Week 06 Good statistical pratice II

Open and reproducible science: dependable computations and statistics

In-class tasks - Solution

Gene expression study of BRCA tumors in breast cancer patients

I Hedenfalk et al. investigated gene-expression profiles in hereditary breast cancer. They looked at 3226 genes, carrying out a two-sample t-test for each gene to see if the expression level of the gene differed between women with one type of breast cancer (BRCA1 mutations) and another type (BRCA2 mutations).

Task 1 Bonferroni adjustment

Question 1
What is the probability that the authors would reject at least one null hypothesis at a level of 5 % even if all null hypotheses were true (assume independence)?

```r
sig.level <- 0.05
1-(1-sig.level)^3226
```

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

What is the Bonferroni adjusted level in this case?

```r
format(round(sig.level/3226,6), scientific = FALSE)
```

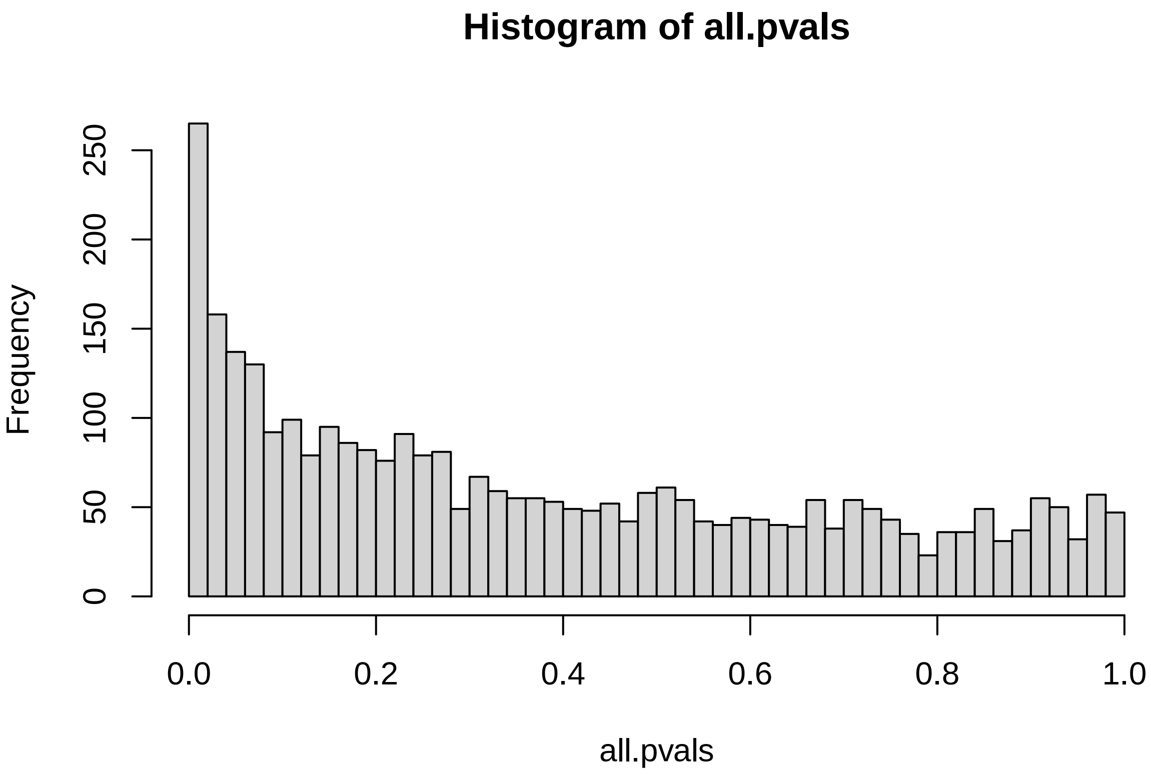
```
## [1] "0.000015"
```

Question 2
The data are provided on the course platform and can be read in and prepared with the following steps:

```r
cancer <- read.table("data/hedenfalk.txt", header = FALSE)
# data from here: https://myweb.uiowa.edu/pbreheny/data/hedenfalk.html
# info on website is transposed:
# one column per gene
# first columns contains mutation information
# first row gen ID
# isolate the gene expression ratios
dat <- cancer[2:dim(cancer)[1],2:dim(cancer)[2]]
# transpose to have one gene per row
dat <- t(dat)
# mutation information
mutations <- cancer[-1,1]
```

Calculate the p-values of the 3226 tests. How many are below 5%? And below the Bonferroni corrected threshold?

```
# calculate all p-values
all.pvals <- apply(dat,1 , function (x) t.test(x[mutations == "BRCA1"],x[mutations == "BRCA2"], var.equa
hist(all.pvals,40)
```



```
sum(all.pvals < sig.level)
```

```
## [1] 498
```

```
sum(all.pvals < sig.level/dim(dat)[1])
```

```
## [1] 2
```

Task 2 False discovery rate

Question 1

A different idea than controlling the FWER (i.e. controlling to make a single mistake with a very low chance) one can also try to limit the number of type I errors.

