Mathematics.

PR - 1.

Descriptive Booster.

Part A.: Theory & Defination.

1. Define the following with examples from the dataset :

- Types of data : Numerical & categorical
Numerical (quantitative)

→ Variable measured as number where arithmetics makes sense, such as Age_of_household_head, Household_income, and Family_size (all ratio scale here, because zero is meaningful and ratio are interpretable).

→ Categorical (qualitative): Variable that describe category or label, such as Education_level (Primary, secondary, Graduate, Post-Grad), Owns_house (Yes/No), Urban_Rural (Urban/Rural), and household_ID (identifiers).

- Types of statistics : Descriptive & Inferential.

→ Descriptive statistics : Methods used to summarize and describe the main features of the data that have been collected, for example computing mean, income, median age, or the proportion of urban household in this dataset.

→ Inferential statistics : Methods used to draw conclusion about a larger populations from sample data, such as using this sample of household to estimate the average income of all households in a region and construct confidence intervals or hypothesis tests.

• What is Descriptive statistics ?

→ Descriptive statistics : Any numerical summary that describes some aspects of the sample data, such as mean, median, standard deviation, skewness, kurtosis, or a pecentile for Household. income.

2. Explain the difference between

• Mean, Median, Mode

→ Mean : The arithmetics average, calculated as total of all values divided by number of observation; for example, mean Household_Income of all household divided by the number of households.

Median : The middle value when income are sorted; 50% of households earns less than or equal to the median income and 50% earn more or equal, which makes it robust to extreme high income in dataset.

Mode : The most frequent value or category, such as the most common Education_level (eg, "Graduate") or the most common Family_Size.

• Range, Variance, standard Deviation.

→ Range : Difference between maximum and minimum value, such as Max Household_income minus min Household_Income, giving the full spread.

Variance : Average of squard deviation from the mean; it measures how far, on average, income deviate from the mean in squared units.

Standard Deviation : Square root of variance, returning the spread to the original unit (e.g., currency units for income), often used to interpret variability more directly.

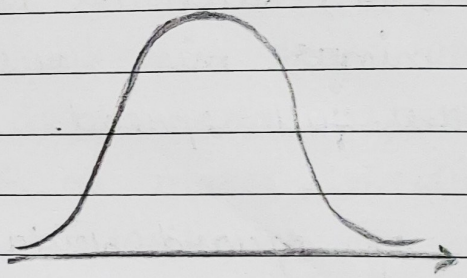
3. Explain the following item with neat and clean diagram along with its formula:

• Gaussian distribution : A symmetric, bell-shaped distribution characterized by mean μ and standard deviation σ; its probability function is :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

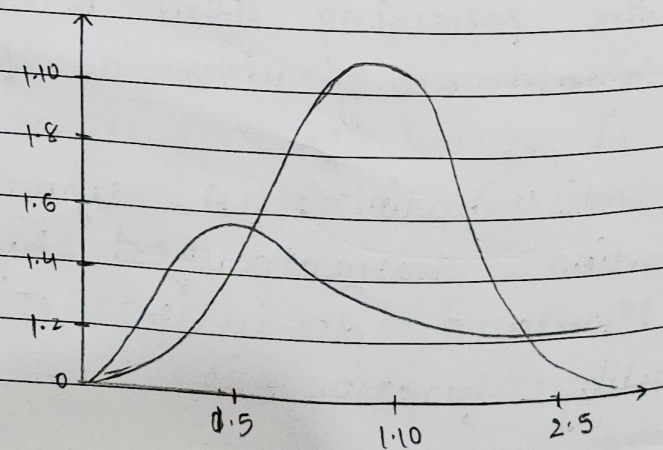
In socio-economic data, variable like standardized tests scores or some measurement errors are often approximated as normal, whereas income typically is not.

