

Statistics and Probability

Define Statistics

-> Statistics is a branch of mathematics that deals with the collection, organization, interpretation, and analysis of data.

What is probability?

-> Probability is the branch of mathematics concerning events and numerical descriptions of how likely they are to occur.

The probability of an event is a number between 0 and 1; the larger the probability, the more likely an event is to occur.

The higher the probability of an event, the more likely it is that the event will occur.

Explain types of statistics briefly

-> There are two kinds of Statistics, which are descriptive Statistics and inferential Statistics.

In descriptive Statistics, the Data or Collection Data are described in a summarized way, whereas in inferential Statistics, we make use of it in order to explain the descriptive kind.

Give examples for measure of central tendency

-> **Mean:** Mean is the average of the given numbers

and is calculated by dividing the sum of given numbers by the total number of numbers.

$$\bar{X} = (\text{Sum of values} \div \text{Total Number of values})$$

example : {1,2,3,4,5,6,7,8}

$$\bar{X} = (1+2+3+4+5+6+7+8)/8 = 36/8 = 4.5$$

-> **Median:** The median is a statistical measure that represents the middle value of a given list of data when arranged in order. The median formula is different for even and odd numbers of observations.

Median for even data set : $((n/2) + ((n/2)+1)) / 2$, ex: {1,2,3,4} median = $((4/2) + ((4/2)+1)) / 2 = (2+3)/2 = 2.5$

Median for odd data set : $(n+1)/2$, ex: {1,2,3,4,5} = $(5+1)/2 = 3/2 = 3$

-> **Mode:** the mode is the value that is repeatedly occurring in a given set.

Explain about quantitative & qualitative data with examples

Quantitative Data: is anything that can be counted or measured; it refers to numerical data

Revenue in dollars

Weight in kilograms or pounds

Age in months or years

Length in centimeters or inches

Distance in kilometers or miles

Statistics and Probability

Height in feet or inches

Qualitative data: is descriptive, referring to things that can be observed but not measured—such as colors or emotions.

Ex: The hair colors of players on a football team

The types of coins in a jar

The letter grades of students in a classroom

The shape of candies in a variety pack

#Frequency Distribution is a tool in statistics that helps us organize the data and also helps us reach meaningful conclusions. It tells us how often any specific values occur in the dataset.

Histogram	Bar Graph
The histogram is a term that refers to a graphical representation that shows data by way of bars to display the frequency of numerical data.	The bar graph is a graphical representation of data that uses bars to compare different categories of data.
Distribution of non-discrete variables.	Comparison of discrete variables.
Bars touch each other, so there are no spaces between bars.	Bars never touch each other, so there are spaces between bars.
In this type of graph, elements are grouped so that they are considered as ranges.	In this type of graph, elements are taken as individual entities.
Histogram width may vary.	The bar chart is mostly of equal width.
To display the frequency of occurrences.	To compare different categories of data.
In Histogram, the data points are grouped and rendered based on its bin value.	In the Bar graph, each data point is rendered as a separate bar.
The items of the Histogram are numbers, which should be categorized to represent data range.	As opposed to the bar graph, items should be considered as individual entities.
In Histogram, we cannot rearrange the blocks.	Bar graph, it is common to rearrange the blocks, from highest to lowest

#Five number Summary

The 5 number summary is an exploratory data analysis tool that provides insight into the distribution of values for one variable. Collectively, this set of statistics describes where data values occur, their central tendency, variability, and the general shape of their distribution.

- Lowest value in the dataset.
- First quartile(Q1) – 25% of the values in the dataset
- Median
- Third Quartile (Q3) – 75% of the values in the dataset

The Multiplication Rule

Statistics and Probability

If A and B are two events defined on a sample space, then:

$$P(A \text{ AND } B) = P(B)P(A | B)$$

This rule may also be written as:

$$P(A | B) = P(A \text{ AND } B) / P(B)$$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are independent, then

$$P(A | B) = P(A).$$

$$P(A \text{ AND } B) = P(A)P(B).$$

The Addition Rule

If A and B are defined on a sample space, then:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

If A and B are mutually exclusive, then

$$P(A \text{ AND } B) = 0.$$

$$P(A \text{ OR } B) = P(A) + P(B).$$

Permutations

Permutation is the distinct interpretations of a provided number of components carried one by one, or some, or all at a time. For example, if we have two components A and B, then there are two likely performances, AB and BA.

$$\text{Permutation : } P(n,r) = n! / (n-r)!$$

Combinations

The number of possible arrangements in a collection of items where the order of selection doesn't matter.

$$\text{Combination: } C(n,r) = (nr) = n! / (r!(n-r)!)$$

Z test:

Z test is used when you need to compare group means. Use the 1-sample analysis to determine whether a population mean is different from a hypothesized value. Or use the 2-sample version to determine whether two population means differ.

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Statistics and Probability

T Test:

A t-test is a kind of inferential statistical test that is extensively used to compare the mean of two groups of samples.

$$T = \frac{m - \mu}{s / \sqrt{n}}$$

Chi Square Test

$$\chi^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

Bernoulli Distribution:

A random variable follows a Bernoulli distribution if it only has two possible outcomes: 0 or 1.

For instance, consider flipping a coin once. Let's denote the probability of it landing heads as (p). Then, the probability of it landing tails is (1-p).

In this case, the random variable (X) follows a Bernoulli distribution.

It can only take on two possible values: 0 (for tails) or 1 (for heads).

When we flip a coin multiple times, the sum of these Bernoulli random variables follows a Binomial distribution.

Binomial Distribution:

Suppose we flip a coin (n) times and want to know the probability of obtaining heads (k) times.

The random variable (X) follows a Binomial distribution.

The probability that (X = k) successes can be found using the following formula: $[P(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}]$

(n): Number of trials

(k): Number of successes

(p): Probability of success on a given trial

$(\binom{n}{k})$: The number of ways to obtain (k) successes in (n) trials

When (n = 1) trial, the Binomial distribution is equivalent to the Bernoulli distribution.