Gov 2001: Problem Set 2

Random Variables, Expectation, and Variance

Spring 2026

Due: Friday, February 27, 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: Random variables, PMFs, expected value, variance, covariance, conditional expectation

Readings: Aronow & Miller §1.2, §2.1–2.2; Blackwell Ch. 1–2

Question 1: Expected Value and Linearity (20 points)

A political scientist studies campaign contributions. Let X be the contribution amount (in dollars) from a randomly selected donor, with the following PMF:

x	25	50	100	250	500
$f(x) = \mathbb{P}(X = x)$	0.40	0.30	0.15	0.10	0.05

(a) (4 points) Verify this is a valid PMF. Calculate $\mathbb{E}[X]$, the expected contribution.

(b) (4 points) The campaign pays a 3% processing fee on each contribution, plus a flat \$2 fee. The net amount received is $Y = 0.97X - 2$. Using the linearity of expectation, calculate $\mathbb{E}[Y]$.

(c) (4 points) Calculate $\mathbb{E}[X^2]$. Then use this to compute $\text{Var}(X)$ using the formula $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

(d) (8 points) **R Simulation:** Verify your calculations.

```
# Simulate 100,000 donors from this distribution
set.seed(2001)
n <- 100000

# Contribution amounts and probabilities
amounts <- c(25, 50, 100, 250, 500)
probs <- c(0.40, 0.30, 0.15, 0.10, 0.05)

# Your code should:
# 1. Sample n contributions from this distribution
# 2. Calculate mean(X) and compare to E[X]
# 3. Calculate mean(0.97*X - 2) and compare to E[Y]
# 4. Calculate var(X) and compare to Var(X)
#    Note: R's var() uses n-1 denominator; for population
#    variance, use mean((X - mean(X))^2)
```

Report your simulated values and confirm they approximately match your analytical answers.

Question 2: Variance of Sums and Dependence (25 points)

This question explores when $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$ holds—and when it doesn't.

Part A: The Formula (10 points)

(a) (5 points) Starting from the definition $\text{Var}(X+Y) = \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2]$, derive the general formula:

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

Show each step.

(b) (5 points) Under what condition does $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$? Prove that if X and Y are independent, this condition holds.

Hint: Use the fact that for independent X, Y: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Part B: A Counterexample (15 points)

Consider a simple example where X and Y are *not* independent.

Let X take values $\{-1, 0, 1\}$ with equal probability (each with probability $1/3$), and let $Y = X^2$.

(c) (3 points) Calculate $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(d) (4 points) Calculate $\text{Var}(X)$ and $\text{Var}(Y)$.

(e) (4 points) Calculate $\text{Cov}(X,Y)$. Are X and Y uncorrelated?

(f) (4 points) **R Simulation:** Verify your calculations.

```
set.seed(2001)
n <- 100000

# Sample X uniformly from {-1, 0, 1}
X <- sample(c(-1, 0, 1), n, replace = TRUE)
Y <- X^2

# Calculate and report:
# mean(X), mean(Y)
# var(X) using population formula, var(Y)
# cov(X, Y) -- does it equal zero?

# Also verify: is Var(X + Y) = Var(X) + Var(Y)?
```

Important: Even though $\text{Cov}(X, Y) = 0$, are X and Y independent? Explain why or why not in one sentence.

Question 3: Covariance and Correlation (25 points)

A researcher collects data on 500 voters, measuring their age (A, in years) and political knowledge score (K, on a 0–100 scale). The data show:

- $\bar{A} = 45$, $s_A = 15$ (mean and standard deviation of age)
- $\bar{K} = 62$, $s_K = 18$ (mean and standard deviation of knowledge)
- $\text{Cov}(A, K) = 81$

(a) (4 points) Calculate the correlation $\rho(A, K) = \text{Cov}(A, K)/(s_A \cdot s_K)$. Interpret this value in one sentence.

(b) (6 points) A research assistant proposes creating a "civic engagement index" defined as:

$$E = 2K - 50$$

This rescales knowledge to a 0–100 scale centered differently.

Calculate $\text{Cov}(A, E)$ and $\text{Corr}(A, E)$. How does the correlation change when you rescale K?

(c) (5 points) Another research assistant wants to measure age in months instead of years. Let $A_m = 12A$. Calculate $\text{Cov}(A_m, K)$ and $\text{Corr}(A_m, K)$. Explain why correlation is "unit-free."

