ERUDITUS

EXECUTIVE EDUCATION

STATISTICS FOR DATA SCIENCE

Contents

MODULE 1

PROBABILITY - JOINT, CONDITIONAL, MARGINAL

OBJECTIVES

By the end of this module you will be able to:

1. Intuitively understand basic, conditional, joint and marginal probabilities
2. Understand the statistical formula used to calculate these probabilities
3. Think about real life data science problems were these concepts are used
4. Get exposed to python code that implements these concepts

PROBABILITY

When you flip a coin, what are the chances that heads show up ? Intuitively you know the chances are 50-50. How do you arrive at that ?

When you flip a coin, only 1 side can show up, but the 2 sides are equally likely to show up. So the possibility that heads show up is 1 outcome out of 2 possible outcomes.

Possibility of heads = Number of desired outcomes (1) / Total number of outcomes (2) = 0.5

There is only 1 desired outcome - heads. There are 2 possible outcomes - heads or tails.

The word 'Probability' is the statistical term for the words we have used above - chance & possibility. The total number of outcomes that can happen is referred to as the "Sample Space".

So the probability of heads in a coin toss is 0.5 or 50-50. Similarly the probability of tails in a coin toss is the same 0.5. This formula is represented as -

P (heads) = Number of desired outcomes / Total number of outcomes

Hence, probability (0.5 in this case) quantifies the likelihood of an event for a random variable. In this example, the flip of the coin is the random variable. It is a variable because the values can change and it is random because we can't control the outcome.

Probability of 0

What is the probability that heads 'and' tails show up at the same time in a coin toss ?

P (heads-tails) = Number of desired outcomes (0) / Total number of outcomes (2) = 0

So a probability of 0 means that the desired outcome is impossible to occur. In a coin toss, you can never have both heads and tails show up at the same time.

Probability of 1

What is the probability that heads 'or' tails shows up in a coin toss ?

P (heads or tails) = Number of desired outcomes (2) / Total number of outcomes (2) = 1

So a probability of 1 means that the desired outcome will happen 100%. In a coin toss, you can be 100% sure that either a heads or a tails will definitely show up.

Another way to look at this is:

P (heads or tails) = P (heads) + P (tails) = 0.5 + 0.5 = 1

So the sum of probabilities for all outcomes of a random variable should equal 1.

Why is probability relevant to data science and machine learning ?

When you use data science and machine learning to solve real business problems, you are basically using past data to predict the future. For example:

- Use historical sales data to predict future sales
- Use historical customer churn data to predict future churn
- Learn from past images to identify new unknown images
- Use past customer purchase data to predict and recommend new products
- Use past credit history to predict which customers will default

While making these predictions, you will always rely on imperfect, uncertain data. From this uncertain

historical data, you will have to make predictions and present those predictions in probabilistic terms - since you can never be certain what the future will be.

So why is historical data imperfect ?
There are 2 reasons:

- Measurement errors: when data is wrongly captured
- Gaps in data: when data for some periods is not captured

Therefore, your understanding of a basic coin toss probability will help you make probabilistic predictions about the future based on uncertain past data.

However machine learning problems are more complex than a simple coin toss - in a coin toss there is only 1 random variable, but in real world problems there are multiple random variables interacting with each other to determine the future.

You need tools such as Joint probability, Conditional probability and Independent probability to predict the future of multiple random variables.

JOINT PROBABILITY

Let's say the following table represents the top 3 shows on TV streaming service Netstar and the demographic breakdown of the number of viewers.

Shows	Male	Female	Total
Hometown	100,000	50,000	150,000
Sita	50,000	200,000	250,000
Cricket	40,000	40,000	80,000
Total	190,000	290,000	**480,000**

What is the probability of a female viewer watching Sita. There are 2 levels of probability here - first the probability that the viewer is female and second the probability that this female viewer is a viewer of Sita.

P (Female and Sita)

= Number desired outcomes (200,000) / Total number of outcomes (480,000) = 0.42

Thus joint probability describes the probability of the outcomes of two simultaneously occurring random variables (Female viewer + Sita Viewer).

Assuming random variable A represents Female viewers and random variable B represents Sita. Calculating the joint probability using the formula would give us the same answer -

P(A and B) = P (A | B) x P (B) = (200,000 / 250,000 = 0.8) x (250,000 / 480,000 = 0.52) = 0.42

How is Joint probability used in Data Science?

The above scenario is a prime example of how entertainment streaming companies deploy data science. By getting insights on what segment of viewers like what shows, they can allocate the right budgets for new shows and for marketing activities.

Another example is in supply chain & logistics. To predict the congestion of Hubs A and B you would consider the joint probability of trucks coming into Hub A during time duration T and trucks leaving Hub A in the same time duration.

CONDITIONAL PROBABILITY

If Reena just got a Netstar subscription, what is the probability that her favorite show will be Hometown ?

This is a conditional probability situation, because you have already been told that your condition is that the viewer Reena is female. Now you have to estimate the probability of one random variable A (favorite show being Hometown) given a conditional random variable B (female Netstar subscriber).

Shows	Male	Female	Total
Hometown	100,000	50,000	150,000
Sita	50,000	200,000	250,000
Cricket	40,000	40,000	80,000
Total	190,000	290,000	**480,000**

Expressed as a formula:

P (A|B) = P(A and B) / P(B) = (50,000 / 480,000 = 0.1) / (290,000 / 480,000 = 0.6) = 0.17

So there is a 0.17 probability that Reena will watch Hometown. The given conditions in this problem is that Reena is female.

How is Conditional probability used in Data Science ?

It is typically used when time series data (data points collected on successive time intervals) is involved. Events occurring on a particular day depend on what happened on previous days, for example - Share Prices, Weather forecast etc

The weatherman might state that your area has a probability of rain of 40 percent. However, this fact is *conditional* on many things, such as the probability of:

- a cold front coming to your area.
- rain clouds forming.
- another front pushing the rain clouds away.

What is the probability that a Netstar viewer watches Hometown ?

What is the probability that a Netstar viewer is male?

Shows	Male	Female	Total
Hometown	100,000	50,000	150,000
Sita	50,000	200,000	250,000
Cricket	40,000	40,000	80,000
Total	190,000	290,000	480,000

0.31 (150,000 / 480,000) is the probability that a Netstar viewer watches Hometown

0.39 (190,000 / 480,000) is the probability that a Netstar viewer is male

Marginal probability is the probability of an event for a random variable irrespective of the outcome of other random variables. In example 1 we only care about if the viewer will watch Hometown or not irrespective of male or female. In example 2 we only care if the viewer is male or female irrespective of what show they watch.

QUIZ 1

1. Tickets with numbers between 1 - 20 (both inclusive) are mixed together. What is the probability that a ticket selected at random is a multiple of 5?
   a. 3/20
   b. 5/20
   c. 1/5
   d. 2/5
2. Find below confirmed Covid cases in a city (Age group & gender distribution):

Age Group	Male	Female	Total
45+	120,000	78,000	198,000
25-45	60,000	65,000	130,000
15-25	25,000	30,000	75,000
Below 15	12,000	22,000	34,000
Total	217,000	195,000	**412,000**

   a. A patient visited the hospital and tested positive. What is the probability that the person is Male?
   b. What is the probability that the person's age is 45+?
   c. Sheetal visited the hospital and tested positive. What is the probability that Sheetal's age is in the range of 15-25?
   d. What is the probability that a person who tested positive is below 15?

PYTHON CODE RECIPE

```
In [23]: import pandas as pd
```

```
In [44]: import numpy as np
```

```
In [24]: from sklearn.datasets import load_breast_cancer
```

```
In [45]: cancer = load_breast_cancer()
```

```
In [56]: df = pd.DataFrame(np.c_[cancer['data'], cancer['target']],
                           columns= np.append(cancer['feature_names'], ['target'
         ]))
```

```
In [59]: df = df[['mean radius','mean texture','target']]
```

```
In [60]: df['radius'] = (df['mean radius']>df['mean radius'].mean())*1
```

```
In [61]: df['texture'] = (df['mean texture']>df['mean texture'].mean())*1
```

```
In [64]: df = df.iloc[:,[2,3,4]]
```

```
In [126]: df['target'].sum()
```

```
Out[126]: 357.0
```

```
In [66]: df.head()
```

Out[66]:

	target	radius	texture
0	0.0	1	0
1	0.0	1	0
2	0.0	1	1
3	0.0	0	1
4	0.0	1	0

In above data, target=1 means benign cancer & target=0 means malignant cancer

radius=1 means radius > mean value of radius for all data

texture= mrean texture > mean value of all texture

only radius & texture has been taken here for purpose of study of 'probablity'

```
In [ ]:
```

```
In [96]: df2 = df.groupby(['radius','texture'],as_index=False).agg('sum')
```

Below is the distribution of radius, texture & number of benign cancer.

```
In [100]: df2
```

Out[100]:

	radius	texture	target
0	0	0	222.0
1	0	1	89.0
2	1	0	32.0
3	1	1	14.0

what is the probablity of being benign cancer when radius < mean radius

```
In [110]: a = df2[df2['radius']==0]['target'].sum()
```

```
In [111]: b = df['target'].sum()
```

```
In [112]: a/b
```

```
Out[112]: 0.8711484593837535
```

What is the probablity of being benign cancer when radius < mean radius & texture > mean texture

```
In [120]: a = df2[(df2['radius']==0) & (df2['texture']==1)]['target'].sum()
```

```
In [121]: b = df['target'].sum()
```

```
In [122]: a/b
```

```
Out[122]: 0.24929971988795518
```

```
In [ ]:
```

PROBABILITY

What is probability ?

Probability is the likelihood that an event will occur.

Types of probability ?

Joint Probability

Is the probability of the events of two simultaneously occurring random variables.

$=P(A) \times P(B)$

Conditional Probability

Is the probability that the event will occur given the knowledge that another event has already occurred.

$P(A/B) = \frac{P(A \cap B)}{P(B)}$

Marginal Probability

Is the probability of an event irrespective of the outcome of another event.

$P(A/B) = \frac{P(A \text{ and } B)}{P(B)}$

Data Science Application

- Predict what profile of customers prefer what kind of products.
- Predict congestion in a supply chain hub based on probability of incoming and outgoing traffic.

MODULE 2

PROBABILITY DISTRIBUTIONS

OBJECTIVE

By the end of this module you will be able to:

1. Understand what a distribution is
2. Decide between the various types of distributions
3. Apply distributions to build data science models

RANDOM VARIABLES AND DISTRIBUTIONS

A random variable can take one of many values that indicate the probability of an event happening. For example, P(tails) in a single coin toss is 0.5.

A probability distribution provides a list of all values that the random variable can take along with the probability of each value occuring.

Discrete and Continuous variables

A discrete random variable can only take a very specific value out of a predefined set of values. For example a throw of a dice can only have 1 of 6 possible values, a coin toss can only have 1 of 2 possible values.

A continuous random variable, can take any value within a certain range, for example a mileage of a car, weight of a person etc.

Applications in Data Science

Understanding your datasets distribution has many decision implications while you build your machine learning model - what kind of hypothesis test you will choose, what kind of machine learning algorithms you will choose, what kind of feature engineering you will have to do etc. Some machine learning models fit well for data with certain distributions, hence it helps to know what distribution your dataset has.

TYPES OF DISTRIBUTION

The most commonly used discrete probability distributions (these summarise the probabilities of a discrete random variable) are:

1. Bernoulli distribution
2. Uniform distribution
3. Binomial distribution
4. Poisson distribution

The most commonly used continuous probability distributions (these summarise the probabilities of a continuous random variable) are:

1. Normal distribution
2. Exponential distribution

1. Bernoulli Distribution

This is the simplest type of distribution. Data that fits this distribution has only 2 outcomes and 1 trial.

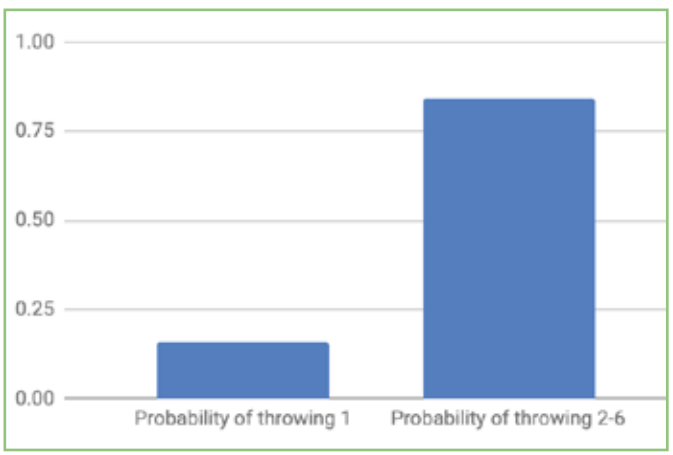
For example, a single coin toss has 2 outcomes (heads or tails), the probability of 1 outcome is p (0.5 or 50%) and the probability of the other outcome is (1-p = 1-0.5 = 0.5 or 50%). This also indicates that the sum of these 2 probabilities = 1 (p+1-p). Intuitively this makes sense, there is a 100% chance (a probability of 1) that a single coin toss will result in a heads or tails.

All the other distributions like binomial, negative binomial etc are variations of Bernoulli distribution.

Here is another example to understand Bernoulli Distribution. Suppose you enter a casino and sit at a game table. You bet $100 that when you throw the dice the number will be 1. What is the probability of your success?

Success for this trial means you should get 1 as the number on your dice throw. There are 6 possible outcomes possible.

P(Success)= Probability of getting number 1 = 1/6
P(Failure)=1-Probability of success= 1- 1/6= 5/6



2. Binomial Distribution

This is a variation of the bernoulli distribution concept. In binomial distribution, only 2 outcomes are possible (like bernoulli) but multiple trials are possible (bernoulli has only 1 trial). The probability of an outcome in each trial is independent of the previous trial and the outcomes do not have to be equally likely.

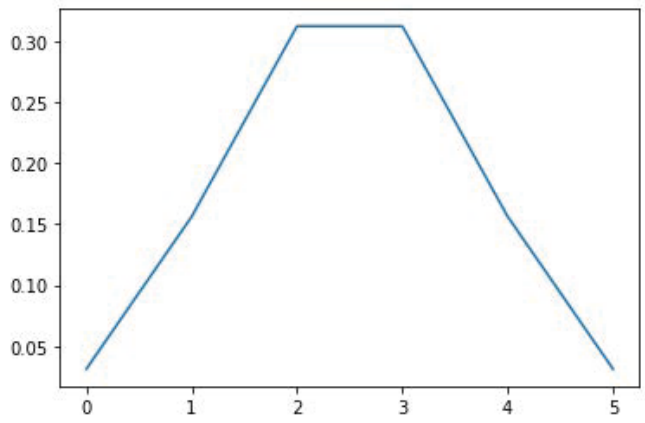
An example of this would be tossing a coin 100 times and finding the probability that a head will show up 5 times consecutively.

