Case study: Van de Vijver and Rosenwald data

Data

To illustrate all the different multiple testing methods, we use two gene expression microarray data sets, both with a survival phenotype.

1. The first one is the data of Van de Vijver. This data set has gene expression profiles of $m = 4919$ probes for 295 breast cancer patients.
2. The second data set, of Rosenwald, has gene expression profiles of $m = 7399$ probes for 240 diffuse B-cell lymphoma patients.

Median follow up was 9 years for the breast cancer patients and 8 years for the lymphoma patients. Although both data sets are by now a decade old, and new technologies, such as RNA sequencing, have since appeared, the multiple testing issues at stake have not changed; only the scale of problems has increased further.

Research question

In a gene expression microarray experiment, we want to test for differential expression of each of the probes on the microarray chip.

The purpose of the experiment is to come up with a list of promising candidates, to be further investigated by the same research group before publication.

The microarray experiment is often not the final experiment before publication of the scientific paper. Separate validation experiments usually follow for some or all of the probes found differentially expressed.

Model

In each data set, we performed a likelihood ratio test in a Cox proportional hazards model for each probe, testing association of expression with survival. The resulting p-values can be imported in R by

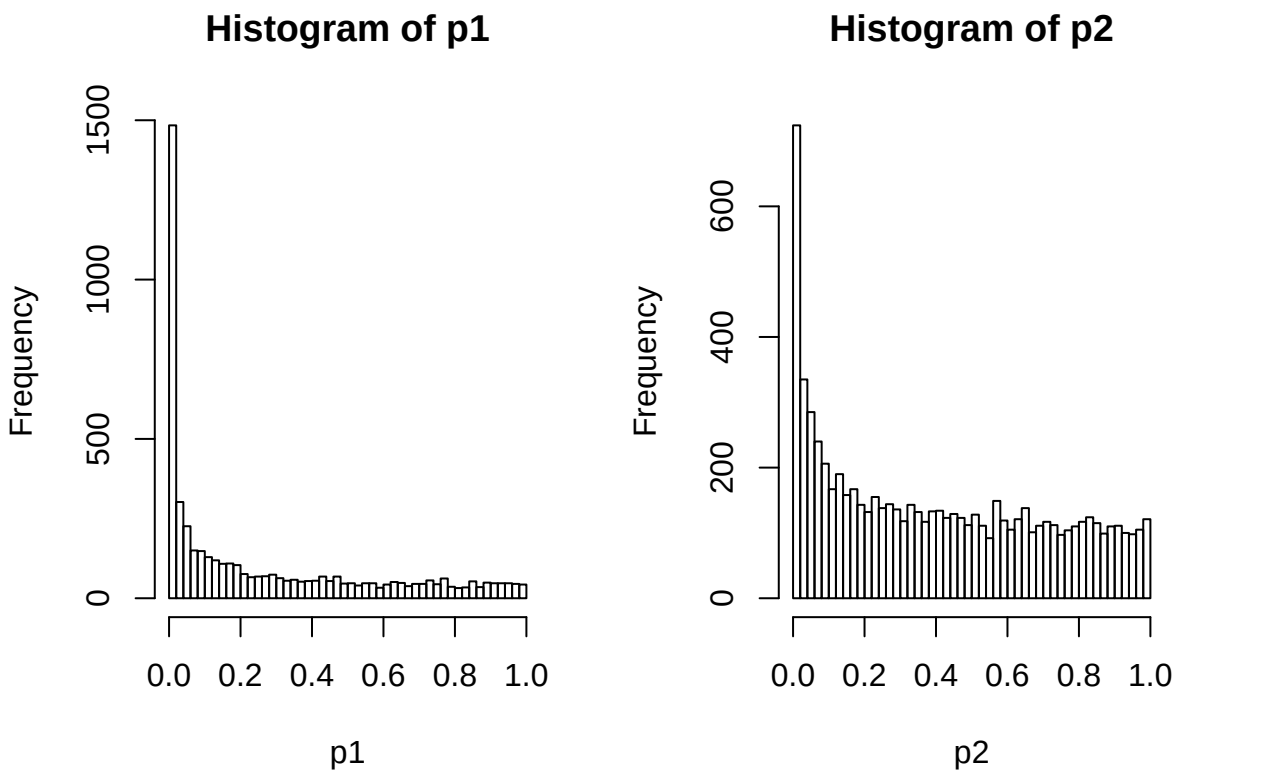
```
load("9_VandeVijver.Rdata")
p1 = p_vandevijver
p2 = p_rosenwald
```

