MT 4113: Computing in Statistics, Example simulation study

Here, I briefly describe and give code for a small simulation study (actually, it's the same study you worked on in Practical 3). I hope this example is useful to you in showing how a simulation study might be put together, as well as giving another example of some modular code.[1]

粗糙的 切边的

The simulation study looks at the performance of a robust estimator of the mean of a sample. This estimator is the *trimmed mean*, and it's more robust to outliers than the standard type of mean. If you've not come across it already, look it up in any stats book. (If you watched the last Olympics, for example, you'll already be familiar with it as it's used in many sports: discard the lowest and highest of the judge's marks and take the mean of those remaining.) We will generate data from a gnarly distribution, and see how the trimmed mean performs. Our measure of performance will be the mean squared error (MSE) – i.e., the average of the squared difference between estimated mean and true mean. The smaller the MSE, the closer, on average, the estimated mean is to the true mean. We will try different levels of trimming, from $\alpha=0$ (i.e., the mean), through 0.1 (i.e., 10% smallest and 10% largest values from the samples deleted) to a maximum of 0.5, which is the sample median.

The distribution we'll simulate from is a mixture distribtuion. We use a a mixture of a normal distribution and a non-central t distribution, i.e., the random variable Z is defined as follows:

$$Z = \eta X + (1-\eta)Y$$

where

$$\begin{aligned} \eta &\sim \text{Bernoulli}(\pi) \\ X &\sim \text{Normal}(\mu_1, \sigma_1) \\ Y &\sim \mu_2 + \text{t}(df) \end{aligned}$$

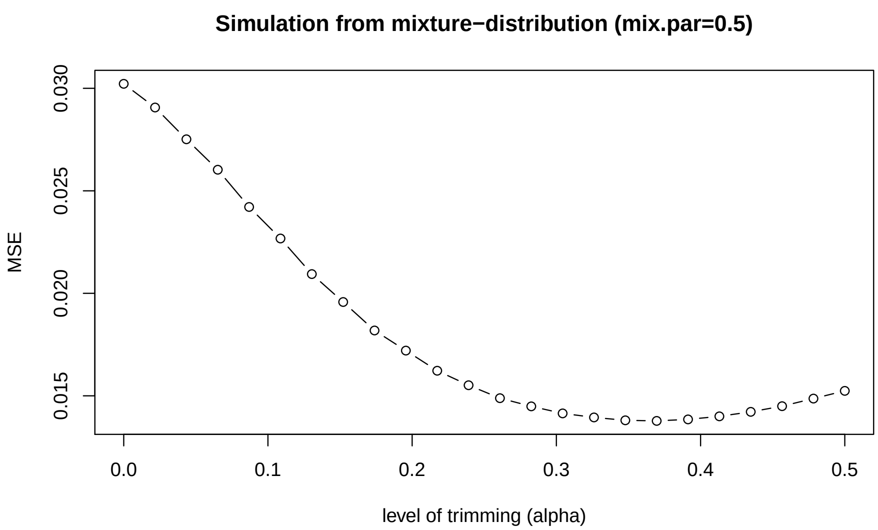
The parameter of interest is $\mu_Z = \text{E}[Z] = \pi\mu_1 + (1-\pi)\mu_2$.[2]

Different levels of trimming will be compared, e.g., α=0, 0.1, 0.2, and the mean square error (MSE) is calculated for each level. The MSE is based on M simulations:

$$MSE(\bar{z}_\alpha) = \frac{1}{M}\sum_{m=1}^{M}(\bar{z}_\alpha(m) - \mu_Z)^2.$$

The other parameters used were M=1,000, μ_1=5, σ_1=2, μ_2=7, df=3.

`R` Functions implementing the simulation are given in the file `ExampleSimulation.r` on the class web site. A file of "driver code", showing some example simulation scenarios, is also on the web site. Here are some example results, obtained using the driver code.



[1] Note that I did not attempt to turn this brief write-up into a "proper" statistical report - if you want advice on how to do that (the advice was written a few years ago for a different course), see https://goo.gl/kMYdDa.

[2] You can derive this result using the "double expectation" formula, i.e., $E[Z] = E_\eta E[Z|\eta]$.