Gov 2001: Problem Set 3

Sampling, Estimation, and Inference

Spring 2026

Due: Friday, March 13, 2026, 11:59 PM Eastern

Submit: PDF to Canvas (we recommend R Markdown or Quarto)

Total: 100 points

Instructions:

- Include all R code and output for simulation problems.
- You may collaborate with classmates, but write your own solutions and list collaborators.
- **Do not use AI assistants (ChatGPT, Claude, Copilot, etc.) on this problem set.** Work with each other instead. The struggle is where learning happens.
- Remember: 70% of your grade comes from in-class exams. Use problem sets to *learn*, not just to get answers.

Topics: Sampling distributions, LLN, CLT, estimation, bias, consistency, confidence intervals, hypothesis testing

Readings: Aronow & Miller §3.1–3.3; Blackwell Ch. 2–4

Question 1: The Central Limit Theorem in Action (25 points)

The Central Limit Theorem (CLT) tells us that the sampling distribution of the mean approaches normality as n grows, regardless of the underlying distribution. Let's see this in action.

Consider a **highly skewed** population: waiting times at a government office follow an exponential distribution with rate $\lambda = 0.1$ (so the mean waiting time is 10 minutes, variance is 100).

(a) (5 points) If you draw a random sample of n people, what does the CLT predict about the distribution of $\bar{X}$? Write down the approximate distribution, including the mean and variance, as a function of n and the population parameters.

(b) (10 points) **R Simulation:** Verify the CLT visually.

```r
set.seed(2001)
n_sims <- 10000

# Population: Exponential with rate 0.1
# Mean = 1/0.1 = 10, Variance = 1/0.1^2 = 100

# For each sample size, simulate 10,000 sample means
sample_sizes <- c(5, 30, 100, 500)

# Your code should:
# 1. For each sample size n:
#    - Draw 10,000 samples of size n from Exp(0.1)
#    - Calculate the sample mean for each
# 2. Create a 2x2 panel of histograms showing
#    the distribution of sample means
# 3. Overlay the predicted normal distribution from CLT
# 4. Comment on how quickly normality emerges
```

(c) (5 points) At what sample size does the sampling distribution look "approximately normal" to your eye? Would your answer change if the population were even more skewed (e.g., a Pareto distribution)?

(d) (5 points) A colleague says: "The CLT means my estimator will be unbiased if my sample is large enough." Is this statement correct? Explain the difference between what the CLT tells us and what unbiasedness means.

Question 2: Bias, Variance, and MSE (25 points)

This question explores the bias-variance tradeoff using a concrete example.

Setup

A researcher wants to estimate the population variance σ^2 from a sample $X_1, \ldots, X_n$. Two estimators are proposed:

Estimator 1 ("divide by n"):

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Estimator 2 ("divide by $n-1$"):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Assume $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ with $\sigma^2 = 25$.

(a) (5 points) It can be shown that $\mathbb{E}[\hat{\sigma}_n^2] = \frac{n-1}{n}\sigma^2$. Using this fact, calculate the bias of $\hat{\sigma}_n^2$. Is it biased upward or downward? Does the bias go to zero as $n \to \infty$?

(b) (5 points) Show that S^2 is unbiased for σ^2. (You can use the fact from part (a).)

(c) (5 points) Even though S^2 is unbiased, might there be situations where $\hat{\sigma}_n^2$ is *preferable*? Explain what would need to be true for the biased estimator to have lower Mean Squared Error (MSE).

Hint: Recall that MSE = Bias^2 + Variance.

