Practice Problem Set
Introduction to Statistical Methods

For My Sec9 Students

August 26, 2025

Contents

1 Descriptive Statistics and Probability

This section covers foundational concepts including measures of central tendency, variability, data distribution shapes, and the fundamentals of probability theory.

Problem 1.1. The following data represents the time (in minutes) taken by 10 employees to complete a task: 39, 29, 43, 52, 39, 44, 40, 31, 44, 35.

(a) Calculate the mean, median, and mode.

(b) Calculate the range, variance, and standard deviation.

(c) Based on the mean and median, comment on the skewness of the data.

Solution. (a) Measures of Central Tendency

Step 1: Calculate the Mean. The mean is the sum of the values divided by the count.

$$\bar{x} = \frac{39+29+43+52+39+44+40+31+44+35}{10} = \frac{396}{10} = 39.6$$

Step 2: Calculate the Median. First, sort the data in ascending order: 29, 31, 35, 39, 39, 40, 43, 44, 44, 52. Since there is an even number of observations (n=10), the median is the average of the two middle values (5th and 6th).

$$\text{Median} = \frac{39+40}{2} = 39.5$$

Step 3: Identify the Mode. The mode is the most frequently occurring value. The values 39 and 44 both appear twice, which is more than any other value. Thus, the data is bimodal.

$$\text{Mode} = 39 \text{ and } 44$$

(b) Measures of Variability

Step 1: Calculate the Range. Range = Maximum Value - Minimum Value.

$$\text{Range} = 52 - 29 = 23$$

Step 2: Calculate the Sample Variance (s^2). The formula for sample variance is $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$.

$$\begin{aligned}\sum(x_i-\bar{x})^2 &= (29-39.6)^2 + (31-39.6)^2 + \cdots + (52-39.6)^2 \\ &= (-10.6)^2 + (-8.6)^2 + (-4.6)^2 + (-0.6)^2 + (-0.6)^2 \\ &\quad + (0.4)^2 + (3.4)^2 + (4.4)^2 + (4.4)^2 + (12.4)^2 \\ &= 112.36 + 73.96 + 21.16 + 0.36 + 0.36 + 0.16 + 11.56 + 19.36 + 19.36 + 153.76 \\ &= 412.4\end{aligned}$$

$$s^2 = \frac{412.4}{10-1} = \frac{412.4}{9} \approx 45.82$$

Step 3: Calculate the Standard Deviation (s). The standard deviation is the square root of the variance.

$$s = \sqrt{45.82} \approx 6.77$$

(c) Comment on Skewness

Step 1: Compare Mean and Median. We calculated Mean = 39.6 and Median = 39.5.

Step 2: Conclusion. Since the mean (39.6) is slightly greater than the median (39.5), the data is slightly **positively skewed** (or right-skewed). This indicates a slight tail towards the higher values.

Problem 1.2. In a university, 60% of students have a laptop. Among students with a laptop, 75% also have a tablet. 45% of all students have a tablet.

(a) What is the probability that a randomly selected student has both a laptop and a tablet?

(b) What is the probability that a student has a laptop or a tablet (or both)?

(c) Are the events "has a laptop" and "has a tablet" independent? Justify your answer.

Solution. Let L be the event that a student has a laptop, and T be the event that a student has a tablet. We are given: $P(L) = 0.60$, $P(T) = 0.45$, and the conditional probability $P(T|L) = 0.75$.

(a) Probability of having both a laptop and a tablet

Step 1: Use the definition of conditional probability. The formula is $P(T \cap L) = P(T|L) \times P(L)$.

Step 2: Calculate the probability.

$$P(T \cap L) = 0.75 \times 0.60 = 0.45$$

The probability that a student has both is 45%.

(b) Probability of having a laptop or a tablet

Step 1: Use the general addition rule for probability. The formula is $P(L \cup T) = P(L) + P(T) - P(L \cap T)$.

Step 2: Calculate the probability.

$$P(L \cup T) = 0.60 + 0.45 - 0.45 = 0.60$$

The probability that a student has either a laptop or a tablet is 60%.

(c) Independence of Events

Step 1: State the condition for independence. Two events L and T are independent if and only if $P(L \cap T) = P(L) \times P(T)$.

Step 2: Check the condition.

- From part (a), we know $P(L \cap T) = 0.45$.
- Calculate the product of individual probabilities: $P(L) \times P(T) = 0.60 \times 0.45 = 0.27$.

Step 3: Conclusion. Since $0.45 \neq 0.27$, the events "has a laptop" and "has a tablet" are **not independent**.

Problem 1.3. A systems manager tracks the number of server failures per day. For the past two weeks, the data is: 1, 3, 0, 3, 2, 3, 4, 3, 0, 1, 3, 2, 6, 4. Construct a five-number summary for this data and identify any potential outliers using the 1.5 × IQR rule.

Solution.

Step 1: Sort the data. First, arrange the 14 data points in ascending order:

$$0, 0, 1, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 6$$

Step 2: Find the Minimum and Maximum.

- Minimum = 0
- Maximum = 6

Step 3: Find the Median (Q2). With $n = 14$ (an even number), the median is the average of the 7th and 8th values.

$$\text{Median (Q2)} = \frac{3+3}{2} = 3$$

Step 4: Find the First Quartile (Q1). Q1 is the median of the lower half of the data: $\{0, 0, 1, 1, 2, 2, 3\}$. The middle value is the 4th value.

$$\text{Q1} = 1$$

Step 5: Find the Third Quartile (Q3). Q3 is the median of the upper half of the data: $\{3, 3, 3, 3, 4, 4, 6\}$. The middle value is the 4th value of this set.

$$\text{Q3} = 3$$

Step 6: State the Five-Number Summary. The five-number summary is: **Minimum = 0, Q1 = 1, Median = 3, Q3 = 3, Maximum = 6**.

Step 7: Identify Potential Outliers.

- Calculate the Interquartile Range (IQR): $IQR = Q3 - Q1 = 3 - 1 = 2$.
- Calculate the lower fence: $Q1 - 1.5 \times IQR = 1 - 1.5 \times 2 = 1 - 3 = -2$.
- Calculate the upper fence: $Q3 + 1.5 \times IQR = 3 + 1.5 \times 2 = 3 + 3 = 6$.

Step 8: Conclusion. Any data point below -2 or above 6 is a potential outlier. In our dataset, the minimum is 0 (which is greater than -2) and the maximum is 6 (which is not greater than the upper fence of 6). Therefore, according to the 1.5 × IQR rule, there are **no outliers** in this dataset.

Problem 1.4. At a factory, three machines (A, B, and C) produce bolts. Machine A produces 40% of the bolts, Machine B produces 35%, and Machine C produces 25%. The defective rates for these machines are 5%, 3%, and 2%, respectively. A bolt is selected at random from the total output and is found to be defective. What is the probability that it was produced by Machine A?

Solution. This is a classic application of Bayes' Theorem. Let A, B, C be the events that a bolt is produced by Machine A, B, C respectively. Let D be the event that a bolt is defective.

Step 1: List the known probabilities.

- Priors: $P(A) = 0.40$, $P(B) = 0.35$, $P(C) = 0.25$.
- Likelihoods (Conditional Probabilities): $P(D|A) = 0.05$, $P(D|B) = 0.03$, $P(D|C) = 0.02$.

Step 2: State the Goal. We want to find the posterior probability $P(A|D)$.

Step 3: Apply Bayes' Theorem Formula.

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

Step 4: Calculate the Total Probability of a Defect, P(D). We use the Law of Total Probability:

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) \\ &= (0.05)(0.40) + (0.03)(0.35) + (0.02)(0.25) \\ &= 0.0200 + 0.0105 + 0.0050 \\ &= 0.0355 \end{aligned}$$

The overall probability of picking a defective bolt is 3.55%.

Step 5: Calculate the Final Posterior Probability.

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{(0.05)(0.40)}{0.0355} = \frac{0.0200}{0.0355} \approx 0.5634$$

Step 6: Conclusion. Given that a defective bolt is selected, the probability that it came from Machine A is approximately **56.34%**.

Problem 1.5. Two dice are thrown. Let A be the event that the sum is greater than 8, and B be the event that at least one die shows a 6.

