Unit 4

Mathematical and statistical Concept using R:

Certainly! Mathematical and statistical concepts are frequently applied in data analysis using the R programming language. Here are some key concepts and corresponding R functions:

Descriptive Statistics:

1. **Mean:**
    - **Math Concept:** The average of a set of numbers.
    - **R Function: mean()**
2. **Median:**
    - **Math Concept:** The middle value of a set of numbers.
    - **R Function: median()**
3. **Standard Deviation:**
    - **Math Concept:** Measures the amount of variation or dispersion of a set of values.
    - **R Function: sd()**
4. **Variance:**
    - **Math Concept:** The average of the squared differences from the Mean.
    - **R Function: var()**
5. **Quantiles:**
    - **Math Concept:** Divides the data into parts of specified probabilities.
    - **R Function: quantile()**

Inferential Statistics:

1. **Hypothesis Testing:**
    - **Math Concept:** Making inferences about a population based on a sample.

- **R Functions: t.test()**, **wilcox.test()**, **chisq.test()**, etc.

2. **Linear Regression:**
    - **Math Concept:** Modeling the relationship between dependent and independent variables.
    - **R Function: lm()**
3. **ANOVA (Analysis of Variance):**
    - **Math Concept:** Comparing means between groups.
    - **R Function: aov()**

Probability Distributions:

1. **Normal Distribution:**
    - **Math Concept:** A continuous probability distribution.
    - **R Functions: dnorm()**, **pnorm()**, **qnorm()**, **rnorm()**
2. **Binomial Distribution:**
    - **Math Concept:** Number of successes in a fixed number of independent Bernoulli trials.
    - **R Functions: dbinom()**, **pbinom()**, **qbinom()**, **rbinom()**

Mathematical Operations:

1. **Matrix Operations:**
    - **Math Concept:** Manipulating matrices.
    - **R Functions: matrix()**, **solve()**, **det()**, **%*%** (matrix multiplication)
2. **Element-wise Operations:**
    - **Math Concept:** Operations performed element-wise.
    - **R Functions: +**, **-**, *****, **/**

Visualization:

1. **Histogram:**

- **Math Concept:** Visual representation of the distribution of a dataset.
- **R Function: hist()**

2. **Boxplot:**
    - **Math Concept:** Graphical representation of the summary of a set of data values.
    - **R Function: boxplot()**
3. **Scatterplot:**
    - **Math Concept:** Displaying the relationship between two continuous variables.
    - **R Function: plot()**

These are just a few examples, and R provides a comprehensive set of functions for various mathematical and statistical concepts.

Frequency Distribution

Frequency Distribution is a tool in statistics that helps us organize the data and also helps us reach meaningful conclusions. Frequency Distribution tells us how often any specific values occur in the dataset. To understand the data easily, we categorize the data into class intervals. The number of items occurring in the specific range or class interval is shown under Frequency against that particular class range to which the item belongs. These distributions allow us to get insights from any data, see the trends, and predict the next values or the direction in which the data will go. Frequency Distribution can be defined for any kind of dataset either grouped or ungrouped. Frequency Distribution is in Class 9 and Class 11 syllabus and hence should be studied by respective students.

What is Frequency Distribution in Statistics?

Frequency distributions tell us how frequencies are distributed over the values. That is how many values lie between different intervals. They give us an idea about the range where most of the values fall and the ranges where values are scarce.

Frequency Distribution Definition

A frequency distribution is an overview of all values of some variable and the number of times they occur.

Frequency Distribution Graphs

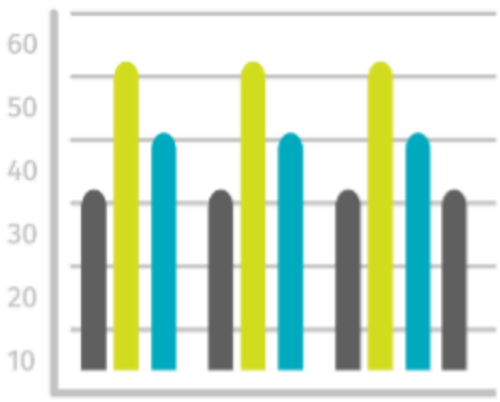
To represent the Frequency Distribution, there are various methods such as Histogram, Bar Graph, Frequency Polygon, and Pie Chart. A brief description of all these graphs is as follows:

Histograms: Histogram is a graphical representation of distribution that represents the frequency of each interval of continuous data generally using bars of equal width.

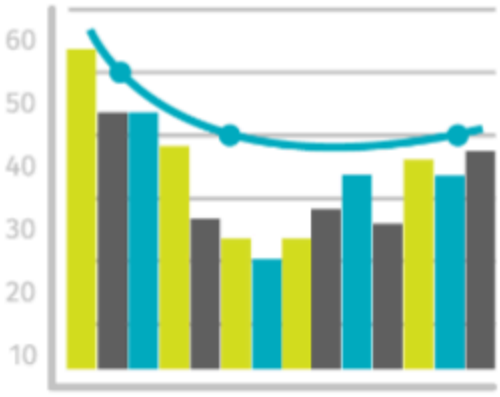
Bar Graph: Bar Graph is also a graphical representation of distribution that represents the frequency of each interval using bars of equal width. It can also represent discrete data, unlike a histogram.

Frequency Polygon: A Frequency Polygon is a type of graphical representation similar to a histogram but instead of using bars it uses a line to connect the midpoints of the frequencies of the class. It helps us to compare the various different datasets.

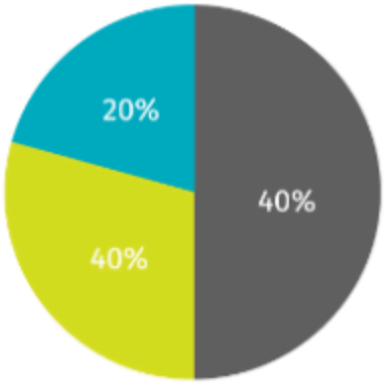
Pie Charts: A Pie Chart is a circular graph that represents the pieces of data in the form of slices of a circle. Each slice represents the proportional size of the data represented by that slice to the complete data set. Pie charts are commonly used to show the relative sizes of different parts of a whole.



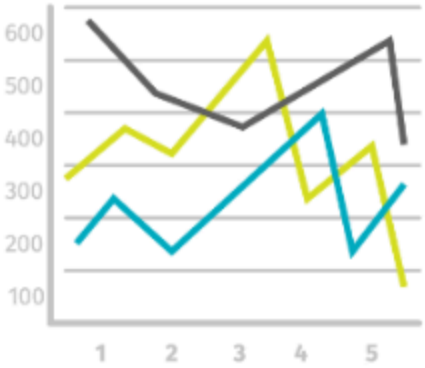
Bar Graph



Histogram



Pie Chart



Frequency Polygon

0 seconds of 25 secondsVolume 0%

This ad will end in 24

Frequency Distribution Table

A frequency distribution table is a way to organize and present data in a tabular form which helps us summarize the large dataset into a concise table. In the frequency distribution table, there are two columns one representing the data either in the form of a range or an individual data set and the other column shows the frequency of each interval or individual. For example, let's say we have a dataset of students' test scores in a class.

Test Score	Frequency
0-20	6
20-40	12
40-60	22
60-80	15
80-100	5

How to Make a Frequency Distribution Table?

To make the Frequency Distribution Table, follow these steps:

Step 1: *Analyze the ungrouped data given and decide what kind of Frequency Distribution Table is needed – grouped, relative, or cumulative.*

Step 2*: For a grouped distribution table, decide on the Class Intervals for the observations or, for an ungrouped distribution table, list the unique observations.*

Step 3: *Count the frequencies of each class or observation.*

Step 4: *Calculate the relative or cumulative frequency for a grouped distribution table if necessary.*

Types of Frequency Distribution Table

Based on the analysis and categorization of the data, there are two types of Frequency Distribution Tables i.e.,

- Grouped Frequency Distribution Table
- Ungrouped Frequency Distribution Table

Let’s learn about these two types in detail.

Grouped Frequency Distribution Table

A grouped frequency distribution table is a table that organizes any given data into intervals or groups, known as class intervals and displays the frequency or number of observations that fall within each interval.

For example, we can consider the table of the number of cattle owned by families in a town.

Number of Cattle	**Number of Families**
10 – 20	5
20 – 30	12
30 – 40	8
40 – 50	15
50 – 60	20

In the above table, we can see there are two columns. The first column represents the number of cattle and the second column represents the number of families who own the associate number of cattle. As the first column is grouped with a certain interval length, thus this table is an example of Grouped Frequency Distribution.

Ungrouped Frequency Distribution Table

An **ungrouped frequency distribution table** is a statistical table that organizes individual data values along with their corresponding frequencies instead of groups or class intervals. For example, consider the number of vowels in any given paragraph.

Vowel	Frequency
a	7
e	10
i	7
o	6
u	3

In the above table, we can the two columns representing a list of vowels and their frequency in any given paragraph. As the first column is a list of some individual elements, thus this table is an example of Ungrouped Frequency Distribution.

Types of Frequency Distribution

There are four types of frequency distributions:

- Grouped Frequency Distribution
- Ungrouped Frequency Distribution
- Relative Frequency Distribution
- Cumulative Frequency Distribution

Grouped Frequency Distribution

In Grouped Frequency Distribution observations are divided between different intervals known as class intervals and then their frequencies are counted for each class interval. This Frequency Distribution is used mostly when the data set is very large.

Example: Make the Frequency Distribution Table for the ungrouped data given as follows:

23, 27, 21, 14, 43, 37, 38, 41, 55, 11, 35, 15, 21, 24, 57, 35, 29, 10, 39, 42, 27, 17, 45, 52, 31, 36, 39, 38, 43, 46, 32, 37, 25

Solution:

As there are observations in between 10 and 57, we can choose class intervals as 10-20, 20-30, 30-40, 40-50, and 50-60. In these class intervals all the observations are covered and for each interval there are different frequency which we can count for each interval. Thus the Frequency Distribution Table for the given data is as follows:

Class Interval	**Frequency**
10 – 20	5
20 – 30	8
30 – 40	12
40 – 50	6
50 – 60	3

Ungrouped Frequency Distribution

In Ungrouped Frequency Distribution, all distinct observations are mentioned and counted individually. This Frequency Distribution is often used when the given dataset is small.

Example: Make the Frequency Distribution Table for the ungrouped data given as follows:

10, 20, 15, 25, 30, 10, 15, 10, 25, 20, 15, 10, 30, 25

Solution:

As unique observations in the given data are only 10, 15, 20, 25, and 30 with each having a different frequency. Thus the Frequency Distribution Table of the given data is as follows:

Value	Frequency
10	4
15	3
20	2
25	3
30	2

Relative Frequency Distribution

This distribution displays the proportion or percentage of observations in each interval or class. It is useful for comparing different data sets or for analyzing the distribution of data within a set.

and Relative Frequency is given by

Relative Frequency = Frequency of the Event/Total Number of Events

Example: Make the Relative Frequency Distribution Table for the following data:

Score Range	0-20	21-40	41-60	61-80	81-100
Frequency	5	10	20	10	5

Solution:

To Create the Relative Frequency Distribution table, we need to calculate Relative Frequency for each class interval. Thus Relative Frequency Distribution table is given as follows:

Score Range	Frequency	Relative Frequency
0-20	5	5/50 = 0.10
21-40	10	10/50 = 0.20
41-60	20	20/50 = 0.40
61-80	10	10/50 = 0.20
81-100	5	5/50 = 0.10
Total	**50**	**1.00**

Cumulative Frequency Distribution

Cumulative frequency is defined as the sum of all the frequencies in the previous values or intervals up to the current one. The frequency distributions which represent the frequency distributions using cumulative frequencies are called cumulative frequency distributions. There are two types of cumulative frequency distributions:

- **Less than type:** We sum all the frequencies before the current interval.
- **More than type:** We sum all the frequencies after the current interval.

Let's see how to represent a cumulative frequency distribution through an example,

Example: The table below gives the values of runs scored by Virat Kohli in the last 25 T-20 matches. Represent the data in the form of less-than-type cumulative frequency distribution:

45	**34**	**50**	**75**	**22**
56	**63**	**70**	**49**	**33**
0	**8**	**14**	**39**	**86**
92	**88**	**70**	**56**	**50**
57	**45**	**42**	**12**	**39**

Solution:

Since there are a lot of distinct values, we'll express this in the form of grouped distributions with intervals like 0-10, 10-20 and so. First let's represent the data in the form of grouped frequency distribution.

Runs	**Frequency**
0-10	2
10-20	2
20-30	1

Runs	Frequency
30-40	4
40-50	4
50-60	5
60-70	1
70-80	3
80-90	2
90-100	1

Now we will convert this frequency distribution into cumulative frequency distribution by summing up the values of current interval and all the previous intervals.

Runs scored by Virat Kohli	Cumulative Frequency
Less than 10	2
Less than 20	4

Runs scored by Virat Kohli	**Cumulative Frequency**
Less than 30	5
Less than 40	9
Less than 50	13
Less than 60	18
Less than 70	19
Less than 80	22
Less than 90	24
Less than 100	25

This table represents the cumulative frequency distribution of less than type.

Runs scored by Virat Kohli	**Cumulative Frequency**
More than 0	25
More than 10	23

Runs scored by Virat Kohli	Cumulative Frequency
More than 20	21
More than 30	20
More than 40	16
More than 50	12
More than 60	7
More than 70	6
More than 80	3
More than 90	1

This table represents the cumulative frequency distribution of more than type.

We can plot both the type of cumulative frequency distribution to make the Cumulative Frequency Curve.

Frequency Distribution Curve

A frequency distribution curve, also known as a frequency curve, is a graphical representation of a data set's frequency distribution. It is used to visualize the distribution and frequency of values or observations within a dataset. Let's understand it's different types based on the shape of it, as follows:

Types of Frequency Distribution Curve

There are various types of Frequency Distribution Curve, some of those are:

Normal Distribution (Gaussian Distribution) Curve: Normal Distribution Curve is the most famous and recognizable curve in all of the Frequency Distribution Curve. It is a symmetric and bell-shaped curve where most of the data is concentrated around the mean and gradually tapers towards the tails.

Skewed Distribution Curve: A distribution is said to be skewed if it is not symmetric. Skewed distributions can be either positively skewed (skewed to the right) or negatively skewed (skewed to the left). In a positively skewed distribution, the tail extends towards higher values, while in a negatively skewed distribution, the tail extends towards lower values.

Bimodal Distribution Curve: A bimodal distribution has two distinct peaks or modes in the frequency distribution. This suggests that the data may arise from two different populations or processes.

