CSDS 313: Introduction to Data Analysis Fall 2020

Assignment 2: Data and Distributions

Instructor: Mehmet Koyutürk, TA: Thomas Varley, Sean Maxwell Due: October 9

Problem 1

The purpose of this exercise is to investigate how different distributions can have similar statistics and/or visualizations. Suppose you are given a normal distribution $\mathcal{N}(\mu, \sigma)$. We would like to estimate a uniform distribution $U(a, b)$ (i.e., the range of the distribution is $[a, b]$) with identical statistics to the given normal distribution. These statistics are specified as follows:

(i) Find the parameters (a and b) of a uniform distribution in terms of μ and σ such that the mean and standard deviation of uniform distribution is the same as the given normal distribution.

(ii) Find the parameters (a and b) of a uniform distribution in terms of μ and σ such that the 25th and 75th percentile points of the uniform distribution and the given normal distribution are the same. Assume you can compute inverse cumulative distribution function $\Phi^{-1}(p, \mu, \sigma)$ of a normal distribution $\mathcal{N}(\mu, \sigma)$ for any $0 \leq p \leq 1$. See probit function for more information. *Hint:* You should estimate the parameters of uniform distribution a and b by simply using $\Phi^{-1}(p, \mu, \sigma)$.

For parts (i) and (ii) separately, obtain a uniform distribution $U(a, b)$ as a function of μ and σ i.e., find $a = f_a(\mu, \sigma)$ and $b = f_b(\mu, \sigma)$. Then, estimate the parameters of uniform distributions $U_1(a_1, b_1)$ and $U_2(a_2, b_2)$ corresponding to parts (i) and (ii) for the normal distribution $\mathcal{N}(\mu = 2, \sigma = 5)$. Simulate 10 000 data points from each of the $U_1(a_1, b_1)$, $U_2(a_2, b_2)$ and $\mathcal{N}(2, 5)$ distributions separately. Visualize the 3 simulated distributions using histograms, error bars, and boxplots. Compare and comment on how the obtained uniform distributions are similar or unsimilar to the given normal distribution. Also, compare and comment on how they are similar or unsimilar to each other.

Note that, you can compute the probit function $\Phi^{-1}(p, \mu, \sigma)$ as follows:

- MATLAB: `norminv` function.
- Python: `norm.ppf` function in `scipy.stats` package.
- R: `qnorm` function.

Problem 2

For this exercise, we will use two datasets that are provided with the assignment:

- The file "airport_routes.csv" contains the number of available routes of 3409 airports all around the world (as of February 2017). Each row indicates an airport (identified with a 3-letter code) and the number of routes. For example, "`CLE, 81`" indicates that Cleveland Hopkins International Airport has outgoing flights to 81 different airports. See data source for more information.
- The file "movie_votes.csv" contains the average rating (between 1 and 10) of 4392 movies in TMDb database sorted in descending order. Each row contains a movie name and the average TMDb vote of that movie. For example, `"The Godfather", 8.4`, `"Interstellar",8.1` etc. See data source for more information.

For each of these datasets, consider the following models:

(a) Suppose the given data points follow a power law distribution. Estimate the corresponding α parameter. You can use the maximum likelihood estimation in Newman's notes on power-law.

(b) Suppose the given data points follow an exponential distribution.
Estimate the corresponding λ parameter.

(c) Suppose the given data points follow a uniform distribution.
Estimate the corresponding range parameters $[a, b]$ of the uniform distribution.

(d) Suppose the given data points follow a normal distribution.
Estimate the corresponding μ and σ parameters.

For each these dataset separately, compare the models you estimated in parts (a) to (d). Which distribution do you think the data follows and why? Explain. For each model, generate random data samples drawn from the respective distribution. Use visualizations of the empirical data and the data you generate to support your conclusions.

Problem 3

Recall the rocket problem from exercise 3: You are working as chief data scientist at a rocket production company. You know that your company's competitor is assigning integer IDs to their rockets. In other words, if the competitor produced M rockets, there is a rocket with ID i for all $1 \leq i \leq M$. Your company's intelligence wasable to collect the IDs of n rockets produced by the competitor and these IDs are $1 \leq x_1 \leq x_2 \leq ... \leq x_n$. You can assume that the IDs collected by the intelligence represent a uniform sampling of the M IDs.

(i) What is the maximum liklihood estimator for M. Simulate the rockets and intelligence reports to show if the maximum liklihood estimator is an unbiased estimator. (hint: make sure to choose a large M and and large number of trials for your simulation)

(ii) Let $\hat{M}_{MVU} = x_n(\frac{n+1}{n}) - 1$. Let $\hat{M}_{MEAN} = 2(\sum_{i=1}^{n} x_i/n) - 1$ Simulate the rockets and intelligence reports to show which of the above unbiased estimators ($\hat{M}_{MVU}$ or $\hat{M}_{MEAN}$) has the lower variance.