Binomial distribution is represented by the following formula:

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

The number of times an experiment is repeated is denoted by n, p represents the probability of a specific outcome(the success), x denotes the number of successes desired and q = 1-p (the probability of a failure)

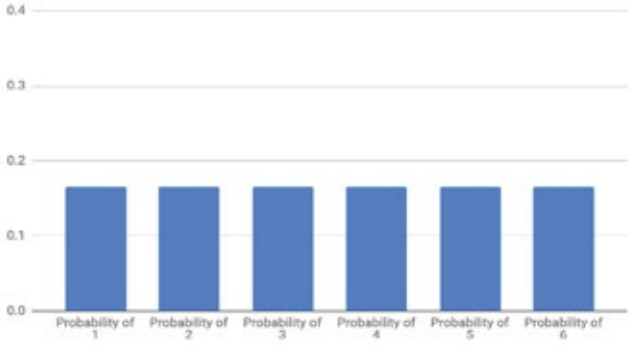
Take an example, where you toss a coin 5 times (n). When you toss a coin, the P(head)=P(tail)=0.5. You need to find the probability of the number of times heads can occur (success, p). Plugging this in the above formula you get the following graph



The x axis has the number of heads that can occur and y axis has the probabilities of success. This graph shows that when a coin is flipped 5 times, the probability of heads occurring 2 out of 5 times and 3 out of 5 times is highest (0.35).

3. Uniform Distribution

Take the same dice example as above. There is an equal probability (1/6) that any of the 6 faces of the dice can show up when you throw. This is called uniform distribution. It is called a rectangular distribution since the outline of the distribution graph below is a rectangle.



Other instances of an uniform distribution are:

a. The probability of taking a card out of a deck (every card has the same probability)

b. The probability of an individual winning a lucky draw contest (every contestant has the same probability to win)

The probability distribution function for intervals [a,b] can be given by:

$$P(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

Which in short means that if you calculate the probability of an event occuring outside intervals a and b is 0. a and b in the above example is 1 and 6.

Consider another example

Assume that a pizza shop owner wants to give a free pizza to a random person who walks past the shop. The lowest number of people who walked past in 1 hour is 4 and the maximum is 8. Find the probability of an individual getting a pizza.

Here a =4 and b=8

Plugging these values in the formula: P(X)= 1/ (b-a) = 1/(8-4) = 0.25

4. Poisson distribution

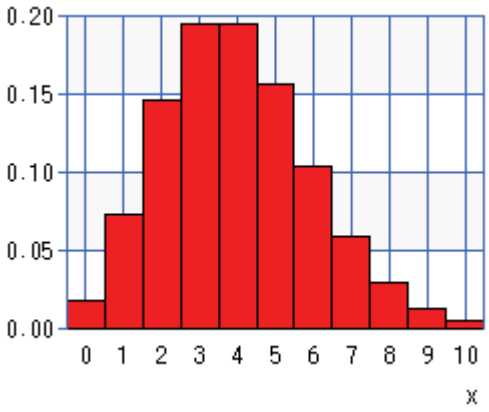
You would typically use Poisson distribution to find out the number of events occurring in a certain time interval. The events can occur at random at any time inside the interval, can occur any number of times and are not dependent on each other. Common scenarios that can be modeled by a Poisson distribution are:

1. Number of tickets handled by a customer support reps in an hour
2. Number of customers coming into a retail shop per day

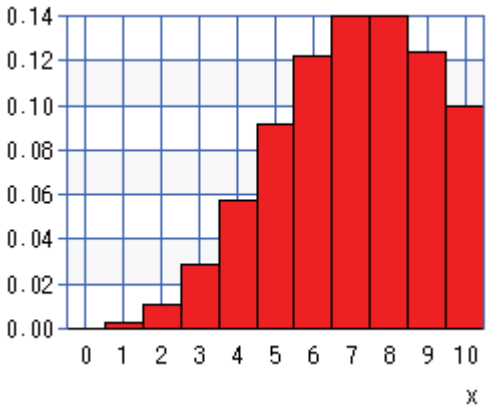
3. Number of games a team can play in a week

A poisson distribution looks like the following for various values of lambda (the rate at which the event occurs).

Lambda = 4



Lambda = 8



Poisson distribution is defined by the formula:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$

Where X is a discrete random variable representing number of events in a certain time period, e is Euler's number, lambda represents the average number of events taking place in a time period.

5. Normal Distribution

This is the most common type of distribution and also called a Gaussian distribution. Most naturally occurring events follow this distribution - for example IQ scores of people, salary of employees in a company, cost of homes etc.

The intuitive way to understand normal distribution is that most values of the above events fall in the middle, there are some very low values and some very high values. Thus normal distribution is symmetric around the mean.

A normal distribution looks like a bell curve like the figure below. The graph shows the mean delivery time for food orders in a city. The mean here is 30 and variance is 2. You will notice that this graph is symmetric in nature - meaning there are the same number of values less than and greater than 0.

This graph shows that half the orders are delivered in less than 30 minutes and half above 30 minutes. About 2.5% of orders take less than 25 minutes to deliver and about 2.5% of orders require more than 33 minutes to deliver. The probability of all the events under the bell curve = 1.

In this graph, the mean tells you the height of the graph and the standard deviation tells you the width of the graph. Smaller the standard deviation, the less spread out the values are and hence a narrower bell curve.

6. Exponential Distribution

Exponential distribution is used to predict the time till when the next event will occur, for example when will the next volcano erupt.

This technique is used commonly in scenarios known as "survival analysis" that predict the expected life of a machine, human etc. Other uses of exponential distribution include:

a. Calculating waiting time and holding time in call centres

b. Repair times (when will next repair event occur)

The equation for exponential distribution is:

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

Where lambda is the rate parameter. Lambda value denotes the rate of decay in the exponential distribution. Higher the value, the faster the decay and the steeper the curve falls down.

With lambda = 0.3

Other distributions include

Negative binomial distribution - This distribution models the number of failures in a series of events.

Standard normal distribution - This is a special case of normal distribution where the mean is 0 and standard deviation is 1.

T-distribution. - The T distribution is similar to a normal distribution, but with having heavier tails. Due to heavier tails the values are far away from mean.

Weibull distribution - This is used to model product lifetimes and time to failure.

QUIZ 2

1. What distribution would the following data fall in - The number of cakes a chef can make in a day

   a. Normal distribution

   b. Poisson distribution

   c. Binomial distribution

   d. Exponential distribution

2. What distribution would the following data fall in - To predict how many miles will an aircraft engine work for

   a) Normal distribution

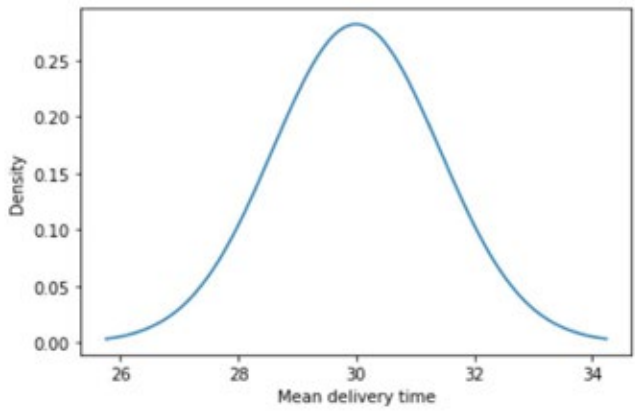
   b) Poisson distribution

   c) Binomial distribution

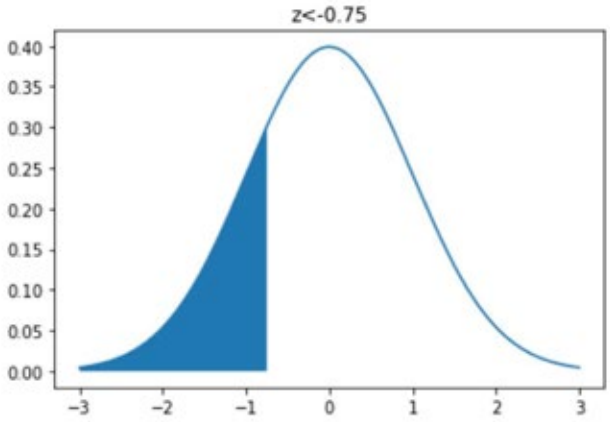
   d) Exponential distribution

PYTHON CODE RECIPE

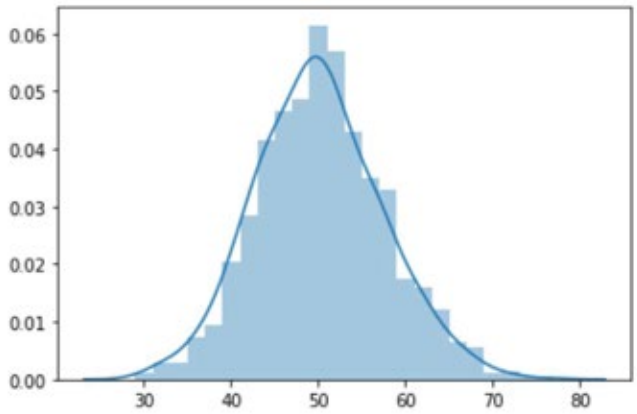
In [1]:
```
#### Distributions Module Code #############

import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
import math



#Normal Distribution
mu = 30
variance = 2
sigma = math.sqrt(variance)
x = np.linspace(mu - 3*sigma, mu + 3*sigma, 100)
plt.plot(x, stats.norm.pdf(x, mu, sigma))
plt.ylabel('Density')
plt.xlabel('Mean delivery time')
plt.show()
```



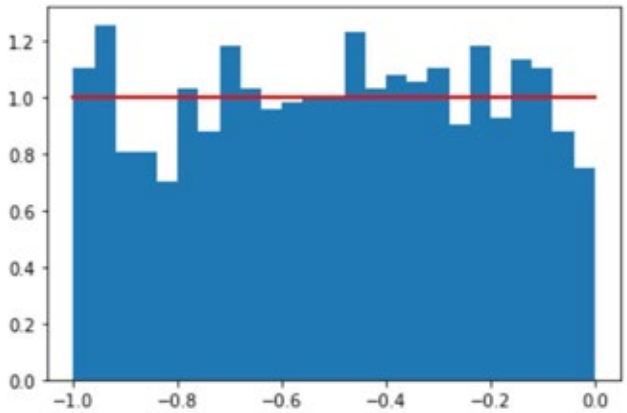
In [2]:
```
#Standard normal dist
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
def draw_z_score(x, cond, mu, sigma, title):
    y = norm.pdf(x, mu, sigma)
    z = x[cond]
    plt.plot(x, y)
    plt.fill_between(z, 0, norm.pdf(z, mu, sigma))
    plt.title(title)
    plt.show()

x = np.arange(-3,3,0.001)
z0 = -0.75
draw_z_score(x, x<z0, 0, 1, 'z<-0.75')
```



In [3]:
```
#Poisson distribution

from numpy import random
import matplotlib.pyplot as plt
import seaborn as sns

sns.distplot(random.poisson(lam=50, size=1000), hist=True, label='poisso
n')

plt.show()
```

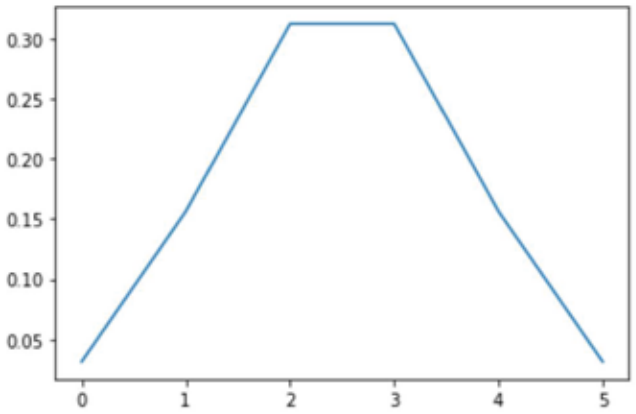


In [4]:
```
#Uniform Distribution

import matplotlib.pyplot as plt
import numpy as np
s = np.random.uniform(-1,0,1000)
count, bins, ignored = plt.hist(s, 25, density=True)
plt.plot(bins, np.ones_like(bins), linewidth=2, color='r')
plt.show()
```



In [5]:
```
#Binomial Distribution
import scipy, scipy.stats
x = scipy.linspace(0,5,6)
pmf = scipy.stats.binom.pmf(x,5,0.5)
import pylab
pylab.plot(x,pmf)
```

Out[5]:
```
[<matplotlib.lines.Line2D at 0x29068bfce48>]
```

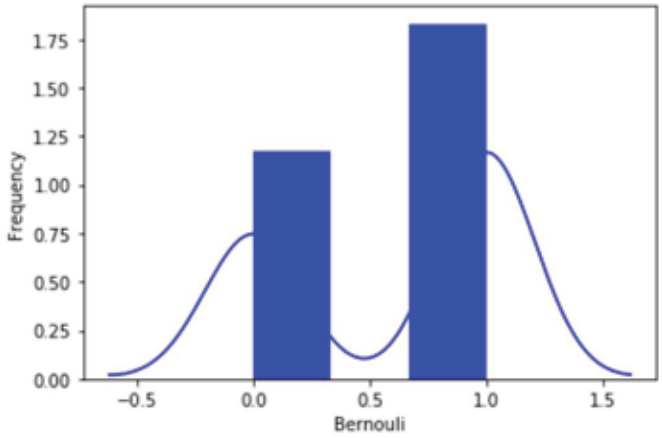


In [6]:
```
#Negative Binomial

s = np.random.negative_binomial(1, 0.1, 100000)

for i in range(1, 11):

   probability = sum(s<i) / 100000.

   print(i, "wells drilled, probability of one success =", probability)
```

```
1 wells drilled, probability of one success = 0.10101
2 wells drilled, probability of one success = 0.19048
3 wells drilled, probability of one success = 0.27355
4 wells drilled, probability of one success = 0.34642
5 wells drilled, probability of one success = 0.41088
6 wells drilled, probability of one success = 0.46954
7 wells drilled, probability of one success = 0.52261
8 wells drilled, probability of one success = 0.57007
9 wells drilled, probability of one success = 0.61342
10 wells drilled, probability of one success = 0.65215
```

