**BASIC OF STATISTICS**

**Assignment Questions:**

**ANS:1**

**QUANTITATIVE:**  This type of data provides only numerical data. It helps or used to make predictions and draw conclusion.

**QUALITATIVE:** This is a information which cannot be counted into numbers. This helps us to understand why, or what happened behind certain behaviour.

**SCALE OF MEASUREMENT:**

1. **Nominal Scale Data:**
2. It is qualitative or categorial data
3. There is no order in the data
4. Examples: Gender, Colour, Location, Percentage
5. **Ordinal Scale Data:**
6. It is classified in order or ranks and are meaningful
7. But, the difference cannot be measured
8. We cannot use operations like subtraction or addition.
9. **Interval Scale Data**
10. The rank and order has meaning
11. Differences can be measured [ Except Ratio ]
12. It doesn’t have ‘Zero’ starting value.
13. Zero starting value means we can do addition or subtraction but not multiplication or division.
14. **Ratio Scale Data**
15. It also has order or rank and have meaning.
16. Differences and ratios are measurable.
17. It also have ‘Zero’ starting point.

**ANS:2** Central tendency represent the centre point of a database or dataset. For example:

1. P.M represents India.
2. 1,2,3,4,5 where 3 is central tendency.

**Measures**

1. Mean
2. Median
3. Mode

**MEAN (Arithmetic Mid Value, Average)**

Mean is summarize of all the observation and dividing by no. of observation.

Mean is calculated when data is in numerical term.

**MEDIAN (Physical Mid Point Of Data)**

Median is middle most value in a set of a numbers. Median is calculated upon outlier. Median is calculated when data or set is ion numerical.

**MODE** (Maximum Frequency)

Mode is Calculated upon highest repeated number or maximum frequency number. Mode is calculated or use when data or set is in categorial.

**ANS:3** Dispersion refers to the extent to which data points in a dataset spread out from the central value (such as the mean). It helps in understanding the variability, consistency, or homogeneity of the data. The higher the dispersion, the more spread out the data points are.

**Variance**

measures how far each data point in the dataset is from the mean. It is the **average of the squared differences** from the mean.

**Standard Deviation**

Standard deviation is the **square root of variance**, making it easier to interpret as it is in the same units as the original data.

**How Variance and Standard Deviation Measure Spread**

* If variance and standard deviation are **low**, the data points are **close to the mean**, meaning **less spread**.
* If variance and standard deviation are **high**, the data points are **far from the mean**, meaning **greater variability** in the data.
* Standard deviation is more commonly used as it is in the same unit as the data, making it easier to interpret.

ANS:4 A **box plot** (or **box-and-whisker plot**) is a graphical representation of a dataset that shows its distribution based on five summary statistics:

1. **Minimum** – The smallest data point (excluding outliers).
2. **First Quartile (Q1)** – The 25th percentile (lower quartile).
3. **Median (Q2)** – The middle value (50th percentile).
4. **Third Quartile (Q3)** – The 75th percentile (upper quartile).
5. **Maximum** – The largest data point (excluding outliers).

It also includes **whiskers** (lines extending from the box) and **outliers** (individual points beyond the whiskers).

**What Can a Box Plot Tell You?**

A box plot provides key insights into the **distribution and spread of the data**, such as:

1. **Central Tendency** – The **median (Q2)** shows the central value of the dataset.
2. **Spread of Data (IQR)** – The **Interquartile Range (IQR)** (Q3−Q1Q3 - Q1Q3−Q1) shows the middle 50% of the data. A larger IQR means more variability.
3. **Skewness (Symmetry of Distribution)**
   * If the **median** is in the center of the box and whiskers are of equal length, the data is **symmetrical**.
   * If the **median** is closer to **Q1** and the upper whisker is longer, the data is **right-skewed** (positive skew).
   * If the **median** is closer to **Q3** and the lower whisker is longer, the data is **left-skewed** (negative skew).
4. **Outliers** – Data points beyond **1.5 × IQR** from the quartiles are considered **outliers**, indicating unusual values in the dataset.

**ANS:5**

**Simple Random Sampling**

1. Every member of the population (N) has equal chance of being selected in the sample.
2. Each person has 1/12 of being selected in the sample by assuming population is about 12

**Disadvantage**

1. Probability of each and every member in a population to be a part of sample is negligible. Part of group or member are being selected for the sample.
2. For example: India has the population of about 140 million people to use each and everyone as a part of sample is not possible therefore some big states like U.P, Maharashtra, Rajasthan population can be taken into consideration for sample.

