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**.
- 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

Solution:

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

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

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

Since expectation is linear:

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\mu}{n} = \mu.$$

Since the X_i are i.i.d., their variances add:

$$\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(b) **Use Chebyshev's inequality to bound $\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon)$.**

Chebyshev's inequality states:

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\mathrm{Var}(Y)}{\varepsilon^2}.$$

Apply it to $\overline{X}_n$:

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

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

From part (b), we have:

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Since the right-hand side tends to 0 as $n \to \infty$, it follows that:

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

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

By definition of convergence in probability:

$$\overline{X}_n \xrightarrow{P} \mu \quad \text{as } n \to \infty,$$

because for any $\varepsilon > 0$, the probability $\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \to 0$.

Weak Law of Large Numbers: Let $\{X_i\}$ be i.i.d. random variables with finite mean μ and finite variance σ^2. Then the sample mean $\overline{X}_n$ converges in probability to μ as $n \to \infty$:

$$\overline{X}_n \xrightarrow{P} \mu.$$

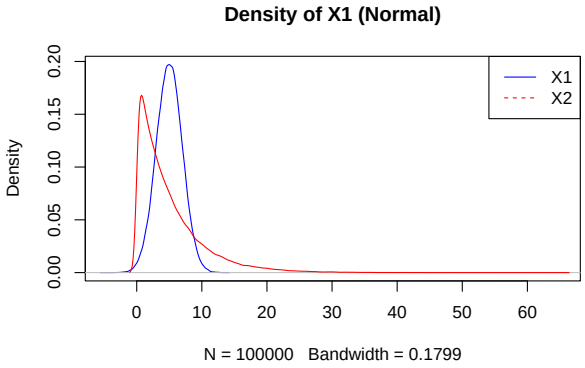
2 WLLN and Uncertainty

1. We plot and describe the distributions:

```
plot(density(X1), main="Density of X1 (Normal)",
col="blue",xlim=c(min(X1,X2),max(X1,X2)))
lines(density(X2), main="Density of X2 (Exponential)", col="red")
```

Findings:

- $X1$ is approximately normally distributed with mean close to 5 and standard deviation close to 2.
- $X2$ is right-skewed (exponential distribution) with mean close to 5 and standard deviation around 5.

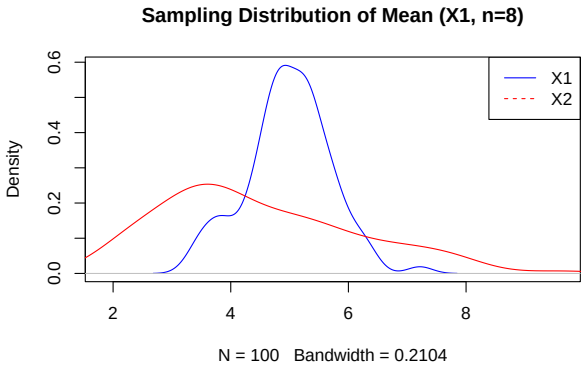


2. We take 100 samples of size 8 from each population and compute their sample means.

```
set.seed(02138)
n <- 8
reps <- 100
X1_means <- numeric(reps)
X2_means <- numeric(reps)

for (i in 1:reps) {
  X1_means[i] <- mean(sample(X1, n))
  X2_means[i] <- mean(sample(X2, n))
}

plot(density(X1_means),
main="Sampling Distribution of Mean (X1, n=8)", col="blue")
plot(density(X2_means),
main="Sampling Distribution of Mean (X2, n=8)", col="red")
```



Findings:

- The sampling distribution of $X1$ means is fairly normal.

- The sampling distribution of $X2$ means is more spread out and slightly skewed, though less than the original distribution.

3. We calculate the standard error and see what proportion of sample means fall within 2 SEs of the population mean.

```
set.seed(02138)
n <- 8
reps <- 100
coverage_X1 <- 0
coverage_X2 <- 0
true_mean_X1 <- mean(X1)
true_mean_X2 <- mean(X2)

for (i in 1:reps) {
  x1_samp <- sample(X1, n)
  x2_samp <- sample(X2, n)

  x1_mean <- mean(x1_samp)
  x1_se <- sd(x1_samp) / sqrt(n)
  x1_ci <- c(x1_mean - 1.96 * x1_se, x1_mean + 1.96 * x1_se)

  x2_mean <- mean(x2_samp)
  x2_se <- sd(x2_samp) / sqrt(n)
  x2_ci <- c(x2_mean - 1.96 * x2_se, x2_mean + 1.96 * x2_se)

  if (true_mean_X1 >= x1_ci[1] && true_mean_X1 <= x1_ci[2]) {
    coverage_X1 <- coverage_X1 + 1
  }
  if (true_mean_X2 >= x2_ci[1] && true_mean_X2 <= x2_ci[2]) {
    coverage_X2 <- coverage_X2 + 1
  }
}

prop_X1 <- coverage_X1 / reps
prop_X2 <- coverage_X2 / reps
prop_X1
prop_X2
```

Findings:

- Proportion for $X1$ is 91%, this is close to 0.95 due to normality.
- For `X2` (exponentially distributed), the coverage is noticeably lower — around 78% — due to:
  - High skew in the distribution.