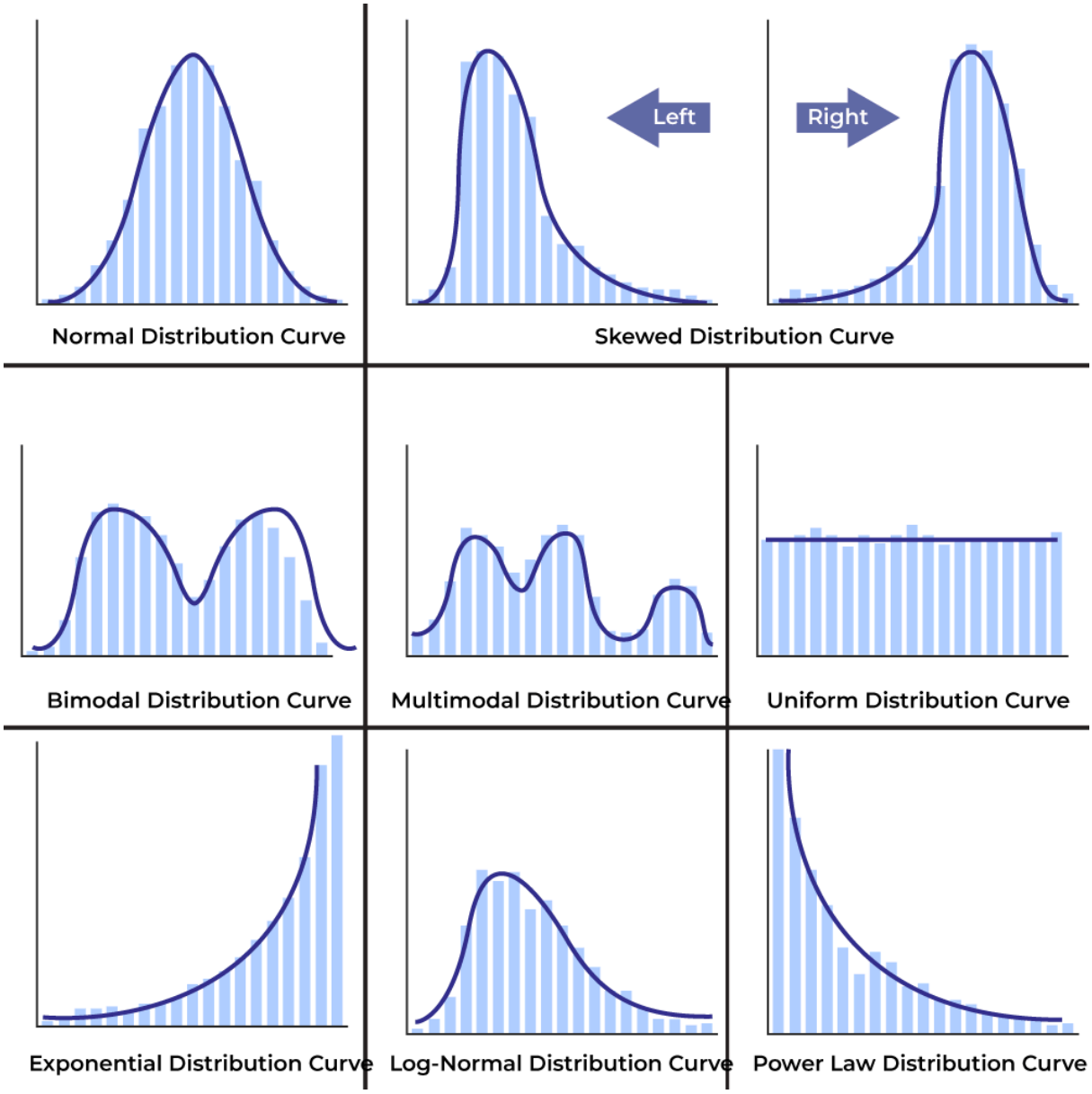
Multimodal Distribution Curve: A multimodal distribution has more than two distinct peaks or modes in the frequency distribution.

Uniform Distribution Curve: In a uniform distribution, all values or intervals have roughly the same frequency. This results in a flat, constant distribution across the range of the data.

Exponential Distribution Curve: An exponential distribution is characterized by a rapid drop-off in frequency as values increase just like an exponential function.

Log-Normal Distribution Curve: In a log-normal distribution, the logarithm of the data follows a normal distribution. The resulting distribution is positively skewed and often used to model data that can be multiplicative in nature.

Power Law Distribution Curve: In a power law distribution, the frequency of an event is inversely proportional to its magnitude. This leads to a heavy-tailed distribution where a few extreme events have much higher frequencies than the majority of events.



Frequency Distribution Formula

There are various formulas which can be learned in the context of Frequency Distribution, one such formula is the coefficient of variation. This formula for Frequency Distribution is discussed below in detail.

Coefficient of Variation

We can use mean and standard deviation to describe the dispersion in the values. But sometimes while comparing the two series or frequency distributions becomes a little hard as sometimes both have different units.

For example: Let's say we have two series, about the heights of students in a class. Now one series measures height in cm and the other one in meters. Ideally, both should have the same dispersion but the out methods of measuring the dispersion are dependent on the units in which we are measuring. This makes such comparisons hard. For dealing with such problems, we define the Coefficient of Variation.

The coefficient of Variation is defined as,

Where,

- *σ represents the standard deviation*
- *represents the mean of the observations*

Note: *The data with greater C.V. is said to be more variable than the other. The series having lesser C.V. is said to be more consistent than the other.*

Comparing Two Frequency Distributions with the Same Mean

We have two frequency distributions. Let's say are the standard deviation and mean of the first series and are the standard deviation and mean of the second series. The **Coefficeint of Variation(CV)** is calculated as follows

C.V of first series =

C.V of second series =

We are given that both series have the same mean, i.e.,

So, now C.V. for both series are,

C.V. of the first series =

C.V. of the second series =

Notice that now both series can be compared with the value of standard deviation only. Therefore, we can say that for two series with the same mean, the series with a larger deviation can be considered more variable than the other one.

Frequency Distribution Calculator

The Frequency Distribution Calculator is the calculator, which gives a curve and distribution table as output when entered with ungrouped data. Let's consider an example of it, how to convert the ungrouped data into a Frequency Distribution Table and Frequency Distribution Curve.

Example: Make a Frequency Distribution Table as well as the curve for the data:

{45, 22, 37, 18, 56, 33, 42, 29, 51, 27, 39, 14, 61, 19, 44, 25, 58, 36, 48, 30, 53, 41, 28, 35, 47, 21, 32, 49, 16, 52, 26, 38, 57, 31, 59, 20, 43, 24, 55, 17, 50, 23, 34, 60, 46, 13, 40, 54, 15, 62}

Answer:

To create the frequency distribution table for given data, let's arrenge the data in ascending order as follows:

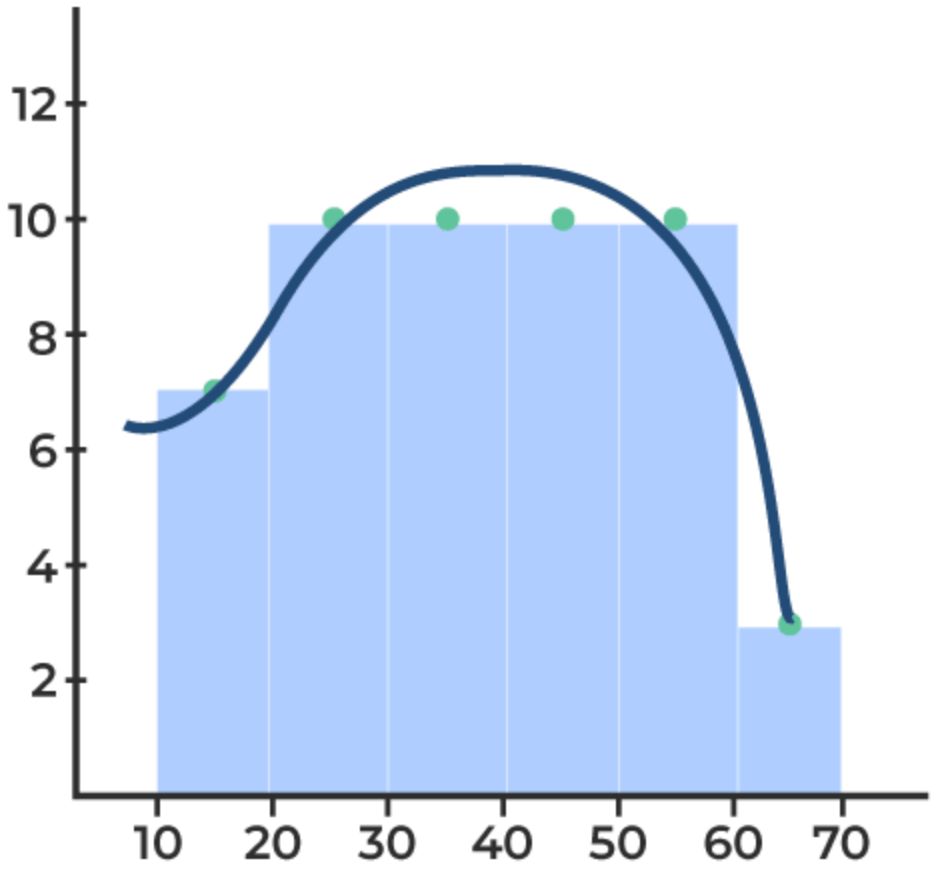
{13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62}

Now, we can count the observations for intervals: 10-20, 20-30, 30-40, 40-50, 50-60 and 60-70.

Interval	**Frequency**
10 – 20	7
20 – 30	10
30 – 40	10

Interval	Frequency
40 – 50	10
50 – 60	10
60 – 70	3

From this data, we can plot the Frequency Distribution Curve as follows:



Sample Problems on Frequency Distribution

Problem 1: Suppose we have a series, with a mean of 20 and a variance is 100. Find out the Coefficient of Variation.

Solution:

We know the formula for Coefficient of Variation,

Given mean = 20 and variance = 100.

Substituting the values in the formula,

Problem 2: Given two series with Coefficients of Variation 70 and 80. The means are 20 and 30. Find the values of standard deviation for both series.

Solution:

In this question we need to apply the formula for CV and substitute the given values.

Standard Deviation of first series.

Thus, the standard deviation of first series = 14.

Standard Deviation of second series.

Thus, the standard deviation of first series = 24.

Problem 3: Draw the frequency distribution table for the following data:

2, 3, 1, 4, 2, 2, 3, 1, 4, 4, 4, 2, 2, 2

Solution:

Since there are only very few distinct values in the series, we will plot the ungrouped frequency distribution.

Value	Frequency
1	2
2	6
3	2
4	4
Total	**14**

Problem 4: The table below gives the values of temperature recorded in Hyderabad for 25 days in summer. Represent the data in the form of less-than-type cumulative frequency distribution:

37	**34**	**36**	**27**	**22**
25	**25**	**24**	**26**	**28**
30	**31**	**29**	**28**	**30**

37	**34**	**36**	**27**	**22**
32	**31**	**28**	**27**	**30**
30	**32**	**35**	**34**	**29**

Solution:

Since there are so many distinct values here, we will use grouped frequency distribution. Let's say the intervals are 20-25, 25-30, 30-35. Frequency distribution table can be made by counting the number of values lying in these intervals.

Temperature	**Number of Days**
20-25	2
25-30	10
30-35	13

This is the grouped frequency distribution table. It can be converted into cumulative frequency distribution by adding the previous values.

Temperature	Number of Days
Less than 25	2
Less than 30	12
Less than 35	25

FAQs on Frequency Distribution

1. Define Frequency Distribution in Statistics.

A frequency distribution is a table or graph that displays the frequency of various outcomes or values in a sample or population. It shows the number of times each value occurs in the data set.

2. What is the Purpose of a Frequency Distribution?

The purpose of frequency distribution is to organize and summarize the data by showing the frequency of the observations. This helps us identify the patterns in any given set of data.

3. What are the Different Types of Frequency Distributions?

There are four types of frequency distributions that are as follows:

- *Grouped Frequency Distribution*
- *Ungrouped Frequency Distribution*
- *Relative Frequency Distribution*
- *Cumulative Frequency Distribution*

4. What is an Ungrouped Frequency Distribution?

An ungrouped frequency distribution is a distribution that shows the frequency of each individual value in a data set.

5. What is a Grouped Frequency Distribution?

A grouped frequency distribution is a distribution that shows the frequency of values within specified intervals or classes.

6. What is a Relative Frequency Distribution?

A relative frequency distribution is a distribution that shows the proportion or percentage of values within each interval or class.

7. What is a Cumulative Frequency Distribution?

A cumulative frequency distribution is a distribution that shows the number or proportion of values that fall below a certain value or interval.

Central Tendency in R Programming

Central Tendency is one of the features of descriptive statistics. Central tendency tells about how the group of data is clustered around the center value of the distribution. Central tendency performs the following measures:

- Arithmetic Mean

- Geometric Mean
- Harmonic Mean
- Median
- Mode

Arithmetic Mean

The arithmetic mean is simply called the average of the numbers which represents the central value of the data distribution. It is calculated by adding all the values and then dividing by the total number of observations. **Formula:** $$X = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$ **where,**

X *indicates the arithmetic mean indicates* i^{th} *value in data vector* ***n*** *indicates total number of observations*

In R language, arithmetic mean can be calculated by the **mean()** function.

Syntax: *mean(x, trim, na.rm = FALSE)* ***Parameters: x:*** *Represents object* ***trim:*** *Specifies number of values to be removed from each side of object before calculating the mean. The value is between 0 to 0.5* ***na.rm:*** *If TRUE then removes the NA value from* ***x***

Example:

- R

```
# Defining vector

x <- c(3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23)



# Print mean

print(mean(x))
```

Output:

[1] 21.5

Geometric Mean

The geometric mean is a type of mean that is computed by multiplying all the data values and thus, shows the central tendency for given data distribution. **Formula:**

$$X = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$ **where,**

X *indicates geometric mean indicates* i^{th} *value in data vector* ***n*** *indicates total number of observations*

prod() and **length()** function helps in finding the geometric mean for a given set of numbers as there is no direct function for geometric mean.

Syntax:

prod(x)^(1/length(x))

where, prod() *function returns the product of all values present in vector* ***x length()*** *function returns the length of vector* ***x***

Example:

- R

```
# Defining vector

x <- c(1, 5, 9, 19, 25)



# Print Geometric Mean

print(prod(x)^(1 / length(x)))
```

Output:

[1] 7.344821

Harmonic Mean

The harmonic mean is another type of mean used as another measure of central tendency. It is computed as the reciprocal of the arithmetic mean of reciprocals of the given set of values. **Formula:** $X = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$ **where,**

X *indicates harmonic mean indicates* i^{th} *value in data vector* ***n*** *indicates total number of observations*

Example: Modifying the code to find the harmonic mean of given set of values.

