tests$p.value
p_bonf <- p.adjust(p_vals, method = "bonf")
q_bh <- p.adjust(p_vals, method = "fdr")
q_st <- qvalue(p_vals)
```

Let's take a look at a histogram of the test statistics with the estimated null and marginal distribution.

```
lfdrPlot
```

```
## function (z, pi0 = 1, delta = 0, sigma = 1, lfdrReturn = TRUE,
##     ...)
## {
##     dens <- density(z, bw = "nrd")
##     f <- approxfun(dens$x, dens$y)
##     f0 <- function(z) pi0 * dnorm(z, mean = delta, sd = sigma)
##     lfdr <- pmin(f0(z)/f(z), 1)
##     h <- hist(z, breaks = seq(min(z), max(z), length = 99), plot = FALSE)
##     zz <- seq(min(z), max(z), len = 299)
##     ylim = c(0, max(c(h$density, dens$y, f0(delta))))
##     plot(h, main = "", border = FALSE, col = "lightgray", freq = FALSE,
##         las = 1, ylim = ylim)
##     lines(dens, col = pal(2)[1], lwd = 2)
##     lines(zz, f0(zz), col = pal(2)[2], lwd = 2)
##     fdr.zz <- f0(h$mids)/f(h$mids)
##     y <- pmax(h$density * (1 - fdr.zz), 0)
##     for (k in 1:length(h$mids)) lines(rep(h$mids[k], 2), c(0,
##         y[k]), lwd = 2, col = "red")
##     if (lfdrReturn)
##         return(lfdr)
## }
```

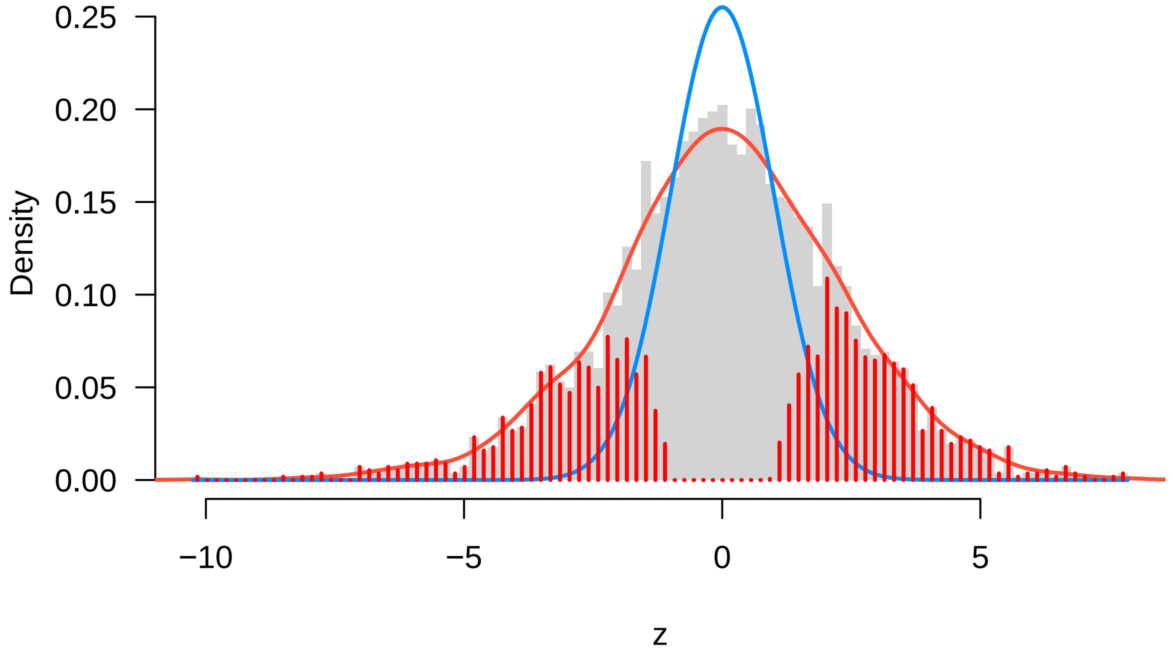
```
lambda = 0.1
m <- nrow(golub)
lfdr1 <- lfdrPlot(test.stat, pi0=1)
```



```
pi0=sum(p_vals> lambda) /((1-lambda)*m)
print(pi0)
```

```
## [1] 0.6394989
```

```
lfdr2 <- lfdrPlot(test.stat, pi0=pi0)
```



Now, we'll estimate the local fdr using the localfdr package.