In [7]:
```
#Bernoulli

from scipy.stats import bernoulli
import seaborn as sb

data_bern = bernoulli.rvs(size=100,p=0.6)
ax = sb.distplot(data_bern,
                 kde=True,
                 color='blue',
                 hist_kws={"linewidth": 25,'alpha':1})
ax.set(xlabel='Bernouli', ylabel='Frequency')
```

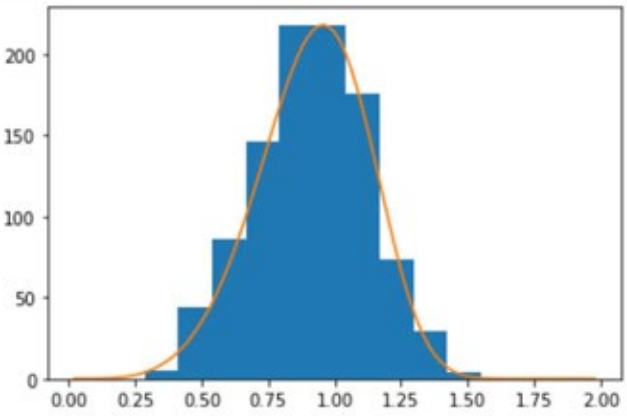
Out[7]:
```
[Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Bernouli')]
```



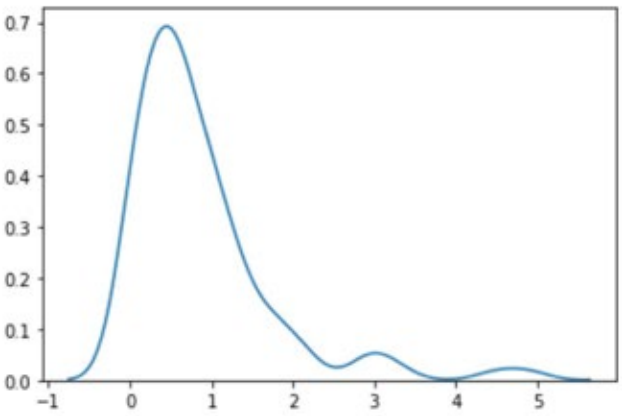
In [8]:
```
# T distribution

from scipy.stats import t
import numpy as np
quantile = np.arange (0.01, 1, 0.1)

numargs = t .numargs
a, b = 4.32, 3.18
rv = t (a, b)
# Random Variates
R = t.rvs(a, b)
print ("Random Variates : \n", R)

# PDF
R = t.pdf(a, b, quantile)
print ("\nProbability Distribution : \n", R)
```

```
Random Variates :
 1.432680001993857

Probability Distribution :
 [0.00663446 0.00721217 0.0078511  0.00855881 0.00934388 0.01021611
 0.01118667 0.01226833 0.01347568 0.01482539]
```

In [9]:
```
#Weibull distribution

a = 5. # shape
s = np.random.weibull(a, 1000)

import matplotlib.pyplot as plt
x = np.arange(1,100.)/50.
def weib(x,n,a):
    return (a / n) * (x / n)**(a - 1) * np.exp(-(x / n)**a)
count, bins, ignored = plt.hist(np.random.weibull(5.,1000))
x = np.arange(1,100.)/50.
scale = count.max()/weib(x, 1., 5.).max()
plt.plot(x, weib(x, 1., 5.)*scale)
plt.show()
```



In [10]:
```
#Exponential Distribution

from numpy import random
import matplotlib.pyplot as plt
import seaborn as sns

sns.distplot(random.exponential(size=100), hist=False)

plt.show()
```



Probability Distributions

Probability distribution provides a list of all values that a random variable can take along with the probability of each value occuring.

Discrete Variables-

A discrete random variable can only take a very specific value out of a predefined set of values.

Continuous Variables-

A continuous random variable, can take any value within a certain range,

Types of Discrete Probability Distributions

1. Bernoulli Distribution
   Used when the variable has only 2 outcomes and 1 trial.
2. Uniform Distribution
   Used when there is equal probability for the variable to take one of the values.
3. Binomial Distribution
   Used when 2 outcomes are possible (like bernoulli) but multiple trials are possible.
4. Poisson Distribution
   Used to find out the number of events occurring in a certain time interval.

Types of Continuous Probability Distributions

1. Normal Distribution
   Used when data is symmetric around the mean.
2. Exponential Distribution
   Used to predict the time till when the next event will occur.

Applications in Data Science

Understanding the datasets distribution helps is choosing the right machine learning model.

MODULE 3

UNIVARIATE ANALYSIS

OBJECTIVES

By the end of this module you will be able to:

1. Understand what univariate analysis is
2. How to analyze data
3. Graphing single variable

UNIVARIATE ANALYSIS

Univariate Analysis explores each variable in the data set, separately.

The analysis of a single variable is the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships.

The main purpose of the analysis is to find patterns that exist within the data and to find outliers if present.

EXAMPLE

Suppose that the weights of 12 students in a class is recorded.

The weight of 12 students are as follows

44, 55,33, 66, 66, 33, 43, 56, 77, 101, 55, 66

There is only one variable that is weight and it is not dealing with any cause or relationship.

We can do a lot of analysis on it like finding mean, median, standard deviation, plotting histograms, finding outliers etc. This kind of analysis using only one variable is called univariate analysis.

Common example of univariate analysis is the mean of a population distribution. Tables, charts, polygons, and histograms are all popular methods for displaying univariate analysis of a specific variable (e.g. mean, median, mode, standard variation, range, etc.).

HOW TO ANALYZE ONE VARIABLE

1. Raw data

Raw data is the data in its purest form. This may require some cleansing but for this example we will assume that we have cleansed data at hand.

For e.g. Raw data for study of number of babies born in a state (per town/hospital)

HOSPITAL NUMBER	TOWN NAME	NUMBER OF GIRLS BORN	NUMBER OF BOYS BORN
1	A	4	1
2	B	4	2
3	C	6	1
4	C	3	1
5	B	5	3
6	A	2	3
7	A	3	3
8	A	2	3
9	A	9	2
10	B	7	2

It is difficult to tell what is going on with each variable in this data set. Similarly, if we had a million records like above it would have been difficult to analyze this dataset

Univariate descriptive statistics can summarize large quantities of numerical data and reveal patterns in the raw data. In order to present the information in a more organized format, start with univariate descriptive statistics for each variable.

For example, the variable "Number of girls born":

NUMBER OF GIRLS BORN									
4	4	6	3	5	2	3	2	9	7

2. Frequency distribution

Another part of univariate analysis is frequency distribution. We do this by identifying the lowest and highest value for the variable and then sorting it in that order.

Next, we count the occurrence of each value in that variable/column.

This is a count of the frequency with which each value occurs in the data set.

For example, for the variable "Number of girls born," the values range from 2 to 9.

NUMBER OF GIRLS BORN	NUMBER OF TIMES EACH VALUE OCCURS
2	2
3	2
4	2
5	1
6	1
7	1
9	1
Total	10

3. Grouped data

While doing any kind of univariate analysis, we need to decide whether the data should be grouped into classes.

The "Number of girls born" can be easily grouped for different values.

One way to construct groups is to have equal class intervals (e.g., 1-3, 4-6, 7-9).

Another way to construct groups is to have about equal numbers of observations in each group. Remember that class intervals must be both mutually exclusive and exhaustive.

NUMBER OF GIRLS BORN	OCCURRENCES IN THAT GROUP
Low (1-3)	4
Medium (3-6)	4
High (6-9)	2
Total	10

Similarly, there are also various ways in which we can do additional univariate analysis. Some of these include cumulative percentages, cumulative distribution etc.

Please note that it is not mandatory to do all of these analyses on the dataset. One can choose to drop some graphs/ charts as needed.

Graphing the Single Variable:

GRAPHS (LINE/BAR/PIE)

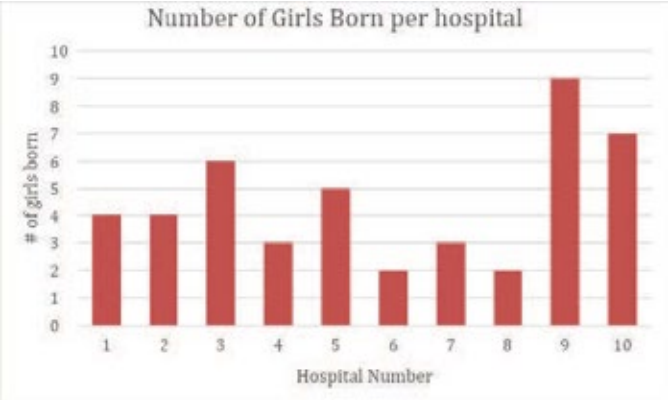
Graphs are used to display the frequency distributions/ percentages/share of each entity/change with respect to time etc.

These are pictorial representations which help us in understanding the dataset better by looking at the figure/diagram instead of the raw numbers/data.

Graphs can be of various kinds. A few of them are:

1. Bar Chart

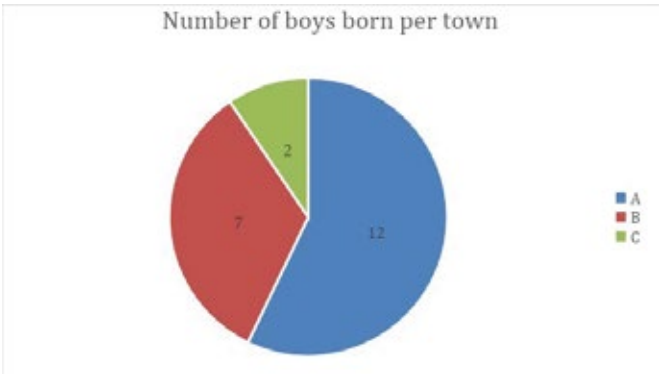
We can plot a bar chart of the number of girls born per hospital here.



The above bar chart shows us in a quick manner which hospitals had the highest and lowest number of girls born across towns.

2. Pie Chart

Pie chart shows the contribution of each value to the total mix. In short, it means if we try to see #of boys born across town and how much each town % is we can use pie chart

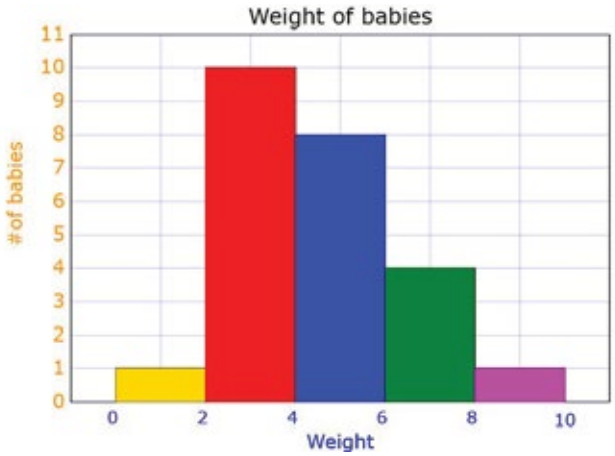


Similarly, we can plot pie charts of the number of boys born across towns, hospitals etc using clustered bar charts, line charts, scatter plot etc.

3. Histogram

A histogram is similar to a bar chart, but instead of representing individual values it represents intervals and the values contained in these intervals. For example, find below is raw data of babies and their weights

Baby Name	Baby weight
A	5
B	3
C	7
D	8
E	2
F	6.
G	7
H	7
I	3
J	2
K	2
L	4
M	1
N	4
O	2
P	4
Q	4
R	4
S	4
T	2
U	2
V	2
W	2
X	4



This histogram reads that there is 1 baby in weight category 0-2, 10 babies in category 2-4kg etc.

QUIZ 3

1. Univariate analysis means:
   a. Analyzing 2 variables at a time
   b. Analyzing time and given data variable
   c. Analyzing 1 variable at a time
   d. Analyzing more than 2 variables at a time
2. Calculating standard deviation of any variable is not considered to be univariate analysis.
   a. True
   b. False
3. What is standard deviation of 4, 2, 3, 2, 9, 11, 14, 1, 9.
   a. 4.4
   b. 4.0
   c. 3.5
   d. 3.1

PYTHON CODE RECIPE

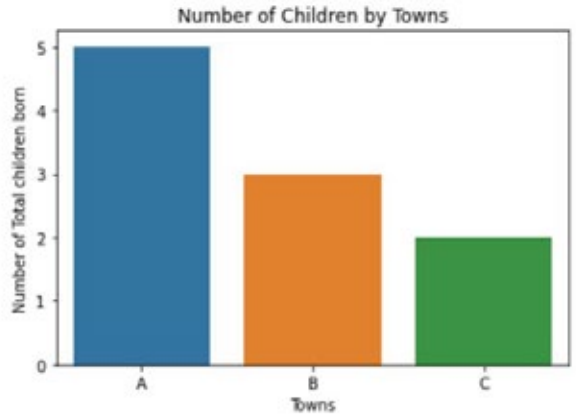
Consider raw data for the study of the number of babies born in a state (per town/hospital). Import the data as a .csv file

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
t=pd.read_csv("univariate_data_table.csv")
```

Let's plot a bar graph showing how many children are born in each town (here you are doing an univariate analysis on "number of children born in each town)

```
sns.countplot(x='Town name', data=t)
plt.title('Number of Children by Towns')
plt.ylabel('Number of Total children born')
plt.xlabel('Towns')
plt.show();
```

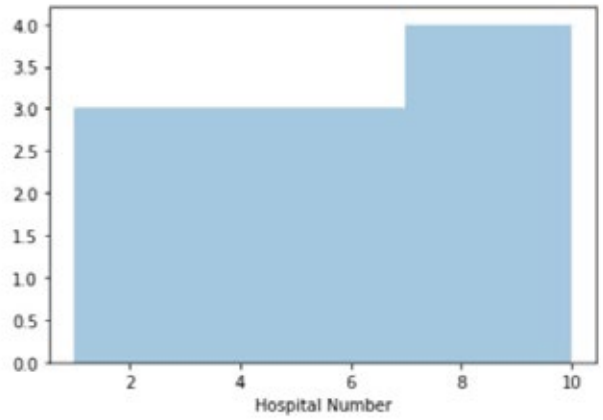


This shows the highest number of children are born in Town A from a given sample.

Let's plot a histogram with default bins calculated as per given data