- R

```
# Defining vector

x <- c(1, 5, 8, 10)



# Print Harmonic Mean

print(1 / mean(1 / x))
```

Output:

[1] 2.807018

Median

The median in statistics is another measure of central tendency which represents the middlemost value of a given set of values. In R language, the median can be calculated by the **median()** function.

Syntax: *median(x, na.rm = FALSE)* ***Parameters: x:*** *It is the data vector* ***na.rm:*** *If TRUE then removes the NA value from* ***x***

Example:

- R

```
# Defining vector
```

```
x <- c(3, 7, 5, 13, 20, 23, 39,
   23, 40, 23, 14, 12, 56, 23)


# Print Median
median(x)
```

Output:

[1] 21.5

Mode

The mode of a given set of values is the value that is repeated most in the set. There can exist multiple mode values in case there are two or more values with matching maximum frequency.

Since many values might occur with the highest frequency in a dataset, more than one mode value can exist in R, making the idea of mode slightly different from the mean and median.

Example 1: Single-mode value In R language, there is no function to calculate the mode. So, modifying the code to find out the mode for a given set of values.

- R

```
# Defining vector
x <- c(3, 7, 5, 13, 20, 23, 39,
   23, 40, 23, 14, 12, 56,
   23, 29, 56, 37, 45, 1, 25, 8)


# Generate frequency table
y <- table(x)
```

```
# Print frequency table
print(y)


# Mode of x
m <- names(y)[which(y == max(y))]


# Print mode
print(m)
```

Output:

```
x
 1  3  5  7  8 12 13 14 20 23 25 29 37 39 40 45 56
 1  1  1  1  1  1  1  1  1  4  1  1  1  1  1  1  2
[1] "23"
```

Example 2: Multiple Mode values

- R

```
# Defining vector
x <- c(3, 7, 5, 13, 20, 23, 39, 23, 40,
   23, 14, 12, 56, 23, 29, 56, 37,
   45, 1, 25, 8, 56, 56)


# Generate frequency table
```

```
y <- table(x)


# Print frequency table

print(y)


# Mode of x

m <- names(y)[which(y == max(y))]


# Print mode

print(m)
```

Output:

```
x

 1  3  5  7  8 12 13 14 20 23 25 29 37 39 40 45 56

 1  1  1  1  1  1  1  1  1  4  1  1  1  1  1  1  4

[1] "23" "56"
```

Hypothesis Testing in R Programming

A hypothesis is made by the researchers about the data collected for any experiment or data set. A hypothesis is an assumption made by the researchers that are not mandatory true. In simple words, a hypothesis is a decision taken by the researchers based on the data of the population collected. **Hypothesis Testing** in R Programming is a process of testing the hypothesis made by the researcher or to validate the hypothesis. To perform hypothesis testing, a random sample of data from the population is taken and testing is performed. Based on the results of the testing, the hypothesis is either selected or rejected. This concept is known as **Statistical Inference**. In this article, we'll discuss the four-step process of hypothesis testing,

One sample T-Testing, Two-sample T-Testing, Directional Hypothesis, one sample -test, two samples -test and correlation test in R programming.

Four Step Process of Hypothesis Testing

There are 4 major steps in hypothesis testing:

- **State the hypothesis-** This step is started by stating null and alternative hypothesis which is presumed as true.
- **Formulate an analysis plan and set the criteria for decision-** In this step, a significance level of test is set. The significance level is the probability of a false rejection in a hypothesis test.
- **Analyze sample data-** In this, a test statistic is used to formulate the statistical comparison between the sample mean and the mean of the population or standard deviation of the sample and standard deviation of the population.
- **Interpret decision-** The value of the test statistic is used to make the decision based on the significance level. For example, if the significance level is set to 0.1 probability, then the sample mean less than 10% will be rejected. Otherwise, the hypothesis is retained to be true.

One Sample T-Testing

One sample T-Testing approach collects a huge amount of data and tests it on random samples. To perform T-Test in R, normally distributed data is required. This test is used to test the mean of the sample with the population. For example, the height of persons living in an area is different or identical to other persons living in other areas.

Syntax: *t.test(x, mu)* ***Parameters: x:*** *represents numeric vector of data* ***mu:*** *represents true value of the mean*

To know about more optional parameters of **t.test()**, try the below command:

help("t.test")

Example:

- r

```
# Defining sample vector

x <- rnorm(100)



# One Sample T-Test

t.test(x, mu = 5)
```

Output:

One Sample t-test

data: x

t = -49.504, df = 99, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:

-0.1910645 0.2090349

sample estimates:

mean of x

0.008985172

- Data: The dataset ‘x’ was used for the test.
- The determined t-value is -49.504.
- Degrees of Freedom (df): The t-test has 99 degrees of freedom.
- The p-value is 2.2e-16, which indicates that there is substantial evidence refuting the null hypothesis.
- Alternative hypothesis: The true mean is not equal to five, according to the alternative hypothesis.

- 95 percent confidence interval: (-0.1910645, 0.2090349) is the confidence interval's value. This range denotes the values that, with 95% confidence, correspond to the genuine population mean.

Two Sample T-Testing

In two sample T-Testing, the sample vectors are compared. If var. equal = TRUE, the test assumes that the variances of both the samples are equal.

Syntax: *t.test(x, y)* ***Parameters: x*** *and* ***y:*** *Numeric vectors*

Example:

- r

```
# Defining sample vector

x <- rnorm(100)

y <- rnorm(100)



# Two Sample T-Test

t.test(x, y)
```

Output:

Welch Two Sample t-test

data: x and y

t = -1.0601, df = 197.86, p-value = 0.2904

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.4362140 0.1311918

sample estimates:

mean of x mean of y

-0.05075633 0.10175478

Directional Hypothesis

Using the directional hypothesis, the direction of the hypothesis can be specified like, if the user wants to know the sample mean is lower or greater than another mean sample of the data.

Syntax: *t.test(x, mu, alternative)* ***Parameters: x:*** *represents numeric vector data* ***mu:*** *represents mean against which sample data has to be tested* ***alternative:*** *sets the alternative hypothesis*

Example:

- r

```
# Defining sample vector

x <- rnorm(100)



# Directional hypothesis testing

t.test(x, mu = 2, alternative = 'greater')
```

Output:

One Sample t-test

data: x

t = -20.708, df = 99, p-value = 1

alternative hypothesis: true mean is greater than 2

95 percent confidence interval:

-0.2307534   Inf

sample estimates:

mean of x

-0.0651628

One Sample -Test

This type of test is used when comparison has to be computed on one sample and the data is non-parametric. It is performed using **wilcox.test()** function in R programming.

Syntax: *wilcox.test(x, y, exact = NULL)* ***Parameters: x*** *and* ***y:*** *represents numeric vector* ***exact:*** *represents logical value which indicates whether p-value be computed*

To know about more optional parameters of **wilcox.test()**, use below command:

help("wilcox.test")

Example:

- r

```
# Define vector

x <- rnorm(100)



# one sample test

wilcox.test(x, exact = FALSE)
```

Output:

Wilcoxon signed rank test with continuity correction

data: x

V = 2555, p-value = 0.9192

alternative hypothesis: true location is not equal to 0

- Data: The dataset ‘x’ was used for the test.
- The calculated test statistic or V value is 2555.
- P-value: The null hypothesis is weakly supported by the p-value of 0.9192.

- The alternative hypothesis asserts that the real location is not equal to 0. This indicates that there is a reasonable suspicion that the distribution's median or location parameter is different from 0.

Two Sample -Test

This test is performed to compare two samples of data. **Example:**

- r

```
# Define vectors

x <- rnorm(100)

y <- rnorm(100)



# Two sample test

wilcox.test(x, y)
```

Output:

Wilcoxon rank sum test with continuity correction

data: x and y

W = 5300, p-value = 0.4643

alternative hypothesis: true location shift is not equal to 0

Correlation Test

This test is used to compare the correlation of the two vectors provided in the function call or to test for the association between the paired samples.

Syntax: *cor.test(x, y)* ***Parameters: x*** *and* ***y:*** *represents numeric data vectors*

To know about more optional parameters in **cor.test()** function, use below command:

help("cor.test")

Example:

- r

```
# Using mtcars dataset in R

cor.test(mtcars$mpg, mtcars$hp)
```

Output:

Pearson's product-moment correlation

data: mtcars$mpg and mtcars$hp

t = -6.7424, df = 30, p-value = 1.788e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.8852686 -0.5860994

sample estimates:

cor

-0.7761684

- Data: The variables'mtcars$mpg' and'mtcars$hp' from the 'mtcars' dataset were subjected to a correlation test.
- t-value: The t-value that was determined is -6.7424.
- Degrees of Freedom (df): The test has 30 degrees of freedom.
- The p-value is 1.788e-07, indicating that there is substantial evidence that rules out the null hypothesis.
- The alternative hypothesis asserts that the true correlation is not equal to 0, indicating that "mtcars$mpg" and "mtcars$hp" are significantly correlated.

- 95 percent confidence interval: (-0.8852686, -0.5860994) is the confidence interval. This range denotes the values that, with a 95% level of confidence, represent the genuine population correlation coefficient.
- Correlation coefficient sample estimate: The correlation coefficient sample estimate is -0.7761684.

Correlation Analysis Using R

Introduction

Can you tell how the prices of gold will change if the stock market goes up or how the prices of gold associated with the stock market? Yes, you can with the help of correlation, one of the most common measures used to associate two variables. It is the most common analytical tool used in analytics.

Table of contents

What is Correlation?

It is a statistical measure that defines the relationship between two variables that is how the two variables are linked with each other. It describes the effect of change in one variable on another variable.

If the two variables are increasing or decreasing in parallel then they have a positive correlation between them and if one of the variables is increasing and another one is decreasing then they have a negative correlation with each other. If the change of one variable has no effect on another variable then they have a zero correlation between them.

It is used to identify the degree of the linear relationship between two variables. It is represented by $\boldsymbol{\rho}$ and calculated as:-

$\rho\ (x, y) = cov(x, y)\ /(\sigma x \times \sigma y\ )$

Where

$cov(x, y)$ = covariance of x and y

σx = Standard deviation of x

σy = Standard deviation of y

$\rho\ (x, y)$ = correlation between x and y

The value of $\rho\ (x, y)$ varies between -1 to +1.

A positive value has a range from 0 to 1 where $\rho\ (x, y)$ = 1 defines the strong positive correlation between the variables.

A negative value has a range from -1 to 0 where $\rho\ (x, y)$ = -1 defines the strong negative correlation between the variables.

No correlation is defined if the value of $\rho\ (x, y)$ = 0

Practical application of correlation using R:-

Determining the association between Fertility and Infant Mortality Rate (Using the existing dataset “swiss”)

<u>Below is the code to compute the correlation</u>

1. Loading the dataset

```
> data1<-swiss
> head(data1, 4)
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2

Franches-Mnt 92.5 39.7 5 5 93.40 20.2

Moutier 85.8 36.5 12 7 33.77 20.3

2. Creating a scatter plot using ggplot2 library

```
> library(ggplot2)

> ggplot(data1, aes(x = Fertility, y = Infant.Mortality)) + geom_point() +

+  geom_smooth(method = "lm", se = TRUE, color = 'black')
```

3. Testing the assumptions (Linearity and Normalcy)

Linearity[#]: Visible from the plot itself (True, the relationship is linear)

Normality[$]: Using Shapiro test (This is a test of normality, here we are checking whether the variables are normally distributed or not)

```
> shapiro.test(data1$Fertility)

        Shapiro-Wilk normality test

data:  data1$Fertility
W = 0.97307, p-value = 0.3449

> shapiro.test(data1$Infant.Mortality)

        Shapiro-Wilk normality test

data:  data1$Infant.Mortality
W = 0.97762, p-value = 0.4978
```

p-value is greater than 0.05, so we can assume the normality

4. Correlation Coefficient

```
> cor(data1$Fertility,data1$Infant.Mortality)
[1] 0.416556
```

5. Checking for the significance

```
> Tes<- cor.test(swiss$Fertility,swiss$Infant.Mortality,method = "pearson")
>
> Tes

        Pearson's product-moment correlation

data:  swiss$Fertility and swiss$Infant.Mortality
t = 3.0737, df = 45, p-value = 0.003585
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1469699 0.6285366
sample estimates:
    cor
0.416556
```

Since the p-value is less than 0.05 (here it is 0.003585, we can conclude that Fertility and Infant Mortality are significantly correlated with a value of 0.41 and a p-value of 0.003585.

Conclusion

As we can see there is a positive value between fertility and infant mortality rate, the point to be noted here is correlation is just a measure of association. It will tell the degree of association along with the direct or indirect proportionality.

Here we discussed only Pearson correlation. There are other types as well such as Kendall, Spearman, and Point-Biserial.

Linearity is a property where the relationship between the variables can be graphically represented as a straight line

Correlation coefficient and correlation test in R

Introduction

Correlations between variables play an important role in a descriptive analysis. A correlation measures the **relationship between two variables**, that is, how they are linked to each other. In this sense, a correlation allows to know which variables

evolve in the same direction, which ones evolve in the opposite direction, and which ones are independent.

In this article, I show how to compute **correlation coefficients**, how to perform **correlation tests** and how to **visualize relationships** between variables in R.

Correlation is usually computed on two quantitative variables, but it can also be computed on two qualitative ordinal variables.[1] See the Chi-square test of independence if you need to study the relationship between two qualitative nominal variables.

If you need to *quantify* the relationship between two variables, I refer you to the article about linear regression.

Data

In this article, we use the `mtcars` dataset (loaded by default in R):

```
# display first 5 observations
head(mtcars, 5)
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

The variables `vs` and `am` are categorical variables, so they are removed for this article:

```
# remove vs and am variables
library(tidyverse)
dat <- mtcars %>%
  select(-vs, -am)

# display 5 first obs. of new dataset
head(dat, 5)
##                    mpg cyl disp  hp drat    wt  qsec gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02    3    2
```

Correlation coefficient

Between two variables

The correlation between 2 variables is found with the `cor()` function.

Suppose we want to compute the correlation between horsepower (`hp`) and miles per gallon (`mpg`):

```
# Pearson correlation between 2 variables
cor(dat$hp, dat$mpg)
## [1] -0.7761684
```

Note that the correlation between variables *X* and *Y* is equal to the correlation between variables *Y* and *X* so the order of the variables in the `cor()` function does not matter.

