Gov 2001: Problem Set 3

Spring 2025

March 31, 2025

Problem Set Instructions:

- This problem set is due on **April 8th, 11:59 pm** Eastern time. Please upload a PDF of your solutions to **Gradescope**. Please match your answers with the questions.
- We will accept hand-written solutions but we strongly advise graduate students to typeset your answers in LaTeX.
- Citing your sources is always a good practice in academia. Please list the names of other students / sources / AI you obtained help from on this problem set.

1 WLLN (20pt)

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with finite expectation $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \text{Var}(X_i) < \infty$. Define the sample mean as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(a) Compute $\mathbb{E}[\overline{X}_n]$ and $\text{Var}(\overline{X}_n)$. (5pt)

(b) Use Chebyshev's inequality to bound $\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon)$ for any $\varepsilon > 0$. (5pt)

(c) Show that $\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$. (5pt)

(d) Conclude that $\overline{X}_n \xrightarrow{P} \mu$ as $n \to \infty$, and state this as the Weak Law of Large Numbers. (5pt)

2 CLT and Uncertainty (50pt)

In this problem, we will explore the implications of the Central Limit Theorem for uncertainty estimation and hypothesis testing. Start by creating two variables, $X1$ and $X2$, using the following code:

```
set.seed(02138)
X1 <- rnorm(100000, 5, 2)
X2 <- rexp(100000, 0.2)
```

For the purposes of this problem, we will treat these variables (each with 100,000 elements) as the full population. We will take samples from these two datasets to evaluate the coverage probability of 95% confidence intervals for the population mean using different types of data and different sample sizes.

1. Plot and describe the full distributions of `X1` and `X2`. What is the population mean and population standard deviation of each random variable (this is just the `mean()` and `sd()` of both of these variables)? (10pt)

   Hint for coding: Consider using the `plot(density())` function for density plots to describe the distributions.

2. Now, create a loop to take 100 samples of size 8 from each dataset and record the sample mean for each sample. Plot the density of sample means for `X1` and `X2` separately and compare these densities. Are they similar or different? Why? (10pt)

   Hint for coding:

   - One way to work on this problem is to use a for loop. If you want to loop over $1, 2, \ldots, n$, then your loop starts with `for(i in 1:n) {…}`.
   - First create an empty matrix (or data frame) to save results. Then save outputs from the for loop to that matrix.

3. Given your answer in part (1), what should be the standard error of the sample mean of `X1` and `X2`? We know that if a random variable is normally distributed, 95% of its distribution will be within 1.96 standard deviations of the mean. What proportion of the sample means for `X1` and `X2` are within 2 standard errors of the population means? (10pt)

4. Repeat the simulation in parts (2) and (3) for samples of size 8, 20, 50, and 500, and increase your number of simulations to 1000. Report the probability of being within 1.96 standard errors of the population mean for each of your eight simulations in a table (you do not need to create additional plots). How do your results change? What differences do you see between `X1` and `X2`? (10pt)

   Hint for coding: One way to work on this problem is to loop over different sample sizes `for(n in c(8,20,50,500))`. Of course, you are free to use other functionalities if you wish.

5. Interpret your findings in parts (2)–(4). How does the Central Limit Theorem explain what you see in the simulations? (10pt)

3 Delta Method (30pt)

The Delta method is a very powerful tool for analyzing asymptotic properties of random variables. Let $X_1, \dots, X_n$ be i.i.d. (continuous random variables) with CDF $F_X(x)$. Consider the random variable

$$Y_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad \text{where } Z_i = I\{X_i \le x\}.$$

Here $I\{\cdot\}$ is the usual indicator function, so $Z_i = 1$ if $X_i \le x$ and 0 otherwise.

(a) What distribution does Z_i follow? Name the distribution. Find its mean and variance in terms of $F_X(x)$. (**Hint:** Fundamental bridge.) (5pt)

(b) What is the asymptotic distribution of $Y_n(x)$? (10pt)

(c) Apply the Delta method to identify the asymptotic distribution of $F_X^{-1}\big(Y_n(x)\big)$. You may use the identity

$$\frac{d}{dx}F_X^{-1}(x) \;=\; \frac{1}{f_X\big(F_X^{-1}(x)\big)},$$

where f_X is the PDF of X_i (**Hint:** Let the quantile function $F_X^{-1}(.)$ be $g(.)$). (10pt)

(d) Let $q_{X,p} = F_X^{-1}(p)$ be the pth quantile of the distribution of X. Given the results from (c), show that the asymptotic distribution of the pth sample quantile, denoted $Q_{X,p} = F_X^{-1}\big(Y_n(x)\big)$ when $Y_n(x)$ is near p, can be written as

$$\sqrt{n}\left(Q_{X,p} - q_{X,p}\right) \;\overset{d}{\to}\; \mathcal{N}\Big(0,\; \frac{p(1-p)}{\big[f_X(q_{X,p})\big]^2}\Big).$$

(5pt)