(d) (10 points) **R Simulation:** Compare the two estimators.

```r
set.seed(2001)
n_sims <- 10000
true_var <- 25  # Population variance

# Try different sample sizes
sample_sizes <- c(5, 10, 30, 100)

# For each sample size:
# 1. Draw 10,000 samples from N(0, 25)
# 2. Calculate both estimators for each sample
# 3. Compute:
#    - Mean of each estimator (compare to 25)
#    - Bias of each estimator
#    - Variance of each estimator
#    - MSE of each estimator

# Create a table showing Bias, Variance, MSE for
# each estimator at each sample size.
# Which estimator has lower MSE at small n?
# Does this change as n grows?
```

Question 3: Confidence Intervals (25 points)

A political scientist surveys $n = 400$ registered voters and finds that 52% support a ballot initiative. She wants to construct a 95% confidence interval for the true population proportion p.

(a) (5 points) Using the standard formula for a confidence interval for a proportion:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

calculate the 95% confidence interval. Show your work.

(b) (5 points) A newspaper headline reads: "Poll shows 52% support initiative; there's a 95% chance the true support is between 47% and 57%." What's wrong with this interpretation? Provide the correct interpretation of the confidence interval.

(c) (5 points) The researcher is asked: "How large would the sample need to be to get a margin of error of ±2% instead of ±5%?" Answer this question. What's the general relationship between sample size and margin of error?

(d) (10 points) **R Simulation:** Verify the "95%" in "95% confidence interval."

```
set.seed(2001)
n_sims <- 10000
n <- 400
true_p <- 0.52  # Pretend we know the truth

# For each simulation:
# 1. Draw a sample of n voters where each supports
#    the initiative with probability true_p
# 2. Calculate p_hat and the 95% CI
# 3. Check whether the CI contains true_p

# Calculate: What proportion of the 10,000 CIs
# actually contain the true parameter?
# Is it close to 95%?

# Repeat with n = 50. Does coverage change?
# Why might small samples have different coverage?
```

Question 4: Hypothesis Testing (25 points)

A nonprofit claims that their get-out-the-vote (GOTV) program increases turnout by at least 5 percentage points. A researcher conducts a randomized experiment:

- Control group ($n_c = 500$): 42% turned out to vote
- Treatment group ($n_t = 500$): 45% turned out to vote

The researcher wants to test whether the program's effect is statistically distinguishable from zero.

(a) (5 points) Set up the null and alternative hypotheses. Calculate the test statistic for the difference in proportions:

$$z = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_t} + \frac{1}{n_c}\right)}}$$

where $\hat{p}$ is the pooled proportion under H_0.

(b) (5 points) Calculate the p-value for a two-sided test. At $\alpha = 0.05$, do you reject the null hypothesis? What do you conclude?

(c) (5 points) The nonprofit is disappointed: "But our program increased turnout by 3 percentage points! How can that not be significant?" Explain to them, in plain language, what "not statistically significant" means and does *not* mean.

(d) (5 points) Calculate the statistical power of this test to detect a true effect of 5 percentage points.

Hint: Power is the probability of rejecting H_0 when the alternative is true. Under the alternative ($p_t - p_c = 0.05$), calculate where the critical value falls in the sampling distribution of $\hat{p}_t - \hat{p}_c$.

(e) (5 points) **R Simulation:** Verify your power calculation.

```
set.seed(2001)
n_sims <- 10000
n_c <- 500
n_t <- 500
p_c <- 0.42
true_effect <- 0.05  # True effect under alternative
p_t <- p_c + true_effect

# For each simulation:
# 1. Draw control outcomes (n_c from Bernoulli(0.42))
# 2. Draw treatment outcomes (n_t from Bernoulli(0.47))
# 3. Conduct the hypothesis test
# 4. Record whether you reject H_0

# Calculate: What proportion of simulations rejected H_0?
# This is your simulated power. Compare to part (d).

# Also simulate under H_0 (true effect = 0)
# What proportion reject? (Should be close to 0.05)
```

Submission Checklist

Before submitting, verify:

- ☐ All analytical work shows clear steps
- ☐ All R code runs without errors
- ☐ Simulation results are compared to analytical answers
- ☐ Collaborators are listed (if any)

This problem set covers material from Weeks 5–7: sampling distributions, LLN, CLT, estimation, confidence intervals, and hypothesis testing.