The Pearson correlation is computed by default with the `cor()` function. If you want to compute the Spearman correlation, add the argument `method = "spearman"` to the `cor()` function:

```
# Spearman correlation between 2 variables
cor(dat$hp, dat$mpg,
  method = "spearman"
)
## [1] -0.8946646
```

The most common correlation methods (Run `?cor` for more information about the different methods available in the `cor()` function) are:

- **Pearson** correlation is often used for quantitative continuous variables that have a linear relationship
- **Spearman** correlation (which is actually similar to Pearson but based on the ranked values for each variable rather than on the raw data) is often used to evaluate relationships involving at least one qualitative ordinal variable or two quantitative variables if the link is partially linear
- **Kendall's tau-b** which is computed from the number of concordant and discordant pairs is often used for qualitative ordinal variables

Note that there exists the *point-biserial correlation* (which can be used to measure the association between a continuous variable and a nominal variable of two levels), but this correlation is not covered here.

Correlation matrix: correlations for all variables

Suppose now that we want to compute correlations for several pairs of variables. We can easily do so for all possible pairs of variables in the dataset, again with the `cor()` function:

```
# correlation for all variables
round(cor(dat),
  digits = 2 # rounded to 2 decimals
)
##        mpg   cyl  disp    hp  drat    wt  qsec  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.49  0.53
```

```
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00 -0.21 -0.66
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66  0.27  1.00
```

This correlation matrix gives an overview of the correlations for all combinations of two variables.

Interpretation of a correlation coefficient

First of all, correlation ranges from **-1 to 1**. It gives us an indication on two things:

1. The direction of the relationship between the 2 variables
2. The strength of the relationship between the 2 variables

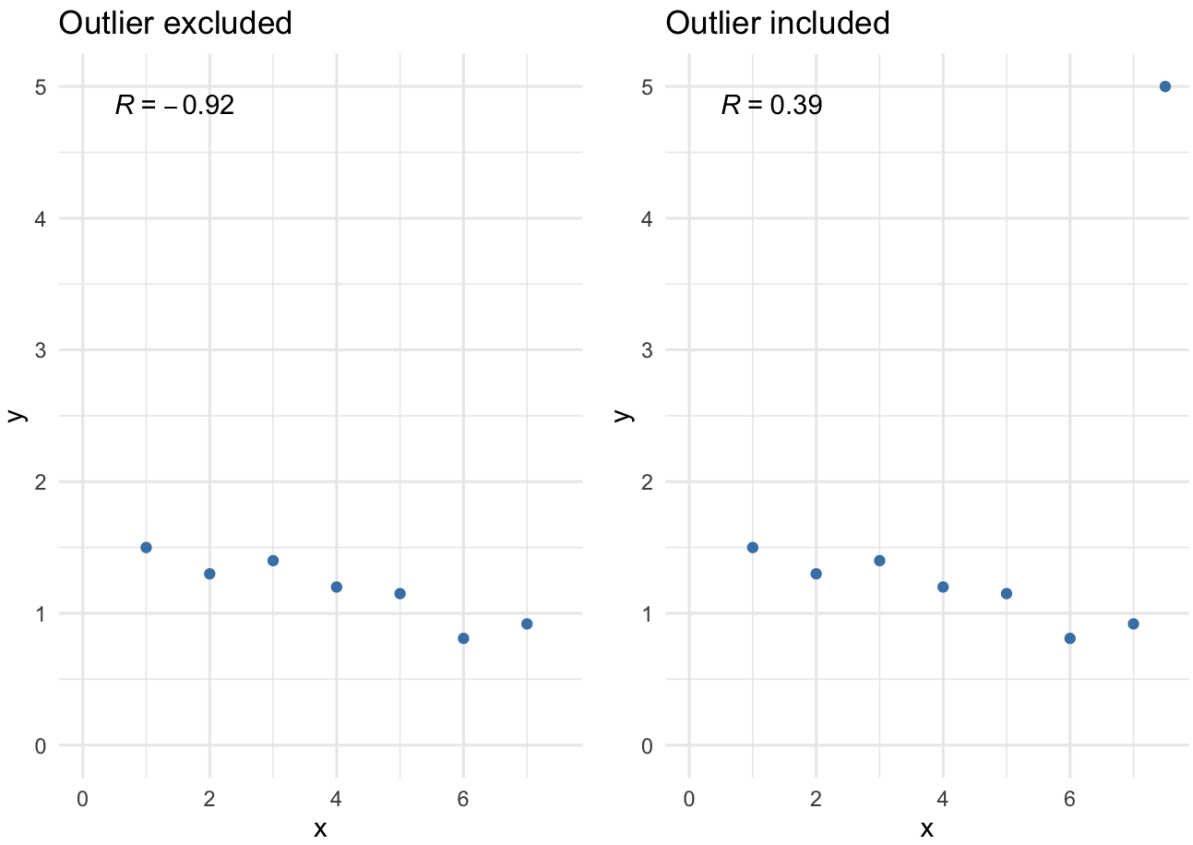
Regarding the direction of the relationship: On the one hand, a **negative correlation** implies that the two variables under consideration vary in **opposite directions**, that is, if a variable increases the other decreases and vice versa. On the other hand, a **positive correlation** implies that the two variables under consideration vary in the **same direction**, i.e., if a variable increases the other one increases and if one decreases the other one decreases as well.

Regarding the strength of the relationship: The **more extreme** the correlation coefficient (the closer to -1 or 1), the **stronger the relationship**. This also means that a **correlation close to 0** indicates that the two variables are **independent**, that is, as one variable increases, there is no tendency in the other variable to either decrease or increase.

As an illustration, the Pearson correlation between horsepower (`hp`) and miles per gallon (`mpg`) found above is -0.78, meaning that the 2 variables vary in opposite direction. This makes sense, cars with more horsepower tend to consume more fuel (and thus have a lower millage par gallon). On the contrary, from the correlation matrix we see that the correlation between miles per gallon (`mpg`) and the time to drive 1/4 of a mile (`qsec`) is 0.42, meaning that fast cars (low `qsec`) tend to have a worse millage per gallon (low `mpg`). This again make sense as fast cars tend to consume more fuel.

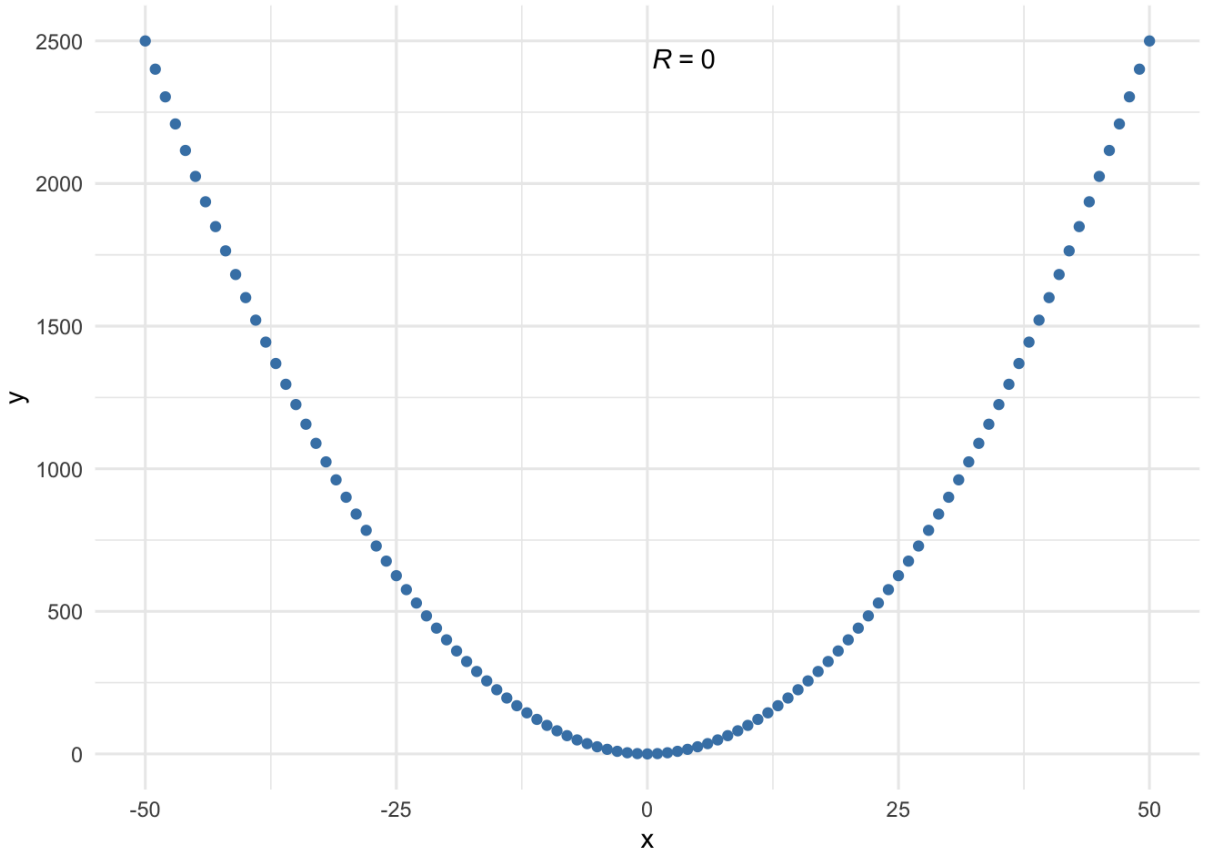
Note that it is a good practice to visualize the type of the relationship between the two variables *before* interpreting the correlation coefficients. The reason is that the correlation coefficient could be biased due to an outlier or due to the type of link between the two variables.

For instance, see the two Pearson correlation coefficients (denoted by `R` in the following plots) when the outlier is excluded and included:



The Pearson correlation coefficient changes drastically due to a single point, and thus the interpretation. It goes from a negative correlation coefficient, indicating a negative relationship between the 2 variables, to a positive coefficient, indicating a positive relationship. We would have missed this insight if we had not visualized the data in a scatterplot (see how to draw a scatterplot in this section).

A correlation coefficient may also miss a non-linear link between two variables:



The Pearson correlation coefficient is equal to 0, indicating no relationship between the two variables, because it measures the **linear** relationship and it is clear from the plot that the link is non-linear.

So to recap, it is a good practice to visualize the data via a scatterplot before interpreting a correlation coefficient (it does not tell the whole story) and see how the correlation coefficient changes when using the parametric (Pearson) or nonparametric version (Spearman or Kendall's tau-b).

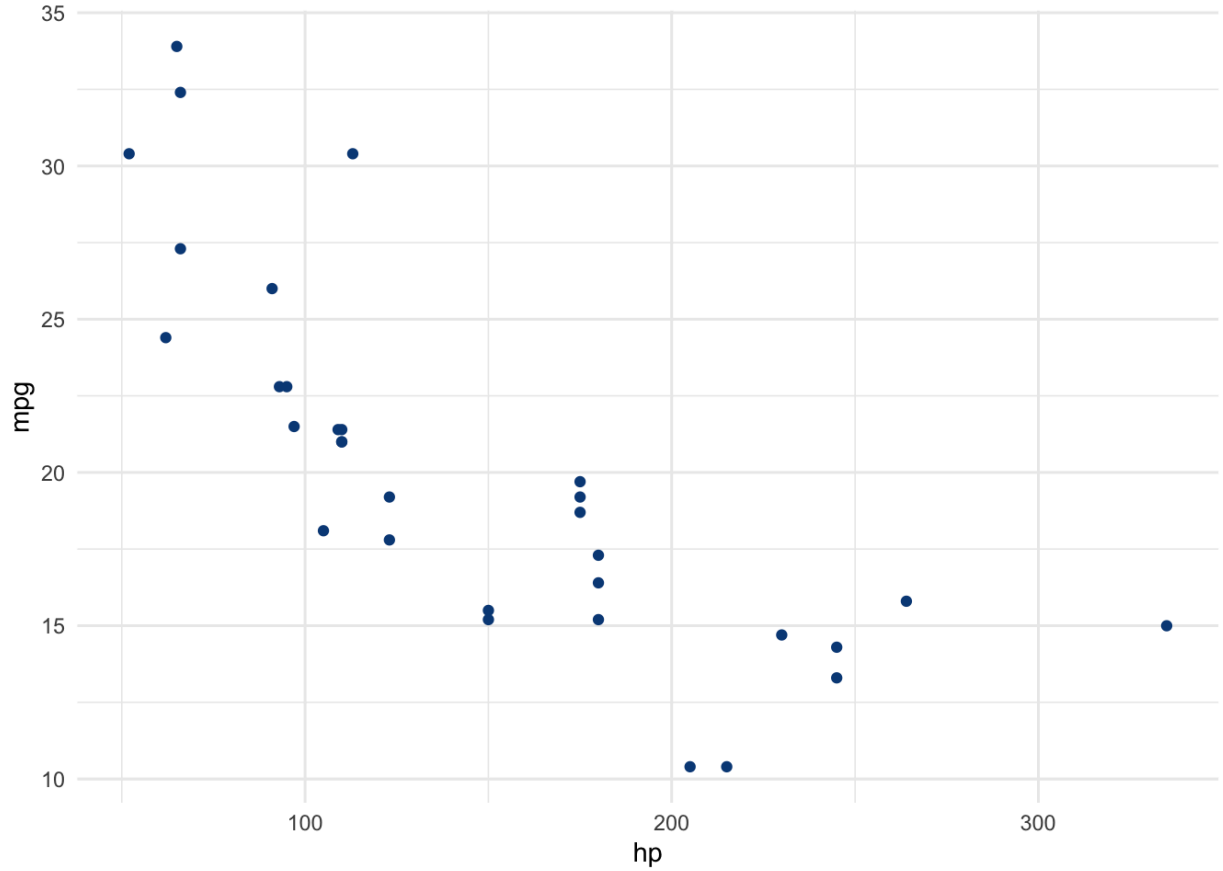
Visualizations

The correlation matrix presented above is not easily interpretable, especially when the dataset is composed of many variables. In the following sections, we present some alternatives to the correlation matrix for better readability.