```
sns.distplot(t['Hospital Number'],kde=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20927fca2e0>
```



This indicates that more children are born in hospitals number 8-10.

Univariate Analysis

Definition

Explores each variable in the data set separately.

Example-

Weight of 12 students

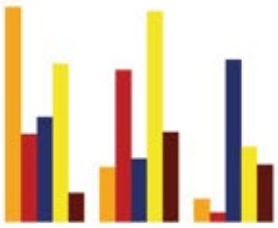
Univariate Analysis -

Mean, median, standard deviation.

Frequency distribution -

Count of the frequency with which each value occurs in the data set.

Ways to represent Univariate Analysis



Bar Graph

Line Chart

Histogram

Data Science Applications -

1. Used in Exploratory Data Analysis (EDA) stage of model building
2. Find outliers
3. Get a sense of the data values

MODULE 4

BIVARIATE ANALYSIS

OBJECTIVE

By the end of this module you will be able to:

1. Understand what bivariate analysis is
2. Difference between two sample data and bivariate data analysis
3. How to analyze bivariate data
4. Positive and negative correlation

BIVARIATE ANALYSIS

Bivariate analysis brings you closer to real Data Science and Machine Learning problem statements. Machine learning problems are all about trying to predict the future using current variables. In order to do this, you need to find out which variables affect (correlate) each other and by how much. Once you know this, you can model out the future for how these variables will behave together.

Hence you will use bivariate analysis to understand the relationship between 2 variables.

This is usually done in the initial EDA (exploratory data analysis) stage of your machine learning model building. Using a scatter plot (X axis and Y axis on a cartesian plane) is a common way to do bivariate analysis during EDA - you will represent the 2 variables X and Y on a scatter plot and their intersection will give you an idea of their correlation. This is useful to find out values of dependent variables based on change in independent variables.

Applications of Bivariate Analysis in Data Science and Machine Learning:

When you work on building your machine learning models you will use bivariate analysis during the EDA phase (Exploratory Data Analysis). This is the phase when you are trying to get a sense of your dataset. EDA is critical as it helps you in understanding the intricacies of your dataset.

In EDA you will first do some univariate analysis to understand the distribution of the data. This will be followed by bivariate analysis to understand the correlations. As part of your EDA you will develop some hypotheses for your target variable (the variable you are predicting or measuring). Bivariate analysis can confirm if such correlations exist for your target variable and other variables in your dataset.

Based on the above you will create or delete features (variables) from your dataset. This step is called feature engineering.

Bivariate analysis and correlation coefficients are also very commonly used in finance and investing. For example, to find how the stock price of an oil company moves based on the price of crude oil. Or how sales numbers affect the stock price. In mutual funds, portfolio managers think about positive and negative correlations before adding a stock to the mutual fund portfolio to make sure there is sufficient diversification.

Example:

The simplest example of bivariate analysis is the correlation between age and weight. Age (X) and weight (Y) are measured for babies in a sample. In this example, age and weight are related to each other and changes in age show changes in weight.

X variable (age in months)	Y variable (weight in kg)
3	4.6
0.5	2.7
2	3.5
4	6.1
6	7.2

Other examples of bivariate data include:

- Ice cream sales and the season
- Number of accidents and the weather
- Number of hours of TV watched and day of the week
- Income and loan approval rate

Difference between bivariate data and two sample data:

- In two sample data, the values of two variables need not be related to each other

Two-sample data:

- Sample 1 (Height): 153,162,155,166,163
- Sample 2 (Income): Rs.1,00,000, Rs.90,000, Rs.3,00,000, Rs.1,50,000, Rs.2,00,000

In bivariate data, values of two variables are related to each other. They are paired together. For example: (x,y) = (height,weight) where there might be a relation

between height and weight of an individual.

(X,Y) =(158,58),(162,84),(159,63),(166,70)

HOW TO IMPLEMENT BIVARIATE ANALYSIS

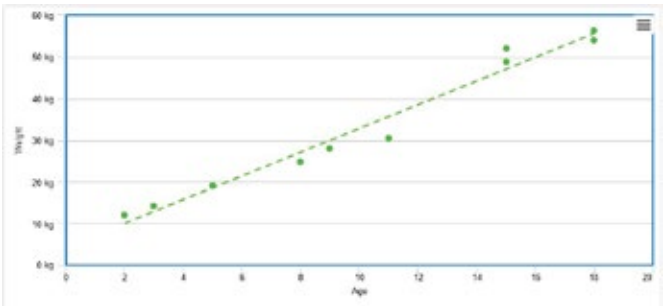
Bivariate analysis is deployed using a combination of statistical techniques including scatter plot, line of best fit, correlation coefficients and regression analysis. All of these are explained below within the context of positive and negative correlations.

1. Positive correlation

The following bivariate data table shows the relationship between age and weight from initial years to teenage. This is an example of positive correlation, because the value of the dependent variable (weight) increases with increasing value of the independent variable (age).

Kid	Age (X)	Weight (Y)
1	2	12.0
2	3	14.2
3	5	19.1
4	8	24.8
5	9	28.0
6	11	30.5
7	15	48.8
8	15	52.0
9	18	56.3
10	18	54.0

If we plot a scatter plot for this data, we see the positive correlation between age and weight of kids. The older the age the greater the weight.



2. Line of best fit

The green dotted line that cuts through the plot is called the "line of best fit" - it simply indicates the relationship between the points. In this case, the line of best fit indicates a very strong correlation, since all the points are densely populated close to the line.

3. Correlation Coefficient

The strength of the relationship between the 2 variables (how strongly are they correlated) is expressed by using a "Correlation Coefficient", denoted by 'r' and having values between -1 to +1. This is calculated using a Pearson Correlation Coefficient Calculator. A 0 value for 'r' means there is no correlation between the 2 variables, +1 means a totally positive correlation and -1 means a totally negative correlation.

4. Regression

In the above example, dependent variable weight (Y axis) is dependent on independent variable age (X axis). This is called a regression of y on x and can be represented by a regression equation.

A simple linear regression equation is: $y = bx + a$

It calculates how much the value of y will change based on the value of x. The best fit line in the graph above is also called a regression line.

5. Negative correlation

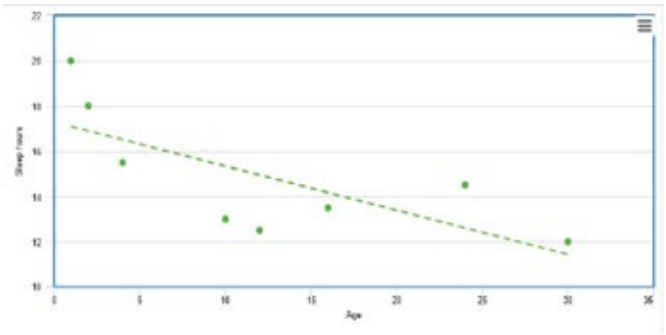
The following bivariate data table shows the relationship between a baby's age and number of hours of sleeping.

Age	Total sleep hours
0-2 months	16-18
2-4 months	14-16
4-6 months	14-15
6-9 months	14

This is an example of negative correlation where value of dependent variable (sleep hours) decreases with increasing value of independent variable (age).

Let's take an sample of 8 kids in age 0-3 years to study relation between age and sleep

Kid	Age (months)	Sleep hours
1	4	15.5
2	10	13
3	24	14.5
4	30	12
5	1	20
6	2	18
7	16	13.5
8	12	12.5



If we plot a scatter plots for this, we see the negative correlation between age and sleep hours of kids. The older the age the lesser the sleep hours.

QUIZ 4

1. Perform bivariate analysis to find out the link between age and basal metabolism (calories/m^2/ hour) for the given sample

Resident	Age	BMR
1	20	39.2
2	45	36.2
3	30	37.2
4	69	35.8
5	17	40.1
6	32	37.3
7	55	35.0
8	25	37.9
9	10	44.0

2. Find the correlation between the following X and y value pairs

   X: 12,33,55,22,11,1

   Y: 22,89,33,22,12,3

   Draw a scatter plot showing how X affects Y.

3. Find the correlation coefficient between the X and Y vectors in question 2 using pearson's coefficient.

PYTHON CODE RECIPE

Export quiz question data in csv.

```
import pandas as pd
```

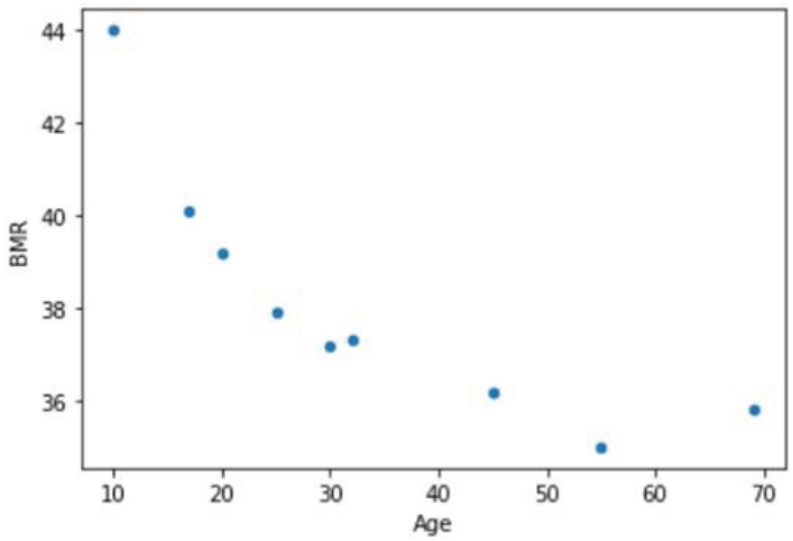
```
t=pd.read_csv("bivariate_data_table.csv")
```

```
t.head()
```

Resident	Age	BMR
1	20	39.2
2	45	36.2
3	30	37.2
4	69	35.8
5	17	40.1

Plot scatter plots

```
ax=t.plot.scatter(x='Age', y='BMR')
```



If we draw the line of best fit we can say there is negative correlation between age and BMR. As age increases, BMR tends to decrease.

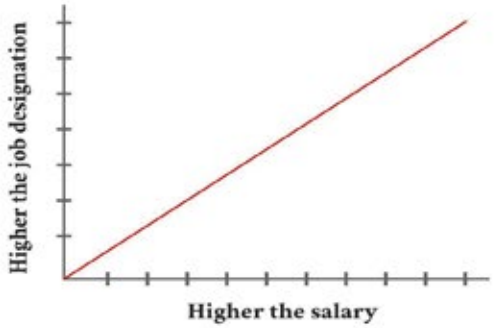
Bivariate Analysis

Definition

Analyse the relationship between 2 variables.

Example-

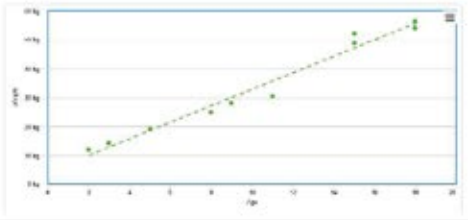
Relationship between salary and job designation.



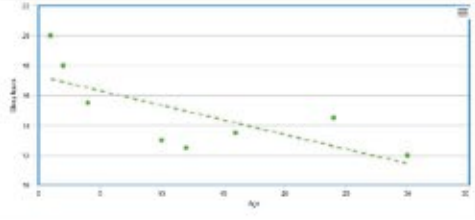
Pearson Correlation Coefficient

Denotes the strength of the relationship between 2 variables (how strongly are they correlated).
It has values between -1 to +1.

Positive Correlation



Negative Correlation



Data Science Applications -

1. Used in Exploratory Data Analysis (EDA) stage of model building
2. Understanding correlations between variables
3. Feature Engineering to delete and create new variables

MODULE 5

MULTIVARIATE ANALYSIS

OBJECTIVES

By the end of this module you will be able to:

1. Understand what multivariate analysis is
2. Understanding independent, dependent, continuous & categorical variables
3. How to do Multivariate analysis
4. Impact & usage of multivariate analysis in Data Science

MULTIVARIATE ANALYSIS

Almost every real world data science problem is a multivariate problem. This means that the problem and the solution are the result of multiple factors interacting together, compared to univariate (only 1 factor) and bivariate (2 factors). Multivariate analysis gives you a set of statistical tools to find the connections between multiple variables simultaneously.

While performing multivariate analysis, you are not only analyzing various variables, but also trying to gauge how much each variable contributes to another variable.

APPLICATIONS IN DATA SCIENCE

Multivariate analysis is one of the most important aspects while performing analysis on data.

While building your model, Bivariate, Multivariate and Univariate analysis fall under your EDA (Exploratory Data Analysis). EDA is a critical step as it helps you in understanding the intricacies of your dataset. Complex relationships between multiple variables in your dataset can be better understood with multivariate analysis.

Based on your understanding of the type of data in the EDA stage, you will decide on a few models to build and test. Examples of real world multivariate situations are:

- A company's stock price being influenced by revenue, profits, team, economy, interest rates etc
- Outcome of a medical vaccine trial being influenced by trial patients age, previous illness, variety of dosage etc
- Weather prediction influence by humidity, windspeed, forest fires etc
- A company's marketing outcomes influenced by marketing channel, budget, collateral etc
- The NASA Mars rover collects soil samples from Mars. Each sample has about 6,000 variables and hence presents a multivariate analysis problem.

Example

Age (X) and weight (Y) are measured for babies in a sample. Along with this we also measure height of the baby according to age. We then analyze all 3 variables together. In this example, age, height and weight are related to each other and changes in age and height can show changes in weight.

X_1 variable (age in months)	X_2 variable (height in centimeters)	Y variable (weight in kg)
3	65	4.6
0.5	40	2.7
2	47	3.5
4	70	6.1
6	74	7.2

Thus if we add the effect of height on weight keeping age constant and add the effect of age on weight keeping height constant, we are doing multivariate analysis.

Multivariate Analysis Techniques

All multivariate techniques are categorized based on the kind of analysis you want done with the variables.

1. Dependence - This is when you are trying to figure out if and how the variables are related and affect each other.
2. Interdependence - This is when you are trying to figure out an underlying pattern to the data.

Within these 2 types of techniques, you will pick the specific technique based on the type of variables your dataset contains. There are 2 broad types of variables that will affect your decision -

1. Continuous variables - Otherwise known as metric variables, are nothing but variables that are numeric and represented by a number. For example, age, salary, sales, humidity etc.

Categorical variables

These are otherwise known as discrete variables, they categorise the data into buckets. For example, Nationality - Indian, Movie genre - Thriller. There are sub-types of categorical variables -

a. Nominal - Nominal variables have 2 or more categories or classes but there is no order amongst them. For example, the variable "Types of homes" can have values / classes such as 2 bhk, 3bhk, 4bhk, Penthouse, Studio, Duplex, Independent house etc.

b. Dichotomous - Dichotomous variables are essentially nominal variables but with 2 categories/ classes only. For example, predicting if Sachin will score a century or not, the answer can only be "Yes" or "No", so only 2 classes.

c. Ordinal - Ordinal variables are just like nominal variables but they have a hierarchy or order to them. For example, the interest level of a sales lead can be 'not interested', 'slightly interested', 'very interested', 'closed'.

Most machine learning models that do multivariate analysis need a numerical (continuous) input. Hence as part of your Feature Engineering stage in your machine learning project, you will convert the categorical variables into continuous variables.

UNDERSTANDING INDEPENDENT AND DEPENDENT VARIABLES

Independent variables (experimental/predictor) are the ones which are manipulated so as to see an effect on another variable (dependent variable). A dependent variable is nothing but a variable which is dependent on multiple independent variables.

For example:

Student ID	Hours of Study	Marks Obtained
1	2	12.0
2	3	14.2
3	5	19.1
4	8	24.8
5	9	28.0

As we can see from this table, the independent variable here is "Hours of Study" and the dependent variable is "Marks" obtained. The more a student studies, the marks go up. Here is a cheat sheet on what kind of popular multivariate analysis to pick based on your dataset.

Category	Dependent Variables	Independent Variables	Multivariate Analysis Technique to use	Applications
Dependence	1 continuous	Any type	Multiple regression	Analyses the degree of impact each independent variable has on the dependent variable. For example predicting sales (dependent variable) based on independent variables such as store timings, number of sales executives etc.
Dependence	1 continuous or categorical	Categorical	Conjoint Analysis	For example analysing the impact on the price of a product (dependent) based on product attributes (independent) such as colour and features.
Dependence	1 categorical	Continuous	Multiple Discriminant Analysis	This is commonly used in classification - to identify the characteristics of the independent variable that make up the class of the categorical variable. For example - Categorising the characteristics (independent variables) of animal pictures to a category of animal 'dog', 'cat' (dependent variable).
Dependence	Many continuous	Categorical	MANOVA	
Interdependence		Continuous	Factor Analysis	Popular techniques here like PCA (Principal Component Analysis) are used in data preprocessing to combine multiple similar variables into a single variable.
Interdependence		Continuous	Cluster Analysis	Groups of values are identified. For example, clusters/groups of customers having the same complaints or buying patterns.

HOW TO PERFORM MULTIVARIATE ANALYSIS

There are many ways to perform Multivariate Analysis on a given dataset. We will discuss one of the most common ways that is used to perform multivariate analysis.

Let us consider the following dataset

Day	Temperature	Humidity	Cycles on Rent
1	37	High	12
2	12	Low	14
3	12	Low	19
4	17	High	7
5	18	High	29
6	20	High	44
7	38	Low	4
8	42	Low	15
9	40	Low	5
10	22	High	40

The above table denotes the data for 10 days in which a store measures temperature and categorizes humidity as high or low. Based on this data they want to check if these variables have an impact on the number of cycles which are given on rent

Step 1 - Identify the variables and types

Temperature & Humidity will have an effect on Number of cycles given on rent. So these two will become independent variables. Number of cycles on rent becomes the dependent variable.

Temperature takes numerical values and it is quantitative in nature. Hence this becomes a continuous variable. The humidity variable can only take 2 values, high and low so it becomes a qualitative variable, hence categorical variable.

Step 2 - Find correlations

Calculate if there are any correlations between Temperature and Number of cycles given on rent. This constitutes a part of Bivariate analysis. Then we can also calculate if humidity affects the number of cycles on rent, using other statistical tests like t test, chi squared etc.

Step 3 - Build a model

Once we establish there is some relationship between these variables we try to quantify it using a model.

Weight/coefficient * Temperature + Weight/coefficient * Humidity = Number of cycles on rent

QUIZ 5

Resident	Age	BMR	Weight
1	20	39.2	52
2	45	36.2	39
3	30	37.2	77
4	69	35.8	55
5	17	40.1	52
6	32	37.3	49
7	55	35.0	50
8	25	37.9	70
9	10	44.0	31

1. Perform multivariate analysis to find out the link between age and basal metabolism (calories/m^2/hour) for given sample along with age and weight. Also decide if weight can be used to predict BMR in this case
2. Find out if Age, Weight & BMR are positively or negatively correlated with Y.
3. If age =20, weight=33 and BMR is 37, what will be the value of Y?

PYTHON CODE RECIPE

In [35]:
```
## Multivariate Analysis Code ##

