Statistics Basics Assignment

Q1: Explain the different types of data (qualitative and quantitative) and provide examples of each. Discuss nominal, ordinal, interval, and ratio scales.

There are two main types of data:

1. Qualitative (Categorical): This type of data describes characteristics or qualities. It can't be measured in numbers.
    - Example: Eye color (blue, brown), Gender (Male, Female), Types of music (Rock, Pop).
2. Quantitative (Numerical): This data is measurable and expressed in numbers.
    - Example: Height (170 cm), Age (25 years), Temperature (30°C).

Scales of Measurement:

- Nominal: Just labels, no order. (Example: Types of cars – Sedan, SUV, Hatchback).
- Ordinal: Has order, but the difference isn't meaningful. (Example: Ratings – Poor, Average, Good).
- Interval: Ordered, equal gaps, but no true zero. (Example: Temperature in Celsius, IQ scores).
- Ratio: Like interval but has a true zero. (Example: Weight, Salary, Distance).

Q2: What are the measures of central tendency, and when should you use each? Discuss the mean, median, and mode with examples and situations where each is appropriate.

There are three main measures of central tendency:

1. Mean (Average): Add all values and divide by the number of values.
    - Example: (10 + 20 + 30) / 3 = 20
    - Use when: Data is normally distributed (e.g., average marks in a class).
2. Median (Middle value): Arrange data in order and pick the middle.
    - Example: [10, 20, 30, 40, 50] → Median = 30
    - Use when: Data is skewed (e.g., house prices).
3. Mode (Most frequent value): The value that appears most.
    - Example: [2, 3, 3, 5, 7] → Mode = 3
    - Use when: Data is categorical (e.g., most common shoe size).

Q3: Explain the concept of dispersion. How do variance and standard deviation measure the spread of data?

Dispersion tells us how spread out the data is. Two key measures:

- Variance: It's the average squared difference from the mean. Higher variance = more spread out.
- Standard Deviation (SD): Square root of variance, making it easier to understand. Low SD = Data is close to the mean. High SD = Data is more spread out.

Example:

- Data: [10, 20, 30] → SD is small (values close).
- Data: [10, 50, 100] → SD is large (values spread out).

Q4: What is a box plot, and what can it tell you about the distribution of data?

A box plot (or whisker plot) helps visualize data distribution. It shows:

- Min & Max (whiskers) → Smallest & largest values.
- Q1 & Q3 (Box edges) → 25th & 75th percentile.
- Median (Line inside the box) → Middle value.
- Outliers (Dots outside whiskers) → Extreme values.

A skewed box plot means the data isn't evenly spread.

Q5: Discuss the role of random sampling in making inferences about populations.

Random sampling helps get unbiased data from a large population. It's used in surveys, research, and experiments to make conclusions about a whole group.

Example:

- Election polls randomly survey 1000 people to predict national results.
- A doctor tests a new medicine on a random group to check effectiveness.

More random = more accurate results!

Q6: Explain the concept of skewness and its types. How does skewness affect the interpretation of data?

Skewness tells if data is symmetrical or lopsided.

1. Positive Skew (Right-skewed):
   - Tail is longer on the right.
   - Example: Income distribution (few rich people pull the avg up).
   - Mean > Median
2. Negative Skew (Left-skewed):
   - Tail is longer on the left.
   - Example: Exam scores (most students score high, a few very low).
   - Mean < Median

Skewness affects which measure of central tendency is best!

Q7: What is the interquartile range (IQR), and how is it used to detect outliers?

IQR = Q3 - Q1 (middle 50% of data). It helps find outliers using:

- Lower Bound = Q1 - 1.5 × IQR
- Upper Bound = Q3 + 1.5 × IQR

Values outside these bounds are outliers!

Example: If IQR = 20, then outliers are anything 1.5 × 20 = 30 beyond Q1 and Q3.

Q8: Discuss the conditions under which the binomial distribution is used.

Used when:

- Fixed no. of trials (n)
- Two possible outcomes (Success/Failure)
- Same probability in each trial (p)

Example:

- Tossing a coin 10 times.
- Checking if a product is defective (Yes/No).

Q9: Explain the properties of the normal distribution and the empirical rule (68-95-99.7 rule).

The normal distribution is a bell-shaped, symmetric curve.

Empirical Rule:

- 68% of data lies within 1 SD of the mean.
- 95% lies within 2 SD.
- 99.7% lies within 3 SD.

Example: IQ scores are normally distributed (most people have an average IQ).

Q10: Provide a real-life example of a Poisson process and calculate the probability for a specific event.

Poisson is used when events happen randomly over time.

Example: A customer service center gets 5 calls per hour. What’s the chance of exactly 3 calls in an hour?

Formula:

P(3)=e−5×533!P(3) = \frac{e^{-5} \times 5^3}{3!}

= 0.14 (14% probability).

Q11: Explain what a random variable is and differentiate between discrete and continuous random variables.

A random variable is a numerical outcome of an experiment.

- Discrete: Countable values (e.g., number of students in class).
- Continuous: Any value in a range (e.g., weight, height).

Example:

- Rolling a die: Discrete (only 1, 2, 3, 4, 5, 6).
- Temperature: Continuous (can be 36.5°C, 36.55°C, etc.).

Q12: Provide an example dataset, calculate both covariance and correlation, and interpret the results.

Dataset:
X = [2, 4, 6, 8]
Y = [3, 6, 7, 10]

Covariance = 4.67 (Positive → when X increases, Y increases).
Correlation = 0.97 (Strong positive relationship).

Higher correlation means a stronger connection between two variables.