Chapter 4: Probability Theory and Distributions - Summary and Exercises

Summary

1. Sample Spaces and Events

Sample Space (

$$\Omega$$

): The complete set of all possible outcomes of a random experiment.

Examples:

- Coin flip: $\Omega = \{H, T\}$
- Die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Temperature: $\Omega = \{t \in \mathrm{R} \mid -50 \leq t \leq 60\}$

Events: Subsets of the sample space. An event A occurs if the outcome lies within that subset.

Set Operations:

- **Union** ($A \cup B$): At least one of A or B occurs
- **Intersection** ($A \cap B$): Both A and B occur
- **Complement** (A^c): A does not occur
- **De Morgan's Laws**: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

2. Conditional Probability and Independence

Conditional Probability: The probability of event A given that B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

Independence: Events A and B are independent if:

$$P(A|B) = P(A) \quad \text{or equivalently} \quad P(A \cap B) = P(A)P(B)$$

3. Bayes' Theorem

The fundamental rule for updating probabilities based on new evidence:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Where:

- **P(B)**: Prior probability
- **P(A | B)**: Likelihood
- **P(A)**: Evidence (total probability)
- **P(B | A)**: Posterior probability

4. Random Variables

Definition: A function that assigns a real number to each outcome in the sample space: $X : \Omega \to \mathrm{R}$

Types:

- **Discrete**: Takes countable values (e.g., dice rolls, coin flips)
- **Continuous**: Takes any value in an interval (e.g., temperature, height)

5. Expected Value and Variance

Expected Value (Mean):

- Discrete: $E[X] = \sum_i x_i P(X = x_i)$
- Continuous: $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$

Variance:

$$\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

Standard Deviation: $\sigma = \sqrt{\text{Var}(X)}$

6. Key Probability Distributions

Bernoulli Distribution

Models a single binary trial (success/failure).

$$X \sim \text{Bernoulli}(p)$$

- PMF: $P(X = k) = p^k(1-p)^{1-k}, \quad k \in \{0, 1\}$
- Mean: $E[X] = p$
- Variance: $\text{Var}(X) = p(1-p)$

Binomial Distribution

Models the number of successes in n independent Bernoulli trials.

$$X \sim \text{Binomial}(n, p)$$

- PMF: $P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$
- Mean: $E[X] = np$
- Variance: $\text{Var}(X) = np(1-p)$

Poisson Distribution

Models the number of events occurring in a fixed interval.

$$X \sim \text{Poisson}(\lambda)$$

- PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Mean: $E[X] = \lambda$

- Variance: $\mathrm{Var}(X) = \lambda$

Normal (Gaussian) Distribution

The bell curve distribution, fundamental in statistics.

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Mean: $E[X] = \mu$
- Variance: $\mathrm{Var}(X) = \sigma^2$

Standard Normal: $Z \sim \mathrm{N}(0, 1)$

Z-score transformation: $Z = \frac{X-\mu}{\sigma}$

Written Exercises

Exercise 1: Sample Spaces and Events

A fair six-sided die is rolled once.

Questions:

1. Define the sample space Ω .
2. Define the following events:
    - A: Rolling an even number
    - B: Rolling a number greater than 4
    - C: Rolling a prime number
3. Find:
    - $A \cup B$
    - $A \cap B$
    - A^c
    - $B - A$

Solution

1. **Sample Space**:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

2. **Events**:

- A = {2, 4, 6} (even numbers)
- B = {5, 6} (greater than 4)
- C = {2, 3, 5} (prime numbers)

3. **Set Operations**:

$A \cup B$ (even OR greater than 4):

$$A \cup B = \{2, 4, 5, 6\}$$

$A \cap B$ (even AND greater than 4):

$$A \cap B = \{6\}$$

A^c (NOT even, i.e., odd):

$$A^c = \{1, 3, 5\}$$

$B - A$ (greater than 4 BUT NOT even):

$$B - A = \{5\}$$

Exercise 2: Conditional Probability

A company produces laptops. From historical data:

- 5% of laptops are defective: $P(D) = 0.05$
- 95% of laptops are non-defective: $P(D^c) = 0.95$
- A quality test correctly identifies 90% of defective laptops: $P(T^+ \mid D) = 0.90$
- The test incorrectly flags 8% of non-defective laptops as defective: $P(T^+ \mid D^c) = 0.08$

Questions:

1. What is the probability that a randomly selected laptop tests positive, $P(T^+)$?
2. If a laptop tests positive, what is the probability it is actually defective, $P(D \mid T^+)$?