A scatterplot for 2 variables

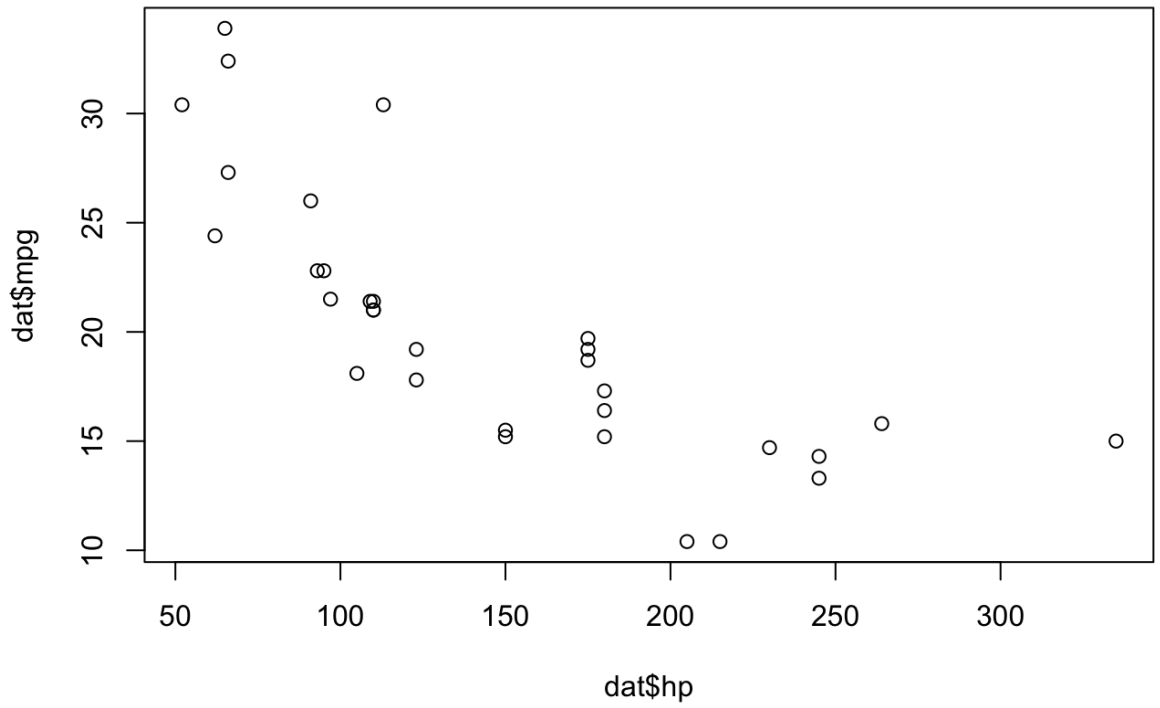
A good way to visualize a correlation between 2 variables is to draw a scatterplot of the two variables of interest. Suppose we want to examine the relationship between horsepower (`hp`) and miles per gallon (`mpg`):

```r
# scatterplot
library(ggplot2)

ggplot(dat) +
  aes(x = hp, y = mpg) +
  geom_point(colour = "#0c4c8a") +
  theme_minimal()
```



If you are unfamiliar with the `{ggplot2}` package, you can draw the scatterplot using the `plot()` function from R base graphics:

```r
plot(dat$hp, dat$mpg)
```



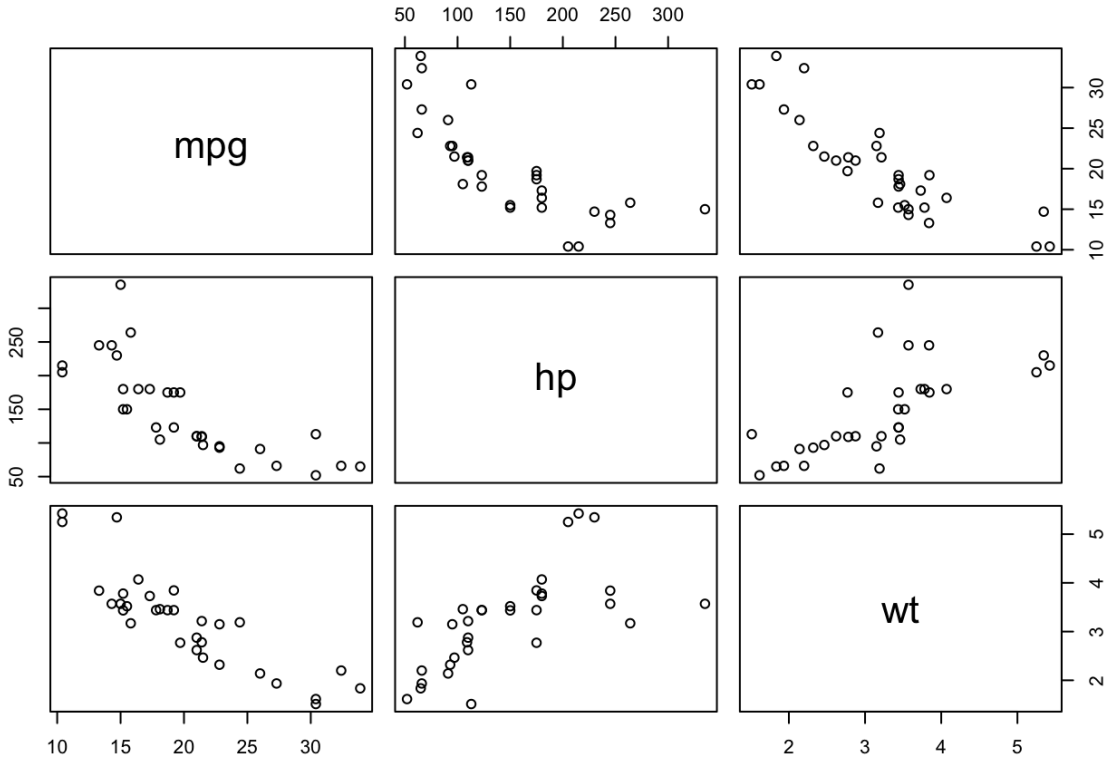
or use the esquisse addin to easily draw plots using the `{ggplot2}` package.

Scatterplots for several pairs of variables

Suppose that instead of visualizing the relationship between only 2 variables, we want to visualize the relationship for several pairs of variables. This is possible thanks to the `pair()` function.

For this illustration, we focus only on miles per gallon (`mpg`), horsepower (`hp`) and weight (`wt`):

```
# multiple scatterplots
pairs(dat[, c("mpg", "hp", "wt")])
```

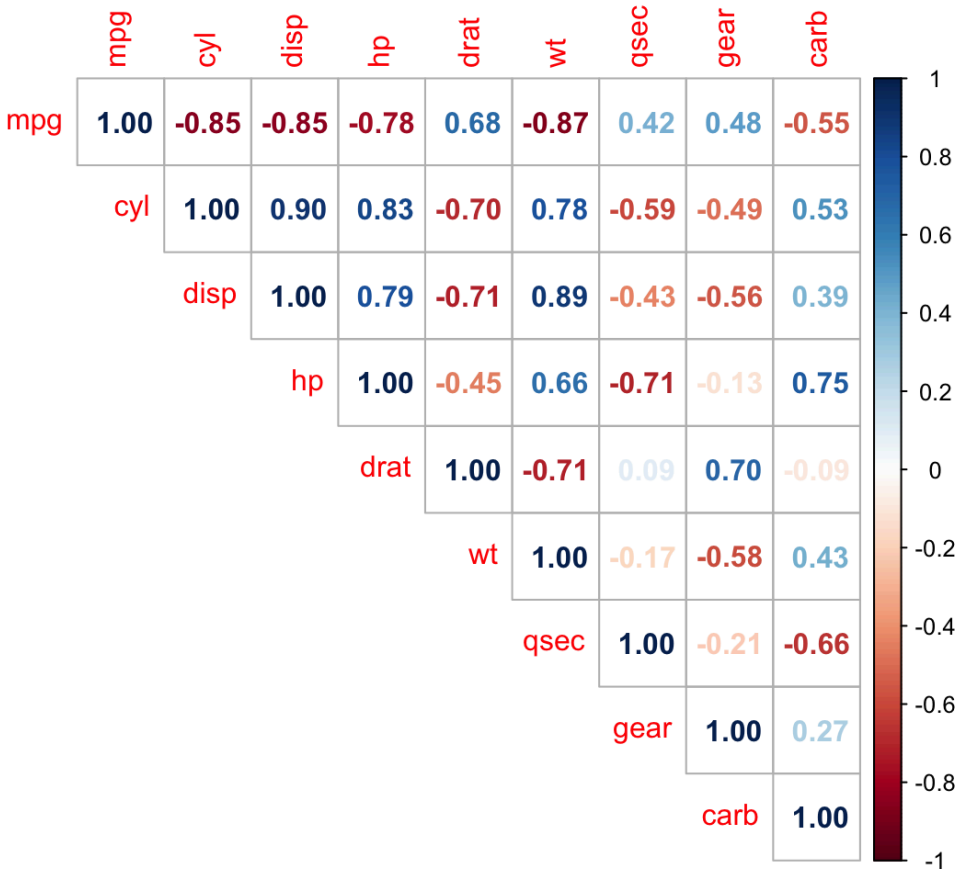


The figure indicates that weight (`wt`) and horsepower (`hp`) are positively correlated, whereas miles per gallon (`mpg`) seems to be negatively correlated with horsepower (`hp`) and weight (`wt`).

Another simple correlation matrix

This version of the correlation matrix presents the correlation coefficients in a slightly more readable way, i.e., by coloring the coefficients based on their sign. Applied to our dataset, we have:

```
# improved correlation matrix
library(corrplot)

corrplot(cor(dat),
  method = "number",
  type = "upper" # show only upper side
)
```



Correlation test

For 2 variables

Unlike a correlation matrix which indicates the correlation coefficients between some pairs of variables in the sample, a correlation test is used to test whether the correlation (denoted ρ) between 2 variables is significantly different from 0 or not in the population.

Actually, a correlation coefficient different from 0 in the sample does not mean that the correlation is **significantly** different from 0 in the population. This needs to be tested with a hypothesis test—and known as the correlation test.

The null and alternative hypothesis for the correlation test are as follows:

- $H_0 0$: $\rho = 0 = 0$ (meaning that there is no linear relationship between the two variables)
- $H_1 1$: $\rho \neq 0 \neq 0$ (meaning that there is a linear relationship between the two variables)

Via this correlation test, what we are actually testing is whether:

- the sample contains sufficient evidence to reject the null hypothesis and conclude that the correlation coefficient does not equal 0, so the relationship exists in the population.
- or on the contrary, the sample does not contain enough evidence that the correlation coefficient does not equal 0, so in this case we do not reject the null hypothesis of no relationship between the variables in the population.

Note that there are 2 assumptions for this test to be valid:

- Independence of the data
- For small sample sizes (usually $n<30$), the two variables should follow a normal distribution

Suppose that we want to test whether the rear axle ratio (`drat`) is correlated with the time to drive a quarter of a mile (`qsec`):

```
# Pearson correlation test
test <- cor.test(dat$drat, dat$qsec)
test
##
##     Pearson's product-moment correlation
##
## data:  dat$drat and dat$qsec
## t = 0.50164, df = 30, p-value = 0.6196
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.265947  0.426340
## sample estimates:
##       cor
## 0.09120476
```

The *p*-value of the correlation test between these 2 variables is 0.62. At the 5% significance level, we do not reject the null hypothesis of no correlation. We therefore conclude that we do not reject the hypothesis that there is no linear relationship between the 2 variables.[2]

This test proves that even if the correlation coefficient is different from 0 (the correlation is 0.09 in the sample), it is actually not significantly different from 0 in the population.

Note that the *p*-value of a correlation test is based on the correlation coefficient **and** the sample size. The larger the sample size and the more extreme the correlation (closer to -1 or 1), the more likely the null hypothesis of no correlation will be rejected.

With a small sample size, it is thus possible to obtain a *relatively* large correlation in the sample (based on the correlation coefficient), but still find a correlation not

significantly different from 0 in the population (based on the correlation test). For this reason, it is recommended to always perform a correlation test before interpreting a correlation coefficient to avoid flawed conclusions.

For several pairs of variables

Similar to the correlation matrix used to compute correlation for several pairs of variables, the `rcorr()` function (from the `{Hmisc}` package) allows to compute *p*-values of the correlation test for several pairs of variables at once. Applied to our dataset, we have:

```
# correlation tests for whole dataset
library(Hmisc)
res <- rcorr(as.matrix(dat)) # rcorr() accepts matrices only

# display p-values (rounded to 3 decimals)
round(res$P, 3)
##        mpg   cyl  disp    hp  drat    wt  qsec  gear  carb
## mpg     NA 0.000 0.000 0.000 0.000 0.000 0.017 0.005 0.001
## cyl  0.000    NA 0.000 0.000 0.000 0.000 0.000 0.004 0.002
## disp 0.000 0.000    NA 0.000 0.000 0.000 0.013 0.001 0.025
## hp   0.000 0.000 0.000    NA 0.010 0.000 0.000 0.493 0.000
## drat 0.000 0.000 0.000 0.010    NA 0.000 0.620 0.000 0.621
## wt   0.000 0.000 0.000 0.000 0.000    NA 0.339 0.000 0.015
## qsec 0.017 0.000 0.013 0.000 0.620 0.339    NA 0.243 0.000
## gear 0.005 0.004 0.001 0.493 0.000 0.000 0.243    NA 0.129
## carb 0.001 0.002 0.025 0.000 0.621 0.015 0.000 0.129    NA
```

Only correlations with *p*-values smaller than the significance level (usually $\alpha=0.05$) should be interpreted.

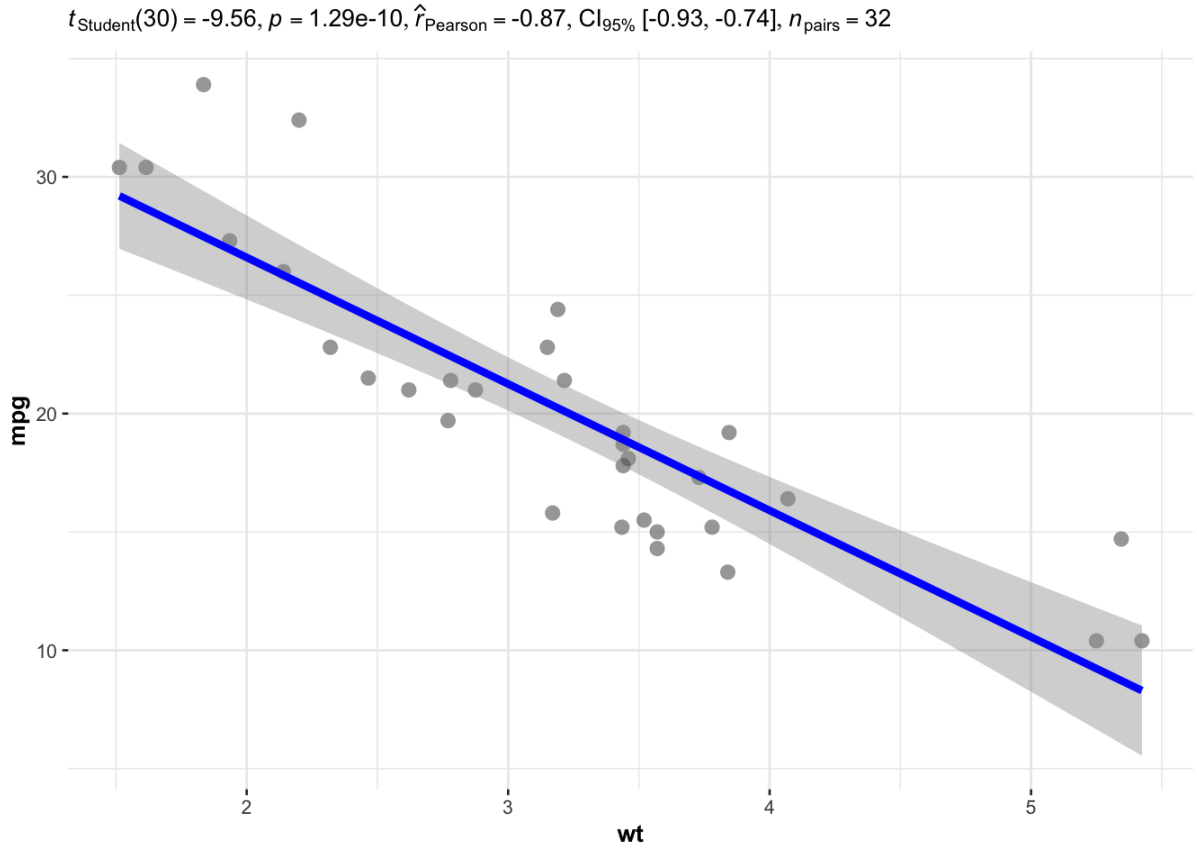
Combination of correlation coefficients and correlation tests

Now that we covered the concepts of correlation coefficients and correlation tests, let see if we can combine the two concepts.