(a) Find $P(A)$.

(b) Find $P(B)$.

(c) Find the conditional probability $P(A|B)$.

Solution. The sample space consists of 36 equally likely outcomes.

(a) Find P(A)

Step 1: List the outcomes for event A (sum ¿ 8). The possible sums are 9, 10, 11, 12.

- Sum=9: (3,6), (4,5), (5,4), (6,3)
- Sum=10: (4,6), (5,5), (6,4)
- Sum=11: (5,6), (6,5)
- Sum=12: (6,6)

There are 4+3+2+1 = 10 favorable outcomes.

Step 2: Calculate the probability.

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Total outcomes}} = \frac{10}{36} = \frac{5}{18}$$

(b) Find P(B)

Step 1: List the outcomes for event B (at least one 6).

- Die 1 is 6: (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
- Die 2 is 6 (excluding (6,6) which is already counted): (1,6), (2,6), (3,6), (4,6), (5,6)

There are 6+5 = 11 favorable outcomes.

Step 2: Calculate the probability.

$$P(B) = \frac{11}{36}$$

(c) Find P(A—B)

Step 1: Use the formula for conditional probability.

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

Step 2: Find the intersection $A \cap B$. This is the event that the sum is greater than 8 AND at least one die is a 6. We look at the outcomes for A and see which ones have a 6. Outcomes from A: (3,6), (4,5), (5,4), (6,3), (4,6), (5,5), (6,4), (5,6), (6,5), (6,6). Outcomes with a 6: (3,6), (6,3), (4,6), (6,4), (5,6), (6,5), (6,6). There are 7 outcomes in the intersection. So, $P(A \cap B) = \frac{7}{36}$.

Step 3: Calculate the conditional probability.

$$P(A|B) = \frac{7/36}{11/36} = \frac{7}{11}$$

Step 4: Conclusion. The probability that the sum is greater than 8, given that at least one die shows a 6, is **7/11**.

Problem 1.6. A committee of 5 is to be chosen from a group of 8 men and 4 women. What is the probability that the committee contains a majority of women?

Solution.

Step 1: Define the total number of possible committees. There are a total of $8 + 4 = 12$ people. The total number of ways to choose a committee of 5 is given by the combination formula:

$$\text{Total Committees} = \binom{12}{5} = \frac{12!}{5!(12-5)!} = \frac{12 \times 11 \times 10 \times 9 \times 8}{5 \times 4 \times 3 \times 2 \times 1} = 792$$

Step 2: Define the favorable outcomes. A majority of women means the committee must have more women than men. Since the committee size is 5, this means having 3, 4, or 5 women. As there are only 4 women available, the possibilities are:

- Case 1: 3 women and 2 men.
- Case 2: 4 women and 1 man.

Step 3: Calculate the number of ways for each case.

- Case 1: $\binom{4}{3} \times \binom{8}{2} = 4 \times \frac{8\times7}{2} = 4 \times 28 = 112$
- Case 2: $\binom{4}{4} \times \binom{8}{1} = 1 \times 8 = 8$

Step 4: Calculate the total number of favorable outcomes. Total favorable outcomes $= 112 + 8 = 120$.

Step 5: Calculate the final probability.

$$P(\text{Majority Women}) = \frac{\text{Favorable Outcomes}}{\text{Total Outcomes}} = \frac{120}{792} = \frac{10}{66} = \frac{5}{33}$$

Problem 1.7. For two events A and B, we have $P(A \cup B) = 0.8$, $P(A) = 0.5$, and $P(B|A) = 0.6$. Find $P(B)$.

Solution.

Step 1: Find the intersection probability, $P(A \cap B)$. Using the conditional probability formula:

$$P(A \cap B) = P(B|A) \times P(A) = 0.6 \times 0.5 = 0.30$$

Step 2: Use the general addition rule to find $P(B)$. The rule is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can rearrange this to solve for $P(B)$.

$$P(B) = P(A \cup B) - P(A) + P(A \cap B)$$

Step 3: Substitute the known values.

$$P(B) = 0.8 - 0.5 + 0.3 = 0.6$$

Step 4: Conclusion. The probability of event B is **0.6**.

Problem 1.8. You are presented with two datasets of salaries for a department.

- Dataset 1: \$50k, \$52k, \$55k, \$58k, \$60k
- Dataset 2: \$50k, \$52k, \$55k, \$58k, \$150k

For each dataset, calculate the mean and median.

Solution.

Step 1: Calculate Mean and Median for Dataset 1.

- Mean: $\frac{50+52+55+58+60}{5} = \frac{275}{5} = \$55k$
- Median: The data is sorted. The middle value (3rd) is \$55k.

Step 2: Calculate Mean and Median for Dataset 2.

- Mean: $\frac{50+52+55+58+150}{5} = \frac{365}{5} = \$73k$
- Median: The data is sorted. The middle value (3rd) is \$55k.

Step 3: Conclusion. For Dataset 1, Mean = \$55k and Median = \$55k. For Dataset 2, Mean = \$73k and Median = \$55k. The median is a more appropriate measure of central tendency for Dataset 2 because it is robust to the outlier (\$150k).

Problem 1.9. Let the sample space be $S = \{1, 2, 3, 4, 5, 6\}$. Let A be the event $\{1, 2, 3\}$ and B be the event $\{3, 4, 5\}$. Find the probabilities of the following events:

(a) $P(A \text{ and } B)$, i.e., $P(A \cap B)$

(b) $P(A \text{ or } B)$, i.e., $P(A \cup B)$

(c) $P(A^c)$, the complement of A

Solution. Assuming each outcome in the sample space is equally likely, the probability of any single outcome is 1/6.

(a) Probability of A and B

Step 1: Find the intersection of A and B. $A \cap B = \{1, 2, 3\} \cap \{3, 4, 5\} = \{3\}$.

Step 2: Calculate the probability. The intersection contains one outcome.

$$P(A \cap B) = \frac{1}{6}$$

(b) Probability of A or B

Step 1: Find the union of A and B. $A \cup B = \{1, 2, 3\} \cup \{3, 4, 5\} = \{1, 2, 3, 4, 5\}$.

Step 2: Calculate the probability. The union contains five outcomes.

$$P(A \cup B) = \frac{5}{6}$$

(c) Probability of the complement of A

Step 1: Find the complement of A. $A^c = S - A = \{1, 2, 3, 4, 5, 6\} - \{1, 2, 3\} = \{4, 5, 6\}$.

Step 2: Calculate the probability. The complement contains three outcomes.

$$P(A^c) = \frac{3}{6} = \frac{1}{2}$$

Problem 1.10. A survey of a city's population showed that 40% read Newspaper A, 25% read Newspaper B, and 15% read both.

(a) What percentage of the population reads at least one of the newspapers?

(b) What percentage of the population reads exactly one newspaper?

(c) Of those who read Newspaper A, what percentage also read Newspaper B?

Solution. Let A be the event of reading Newspaper A and B be the event of reading Newspaper B. Given: $P(A) = 0.40$, $P(B) = 0.25$, $P(A \cap B) = 0.15$.

(a) Percentage reading at least one newspaper

Step 1: Identify the required probability. "At least one" corresponds to the union of the two events, $P(A \cup B)$.

Step 2: Apply the addition rule.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.40 + 0.25 - 0.15 = 0.50$$

Step 3: Conclusion. 50% of the population reads at least one newspaper.

(b) Percentage reading exactly one newspaper

Step 1: Formulate the probability. This corresponds to the outcomes in the union minus the outcomes in the intersection.

$$P(\text{Exactly One}) = P(A \cup B) - P(A \cap B)$$

Step 2: Calculate the value.

$$P(\text{Exactly One}) = 0.50 - 0.15 = 0.35$$

Step 3: Conclusion. 35% of the population reads exactly one newspaper.

(c) Percentage who read B, given they read A

Step 1: Identify the required probability. This is the conditional probability $P(B|A)$.

Step 2: Apply the conditional probability formula.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.15}{0.40} = \frac{15}{40} = \frac{3}{8} = 0.375$$

Step 3: Conclusion. Of those who read Newspaper A, 37.5% also read Newspaper B.

2 Probability Distributions

This section explores various discrete and continuous probability distributions, including their properties, applications, and the calculation of key metrics like mean and variance.