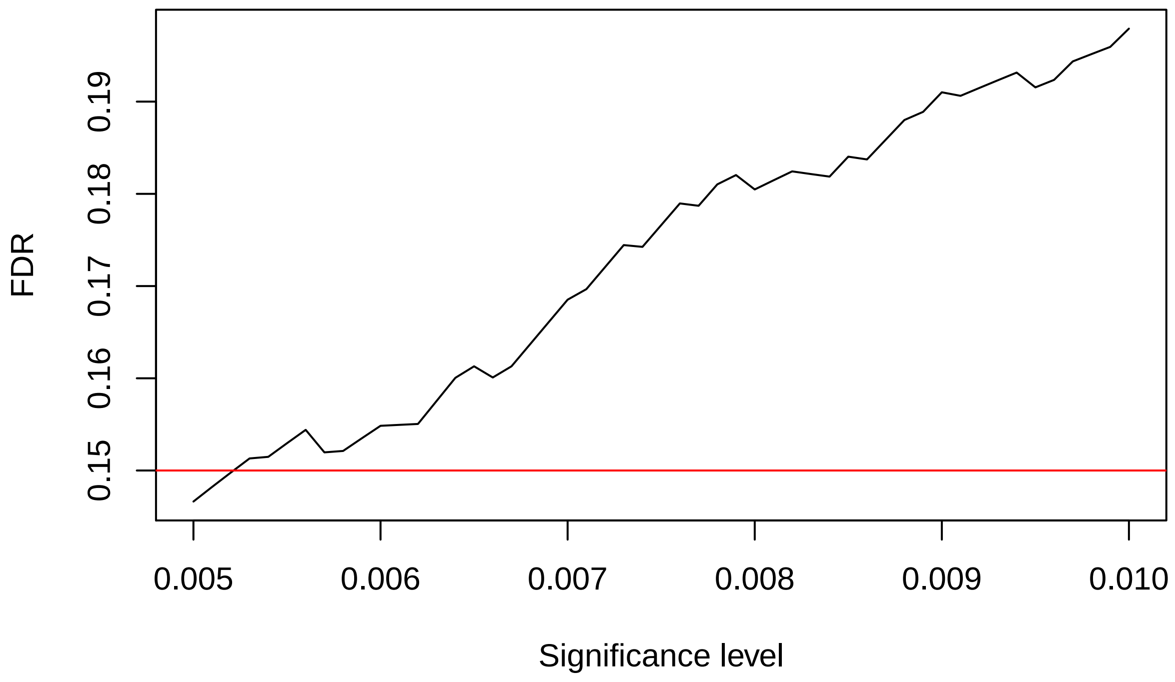
The so-called false discovery rate estimates the proportion of significant findings that are type I errors. It can be estimated by the nominal number of type errors, here 3226*0.05 divided by the number of times the p-value was indeed below 5 %. Estimate the FDR in our case.

```
round(dim(dat)[1]*sig.level/sum(all.pvals < sig.level),3)
```

```
## [1] 0.324
```

Question 2
If you now want to fix the FDR at 15%, what significance level should you consider? Calculate FDR for a sequence of possible significance levels, the FDR for 5% gives you an indication that the target needs to be rather small. Compare with the Bonferroni corrected significance level by indicating the number of positive tests with both methods. You can also use the function `p.adjust`.

```
targetFDR <- 0.15
sig.levels <- seq(0.005,0.01, by=0.0001)
allfdr <- NULL
for ( i in 1: length(sig.levels)){
  allfdr <- c(allfdr,dim(dat)[1]*sig.levels[i]/sum(all.pvals < sig.levels[i]))}
plot(sig.levels,allfdr,type='l', xlab = "Significance level", ylab = "FDR") +
abline(h = 0.15, col = "red")
```



```
## integer(0)
```

```
allfdr[1:5]
```

```
## [1] 0.1466364 0.1482216 0.1497786 0.1513080 0.1514817
```

```
sig.levelFDR <- sig.levels[3]
sig.levelFDR
```

```
## [1] 0.0052
```

```
sum(all.pvals < sig.level/dim(dat)[1])
```

```
## [1] 2
```

```
sum(all.pvals < sig.levelFDR)
```

```
## [1] 112
```

```
sum(p.adjust(all.pvals, method ="bonferroni") < sig.level)
```

```
## [1] 2
```

```
sum(p.adjust(all.pvals, method ="BH") < .15)
```

```
## [1] 112
```