**Assignment - Statistics**

**Statistics** - branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data.

They include descriptive and inferential statistics.

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Descriptive Statistics** | **Inferential Statistics** |
| Purpose | Summarizes and describes features of a dataset without making broader inferences. | Makes inferences, predictions, or generalizations about a population based on sample data |
| Objective | Describes characteristics of the data without generalizing | Generalizes findings from sample to population |
| Examples | Measures of central tendency, dispersion, frequency distributions, graphical representations | Hypothesis testing, regression analysis, confidence intervals |
| Statistical Techniques | Mean, median, mode, range, variance, standard deviation, histograms, box plots, etc. | Hypothesis testing, regression analysis, confidence intervals |
| Goal | To provide insights into the characteristics of a dataset | To make predictions or draw conclusions about a population |

**MEASURES OF CENTRAL TENDENCY 🡪** These measures indicate where most values in a distribution fall.

* **Mean (Arithmetic Average)**: The sum of all values divided by the number of values.
* **Median**: The middle value when the data are arranged in ascending or descending order. If there is an even number of observations, the median is the average of the two middle numbers.
* **Mode**: The value that appears most frequently in the data set. A data set may have one mode, more than one mode, or no mode at all.

**MEASURES OF DISPERSION** 🡪 These measures indicate the spread or variability among values.

* **Range**: The difference between the maximum and minimum values in a data set.
* **Variance**: The average of the squared differences from the mean. It gives an idea of how much the values in a data set deviate from the mean.
* **Standard Deviation**: The square root of the variance. It is expressed in the same units as the data and provides a measure of the average distance from the mean.
* **Interquartile Range (IQR)**: The quartiles divide a dataset into four equal parts, each representing 25% of the data.

**First Quartile (Q1):** Q1 divides the lowest 25% of the data from the rest.

Calculation: It is the median of the lower half of the dataset. If the dataset is arranged in ascending order, Q1 is at the 25th percentile.

**Second Quartile (Q2):** Q2 is the median of the dataset.

Calculation: It divides the dataset into two halves, with 50% of the data below it and 50% above it. It is at the 50th percentile.

**Third Quartile (Q3):** Q3 divides the upper 25% of the data from the rest.

Calculation: It is the median of the upper half of the dataset. Q3 is at the 75th percentile when the dataset is ordered.

**Fourth Quartile (Q4):** Q4 represents the highest 25% of the data.

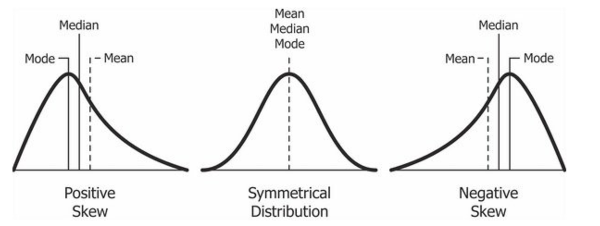
Calculation: While not commonly referred to as Q4, it would conceptually represent the top 25% of the dataset, above Q3.

These quartiles are useful for understanding the distribution and spread of data, especially in statistical analysis and box plots.

**IQR:** The range between the first quartile (Q1) and the third quartile (Q3). It measures the spread of the middle 50% of values. IQR=Q3−Q1

**SHAPE OF THE DISTRIBUTION 🡪** These measures describe the symmetry and peak of the data distribution.

* **Skewness**: Measures the asymmetry of the data distribution. A distribution is skewed if one of its tails is longer or fatter than the other.
  + **Positive Skew**: Tail on the right side is longer.
  + **Negative Skew**: Tail on the left side is longer.



* **Kurtosis**: Measures the "tailed ness" of the distribution. It indicates whether the data are heavy-tailed or light-tailed relative to a normal distribution.

**DATA VISUALISATION 🡪** These tools help to visually summarize data.

* **Histograms**: Show the frequency distribution of a data set. They are useful for understanding the shape of the data distribution.
* **Box Plots**: Show the median, quartiles, and potential outliers in a data set. They are useful for comparing distributions.
* **Bar Charts**: Represent categorical data with rectangular bars. Each bar's height is proportional to the value it represents.
* **Pie Charts**: Show the relative proportions of different categories in a data set as slices of a circle.
* **Scatter Plots**: Display values for two variables as points on a Cartesian plane. They are useful for identifying relationships between variables.

df.describe() will show the summary of descriptive statistics.

**Inferential Statistics**

Inferential statistics includes hypothesis testing, confidence intervals, and regression analysis, among other techniques.

These methods help researchers determine whether their findings are statistically significant and whether they can generalize their results to the larger population.

**Hypothesis testing**

**Null Hypothesis (H₀)**

The null hypothesis is a statement that there is no effect or no difference, and it serves as the default or starting assumption in hypothesis testing. It is the hypothesis that the researcher aims to test against.

* **Example**: In a clinical trial testing a new drug, the null hypothesis might state that the drug has no effect on patients compared to a placebo.

**Alternative Hypothesis (H₁ or Ha)**

The alternative hypothesis is a statement that indicates the presence of an effect or a difference. It is what the researcher wants to prove.