If you need to do this for a few pairs of variables, I recommend using the `ggscatterstats()` function from the `{ggstatsplot}` package. Let's see it in practice with one pair of variables—`wt` and `mpg`:

```
## plot with statistical results
library(ggstatsplot)

ggscatterstats(
  data = dat,
```

```
  x = wt,
  y = mpg,
  bf.message = FALSE,
  marginal = FALSE # remove histograms
)
```



Based on the result of the test, we conclude that there is a negative correlation between the weight and the number of miles per gallon ($r=-0.87$, value < 0.001).

If you need to do it for many pairs of variables, I recommend using the the `correlation` function from the [easystats `{correlation}` package].

This function allows to combine correlation coefficients and correlation tests for *several pairs* of variables, all in a single table (thanks to krzysiektr for pointing it out to me):

```
library(correlation)

correlation::correlation(dat,
  include_factors = TRUE, method = "auto"
)
## # Correlation Matrix (auto-method)
##
```

```
## Parameter1 | Parameter2 |     r |         95% CI | t(30) |         p
## ------------------------------------------------------------------
## mpg        |        cyl | -0.85 | [-0.93, -0.72] | -8.92 | < .001***
## mpg        |       disp | -0.85 | [-0.92, -0.71] | -8.75 | < .001***
## mpg        |         hp | -0.78 | [-0.89, -0.59] | -6.74 | < .001***
## mpg        |       drat |  0.68 | [ 0.44,  0.83] |  5.10 | < .001***
## mpg        |         wt | -0.87 | [-0.93, -0.74] | -9.56 | < .001***
## mpg        |       qsec |  0.42 | [ 0.08,  0.67] |  2.53 | 0.137
## mpg        |       gear |  0.48 | [ 0.16,  0.71] |  3.00 | 0.065
## mpg        |       carb | -0.55 | [-0.75, -0.25] | -3.62 | 0.016*
## cyl        |       disp |  0.90 | [ 0.81,  0.95] | 11.45 | < .001***
## cyl        |         hp |  0.83 | [ 0.68,  0.92] |  8.23 | < .001***
## cyl        |       drat | -0.70 | [-0.84, -0.46] | -5.37 | < .001***
## cyl        |         wt |  0.78 | [ 0.60,  0.89] |  6.88 | < .001***
## cyl        |       qsec | -0.59 | [-0.78, -0.31] | -4.02 | 0.007**
## cyl        |       gear | -0.49 | [-0.72, -0.17] | -3.10 | 0.054
## cyl        |       carb |  0.53 | [ 0.22,  0.74] |  3.40 | 0.027*
## disp       |         hp |  0.79 | [ 0.61,  0.89] |  7.08 | < .001***
## disp       |       drat | -0.71 | [-0.85, -0.48] | -5.53 | < .001***
## disp       |         wt |  0.89 | [ 0.78,  0.94] | 10.58 | < .001***
## disp       |       qsec | -0.43 | [-0.68, -0.10] | -2.64 | 0.131
## disp       |       gear | -0.56 | [-0.76, -0.26] | -3.66 | 0.015*
## disp       |       carb |  0.39 | [ 0.05,  0.65] |  2.35 | 0.177
## hp         |       drat | -0.45 | [-0.69, -0.12] | -2.75 | 0.110
## hp         |         wt |  0.66 | [ 0.40,  0.82] |  4.80 | < .001***
## hp         |       qsec | -0.71 | [-0.85, -0.48] | -5.49 | < .001***
## hp         |       gear | -0.13 | [-0.45,  0.23] | -0.69 | > .999
## hp         |       carb |  0.75 | [ 0.54,  0.87] |  6.21 | < .001***
## drat       |         wt | -0.71 | [-0.85, -0.48] | -5.56 | < .001***
## drat       |       qsec |  0.09 | [-0.27,  0.43] |  0.50 | > .999
## drat       |       gear |  0.70 | [ 0.46,  0.84] |  5.36 | < .001***
## drat       |       carb | -0.09 | [-0.43,  0.27] | -0.50 | > .999
## wt         |       qsec | -0.17 | [-0.49,  0.19] | -0.97 | > .999
## wt         |       gear | -0.58 | [-0.77, -0.29] | -3.93 | 0.008**
## wt         |       carb |  0.43 | [ 0.09,  0.68] |  2.59 | 0.132
## qsec       |       gear | -0.21 | [-0.52,  0.15] | -1.19 | > .999
## qsec       |       carb | -0.66 | [-0.82, -0.40] | -4.76 | < .001***
## gear       |       carb |  0.27 | [-0.08,  0.57] |  1.56 | 0.774
##
## p-value adjustment method: Holm (1979)
## Observations: 32
```

As you can see, it gives, among other useful information, the correlation coefficients (column `r`) and the result of the correlation test (column `95% CI` for the confidence interval or `p` for the p-value) for all pairs of variables.

Correlograms

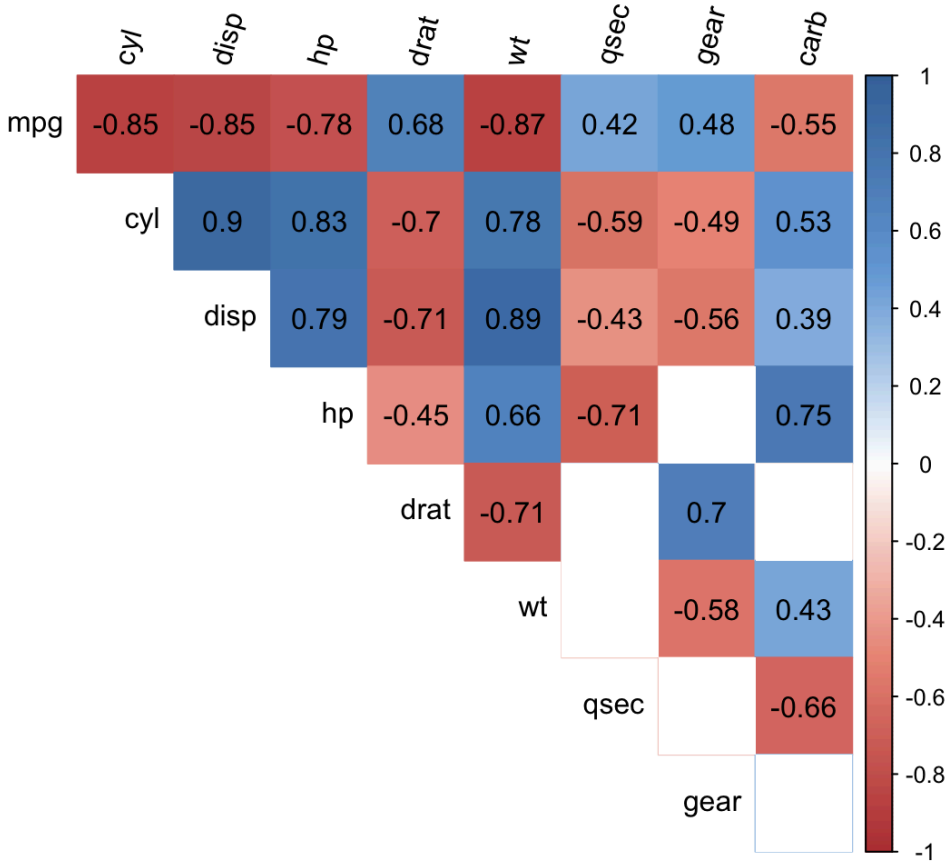
The table above is very useful and informative, but let see if it is possible to combine the concepts of correlation coefficients and correlations test in one single visualization. A visualization that would be easy to read and interpret.

Ideally, we would like to have a concise overview of correlations between all possible pairs of variables present in a dataset, with a clear distinction for correlations that are significantly different from 0.

The figure below, known as a correlogram and adapted from the `corrplot()` function, does precisely this:

```
# do not edit
corrplot2 <- function(data,
                      method = "pearson",
                      sig.level = 0.05,
                      order = "original",
                      diag = FALSE,
                      type = "upper",
                      tl.srt = 90,
                      number.font = 1,
                      number.cex = 1,
                      mar = c(0, 0, 0, 0)) {
  library(corrplot)
  data_incomplete <- data
  data <- data[complete.cases(data), ]
  mat <- cor(data, method = method)
  cor.mtest <- function(mat, method) {
    mat <- as.matrix(mat)
    n <- ncol(mat)
    p.mat <- matrix(NA, n, n)
    diag(p.mat) <- 0
    for (i in 1:(n - 1)) {
      for (j in (i + 1):n) {
        tmp <- cor.test(mat[, i], mat[, j], method = method)
        p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
      }
    }
    colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
    p.mat
  }
  p.mat <- cor.mtest(data, method = method)
    col  <-  colorRampPalette(c("#BB4444",  "#EE9988",  "#FFFFFF",  "#77AADD",
"#4477AA"))
  corrplot(mat,
    method = "color", col = col(200), number.font = number.font,
    mar = mar, number.cex = number.cex,
    type = type, order = order,
    addCoef.col = "black", # add correlation coefficient
    tl.col = "black", tl.srt = tl.srt, # rotation of text labels
    # combine with significance level
```

```
    p.mat = p.mat, sig.level = sig.level, insig = "blank",
    # hide correlation coefficients on the diagonal
    diag = diag
  )
}

# edit from here
corrplot2(
  data = dat,
  method = "pearson",
  sig.level = 0.05,
  order = "original",
  diag = FALSE,
  type = "upper",
  tl.srt = 75
)
```



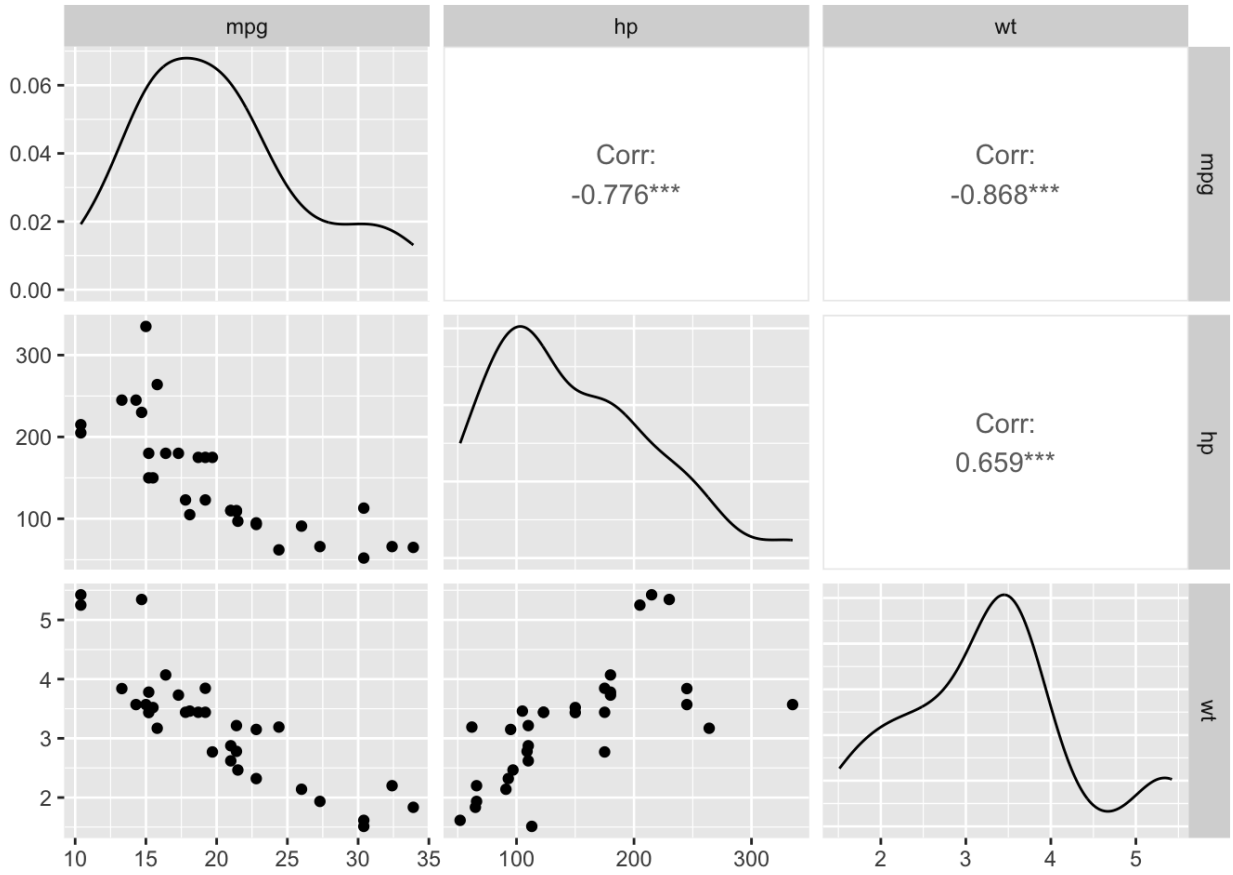
The correlogram shows correlation coefficients for all pairs of variables (with more intense colors for more extreme correlations), and correlations not significantly different from 0 are represented by a white box.

