PS1

September 2, 2025

Note: Write your code in the code cells, and your responses in markdown. Run the entire script and display the outputs of your code.

Due: **11:59PM Central Time on Friday, 09/12**. Upload both your code (.ipynb) and responses (html or pdf) to Canvas by then.

Prep

Please make sure you have installed Python (3.10+) and jupyter notebook before you start the assignment. - e.g., pip install jupyter notebook - Type "jupyter notebook" or "python -m jupyter notebook" in command prompt (Terminal on Mac, Powershell on Windows): this should generate a link for you to open in the browser and create a Jupyter notebook. - Alternatively, you can also use Jupyter in IDE like Visual Studio, Cursor (with AI chats), or Google Colab.

Jupyter/Colab allow you to run the entire script and display the outputs (including figures or tables). Make sure to upload both the original notebook (.ipynb) and the output in HTML/PDF format: - HTML: HTML first. Use "Export" in Jupyter/Colab or in bash:(py -m) *jupyter nbconvert PS1.ipynb –to html* to save a HTML. You may print the HTML as PDF. - PDF: Export the notebook as PDF directly via *jupyter nbconvert PS1.ipynb –to pdf*. This requires you have installed LaTex, Pandoc, and maybe nb_pdf_template to make the PDF look like the original notebook.

I recommend writing the code on your own, or at least a pseudocode before you ask for help from AI assistants. Please complete the Disclaimer on any use of AI at the end of the problem set.

[ ]:
```
# Packages you might need: (pip install ... if you don't have them)
import pandas as pd
import numpy as np
import matplotlib
```

1 Refresher

Consider a random variable X with (population) mean μ and variance σ^2. $(x_1, x_2, ..., x_n)$ is a random sample of X from the population, in which the data points are independent and identically distributed (iid).

1. Prove the following: (a) $E[(x_i - \mu)x_i] = \sigma^2$, and (b) $E[x_i^2] - E[x_i]^2 = \sigma^2$.
2. Define the sample variance as $s_n = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$, where $\bar{x} = \frac{1}{n}\sum_i x_i$. Show s_n is an unbiased estimator for population variance, $E[s_n] = \sigma^2$.

3. Suppose we run a regression of x_i on a constant 1, $x_i = \beta * 1 + \epsilon_i$. What is β in the population regression, and why? What is the OLS coefficient $\hat{\beta}$ in the sample regression?

2 Simulation - Bernoulli

Set up an Python program to conduct a simulation of drawing a sample size of n from a Bernoulli distribution with mean p. In each "replication" r you will draw a sample, construct the estimated mean from that replication (which we will denote as $\overline{Y}_r$) , and calculate the 95% confidence interval $(\overline{Y}_r - 1.96 * s_r/\sqrt{n}, \overline{Y}_r - 1.96 * s_r/\sqrt{n})$ where s_r is the estimated standard deviation in that replication, $s_r = \sqrt{\overline{Y}_r * (1 - \overline{Y}_r)}$. In each replication, record the length of the confidence interval, and whether or not the true mean is inside the interval.

For given (n, p), conduct $1,000$ replications are report the following statistics:

1. the mean estimate of p;
   - plot the distribution of $\overline{Y}_r$ (sample estimates of p) across replications.
2. the mean estimate of the true standard deviation.
3. the fraction of time that the confidence interval contains the true p. This is called the *coverage rate*.

Conduct the analysis for the cases $n = 30$ using $p = 0.05$ and $p = 0.25$, and again for $n = 60$ using $p = 0.05$ and $p = 0.25$ (a total of 4 cases). It is often claimed that n of 30 or larger is enough to insure that asymptotic confidence intervals work well. Do you agree or not?

[ ]:
```
# Write your code here
```

3 Data analysis: Age Profile and Gender Gap in the Use of Cell Phone

Download the dataset "october_cps", which is an extract of data from the October 2012 CPS.

There are several variables describing each person in the survey, including age, sex, "educ" (which is education, coded in a certain way) as well as variables about how someone uses their cell phone (if they have one): cell_use_phone, cell_use_msg, cell_use_video, cell_use_browse_web, cell_use_email, cell_use_games, cell_use_social_media, cell_use_download_apps, and cell_use_music. Each of these is coded "1" if the person uses the cell phone for that use, "2" if not, and (as a phone; for text messages, for video, to browse the web, for email for games, to access social media, to download apps, or to listen to music).

1. Develop a graph to show how people of different ages use their cell phone. Be creative.
2. In your sample, test whether males (sex=1) and females (sex=2) use their cell phone at the same rate for each of the 9 different uses.

[ ]:
```
# Write your code here
```