Statistics Homework Assignment

Based on Lectures 1, 2, and 3 by Sulaiman Ahmed

Question 1: Basic Measures of Central Tendency and Dispersion (Beginner)

Points: 10

A retail store manager collected data on the number of customers per day for the past week: **Data:** 45, 52, 38, 47, 55, 41, 49

Part A (4 points): Calculate the mean, median, and mode for this dataset.

Part B (3 points): Calculate the sample variance and sample standard deviation. Show your work step by step.

Part C (3 points): Explain which measure of central tendency (mean, median, or mode) would be most appropriate if there was an outlier day with 150 customers, and why?

Question 2: Population vs Sample and Variance Division (Beginner-Intermediate)

Points: 12

A quality control manager wants to analyze the weight of chocolate bars produced in a factory. The population of all chocolate bars has a mean weight of $\mu = 100$g and standard deviation $\sigma = 5$g.

Part A (4 points): Explain the difference between population and sample in this context. Why might the manager choose to work with a sample rather than the entire population?

Part B (5 points): The manager takes a sample of 8 chocolate bars with weights: 98, 102, 95, 105, 99, 103, 97, 101g. Calculate the sample variance using the formula with (n-1) in the denominator.
Part C (3 points): Explain why we divide by (n-1) instead of n when calculating sample variance. What statistical concept does this address?

Question 3: Variables and Measurement Scales (Beginner-Intermediate)

Points: 15

A university is conducting a student satisfaction survey with the following variables:

1. Student ID number
2. Age in years
3. Major field of study
4. Satisfaction rating (Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied)
5. Monthly income in dollars
6. Number of courses enrolled
7. GPA on a 4.0 scale

Part A (7 points): For each variable above, identify whether it is:

Quantitative or Qualitative
If quantitative, whether it's discrete or continuous
The measurement scale (Nominal, Ordinal, Interval, or Ratio)

Part B (4 points): Create a frequency distribution table for the satisfaction rating data: Very Unsatisfied: 5 students, Unsatisfied: 12 students, Neutral: 25 students, Satisfied: 35 students, Very Satisfied: 23 students

Part C (4 points): Calculate the cumulative frequency for the satisfaction rating data and explain what the cumulative frequency tells us.

Question 4: Distribution Types and Skewness (Intermediate)

Points: 18

An online learning platform analyzed the completion times (in minutes) for a statistics course. They found three different patterns in their data:

Dataset A: Mean = 120 min, Median = 115 min, Mode = 110 min
Dataset B: Mean = 85 min, Median = 90 min, Mode = 95 min
Dataset C: Mean = 95 min, Median = 95 min, Mode = 95 min

Part A (9 points): For each dataset, determine the type of skewness (left skewed, right skewed, or symmetrical) and explain your reasoning based on the relationship between mean, median, and mode.

Part B (6 points): Give two real-world examples of data that would naturally follow each type of distribution (left skewed, right skewed, symmetrical) and explain why.

Part C (3 points): Which measure of central tendency would be most reliable for making business decisions in each of the three datasets? Justify your answer.

Question 5: Z-Scores and Standardization (Intermediate)

Points: 13

A university professor finds that exam scores follow a normal distribution with $\mu = 78$ and $\sigma = 12$.

Part A (8 points): Calculate the z-scores for the following students:

Student A: 90 points
Student B: 65 points
Student C: 78 points
Student D: 95 points

Show your calculations and interpret what each z-score means.

Part C (5 points): If the professor wants to convert all scores to a standard normal distribution (mean = 0, standard deviation = 1), what would be the new standardized scores for Students A and B? Explain the advantages of standardization in this context.

Question 6: Outlier Detection Using Z-Scores (Intermediate)

Points: 25

A data analyst is examining daily website traffic for an e-commerce company. The data for the past 15 days shows the following visitor counts:

Data: 2400, 2350, 2500, 2420, 2380, 2450, 2390, 4200, 2460, 2430, 2510, 2370, 2440, 2480, 2400

Part A (8 points):

Calculate the mean and standard deviation for this dataset
Calculate the z-score for each data point
Show your calculations for at least 3 data points

Part B (8 points):

Identify any outliers using the z-score method (threshold of $|z| > 3$)
If no outliers exist using $|z| > 3$, use $|z| > 2$ as the threshold
Explain why this data point(s) might be considered an outlier

Part C (9 points):

Calculate the mean and standard deviation after removing the outlier(s)
Discuss how the outlier affected the original mean and standard deviation
Suggest two possible real-world explanations for why this outlier might have occurred in website traffic data
Should the outlier be removed or investigated further? Justify your recommendation.

Question 7: Comprehensive Application - Five Number Summary and Data Analysis (Advanced Intermediate)

Points: 30

A fitness app company collected data on daily steps for 20 users over a month. Here's a sample of daily step counts:
Data: 8500, 12000, 6500, 15500, 9200, 11800, 7300, 13400, 10500, 8900, 14200, 5800, 12600, 9800, 16500, 7800, 11200, 8400, 13800, 10200

Part A: Five Number Summary (10 points)

Calculate the five number summary (Min, Q1, Median, Q3, Max)

Show your step-by-step calculation for Q1 and Q3
Calculate the Interquartile Range (IQR)

Part B: Outlier Detection using IQR Method (8 points)

Use the IQR method to identify outliers (using 1.5 × IQR rule)
Calculate the lower and upper bounds for outliers
Identify any outliers in the dataset

Part C: Standardization and Z-Score Analysis (12 points)

Calculate the z-scores for all data points that you identified as outliers (if any) or the 3 highest values
Compare the IQR method results with z-score method results (using $|z| > 2$ threshold)
Create a brief analysis report discussing:
The distribution characteristics of the data
Which outlier detection method seems more appropriate for this dataset and why Practical implications for the fitness app company (should they investigate these users, are these realistic values, etc.)

Submission Guidelines

1. **Show all calculations** - Partial credit will be given for correct methodology even if final answers are incorrect
2. **Round final answers** to 2 decimal places unless otherwise specified
3. **Provide clear explanations** for conceptual questions
4. **Use proper statistical notation** (μ, σ, $\bar{x}$, s, etc.)
5. **Include units** where applicable

This assignment covers key concepts from Statistics Lectures 1, 2, and 3, focusing on descriptive statistics, measures of central tendency and dispersion, variable types, normal distribution, z-scores,
standardization, and outlier detection methods.