import numpy as np
import pandas as pd
from sklearn.datasets import load_boston
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [36]:
```
data = load_boston()
boston = pd.DataFrame(data.data, columns=data.feature_names)
boston.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 13 columns):
CRIM        506 non-null float64
ZN          506 non-null float64
INDUS       506 non-null float64
CHAS        506 non-null float64
NOX         506 non-null float64
RM          506 non-null float64
AGE         506 non-null float64
DIS         506 non-null float64
RAD         506 non-null float64
TAX         506 non-null float64
PTRATIO     506 non-null float64
B           506 non-null float64
LSTAT       506 non-null float64
dtypes: float64(13)
memory usage: 51.5 KB
```

In [23]:
```
correlation_matrix = boston.corr().round(2)
sns.heatmap(data=correlation_matrix, annot=True)
```

Out[23]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x217cb114148>
```

In [24]:
```
boston['MEDV'] = data.target
```

In [26]:
```
X = pd.DataFrame(np.c_[boston['LSTAT'], boston['RM']], columns=['LSTAT',
'RM'])
Y = boston['MEDV']
```

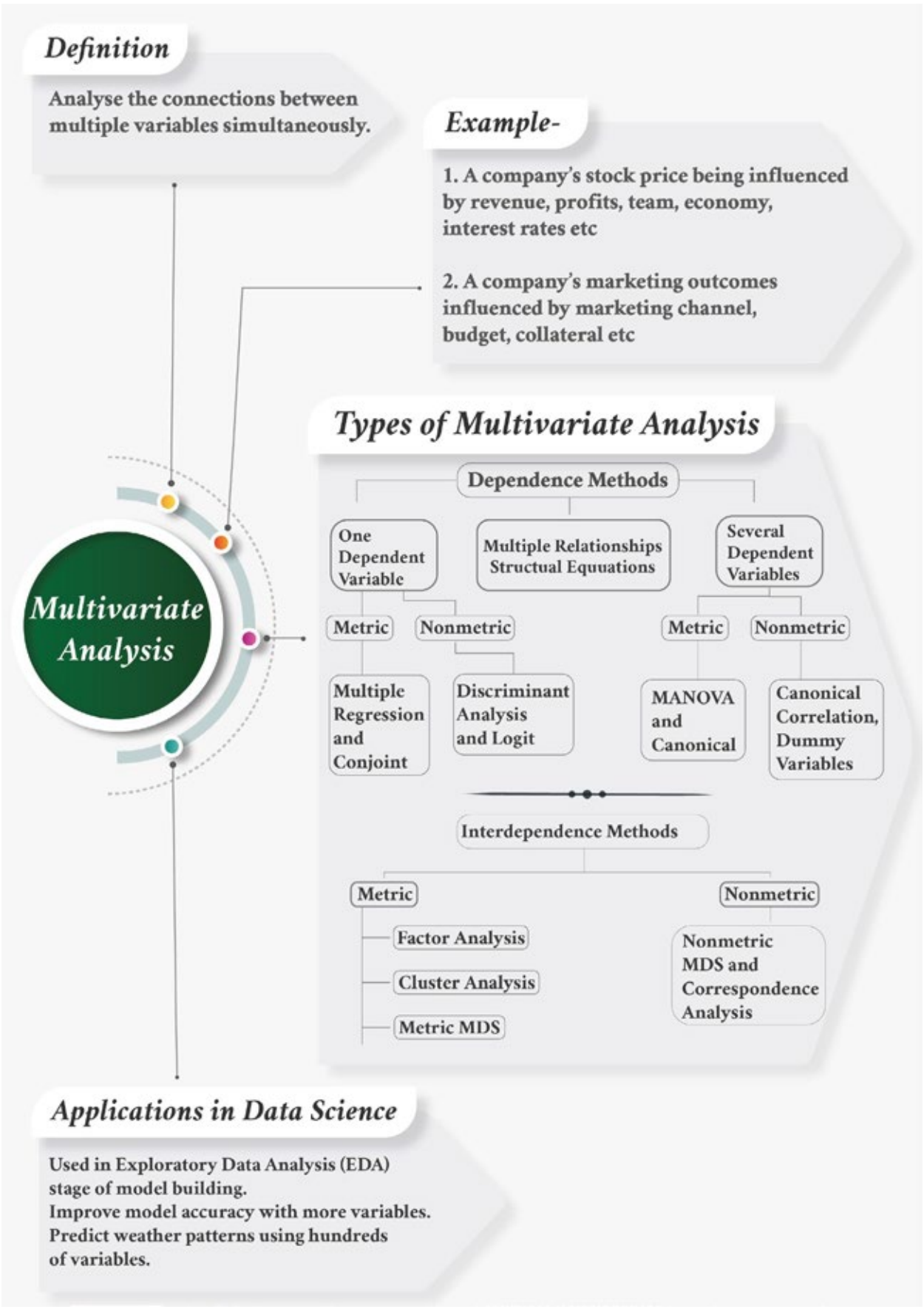
In [30]:
```
lin_reg_mod = LinearRegression()

lin_reg_mod.fit(X, Y)
```

Out[30]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normaliz
e=False)
```

In [34]:
```
lin_reg_mod.coef_
```

Out[34]:
```
array([-0.64235833,  5.09478798])
```

In [ ]:
```
# Thus the equation becomes

#MEDV= -0.64 * LSTAT + 5.09 * RM
```


Definition
Analyse the connections between multiple variables simultaneously.
Example-
1. A company's stock price being influenced by revenue, profits, team, economy, interest rates etc
2. A company's marketing outcomes influenced by marketing channel, budget, collateral etc
Multivariate Analysis
Types of Multivariate Analysis
Dependence Methods
One Dependent Variable
Multiple Relationships Structual Equuations
Several Dependent Variables
Metric
Nonmetric
Metric
Nonmetric
Multiple Regression and Conjoint
Discriminant Analysis and Logit
MANOVA and Canonical
Canonical Correlation, Dummy Variables
Interdependence Methods
Metric
Nonmetric
Factor Analysis
Cluster Analysis
Metric MDS
Nonmetric MDS and Correspondence Analysis
Applications in Data Science
Used in Exploratory Data Analysis (EDA) stage of model building.
Improve model accuracy with more variables.
Predict weather patterns using hundreds of variables.

MODULE 6

SAMPLE AND POPULATION

OBJECTIVES

By the end of this module you will be able to:

1. Understand what sample and population are
2. Differentiate between sample and population
3. Understand parameters and statistics
4. Understand Standard deviation and Standard error

POPULATION

Assume you want to run an analysis on all credit card transactions of a popular crest card 'BankWorldCard'. That entire list of transactions is called the 'population'. A population is the entire group that you want to draw conclusions about.

A statistical population is a set of entities (people or things) from which statistical inferences can be drawn. So the term 'population' in statistics does not refer to only humans - it can be a collection of any phenomena (people, events, objects etc) having something in common that needs to be analysed.

SAMPLE

Many times when you want to analyse an entire population, there are restricting factors such as data size, incomplete data etc. In such cases, you will choose a small subset of the population to analyse. This subset will have characteristics that are representative of the entire population.

This smaller subset of members selected to characterise the population is known as a 'sample'.

Since it is impossible to observe the entire statistical population, due to various constraints, a statistical sample is selected from the population in order to understand the population in its entirety.

For example, the sample may be some of BankWorldCard's credit card transactions.

Parameter

A parameter is any number that conveys information about the population. For example the mean or standard deviation of the population.

Statistic

The statistical measurement of a sample is known as a statistic For example, mean of a sample, standard deviation of a sample etc.

Inferential statistics lets you make an educated guess about a population's parameter based on random samples from the population.

For instance, let's say you need to calculate the mean income of subscribers to TV streaming service NetStar. Mean income here is a parameter of the NetStar subscribers population. You draw a random sample of 1,000 subscribers and calculate that their mean income is $10,000 (a statistic). You learn that the mean income r of the population is close to $10,000. This example is one of statistical inference.

Standard Deviation and Standard Error

When we are given a dataset for analysis, one of the first things we do to get an idea of the "typical" element in the dataset is find out the mean and median. This is a good starting point, but we also need to know the overall shape of our data or how spread out our data is.

For example, consider the closing stock prices of 2 companies across 3 days:

Company A - $42, $39, $45

Company B - $34, $39, $52

Even though the average of both the companies closing stock prices is the same $42, Company B's stock price has a wider range $34 - $52 and hence a higher standard deviation from the mean and hence higher volatility.

Standard deviation indicates how far away each measurement is from the mean of a dataset. It is popularly used in finance where standard deviation of price data is used to understand volatility.

A low standard deviation indicates that there is less variability in the data and that the dataset is more reliable. A high standard deviation indicates that there is nigh variability in the data and that the dataset is less reliable. The standard deviation is a reliable metric of the variability in a dataset as long as there are not too many outliers.

The standard deviation is calculated as follows:

- Step 1 - Calculate the mean of the sample.
- Step 2 - For each value calculate its deviation from the mean (by subtracting the mean from the data value).

- Step 3 - Square each of the deviations calculated in step 2.
- Step 4 - Calculate the average of the squares from step 3. This is called the variance.
- Step 5 - The square root of the variance from step 4 is the standard deviation.

The standard deviation metrics is usually represented along with the mean metric for a sample.

If you divide the standard deviation by the square root of the number of observations you get an estimate of the standard error of the mean.

$$SEM = SD / \sqrt{n}$$

A standard error is a measure of precision of an estimate of a population parameter. A standard error is attached to a parameter. There can be standard errors of any estimate - mean, median, etc.

Applications of Sample and Population in data science and machine learning

Data science requires statistical analysis and your first step is to find out if your dataset is a population or a sample. The main reasons to use sampling are:

- Necessity: In certain scenarios, it's not possible to study the full population due to its large size or difficulty to get the data.
- Practicality: Lesser time and resources are needed to get small amounts of data from a sample.
- Cheaper: Far fewer infrastructure and team is needed to get this done.
- Complexity: It's far less complicated and less technical to run analysis on small datasets.

Handling Imbalanced Data

When working on classification problems (mostly in fraud detection). The number of occurrences of one class (non fraud transactions) might be far higher than another class (fraud transactions). This makes it difficult for most machine learning models to predict new instances of the minority class.