Problem 2.1. A call center receives an average of 4 calls per minute. Assuming the calls follow a Poisson distribution, find the probability that in a given minute:

(a) Exactly 2 calls are received.

(b) At least 3 calls are received.

(c) No calls are received in a 30-second interval.

Solution. The Poisson probability mass function is given by $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, where λ is the average rate of events.

(a) Exactly 2 calls are received

Step 1: Identify parameters. For a one-minute interval, $\lambda = 4$. We want to find $P(X = 2)$.

Step 2: Apply the formula.

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

(b) At least 3 calls are received

Step 1: Formulate the probability. ”At least 3” means $P(X \geq 3)$. It’s easier to calculate the complement: $1 - P(X < 3)$.

$$P(X \geq 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

Step 2: Calculate the individual probabilities. We have $\lambda = 4$.

- $P(X = 0) = \frac{e^{-4}4^0}{0!} = e^{-4} \approx 0.0183$
- $P(X = 1) = \frac{e^{-4}4^1}{1!} = 4e^{-4} \approx 0.0733$
- $P(X = 2) \approx 0.1465$ (from part a)

Step 3: Calculate the final probability.

$$P(X \geq 3) = 1 - (0.0183 + 0.0733 + 0.1465) = 1 - 0.2381 = 0.7619$$

(c) No calls in a 30-second interval

Step 1: Adjust the rate λ for the new interval. The original rate is 4 calls per 60 seconds. For a 30-second interval, the new average rate is:

$$\lambda' = 4 \times \frac{30}{60} = 2$$

Step 2: Calculate the probability of zero events. We want to find $P(X = 0)$ with $\lambda' = 2$.

$$P(X = 0) = \frac{e^{-2}2^0}{0!} = e^{-2} \approx 0.1353$$

Problem 2.2. A machine produces items with a 10% defect rate. A quality inspector randomly selects 8 items. Let X be the number of defective items found.

(a) What is the probability that exactly 2 items are defective?

(b) What is the probability that at most 1 item is defective?

(c) What are the mean and variance of the number of defective items?

Solution. This scenario follows a Binomial distribution. The probability mass function is $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$.

(a) Probability of exactly 2 defects

Step 1: Identify parameters. $n = 8, p = 0.1, k = 2$.

Step 2: Apply the formula.

$$P(X = 2) = \binom{8}{2}(0.1)^2(0.9)^{8-2} = \frac{8!}{2!6!}(0.01)(0.9)^6 = 28 \times 0.01 \times 0.531441 \approx 0.1488$$

(b) Probability of at most 1 defect

Step 1: Formulate the probability. "At most 1" means $P(X \leq 1) = P(X = 0) + P(X = 1)$.

Step 2: Calculate the individual probabilities.

- $P(X = 0) = \binom{8}{0}(0.1)^0(0.9)^8 = 1 \times 1 \times 0.430467 \approx 0.4305$
- $P(X = 1) = \binom{8}{1}(0.1)^1(0.9)^7 = 8 \times 0.1 \times 0.478297 \approx 0.3826$

Step 3: Calculate the final probability.

$$P(X \leq 1) = 0.4305 + 0.3826 = 0.8131$$

(c) Mean and Variance

Step 1: Use the formulas for the Binomial distribution. Mean (μ) = np; Variance (σ^2) = $np(1-p)$.

Step 2: Calculate the values.

- $\mu = 8 \times 0.1 = 0.8$
- $\sigma^2 = 8 \times 0.1 \times 0.9 = 0.72$

Problem 2.3. The height of adult males in a city is normally distributed with a mean of 175 cm and a standard deviation of 7 cm.

(a) What is the probability that a randomly selected male is taller than 185 cm?

(b) What is the probability that a randomly selected male has a height between 170 cm and 180 cm?

(c) What height is the 90th percentile?

Solution. We use the standard normal distribution by converting the height X to a Z-score using the formula $Z = \frac{X-\mu}{\sigma}$. Here, $\mu = 175$ and $\sigma = 7$.

(a) Probability of being taller than 185 cm

Step 1: Calculate the Z-score for X = 185.

$$Z = \frac{185 - 175}{7} = \frac{10}{7} \approx 1.43$$

Step 2: Find the probability from the Z-table. We want $P(Z > 1.43) = 1 - P(Z < 1.43)$.

$$P(Z > 1.43) = 1 - 0.9236 = 0.0764$$

(b) Probability of height between 170 and 180 cm

Step 1: Calculate the Z-scores for X=170 and X=180.

$$Z_1 = \frac{170 - 175}{7} \approx -0.71 \quad , \quad Z_2 = \frac{180 - 175}{7} \approx 0.71$$

Step 2: Find the probability from the Z-table. We want $P(-0.71 < Z < 0.71) = P(Z < 0.71) - P(Z < -0.71)$.

$$P(-0.71 < Z < 0.71) = 0.7611 - 0.2389 = 0.5222$$

(c) Height at the 90th percentile

Step 1: Find the Z-score corresponding to the 90th percentile. We look for the Z-score such that $P(Z < z) = 0.90$. From the Z-table, the closest Z-score is approximately 1.28.

Step 2: Convert the Z-score back to the original scale (X). Use $X = \mu + Z\sigma$.

$$X = 175 + (1.28)(7) = 175 + 8.96 = 183.96$$

The 90th percentile for height is approximately 183.96 cm.

Problem 2.4. Let X and Y be two discrete random variables with the following joint probability mass function $f(x, y)$:

f(x,y)	**y=0**	**y=1**	**y=2**
x=0	0.1	0.2	0.1
x=1	0.3	0.1	0.2

(a) Find the marginal probability distributions of X and Y.

(b) Calculate the expected value of X, E(X), and the expected value of Y, E(Y).

(c) Are X and Y independent? Justify your answer.

Solution. (a) Marginal Distributions

Step 1: Find the marginal distribution of X, g(x). Sum probabilities across rows.

- $g(0) = P(X = 0) = 0.1 + 0.2 + 0.1 = 0.4$
- $g(1) = P(X = 1) = 0.3 + 0.1 + 0.2 = 0.6$

Step 2: Find the marginal distribution of Y, h(y). Sum probabilities down columns.

- $h(0) = P(Y = 0) = 0.1 + 0.3 = 0.4$

- $h(1) = P(Y = 1) = 0.2 + 0.1 = 0.3$
- $h(2) = P(Y = 2) = 0.1 + 0.2 = 0.3$

(b) Expected Values

Step 1: Calculate E(X). $E(X) = \sum x \cdot g(x) = (0 \times 0.4) + (1 \times 0.6) = 0.6$.

Step 2: Calculate E(Y). $E(Y) = \sum y \cdot h(y) = (0 \times 0.4) + (1 \times 0.3) + (2 \times 0.3) = 0 + 0.3 + 0.6 = 0.9$.

(c) Independence of X and Y

Step 1: State the condition for independence. X and Y are independent if $f(x, y) = g(x) \times h(y)$ for all (x,y).

Step 2: Test a point, e.g., (x=0, y=0).

- Joint probability: $f(0, 0) = 0.1$.
- Product of marginals: $g(0) \times h(0) = 0.4 \times 0.4 = 0.16$.

Step 3: Conclusion. Since $0.1 \neq 0.16$, the random variables X and Y are **not independent**.

Problem 2.5. The time (in hours) a user spends on a social media app is a continuous random variable X with the probability density function (PDF): $f(x) = \frac{1}{4}x$ for $0 \leq x \leq \sqrt{8}$, and $f(x) = 0$ otherwise.

(a) Verify that this is a valid PDF.

(b) Find the probability that a user spends less than 1 hour on the app, $P(X < 1)$.

(c) Find the expected time spent on the app, E(X).

Solution. (a) Verify the PDF

Step 1: Check for non-negativity. For $0 \leq x \leq \sqrt{8}$, $f(x) = \frac{1}{4}x \geq 0$. The condition is met.

Step 2: Check that the total area under the curve is 1.

$$\int_0^{\sqrt{8}} \frac{1}{4}x\,dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_0^{\sqrt{8}} = \frac{1}{8}[(\sqrt{8})^2 - 0^2] = \frac{8}{8} = 1$$

Since both conditions are met, it is a valid PDF.

