Exercise Sheet 2: Fundamental Probability Theory

Bayes' Theorem, Transformations, and Conditional Probability

Dr. Florian Herzog, FHGR

CDS120 - Uncertainty Quantification

Instructions

This exercise sheet applies the fundamental probability theory concepts introduced in Chapter 1. You will work with Bayes' theorem, probability transformations, linear transformations of distributions, and conditional probability through both theoretical calculations and simulation studies.

Learning Objectives:

- Apply Bayes' theorem to real-world probability problems
- Understand probability transformations and their applications
- Work with moments of Student-t distributions and linear transformations
- Analyze conditional probability through urn experiments
- Connect theoretical calculations with simulation results

Notation Reminder:

- $P(A|B)$ - Conditional probability of event A given event B
- $f_X(x)$ - Probability density function of random variable X
- $F_X(x)$ - Cumulative distribution function of random variable X
- $\mathbb{E}[X]$ - Expectation of random variable X
- $\text{Var}(X)$ - Variance of random variable X

1 Exercise 1: Bayes' Theorem Applications

Bayes' theorem is fundamental for updating beliefs with new evidence and forms the foundation of Bayesian statistics and machine learning.

1.1 Part (a): Quality Control in Manufacturing

Task: A semiconductor manufacturing company uses three different production lines (A, B, and C) to produce microchips. Apply Bayes' theorem to analyze defect patterns and production optimization.

Given Information:

- Production line A produces 50% of all chips
- Production line B produces 30% of all chips

- Production line C produces 20% of all chips
- Defect rates: Line A has 2% defective chips, Line B has 3% defective, Line C has 1% defective
- A quality inspector randomly selects chips for testing without knowing which line produced them

Questions:

1. If a randomly selected chip is defective, what is the probability it came from line A?
2. If a randomly selected chip is defective, what is the probability it came from line B?
3. If a randomly selected chip is defective, what is the probability it came from line C?
4. Which production line should be investigated first when defective chips are found?
5. What is the overall defect rate across all production lines?

Requirements:

- Write out the complete Bayes' theorem setup with proper notation
- Show all intermediate calculations
- Interpret the results for production management decisions
- Explain why the line with highest defect rate might not be the most likely source

1.2 Part (b): Network Security Intrusion Detection

Task: A cybersecurity system monitors network traffic to detect potential intrusions. The system uses Bayesian analysis to classify network connections as either normal traffic or potential attacks.

Given Information:

- In a typical day, 5% of network connections are actual attack attempts
- The intrusion detection system has three alert levels:
  - **High Alert**: Triggered by 85% of actual attacks, but also by 2% of normal traffic
  - **Medium Alert**: Triggered by 60% of actual attacks, but also by 8% of normal traffic
  - **Low Alert**: Triggered by 30% of actual attacks, but also by 15% of normal traffic
- Multiple alerts can be triggered simultaneously for the same connection

Questions:

1. If the system triggers a High Alert, what is the probability that it's an actual attack?
2. If the system triggers a Medium Alert, what is the probability that it's an actual attack?
3. If the system triggers both High and Medium alerts simultaneously, what is the probability of an actual attack? (Assume alert independence given the connection type)
4. A security analyst wants to minimize false alarms while catching most attacks. Which alert level provides the best balance?
5. If no alerts are triggered, what is the probability the connection is actually an attack?

Requirements:

- Apply Bayes' theorem systematically for each alert level
- Calculate precision and recall metrics for each alert type
- Discuss the trade-off between security and operational efficiency
- Explain how base rate (5% attacks) affects the interpretation of alerts

Hints:

- Define events clearly: A = attack, H = High Alert, M = Medium Alert, L = Low Alert
- For simultaneous alerts: $P(A|H \cap M) = \frac{P(H \cap M|A) \times P(A)}{P(H \cap M)}$
- Consider both sensitivity (true positive rate) and specificity (true negative rate)
- Think about practical implications: What happens if you act on every alert?

2 Exercise 2: Probability Transformations

Understanding how transformations affect probability distributions is crucial for statistical modeling and machine learning.

2.1 Part (a): Logarithmic Transformation of Exponential Distribution

Task: Find the probability distribution of $Y = \log(X)$ when $X \sim \text{Exp}(\lambda)$.
Mathematical Setup:

- Start with $X \sim \text{Exp}(\lambda)$, so $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$
- Define the transformation $Y = \log(X)$
- Find the probability density function $f_Y(y)$

Requirements:

1. Use the transformation formula: $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$
2. Show that the inverse transformation is $X = e^Y$
3. Calculate the Jacobian of the transformation
4. Derive the final PDF of Y
5. Identify what well-known distribution this represents

Hints:

- If $Y = \log(X)$, then $X = e^Y$ and $\frac{dx}{dy} = e^y$
- The support of Y will be $(-\infty, \infty)$ since $X > 0$
- This transformation creates the Gumbel distribution

2.2 Part (b): Properties of the Transformed Distribution

Task: Analyze the properties of the transformed distribution.

Requirements:

1. Calculate $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ for $Y = \log(X)$ where $X \sim \mathrm{Exp}(\lambda)$
2. Compare the shape of the original exponential distribution with the transformed distribution
3. Explain why this transformation might be useful in practice

Hints:

- For exponential distribution: $\mathbb{E}[X] = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$
- You may need to use properties of the Euler-Mascheroni constant $\gamma \approx 0.5772$
- Consider applications in extreme value theory and survival analysis

3 Exercise 3: Linear Transformations of Student-t Distribution

The Student-t distribution is fundamental in statistical inference, especially when dealing with small samples or unknown population variance.

3.1 Part (a): Student-t Distribution Properties

Task: Work with the moments of the Student-t distribution and linear transformations.