This problem can be solved either by OverSampling - adding more copies of the minority class or UnderSampling - reducing copies of the minority class.

Handling Large Datasets

Most times, We are faced with large amounts of data with millions of rows containing gigabytes. Analysing this data might be computationally expensive, especially when we don't have a powerful computer to crunch this huge amount of data. In such circumstances, sampling of this data might be a good option to resolve to.

Imagine you want to do a survey of cell phone users in Nigeria. You certainly need to collect data maybe through the distribution of questionnaires. Nigeria has a population of about 202,844,689 million people. Therefore, distributing these questionnaires to each individual in Nigeria would be impossible. This could be due to limited-time, resources and other social factors .

We resolve to sampling. So instead of carrying out a survey on the entire population, you can pick a subset of the population, about 1 million people ,and carry out this survey. But when picking this subset, you must endeavour that this subset captures the characteristics/ distribution of the entire population. You would observe that distributing questionnaires to 1 million people is more realistic and easier than distributing to about 200 million people.

QUIZ 6

1. Identify the Population:

   Telcostar wants to understand what apps do customers in tier 3 cities prefer. There are a total of 45 million customers in tier 3 cities. They send surveys to 15,000 customers, 3,500 like education apps, 10,000 like entertainment apps and 1,500 like news apps.

   a. 45 million

   b. 15,000

   c. 3,500

   d. 10,000

   e. 1,500

2. Identify the Sample:

   Telcostar wants to understand what apps do customers in tier 3 cities prefer. There are a total of 45 million customers in tier 3 cities. They send surveys to 15,000 customers, 3,500 like education apps, 10,000 like entertainment apps and 1,500 like news apps.

   a. 45 million

   b. 15,000

   c. 3,500

   d. 10,000

   e. 1,500

3. For the following dataset find mean and variance

   10, 12, 23, 23, 16, 23, 21, 16

PYTHON CODE RECIPE

Download the iris data set from the web.

Let's check size of the data

```
import pandas as pd
```

```
t=pd.read_csv("iris_csv.csv")
```

```
t.shape
```

```
(150, 5)
```

This shows there are 150 rows and 5 columns meaning a total 150 entries with 5 properties each. This is the population size

```
t.sample(10)
```

	sepallength	sepalwidth	petallength	petalwidth	class
91	6.1	3.0	4.6	1.4	Iris-versicolor
50	7.0	3.2	4.7	1.4	Iris-versicolor
5	5.4	3.9	1.7	0.4	Iris-setosa
89	5.5	2.5	4.0	1.3	Iris-versicolor
106	4.9	2.5	4.5	1.7	Iris-virginica
142	5.8	2.7	5.1	1.9	Iris-virginica
2	4.7	3.2	1.3	0.2	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
109	7.2	3.6	6.1	2.5	Iris-virginica

The sample function is randomly selecting 10 entries.

Generating 5% sample of data frame:

```
t.sample(frac=0.05)
```

	sepallength	sepalwidth	petallength	petalwidth	class
37	4.9	3.1	1.5	0.1	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
46	5.1	3.8	1.6	0.2	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
127	6.1	3.0	4.9	1.8	Iris-virginica
132	6.4	2.8	5.6	2.2	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica

Sample and Population

Sample

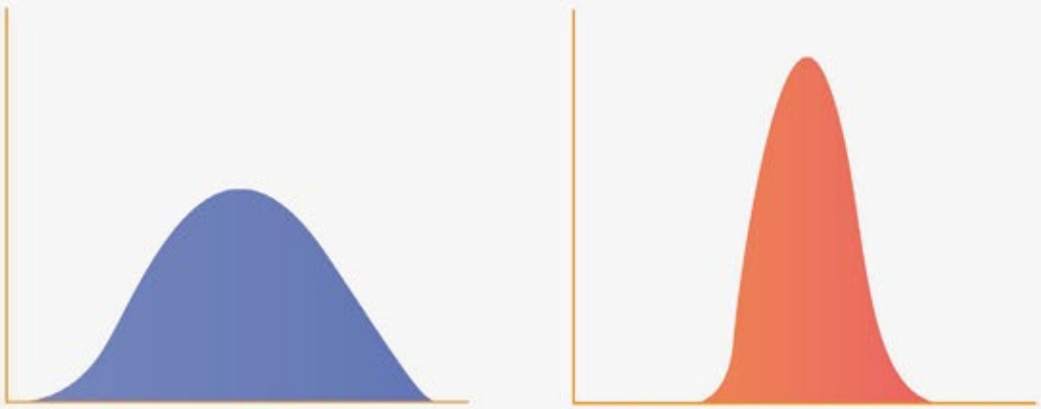
A sample is a subset of the population.

Population

A population is the entire group that you want to draw conclusions about.

Standard Deviation

Indicates how far away each data value is from the mean.



Data Science Applications:

1. To run analysis on large datasets.
2. Cheaper and faster to analyse a small sample instead of the full dataset.

MODULE 7

SAMPLING TECHNIQUES

OBJECTIVES

By the end of this module you will be able to:

1. Understand why there are different sampling techniques.
2. Detail the various sampling techniques.
3. Understand why re-sampling is used.

SAMPLING

As discussed in the last module, a sample is a subset of a larger 'population'. You sample in order to reduce the time and resources needed for your data science analysis. The entire population could be either too big, or incomplete or takes too much time to analyse.

The key requirement for sampling is that the sample should be representative of the larger population. If this sampling process goes wrong, even the best machine learning algorithms are useless since your model's conclusions will not represent the population's characteristics.

For example consider the case of a TV streaming company Starflix - if you want to understand the characteristics of Starflix customer churn, your sample should have males, females, viewers of different genres etc. If your sample leaves out males, your churn analysis will not be very accurate.

There are many established methods to sample. They are broadly classified into:

1. Probability sampling
2. Non-probability sampling

PROBABILITY SAMPLING

These methods use randomization to ensure that every element of the population has an equal chance of being in the sample. This increases the likelihood that you will pick a balanced sample. This is also known as random sampling. Probability sampling tends to be more expensive and time consuming compared to other methods but has the best chance of generalising the analysis results to the larger population.

THERE ARE 4 TYPES OF PROBABILITY SAMPLING

1. Simple Random Sampling

A sample is generated from the population purely at random. This is like assigning a number to every Starflix customer (a population of 5 million customers) and picking 1,000 random numbers/customers as your sampling frame. The advantage is that it is simple and fast. The disadvantage is that certain characteristics of the population might be left out of the sample, especially if those characteristics are rare.

2. Systematic Sampling

Rather than randomly picking sample members, you will pick sample members at regular intervals. So from the previous example, rather than picking 1,000 random Starflix customers, you will pick 1 customer every 5,000 (5 million divided by 1,000) customers.

This is easier to do than random sampling since it is computationally less intensive, but could lead to bias if the underlying customer seres has a pattern that is prevalent in certain sequences - like for example if Starflix customers are listed by genre and the genre is ordered by descending age, you run the risk of your interval might skip over younger customers in each genre.

3. Stratified Sampling

The previous 2 methods run the risk that your sample may miss out on some characteristics of the population. Stratified sampling fixes that issue by first making a list of 'strata' or characteristics that you want represented in your sample.

For example, you will split Starflix customers into different 'strata' such as comedy, thriller, drama, romance and kids. Then you will use either random or systematic sampling to pick from each of the strata. While doing this, you can adjust the number of samples from each strata in proportion to the number of elements in that strata.

This way you can be sure that all characteristics are there in your sample. However to effectively implement this, you will need to have domain knowledge of your dataset in order to pick the right stratas.

4. Cluster Sampling

In this method you will split the population into clusters randomly and pick 1 or more clusters for your sample. The clusters are not split according to any specific characteristics. There is a higher risk of sampling errors in this method because of substantial differences between the clusters.

Clustered sampling is usually used when trying to sample a group of people spread over a wide geographical area. Then it is cheaper and faster to pick a couple of smaller areas to sample. The risk is that these smaller areas might be alike and the non sampled areas might be widely different.

NON-PROBABILITY SAMPLING

These sampling techniques use non-random ways to pick the samples. Due to this they are prone to bias and might need domain knowledge.

Probability sampling is used to understand the general characteristics of a large population. However non-probability sampling is used for exploratory data analysis to get an initial understanding of the data.

1. Convenience Sampling

This is a quick and dirty way to sample. You sample based on who or what is immediately available to sample. For example, if the Starflix customer support team wanted customer feedback, by using convenience sampling they would quickly call a few friends. This will give them some feedback but there is no way to be sure if this is representative of their population (their customer base).

2. Quota Sampling

A quota is made whereby the proportion of elements in the sample is the same as the population. Then these elements are analysed. For example, if 60% of Starflix customers watch on the weekend and 40% on weekdays, then that;s your quota. Your sample will also have 60% weekend viewers and 40% weekday viewers.

While this is easy to accomplish and might be representative of the population, you run the risk of not covering other characteristics in the population -like for example not covering kids customers.

3. Judgement or Purposive Sampling

This type of sampling is guided by the purpose or the intention of the analysis. If your analysis is to find out churn among Starflix customers who watch thriller movies on weekends, then you will ignore all other customers and only pick from the universe of thriller customers.

Though this is quick and easy to get done, it is prone judgement errors from the analyst

4. Snowball Sampling

This method is something like multi-level marketing. You first get the initial cohort of sample elements / people. Then you get the next cohort of people for your sample through the people from the first cohort.

This is usually deployed to sample groups of people that are hard to reach.

RESAMPLING

Once you have a data sample you can estimate the parameter of a population. The challenge is that you have only 1 estimate and the variability of the population is no knowd. Resampling solves this problem by estimating the same parameter multiple times.

2 commonly used resampling methods in machine learning are ‘Bootstrap’ and ‘k-fold cross-validation’.

APPLICATION OF SAMPLING TECHNIQUES IN DATA SCIENCE

Sampling is one of the most commonly used techniques in the world of data science and is popularly used in predictive modelling.

A prime example of this is a credit card use case. Banks do not give credit cards to most people, they select a few people who meet their criteria and only give credit cards to them. In this case because of a data imbalance we use stratified sampling so that we get an even sample of all the people who have and have not got the credit card.

Let us assume that the bank has 1,000 applicants for a credit card and approved only 10 applicants Thus the ratio is 99:1 for rejected applications vs accepted applications.

In this scenario we will use stratified sampling to maintain this ratio (99:1) for our sample i.e if we pick a sample of size 100 rows, then 99 would be that of rejected applications and 1 of the accepted one.

If the data is balanced (the ratio is 50:50), we will use random sampling to get an even sample of the data.

QUIZ 7

1. In order to use samples to estimate something from the population, the sample should be _____ the population.

   a. exactly the same as

   b. nothing like

   c. representative of

   d. larger than

2. If every individual in a population has the same chance of being included in a sample, the sample is a _____ sample.

   a. Biased Sample

   b. Convenience Sample

   c. Random Sample

   d. Stratified Sample

3. Mr. ABC samples his class by selecting 5 girls and 7 boys. This type of sampling is called?

   a. Stratified

   b. Systematic

   c. Simple

   d. Cluster

4. The school librarian wants to determine how many students use the library on a regular basis. What type of sampling method would she use if she chose to use a random number generator to randomly select 50 students from the school's attendance roster.

   a. Convenience Sample

   b. Simple Random Sample

   c. Stratified Random Sample

   d. Interval Sample

PYTHON CODE RECIPE

Download the iris dataset

```
import pandas as pd
```

```
t=pd.read_csv("iris_csv.csv")
```

```
t.shape
```

```
(150, 5)
```

1. Simple random sample

   Let's select 8 items randomly and sort them according to 'petalwidth'

```
random_sample = t.sample(n=8).sort_values(by='petalwidth')
```

```
random_sample
```

	sepallength	sepalwidth	petallength	petalwidth	class
8	4.4	2.9	1.4	0.2	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
62	6.0	2.2	4.0	1.0	Iris-versicolor
76	6.8	2.8	4.8	1.4	Iris-versicolor
84	5.4	3.0	4.5	1.5	Iris-versicolor
56	6.3	3.3	4.7	1.6	Iris-versicolor

2. Systematic sampling:

   Let's select entries based on a fixed sampling interval, say 10.

```
import numpy as np
```

```
systematic_sample = t.iloc[np.arange(0,len(t),step=10)]
```

```
systematic_sample
```

	sepallength	sepalwidth	petallength	petalwidth	class
0	5.1	3.5	1.4	0.2	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
20	5.4	3.4	1.7	0.2	Iris-setosa
30	4.8	3.1	1.6	0.2	Iris-setosa
40	5.0	3.5	1.3	0.3	Iris-setosa
50	7.0	3.2	4.7	1.4	Iris-versicolor
60	5.0	2.0	3.5	1.0	Iris-versicolor
70	5.9	3.2	4.8	1.8	Iris-versicolor
80	5.5	2.4	3.8	1.1	Iris-versicolor
90	5.5	2.6	4.4	1.2	Iris-versicolor
100	6.3	3.3	6.0	2.5	Iris-virginica
110	6.5	3.2	5.1	2.0	Iris-virginica
120	6.9	3.2	5.7	2.3	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
140	6.7	3.1	5.6	2.4	Iris-virginica

3. Cluster Sampling:

Divide data into cluster of equal size. Lets say each cluster has 10 entries.

```
# Cluster Sampling
no_of_clusters=15
```

```
t['cluster_id'] = np.repeat([range(1,no_of_clusters+1)],len(t)/no_of_clusters)
t
```

	sepallength	sepalwidth	petallength	petalwidth	class	cluster_id
0	5.1	3.5	1.4	0.2	Iris-setosa	1
1	4.9	3.0	1.4	0.2	Iris-setosa	1
2	4.7	3.2	1.3	0.2	Iris-setosa	1
3	4.6	3.1	1.5	0.2	Iris-setosa	1
4	5.0	3.6	1.4	0.2	Iris-setosa	1
...	...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica	15
146	6.3	2.5	5.0	1.9	Iris-virginica	15
147	6.5	3.0	5.2	2.0	Iris-virginica	15
148	6.2	3.4	5.4	2.3	Iris-virginica	15
149	5.9	3.0	5.1	1.8	Iris-virginica	15

150 rows × 6 columns

Append the indexes from the clusters that meet the criteria. We will select the cluster with cluster ID divisible by 7.

```
index = []
for i in range(0,len(t)):
    if t['cluster_id'].iloc[i]%7 == 0:
        index.append(i)
```

```
cluster_sample = t.iloc[index]
```