Solution

1. $P(T^+)$ using the law of total probability:

$$P(T^+) = P(T^+|D) \cdot P(D) + P(T^+|D^c) \cdot P(D^c)$$

$$P(T^+) = (0.90)(0.05) + (0.08)(0.95)$$

$$P(T^+) = 0.045 + 0.076 = 0.121$$

2. $P(D|T^+)$ using Bayes' Theorem:

$$P(D|T^+) = \frac{P(T^+|D) \cdot P(D)}{P(T^+)}$$

$$P(D|T^+) = \frac{(0.90)(0.05)}{0.121}$$

$$P(D|T^+) = \frac{0.045}{0.121} \approx 0.372$$

Interpretation: Even though the test is 90% accurate for defective laptops, only about 37.2% of laptops that test positive are actually defective. This is because defective laptops are rare (only 5%), so false positives dominate.

Exercise 3: Expected Value and Variance

A discrete random variable X has the following probability distribution:

X	1	2	3	4
P(X)	0.1	0.3	0.4	0.2

Questions:

1. Verify that this is a valid probability distribution.
2. Calculate $E[X]$.
3. Calculate $E[X^2]$.
4. Calculate $\text{Var}(X)$.
5. Calculate the standard deviation σ .

Solution

1. **Validity Check**:

$$\sum P(X = x_i) = 0.1 + 0.3 + 0.4 + 0.2 = 1.0 \quad \checkmark$$

2. **Expected Value**:

$$E[X] = \sum x_i P(X = x_i)$$

$$E[X] = (1)(0.1) + (2)(0.3) + (3)(0.4) + (4)(0.2)$$

$$E[X] = 0.1 + 0.6 + 1.2 + 0.8 = 2.7$$

3. $E[X^2]$:

$$E[X^2] = \sum x_i^2 P(X = x_i)$$

$$E[X^2] = (1)^2(0.1) + (2)^2(0.3) + (3)^2(0.4) + (4)^2(0.2)$$

$$E[X^2] = 0.1 + 1.2 + 3.6 + 3.2 = 8.1$$

4. Variance:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

$$\mathrm{Var}(X) = 8.1 - (2.7)^2 = 8.1 - 7.29 = 0.81$$

5. Standard Deviation (σ):

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{0.81} = 0.9$$

Exercise 4: Binomial Distribution

A student takes a multiple-choice exam with 10 questions. Each question has 4 choices, and the student guesses randomly on every question.

Questions:

1. What is the probability of success (correct answer) on a single question?
2. Let X = number of correct answers. What distribution does X follow?
3. What is the probability the student gets exactly 3 questions correct?
4. What is the expected number of correct answers?
5. What is the variance of the number of correct answers?

Solution

1. **Probability of Success**:

$$p = \frac{1}{4} = 0.25$$

2. **Distribution**:

$$X \sim \mathrm{Binomial}(n = 10, p = 0.25)$$

3. $P(X = 3)$:

$$P(X = 3) = \binom{10}{3}(0.25)^3(0.75)^7$$

Calculate binomial coefficient:

$$\binom{10}{3} = \frac{10!}{3!7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

Calculate probability:

$$P(X = 3) = 120 \times (0.25)^3 \times (0.75)^7$$

$$P(X = 3) = 120 \times 0.015625 \times 0.1335 \approx 0.2503$$

4. **Expected Value**:

$$E[X] = np = (10)(0.25) = 2.5$$

5. **Variance**:

$$\mathrm{Var}(X) = np(1 - p) = (10)(0.25)(0.75) = 1.875$$

Exercise 5: Poisson Distribution

A call center receives an average of 4 calls per hour.

Questions:

1. What distribution models the number of calls in one hour?
2. What is the probability of receiving exactly 5 calls in one hour?
3. What is the probability of receiving at most 2 calls in one hour?
4. What is the expected number of calls and the variance?

Solution

1. Distribution:

$$X \sim \text{Poisson}(\lambda = 4)$$

2. $P(X = 5)$:

$$P(X = 5) = \frac{\lambda^5 e^{-\lambda}}{5!} = \frac{4^5 e^{-4}}{120}$$

$$P(X = 5) = \frac{1024 \times 0.0183}{120} \approx \frac{18.75}{120} \approx 0.156$$

3. $P(X \leq 2)$:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$P(X = 0) = \frac{4^0 e^{-4}}{0!} = e^{-4} \approx 0.0183$$

$$P(X = 1) = \frac{4^1 e^{-4}}{1!} = 4e^{-4} \approx 0.0733$$

$$P(X=2) = \frac{4^2 e^{-4}}{2!} = \frac{16e^{-4}}{2} = 8e^{-4} \approx 0.1465$$

$$P(X \leq 2) \approx 0.0183 + 0.0733 + 0.1465 = 0.2381$$

4. **Expected Value and Variance**:

   For Poisson distribution:

$$E[X] = \lambda = 4$$

$$\text{Var}(X) = \lambda = 4$$

Exercise 6: Normal Distribution and Z-scores

The heights of adult men in a population follow a normal distribution with mean $\mu = 175$ cm and standard deviation $\sigma = 8$ cm.