- ~~tog~~

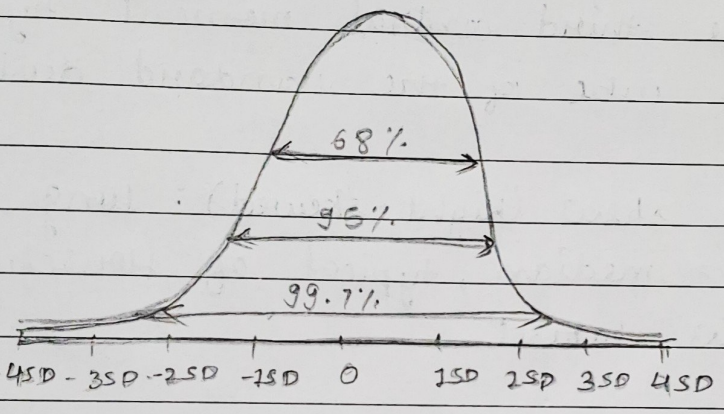


- log-Normal distribution : A positive, right-skewed distribution where $\ln(X)$ is normally distributed; the density involve parameters μ and σ of the underlying normal for $\ln(X)$.

Household-Income is usually better modelled as log-normal because most households cluster at lower income with a long right tail of high-income households.



- 3-sigma Rule or Empirical Rule : For a normal distribution, about 68% of observation fall within $\mu \pm 2\sigma$, about 95% with $\mu \pm 2\sigma$, and about 99.7% within $\mu \pm 3\sigma$.
When a variable in this dataset is approximately normal (e.g., maybe Age_of_Household_Head), values beyond about 3 standard deviations from the mean can be treated as extreme.



- Percentiles : Values that split ordered data into 100 equal parts ; for example, the 90th percentile of Household_Income is the income level above which the top 10% of households lie.

$$R = \frac{P}{100}(N+1)$$

- Quartiles : Special percentiles that divide ordered data into four parts :

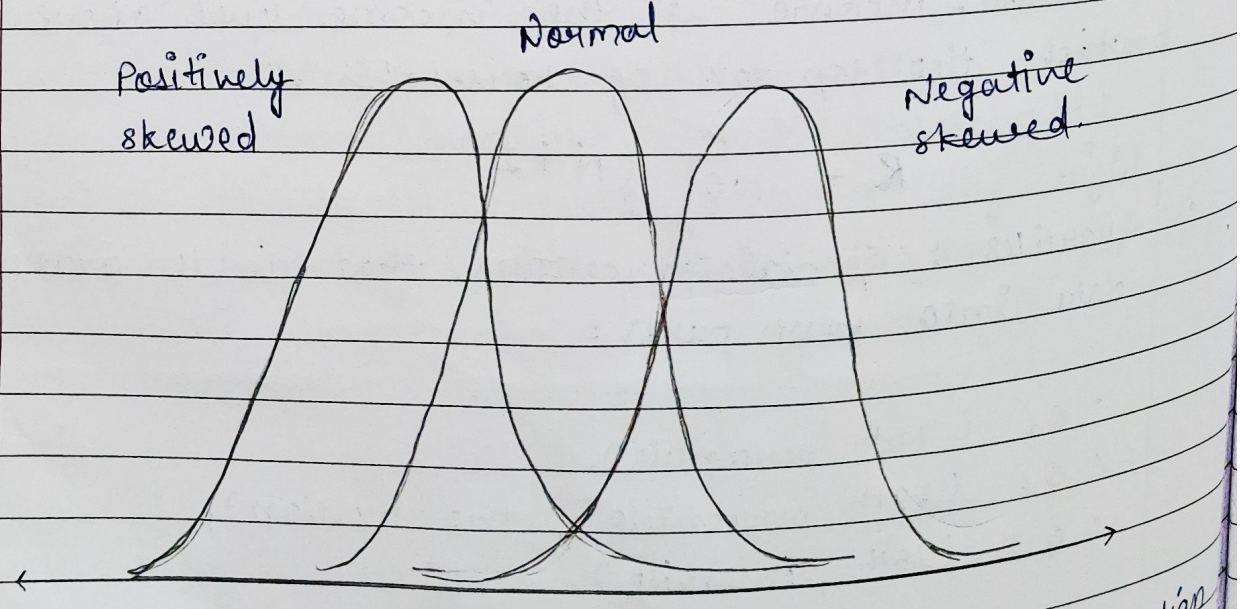
  - Q_1 (25th percentile),
  - Q_2 (50th percentile, the median),
  - Q_3 (75th percentile).