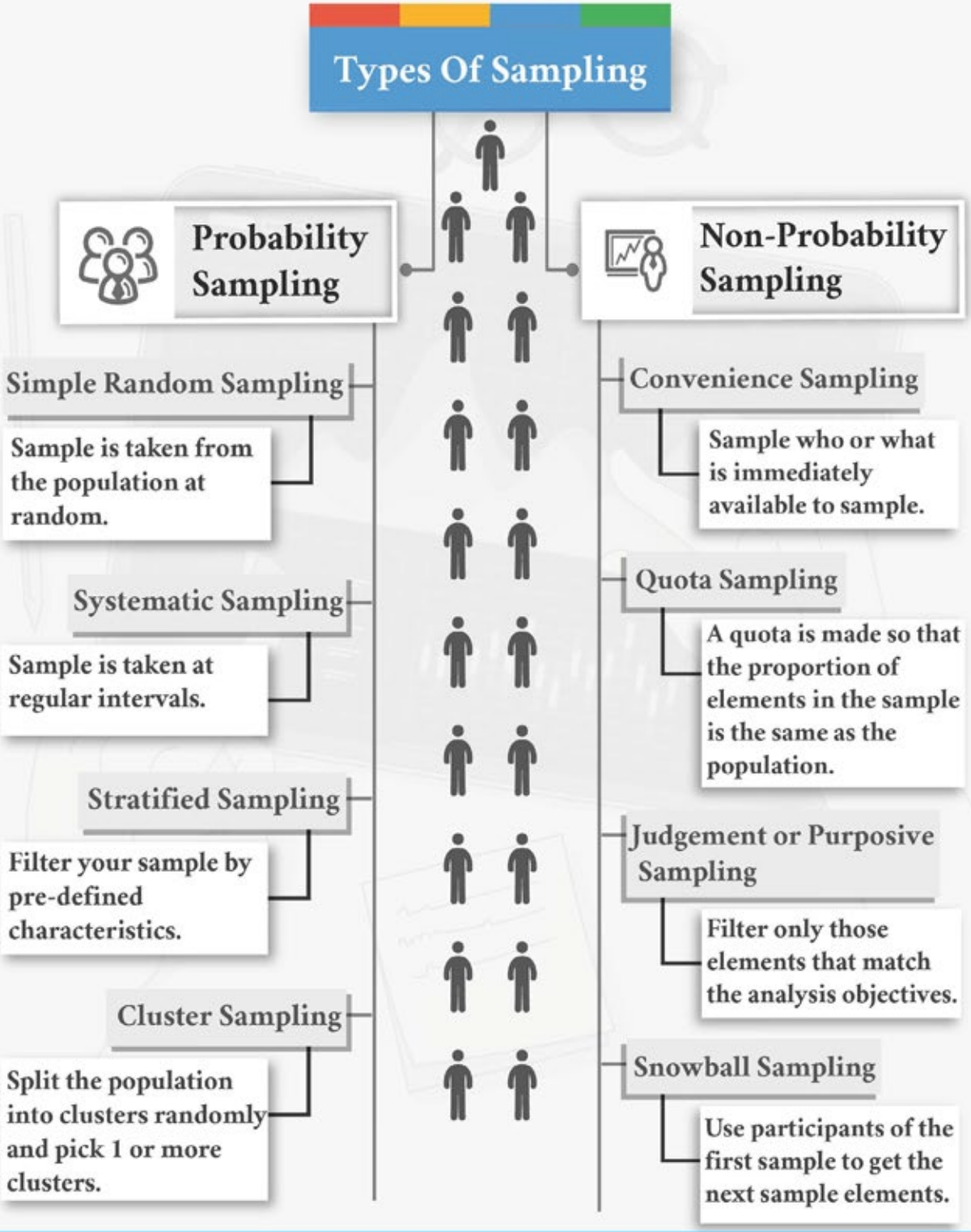
```
cluster_sample
```

```
cluster_sample
```

	sepallength	sepalwidth	petallength	petalwidth	class	cluster_id
60	5.0	2.0	3.5	1.0	Iris-versicolor	7
61	5.9	3.0	4.2	1.5	Iris-versicolor	7
62	6.0	2.2	4.0	1.0	Iris-versicolor	7
63	6.1	2.9	4.7	1.4	Iris-versicolor	7
64	5.6	2.9	3.6	1.3	Iris-versicolor	7
65	6.7	3.1	4.4	1.4	Iris-versicolor	7
66	5.6	3.0	4.5	1.5	Iris-versicolor	7
67	5.8	2.7	4.1	1.0	Iris-versicolor	7
68	6.2	2.2	4.5	1.5	Iris-versicolor	7
69	5.6	2.5	3.9	1.1	Iris-versicolor	7
130	7.4	2.8	6.1	1.9	Iris-virginica	14
131	7.9	3.8	6.4	2.0	Iris-virginica	14
132	6.4	2.8	5.6	2.2	Iris-virginica	14
133	6.3	2.8	5.1	1.5	Iris-virginica	14
134	6.1	2.6	5.6	1.4	Iris-virginica	14
135	7.7	3.0	6.1	2.3	Iris-virginica	14
136	6.3	3.4	5.6	2.4	Iris-virginica	14
137	6.4	3.1	5.5	1.8	Iris-virginica	14
138	6.0	3.0	4.8	1.8	Iris-virginica	14
139	6.9	3.1	5.4	2.1	Iris-virginica	14

Sampling Techniques

A sample should be representative of the larger population.



Data Science applications:

Ensure that the sample chosen reflects the population and the needs of the analysis.

MODULE 8

SAMPLING SIZE CALCULATIONS

OBJECTIVE

By the end of this module you will be able to:

1. Understand what sampling size is
2. Find sampling error and confidence intervals
3. Perform sample size calculations

SAMPLING SIZE

Sampling size is nothing but selecting a few elements/ parts from a larger group (population) for your analysis. For example, if you want to do a customer satisfaction survey in a telecom company to understand what types of customers like what kind of products. You will not survey every single customer, since that takes time and money. You will pick a small set of customers (the sample), whose characteristics match the larger group (the population).

The important part here is choosing the right sample i.e selecting the customers in the right manner. If you bias your customer sample, with either too few or too many customers of a certain type, then your analysis will be skewed and not accurate.

Thus, it is imperative for us to get a good representation of the overall population at play.

APPLICATIONS IN DATA SCIENCE

Sampling Size is one the most integral parts of any data science model building activity.

For example, let us assume that you are building a model to find fraud transactions for a bank. This bank receives nearly 20 million credit card transactions every hour.

As the rate of transactions is very high, you cannot even use a complete days data for building your statistical model. Hence you will try to mirror the population or the overall transactions by taking a stratified sample.

SAMPLING ERROR & CONFIDENCE INTERVALS

Sampling error and Confidence Intervals are among the most important metrics in statistics.

These two metrics alone can determine how good your analysis is. If you pick the wrong sample, no amount of sophisticated machine learning algorithms and frameworks will help.

Hence it is important to get this stage in your analysis right.

In the example above, let us take a scenario where you took a very small sample. Meaning your population had 10,000 customers, but you took only 5 customers of Product A and 5 customers of Product B for the customer satisfaction survey.

Since you are looking into a very small part of the population, your confidence in the survey responses drops. Thus, uncertainty creeps in. You are essentially not capturing the essence of the population in your sample. This is known as sampling error. Sampling error is measured by the **Confidence Interval metric (C.I).** It tells you how much uncertainty there is in any metric.

CONFIDENCE LEVEL & MARGIN OF ERROR

Let's take another scenario when you picked a sample of 2,500 customers for your survey. This is obviously a much better sample size. If you convey the survey results by saying that 71% of the customers like Product A with a Confidence Interval of 4%, this means that between 67% (71-4) to 75% (71+4) of all customers (the entire 10,000) will like Product A. The 4% is also known as the **Margin of error.** The smaller your margin of error, the more accurate your results are. The margin of error indicates how much your sample mean differs from your population's mean.

If your **confidence level** of the above survey is 95%, that means that if that survey was repeated multiple times using the same techniques, 95% of the times the results would be similar to the published results (71% with a margin of error of 7%). So confidence level basically conveys how confident you are about the results of the survey.

Margin of Error (C.I) is generally calculated in terms of means. For this, generally the mean of the population and mean of sample is calculated and then its difference is found out.

If that difference is within your limits or boundaries, then it is fine, if not then you might have to change the sample.

STEPS IN SAMPLE SIZE CALCULATIONS

- Step 1 - Identify your population size (10,000 customers in our example)
- Step 2 - Lock down your margin of error / C.I. Let's assume 5.
- Step 3 - Determine the confidence level. 95% is most often used. This has to be converted to a z-score from a z-table. The z-score for 95% is 1.96.
- Step 4 - Identity the standard deviation you are comfortable with. This number indicates how much variance there is between each sample unit and the average population. Let's assume 0.75.

There are 2 scenarios in calculating the sample size.

1. Calculating sample size with unknown population size:

Sample Size = $z^2 * p * (1-p) / e^2$

Sample Size = ((Z-score)^2 * Standard Deviation * (1-Standard Deviation)) / (Margin of error)^2

Your values are 95% confidence level (1.96 Z-score), standard deviation of 0.75 and margin of error 5. Plug these into the formula -

Sample size= $(1.96)^2 * 0.75 * 0.25 / (0.05)^2 = 288$

This means that to have a good sample, you need a sample size of 288 customers to mirror the overall population parameters.

2. Calculating sample size with a known population size:

Sample size = $(z^2 * p * (1-p))/e^2 / 1+ (z^2 * p * (1-p) / e^2 * N)$

z = Z score = 1.96
P = standard deviation = 0.75
e = margin of error = 5%
N = population size = 10,000 (the total number of customers in our example)

So sample size = $(1.96^2 * 0.75 * (1-0.75))/0.05^2 / 1+ (1.96^2 * 0.75 * (1-0.75) / 0.05^2 * 10,000)$ Sample Size = 370

This means that to have a good sample, you need a sample size of 370 customers to mirror the overall population parameters.

QUIZ 8

1) Find sample size required for the following Std Dev: 0.5; Confidence Level: 95%; Confidence Interval: 5

2) Find sample size required for the following

   Population: 1000; Std Dev: 0.5; Confidence Level: 95%; Confidence Interval: 5

In [7]:
```
#Sampling Size

from scipy.stats import norm
```

In [3]:
```
def sampling_size(population_size,
    margin_error=.05,
    confidence_level=.99,
    sigma=1/2 #sigma is std dev
):
    alpha = 1 - (confidence_level)
    #Creating dictionary for confidence levels and corresponding z-score
s
    #computed via norm.ppf(1 - (alpha/2))
    zdict = {
        .90: 1.645,
        .91: 1.695,
        .99: 2.576,
        .97: 2.17,
        .94: 1.881,
        .93: 1.812,
        .95: 1.96,
        .98: 2.326
    }
    if confidence_level in zdict:
        z = zdict[confidence_level]
    else:

        z = norm.ppf(1 - (alpha/2))
    N = population_size
    M = margin_error
    numerator = z**2 * sigma**2 * (N / (N-1))
    denom = M**2 + ((z**2 * sigma**2)/(N-1))
    return numerator/denom
```

In [5]:
```
sampling_size(1000,margin_error=.05,confidence_level=.99,sigma=1/2)
```

Out[5]:
```
399.12579118111546
```

In [6]:
```
#This is the sampling size for population size 1000
```

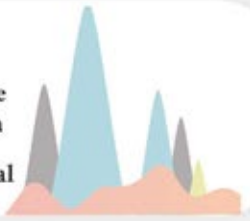
In [ ]:
```

```

Sampling Size Calculations

Sampling errors occur when your sample does not capture the population's characteristics.

Sampling error is measured by the Confidence Interval metric (C.I.).

95% confidence that
85% of customers like Product A with a
4% confidence interval



Calculating Sample Size

1) With unknown population size

$$\text{Sample Size} = \frac{z^2 * p * (1-p)}{e^2}$$

z - z score, p - standard deviation, e - margin of error

2) With known population size

$$\text{Sample size} = \frac{\frac{(z^2 * p * (1-p))}{e^2}}{1 + \frac{(z^2 * p * (1-p)}{e^2 * N)}}$$

N - population size

Applications in Data Science

Picking the right sample size is critical to ensure that the sample analysis reflects the characteristics of the larger population.

MODULE 9

CORRELATION, COVARIANCE AND INTERQUARTILE RANGES

OBJECTIVES

By the end of this module you will be able to:

1. Understand what correlation is
2. When to use covariance and what it means
3. Understand quartiles and IQR's
4. Significance of correlation, covariance & IQR in Data Science

APPLICATION IN DATA SCIENCE

The whole idea behind data science and machine learning is to understand how the different variables in your dataset relate to each other. Once you understand that relationship, its strength and direction, you can choose or build a model, train it and then use the model to predict the future.

Correlation and Covariance are 2 statistical techniques that help you do the above. Both of them almost meant the same thing.

COVARIANCE

Covariance indicates how 2 or more variables vary together. It quantifies the extent to which one variable deviates from its mean compared to another variable deviating from its mean. Simply put, covariance measures how 2 variables change together

Covariance is used widely in the following scenarios:

1. In finance, for asset allocation also known as portfolio theory
2. In genetics and biology, covariance is used to track gene development and gene changes across 2 different genetic mutations.

Digging deeper into how covariance is used in stock asset allocation. A primary example is to measure which 2 stocks move similar to each other. They may not necessarily move due to each other and there might be another reason that is triggering this movement. Finding stocks that have positive or negative covariance will tell you how they will behave under certain market conditions and also help you identify why the stocks are moving. While constructing a stock portfolio, you ideally want a mix of stocks that have positive and negative covariance.

Covariance of 2 variables X and Y can be calculated with the following formula:

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$

E[X] and E[Y] are the average values of X and Y from the 2 samples. So (X-E[X]) and (Y-E[Y]) tells you how far away from the samples average the specific X and Y value falls.

Multiplying this gives the covariance number. A positive value to this means that X and Y move in the same direction (for example precious metal prices like gold and silver) . A negative value to this means that X and Y move in the opposite direction (for example airline stocks and fuel prices).

For example, take the following pricing data for shares of 2 companies A & B

Week	Share A	Share B
1	70	72
2	66	74
3	72	58
4	65	56

Step 1 - Calculate the mean of both these groups.

