Tutorial 5

ST2137-2420

Material

This tutorial covers the topics and concepts from chapter 5 of the course textbook: robust statistics. The first question is for you to grasp the value of robust statistics. It emphasises that we can use a computer to understand/test new methodologies out before diving deeper into them.

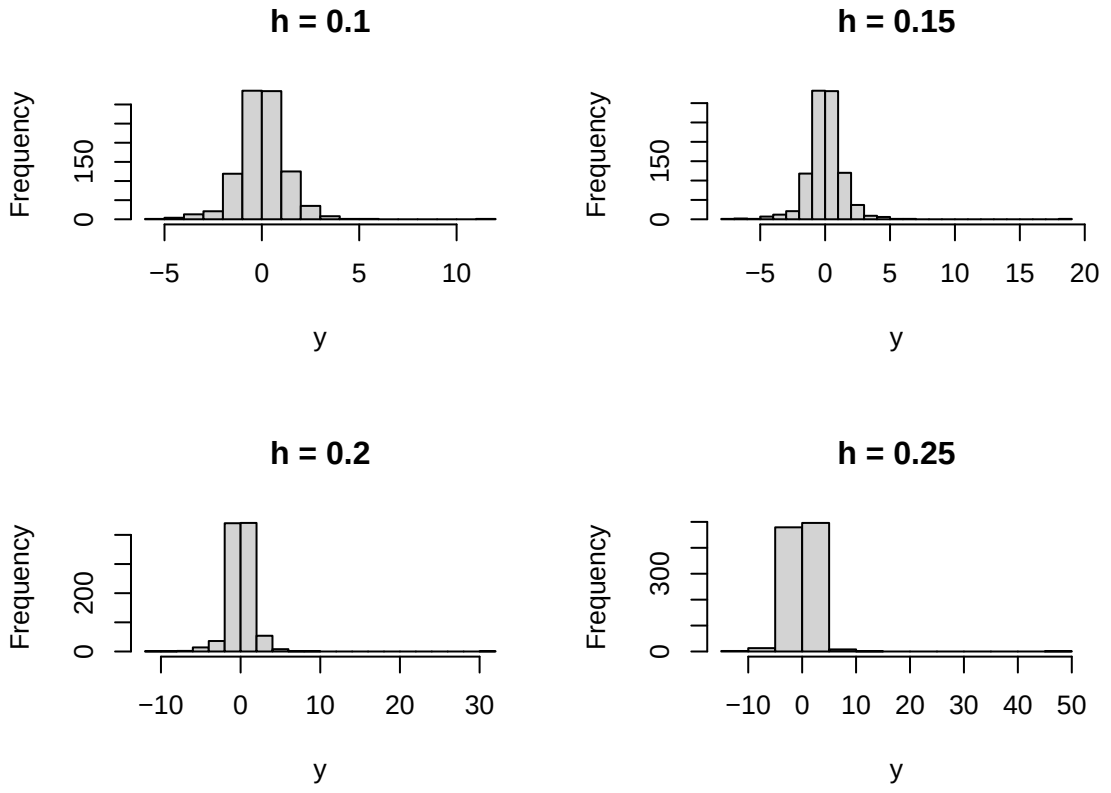
Remember to code up the answers in both R and Python.

h-Distributions

Let us define a family of distributions that will enable us to study the value of robust statistics, in terms of dealing with distributions with fat tails. If $Z \sim N(0,1)$, then for a fixed h, define the random variable Y:

$$Y = Ze^{hZ^2/2}$$

Here are some sample histograms of observations for various values of h:



As we can see, the value of h can be used to "control" the amount of elongation in the tails, leading to very large observations.

1. Perform the following simulation experiment:

A. Repeat 50 times:
   1. Generate 30 observations from $h = 0.3$.
   2. Compute the sample mean and the trimmed mean.

B. Compute the average and s.d. of the 50 sample means and trimmed means.

What is your view/opinion of trimmed mean vs. mean in this case?

Outlier Detection

The following three methods are sometimes used to detect outliers:

Rule based on means and variances:

The rule is to declare an outlier if

$$\frac{|X_i - \bar{X}|}{s} > K$$

Usually, K is taken to be 2.24. If X_i was truly Normal, 2.5% of observations would be classified as outliers, on average.

Rule based on IQR

The boxplot uses the following rule.

$$X_i < q_{0.25} - 1.5 \times IQR(X), \text{ or } X_i > q_{0.75} + 1.5 \times IQR(X)$$

where q_1 and q_3 are the sample quartiles.

Rule based on median and MAD(X):

X_i is declared an outlier if

$$\frac{|X_i - \text{median}(X)|}{MAD(X)/0.6745} > K$$

In this K is taken to be the square root of the 0.975 quantile of a χ^2 distribution. It is approximately 2.24 once again.

2. Use the above three methods to detect the outliers in the following dataset:

$$2, 2, 3, 3, 3, 4, 4, 4, 100000, 100000$$

Student Performance Dataset

3. Use the Winsorized and trimmed means with $\gamma = 0.1$ to estimate the location parameter for `G3`, grouped by `Medu`. Compare these estimates of location to the usual sample mean. Why are the estimates different? Which would you use?
4. Use the three techniques in the previous sections to identify any potential outliers in `G3`, grouped by `Medu`.