(b) Probability of spending less than 1 hour

Step 1: Evaluate the integral.

$$P(X < 1) = \int_0^1 \frac{1}{4}x\,dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_0^1 = \frac{1}{8}(1^2 - 0^2) = \frac{1}{8} = 0.125$$

(c) Expected time spent

Step 1: Evaluate the integral for E(X). $E(X) = \int_0^{\sqrt{8}} x \cdot f(x)dx$.

$$E(X) = \int_0^{\sqrt{8}} \frac{1}{4}x^2 dx = \frac{1}{4}\left[\frac{x^3}{3}\right]_0^{\sqrt{8}} = \frac{1}{12}[(\sqrt{8})^3] = \frac{16\sqrt{2}}{12} = \frac{4\sqrt{2}}{3} \approx 1.886$$

Problem 2.6. A factory produces light bulbs, and 2% of them are defective. A large shipment of 500 bulbs is sent to a retailer.

(a) Write the formula for the exact probability that exactly 10 bulbs are defective.

(b) Use the Poisson approximation to the Binomial distribution to estimate this probability.

Solution. **(a) Exact Binomial Probability** The exact probability is given by the Binomial formula with $n = 500, p = 0.02, k = 10$:

$$P(X = 10) = \binom{500}{10}(0.02)^{10}(0.98)^{490}$$

(b) Poisson Approximation The Poisson approximation is appropriate since $n = 500$ is large and $p = 0.02$ is small.

Step 1: Calculate the rate parameter λ.

$$\lambda = np = 500 \times 0.02 = 10$$

Step 2: Apply the Poisson formula.

$$P(X = 10) \approx \frac{e^{-\lambda}\lambda^{10}}{10!} = \frac{e^{-10}10^{10}}{10!} \approx 0.1251$$

Problem 2.7. A continuous random variable X has the cumulative distribution function (CDF):

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^3/8 & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

(a) Find the probability density function (PDF), $f(x)$.

(b) Calculate $P(X > 1)$.

(c) Find the median of this distribution.

Solution. **(a) Find the PDF** The PDF is the derivative of the CDF, $f(x) = F'(x)$.

$$f(x) = \frac{d}{dx}\left(\frac{x^3}{8}\right) = \frac{3x^2}{8} \quad \text{for } 0 \leq x \leq 2$$

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(b) Calculate $P(X > 1)$ Using the CDF, $P(X > 1) = 1 - P(X \leq 1) = 1 - F(1)$.

$$P(X > 1) = 1 - \frac{1^3}{8} = 1 - \frac{1}{8} = \frac{7}{8}$$

(c) Find the Median The median (m) is the value where the CDF is 0.5.

$$F(m) = 0.5 \implies \frac{m^3}{8} = 0.5 \implies m^3 = 4 \implies m = \sqrt[3]{4} \approx 1.587$$

Problem 2.8. Let X be a random variable with mean $E(X) = 10$ and variance $Var(X) = 4$. Let Y be a new random variable defined as $Y = 3X - 5$. Find the mean and variance of Y.

Solution.

Step 1: Calculate E(Y). Using the property $E(aX + b) = aE(X) + b$:

$$E(Y) = E(3X - 5) = 3E(X) - 5 = 3(10) - 5 = 25$$

Step 2: Calculate Var(Y). Using the property $Var(aX + b) = a^2 Var(X)$:

$$Var(Y) = Var(3X - 5) = 3^2 Var(X) = 9 \times 4 = 36$$

Step 3: Conclusion. The mean of Y is **25** and the variance of Y is **36**.

Problem 2.9. For the joint PDF $f(x, y) = \frac{x+y}{3}$ for $0 < x < 2, 0 < y < 1$, find the conditional PDF of X given $Y = y$, denoted $f(x|y)$, and calculate $P(X > 1|Y = 0.5)$.

Solution.

Step 1: Find the marginal PDF of Y, $h(y)$.

$$h(y) = \int_0^2 \frac{x+y}{3} dx = \frac{1}{3} \left[\frac{x^2}{2} + yx \right]_0^2 = \frac{1}{3}(2 + 2y) = \frac{2(1+y)}{3}$$

Step 2: Find the conditional PDF $f(x|y)$.

$$f(x|y) = \frac{f(x, y)}{h(y)} = \frac{(x + y)/3}{2(1 + y)/3} = \frac{x + y}{2(1 + y)}$$

Step 3: Calculate $P(X > 1|Y = 0.5)$. First, find the specific conditional PDF for $y = 0.5$: $f(x|Y = 0.5) = \frac{x+0.5}{3}$. Now integrate this PDF from 1 to 2.

$$P(X > 1|Y = 0.5) = \int_1^2 \frac{x + 0.5}{3} dx = \frac{1}{3} \left[\frac{x^2}{2} + 0.5x \right]_1^2$$

$$= \frac{1}{3} \left[(2 + 1) - (0.5 + 0.5)\right] = \frac{1}{3}[3 - 1] = \frac{2}{3}$$

3 Hypothesis Testing

This section focuses on the procedures for making statistical inferences about population parameters based on sample data, including Z-tests, t-tests, F-tests, and ANOVA.

Problem 3.1. A car manufacturer claims that its new model has an average fuel efficiency of at least 35 miles per gallon (mpg). A consumer advocacy group tests a random sample of 49 cars and finds a sample mean of 34.2 mpg. Assuming the population standard deviation is known to be 2.8 mpg, test the manufacturer's claim at a 5% level of significance ($\alpha = 0.05$).

Solution. We perform a one-sample Z-test for the mean.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\mu \geq 35$ (The claim is true)
- Alternative Hypothesis (H_1): $\mu < 35$ (The claim is false)

Step 2: Calculate the Test Statistic.

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{34.2 - 35}{2.8/\sqrt{49}} = \frac{-0.8}{0.4} = -2.00$$

Step 3: Determine the Rejection Region. For a left-tailed test at $\alpha = 0.05$, the critical Z-value is $Z_{critical} = -1.645$. We will reject H_0 if $Z < -1.645$.

Step 4: Make a Decision. Since $-2.00 < -1.645$, the test statistic falls into the rejection region.

Step 5: Conclusion. At the 5% level of significance, we reject the null hypothesis. There is sufficient evidence to contradict the manufacturer's claim.

Problem 3.2. To compare the effectiveness of two teaching methods (A and B), two groups of students are randomly selected. Group A (12 students) has a mean score of 85 with standard deviation 4. Group B (10 students) has a mean score of 81 with standard deviation 5. Assume normal populations with equal variances. Test for a significant difference in mean scores at $\alpha = 0.05$.

Solution. We use an independent samples t-test.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\mu_A = \mu_B$
- Alternative Hypothesis (H_1): $\mu_A \neq \mu_B$

Step 2: Calculate the Pooled Standard Deviation (s_p).

$$s_p^2 = \frac{(12-1)(4^2) + (10-1)(5^2)}{12 + 10 - 2} = \frac{176 + 225}{20} = 20.05 \implies s_p \approx 4.478$$

Step 3: Calculate the Test Statistic.

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{(85 - 81)}{4.478\sqrt{\frac{1}{12} + \frac{1}{10}}} \approx 2.086$$

Step 4: Determine the Rejection Region. Degrees of freedom $df = 20$. For a two-tailed test at $\alpha = 0.05$, $t_{critical} = \pm 2.086$.

Step 5: Make a Decision. The calculated t-statistic (2.086) equals the critical value.

Step 6: Conclusion. We reject the null hypothesis. There is a statistically significant difference between the mean scores.

Problem 3.3. A new diet program is tested on 9 volunteers. Their weights (in kg) were recorded before and after. Test if the diet program is effective in reducing weight at the 1% significance level.

Weights:

Before	75	70	46	68	68	43	55	68	77
After	70	77	57	60	79	64	55	77	76

Solution. This requires a paired t-test on the differences, $d = \text{Before} - \text{After}$.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\mu_d \leq 0$
- Alternative Hypothesis (H_1): $\mu_d > 0$ (The diet reduces weight)

Step 2: Calculate the Differences and Summary Statistics.