Analysis

Histograms and plot of the sorted p-values

Histogram of p-values:

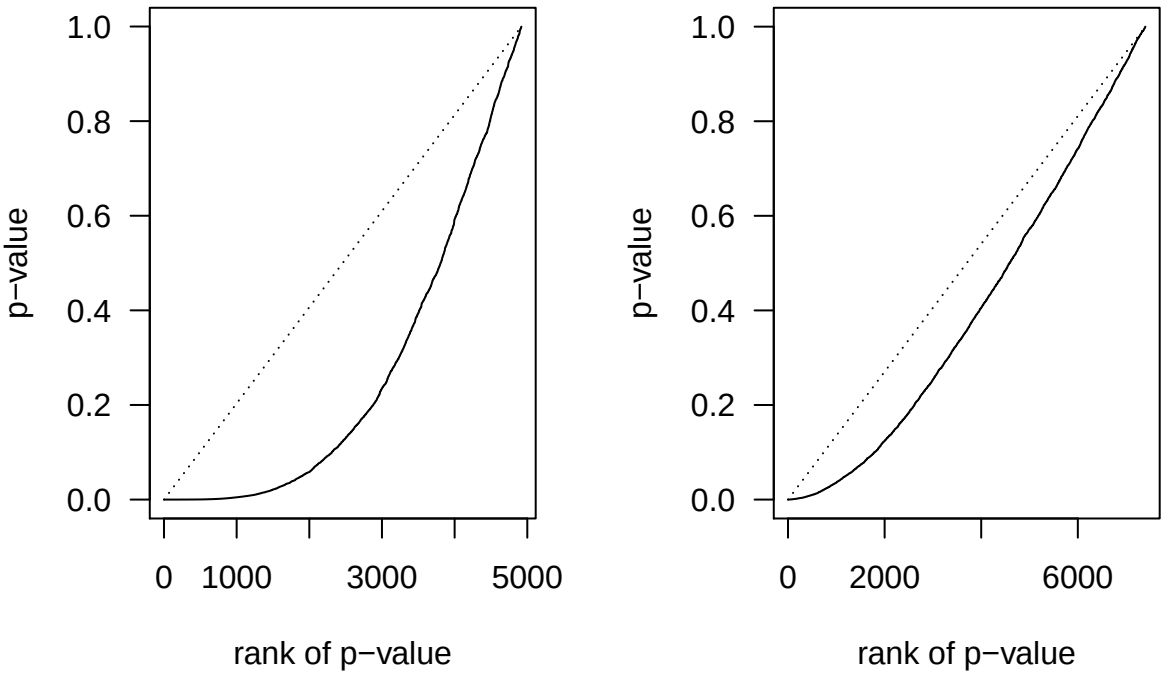
```
op <- par(mfrow = c(1, 2))
hist(p1, 50)
hist(p2, 50)
```




```
par(op)
```

Plot of the sorted p-values:

```
op <- par(mfrow = c(1, 2))
plot(sort(p1), type="l", xlab="rank of p-value", ylab="p-value", las=1)
lines(c(0,4919), c(0,1), lty=3)
plot(sort(p2), type="l", xlab="rank of p-value", ylab="p-value", las=1)
lines(c(0,7399), c(0,1), lty=3)
```




```
par(op)
```

From the plots of the Van de Vijver data, we may suspect that many of the hypotheses that are tested are false. If the overwhelming majority of the hypotheses were true, we would expect the histogram of the p-values to be approximately uniform and the plot of the sorted p-values approximately to follow the dotted line, because p-values of true hypotheses follow a uniform distribution.

There is less immediate evidence for differential expression in the Rosenwald data, although there seems to be enrichment of low p-values here too.

