Statistics 215B Assignment 5 Due: 5pm Wed 28 Feb 2024

James-Stein estimation and empirical Bayes

Math stats
Work the following exercises in Efron (2010): 1.1, 1.2, 1.4, 1.5.

Simulation
Produce your own version of Table 1.2 in Efron (2010) by repeating the simulation study described on pp. 7-9. Use the same μ_i's as Efron. Explain how many decimal places of agreement one would expect to see between your results and Efron's. How well did you meet this expectation?

Shrinking radon
The file `srrs2.dat` contains 12,777 observed radon levels from households throughout the United States. This data file comes from Andrew Gelman's website, `http://www.stat.columbia.edu/~gelman/arm/software/`. We will focus on the 766 measurements taken in the basements of the Minnesota homes. These homes are spread across 85 counties in Minnesota; the data set tells us which observations came from which counties.

- Load the data into R. Extract the subset of observations taken in Minnesota basements. Although there is a `basement` variable, you should instead use the `floor` variable—a zero value means a basement. (Don't ask.)
- Reduce the data set further: keep only the data for counties with at least 10 observations. You should find 17 such counties, with a total of 511 observations.
- Now split the data into two sets: a training set with five randomly chosen observations from each county, and a test set with the other observations.
- Compute μ, the vector of mean radon levels by county in the test data. Radon levels are given in the variable `activity`. From now on we will treat μ as a population-level parameter to be estimated.
- Make the standard James-Stein independent-normals assumption: the five observations in county i are iid draws from a $\mathcal{N}(\mu_i, \tau^2)$ distribution; these five draws are independent of the draws from every other county. Compute $\hat{\mu}^{(\text{MLE})}$, the maximum-likelihood estimate of μ based on the training data.
- Now compute $\hat{\mu}^{(\text{JS})}$, the James-Stein estimator, using the average value in $\hat{\mu}^{(\text{MLE})}$ as the shrinkage target. We are assuming that the components of $\hat{\mu}^{(\text{MLE})}$ share a common SE. Using the same number of observations in each county tends to aid this assumption. To estimate this shared SE, you must estimate τ^2, using the pooled-variance technique: add up all the within-county squared residuals, and divide by the total degrees of freedom.
  Caution: The SE of $\hat{\mu}_i^{(\text{MLE})}$ is not τ. If you proceed as though it is, you will over-shrink.
- What is the total squared error of $\hat{\mu}^{(\text{MLE})}$? Of $\hat{\mu}^{(\text{JS})}$? What is the ratio of the larger to the smaller? What do you conclude about Stein shrinkage in this application?