Stat 624: Final

Complete all problems in a folder in the git repository labeled "Final". In the end you will need to create 4 files: (1) a python file named GetData.py, (2) a data file named FinalData.csv, (3) a C++ final named ks.cpp, (4) and finally an R script named ForFInal.R. No formal write-up is required.

Background

The Kolmogorov Smirnov Test-Statistic is a tool that can be used to test distributional assumptions for 1 dimensional data. The idea is to compare the theoretical cumulative distribution function of a distribution with the empirical distribution function. Suppose your data has n ordered elements, $y_1, ..., y_n$ where $y_i < y_{i+1}$ for all i, and the theoretical cumulative distribution function of the data is F. The formula for the Kolmogorov Smirnov test statistic is

$$KS(F, y) = \underset{i=1,...n}{max} \left(\left| F(y_i) - \frac{i-1}{n} \right|, \left| F(y_i) - \frac{i}{n} \right| \right)$$

In words, this means you find the absolute value of $F(y_i) - \frac{i-1}{n}$ and the absolute value of $F(y_i) - \frac{i}{n}$ for every i. The Kolmogorov Smirnov test statistic is the maximum of all of $2n$ of those.

If you are testing different distributions on the same data set, it is likely that the distribution with the lowest KS statistic will be a better fit for the data.

(a) The data to be used for this final is on the website `richardson.byu.edu/624/Final2019Data` in an html table. Use python webscraping to create a data file inside the Final directory of this data. Use any method you would like for the webscraping but make sure the final data is what you think it is.

If you are unable to complete this problem correctly, you will still need data. Inside the Final directory is the data set BackupData.csv. This can be used to complete the rest of the problems is needed.

(b) Each column in the Final data frame was generated from one of the following distributions. These will serve as the theoretical CDF's for the Kolmogorov Smirnov Test Statistic.

Label	Distribution
G	Frechet with $s = 3$ and $\alpha = 3$
H	Weibull with $\lambda = 4$ and $k = 2$
K	Gumbel with $\mu = 3$ and $\beta = 2$

- The Frechet distribution has a CDF of

$$G(x|s,\alpha) = e^{-(x/s)^{-\alpha}},$$

see `https://en.wikipedia.org/wiki/Frechet_distribution`

- The Weibull distribution has a CDF of

$$H(x|\lambda,k) = 1 - e^{-(x/\lambda)^k},$$

see `https://en.wikipedia.org/wiki/Weibull_distribution`

- The Gumbel distribution has CDF of

$$K(x|\mu,\beta) = e^{-e^{-(x-\mu)/\beta}},$$

see `https://en.wikipedia.org/wiki/Gumbel_distribution`

Write a C++ function that can be imported into R that takes as an argument a data vector and returns the Kolmogorov test statistic for distributions G, H, and K. The function will need to order the data vector it is given.

(c) For each column in the data set, determine which distribution it would be assigned using the Kolmogorov test statistic as comparison (lower is best). Create a vector of labels, G, H, and K, and determine how many there are assigned to each distribution. Work can be done in R, but use the function you created in (b) and the data you collected in (a).

(d) Random values can be simulated from the Weibull distribution using the inverse CDF trick according to

$$y_i = \lambda(-\ln(1-U))^{1/k},$$

where U is a random uniform variable. Create 1,000 data sets of size 150 each with the parameters $\lambda = 4$ and $k = 2$. Apply the function created in part (b) to each one. How often is the Weibull correctly identified as the correct distribution?

(e) Use just the first column of the data set for problems (e) through (g). Determine the optimal parameters s and α of the Frechet distribution that minimizes the Kolmogorov-Smirnov test statistic. You can use any software for this and the remainder of the problems. Both s and α are restricted to be positive.

(f) The probability density function of the Gumbel distribution is

$$k(x|\mu,\beta) = \frac{1}{\beta}e^{-(z+e^{-z})},$$

where $z = \frac{x-\mu}{\beta}$. The mean of the Gumbel distribution with parameters μ and β is approximately $\mu + 0.5772\beta$. Find the maximum likelihood estimate of μ and β in the Gumbel distribution using the first column of the data frame while restricting the mean of the distribution to be equal to 4. β is restricted to be positive and μ has no domain restrtictions.

(g) Repeat part (e) with the change that the mean must be greater than or equal to 4, not strictly equal to it.