**Statistics - Statistical and Non-Statistical Analysis**

Statistics is a form of mathematical analysis that concerns the collection, organization, analysis, interpretation, and presentation of data. Statistical analysis helps to make the best use of the vast data available and improves the efficiency of solutions.

R – Statistics

R Programming Language is used for environment statistical computing and graphics. The following is an introduction to basic R Statistics concepts like normal distribution (bell curve), central tendency (the mean, median, and mode), variability (25%, 50%, 75% quartiles), variance, standard deviation, modality, and skewness.

Data Concepts

Data can be formed in different structures and different formats, before starting the concepts of R Statistics we need to know the data formats.

Basic Calculations

In R, you can perform basic calculations using arithmetic operators. Here are the basic arithmetic operators in R:

Addition: +

Subtraction: -

Multiplication: \*

Division: /

Exponentiation: ^

Modulus (remainder after division): %%

Integer division (quotient after division): %/%

How to do Addition in R?

You can do addition in R using the + operator. Here’s an example:

**# Addition of two numbers**

2 + 3

# Output: 5

# Addition of two variables

x <- 4

y <- 5

x + y

# Output: 9

**Exponents in R**

You can perform exponentiation in R using the ^ operator or the exp() function. Here’s an example:

# Exponentiation using the ^ operator

2^3

# Output: 8

# Exponentiation using the exp() function

exp(2)

# Output: 7.389056

# Exponentiation of a variable

x <- 2

y <- 3

x^y

# Output: 8

**Mathematical Constants in R**

R provides several built-in mathematical constants that you can use in your calculations. Here are some examples:

# Value of pi

pi

# Output: 3.141593

# Value of Euler's number (e)

exp(1)

# Output: 2.718282

# Value of the golden ratio (phi)

phi <- (1 + sqrt(5))/2

phi

# Output: 1.618034

**Logarithms in R**

The basic logarithmic function in R is log(), which computes the natural logarithm (base e) of a given number. The general syntax for using the log() function is:

log(x, base)

where x is the input value and base is the optional base of the logarithm. If base is not specified, it defaults to base = e.

Here are some examples:

# Compute the natural logarithm of 2

log(2)

# Compute the logarithm base 10 of 100

log(100, base = 10)

# Compute the logarithm base 2 of 8

log(8, base = 2)

**Trigonometry in R**

R is a programming language that provides a variety of functions for performing trigonometric calculations. The most commonly used trigonometric functions in R are:

sin(x): returns the sine of x, where x is in radians

cos(x): returns the cosine of x, where x is in radians

tan(x): returns the tangent of x, where x is in radians

asin(x): returns the inverse sine of x, where x is in radians

acos(x): returns the inverse cosine of x, where x is in radians

atan(x): returns the inverse tangent of x, where x is in radians

atan2(y, x): returns the angle between the positive x-axis and the point (x,y) in radians

Here’s an example of using some of these functions in R:

# calculate the sine and cosine of pi/4

sin(pi/4)

cos(pi/4)

# calculate the tangent of pi/3

tan(pi/3)

# calculate the inverse sine of 0.5

asin(0.5)

# calculate the angle between (1,1) and the positive x-axis

atan2(1,1)

**Basic statistics** is the branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It includes a range of methods used to describe and summarize numerical data, including

measures of central tendency (such as mean, median, and mode)

measures of variability (such as range, variance, and standard deviation)

graphical representations (such as histograms and box plots)

**Mean Median and Mode – Definition, Formulas, Examples**

Mean, median, and mode are the three measures of central tendency used in statistics to describe the central or typical value of a set of data.

In “Graphs and shapes of Distributions” section it is explained how to summarizing a distribution of your data in terms of graphs. Now it’s time to measure the center of your distribution. Once we talk about measuring central tendency of a variable then 3 M’s come into picture.

Mode

Median

Mean



**Mode**

The mode is the value that occurs most frequently in a dataset. It is the most common value in the dataset. A dataset can have multiple modes, or no mode at all if there is no value that occurs more than once.

If your variable of interest is measured in nominal or ordinal (Categorical) level then Mode is the most often used technique to measure the central tendency of your data.

Finding the mode is easy. Basically, it is the value that occurs most frequently. In other words, mode is the most common outcome. Mode is the name of the category that occurs more often.

There is a chance of having more than one mode in your variable.



**Median**

The second measure of central tendency is the median. The median is nothing more than the middle value of your observations when they are order from the smallest to the largest.



**Mean**

The mean, also known as the arithmetic mean or average, is the sum of all values in a dataset divided by the number of values. It is a measure of the central tendency that reflects the center of mass of the data. The formula for the mean is:

Mean = (sum of values) / (number of values)

It’s the third measure of central tendency is the most often used one, and also the one you most probably already know quite well: the mean. The mean is the sum of all the values divided by the number of observations. It is nothing but the average value.



Example:

Dataset-https://www.kaggle.com/datasets/saurav9786/cardiogoodfitness

# R program to import data into R

# Import the data using read.csv()

myData = read.csv("CardioGoodFitness.csv",

stringsAsFactors=F)

# Print the first 6 rows

print(head(myData))

# R program to illustrate

# Descriptive Analysis

# Import the data using read.csv()

myData = read.csv("CardioGoodFitness.csv",

stringsAsFactors=F)

# Compute the mean value

**mean** = mean(myData$Age)

print(mean)

# R program to illustrate

# Descriptive Analysis

# Import the data using read.csv()

myData = read.csv("CardioGoodFitness.csv",

stringsAsFactors=F)

# Compute the median value

**median** = median(myData$Age)

print(median)

# Import the data using read.csv()

myData = read.csv("CardioGoodFitness.csv",

stringsAsFactors=F)

**mode** = function(){

return(sort(-table(myData$Age))[1])

}

mode()

**More Examples**

**# R program to get average of a list**

**# Taking a list of elements**

list = c(2, 4, 4, 4, 5, 5, 7, 9)

# Calculating average using mean()

print(mean(list))

**Variance in R**

Variance is the sum of squares of differences between all numbers and means. The mathematical formula for variance is as follows,

Formula: \sigma^{2}= \frac { \sum\_{i=1}^{N} (x\_{i}-\mu)^{2}}{N}

# R program to get variance of a list

# Taking a list of elements

list = c(2, 4, 4, 4, 5, 5, 7, 9)

# Calculating variance using var()

print(var(list))

**Standard Deviation** in R Programming Language

Standard Deviation is the square root of variance. It is a measure of the extent to which data varies from the mean. The mathematical formula for calculating standard deviation is as follows,

Standard Deviation = \sqrt{ variance }

# R program to get

# standard deviation of a list

# Taking a list of elements

list = c(290, 124, 127, 899)

# Calculating standard

# deviation using sd()

print(sd(list))