To learn more about this plot and the code used, I invite you to read the article entitled "Correlogram in R: how to highlight the most correlated variables in a dataset".

For those of you who are still not completely satisfied, I recently found two alternatives—one with the `ggpairs()` function from the `{GGally}` package and one with the `ggcormat()` function from the `{ggstatsplot}` package.

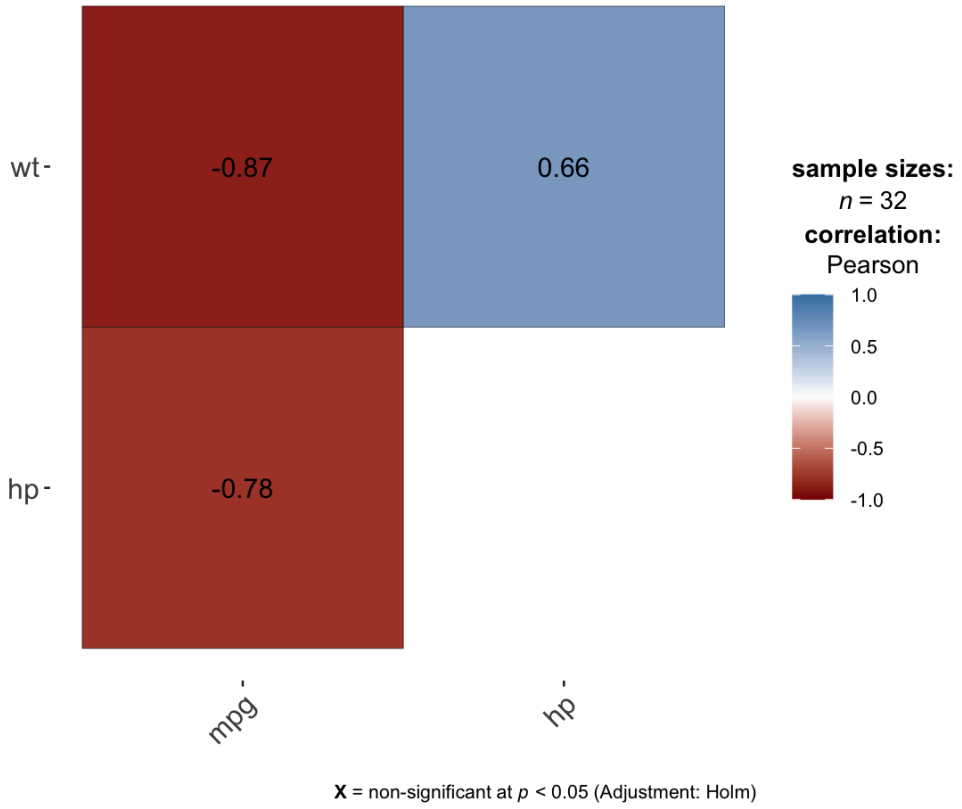
The two functions are illustrated with the variables `mpg`, `hp` and `wt`:

```
library(GGally)

ggpairs(dat[, c("mpg", "hp", "wt")])
```



The plot above combines correlation coefficients, correlation tests (via the asterisks next to the coefficients[3]) and scatterplots for all possible pairs of variables present in a dataset.

```
library(ggstatsplot)

ggcorrmat(
  data = dat[, c("mpg", "hp", "wt")],
  type = "parametric", # parametric for Pearson, nonparametric for Spearman's correlation
  colors = c("darkred", "white", "steelblue") # change default colors
)
```



The plot above also shows the correlation coefficients and if any, the non-significant correlations (by default at the 5% significance level with the Holm adjustment method) are shown by a big cross on the correlation coefficients.

The advantage of these two alternatives compared to the first one is that it is directly available within a package, so you do not need to run the code of the function first in order to draw the correlogram.

Correlation does not imply causation

I am pretty sure you have already heard the statement "Correlation does not imply causation" in statistics. An article about correlation would not be complete without discussing about causation.

A non-zero correlation between two variables does not necessarily mean that there is a cause and effect relationship between these two variables!

Indeed, a significant correlation between two variables means that changes in one variable are associated (positively or negatively) with changes in the other variable. Nonetheless, a significant correlation *does not* indicates that variations in one variable *cause* the variations in the other variable.

A non-zero correlation between X and Y can appear in several cases:

- X causes Y
- Y causes X
- a third variable cause X and Y
- a combination of these three reasons

Sometimes it is quite clear that there is a causal relationship between two variables. Take for example the correlation between the price of a consumer product such as milk and its consumption. It is quite obvious that there is a causal link between the two: if the price of milk increases, it is expected that its consumption will decrease.

However, this causal link is not always present even if the correlation is significant. Maurage, Heeren, and Pesenti (2013) showed that, although there is a positive and significant correlation between chocolate consumption and the number of Nobel laureates, this correlation comes from the fact that a third variable, Gross Domestic Product (GDP), causes chocolate consumption and the number of Nobel laureates. They found that countries with higher GDP tend to have a higher level of chocolate consumption and scientific research (leading to more Nobel laureates).

This example shows that one must be very cautious when interpreting correlations and avoid over-interpreting a correlation as a causal relationship.

Probability Distributions:

What is Probability Distribution?

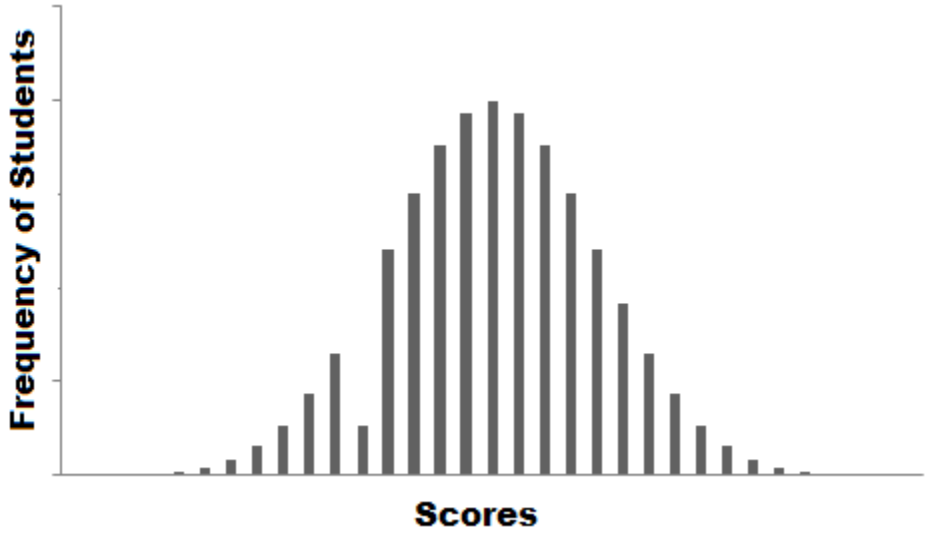
A probability distribution is a mathematical function that defines the likelihood of different outcomes or values of a variable. This function is commonly represented by a graph or probability table, and it provides the probabilities of various possible results of an experiment or random phenomenon based on the sample space and the probabilities of events. Probability distributions are fundamental in probability theory and statistics for analyzing data and making predictions.

Example of Probability Distribution

Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students. You gave these graded papers to a data entry guy in the university and told him to create a spreadsheet containing the grades of all the students. But the guy only stores the grades and not the corresponding students.

S. No.	Scores
1	25
2	27
3	38
4	42
5	
6	16
7	35
8	46
9	48
10	31

He made another blunder; he missed a few entries in a hurry, and we have no idea whose grades are missing. One way to find this out is by visualizing the grades and seeing if you can find a trend in the data.



The graph you plotted is called the frequency distribution of the data. You see that there is a smooth curve-like structure that defines our data, but do you notice an anomaly? We have an abnormally low frequency at

a particular score range. So the best guess would be to have missing values that remove the dent in the distribution.

Need of Probability Distribution

Probability distributions are versatile tools used in various fields and applications. They primarily model and quantify uncertainty and variability in data, making them fundamental in data science, statistics, and decision-making processes. Probability distributions enable us to analyze data and draw meaningful conclusions by describing the likelihood of different outcomes or events.

In statistical analysis, these distributions play a pivotal role in parameter estimation, hypothesis testing, and data inference. They also find extensive use in risk assessment, particularly in finance and insurance, where they help assess and manage financial risks by quantifying the likelihood of various outcomes.

Machine learning algorithms leverage probability distributions to model uncertainty in predictions, enhancing their ability to make accurate forecasts. Additionally, probability distributions support quality control efforts, allowing for the monitoring and controlling processes by identifying deviations from expected values.

Probability distributions are not confined to data analysis alone; they also play crucial roles in fields like engineering, environmental science, epidemiology, and physics. In these diverse domains, probability distributions enable reliable modeling, simulation, and prediction, ultimately contributing to informed decision-making and problem-solving.

Common Data Types

Before we jump on to the explanation of distributions, let’s see what kind of data we can encounter. The data can be discrete or continuous.

- **Discrete Data**, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5, or 6, not 1.5 or 2.45. (Discrete Probability Distribution)
- **Continuous Data** can take any value within a given range. The range may be finite or infinite. For example, a girl's weight or height, the length of the road. The weight of a girl can be any value – 54 kgs, 54.5 kgs, or 54.5436kgs. (Continuous Probability Distribution)

Now let us start with the types of distributions.

Types of Distributions

Here is a list of distributions types

1. Bernoulli Distribution
2. Uniform Distribution
3. Binomial Distribution
4. Normal or Gaussian Distribution
5. Exponential Distribution
6. Poisson Distribution

Bernoulli Distribution

Let's start with the easiest distribution, which is Bernoulli Distribution. It is actually easier to understand than it sounds!

All you cricket junkies out there! At the beginning of any cricket match, how do you decide who will bat or ball? A toss! It all depends on whether you win or lose the toss, right? Let's say if the toss results in a head, you win. Else, you lose. There's no midway.

A **Bernoulli distribution** has only two bernoulli trials or possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X with a Bernoulli distribution can take the value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure.
Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is given by: px(1-p)1-x where x € (0, 1)

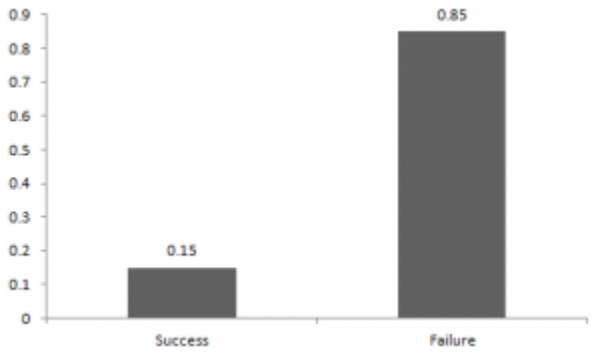
It can also be written as:

$$P(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

The probabilities of success and failure need not be equally likely, like the result of a fight between Undertaker and me. He is pretty much certain to win. So, in this case probability of my success is 0.15, while my failure is 0.85

Bernoulli Distribution Example

Here, the probability of success(p) is not the same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.



Here, the probability of success = 0.15, and the probability of failure = 0.85. The expected value is exactly what it sounds like. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows:

E(X) = 1*p + 0*(1-p) = p

The variance of a random variable from a bernoulli distribution is:

$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p)$

There are many examples of Bernoulli distribution, such as whether it will rain tomorrow or not, where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely, which is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$

The graph of a uniform distribution curve looks like

You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

Uniform Distribution Example

The number of bouquets sold daily at a flower shop is uniformly distributed, with a maximum of 40 and a minimum of 10.

Let’s try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution are:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = $(b-a)^2/12$

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Binomial Distribution

Let’s get back to cricket. Suppose you won the toss today, indicating a successful event. You toss again, but you lose this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let’s assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, the probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is the same for all the trials is called a Binomial Distribution.

Binomial Distribution Example

The outcomes need not be equally likely. Remember the example of a fight between Undertaker and me? So, if the probability of success in an

experiment is 0.2, then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p, where n is the total number of trials and p is the probability of success in each trial.

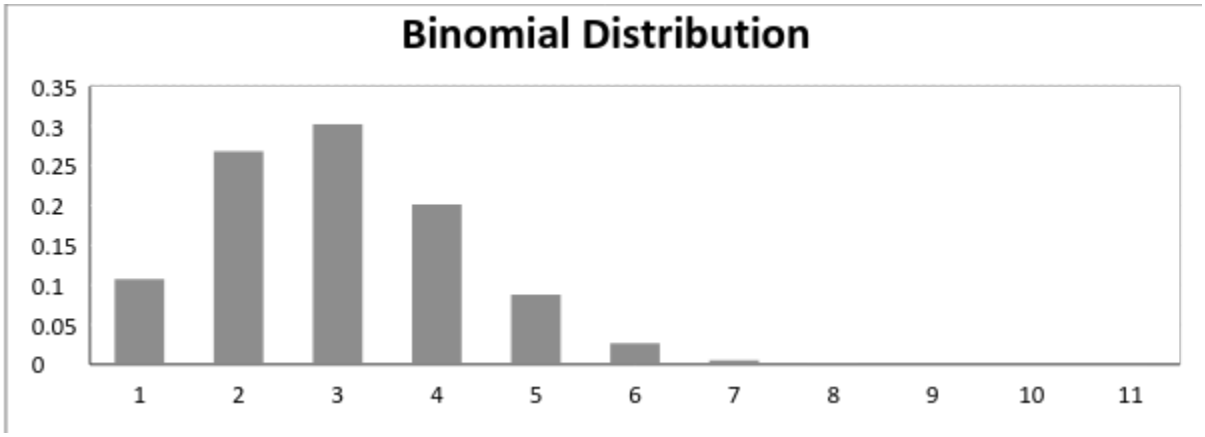
Based on the above explanation, the properties of a Binomial Distribution are:

1. Each trial is independent.
2. There are only two possible outcomes in a trial – success or failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is the same for all trials. (Trials are identical.)