* **Example**: In the same clinical trial, the alternative hypothesis might state that the drug has a significant effect on patients compared to a placebo.

**Probability Distribution**

Probability is a measure of the likelihood of an event occurring. Probability distributions describe how the probabilities are distributed over the values of the random variable.

There are two main types of probability distributions: discrete and continuous.

**Discrete Probability Distributions**

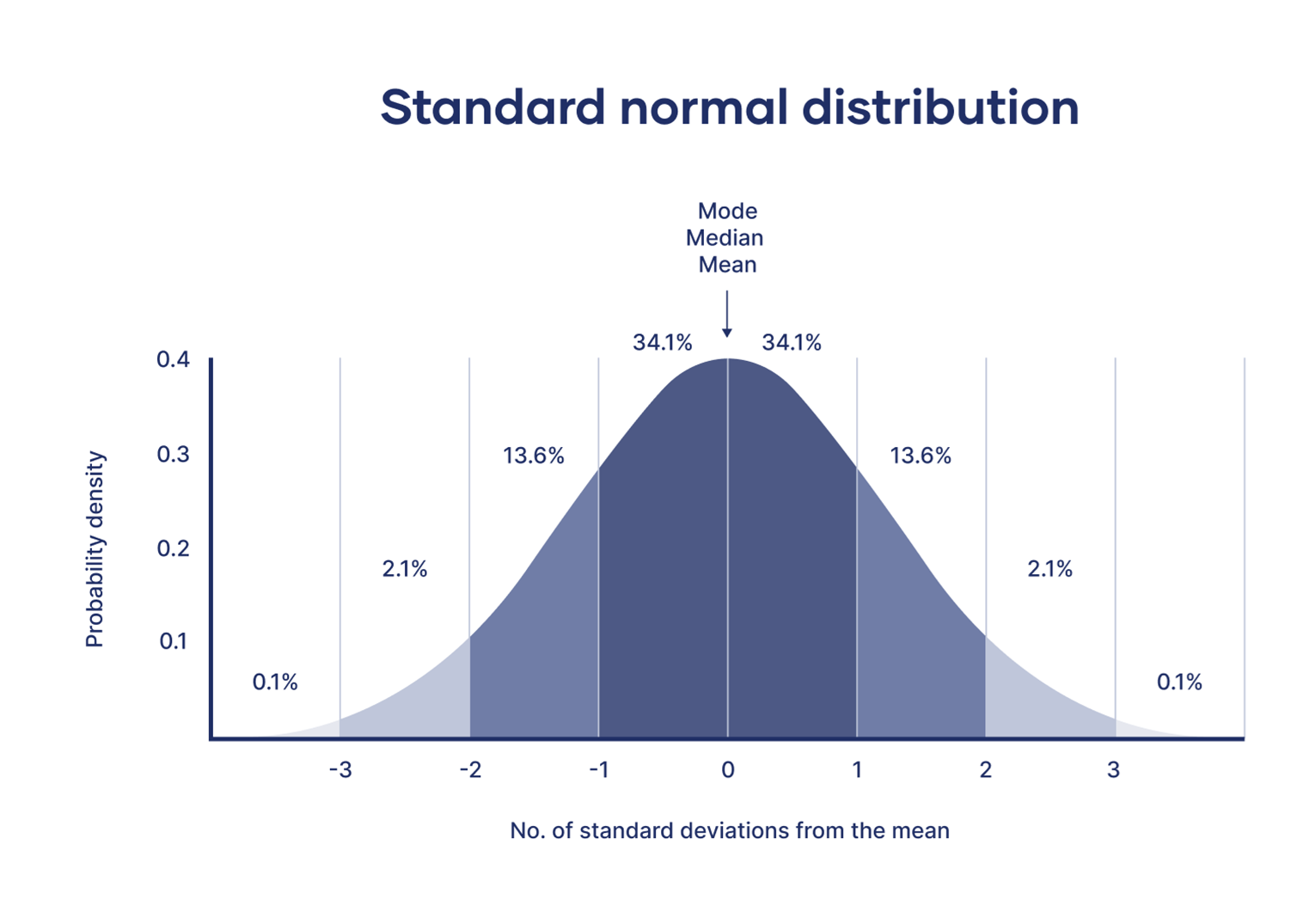
* Binomial Distribution
* Poisson Distribution

**Continuous Probability Distributions**

* Normal Distribution
* Exponential Distribution
* Uniform Distribution

**Normal Distribution**

The normal distribution is a continuous probability distribution that is symmetric about the mean, representing the distribution of many types of data (e.g., heights, test scores).



In the above normal distribution, it can be seen that almost 68% data lies within the 1st standard deviation which means 68% data has probability to lie within 1st std deviation.

**Frequency Distribution**

A frequency distribution is an overview of all values of some variable and the number of times they occur. They give us an idea about the range where most values fall and the ranges where values are scarce.

To represent the Frequency Distribution, there are various methods such as Histogram, Bar Graph, Frequency Polygon, and Pie Chart.

A graph of ageing

Description automatically generated with medium confidence

From the above frequency distribution, it can be inferred that between age 29-39 years has maximum frequency of occurrence followed by age group 19-29 years.