**ANS:6**

Skewness is a measure of **asymmetry** in the distribution of data. It tells us whether the data is **symmetrically distributed** or **leaning more toward one side**.

* A **symmetrical** distribution has **zero skewness**.
* A **positive** or **negative skew** indicates that the data is concentrated more on one side.

**Types of Skewness**

1. **Symmetrical Distribution (Zero Skewness)**
   * The mean, median, and mode are **approximately equal**.
   * The distribution has a **bell shape** (like a normal distribution).
   * Example: Heights of people in a large population.
2. **Positive Skewness (Right-Skewed Distribution, Skewness > 0)**
   * The **right tail** of the distribution is **longer**.
   * The **mean** is greater than the **median**, and the **mode** is the smallest value.
   * The data has **more smaller values**, with a few large values pulling the tail to the right.
   * Example: **Income distribution** (a few people earn very high salaries, shifting the mean).
3. **Negative Skewness (Left-Skewed Distribution, Skewness < 0)**
   * The **left tail** of the distribution is **longer**.
   * The **mean** is less than the **median**, and the **mode** is the largest value.
   * The data has **more larger values**, with a few small values pulling the tail to the left.
   * Example: **Age at retirement** (most people retire around 60, but some retire early at 40 or 50).

**How Skewness Affects Data Interpretation**

1. **Affects Measures of Central Tendency**
   * In **positively skewed data**, the **mean is greater** than the median, which can **overestimate** the "typical" value.
   * In **negatively skewed data**, the **mean is lower** than the median, which can **underestimate** the "typical" value.
2. **Impacts Decision Making**
   * If a company's sales data is **right-skewed**, the average sales figure may not represent most sales. Instead, the **median** is a better measure.
   * If exam scores are **left-skewed**, most students performed well, and a few poor scores bring down the mean.
3. **Influences Statistical Analysis**
   * Many statistical tests assume **normality** (zero skewness). If data is highly skewed, **log transformation** or other techniques might be needed before applying models like regression.

**ANS:7**

Interquartile Range (IQR) & Outlier Detection

What is IQR?

The Interquartile Range (IQR) measures the spread of the middle 50% of data. It is calculated as:

Where:

* Q1 (First Quartile) = 25th percentile (lower quartile)
* Q3 (Third Quartile) = 75th percentile (upper quartile)

How is IQR Used to Detect Outliers?

Outliers are values that fall too far from the middle range of the data. The standard rule:

* Lower Bound = Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR
* Upper Bound = Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR

Any value below the lower bound or above the upper bound is considered an outlier.

ANS:8

**Conditions for Using Binomial Distribution**

The **Binomial Distribution** is used when an experiment meets the following **four conditions**:

1. **Fixed Number of Trials (n)**
   * The experiment consists of a **fixed number** of independent trials (e.g., flipping a coin **10 times**).
2. **Only Two Possible Outcomes**
   * Each trial has only **two possible outcomes**: **Success (S)** or **Failure (F)** (e.g., heads or tails, pass or fail).
3. **Constant Probability (p)**
   * The probability of success **remains the same** for each trial. (e.g., the probability of getting heads in a fair coin toss is **0.5**).
4. **Independent Trials**
   * Each trial is **independent**; the outcome of one trial **does not affect** the next.

