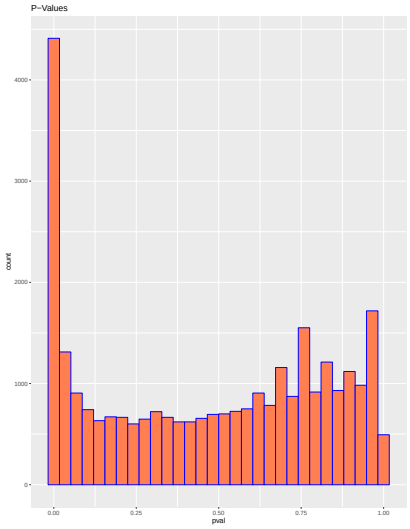
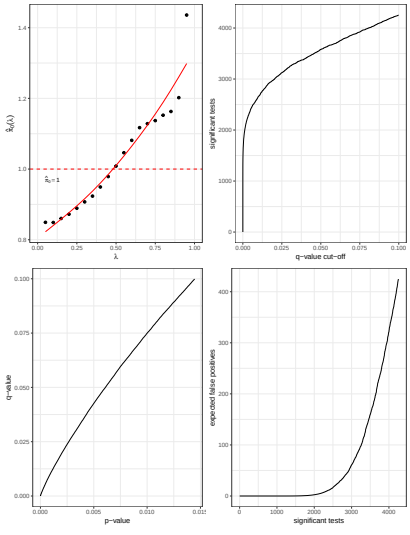
1 Exercise 1: Airways Power Diagnostics

In this problem we are looking to evaluate the power of the airway study - a study with 4 treated and 4 untreated individuals each with 64k gene counts. Here we use DESeq2 to clean the assays matrix. Moreover, we group genes by their cell lines to adjust for correlated genes. This will greatly increase the strength of our tests to find signal in each gene. After adjusting for this affect, we run the DESeq2 analysis that gives us the resulting p-value histogram Here we see that there are



several genes that appear to be explaining a non-negligible amount of variance in the presence of the disease in these 8 patients. From here we can use the p-values to estimate the local false discovery rate CDF. For this, we use the Bioconductor package q-value which does just this. The results of this analysis can be found in the figure below. The estimated local false discovery rate distribution



is given in the top right panel. Here we see that this study has good power. We see that for small values of q that the estimated number of significant tests is quite high. As this tolerance increases,

we also see that we have a steady increase in significant tests. In either case, even when controlling for local false discovery rates, we see that several of these tests are still significant. These findings suggest that this study has good power.

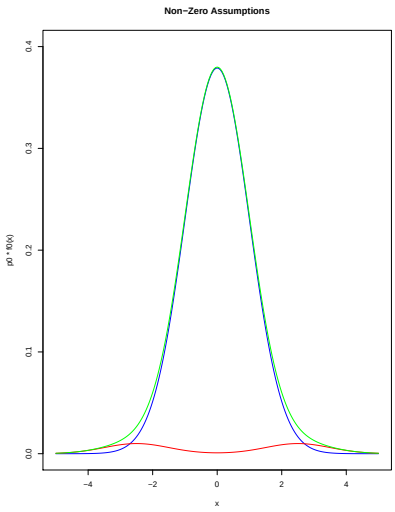
2 Exercise 2: Book Problems

2.1 6.2

Suppose that $\pi_0 = 0.95$ and $f_0 \sim N(0,1)$ and the non-null distribution of $f_1 \sim \frac{1}{2}N(-2.5,1) + \frac{1}{2}N(-2.5,1)$. This then gives the three way mixture

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

Here we see that $f(x)$ is unidentifiable. For this reason we take the *zero - assumption* on $\mathcal{A}_0 = [-1,1]$. Therefore, we wish to redefine the non-null and null distribution to follow this assumption.



$$\widetilde{\pi}_0 \widetilde{f}_0(x) = \begin{cases} \pi_0 f_0(x) + (1-\pi_0) f_1(x) & x \in \mathcal{A}_0 \\ \pi_0 \phi_0(x) & x \notin \mathcal{A}_0 \end{cases}$$
$$= \begin{cases} \pi_0 \left\{ f_0(x) + \frac{1-\pi_0}{\pi_0} f_1(x) \right\} & x \in \mathcal{A}_0 \\ \pi_0 \phi_0(x) & x \notin \mathcal{A}_0 \end{cases}$$

Hence we see that on $\mathcal{A}_0$ the adjusted null distribution is given by the mixture $\widetilde{f}_0(x) = f_0(x) + \frac{1-\pi_0}{\pi_0} f_1(x)$ with $\widetilde{\pi}_0 = \pi_0$

