Whai is statistics?

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data to make informed decisions or predictions. It involves various methods to summarize and infer properties about the data set.

Types

1 Descriptive Statistics: These are methods used to summarize and describe the main features of a data set. Examples include measures of central tendency (mean, median, mode) .

2 Inferential Statistics: These are methods used to make predictions or inferences about a population based on a sample of data. Techniques include hypothesis testing, confidence intervals, and regression analysis.

Data:

Data is a collection of facts, figures, or information that can be used for analysis and decision-making.

Population:

* A population is the entire group of individuals or items that you want to study or collect information from.
* It includes every member of the group that fits your criteria of interest.
* For example, if you're studying the height of all students in a school, the population would be all students in that school.

Sample:

* A sample is a subset of the population that you actually collect data from.
* It is used to make inferences or generalizations about the entire population.
* For example, if you randomly select 50 students from the school to measure their height, those 50 students represent a sample of the entire student population.

Sampling techniques:

Simple Random Sampling:

* Every member of the population has an equal chance of being selected.
* Example: Drawing names from a hat.
* Stratified Sampling:
* The population is divided into strata (groups) based on a specific characteristic, and random samples are taken from each stratum.
* Example: Sampling students from different grades in a school.

Systematic Sampling:

* Selecting every nth member of the population.
* Example: Choosing every 10th person from a list.

Convenience Sampling:

* Selecting members of the population that are easily accessible.
* Example: Surveying people at a shopping mall

Variable:

In statistics, a variable is any characteristic, number, or quantity that can be measured or quantified. Variables can vary or change from one entity to another or over time. Here's a breakdown of the types of variables:

1. Qualitative (Categorical) Variables:
   * These variables describe qualities or categories and cannot be measured numerically.
   * Example: Gender (male, female), Blood type (A, B, AB, O).
2. Quantitative (Numeric) Variables:
   * These variables represent numerical values and can be measured.
   * There are two subtypes:
     1. Discrete Variables: Countable values (e.g., number of students in a class).
     2. Continuous Variables: Values that can take any value within a range (e.g., height, weight).

variable measurement scales :

Variable measurement scales are systems for classifying data based on the type and level of measurement. Here are the four main types of scales used to measure variables:

1. Nominal Scale:
   * Used for qualitative data where the categories do not have any logical order.
   * Example: Gender (male, female), Blood type (A, B, AB, O).
2. Ordinal Scale:
   * Used for qualitative data where the categories have a logical order but the differences between categories are not meaningful.
   * Example: Satisfaction levels (satisfied, neutral, dissatisfied), Education levels (high school, bachelor's, master's).
3. Interval Scale:
   * Used for quantitative data where the differences between values are meaningful, but there is no true zero point.
   * Example: Temperature in Celsius or Fahrenheit, IQ scores.
4. Ratio Scale:
   * Used for quantitative data where there is a true zero point, and both differences and ratios between values are meaningful.
   * Example: Weight, height, age.

Frequency:

In statistics, frequency refers to the number of times a specific value or category appears in a data set.

Types of frequency:

1. Absolute Frequency:
   * The actual number of times a particular value appears in a data set.
   * Example: If a survey records 15 people owning a car, the absolute frequency of car owners is 15.

2. Cumulative Frequency:

* The sum of the frequencies of all categories up to a certain point.
* It helps understand the accumulation of data values.
* Example: If a class has test scores categorized into ranges, the cumulative frequency for the range 70-80 includes all scores below 80.

Measures of Central Tendency

Measures of central tendency are statistical tools that describe the center or typical value of a data set. There are three main measures of central tendency:

1. Mean :
   * The sum of all the values in the data set divided by the number of values.
2. Median:
   * The middle value of a data set when the values are arranged in ascending or descending order.
   * If the data set has an odd number of values, the median is the middle value.
   * If the data set has an even number of values, the median is the average of the two middle values.
3. Mode:
   * The value that appears most frequently in the data set.
   * A data set can have more than one mode if multiple values have the highest frequency (bimodal or multimodal).
   * Example: For the data set [3, 5, 5, 7], the mode is 5. For [3, 3, 5, 5, 7], the data set is bimodal with modes 3 and 5.

Variance:

Variance is a measure of the spread or dispersion of a set of data points. It indicates how much the values in a data set differ from the mean (average) value. Variance gives you an idea of the degree to which individual observations deviate from the mean of the data set.

Here's how to calculate variance:

1. Find the mean of the data set.
2. Subtract the mean from each data point and square the result.
3. Sum all the squared results.
4. Divide the sum by the number of data points (for a population) or by the number of data points minus one (for a sample).

### **Standard Deviation:**

The standard deviation is the square root of the variance. It provides a measure of spread in the same unit as the original data, making it easier to interpret. It is one of the most commonly used measures of dispersion.

1. Percentage

Percentage is a way of expressing a number as a fraction of 100. It is used to describe how much something is relative to a whole. Percentages are widely used in many fields like finance, education, and health to express proportions or parts of a total.

2. Percentile

Percentile is a measure used in statistics to describe the position of a particular value in a dataset. It indicates the percentage of data points that fall below a given value.

*Formula for Percentile:*

Percentiles are calculated based on the rank of a value in a sorted dataset. The general process for finding the percentile rank of a value is:

Five Number Summary

A Five-Number Summary is a way to describe a dataset using five key values. It's a quick and effective method to understand the distribution and spread of the data. The five values included in this summary are:

1. Minimum: The smallest value in the dataset.
2. First Quartile (Q1): The median of the lower half of the dataset (25th percentile).
3. Median (Q2): The middle value of the dataset (50th percentile).
4. Third Quartile (Q3): The median of the upper half of the dataset (75th percentile).
5. Maximum: The largest value in the dataset.