FWER controlling methods

Bonferroni and Holm

In the Van de Vijver data, the Bonferroni method rejects 203 hypotheses at a critical value of 0.05/4919.

```
alpha = 0.05

# Bonferroni
alpha/4919 # Bonferroni critical value
```

```
## [1] 1.016467e-05
```

```
sum(p1 <= alpha/4919) # number of rejections
```

```
## [1] 203
```

```
p1.Bonf = p.adjust(p1,"bonf") # adjusted p-values
sum(p1.Bonf <= alpha)  # number of rejections
```

```
## [1] 203
```

This allows the critical value in the second step of Holm's procedure to be adjusted to $0.05/(4919 - 203)$, which allows three more rejections. The increase in the critical value resulting from these three rejections is not sufficient to allow any additional rejections, giving a total of 206 rejections for Holm's procedure.

```
# Holm: sequential Bonferroni
R = 0
Rnew = sum(p1 <= alpha/4919)
while(Rnew > R){
print(Rnew)
R = Rnew
Rnew = sum(p1 <= alpha/(4919-R) )
}
```

```
## [1] 203
## [1] 206
```

```
p1.Holm <- p.adjust(p1, "holm")  # adjusted p-values
sum(p1.Holm <= alpha)  # number of rejections
```

```
## [1] 206
```

In the Rosenwald data, Bonferroni allows only four rejections at its critical value of $0.05/7399$, but the resulting increase in the critical value in Holm's method to $0.05/(7399-4)$ is insufficient to make a difference.

```
sum(p.adjust(p2, "bonf") <= alpha)  # Bonferroni: number of rejections
```

```
## [1] 4
```

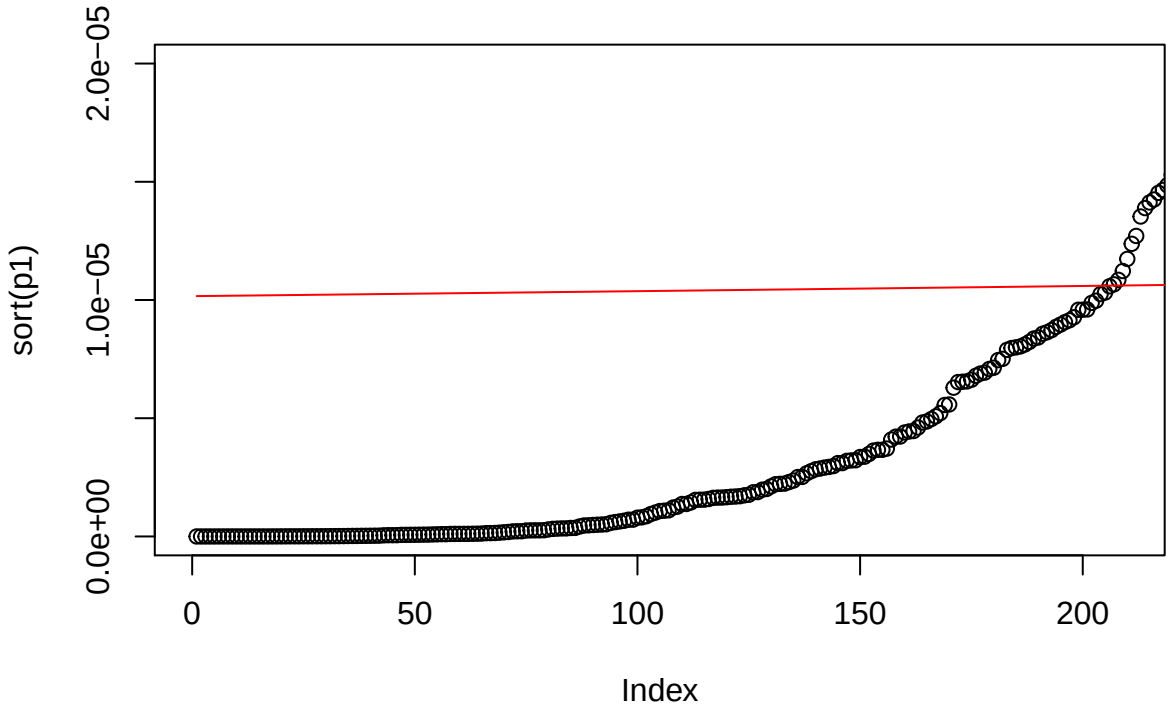
```
sum(p.adjust(p2, "holm") <= alpha)  # Holm: number of rejections
```

```
## [1] 4
```