```
?locfdr
```

```
Local False Discovery Rate Calculation

Description:
```

Compute local false discovery rates, following the definitions and description in references listed below.

Usage:

```
     locfdr(zz, bre = 120, df = 7, pct = 0, pct0 = 1/4, nulltype = 1, type =
     0, plot = 1, mult, mlests, main = " ", sw = 0)
```

Arguments:

zz: A vector of summary statistics, one for each case under simultaneous consideration. The calculations assume a large number of cases, say 'length(zz)' exceeding 200. Results may be improved by transforming zz so that its elements are theoretically distributed as N(0,1) under the null hypothesis. See the locfdr vignette for tips on creating zz.

bre: Number of breaks in the discretization of the z-score axis, or a vector of breakpoints fully describing the discretization. If 'length(zz)' is small, such as when the number of cases is less than about 1000, set bre to a number lower than the default of 120.

df: Degrees of freedom for fitting the estimated density f(z).

pct: Excluded tail proportions of zz's when fitting f(z). 'pct=0' includes full range of zz's. pct can also be a 2-vector, describing the fitting range.

pct0: Proportion of the zz distribution used in fitting the null density f0(z) by central matching. If a 2-vector, e.g. 'pct0=c(0.25,0.60)', the range [pct0[1], pct0[2]] is used. If a scalar, [pct0, 1-pct0] is used.

nulltype: Type of null hypothesis assumed in estimating f0(z), for use in the fdr calculations. 0 is the theoretical null N(0,1), 1 is maximum likelihood estimation, 2 is central matching estimation, 3 is a split normal version of 2.

type: Type of fitting used for f; 0 is a natural spline, 1 is a polynomial, in either case with degrees of freedom df [so total degrees of freedom including the intercept is 'df+1'.]

plot: Plots desired. 0 gives no plots. 1 gives single plot showing the histogram of zz and fitted densities f and p0*f0. 2 also gives plot of fdr, and the right and left tail area Fdr curves. 3 gives instead the f1 cdf of the estimated fdr curve; plot=4 gives all three plots.

mult: Optional scalar multiple (or vector of multiples) of the sample size for calculation of the corresponding hypothetical Efdr value(s).

mlests: Optional vector of initial values for (delta0, sigma0) in the maximum likelihood iteration.

main: Main heading for the histogram plot when 'plot>0'.

sw: Determines the type of output desired. 2 gives a list consisting of the last 5 values listed under Value below. 3 gives the square matrix of dimension bre-1 representing the influence function of log(fdr). Any other value of sw returns a list consisting of the first 5 (6 if mult is supplied) values listed below.

Details:

See the locfdr vignette for details and tips.

Value:

fdr: the estimated local false discovery rate for each case, using the selected type and nulltype.

fp0: the estimated parameters delta (mean of f0), sigma (standard deviation of f0), and p0, along with their standard errors.

Efdr: the expected false discovery rate for the non-null cases, a measure of the experiment's power as described in Section 3 of the second reference. Overall Efdr and right and left values are given, both for the specified nulltype and for nulltype 0. If 'nulltype==0', values are given for nulltypes 1 and 0.

cdf1: a 99x2 matrix giving the estimated cdf of fdr under the non-null distribution f1. Large values of the cdf for small fdr values indicate good power; see Section 3 of the second reference. Set plot to 3 or 4 to see the cdf1 plot.

mat: A matrix of estimates of f(x), f0(x), fdr(x), etc. at the bre-1 midpoints "x" of the break discretization, convenient for comparisons and plotting. Details are in the locfdr vignette.

z.2: the interval along the zz-axis outside of which $fdr(z)<0.2$, the locations of the yellow triangles in the histogram plot. If no elements of zz on the left or right satisfy the criterion, the corresponding element of z.2 is NA.

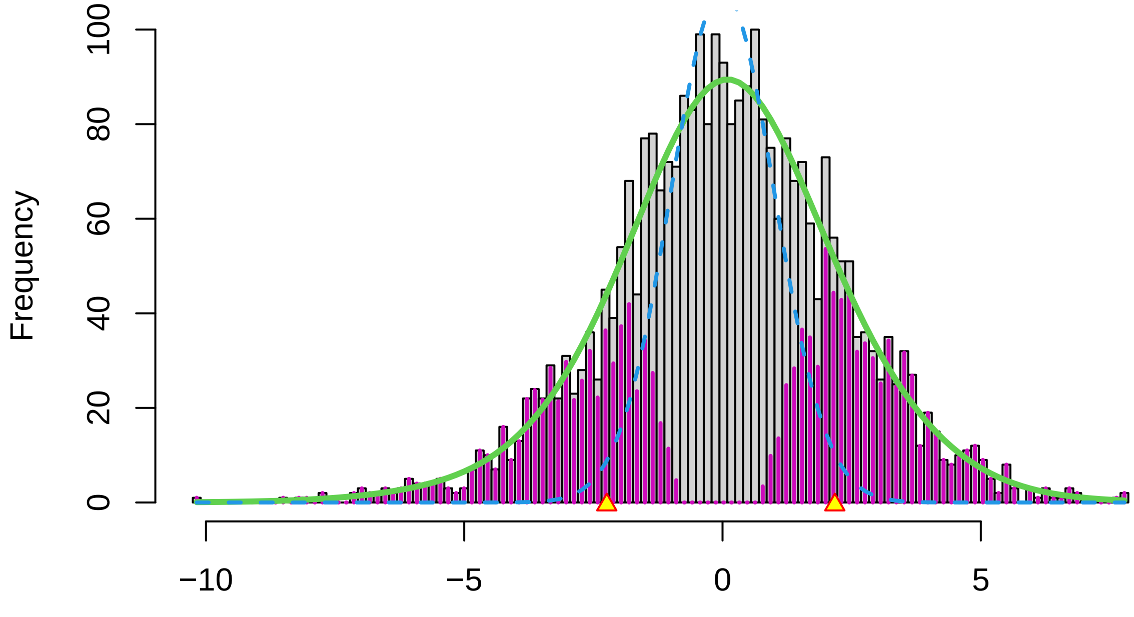
call: the function call.

mult: If the argument mult was supplied, vector of the ratios of hypothetical Efdr for the supplied multiples of the sample size to Efdr for the actual sample size.

pds: The estimates of p0, delta, and sigma.

