8.6 Empirical Bayesian Statistics

Example 8.6.1 (Baseball example)**.**

Let X_i be the bating average in the first 45 at-bats. Then $45X_i \sim Bin(n, p_i)$. Apply the variance stabilization transformation,

$$Y_i = 2\sqrt{45}\arcsin\sqrt{X_i}.$$

Then it is known that $Y_i \sim N(\theta_i, 1)$ where $\theta_i = 2\sqrt{45}\arcsin\sqrt{p_i}$. Assume that $\theta_i \sim N(\mu, \tau^2)$ where μ, τ^2 are two hyper-parameters. We thus have the following Normal-Normal model

$$\begin{cases} Y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1) \\ \theta_i \overset{\text{iid}}{\sim} N(\mu, \tau^2). \end{cases}$$

Based on the Normal-Normal model, then the Bayes estimator for θ_i are

$$\hat{\theta}_i = MY_i + (1 - M)\mu.$$

How to choose the hyper-parameters?

8.6.1 James Stein Estimator

Assume that $Y_i \overset{\text{iid}}{\sim} N(\theta_i, \sigma^2)$. Consider the estimator

$$\hat{\theta}_i = (1 - \frac{(p-2)\sigma^2}{\sum Y_i^2})Y_i.$$

When $p > 2$, then

$$E\sum_i(\hat{\theta}_i - \theta_i)^2 \leq E\sum_i(X_i - \theta_i)^2.$$

8.6.2 Two-Group Model

The two-group model has been very widely used in the high dimensional data analysis. [Efron(2008), Efron(2010)], [He et al.(2013)He, Sarkar, and Zhao].

Example 8.6.2. Prostate cancer

Use θ_i, an indicator function, to denote whether a gene is differential expressed. If we had only data from gene i to consider, we should use t_i in the usual way to test the null hypothesis,

$$H_{0i} : \theta_i = 0.$$

One can transform the t-statistic t_i to a z-statistic z_i. Two groups model

$$\begin{cases} z_i|\theta_i \sim (1-\theta_i)f_0(z) + \theta_i f_1(z) \\ \theta_i \overset{\text{iid}}{\sim} Bernoulli(\pi_1) \end{cases}$$

Which gene is more important, or more likely to be rejected? From the Bayesian perspective, we should look at the Bayes posterior:

$$fdr_i(\boldsymbol{z}) = P(\theta_i = 0|\boldsymbol{z});$$

How to calculate the local fdr score?