- More variability in the sample standard deviation.
- The Central Limit Theorem not yet fully kicking in for small n.

4. We repeat the process for sample sizes 8, 20, 50, and 500.

```
set.seed(02138)
X1 <- rnorm(100000, 5, 2)
X2 <- rexp(100000, 0.2)
sample_sizes <- c(8, 20, 50, 500)
reps <- 1000
results <- data.frame(SampleSize = integer(),
Var = character(), Coverage = numeric())

true_mean_X1 <- mean(X1)
true_mean_X2 <- mean(X2)

for (n in sample_sizes) {
  coverage_X1 <- 0
  coverage_X2 <- 0

  for (i in 1:reps) {
    x1_samp <- sample(X1, n)
    x2_samp <- sample(X2, n)

    x1_mean <- mean(x1_samp)
    x1_se <- sd(x1_samp) / sqrt(n)
    x1_ci_low <- x1_mean - 1.96 * x1_se
    x1_ci_high <- x1_mean + 1.96 * x1_se
    if (true_mean_X1 >= x1_ci_low && true_mean_X1 <= x1_ci_high) {
      coverage_X1 <- coverage_X1 + 1
    }

    x2_mean <- mean(x2_samp)
    x2_se <- sd(x2_samp) / sqrt(n)
    x2_ci_low <- x2_mean - 1.96 * x2_se
    x2_ci_high <- x2_mean + 1.96 * x2_se
    if (true_mean_X2 >= x2_ci_low && true_mean_X2 <= x2_ci_high) {
      coverage_X2 <- coverage_X2 + 1
    }
  }

  results <- rbind(results,
                   data.frame(SampleSize = n, Var = "X1",
                   Coverage = coverage_X1 / reps),
                   data.frame(SampleSize = n, Var = "X2",
```

```
                    Coverage = coverage_X2 / reps))
}
print(results)
stargazer::stargazer(results,type="latex",summary=FALSE)
```

Results Table:

Table 1:

	SampleSize	Var	Coverage
1	8	X1	0.910
2	8	X2	0.864
3	20	X1	0.943
4	20	X2	0.909
5	50	X1	0.948
6	50	X2	0.927
7	500	X1	0.941
8	500	X2	0.942

2.1 Interpretation

The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean will approximate normality as sample size increases, regardless of the underlying distribution. Our findings support this:

- For `X1`, the sampling distribution is approximately normal even for small n due to the underlying normality.
- For `X2`, the distribution of sample means is not normal when n is small (hence low coverage), but becomes normal as n increases.
- As sample size increases, coverage probabilities for both variables converge to the nominal 95% level.

This shows the CLT in action and highlights the importance of large sample sizes, especially when working with skewed data.

3 Delta Method

The Delta method is a very powerful tool for analyzing asymptotic properties of random variables. Let $X_1, \ldots, X_n$ 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 $X_i \leq x$ 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.)

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

(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 $g(.)$ be the quantile function $F_X^{-1}(.)$).

(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

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

Solution

(a) Each $Z_i = I\{X_i \leq x\}$ is a Bernoulli random variable with parameter $p = F_X(x)$. Thus,

$$\mathbb{E}[Z_i] = F_X(x), \quad \mathrm{Var}(Z_i) = F_X(x)\big(1 - F_X(x)\big).$$

(b) Since $Y_n(x) = \frac{1}{n}\sum_{i=1}^n Z_i$, by the Central Limit Theorem,

$$\sqrt{n}\,\big(Y_n(x) - F_X(x)\big) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\; F_X(x)\big(1 - F_X(x)\big)\Big).$$

(c) We want the asymptotic distribution of $F_X^{-1}\big(Y_n(x)\big)$. Set $g(u) = F_X^{-1}(u)$. Then $g'(u) = \frac{1}{f_X\left(F_X^{-1}(u)\right)}$. By the Delta Method:

$$\sqrt{n}\big(g(Y_n(x)) - g(F_X(x))\big) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\; \big[g'(F_X(x))\big]^2 F_X(x)\big(1 - F_X(x)\big)\Big).$$

Substituting $g'(F_X(x)) = \frac{1}{f_X(x)}$ and noting that $g(F_X(x)) = x$, we get

$$\sqrt{n}\big(F_X^{-1}(Y_n(x)) - x\big) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\; \frac{F_X(x)\,\big(1 - F_X(x)\big)}{[f_X(x)]^2}\Big).$$

(d) For the pth quantile, let $x = q_{X,p} = F_X^{-1}(p)$. Then $Y_n(q_{X,p})$ is approximately p, and

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

This is the classic asymptotic result for sample quantiles.