Questions:

1. What is the Z-score for a man who is 183 cm tall?
2. What is the Z-score for a man who is 160 cm tall?
3. Using the standard normal table (or knowing that $P(Z \leq 1) \approx 0.8413$), what proportion of men are taller than 183 cm?
4. What height corresponds to the 95th percentile? (Use $P(Z \leq 1.645) \approx 0.95$)

Solution

1. **Z-score for 183 cm**:

$$Z = \frac{X - \mu}{\sigma} = \frac{183 - 175}{8} = \frac{8}{8} = 1$$

2. Z-score for 160 cm:

$$Z = \frac{160 - 175}{8} = \frac{-15}{8} = -1.875$$

3. Proportion taller than 183 cm:

We need $P(X > 183) = P(Z > 1)$

$$P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.8413 = 0.1587$$

About 15.87% of men are taller than 183 cm.

4. 95th Percentile Height:

We need X such that $P(X \leq x) = 0.95$

This corresponds to Z = 1.645

$$Z = \frac{X - \mu}{\sigma}$$

$$1.645 = \frac{X - 175}{8}$$

$$X = 175 + 1.645 \times 8 = 175 + 13.16 = 188.16 \text{ cm}$$

The 95th percentile height is approximately 188 cm.

Exercise 7: Bayes' Theorem - Medical Diagnosis

A disease affects 1% of the population. A diagnostic test has the following properties:

- Sensitivity (true positive rate): $P(T^+ \mid D) = 0.95$
- Specificity: $P(T^- \mid D^c) = 0.90$

Questions:

1. What is the false positive rate $P(T^+ \mid D^c)$?
2. What is the probability a randomly selected person tests positive?
3. If a person tests positive, what is the probability they actually have the disease?

Solution

1. **False Positive Rate**:

$$P(T^+ \mid D^c) = 1 - P(T^- \mid D^c) = 1 - 0.90 = 0.10$$

2. $P(T^+)$ using total probability:

$$P(T^+) = P(T^+|D) \cdot P(D) + P(T^+|D^c) \cdot P(D^c)$$

$$P(T^+) = (0.95)(0.01) + (0.10)(0.99)$$

$$P(T^+) = 0.0095 + 0.099 = 0.1085$$

3. $P(D \mid T^+)$ using Bayes' Theorem:

$$P(D \mid T^+) = \frac{P(T^+ \mid D) \cdot P(D)}{P(T^+)}$$

$$P(D|T^+) = \frac{(0.95)(0.01)}{0.1085} = \frac{0.0095}{0.1085} \approx 0.0876$$

Interpretation: Even with a 95% sensitive test, only about 8.76% of people who test positive actually have the disease. This is because the disease is rare (1%), so false positives outnumber true positives.

Python Code Examples

Example 1: Binomial Probability Calculation

```
import scipy.stats as stats

# Parameters
n = 10
p = 0.25
k = 3

# Calculate P(X = k)
prob = stats.binom.pmf(k, n, p)
print(f"P(X = {k}) = {prob:.4f}")

# Expected value and variance
mean = n * p
variance = n * p * (1 - p)
print(f"E[X] = {mean}")
print(f"Var(X) = {variance}")
```

Example 2: Normal Distribution Calculations

```
import scipy.stats as stats

# Parameters
mu = 175
sigma = 8
x = 183

# Calculate Z-score
z = (x - mu) / sigma
print(f"Z-score for {x} cm: {z:.2f}")

# Calculate probability
prob_less = stats.norm.cdf(z)
prob_greater = 1 - prob_less
print(f"P(X > {x}) = {prob_greater:.4f}")

# Find 95th percentile
percentile_95 = stats.norm.ppf(0.95, loc=mu, scale=sigma)
print(f"95th percentile: {percentile_95:.2f} cm")
```

Example 3: Poisson Probability

```
import scipy.stats as stats

# Parameter
lambda_val = 4

# P(X = 5)
prob_5 = stats.poisson.pmf(5, lambda_val)
print(f"P(X = 5) = {prob_5:.4f}")

# P(X <= 2)
prob_le_2 = stats.poisson.cdf(2, lambda_val)
print(f"P(X <= 2) = {prob_le_2:.4f}")
```