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

Answer:

Data can be broadly categorized into two types:

- **Qualitative (Categorical) Data**: Describes characteristics or qualities that cannot be measured with numbers.
  - *Example*: Colors, gender, nationality.
  - **Nominal Scale**: Categories without a meaningful order.
    - *Example*: Hair color (black, brown, blonde).
  - **Ordinal Scale**: Categories with a meaningful order but no consistent difference between levels.
    - *Example*: Movie ratings (poor, average, good, excellent).

- **Quantitative (Numerical) Data**: Represents measurable quantities.
  - *Example*: Height, weight, temperature.
  - **Interval Scale**: Numeric scale with equal intervals, but no true zero.
    - *Example*: Temperature in Celsius.
  - **Ratio Scale**: Numeric scale with equal intervals and a true zero.
    - *Example*: Weight, income.

---

2. 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.

Answer:

Measures of central tendency describe the center of a data set:

- **Mean** (Average): Sum of all values divided by the number of values.
  - *Use when*: Data is symmetrically distributed.
  - *Example*: Average score in an exam.

- **Median**: Middle value when data is ordered.
  - *Use when*: Data contains outliers or is skewed.
  - *Example*: Median salary to avoid distortion by very high incomes.

- **Mode**: Most frequently occurring value.
  - *Use when*: Analyzing categorical data.
  - *Example*: Most common shoe size sold in a store.

---

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

Answer:

Dispersion measures how much the data values vary from the central tendency.

- **Variance**: Average of the squared differences from the mean. A higher variance indicates more spread.
- **Standard Deviation**: Square root of variance; gives spread in original units.
  - *Example*: If the average test score is 70 with a high standard deviation, students' scores vary greatly.

---

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

Answer:

A **box plot** (box-and-whisker plot) is a graphical representation of the distribution of data through five-number summary: minimum, Q1, median (Q2), Q3, and maximum.

It shows:
- Central tendency (median)
- Spread (IQR)
- Skewness
- Outliers (as individual points beyond whiskers)

---

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

Answer:

Random sampling ensures each member of a population has an equal chance of being selected. It helps:
- Avoid bias
- Make valid generalizations
- Support the accuracy of statistical inference
Example: Using a random sample of voters to predict election outcomes.

---

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

Answer:

Skewness measures asymmetry in the distribution of data.

- **Positive skew**: Tail on the right; mean > median.
- **Negative skew**: Tail on the left; mean < median.
- **Zero skew**: Symmetrical distribution; mean = median.

Effect: Skewed data can mislead mean values; median is more reliable in such cases.

---

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

Answer:

IQR = Q3 - Q1

It measures the spread of the middle 50% of data.

Outlier detection:
- Lower bound = Q1 - 1.5 × IQR
- Upper bound = Q3 + 1.5 × IQR
Values outside these bounds are considered outliers.

---

8. Discuss the conditions under which the binomial distribution is used.

Answer:

Binomial distribution applies when:
- Fixed number of trials (n)
- Two possible outcomes (success/failure)
- Constant probability of success (p)
- Independent trials

Example: Tossing a coin 10 times and counting heads.

---

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

Answer:

Normal distribution is a bell-shaped, symmetric distribution.

Properties:
- Mean = median = mode
- Symmetrical about the mean

Empirical Rule:
- 68% of data within 1 SD of mean
- 95% within 2 SDs
- 99.7% within 3 SDs

Used for predicting probabilities in normal data.

---

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

Answer:

Example: Number of emails received per hour.

If the average is λ = 4 emails/hour, what is the probability of receiving exactly 2 emails?

Using the Poisson formula:
\[P(X = k) = \frac{e^{-λ} λ^k}{k!} \]

\[P(2) = \frac{e^{-4} × 4^2}{2!} = \frac{e^{-4} × 16}{2} ≈ \frac{0.0183 × 16}{2} = 0.146 \]

Answer: Approximately 14.6% chance of receiving exactly 2 emails.

---

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

Answer:

A **random variable** is a numerical outcome of a random process.

- **Discrete**: Finite/countable outcomes.
  - *Example*: Number of cars in a parking lot.

- **Continuous**: Infinite/unmeasurable outcomes within a range.
  - *Example*: Weight of a person.

---

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

Answer:

Dataset:
X: [2, 4, 6, 8]
Y: [1, 3, 5, 7]

Step 1: Means
Mean(X) = 5, Mean(Y) = 4

Step 2: Covariance
$$\text{Cov}(X, Y) = \frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$
= [(2−5)(1−4) + (4−5)(3−4) + (6−5)(5−4) + (8−5)(7−4)] / 4
= (−3×−3 + −1×−1 + 1×1 + 3×3) / 4 = (9 + 1 + 1 + 9) / 4 = 20 / 4 = **5**

Step 3: Correlation
$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

SD of X = SD of Y = $\sqrt{5} \approx 2.24$
$$r = \frac{5}{2.24 \times 2.24} = \frac{5}{5.02} \approx 0.996$$

Interpretation: Strong positive correlation between X and Y.