Hochberg and Hommel

In the Van de Vijver data set, the curve of the ordered p-values crosses the curve of the Holm and Hochberg critical values only once, so the number of rejections is identical to 206 in both methods.

```
plot(sort(p1), ylim=c(0,0.00002), xlim=c(0,210))
lines(1:4919, alpha/(4919-1:4919 +1 ) , col=2) # Holm/Hochberg critical values
```



```
sum(p.adjust(p1, "hoch") <= alpha)  # Hochberg: number of rejections
```

```
## [1] 206
```

Hommel's method, however, is able to improve upon this by further three rejections, making a total of 209.

```
sum(p.adjust(p1, "hommel") <= alpha)  # Hommel: number of rejections
```

```
## [1] 209
```

In the Rosenwald data, neither method is able to improve upon the four rejections that were already found by the Bonferroni procedure.

```
sum(p.adjust(p2, "hochberg") <= alpha)  # Hochberg: number of rejections
```

```
## [1] 4
```

```
sum(p.adjust(p2, "hommel") <= alpha)  # Hommel: number of rejections
```

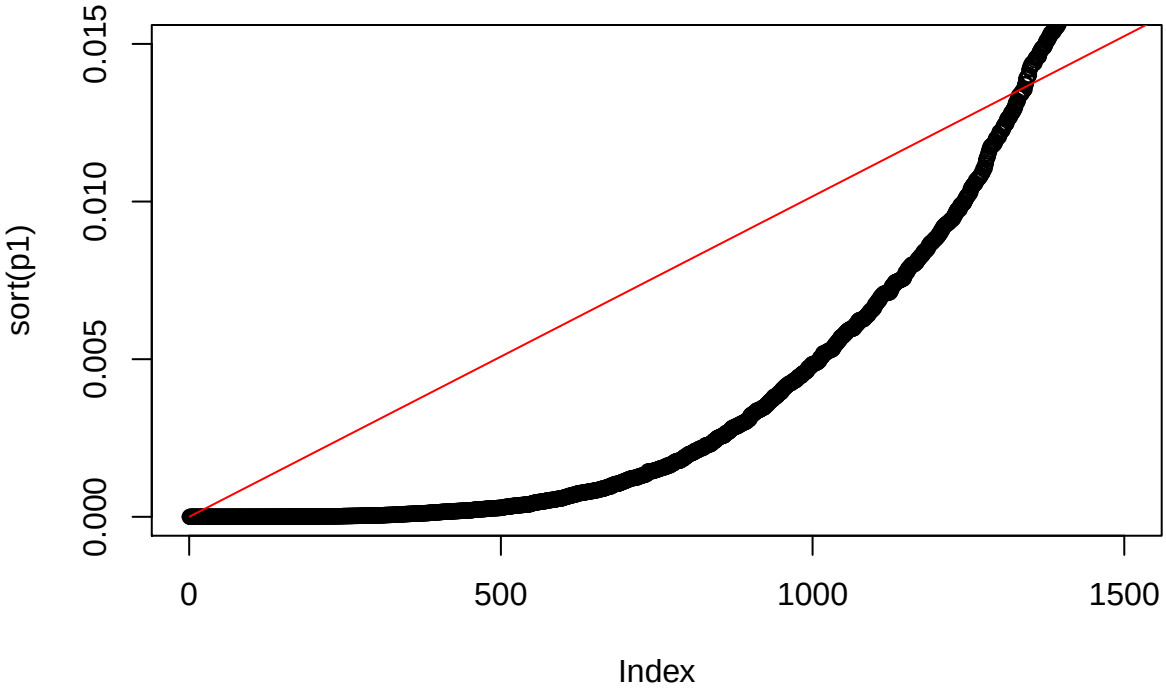
```
## [1] 4
```

FDR controlling methods

Benjamini and Hochberg

The critical values of the Benjamini & Hochberg procedure are much larger than those of Hochberg or Hommel, so that many more rejections can usually be made. In the Van de Vijver data, we reject 1340 hypotheses with p-values below 0.0136 at an FDR of 0.05, compared with 209 with p-values below 0.0000112 for Hommel's method.

```
plot(sort(p1), ylim=c(0,0.015), xlim=c(0,1500))
lines(1:4919, alpha*(1:4919)/4919, col=2) # BH critical values
```




```
sum(p.adjust(p1, "BH") <= alpha)  # BH: number of rejections
```

```
## [1] 1340
```

