**Statistics assignment**

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

**Quantitative data** refers to numerical values:

* **Discrete**: Countable values (e.g., number of students: 1, 2, 3…).
* **Continuous**: Measurable values (e.g., height: 170.5 cm).

**Qualitative data** refers to categorical values:

* **Nominal**: Categories without order (e.g., gender: male/female).
* **Ordinal**: Categories with order (e.g., satisfaction level: satisfied < very satisfied).

**Scales of Measurement**:

* **Nominal**: Labels (e.g., colors).
* **Ordinal**: Ranked data (e.g., customer feedback: bad < good).
* **Interval**: Numeric data without true zero (e.g., temperature in Celsius).
* **Ratio**: Numeric data with true zero (e.g., weight, age).

**2. What are the measures of central tendency, and when should you use each?**

* **Mean**: Average value. Best for symmetrical data without outliers.  
  *Example*: Average test score.
* **Median**: Middle value. Use when data is skewed or has outliers.  
  *Example*: Median income.
* **Mode**: Most frequent value. Use for categorical or repeated values.  
  *Example*: Most common shirt size sold.

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

**Dispersion** measures the spread of data around the central value.

* **Variance**: Average of squared differences from the mean.
* **Standard Deviation**: Square root of variance; shows average distance from the mean.

Smaller values mean data is tightly clustered; larger values mean more spread.

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

A **box plot** shows the five-number summary of a dataset:

* Minimum, Q1, Median (Q2), Q3, Maximum.

It displays **spread, center, and outliers** clearly and is useful for comparing distributions.

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

**Random sampling** ensures that every individual has an equal chance of selection, which:

* Reduces bias,
* Makes results generalizable,
* Enables valid statistical inference.

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

**Skewness** indicates asymmetry in a dataset:

* **Positive skew**: Tail on the right (e.g., income data).
* **Negative skew**: Tail on the left (e.g., age at retirement).
* **Zero skew**: Symmetric data.

Skewness affects choice of central tendency (use median for skewed data).

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

**IQR = Q3 − Q1**  
It measures the middle 50% spread of data.

**Outliers** are values:

* Below Q1 − 1.5 × IQR
* Above Q3 + 1.5 × IQR

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

Use the **binomial distribution** when:

* Fixed number of trials,
* Each trial is independent,
* Only two outcomes (success/failure),
* Constant probability of success.

*Example*: Flipping a coin 10 times.

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

**Normal distribution** is:

* Symmetrical, bell-shaped,
* Mean = median = mode.

**Empirical Rule**:

* 68% within 1 SD,
* 95% within 2 SDs,
* 99.7% within 3 SDs.

Useful for identifying outliers and probability estimates.

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

**Example**: Number of emails received per hour.

If λ = 5 emails/hour, find probability of receiving exactly 3 emails:

P(X=3)=e−5⋅533!=e−5⋅1256≈0.1404P(X = 3) = \frac{e^{-5} \cdot 5^3}{3!} = \frac{e^{-5} \cdot 125}{6} \approx 0.1404P(X=3)=3!e−5⋅53​=6e−5⋅125​≈0.1404

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

A **random variable** assigns numerical values to outcomes of a random process.

* **Discrete**: Countable outcomes (e.g., number of heads in coin flips).
* **Continuous**: Infinite values in a range (e.g., weight of a person).

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

**Dataset** (x: study hours, y: scores):  
x = [2, 4, 6], y = [50, 60, 80]

* **Covariance**:

Cov(x,y)=(2−4)(50−63.33)+(4−4)(60−63.33)+(6−4)(80−63.33)3=(−2)(−13.33)+0+2(16.67)3≈20.00\text{Cov}(x, y) = \frac{(2-4)(50-63.33) + (4-4)(60-63.33) + (6-4)(80-63.33)}{3} = \frac{(-2)(-13.33) + 0 + 2(16.67)}{3} ≈ 20.00Cov(x,y)=3(2−4)(50−63.33)+(4−4)(60−63.33)+(6−4)(80−63.33)​=3(−2)(−13.33)+0+2(16.67)​≈20.00

* **Correlation (r)**:

r=Cov(x,y)σxσy≈0.98r = \frac{\text{Cov}(x,y)}{\sigma\_x \sigma\_y} ≈ 0.98r=σx​σy​Cov(x,y)​≈0.98

**Interpretation**: Strong positive correlation between study hours and scores.