(d) (10 points) **R Simulation:** Generate synthetic data to verify your understanding.

```
set.seed(2001)
n <- 500

# Generate correlated data with approximately the
# specified means, SDs, and correlation
# Use the mvrnorm function from MASS package
```

```r
library(MASS)

# Target: mean_A = 45, sd_A = 15, mean_K = 62, sd_K = 18
# Cov(A,K) = 81, so Corr = 81/(15*18) = 0.30

mu <- c(45, 62)
# Covariance matrix: [[var_A, cov], [cov, var_K]]
Sigma <- matrix(c(15^2, 81, 81, 18^2), nrow = 2)

data <- mvrnorm(n, mu, Sigma)
A <- data[, 1]
K <- data[, 2]

# Your code should:
# 1. Verify mean(A), sd(A), mean(K), sd(K), cov(A,K), cor(A,K)
# 2. Create E = 2*K - 50 and verify cov(A, E), cor(A, E)
# 3. Create A_m = 12*A and verify cov(A_m, K), cor(A_m, K)
```

Do your simulation results match your analytical predictions from parts (b) and (c)?

Question 4: Conditional Expectation and the CEF (30 points)

This question builds intuition for conditional expectation and the law of iterated expectations.

Setup

A survey asks voters about their party identification (P) and support for a policy (S, on a 1–10 scale). The joint distribution is:

	$P = D$	$P = I$	$P = R$	Marginal
$\mathbb{P}(P)$	0.35	0.30	0.35	1.00
$\mathbb{E}[S \mid P]$	7.2	5.0	3.1	—
$\text{Var}(S \mid P)$	2.5	4.0	2.0	—

That is: 35% are Democrats with average policy support 7.2; 30% are Independents with average support 5.0; 35% are Republicans with average support 3.1.

Part A: Law of Iterated Expectations (12 points)

(a) (6 points) Using the Law of Iterated Expectations, calculate $\mathbb{E}[S]$:

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S \mid P]] = \sum_p \mathbb{E}[S \mid P = p] \cdot \mathbb{P}(P = p)$$

Show your calculation.

(b) (6 points) Using the Law of Total Variance, calculate $\text{Var}(S)$:

$$\text{Var}(S) = \mathbb{E}[\text{Var}(S \mid P)] + \text{Var}(\mathbb{E}[S \mid P])$$

The first term is the average "within-group" variance; the second is the "between-group" variance. Calculate each term and interpret what they measure.

Part B: The CEF as Best Predictor (8 points)

(c) (4 points) Suppose you want to predict a voter's policy support S using only their party P. The CEF says: predict $\mathbb{E}[S \mid P = p]$ for each party.

What would you predict for:

- A Democrat?
- An Independent?
- A Republican?

(d) (4 points) Alternatively, suppose you ignore party and just predict $\mathbb{E}[S]$ for everyone. Using the numbers from (a), explain why the CEF-based prediction (using party) is better than the constant prediction (ignoring party).

Hint: Think about mean squared error. The MSE of the constant prediction is $\text{Var}(S)$. The MSE of the CEF prediction is $\mathbb{E}[\text{Var}(S \mid P)]$.

Part C: Simulation (10 points)

(e) (10 points) **R Simulation:** Generate data consistent with this setup and verify your calculations.

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

# Step 1: Generate party affiliation
party <- sample(c("D", "I", "R"), n, replace = TRUE,
                prob = c(0.35, 0.30, 0.35))

# Step 2: Generate policy support conditional on party
# For each party, draw from N(mean, var) then clip to [1,10]
S <- numeric(n)
S[party == "D"] <- rnorm(sum(party == "D"), mean = 7.2, sd = sqrt(2.5))
S[party == "I"] <- rnorm(sum(party == "I"), mean = 5.0, sd = sqrt(4.0))
S[party == "R"] <- rnorm(sum(party == "R"), mean = 3.1, sd = sqrt(2.0))

# Your code should:
# 1. Calculate mean(S) and compare to E[S] from part (a)
# 2. Calculate var(S) and compare to Var(S) from part (b)
# 3. Calculate E[S|P] for each party (group means)
# 4. Calculate the "within" and "between" variance components

# Bonus: Calculate MSE for constant vs. CEF prediction
# MSE_constant = mean((S - mean(S))^2)
# MSE_cef = mean((S - group_mean_for_each_obs)^2)
```

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 3–4: random variables, expectation, variance, covariance, and the conditional expectation function.