- d: 5, -7, -11, 8, -11, -21, 0, -9, 1
- Mean difference $\bar{d} = -5$
- Standard deviation of differences $s_d \approx 9.206$

Step 3: Calculate the Test Statistic.

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{-5}{9.206/\sqrt{9}} \approx -1.629$$

Step 4: Determine the Rejection Region. $df = 8$. For a right-tailed test at $\alpha = 0.01$, $t_{critical} = 2.896$. Reject H_0 if $t > 2.896$.

Step 5: Make a Decision. Since -1.629 is not greater than 2.896, we fail to reject H_0.

Step 6: Conclusion. At the 1% significance level, there is not sufficient evidence to conclude that the diet program is effective.

Problem 3.4. A company uses two different machines to produce resistors. A sample of 25 resistors from machine 1 has a variance of $s_1^2 = 1.04$ and a sample of 25 from machine 2 has $s_2^2 = 0.51$. Test at a 5% level of significance if there is a significant difference in the variances.

Solution. We use an F-test to compare two population variances.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$
- Alternative Hypothesis (H_1): $\sigma_1^2 \neq \sigma_2^2$

Step 2: Calculate the Test Statistic.

$$F = \frac{s^2_{larger}}{s^2_{smaller}} = \frac{1.04}{0.51} \approx 2.04$$

Step 3: Determine the Rejection Region. $df_1 = 24$, $df_2 = 24$. For a two-tailed test at $\alpha = 0.05$, we use $\alpha/2 = 0.025$. The critical F-value is $F_{critical}(0.025, 24, 24) \approx 2.27$. Reject H_0 if $F > 2.27$.

Step 4: Make a Decision. Since $2.04 < 2.27$, we fail to reject the null hypothesis.

Step 5: Conclusion. There is not sufficient evidence to conclude a difference in variability between the two machines.

Problem 3.5. A manager wants to know if there is a significant difference in the mean daily sales of three different stores. She collects the sales data (in \$100s) for 5 days from each store. Perform a one-way ANOVA to test the hypothesis at $\alpha = 0.05$.

Store 1	Store 2	Store 3
8	7	12
10	5	9
7	10	13
14	9	12
11	9	14

Solution. We will use a one-way ANOVA to compare the means of the three stores.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3$ (The true mean sales for all three stores are equal).
- Alternative Hypothesis (H_1): At least one of the store means is different.

Step 2: Preliminary Calculations. First, we calculate the sum for each group (column) and the grand total.

- Store 1 Total (C_1): $8 + 10 + 7 + 14 + 11 = 50$
- Store 2 Total (C_2): $7 + 5 + 10 + 9 + 9 = 40$
- Store 3 Total (C_3): $12 + 9 + 13 + 12 + 14 = 60$
- Grand Total (G): $50 + 40 + 60 = 150$
- Number of groups (k)=3. Observations per group (n_j)=5. Total observations (N)=15.

Step 3: Calculate the Correction Factor (CF). The correction factor is used to simplify the SS calculations.

$$CF = \frac{G^2}{N} = \frac{150^2}{15} = \frac{22500}{15} = 1500$$

Step 4: Calculate the Sums of Squares (SS).

- **Total Sum of Squares (SST):** This measures the total variation in the data.

$$SST = \sum x^2 - CF$$

$$\sum x^2 = (8^2+10^2+7^2+14^2+11^2)+(7^2+5^2+10^2+9^2+9^2)+(12^2+9^2+13^2+12^2+14^2)$$

$$\sum x^2 = (530) + (336) + (734) = 1600$$

$$SST = 1600 - 1500 = 100$$

- **Sum of Squares Between Groups (SSB):** This measures the variation *between* the store means.

$$SSB = \sum \frac{C_j^2}{n_j} - CF = \left(\frac{50^2}{5} + \frac{40^2}{5} + \frac{60^2}{5}\right) - 1500$$

$$= \left(\frac{2500}{5} + \frac{1600}{5} + \frac{3600}{5}\right) - 1500 = (500 + 320 + 720) - 1500$$

$$= 1540 - 1500 = 40$$

- **Sum of Squares Within Groups (SSW):** This measures the variation *within* each store's data (error).

$$SSW = SST - SSB = 100 - 40 = 60$$

Step 5: Construct the ANOVA Table. We use the SS values to calculate the mean squares (MS) and the F-statistic.

- Degrees of freedom between: $df_B = k - 1 = 3 - 1 = 2$.
- Degrees of freedom within: $df_W = N - k = 15 - 3 = 12$.
- Mean Square Between: $MSB = SSB/df_B = 40/2 = 20$.
- Mean Square Within: $MSW = SSW/df_W = 60/12 = 5$.
- F-statistic: $F = MSB/MSW = 20/5 = 4.0$.

Source of Variation	**SS**	**df**	**MS**	**F**
Between Groups	40	2	20	4.0
Within Groups (Error)	60	12	5	
Total	100	14		

Step 6: Determine the Critical Value and Make a Decision. For a significance level of $\alpha = 0.05$ with degrees of freedom $df_1 = 2$ (numerator) and $df_2 = 12$ (denominator), we find the critical F-value from an F-distribution table.

$$F_{critical} = 3.89$$

Since our calculated F-statistic ($F = 4.0$) is greater than the critical value ($F_{crit} = 3.89$), we reject the null hypothesis.

Step 7: Conclusion. At the 5% significance level, there is sufficient statistical evidence to conclude that there is a significant difference in the mean daily sales among the three stores.

Problem 3.6. A political polling agency wants to determine if the approval rating of a mayor is different in two districts of a city. In District A, a sample of 200 voters showed 110 approved of the mayor. In District B, a sample of 250 voters showed 120 approved. Test at $\alpha = 0.05$ if there is a significant difference in the approval ratings between the two districts.

Solution. This problem requires a Z-test for the difference between two population proportions.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $p_A = p_B$ (The true approval ratings are the same).
- Alternative Hypothesis (H_1): $p_A \neq p_B$ (The approval ratings are different).

Step 2: Calculate Sample Proportions.

- $\hat{p}_A = 110/200 = 0.55$
- $\hat{p}_B = 120/250 = 0.48$

Step 3: Calculate the Pooled Proportion ($\hat{p}$).

$$\hat{p} = \frac{x_A + x_B}{n_A + n_B} = \frac{110 + 120}{200 + 250} = \frac{230}{450} \approx 0.511$$

Step 4: Calculate the Test Statistic. The formula is $Z = \frac{(\hat{p}_A - \hat{p}_B) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_A} + \frac{1}{n_B})}}$.

$$Z = \frac{0.55 - 0.48}{\sqrt{0.511(1 - 0.511)(\frac{1}{200} + \frac{1}{250})}} = \frac{0.07}{\sqrt{0.511(0.489)(0.005 + 0.004)}} \approx 1.476$$

Step 5: Determine the Rejection Region. For a two-tailed test at $\alpha = 0.05$, the critical Z-values are ± 1.96.

Step 6: Make a Decision. Since $|1.476|$ is not greater than 1.96, we fail to reject the null hypothesis.

Step 7: Conclusion. At the 5% significance level, there is not sufficient evidence to conclude that there is a difference in the mayor's approval ratings between the two districts.

Problem 3.7. A study is conducted to test the effects of fertilizer type (A, B, C) and soil type (Sandy, Clay) on the yield of a crop. Two plots are tested for each combination of factors. The yields are recorded in the table below. Test for significant effects of fertilizer, soil type, and their interaction on yield at $\alpha = 0.05$.

Soil Type	**Fertilizer A**	**Fertilizer B**	**Fertilizer C**
Sandy	6, 8	7, 6	8, 9
Clay	7, 9	8, 9	10, 11

Solution. We perform a two-way ANOVA with replication, which allows us to test the main effects of each factor (Soil, Fertilizer) and their interaction effect.

Step 1: State the Hypotheses.

- For Interaction: $H_{0,int}$: There is no interaction effect between soil and fertilizer. $H_{1,int}$: There is an interaction effect.

- For Soil Type (Rows): $H_{0,R}$: There is no main effect of soil type. $H_{1,R}$: There is a main effect of soil type.
- For Fertilizer (Cols): $H_{0,C}$: There is no main effect of fertilizer type. $H_{1,C}$: There is a main effect of fertilizer type.

