San José State University
Department of Applied Data Science

DATA 220
Mathematical Methods for Data Analysis

Spring 2021
Instructor: Ron Mak

Assignment #6

Assigned: Thursday, March 4, 2021
Due: Thursday, March 11 at 5:30 pm
Team assignment, 100 points max

The Central Limit Theorem
In this assignment, you will conduct experiments on the Central Limit Theorem of statistics.

Random sampling
Choose a dataset that has a fairly large number (several hundred) values that you can use. Calculate three population parameters (e.g., mean, median, standard deviation, Q1, interquartile range, etc.). Perform random sampling and use the samples to estimate those population parameters. For each distribution:

- Experiment with different numbers of samples and different sample sizes.
- Create Matplotlib or Seaborn charts to illustrate your results.

Report
Incorporate a short report in your notebook's markdowns:

- Describe which distributions and parameters you used for this assignment.
- What were the results? What inferences were you able to make? Did you notice anything unusual that needed further exploration or research to explain?
- Some possible topics to discuss in your report based on the results of your experiments (feel free to discuss other topics):
    - What is a good compromise between the number of samples and the sample size? In the real world (such as professional polling services), it is expensive to do sampling. But you still want the most accurate estimates.
    - Which population probability distributions are more challenging to estimate their parameters with random sampling? Given what the Central Limit Theorem says about the distributions of the sample statistics, would a normally distributed population be easiest? Would other distributions be harder to make good parameter estimates? If so, why?

- From the charts in the March 4th set of lecture notes, it appears that estimating the standard deviation parameter is more challenging than estimating the population mean and median. Is that generally true for all population distributions? If so, why?
- For a fixed sample count, what sample sizes are required for the sample distributions to approach normal for each parameter estimation? Is it possible to rank the parameters by the required sample sizes when estimating using random sampling?

Create one or more Jupyter notebooks that contain your experiments.

What to submit to Canvas
Submit your Jupyter notebook(s) into Canvas:
Assignment #6.

Rubric
Your program will be graded according to these criteria:

Criteria	Max points
Population (dataset values) • Random samples generated to estimate three population parameters. • Graphs of the sample statistics.	**20** • 10 • 10
Random sampling • Estimates of three population parameters of the dataset. • Graphs of different sample counts and sample sizes.	**40** • 20 • 20
Report (incorporated in the notebooks) • Description of the experiments (dataset, samples, etc.) • Discussions of experimental results.	**40** • 15 • 25