- 5-number summary : A concise description of distributing using Minimum, Q_1, Median Q_3, and Maximum; for e.g, for household_income you would list these five values to understand overall spread and central tendency including potential extremes.

- Skewness : A numerical measure of asymmetry of a distribution; a common sample formula uses the third central moment by the divided by the cube of the standard deviation.

  - Positive skew (Right-skewed) : long right tail, mean > median, typical of Household_Income in this dataset.

  - Negative skew (left-skewed) : long left tail, mean < median, which could appear in capped variables.

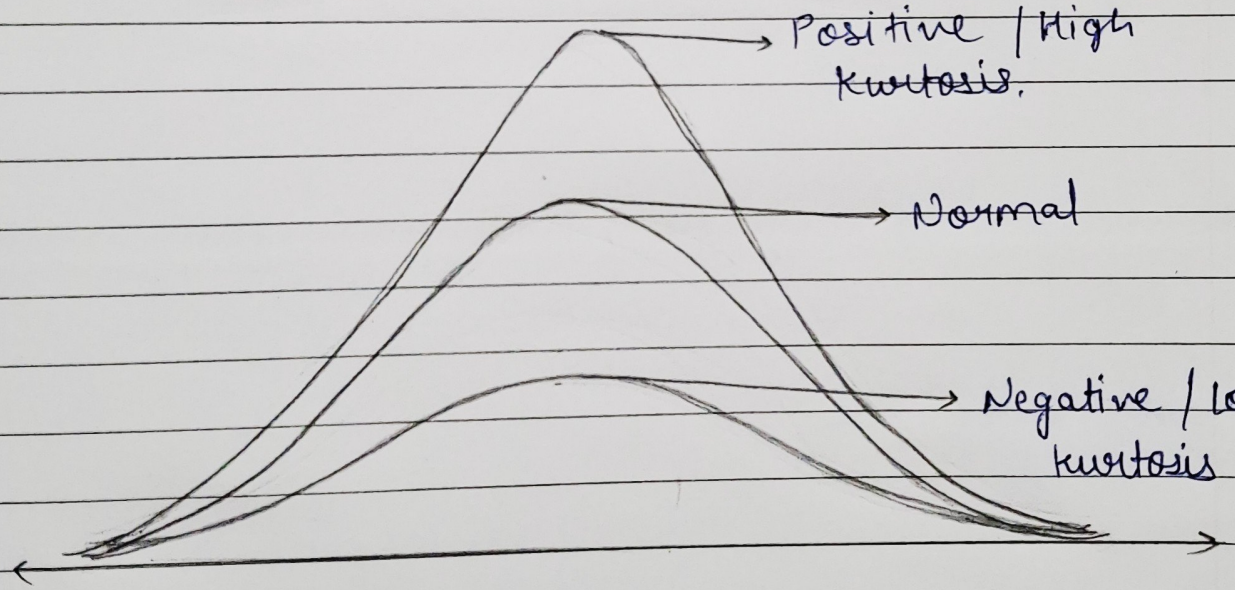


$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3, \quad \frac{\bar{x} - \text{median}}{s}$$

- Kurtosis : A numerical measure of tail heaviness and peakedness ; the sample coefficient is based on the fourth central moment divided by the fourth ~~central moment~~ power of the standard deviation, often expressed as "excess kurtosis" (kurtosis minus 3).

  - High kurtosis (leptokurtic) : Heavier tails than Normal, indicating more extreme incomes (very poor and very rich) relative to a normal model.

  - Low kurtosis (Platykurtic) : Lighter tails than normal, indicating fewer extreme values.



$$\text{Excess kurtosis} = \frac{\mu_4}{\sigma^4} - 3.$$