# STAT 5372 Assignment 2

(Lectures 4–8)

## Mathematical concepts and derivations

1. Prove the Theorem in Lecture 4 (Slide #4). That is, for a parametric model where $\theta = T(F)$ is defined implicitly by solving the score equation $\mathbb{E}U_\theta(X) = 0$, the influence function is:

$$L_\theta(x) = \frac{U_\theta(x)}{i(\theta)}, \qquad \text{where} \quad i(\theta) = -\int \dot{U}_\theta(y)dF(y), \quad \text{and} \quad \dot{U}_\theta(y) = \frac{\partial U_\theta(y)}{\partial\theta}.$$

   Do this in sequence as follows.

   (a) Make a slight change in notation by writing the score equation as:

$$0 = \mathbb{E}U(X;\theta) = \int U(y;T(F))dF(y).$$

   Now replace $F$ with $F_\epsilon = (1-\epsilon)F + \epsilon\delta_\epsilon$ to get:

$$0 = \int U(y;T(F_\epsilon))d((1-\epsilon)F + \epsilon\delta_x)(y).$$

   (b) Use the properties of the Dirac $\delta$-function to simplify this and show that it amounts to the equation $0 = \alpha(\epsilon) + \beta(\epsilon)$, where $\alpha(\epsilon)$ and $\beta(\epsilon)$ are functions that explicitly (and implicitly) depend on $\epsilon$.

   (c) Differentiate $\alpha(\epsilon)$ and $\beta(\epsilon)$ and evaluate them at $\epsilon = 0$ to obtain the final result, not forgetting that by definition:

$$L_\theta(x) = \left.\frac{\partial T(F_\epsilon))}{\partial\epsilon}\right|_{\epsilon=0}.$$

2. Consider bootstrap samples $\{x_i^*\}_{i=1}^n$ to be constructed from the $n$ observations $\{x_i\}_{i=1}^n$.

   (a) Argue that, if we care about the order of the $\{x_i^*\}_{i=1}^n$, then there are a total of $n^n$ possible bootstrap samples.

   (b) Now argue that, among the $n^n$ total bootstrap samples in (a), only $\binom{2n-1}{n}$ are actually *distinct*. (Note: *distinct* bootstrap samples are defined as those having different order statistics.)

3. Let $X$ and $Y$ be continuous random variables with cdfs $F_X(x)$ and $F_Y(y)$, and quantile functions $Q_X(\alpha) = F_X^{-1}(\alpha)$ and $Q_Y(\alpha) = F_Y^{-1}(\alpha)$. If $X$ and $Y$ are related through the equation $Y = g(X)$, where $g(x)$ is a monotone increasing function, then show that the quantile functions are related as follows:

$$Q_Y(\alpha) = g\left(Q_X(\alpha)\right).$$

# Simulation

4. Here we want to compare 3 nonparametric methods for constructing confidence intervals for the variance of a random variable: the functional delta method (Lecture 3), the bootstrap studentized interval (called bootstrap-$t$ in Lecture 8), and the BCa interval (Lecture 8). Specifically, conduct a simulation study to determine how the coverage probability and average interval width of these intervals varies with the sample size $n$. For each of the distributions below, produce a plot of coverage probability versus sample size, with lines representing the various methods, as well as a corresponding plot for interval width.

   (a) Carry out the simulation with data generated from the standard normal distribution.

   (b) Repeat using data generated from an exponential distribution with rate 1.

   (c) Submit the code for your simulation, and comment on the performance of these 3 methods in each of (a) and (b).

# Implementation

5. Write an R function which implements the jackknife. The function should accept two arguments: $x$ (the data) and `theta` (a function which, when applied to $x$, produces the estimate). The function should return a named list with the following components:

   **bias:** the jackknife estimate of bias.

   **se:** the jackknife estimate of standard error.

   **values:** the leave-one-out estimates $\{\hat{\theta}_{(i)}\}$.

   Submit the code for your function.