ANS:8

**Properties of Normal Distribution**

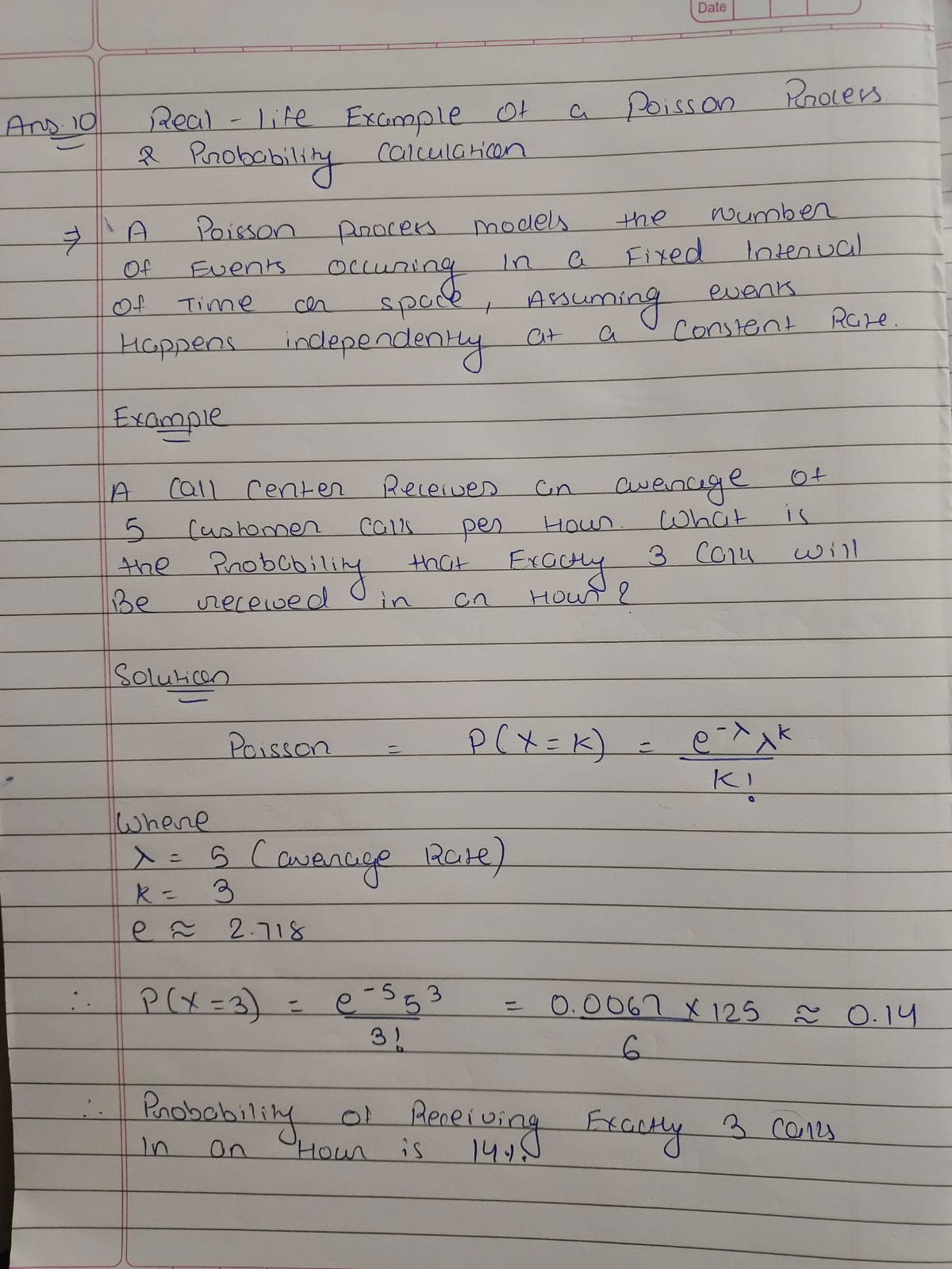
1. **Bell-Shaped & Symmetric** – The curve is symmetric around the mean (μ).
2. **Mean = Median = Mode** – All are equal and located at the center.
3. **Asymptotic** – The tails extend infinitely without touching the x-axis.
4. **Defined by Mean (μ) & Standard Deviation (σ)** – These parameters determine the shape and spread.
5. **Total Area = 1** – The total probability under the curve is **100% (1)**.

**Empirical Rule (68-95-99.7 Rule)**

This rule describes how data is distributed in a **normal distribution**:

* **68%** of data lies **within 1σ** of the mean (μ ± 1σ).
* **95%** of data lies **within 2σ** of the mean (μ ± 2σ).
* **99.7%** of data lies **within 3σ** of the mean (μ ± 3σ).

**ANS:10**



**ANS:11**

Random Variable & Its Types

A random variable (RV) is a numerical outcome of a random phenomenon.

Types:

1. Discrete Random Variable – Takes countable values (e.g., number of students in a class, dice roll outcomes).
2. Continuous Random Variable – Takes infinite values in a range (e.g., height of students, temperature readings).

📌 Key Difference: Discrete variables arise from counting, while continuous variables arise from measuring.

**ANS:12**