2.2 6.10

In our case our general two sample t statistic can be written as follows

$$t_i = \sqrt{n-2} \frac{(\overline{x}_A - \mu_A) - (\overline{x}_B - \mu_B)}{\sqrt{\sum_{j=1}^{n/2}(x_{ij} - \overline{x}_A)^2 + \sum_{j=n/2+1}^{n}(x_{ij} - \overline{x}_B)^2}}$$

Under the null distribution $\mu_A = \mu_B = 0$. Hence the above reduces to the following.

$$t_i = \sqrt{n-2} \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\sum_{j=1}^{n/2}(x_{ij} - \overline{x}_A)^2 + \sum_{j=n/2+1}^{n}(x_{ij} - \overline{x}_B)^2}}$$

Now, notice that for each x_{ij} we have that $\mathbb{E}(x_{ij}) = 0$ and $\text{Var}(x_{ij}) = 1 + \frac{\sigma_B^2}{4}$. If we now scale the denominator so that it has a χ^2 distribution we can find the distribution of t_i

$$t_i = \frac{\sqrt{n-2}}{\sqrt{1+\sigma_B^2/4}} \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\sum_{j=n/2+1}^{n}(\frac{x_{ij}-\overline{x}_A}{\sqrt{1+\sigma_B^2/4}})^2 + \sum_{j=n/2+1}^{n}(\frac{x_{ij}-\overline{x}_B}{\sqrt{1+\sigma_B^2/4}})^2}}$$

Now notice here that the denominator has a $\chi^2_{n/2-1} + \chi^2_{n/2-1} \overset{D}{=} \chi^2_{n-2}$ distribution. Hence we see that

$$t_i \overset{D}{=} \frac{\overline{x}_A - \overline{x}_b}{\sqrt{1+\sigma_B^2/4}} \frac{1}{\sqrt{\chi^2_{n-2}/n-2}}$$

Moreover, as $\overline{x}_A - \overline{x}_B \sim N(0, \frac{2}{n}\left(1+\sigma_B^2/4\right))$. With this we see that

$$t_i \overset{D}{=} \sqrt{\frac{2}{n}} \frac{Z}{\sqrt{\chi^2_{n-2}}/n-2} \sim \sqrt{\frac{2}{n}} \cdot t_{n-2}$$

To find an expression for the local false discovery rate, $fdr(t_i)$ with $\pi_0 = 1$ and $\sigma_B^2 = \frac{2}{\sqrt{n}}$ we first see that $t_i \sim \sqrt{2} \cdot t_{n-2}$. As we just argued, the new null distribution is over dispersed by a factor of $\sqrt{2}$ as compared to the normal theoretical null. That is, under the normal theoretical null we assume that $t_i \sim t_{n-2}$. In this model we see that the variance $\text{Var}(t_i) = \frac{n-2}{n-4} \to 1$ while in reality however, this variance we know by the previous discussion is actually closer to 2. Hence to correct for this, we see that $\frac{t_i}{\sqrt{2}}$ will have the proper large N variance. Therefore, if we were to calculate this local fdr, we could approximate it as follows

$$fdr(t_i) = \frac{\text{null}}{\text{theoretical null}} \approx \frac{\sqrt{2} f_{n-2}(t_i)}{f_{n-2}(t_i/\sqrt{2})}$$

Lastly, we can find $\mathbb{P}(fdr(t_i) \leq 0.20)$ using the previous result.

$$\begin{aligned}
\mathbb{P}(fdr(t_i) \leq 0.2) &= \mathbb{P}\left[f_{n-2}(t_i) \leq \frac{0.2}{\sqrt{2}} f_{n-2}(t_i/\sqrt{2}) \right] \\
&= \mathbb{P}\left[\left(1 + \frac{t_i^2}{18}\right)^{-19/2} \leq \frac{0.2}{\sqrt{2}} \left(1 + \frac{t_i^2}{36}\right)^{-19/2} \right] \\
&= \mathbb{P}\left[t_i^2 \geq \frac{(1/5\sqrt{2})^{-2/19} 1/2 * 36 - 18}{1 - (1/5\sqrt{2})^{-2/19} 1/2} \right]
\end{aligned}$$

Now recall that $t_i \sim t_{n-2}$. Hence $t_i^2 \sim F_{1,n-2}$. Therefore,

$$\mathbb{P}(fdr(t_i) \leq 0.2) = 1 - F_{1,n-2}(10.667) \approx 0.0043$$

Therefore we see that the probability of the local false discovery being less than 0.2 is considerably low.