Given Information: For $X \sim t_\nu$ (Student-t with ν degrees of freedom):

- $\mathbb{E}[X] = 0$ (for $\nu > 1$)
- $\mathrm{Var}(X) = \frac{\nu}{\nu-2}$ (for $\nu > 2$)
- The distribution is undefined for $\nu \leq 0$, has infinite variance for $\nu \leq 2$
- As $\nu \to \infty$, $t_\nu \to \mathcal{N}(0, 1)$

Questions:

1. For $X \sim t_5$, calculate $\mathbb{E}[X]$ and $\mathrm{Var}(X)$
2. For $X \sim t_{10}$, calculate $\mathbb{E}[X]$ and $\mathrm{Var}(X)$
3. Compare these results with the standard normal distribution $\mathcal{N}(0, 1)$

3.2 Part (b): Linear Transformation Analysis

Task: Analyze the linear transformation $Y = aX + b$ where $X \sim t_\nu$.

Requirements:

1. For $Y = 3X + 5$ where $X \sim t_8$, calculate:
   - $\mathbb{E}[Y]$
   - $\mathrm{Var}(Y)$
   - $\mathrm{SD}(Y)$
2. For $Y = -2X + 10$ where $X \sim t_{15}$, calculate:

- $\mathbb{E}[Y]$
- $\text{Var}(Y)$
- $\text{SD}(Y)$

3. Explain why Y does not follow a Student-t distribution (except in special cases)

Hints:

- Use the linear transformation rules: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ and $\text{Var}(aX + b) = a^2\text{Var}(X)$
- Remember that $\text{SD}(aX + b) = |a|\text{SD}(X)$
- The Student-t distribution is only preserved under transformations of the form $Y = aX$ where $a > 0$

3.3 Part (c): Degrees of Freedom Impact

Task: Investigate how degrees of freedom affect the linear transformation results.

Requirements:

1. For the transformation $Y = 2X + 3$, calculate $\mathbb{E}[Y]$ and $\text{Var}(Y)$ for:
   - $X \sim t_3$ (if variance exists)
   - $X \sim t_5$
   - $X \sim t_{20}$
   - $X \sim t_{100}$
2. Compare these results with $Y = 2Z + 3$ where $Z \sim \mathcal{N}(0, 1)$
3. Discuss the practical implications for statistical inference

4 Exercise 4: Urn Experiment - Conditional Probability and Simulation

This exercise combines theoretical probability calculations with simulation to understand sampling without replacement.

4.1 Part (a): Experimental Setup

Task: Set up the mathematical framework for an urn experiment.

Urn Contents:

- 5 white balls
- 3 red balls
- 2 green balls
- Total: 10 balls

Experiment: Draw balls one by one without replacement.

Requirements:

1. Define the sample space for the first three draws
2. Set up notation for events (e.g., W_1 = "white ball on first draw")
3. Write the probability framework for sampling without replacement

4.2 Part (b): Theoretical Calculations

Task: Calculate the following probabilities using theoretical methods.

Questions:

1. What is the probability of drawing a green ball on the second draw?
2. What is the probability of drawing a white ball on the first draw, given that a red ball was drawn on the second draw?
3. What is the probability of drawing a white ball on the third draw (marginal probability)?

Requirements:

- Show complete probability tree diagrams or systematic enumeration
- Use proper conditional probability notation
- Apply the law of total probability where appropriate
- Verify that probabilities sum to 1 where expected

Hints:

- For question 1: Consider all possible outcomes for the first draw
- For question 2: Use Bayes' theorem - this is a "reverse" conditional probability
- For question 3: By symmetry, this should equal the probability for any other draw position
- Remember that sampling without replacement creates dependencies between draws

4.3 Part (c): Simulation Study

Task: Create a simulation to verify the theoretical calculations.

Requirements:

1. Implement the urn experiment in Python/R
2. Run the simulation for $N = 10,000$ experiments
3. For each simulation:
   - Record the color of each ball drawn
   - Track the events of interest
4. Calculate empirical probabilities for all three questions
5. Compare simulation results with theoretical calculations
6. Provide confidence intervals for the simulation estimates

Programming Hints:

- Use random sampling without replacement
- Consider using numpy.random.choice with replace=False
- Track conditional events carefully (e.g., count cases where second draw is red)
- Use proper statistical testing to compare theoretical vs. empirical results

4.4 Part (d): Extended Analysis

Task: Extend the analysis to explore the properties of sampling without replacement.

Requirements:

1. Calculate the probability distribution of the number of white balls in the first three draws
2. Compare this with what the distribution would be if sampling were with replacement
3. Analyze how the dependency structure affects uncertainty quantification
4. Create visualizations showing the difference between with and without replacement scenarios

Questions to Address:

- How does sampling without replacement affect the variance of estimates?
- What happens to the dependency structure as the urn size increases?
- How does this relate to finite population corrections in survey sampling?

Deliverables

Submit the following:

1. **Theoretical solutions**: Complete mathematical solutions for all parts showing:
   - Proper use of Bayes' theorem and conditional probability
   - Correct transformation calculations with all steps
   - Student-t distribution moment calculations
   - Urn experiment probability calculations
2. **Simulation code**: Well-documented implementation for the urn experiment
3. **Results comparison**: Analysis comparing theoretical and simulation results
4. **Interpretation report**: Discussion of practical implications including:
   - When Bayes' theorem gives counterintuitive results
   - Applications of probability transformations
   - Importance of degrees of freedom in Student-t distributions
   - Impact of sampling without replacement on uncertainty

Grading Criteria:

- Mathematical accuracy and completeness
- Proper use of probability notation and concepts
- Quality of simulation implementation and validation
- Clarity of explanations and practical interpretations
- Connection to broader uncertainty quantification principles