Step 2: Preliminary Calculations. We calculate cell, row, column, and grand totals.

- Cell Totals (T_{ij}): $T_{11} = 6 + 8 = 14$, $T_{12} = 7 + 6 = 13$, $T_{13} = 8 + 9 = 17$. $T_{21} = 7 + 9 = 16$, $T_{22} = 8 + 9 = 17$, $T_{23} = 10 + 11 = 21$.
- Row Totals (R_i): R_1(Sandy) $= 14+13+17 = 44$, R_2(Clay) $= 16+17+21 = 54$.
- Column Totals (C_j): C_1(A) $= 14 + 16 = 30$, C_2(B) $= 13 + 17 = 30$, C_3(C) $= 17 + 21 = 38$.
- Grand Total (G): $44 + 54 = 98$.
- $r = 2$ (rows), $c = 3$ (columns), $k = 2$ (replicates). Total observations $N = rck = 12$.

Step 3: Calculate the Correction Factor (CF).

$$CF = \frac{G^2}{N} = \frac{98^2}{12} = \frac{9604}{12} \approx 800.33$$

Step 4: Calculate the Sums of Squares (SS).

- **Total Sum of Squares (SST):**

$$\sum x^2 = 6^2 + 8^2 + 7^2 + 6^2 + 8^2 + 9^2 + 7^2 + 9^2 + 8^2 + 9^2 + 10^2 + 11^2 = 852$$

$$SST = \sum x^2 - CF = 852 - 800.33 = 51.67$$

- **Sum of Squares for Rows (SSR):**

$$SSR = \frac{\sum R_i^2}{ck} - CF = \frac{44^2 + 54^2}{3 \times 2} - 800.33 = \frac{1936 + 2916}{6} - 800.33 = 808.67 - 800.33 = 8.34$$

- **Sum of Squares for Columns (SSC):**

$$SSC = \frac{\sum C_j^2}{rk} - CF = \frac{30^2 + 30^2 + 38^2}{2 \times 2} - 800.33 = \frac{900 + 900 + 1444}{4} - 800.33 = 811 - 800.33 =$$

- **Sum of Squares for Subtotals (Cells):**

$$SS(\text{Cells}) = \frac{\sum T_{ij}^2}{k} - CF = \frac{14^2 + 13^2 + 17^2 + 16^2 + 17^2 + 21^2}{2} - 800.33 = 19.67$$

- **Sum of Squares for Interaction (SSRC):**

$$SSRC = SS(\text{Cells}) - SSR - SSC = 19.67 - 8.34 - 10.67 = 0.66$$

- **Sum of Squares for Error (SSE):**

$$SSE = SST - SS(\text{Cells}) = 51.67 - 19.67 = 32.00$$

Step 5: Construct the ANOVA Table.

- $df_R = r - 1 = 1$, $df_C = c - 1 = 2$, $df_{RC} = (r-1)(c-1) = 2$, $df_E = rc(k-1) = 6$.
- $MSR = SSR/df_R = 8.34/1 = 8.34$.
- $MSC = SSC/df_C = 10.67/2 = 5.335$.
- $MSRC = SSRC/df_{RC} = 0.66/2 = 0.33$.
- $MSE = SSE/df_E = 32.00/6 \approx 5.333$.

Source	**SS**	**df**	**MS**	**F**
Soil (Rows)	8.34	1	8.34	1.56
Fertilizer (Cols)	10.67	2	5.335	1.00
Interaction	0.66	2	0.33	0.06
Error	32.00	6	5.333	
Total	51.67	11		

Step 6: Determine Critical Values and Make Decisions ($\alpha = 0.05$).

- **Interaction Effect:** $df = (2, 6)$, $F_{crit} = 5.14$. Calculated $F_{int} = 0.06$. Since $0.06 < 5.14$, we fail to reject $H_{0,int}$. There is **no significant interaction effect**.
- **Soil Effect (Rows):** $df = (1, 6)$, $F_{crit} = 5.99$. Calculated $F_R = 1.56$. Since $1.56 < 5.99$, we fail to reject $H_{0,R}$. There is **no significant main effect** of soil type.
- **Fertilizer Effect (Cols):** $df = (2, 6)$, $F_{crit} = 5.14$. Calculated $F_C = 1.00$. Since $1.00 < 5.14$, we fail to reject $H_{0,C}$. There is **no significant main effect** of fertilizer type.

Step 7: Conclusion. At the 5% significance level, we conclude that there is no significant interaction between soil type and fertilizer type. Furthermore, there is not sufficient evidence to conclude that either soil type or fertilizer type has a significant main effect on the crop yield.

Problem 3.8. A lab is testing a new analytical instrument. They measure a standard sample with a known concentration of 10.0 mg/L. They perform 10 measurements and get a sample mean of 10.2 mg/L with a sample standard deviation of 0.3 mg/L. Is there evidence at the 5% significance level that the instrument has a systematic bias (i.e., its mean measurement is different from 10.0)?

Solution. This requires a one-sample t-test.

Step 1: State the Hypotheses.

- Null Hypothesis (H_0): $\mu = 10.0$ (No bias).
- Alternative Hypothesis (H_1): $\mu \neq 10.0$ (Bias exists).

Step 2: Calculate the Test Statistic.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{10.2 - 10.0}{0.3/\sqrt{10}} \approx 2.108$$

Step 3: Determine the Rejection Region. $df = n - 1 = 9$. For a two-tailed test at $\alpha = 0.05$, $t_{critical} = \pm 2.262$.

Step 4: Make a Decision. Since $|2.108| < 2.262$, the test statistic does not fall in the rejection region.

Step 5: Conclusion. At the 5% significance level, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that the instrument has a systematic bias.

Problem 3.9. A political poll of 1200 voters is conducted. 630 voters say they will vote for Candidate A. Construct a 95% confidence interval for the true proportion of voters who support Candidate A.

Solution. We construct a confidence interval for a population proportion.

Step 1: Calculate the sample proportion ($\hat{p}$).

$$\hat{p} = \frac{x}{n} = \frac{630}{1200} = 0.525$$

Step 2: Find the critical Z-value. For 95% confidence, $Z_{\alpha/2} = 1.96$.

Step 3: Calculate the Margin of Error (ME).

$$ME = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.525(0.475)}{1200}} \approx 0.0282$$

Step 4: Construct the Confidence Interval.

$$\text{CI} = \hat{p} \pm ME = 0.525 \pm 0.0282 = (0.4968, 0.5532)$$

Step 5: Conclusion. We are 95% confident that the true proportion of voters who support Candidate A is between **49.68% and 55.32%**.

4 Regression and Correlation

This section deals with modeling the relationship between variables, from simple linear regression to multiple and non-linear models.

Problem 4.1. Given the following data on advertising expenditure (X, in \$1000s) and sales revenue (Y, in \$10,000s):

X	2	3	5	6	8
Y	5	7	10	11	15

(a) Compute the least squares regression line $\hat{y} = \beta_0 + \beta_1 x$.

(b) Calculate the coefficient of determination, r^2, and interpret its value.

(c) Predict the sales revenue if the advertising expenditure is \$7,000.

Solution. (a) Least Squares Regression Line

Step 1: Calculate necessary sums.

- $n = 5$, $\sum x_i = 24$, $\sum y_i = 48$
- $\bar{x} = 4.8$, $\bar{y} = 9.6$
- $\sum x_i y_i = 267$, $\sum x_i^2 = 138$

Step 2: Calculate the slope, β_1.

$$\beta_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} = \frac{5(267) - (24)(48)}{5(138) - (24)^2} = \frac{183}{114} \approx 1.605$$

Step 3: Calculate the intercept, β_0.

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 9.6 - (1.605)(4.8) \approx 1.896$$

Step 4: State the regression line.

$$\hat{y} = 1.896 + 1.605x$$

(b) Coefficient of Determination (r^2)

Step 1: Calculate r^2.

$$r^2 = \frac{[\,n(\sum xy) - (\sum x)(\sum y)\,]^2}{[\,n\sum x^2 - (\sum x)^2\,][\,n\sum y^2 - (\sum y)^2\,]}$$

We need $\sum y_i^2 = 5^2 + 7^2 + 10^2 + 11^2 + 15^2 = 520$.

$$r^2 = \frac{(183)^2}{(114)(5(520) - 48^2)} = \frac{33489}{(114)(296)} \approx 0.9924$$

Step 2: Interpretation. An r^2 of 0.9924 means that approximately **99.24% of the total variation** in sales revenue can be explained by the linear relationship with advertising expenditure.