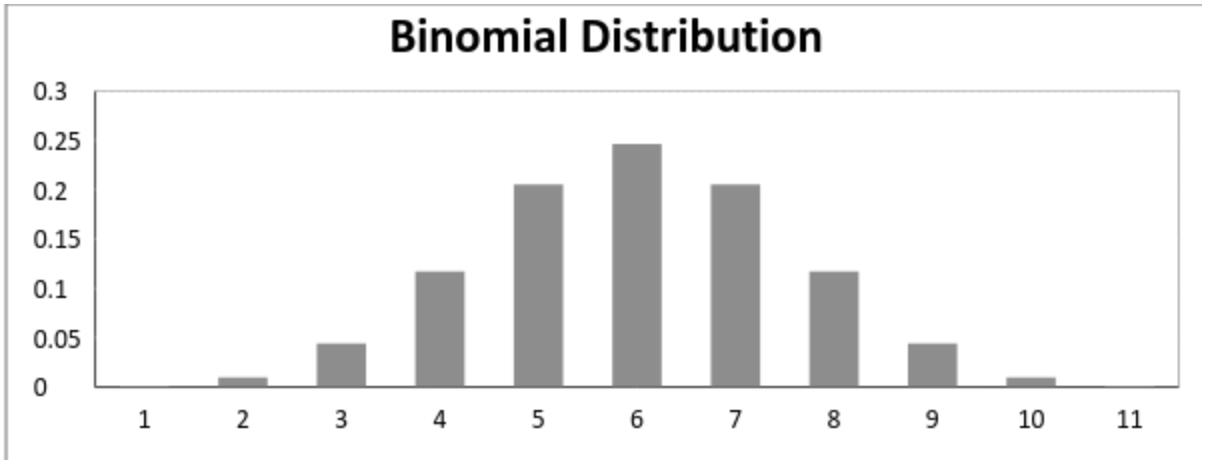
The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like this.



Now, when the probability of success = probability of failure, in such a situation, the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

- Mean -> μ = n*p
- Variance -> Var(X) = n*p*q

Normal Distribution vs Gaussian Distribution

The **normal distribution** represents the behavior of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median, and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center, and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity, then the shapes will be quite similar.

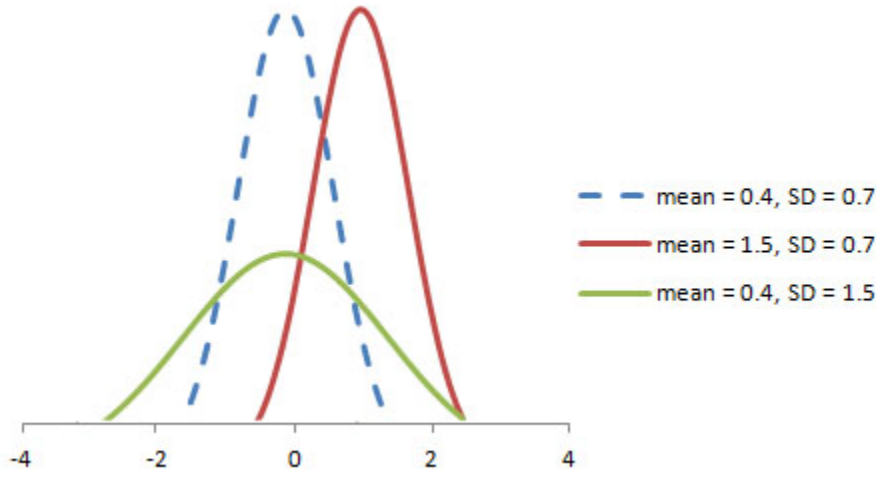
The PDF of a random variable X, following a normal distribution, is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

The mean and variance of a random variable X, which is said to be normally distributed, is given by:

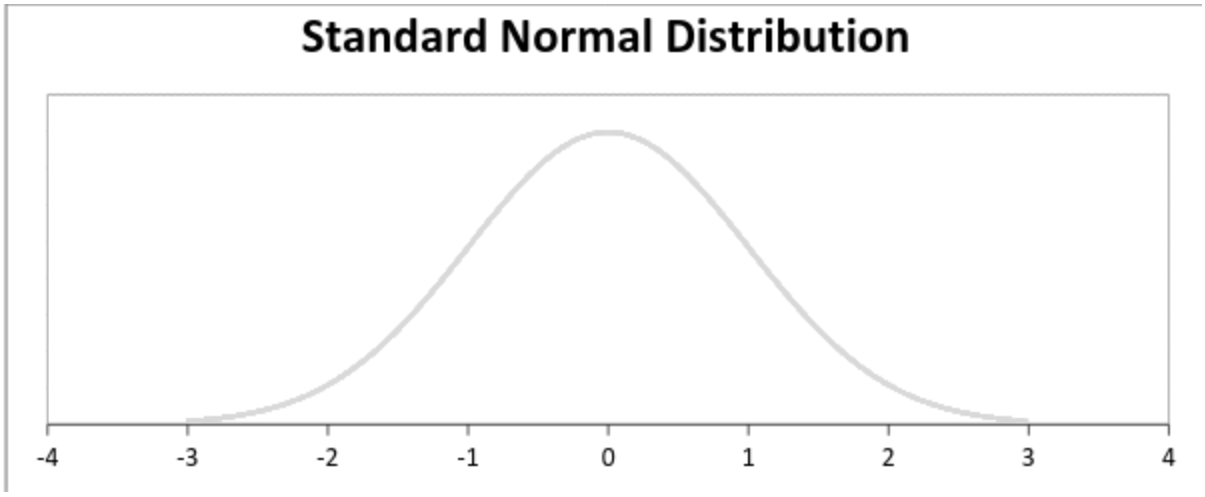
- Mean -> E(X) = μ
- Variance -> Var(X) = σ^2

Here, μ (mean) and σ (standard deviation) are the parameters.
The graph of a random variable X ~ N (μ, σ) is shown below.



A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1. For such a case, the PDF becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad for -\infty < x < \infty$$



Poisson Distribution

Suppose you work at a call center; approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are:

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area in a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors on each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

Poisson Distribution Example

A distribution is called a **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.

2. The probability of success over a short interval must equal its probability over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions, then it is a Poisson distribution. Some notations used in Poisson distribution are:

- λ is the rate at which an event occurs,
- t is the length of a time interval,
- And X is the number of events in that time interval.

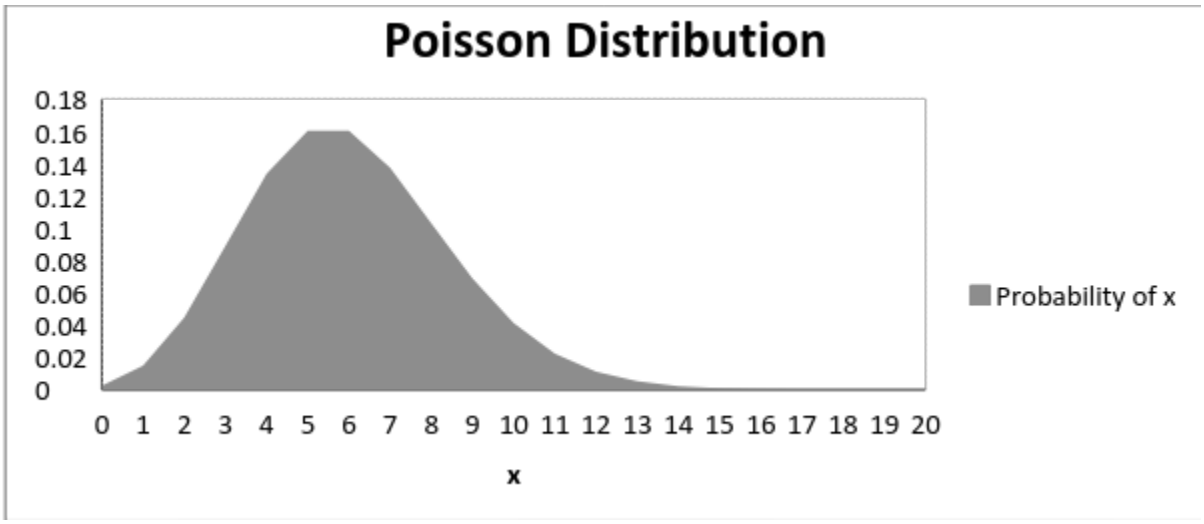
Here, X is called a Poisson Random Variable, and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

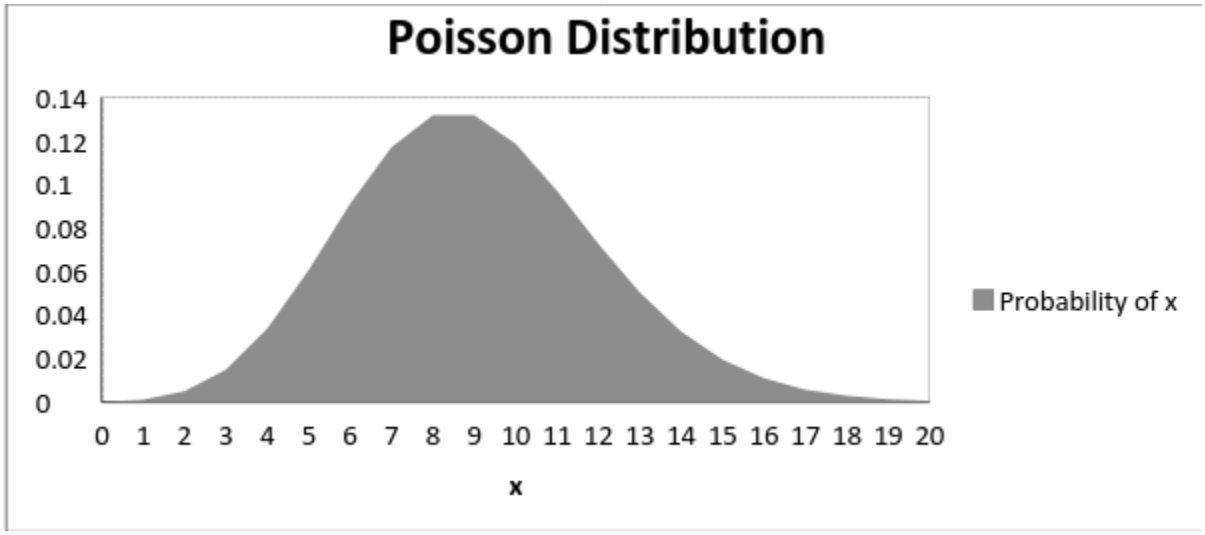
The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots..$$

The mean μ is the parameter of this distribution. μ is also defined as the λ times the length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to the increase in the mean.



It is perceptible that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution:

Mean -> E(X) = μ
Variance -> Var(X) = μ

Exponential Distribution

Let's consider the call center example one more time. What about the interval of time between the calls? Here, the exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time between metro arrivals
2. Length of time between arrivals at a gas station
3. The life of an air conditioner

The exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an **exponential distribution** with PDF:

$f(x) = \{ \lambda e^{-\lambda x}, \quad x \geq 0$

And parameter **$\lambda > 0$,** which is also called the rate.

For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

- Mean -> $E(X) = 1/\lambda$
- Variance -> $Var(X) = (1/\lambda)^2$

Also, the greater the rate, the faster the curve drops, and the lower the rate, the flatter the curve. This is explained better with the graph shown below.

To ease the computation, there are some formulas given below:

- $P\{X \leq x\} = 1 - e^{-\lambda x}$ corresponds to the area under the density curve to the left of x
- $P\{X > x\} = e^{-\lambda x}$ corresponds to the area under the density curve to the right of x
- $P\{x1 < X \leq x2\} = e^{-\lambda x1} - e^{-\lambda x2}$, corresponds to the area under the density curve between x1 and x2.

Distribution Function in Probability

In probability, the probability density function (PDF) of a continuous random variable serves as a function that interprets the relative likelihood of the random variable matching a given sample within the dataset or sample space. The PDF represents the probability per unit length. Essentially, it allows us to gauge the higher likelihood of the random variable being near one sample compared to another by comparing the values of the PDF at these two samples.

Relations Between the Distributions

Relation Between Bernoulli and Binomial Distribution

- Bernoulli Distribution is a special case of Binomial Distribution with a single trial.
- There are only two possible outcomes of a Bernoulli and Binomial distribution, namely success and failure.
- Both Bernoulli and Binomial Distributions have independent trials.

Relation Between Poisson and Binomial Distribution

Poisson Distribution is a limiting case of binomial distribution under the following conditions:

- The number of trials is indefinitely large or $n \to \infty$.
- The probability of success for each trial is the same and indefinitely small or $p \to 0$.
- $np = \lambda$, is finite.

Relation Between Normal and Binomial Distribution & Normal and Poisson Distribution

A normal distribution is another limiting form of binomial distribution under the following conditions:

- The number of trials is indefinitely large, $n \to \infty$.
- Both p and q are not indefinitely small.

The normal distribution is also a limiting case of Poisson distribution with the parameter $\lambda \to \infty$.

Relation Between Exponential and Poisson Distribution

If the times between random events follow an exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

Test Your Knowledge

You have come this far. Now, are you able to answer the following questions? Let me know in the comments below:

1. The formula to calculate standard normal random variable is:

a. (x+μ) / σ
b. (x-μ) / σ
c. (x-σ) / μ

2. In Bernoulli Distribution, the formula for calculating standard deviation is given by:

a. p (1 – p)
b. SQRT(p(p – 1))
c. SQRT(p(1 – p))

3. For a normal distribution, an increase in the mean will:

a. shift the curve to the left
b. shift the curve to the right
c. flatten the curve

4. The lifetime of a battery is exponentially distributed with λ = 0.05 per hour. The probability for a battery to last between 10 and 15 hours is:

a.0.1341
b.0.1540
c.0.0079

Conclusion

Probability Distributions are prevalent in many sectors, including insurance, physics, engineering, computer science, and even social science, wherein students of psychology and medicine are widely using probability distributions. It has an easy application and widespread use. This article highlighted and explained the application of six important distributions observed in daily life. Now you will be able to identify, relate and differentiate among these distributions.

For a more in-depth write up of these distributions, you can refer this resource.

Key Takeaways

- Probability is commonly used by data scientists to model situations where experiments, independent events conducted during similar circumstances, yield different results, such as throwing dice or a coin.
- Discrete random variables and continuous random variables are two types of quantitative variables. Discrete variables represent counts, for example, the number of objects in a collection, whereas continuous variables represent measurable amounts, for example, water volume or weight.
- Normal distribution, chi-square distribution, binomial distribution, poisson distribution, and uniform distribution are some of the many different classifications of probability distributions.

Frequently Asked Questions

Q1. What distribution is the most commonly used in data science?

A. Gaussian distribution (normal distribution) is famous for its bell-like shape, and it's one of the most commonly used distributions in data science or for Hypothesis Testing.

Q2. What are the 6 common probability distributions every data science professional should know?

A. The 6 common probability distributions are Bernoulli, Uniform, Binomial, Normal, Poisson, and Exponential Distribution.

Q3. What is the difference between a discrete and continuous distribution?

A. A discrete distribution is one in which the data can only take on certain values, and a continuous distribution is one in which data can take on any value within a specified range.

Q4. What are typical types of distribution?

A. Typical types of distribution in data science include normal (Gaussian), uniform, exponential, Poisson, and binomial distributions, each characterizing the probability patterns of different types of data.