Outliers:

An outlier is a data point that significantly differs from other observations in a dataset. Outliers can result from variability in the data, measurement errors, or experimental errors, and they can sometimes indicate something significant about the data being studied.

Data Distribution:

Data distribution refers to the way data values are spread or distributed across a range of possible values. Understanding data distribution helps in identifying patterns, trends, and making inferences about the dataset. Here are some common types of data distributions:

1. Normal Distribution:
   * Also known as Gaussian distribution or bell curve.
   * Data is symmetrically distributed around the mean.
   * Most values cluster around the mean, with fewer values appearing as you move away from the mean.
   * Example: Heights of people in a population.
2. Skewed Distribution:
   * Positively Skewed (Right-Skewed): The tail on the right side of the distribution is longer or fatter.
     + Example: Income distribution in many countries.
   * Negatively Skewed (Left-Skewed): The tail on the left side of the distribution is longer or fatter.
     + Example: Age at retirement in certain professions.
3. Uniform Distribution:
   * All values have the same frequency or probability.
   * Data is evenly distributed across the range.
   * Example: Rolling a fair die.
4. Bimodal Distribution:
   * Contains two distinct peaks or modes.
   * Can indicate the presence of two different groups within the dataset.
   * Example: Test scores with two different levels of performance.
5. Exponential Distribution:
   * Describes the time between events in a Poisson process.
   * Data is skewed to the right with a rapid drop-off.
   * Example: Time between arrivals at a service station.

Z-Score:

The Z-score (or standard score) is a measure that describes how many standard deviations a data point is from the mean of the data set. It helps standardize different data points, allowing you to compare them even if they come from different distributions.

Z=(X−μ)σZ = \frac{(X - \mu)}{\sigma}

Where:

* XX is the data point.
* μ\mu (mu) is the mean of the data set.
* σ\sigma (sigma) is the standard deviation of the data set.

Common Types of Normalization

There are several methods of normalization, with the choice depending on the nature of the data and the requirements of the model or analysis. Below are the most common techniques:

*Min-Max Normalization (Rescaling)*

* Description: Min-Max normalization scales the data to a fixed range, usually 0 to 1, by transforming each feature to a specific range.

*Binomial Distribution*

* Description: The binomial distribution models the number of successes in a fixed number of independent trials, each with two possible outcomes (success or failure). It is used when the trials are independent, and the probability of success remains constant across trials.
* Example: If you flip a fair coin 5 times, the probability of getting exactly 3 heads (successes) is given by the binomial distribution

Bernoulli Distribution

The Bernoulli distribution is one of the simplest and most fundamental discrete probability distributions in statistics. It models a random experiment or trial that has only two possible outcomes: a success (usually denoted as 1) or a failure (denoted as 0). This makes it a special case of the binomial distribution, where the number of trials is 1.

Confidence Level and Confidence Interval

Let's dive into the concepts of confidence level and confidence interval, which are fundamental in inferential statistics.

Confidence Level

Confidence level refers to the percentage of times that a confidence interval would contain the true population parameter if you were to repeat an experiment or sampling procedure multiple times. Common confidence levels include 90%, 95%, and 99%.

* A 95% confidence level means that if you were to take 100 different samples and compute a confidence interval for each sample, approximately 95 of those intervals would contain the true population parameter.

Confidence Interval

Confidence interval is a range of values, derived from the sample data, that is likely to contain the true population parameter. It provides an estimated range of values which is likely to include an unknown population parameter.

Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on sample data. It involves testing an assumption (hypothesis) about a population parameter. Here are the main steps involved in hypothesis testing:

1. State the Hypotheses

* Null Hypothesis (H0H\_0): The statement that there is no effect or no difference. It's the hypothesis that we assume to be true unless there is strong evidence against it.
* Alternative Hypothesis (H1H\_1 or HAH\_A): The statement that there is an effect or a difference. It's what you want to prove.
* The p-value indicates the probability of obtaining the observed result (or more extreme) if the null hypothesis is true. A low p-value suggests that the null hypothesis is unlikely.

Types of Errors in Hypothesis Testing

In hypothesis testing, there are two main types of errors that can occur:

Type I Error (False Positive)

* Definition: This occurs when the null hypothesis (H0H\_0) is true, but you incorrectly reject it.
* Example: Suppose you are testing a new drug's effectiveness. A Type I error would occur if you conclude that the drug is effective when it actually isn't.

Type II Error (False Negative)

* Definition: This occurs when the null hypothesis (H0H\_0) is false, but you fail to reject it.
* Example: In the same drug effectiveness study, a Type II error would occur if you conclude that the drug is not effective when it actually is.

t test

A t-test is a statistical test used to compare the means of two groups. It helps determine if the difference between the groups is statistically significant. There are different types of t-tests, each suited for specific scenarios:

Types of t-tests:

One-Sample t-test:

* + Compares the mean of a single sample to a known value (usually a population mean).
  + Example: Testing if the average height of a sample of students is equal to the national average height.

Independent Two-Sample t-test (or Unpaired t-test):

* + Compares the means of two independent samples.
  + Example: Testing if the average test scores of students from two different schools are different.

Paired Sample t-test (or Dependent t-test):

* + Compares the means of two related samples, such as the same group of individuals measured at two different times.

chi square test:

The chi-square test is a statistical test used to determine whether there is a significant association between categorical variables. It's commonly used in hypothesis testing to assess how likely it is that an observed distribution is due to chance. There are two main types of chi-square tests:

Chi-Square Test of Independence

* Purpose: To determine if there is a significant association between two categorical variables.
* Example: Testing if there is an association between gender (male/female) and preference for a product (like/dislike).