Gov 2001: Problem Set 1

Probability Foundations

Spring 2026

Due: Friday, February 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: Conditional probability, Bayes' Rule, independence, Law of Total Probability

Readings: Aronow & Miller §1.1; Blackwell Ch. 2.1

Question 1: Conditional Probability and Polling (25 points)

A polling firm surveys 1,200 likely voters in a swing state before the 2024 election. They record party registration and candidate preference:

	Harris	Trump	Undecided	**Total**
Democrat	336	18	30	384
Republican	24	372	36	432
Independent	114	168	102	384
Total	474	558	168	1,200

(a) (4 points) Calculate $\mathbb{P}(\text{Trump} \mid \text{Republican})$. Interpret this probability in one sentence.

(b) (4 points) Calculate $\mathbb{P}(\text{Republican} \mid \text{Trump})$. Interpret this probability in one sentence.

(c) (5 points) A cable news pundit says: "86% of Trump voters are Republicans, so if you meet a Republican, they're almost certainly a Trump voter." Evaluate this reasoning using your answers from (a) and (b). What error is the pundit making?

(d) (12 points) **R Simulation:** Write R code to verify your calculations.

```
# Create a data frame with all 1,200 voters
# Each row is one voter with party and preference

# Your code should:
# 1. Create the population matching the table above
# 2. Calculate P(Trump | Republican) from the data
# 3. Calculate P(Republican | Trump) from the data
# 4. Verify these match your analytical answers
```

Report your simulated proportions and confirm they match (a) and (b).

Question 2: Independence vs. Mutually Exclusive (20 points)

A fellow PhD student makes the following claim during a study session:

> "If two events are mutually exclusive, they must be independent. After all, if A and B can't both happen, then whether A happened tells you nothing about B—you already know B didn't happen."

(a) (6 points) Is this claim correct? If not, identify the specific error in the reasoning.

(b) (6 points) Prove mathematically that if A and B are mutually exclusive events with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then A and B **cannot** be independent.

(c) (8 points) **R Simulation:** Create a simulation to demonstrate the difference.

```
# Simulation 1: Mutually exclusive events
# - Roll a fair die
# - Event A: roll is 1, 2, or 3
# - Event B: roll is 4, 5, or 6
# Simulate 10,000 rolls and compute:
#   P(A), P(B), P(A and B), P(A)*P(B)
# Are A and B independent?

# Simulation 2: Independent events
# - Flip two fair coins
# - Event A: first coin is heads
# - Event B: second coin is heads
# Simulate 10,000 flips and compute:
#   P(A), P(B), P(A and B), P(A)*P(B)
# Are A and B independent?
```

Explain what your simulations demonstrate about the relationship between mutual exclusivity and independence.

Question 3: Bayes' Rule and Medical Testing (30 points)

A rapid diagnostic test for a rare disease has the following characteristics:

- **Sensitivity** (true positive rate): $\mathbb{P}(+ \mid \text{Disease}) = 0.95$
- **Specificity** (true negative rate): $\mathbb{P}(- \mid \text{No Disease}) = 0.98$
- **Prevalence**: $\mathbb{P}(\text{Disease}) = 0.002$ (2 in 1,000 people have the disease)

(a) (6 points) Using Bayes' Rule, calculate $\mathbb{P}(\text{Disease} \mid +)$, the probability that a person who tests positive actually has the disease. Show your work step by step.

(b) (4 points) Your answer to (a) is probably much lower than most people expect. Explain intuitively, in 2–3 sentences, why a positive test doesn't mean you probably have the disease.

(c) (10 points) **R Simulation:** Verify your calculation.

```
# Simulate a population of 100,000 people
# - 0.2% have the disease
# - Apply the test (sensitivity = 0.95, specificity = 0.98)
# Calculate:
#   1. Number of true positives
#   2. Number of false positives
#   3. Among all positive tests, what proportion are true positives?

set.seed(2001)  # For reproducibility
n <- 100000

# Your code here
```

Does your simulation result match your analytical answer from (a)?

(d) (6 points) Now suppose a second, independent test is administered to everyone who tested positive on the first test. This second test has the same sensitivity (0.95) and specificity (0.98).

Calculate $\mathbb{P}(\text{Disease} \mid \text{both tests positive})$. How does this compare to your answer in (a)?

(e) (4 points) A policy question: Given your answers above, should this test be used for mass screening of the general population? What if it were used only for people already showing symptoms (where prevalence might be 10% instead of 0.2%)? Briefly explain your reasoning.

Question 4: Law of Total Probability (25 points)

A political scientist studies voter turnout in a state with three types of counties:

County Type	Share of Registered Voters	Turnout Rate
Urban	50%	58%
Suburban	30%	72%
Rural	20%	64%

(a) (5 points) Using the Law of Total Probability, calculate the overall turnout rate for the state.

(b) (6 points) A voter is selected at random from those who voted. What is the probability they are from a suburban county? (Use Bayes' Rule.)

(c) (10 points) **R Simulation:** Verify your calculations.

```r
# Create a population of 10,000 registered voters
# - 50% urban, 30% suburban, 20% rural
# - Each voter turns out according to their county's rate

set.seed(2001)
n <- 10000

# Your code should:
# 1. Assign each voter to a county type
# 2. Simulate whether each voter turns out
# 3. Calculate overall turnout rate
# 4. Among those who voted, calculate proportion from suburban

# Compare to your analytical answers
```

(d) (4 points) **Simpson's Paradox:** Suppose in a subsequent election:

- Turnout *increases* in every county (Urban: 62%, Suburban: 75%, Rural: 68%)
- But the population shifts toward urban areas (Urban: 60%, Suburban: 25%, Rural: 15%)

Calculate the new overall turnout rate. Is it possible for overall turnout to *decrease* even though every county's turnout increased? Explain what's happening.

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 1–2: probability axioms, conditional probability, Bayes' Rule, independence, and the Law of Total Probability.