```
       x: The bin midpoints.

       f: The values of f(z) at the bin midpoints.

    pds.: The derivative of the estimates of p0, delta, and sigma with
          respect to the bin counts.

   stdev: The delta-method estimates of the standard deviations of the
          p0, delta, and sigma estimates.

References:

     Efron, B. (2004) "Large-scale simultaneous hypothesis testing: the
     choice of a null hypothesis", Jour Amer Stat Assoc, *99*, pp.
     96-104

     Efron, B. (2006) "Size, Power, and False Discovery Rates"

     Efron, B. (2007) "Correlation and Large-Scale Simultaneous
     Significance Testing", Jour Amer Stat Assoc, *102*, pp. 93-103

     <URL: http://statweb.stanford.edu/~ckirby/brad/papers/>
```

```
res <- locfdr(test.stat, nulltype = 0)
```



```
lfdr_res <- res$fdr
summary(lfdr_res)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0.1162583	0.5824143	0.5470777	1	1

```r
s_lfdr <- sort(lfdr_res)
Fdr <- cumsum(s_lfdr)/(1:m)
lfdr_cut <- max(s_lfdr[Fdr < 0.05])
lfdr_cut
```

```
[1] 0.1967217
```

```r
data.frame("Bonf" = sum(p_bonf <= 0.05), "BH" = sum(q_bh <= 0.05),
           "Storey" = sum(q_st$qvalues <= 0.05), "lfdr" = sum(lfdr_res <= lfdr_cut))
```

Bonf	BH	Storey	lfdr
98	681	876	919

```r
summary(test.stat[q_st$qvalues <= 0.05 & test.stat<0])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-10.25597	-4.214273	-3.329132	-3.682799	-2.786718	-2.295487

```r
summary(test.stat[lfdr_res <= lfdr_cut & test.stat<0])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-10.25597	-4.204883	-3.325732	-3.665462	-2.776935	-2.253164

```r
summary(test.stat[q_st$qvalues <= 0.05 & test.stat>0])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.281612	2.683075	3.245655	3.49488	3.997593	7.855191

```r
summary(test.stat[lfdr_res <= lfdr_cut & test.stat>0])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.184034	2.557896	3.151972	3.400174	3.91658	7.855191