# Application

6. The course dataset `testscores.txt` contains data consisting of test scores of 88 students in 5 subjects: Mechanics, Vectors, Algebra, Analysis, and Statistics. One natural question about this data is the extent to which these tests measure separate skills vs. general tests of quantitative ability. One way to quantify this is via the ratio of the largest eigenvalue of the sample correlation matrix to the sum of the eigenvalues:

$$\hat{\theta} = \frac{\hat{\lambda}_1}{\sum_{i=1}^{5} \hat{\lambda}_i},$$

   where $\hat{\lambda}_5 \leq \cdots \leq \hat{\lambda}_1$ are the eigenvalues sorted from smallest to largest.

   (a) Use the bootstrap to estimate the standard error of $\hat{\theta}$.

   (b) Plot a histogram of your bootstrap replications $\{\hat{\theta}_b^*\}$. Does the sampling distribution appear to be normally distributed?

7. The standardized test used by law schools is called the Law School Admission Test (LSAT), and it has a reasonably high correlation with undergraduate GPA. The course dataset `lsat.txt` contains data on the average LSAT score and average undergraduate GPA for the 1973 incoming class of 15 law schools. (Note: the sample correlation coefficient is given by:

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\hat{\sigma}_x \hat{\sigma}_y},$$

where $\bar{x}$ and $\hat{\sigma}_x^2$ are the plug-ins for the mean and variance of $X$, and similarly for $Y$.)

   (a) Use your function from Question 5 to obtain a jackknife estimate of the bias and standard error of the sample correlation coefficient $\hat{\rho}$ between GPA and LSAT scores. Comment on whether this estimate of $\rho$ is biased upward or downward.

   (b) If $x$ and $y$ are drawn from a bivariate normal, where $X$ has mean $\mu_x$ and variance $\sigma_x^2$ (and similalrly for $Y$), then $n\mathbb{V}(\hat{\rho}) \xrightarrow{P} (1-\rho^2)^2$. Use this fact to estimate the standard error of $\hat{\rho}$.

   (c) On page 21 of our textbook, the author gives the influence function for the correlation coefficient $\rho$ ($\rho = \theta = T(F)$):

$$L(x, y) = \tilde{x}\tilde{y} - \frac{\theta}{2}(\tilde{x}^2 + \tilde{y}^2),$$

   where

$$\tilde{x} = \frac{x - \mu_x}{\sigma_x}, \qquad \text{and} \qquad \tilde{y} = \frac{y - \mu_y}{\sigma_y}.$$

   Use this to estimate the standard error of $\hat{\rho}$.

   (d) Use the bootstrap to estimate the standard error of $\hat{\rho}$.

   (e) Plot a histogram of your bootstrap replications $\{\hat{\rho}_b^*\}$. Does the sampling distribution appear to be normally distributed?

   (f) Compare the four estimates (a)–(d).

8. Consider the following dataset of 6 observations:

$$0.28,\ 0.98,\ 1.36,\ 1.38,\ 2.4,\ 7.42.$$

We wish to fit a Gamma($\alpha = 1.2, \beta$) model to these data, where $\alpha$ is the shape parameter, and $\beta$ is the scale. The purpose here is to estimate the entire sampling distribution of the MLE of $\beta$, when it is known that $\alpha = 1.2$. In this Question you must write your own code, i.e., **do not use a package!**

   (a) Estimate $\beta$ via maximum likelihood (MLE), call it $\hat{\beta}$, and state what its asymptotic distribution is. (Note: this can all be done analytically!)

   (b) Estimate the distribution $\hat{\beta}$ using the *parametric* bootstrap with $B = 10,000$ replicates. Submit your code.

   (c) Plot a histogram of the bootstrap distribution of $\hat{\beta}$ from (b), and superimpose on it the asymptotic distribution from (a).

   (d) Comment on the shapes of these two distributions. Are they close? Why, or why not?