STAT 830

Problems: Assignment 2

1. Consider the empirical distribution function $\hat{F}_n(x)$ for a sample $X_1, \ldots, X_n$ from a cdf F. In this problem I want you to do a Monte Carlo study to compare several confidence limits for $F(x)$:

   - The pointwise interval based on the approximately normal pivot

   $$\frac{\sqrt{n}(\hat{F}_n(x) - F(x))}{\sqrt{\hat{F}_n(x)[1 - \hat{F}_n(x)]}}$$

   - The pointwise interval based on the approximately normal pivot

   $$\frac{\sqrt{n}(\hat{F}_n(x) - F(x))}{\sqrt{F(x)[1 - F(x)]}}$$

   - The simultaneous interval based on the Dvoretsky-Kiefer-Wolfowitz inequality as described in my notes.
   - The simultaneous interval based on the assertion that

   $$\sup_x \{\sqrt{n}|\hat{F}_n(x) - F(x)|\} \xrightarrow{d} \sup_x \{|B_0(x)|\}$$

   where B_0 is a Brownian Bridge.

   (a) Begin by writing out the confidence intervals for the two pointwise methods as a function of α when the confidence level is $1 - \alpha$.

   (b) Find (on-line for instance) tables of the asymptotic critical points of the Kolmogorov-Smirnov test. The level α critical point, c_α has the property

   $$P(\sup_{0 \le t \le 1} \{|B_0(t)|\} > c_\alpha) = \alpha.$$

   Warning: Some tables record $c_\alpha/\sqrt{n}$ for a variety of small n and then give a formula for larger n from which you can find c_α. Use the tables to compare the Dvoretsky et al limits to the Brownian Bridge limits.

(c) Generate a sample of size 20 from the Uniform[0,1] distribution. Plot, on one graph, the 3 intervals above, excluding the Brownian Bridge method, along with the true cdf F for x running from 0 to 1.

(d) Generate 1000 samples of size 20 from the same distribution and for each x in $\{0.1, 0.2, \ldots, 0.9\}$ estimate the pointwise coverage probability for each procedure.

(e) For the same samples estimate the simultaneous coverage probability of all 3 intervals for the set of 9 x values in the previous problem. Please do the same for the 99 values $i/100$ for $i = 1, 2, \ldots, 99$. Attach standard errors to each of these estimates.

Then I want you to summarize in a paragraph the conclusions of the comparisons. In making the comparisons you need to know that

$$P\left(\sup_x\{|B_0(x)|\} \geq 1.358\right) = 0.05.$$

Your summary will take the form of a paragraph or two written in LaTeXin which you discuss the comparisons as if you were advising people on which procedure to use in which circumstances.

You might like to look at the `R` function `ecdf` in case you find it useful.

2. Suppose you have a sample of size n from the uniform distribution on $[0, \theta]$ where $\theta > 0$ is unknown.

(a) Find mean, standard error and mean squared error of two estimates θ: the largest observation, and twice the sample mean. Compare the two estimators on the basis of these answers. Suppose $n = 4$ and you observe 1,2,3, and 10. Which estimator is definitely better and why.

(b) Find the cumulative distribution function and the density of the largest observation by writing the event that the largest observation is less than or equal to x as an intersection of n independent events. (Be sure to be explicit about this step.)

(c) Use `setseed(1943)` to generate a sample of size 50 from this distribution with $\theta = 10$. Then generate 500 bootstrap samples.

For i from 1 to 500 let $\hat{\theta}_i^*$ be the maximum of the ith bootstrap sample. Plot a histogram of these 500 values using `prob=TRUE` and superimpose a graph of the true density of $\hat{\theta}$. Is the density well approximated by the histogram?

(d) Show that $P_\theta(\hat{\theta} = \theta) = P_\theta(\hat{\theta} = \theta(F)) = 0$ and compare this to

$$\lim_{n\to\infty} P(\hat{\theta}^* = \max X_1, \ldots, X_n | X_1, \ldots, X_n) = \lim_{n\to\infty} P(\hat{\theta}^* = \theta(\hat{F}_n)) | X_1, \ldots, X_n).$$

3. If $S(t)$ is a random variable for each $t \in [0,1]$ we call S a stochastic process (the term has much broader application than this). The Covariance function of S is

$$\rho_S(s,t) = \mathrm{Cov}(S(s), S(t)).$$

Find the covariance function for $\hat{F}_n(x)$. The answer will involve the true distribution function F.

4. If $a < b$ then $\theta = T(F) = F(b) - F(a)$ is a statistical functional. Find the standard error and plug-in estimated standard error of the plug-in estimate

$$\hat{\theta} = T(\hat{F}_n).$$

Use this to derive an approximate level $1-\alpha$ confidence interval for θ.

5. Suppose we observe a sample $X_1, \ldots, X_n$ and that the n numbers are all different. When you draw a bootstrap sample from these data you get a certain number of copies of X_1, of X_2, and so on. Two bootstrap samples are only really different if they get different numbers of copies of some data point. Show that there are

$$\binom{2n-1}{n}$$

different bootstrap samples. Use this idea to compute the exact distribution of the "pivots", $\bar{X}^* - \bar{X}$ and $\sqrt{n}(\bar{X}^* - \bar{X})/s^*$ for the particular data set

x = c(0.9575,0.4950,0.1080,0.9359,0.6326).

Each of these two quantities has 126 possible values and you should plot the probabilities of each of the 126 values of the statistic against the corresponding value. (Some values involve division by 0 and may be omitted from the graphs.) The distribution I have in mind is discrete; it is a conditional distribution given the original data set – the 5 numbers above.

6. Suppose we have a sample $X_1, \ldots, X_n$ and a bootstrap sample $X_1^*, \ldots, X_n^*$. Find the bootstrap expected value and standard deviation of the bootstrap mean. That is find

$$\mathrm{E}(\bar{X}^* | X_1, \ldots, X_n)$$

and

$$\sqrt{\mathrm{Var}(\bar{X}^* | X_1, \ldots, X_n)}.$$

Due date: 1 October 2020. Please email me a pdf file with your answers whose file name starts with your family name, then given name, like LockhartRichard. It helps me sort them in my folders.