STAT 5372 Assignment 1

Mathematical concepts and derivations

1. Show that out of all possible CDFs, the empirical CDF $\hat{F}$ maximizes the empirical likelihood:

$$L(F) = \prod_{i=1}^{n} \left[F(x_i) - F(x_i^-) \right].$$

   Note that w.l.o.g. we can take F to be discrete, since a continuous F would yield $F(x_i) - F(x_i^-) = 0$ for every x, and since $\widehat{F}(x) \geq 0$, this would imply $F(x) \leq \widehat{F}(x)$. To proceed with a discrete F, let $z_1 < z_2 < \cdots < z_m$ be the distinct values of $\{x_1, \ldots, x_n\}$, with corresponding multiplicities $\{n_1, \ldots, n_m\}$. Now, with $p_j = F(z_j) - F(z_j^-)$ and $\hat{p}_j = n_j/n$, note that we can write

$$\widehat{F}(x) = \sum_{j=1}^{m} \hat{p}_j I(z_j \leq x).$$

   Conclude the argument from here by showing that:

$$\log \left(\frac{L(F)}{L(\widehat{F})} \right) < 0.$$

   To get the inequality, use the fact that $\log(x) < x - 1$ for all $x > 0$ provided that $x \neq 1$.

2. Consider the quantile functional $\theta = T(F) = F^{-1}(p)$. Suppose that F is continuous at θ with positive density $f(\theta) > 0$. Show that the influence function for $T(F)$ is given by:

$$L(x) = \begin{cases} \frac{p-1}{f(\theta)}, & x \leq \theta \\ \frac{p}{f(\theta)}, & x > \theta. \end{cases}$$

   [Hint: note that $\theta_\epsilon = T(F_\epsilon) = F_\epsilon^{-1}(p)$, where $F_\epsilon(y) = (1-\epsilon)F(y) + \epsilon\delta_x(y)$, implies that $F_\epsilon(\theta_\epsilon) = p$. Now differentiate both sides of this last expression.]

3. With $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$ as before, define the following function w.r.t. a generic estimator $T(F)$:

$$b(\epsilon) = \sup_x |T(F) - T(F_\epsilon)|.$$

   From here, the *breakdown point* of $T(F)$ is defined to be $\epsilon^* = \inf\{\epsilon : b(\epsilon) = \infty\}$. Find the breakdown point of the sample mean.

4. Consider a positive random variable X, and suppose we are interested in the two functionals

$$\theta = \int \log(x) dF(x), \qquad \text{and} \qquad \lambda = \log(\mu),$$

where $\mu = \mathbb{E}(X)$.

   (a) What is the plug-in estimator of θ?
   (b) Derive the influence function and empirical influence function for θ.
   (c) What is the plug-in estimator of λ?
   (d) Derive the influence function and empirical influence function for λ.
   (e) Derive an asymptotic $1-\alpha$ nonparametric confidence interval for $\hat{\lambda}$.
   (f) Do $\hat{\theta}$ and $\hat{\lambda}$ converge to the same number? Justify.
   (g) Plot the empirical influence functions from parts (b) and (d). In each case, label the point x on the horizontal axis where $\hat{L}(x) = 0$.

5. Suppose there exists a constant C such that the following relation holds for all G:

$$|T(F) - T(G)| \leq C \sup_x |F(x) - G(x)|.$$

Show that $T(\hat{F}) \xrightarrow{a.s.} T(F)$.

Simulation

.6 Generate a random sample $X_1, \ldots, X_{100}$ and compute a 95% global confidence band for the CDF F based on the DKW inequality. Repeat this 1,000 times and report the proportion of data sets for which the confidence band contained the true distribution function.

   (a) Carry out the above simulation with data coming from the standard normal distribution.
   (b) Repeat using data generated from the standard Cauchy distribution.

7. Compare the nonparametric confidence interval for the variance obtained from using the functional delta method to the normal-theory interval:

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \right),$$

where s^2 is the usual unbiased estimator of the variance. Conduct a simulation study to determine the coverage probability and average interval width of these two intervals.

   (a) Carry out the above simulation with data generated from the standard normal distribution.
   (b) Repeat using data generated from an exponential distribution with rate 1.
   (c) Briefly, comment on the strengths and weaknesses of these two methods.

Application

8. The R data set `quakes` contains (among other information) the magnitude of 1,000 earthquakes that have occurred near the island of Fiji.

   (a) Estimate the CDF for the magnitude of earthquakes in this region, along with a 95% confidence interval. Plot your results.

   (b) Estimate and provide a 95% confidence interval for $F(4.9) - F(4.3)$.

   (c) Estimate the variance of the magnitude, and provide a nonparametric 95% confidence interval for its value.