(c) Predict Sales Revenue

Step 1: Use the regression equation with $x = 7$.

$$\hat{y} = 1.896 + 1.605(7) \approx 13.131$$

Step 2: Conclusion. The predicted sales revenue is 13.131 in units of \$10,000, which is **\$131,310**.

Problem 4.2. A researcher believes the relationship between chemical concentration (x) and reaction rate (y) is $y = \alpha e^{\beta x}$. Data:

x	1	2	3	4
y	3.1	8.2	22.1	60.3

Transform the variables to linearize the model and find the estimates for α and β.

Solution.

Step 1: Linearize the model. Take the natural log of both sides:

$$\ln(y) = \ln(\alpha) + \beta x$$

This is a linear model $y' = \beta_0 + \beta_1 x'$, where $y' = \ln(y)$, $x' = x$, $\beta_0 = \ln(\alpha)$, and $\beta_1 = \beta$.

Step 2: Transform the data and perform linear regression.

- Transformed y-values, $y' = \ln(y)$: 1.131, 2.104, 3.096, 4.099.
- Using linear regression calculations on (x, y'), we find the slope and intercept.
- Slope: $\hat{\beta}_1 \approx 0.99$.
- Intercept: $\hat{\beta}_0 \approx 0.1335$.

Step 3: Estimate the original parameters.

- $\hat{\beta} = \hat{\beta}_1 \approx 0.99$
- $\hat{\alpha} = e^{\hat{\beta}_0} = e^{0.1335} \approx 1.143$

Step 4: Conclusion. The estimated non-linear model is $y = 1.143e^{0.99x}$.

Problem 4.3. A multiple linear regression model is $\hat{Y} = 50 + 0.15X_1 - 2.5X_2$ where Y is price (\$1000s), X1 is size (sq. ft.), and X2 is age (years).

(a) Predict the price of a 2000 sq. ft. house that is 10 years old.

(b) For houses of the same age, what is the expected price difference for a 100 sq. ft. increase in size?

Solution. (a) Predict the price of a specific house Plug $X_1 = 2000$ and $X_2 = 10$ into the model:

$$\hat{Y} = 50 + 0.15(2000) - 2.5(10) = 50 + 300 - 25 = 325$$

The predicted price of the house is \$325,000.

(b) Expected difference in price The coefficient $\beta_1 = 0.15$ gives the change in price for a one sq. ft. change in size. For a 100 sq. ft. increase, the change is:

$$\text{Expected Difference} = 100 \times \beta_1 = 100 \times 0.15 = 15$$

The expected price increase is \$15,000.

Problem 4.4. For a logistic regression model, $\ln\left(\frac{P}{1-P}\right) = -2.5 + 0.05 \times \text{Age}$, find the probability of the condition for a person who is 50 years old.

Solution.

Step 1: Calculate the log-odds (logit) for Age = 50.

$$\text{Log-odds} = -2.5 + 0.05(50) = 0$$

Step 2: Convert log-odds to probability. The formula is $P = \frac{e^{\text{log-odds}}}{1+e^{\text{log-odds}}}$.

$$P = \frac{e^0}{1+e^0} = \frac{1}{1+1} = 0.5$$

The probability of a 50-year-old having the condition is 50%.

Problem 4.5. A regression of crop yield on rainfall for 20 years of data yields SSE = 150 and SST = 600. The model is $\hat{Y} = 2.5 + 0.8X$.

(a) Calculate the coefficient of determination, r^2.

(b) Calculate the estimate of the model variance, s^2.

(c) What is the value of the sample correlation coefficient, r?

Solution. (a) Coefficient of Determination (r^2)

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{150}{600} = 0.75$$

(b) Estimate of Model Variance (s^2)

$$s^2 = \frac{SSE}{n-2} = \frac{150}{20-2} = \frac{150}{18} \approx 8.33$$

(c) Sample Correlation Coefficient (r) $r = \sqrt{r^2} = \sqrt{0.75} \approx 0.866$. The sign is positive because the slope (0.8) is positive.

Problem 4.6. For the model $\hat{y} = 1.896 + 1.605x$ (from Problem 4.1, n=5), test if the slope is significantly different from zero at $\alpha = 0.05$. The standard error of the slope, $SE(\beta_1)$, is 0.15.

Solution.

Step 1: State the Hypotheses. $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

Step 2: Calculate the Test Statistic.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{1.605}{0.15} = 10.7$$

Step 3: Determine the Rejection Region. $df = n - 2 = 3$. For a two-tailed test at $\alpha = 0.05$, $t_{critical} = \pm 3.182$.

Step 4: Make a Decision. Since $|10.7| > 3.182$, we reject the null hypothesis.

Step 5: Conclusion. There is a significant linear relationship between advertising and sales.

Problem 4.7. For the regression model in Problem 4.1 ($\hat{y} = 1.896 + 1.605x$), calculate the residual for the data point ($x = 5, y = 10$).

Solution.

Step 1: Calculate the Predicted Value ($\hat{y}$).

$$\hat{y} = 1.896 + 1.605(5) = 1.896 + 8.025 = 9.921$$

Step 2: Calculate the Residual.

$$e = y - \hat{y} = 10 - 9.921 = 0.079$$

The residual is 0.079.

5 Time Series and Forecasting

This section covers methods for analyzing data points collected over time, including decomposition, smoothing, and building predictive models like ARIMA.

Problem 5.1. Given the following quarterly sales data (in \$ thousands):

Year	**2023**				**2024**			
Quarter	1	2	3	4	1	2	3	4
Sales	210	250	230	290	240	280	260	320

(a) Calculate the 4-quarter centered moving average.

(b) Estimate the seasonal factors for each quarter using a multiplicative model.

Solution. (a) 4-Quarter Centered Moving Average

Step 1: Calculate 4-Quarter Moving Averages.

- MA1 (Q1-Q4 '23): $(210 + 250 + 230 + 290)/4 = 245.0$
- MA2 (Q2 '23-Q1 '24): $(250 + 230 + 290 + 240)/4 = 252.5$
- MA3 (Q3 '23-Q2 '24): $(230 + 290 + 240 + 280)/4 = 260.0$
- MA4 (Q4 '23-Q3 '24): $(290 + 240 + 280 + 260)/4 = 267.5$
- MA5 (Q1-Q4 '24): $(240 + 280 + 260 + 320)/4 = 275.0$

Step 2: Center the Averages (CMA).

- Q3 2023: $(245.0 + 252.5)/2 = 248.75$
- Q4 2023: $(252.5 + 260.0)/2 = 256.25$
- Q1 2024: $(260.0 + 267.5)/2 = 263.75$
- Q2 2024: $(267.5 + 275.0)/2 = 271.25$

(b) Estimate Seasonal Factors

Step 1: Calculate Seasonal Ratios (Sales / CMA).

- Q3 2023: $230/248.75 \approx 0.9246$
- Q4 2023: $290/256.25 \approx 1.1317$
- Q1 2024: $240/263.75 \approx 0.9099$
- Q2 2024: $280/271.25 \approx 1.0323$

Step 2: Average and Normalize the Factors. Initial Sum $= 0.9099 + 1.0323 + 0.9246 + 1.1317 = 3.9985$. Normalization constant $= 4/3.9985 \approx 1.000375$.

Step 3: Final Seasonal Factors.

- Q1: $0.9099 \times 1.000375 \approx 0.910$
- Q2: $1.0323 \times 1.000375 \approx 1.033$
- Q3: $0.9246 \times 1.000375 \approx 0.925$
- Q4: $1.1317 \times 1.000375 \approx 1.132$

Problem 5.2. The number of weekly visitors over the last 5 weeks was: 1500, 1750, 1600, 1800, 1700.

(a) Forecast visitors for week 6 using exponential smoothing with $\alpha = 0.3$.

(b) If actual visitors in week 6 was 1850, what is the forecast for week 7?