In the Rosenwald data, we reject 72 hypotheses with p-values below 0.000486 with Benjamini & Hochberg, compared with four with p-values below 0.00000676 for Hommel's. Clearly, without changing the assumptions, relaxing the criterion from FWER to FDR can make a huge difference in terms of power.

```
sum(p.adjust(p2, "BH") <= alpha)  # BH: number of rejections
```

```
## [1] 72
```

Adaptative procedures

The adaptative procedure of Benjamini, Krieger and Yekutieli, adjusts the α- level slightly from α to $\alpha^* = \alpha/(1+\alpha)$ to adjust for the additional variance from estimation of π_0. FDR control for the adaptive Benjamini, Krieger and Yekutieli procedure has only yet been proven under independence, although simulations suggest FDR control under positive dependence as well.

This procedure estimates π_0 by first performing an initial Benjamini & Hochberg procedure at the slightly reduced level α^*, estimating π_0 by $\hat{\pi}_0 = (m - R_0)/m$, where R_0 is the number of rejections obtained in this first step. In the second and final step, a subsequent Benjamini & Hochberg procedure is performed at level $\alpha^*/\hat{\pi}_0$.

Note that, unlike simpler plug-in procedures, this procedure is not guaranteed to give more rejections than the regular, nonadaptive Benjamini & Hochberg procedure, because $\alpha^*/\hat{\pi}_0$ may be smaller than α. This reflects the additional risk incurred in estimating π_0.

The adaptive procedure estimates $\hat{\pi}_0 = 0.73$ for the Van de Vijver data, resulting in 1468 rejections, compared with 1340 for the nonadaptive procedure.

```
alphastar = alpha/(1+alpha)  # alphastar
alphastar
```

```
## [1] 0.04761905
```

```
R0 = sum(p.adjust(p1, "BH") <= alphastar) # R0
R0
```

```
## [1] 1317
```

```
hatpi0 = (4919 - R0)/4919 # hatpi0
hatpi0
```

```
## [1] 0.7322627
```

```
sum(p.adjust(p1, "BH") <= alphastar/hatpi0) # number of rejections: adaptative
```

```
## [1] 1468
```

In the Rosenwald data, the same procedure finds a disappointing $\hat{\pi}_0 = 0.99$, so that the critical value for the second stage is increased rather than decreased. A number of 69 hypotheses are rejected, compared with 72 for the nonadaptive Benjamini & Hochberg procedure.

```
R0 = sum(p.adjust(p2, "BH") <= alphastar) # R0
hatpi0 = (7399 - R0)/7399 # hatpi0
sum(p.adjust(p2, "BH") <= alphastar/hatpi0) # number of rejections: adaptative
```

```
## [1] 69
```

Benjamini and Yekutieli

In the Rosenwald data, where Holm's method rejected four hypotheses and Benjamini & Hochberg rejected 72, the procedure of Benjamini & Yekutieli, allows no rejections at all. In this case, Holm's method may be superior in terms of power to Benjamini & Yekutieli, and a better choice for FDR control.

```r
sum(1/1:7399) # factor
```

```
## [1] 9.486383
```

```r
sum(p.adjust(p2, "BY") <= alpha) # number of rejections: BY
```

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

In the Van de Vijver data, where m is smaller and there are more false hypotheses, Benjamini & Yekutieli do reject substantially more than Holm, namely 614 hypotheses against 206 for Holm.

```r
sum(p.adjust(p1, "BY") <= alpha) # number of rejections: BY
```

```
## [1] 614
```

References

- Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine 2014 33:1946-78