Solution. The formula is $F_{t+1} = \alpha Y_t + (1-\alpha)F_t$. We initialize with $F_2 = Y_1 = 1500$.

(a) Forecast for Week 6

- $F_3 = 0.3(1750) + 0.7(1500) = 1575$
- $F_4 = 0.3(1600) + 0.7(1575) = 1582.5$
- $F_5 = 0.3(1800) + 0.7(1582.5) = 1647.75$
- $F_6 = 0.3(1700) + 0.7(1647.75) \approx 1663.43$

The forecast for week 6 is approximately **1663** visitors.

(b) Forecast for Week 7 Using $Y_6 = 1850$ and $F_6 \approx 1663.43$:

$$F_7 = 0.3(1850) + 0.7(1663.43) \approx 555 + 1164.4 \approx 1719.4$$

The forecast for week 7 is approximately **1719** visitors.

Problem 5.3. A stationary time series model is $Y_t = 10 + 0.7Y_{t-1} - 0.2Y_{t-2} + \epsilon_t$. If $Y_{t-1} = 12$ and $Y_{t-2} = 11$, what is the forecast for Y_t?

Solution. This is an Autoregressive model of order 2, AR(2). The forecast $\hat{Y}_t$ is the expected value of Y_t, and the expected value of the error term ϵ_t is 0.

$$\hat{Y}_t = 10 + 0.7Y_{t-1} - 0.2Y_{t-2}$$

$$\hat{Y}_t = 10 + 0.7(12) - 0.2(11) = 10 + 8.4 - 2.2 = 16.2$$

The forecast for Y_t is **16.2**.

Problem 5.4. Given the time series data: 2, 5, 10, 17, 26, 37, perform first-order differencing.

Solution. First-order differencing calculates the change between consecutive observations, $Z_t = Y_t - Y_{t-1}$.

- $Z_2 = 5 - 2 = 3$
- $Z_3 = 10 - 5 = 5$
- $Z_4 = 17 - 10 = 7$
- $Z_5 = 26 - 17 = 9$
- $Z_6 = 37 - 26 = 11$

The first-differenced series is **3, 5, 7, 9, 11**.

Problem 5.5. The ACF and PACF plots of a stationary (differenced) time series are examined.

- The ACF plot shows a significant spike at lag 1 and then cuts off to zero.
- The PACF plot shows a pattern of exponential decay.

Based on these plots, what ARIMA(p,d,q) model would you suggest?

Solution. The patterns (ACF cuts off at lag 1, PACF tails off) are the classic signs of a Moving Average process of order 1, i.e., **MA(1)**. Since this is for the *differenced* data, the parameters for the stationary series are $p = 0$ and $q = 1$. If we assume one difference was taken to achieve stationarity ($d = 1$), the suggested model for the original series is **ARIMA(0,1,1)**.

6 Gaussian Mixture Models (GMM) and EM

This section covers probabilistic clustering using GMMs and the Expectation-Maximization algorithm used to fit them.

Problem 6.1. A GMM with two components is defined by:

- Comp 1: $\mu_1 = 5, \sigma_1 = 1, \pi_1 = 0.6$
- Comp 2: $\mu_2 = 10, \sigma_2 = 2, \pi_2 = 0.4$

Perform the E-step for a single data point $x_i = 6$ by calculating its responsibilities.

Solution. The E-step calculates the responsibility of each component k for data point i, $\gamma(z_{ik})$.

Step 1: Calculate the likelihood of the data point under each component.

- Comp 1: $\mathcal{N}_1 = \frac{1}{\sqrt{2\pi(1)}} e^{-\frac{(6-5)^2}{2(1)}} \approx 0.2420$
- Comp 2: $\mathcal{N}_2 = \frac{1}{\sqrt{2\pi(4)}} e^{-\frac{(6-10)^2}{2(4)}} \approx 0.0270$

Step 2: Calculate the weighted likelihoods.

- Numerator 1: $\pi_1 \mathcal{N}_1 = 0.6 \times 0.2420 = 0.1452$
- Numerator 2: $\pi_2 \mathcal{N}_2 = 0.4 \times 0.0270 = 0.0108$

Step 3: Calculate the total evidence (denominator). Denom $= 0.1452 + 0.0108 = 0.1560$.

Step 4: Calculate the final responsibilities.

- $\gamma(z_{i1})$ (Resp. of Comp 1): $\frac{0.1452}{0.1560} \approx 0.9308$
- $\gamma(z_{i2})$ (Resp. of Comp 2): $\frac{0.0108}{0.1560} \approx 0.0692$

Step 5: Conclusion. Responsibility for Comp 1 is **93.08%**; for Comp 2 is **6.92%**.

Problem 6.2. After an E-step, the responsibilities for a dataset $\{2, 8, 9\}$ are:

Data Point (x_i)	**Resp. Comp 1 (γ_{i1})**	**Resp. Comp 2 (γ_{i2})**
2	0.9	0.1
8	0.2	0.8
9	0.1	0.9

Perform the M-step: update all parameters (μ_1, σ_1^2, π_1 and μ_2, σ_2^2, π_2).

Solution.

Step 1: Calculate effective number of points (N_k).

Step 2: $N_1 = 0.9 + 0.2 + 0.1 = 1.2$

Step 3: $N_2 = 0.1 + 0.8 + 0.9 = 1.8$

Step 4: Update the mixing coefficients (π_k).

Step 5: $\pi_1^{new} = \frac{1.2}{3} = 0.4$

Step 6: $\pi_2^{new} = \frac{1.8}{3} = 0.6$

Step 7: **Update the means (μ_k).**

Step 8: $\mu_1^{new} = \frac{1}{1.2}(0.9 \times 2 + 0.2 \times 8 + 0.1 \times 9) = \frac{4.3}{1.2} \approx 3.583$

Step 9: $\mu_2^{new} = \frac{1}{1.8}(0.1 \times 2 + 0.8 \times 8 + 0.9 \times 9) = \frac{14.7}{1.8} \approx 8.167$

Step 10: **Update the variances (σ_k^2).**

Step 11: $(\sigma_1^2)^{new} = \frac{1}{1.2}[0.9(2-3.583)^2 + 0.2(8-3.583)^2 + 0.1(9-3.583)^2] \approx 7.576$

Step 12: $(\sigma_2^2)^{new} = \frac{1}{1.8}[0.1(2-8.167)^2 + 0.8(8-8.167)^2 + 0.9(9-8.167)^2] \approx 2.472$

Step 13: **Conclusion.** The updated parameters are:

- Component 1: $\pi_1 = 0.4$, $\mu_1 \approx 3.583$, $\sigma_1^2 \approx 7.576$
- Component 2: $\pi_2 = 0.6$, $\mu_2 \approx 8.167$, $\sigma_2^2 \approx 2.472$

Problem 6.3. A GMM has been fit to a 2D dataset with two components. How would you use this model for anomaly detection?

Solution. A trained GMM learns the probability density of "normal" data. Anomalies can be detected as points with a very low probability density under this model.

Step 1: **Train the GMM** on a dataset consisting of normal, non-anomalous data points.

Step 2: **Calculate the Probability Density for a New Point.** For any new data point, x_{new}, calculate its probability density under the fitted GMM:

$$p(x_{new}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_{new} | \mu_k, \Sigma_k)$$

Step 3: **Set a Probability Threshold (ϵ).** Establish a threshold, ϵ. Any point where $p(x_{new}) < \epsilon$ is classified as an anomaly. This threshold is typically determined using a validation set to balance false positives and false negatives.

Problem 6.4. What is the role of the covariance matrix in a GMM? Describe what is implied by using a spherical, diagonal, or full covariance matrix for the components.

Solution.

Step 1: **Role of the Covariance Matrix.** In a GMM, the covariance matrix (Σ_k) for each component k describes the **shape, orientation, and size** of the cluster.

Step 2: **Spherical Covariance.** Implies the clusters are hyperspherical (circular in 2D), not elongated or rotated.

Step 3: **Diagonal Covariance.** Implies the clusters are axis-aligned ellipses, stretched along the feature axes but not rotated.

Step 4: **Full Covariance.** The most flexible option. It implies clusters can have any elliptical shape, size, and orientation (rotation).