Mean of share A= 68.25 and Mean of share B = 65

Step 2 - Plug the values into the Covariance formula

$$Cov(A,B) = [(70-68.25) \times (72-65) + (66-68.25) \times (74-65) + (72-68.25) \times (58-65) + (65-68.25) \times (56-65)] / 4 = -1.25$$

As the covariance is negative, we would say that the shares move in opposite directions

CORRELATION

Correlation is the mutual relationship between 2 or more variables. Correlations can be seen across many examples in our society:

1. Carbon emissions are highly correlated with the number of vehicles on road
2. Work experience is highly correlated with salary
3. Life expectancy is highly correlated with tobacco consumption

Correlation being a statistical measure is represented with a number between the range -1 to +1. This value measures both the strength and the direction of the correlations. Higher the number, stronger the correlation.

A value of +1 indicates a 100% **positive correlation**, meaning the 2 variables move directly proportional to each other. A value of -1 indicates a 100% **negative correlation**, meaning the 2 variables move inversely proportional to each other. A value of 0 means there is no correlation between the variables.

When you are given a large amount of data, the best way to find out correlations is by plotting a scatter plot. The scatter plot below depicts "Years of work experience" vs "Salary".

As you can see, the salary keeps increasing when the experience increases. This means that when one variable is increased the other variable also increases. This is called a positive correlation. This does not necessarily mean it is a +1 (100% positive) correlation, it only means that the correlation value is greater than 0.

In the next scatter plot example below, we plot "Life expectancy" vs "Tobacco usage".

As you can see the two variables are inversely proportional to each other, meaning if you increase one you lower the other. As the tobacco usage increases, life expectancy decreases. This is negative correlation. Again, this does not necessarily mean it is a -1 (100% negative) correlation, it only means that the correlation value is lesser than 0.

If the correlation value is between 0 and 1, the variables are positively correlated and between 0 and -1 the variables are negatively correlated.

There are various methods to calculate correlation between 2 variables. The most commonly used is the Pearson's Correlation Coefficient. The correlation between 2 variables X and Y is calculated using:

p (X,Y) = Cov (X,Y) / Standard deviation (X) Standard deviation (Y)

An important point to note here is correlation does not mean causation, which means if 2 variables are highly correlated, that essentially doesn't mean that the change in one variable is caused by change in another. It could be but does not have to be. This interpretation mistake is a common cause of errors while building data science models.

INTER QUARTILE RANGE

When you are given a dataset to build a model. One of the first things you need to know is the range of your data. The full range of your data will tell you the spread from the smallest to the largest value. Though this information is helpful, it factors in outliers and may not necessarily tell you the true spread of your data, or the range in which most of your data values lie.

This problem is solved by finding out the inter quartile range of your dataset. This will tell you the range in which most of your data lies.

A quartile refers to 25% of your dataset. The first quartile means the first 25%. So every dataset will have 4 quartiles. The interquartile is calculated by subtracting the first quartile from the third quartile. This gives you the range for the the middle 2 quartiles of your dataset.

For example, let us take the below example and divide the data into quartiles.

Data = [5, 6, 4, 4, 8, 2, 9]

The first step is to sort the data in ascending order

Sorted_data= [2, 4, 4, 5, 6, 8, 9]

Each quartile length= n/4= 7/4= 2 approx

Q1 contains 2, 4
Q2 contains 4, 5
Q3 contains 6, 8
Q4 contains 9

Interquartile range is from quartiles 1 to 3 (Q1 to Q3)

IQR = Q3-Q1

So in the above example, Q3 values ends at 8 and Q1 values ends at 4. So the IQR becomes

IQR = 8-4 = 4

This tells you that the spread between the middle values of this dataset is 4. The IQR is usually represented using a Box Plot.

QUIZ 9

1. Find covariance between these 2 datasets

   A=[5,2,3,1,2,2]

   B=[2,2,12,3,4,5]

2. Find the IQR of following list

   A= [5, 7, 4, 4, 6, 2, 8]

3. Find the correlation in following 2 datasets

   A:[12,24,35,22,44,77]

   B:[100,122,129,100,190,191]

PYTHON CODE RECIPE

```
In [3]: import numpy as np
        import pandas as pd
        a=np.array([12,33,44,66,33])
        b=np.array([22,44,66,88,44])

In [4]: r = np.corrcoef(a, b)

In [5]: r

Out[5]: array([[1.        , 0.99064704],
               [0.99064704, 1.        ]])

In [6]:  r[0, 1]

Out[6]: 0.9906470354350158

In [8]: #Correlation is 0.99! Lets look at covariance

        np.cov(a,b)[0][1]

Out[8]: 488.40000000000003

In [9]: #So covariance is 488 units which is highly positive meaning both arrays
        move in same direction

In [13]: sample = pd.DataFrame([[500],[200],[300],[400],[500],[100],[700],[800],[
         900],[2000]],columns=["Salary"])

In [20]: sorted(sample["Salary"])
         Q1,Q3 = np.percentile(sample["Salary"] , [25,75])
         IQR = Q3 - Q1
         print(IQR)

         450.0

In [ ]: #here thus IQR is 450
```

Covariance, Correlation & Interquartile Range

COVARIANCE

Indicates how 2 or more variables vary together.
It quantifies the extent to which one variable deviates from its mean compared to another variable deviating from its mean.

Cov (X,Y) = E[(X-E[X]) (Y-E[Y])]
E[X] and E[Y] are the average values of X and Y from the 2 samples.

(X-E[X]) and (Y-E[Y])
indicates how far away from the samples average the specific X and Y value falls.

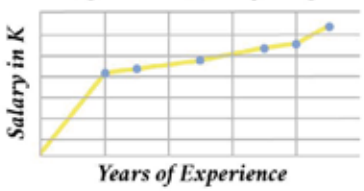
CORRELATION

Correlation is the mutual relationship between 2 or more variables.

$$p(X,Y) = \frac{\text{Cov}(X,Y)}{\text{Standard deviation (X)} * \text{Standard deviation (Y)}}$$

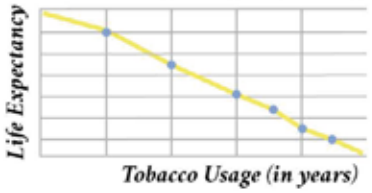
Positive Correlation

Experience vs Salary Graph

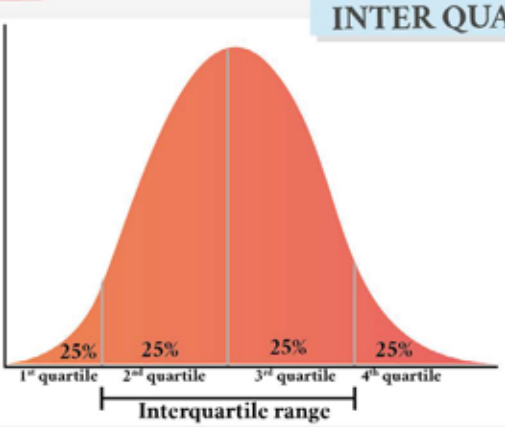


Negative Correlation

Tobacco Usage vs Life Expectancy



INTER QUARTILE RANGE



This tells you the range in which most of your data lies.

IQR = Q3-Q1

Applications in Data Science

Used during EDA (exploratory data analysis) to understand how the variables behave and their range.

MODULE 10

INFERENTIAL ANALYSIS-HYPOTHESIS TESTING

OBJECTIVE

By the end of this module you will be able to:

1. Understand what hypothesis testing is
2. Parametric Tests
3. Non parametric tests
4. Significance of Hypothesis Testing in Data Science

HYPOTHESIS TESTING

Inferential analysis refers to techniques that help you infer and understand the behaviour of a larger population using learnings from a small sample. Hypothesis testing is one such technique.

The term hypothesis means assumption - it is your belief about a phenomenon or an outcome about the population you are dealing with. A hypothesis remains an assumption unless you can prove it with data. Hypothesis testing is a systematic way in which you will make an assumption about a parameter in your population and try to prove whether it is correct or wrong based on the data given.

APPLICATIONS IN DATA SCIENCE

Once you have picked your samples using the various sampling techniques, you will need to draw conclusions about the population from the sample data. Hence, before starting any problem, you create hypotheses in data science. You try to form your null hypothesis and use the data to accept or reject it.

When you want to make claims about data distribution or compare 2 sets of results in machine learning, you rely on hypothesis testing.

You will constantly use hypothesis testing to evaluate how robust your machine learning models are.

When you start building your data science models, you will make multiple hypotheses about possible outcomes and test them. Before starting any problem you will do hypothesis testing, so that you can either confirm or deny the solution approach to the problem.

For example, let's say you are building a model for your marketing team. From market research reports you know that the average customer acquisition cost for your vertical is less than 25% of sales. That is a hypothesis until it is tested on your specific dataset. To test whether this claim is true or not, you will take the customer acquisition cost of a small sample of your dataset. Then you will calculate this sample mean and compare it to the population mean for the same parameter (customer acquisition cost).

TYPES OF HYPOTHESIS

There are 2 types of hypothesis you can make:

1. Null hypothesis (H_0) - In the above example, the hypothesis that "customer acquisition cost is less than 25% of sales" is called the null hypothesis (H_0). At the start of the testing, you assume that the null hypothesis is true. Technically speaking, the null hypothesis states that the difference between the two samples/populations (the market research report and your own dataset) is not statistically significant.
2. Alternate hypothesis (H_1) - Alternate hypothesis is the opposite of a null hypothesis. In the above example, the alternate hypothesis would be that "customer acquisition cost is greater than 25% of sales". This contradicts the null hypothesis.

In hypothesis testing, you try to prove that the null hypothesis is not true, which in turn makes you accept the alternate hypothesis.

STEPS IN HYPOTHESIS TESTING

Step 1 - Define the hypothesis

From the above example, the null hypothesis is that customer acquisition cost is less than 25% of sales. For the purpose of the test, you assume that this statement is true and set out to prove it.

Step 2 - Define the decision criteria

This is done using the "significance level". This is the criteria using which you accept or fail to accept the null hypothesis.

Step 3 - Calculate the hypothesis test

Suppose you took a sample of customer acquisition costs from your dataset and calculated the mean and that number was 30% of revenue. You have to figure out how likely is this sample mean, given that the mean of your population is 25% (as defined by your null hypothesis). This likelihood is represented by the p-value (values between 0 to 1). The larger this

p-value the larger is your sample mean away (in terms of standard deviation) from the population mean.

STEP 4 - MAKE THE DECISION

Compare the p-value from step 3 to the significance level from step 2, to decide. If the probability to obtain this sample mean (30% of revenue) is less than the significance level (usually kept at 5% or 0.05), then the null hypothesis is rejected (because this means that for your population the customer acquisition cost is more than 25% of sales).

p-value and Significance level

The p value is the output of a hypothesis test (in step 3) and helps to quantify the result by accepting or rejecting the null hypothesis. You can achieve this by comparing the p-value to an acceptable threshold value (decided beforehand) called the significance level (from step 2 and denoted as alpha).

TYPES OF HYPOTHESIS TESTS

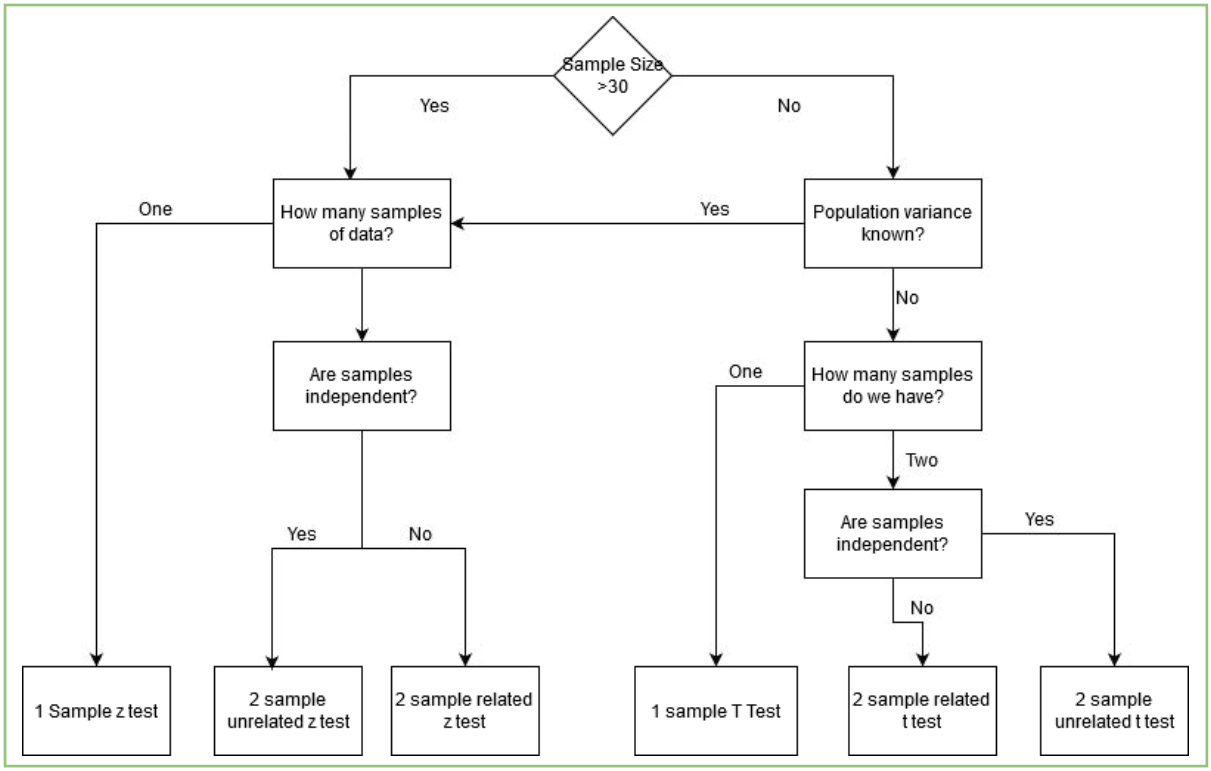
To achieve Step 3 and calculate the p-value there are 2 categories of tests:

1. Parametric Tests:

Parametric tests assume that your dataset follows a normal distribution (also called a gaussian distribution) represented by a bell shaped curve. And as normal distributions are symmetrical, the test also expects data to be symmetrical in nature. However, most real world data sets do not follow normal distribution.

To establish a normal distribution, you need to have a large enough dataset (typically more than 30 samples and ensure data is not skewed due to large outliers.

To decide which test to choose you have to ask a few questions as depicted by this flowchart below.



5% is a common significance level that is used. This means that of all the times the analysis was repeated, a maximum of 5% of the times the output can be due to chance. A p-value less than 5% indicates that your analysis produced results that are not by chance (defined by the 5% threshold) and are statistically significant.

If p-value <= alpha, results are statistically significant and reject the null hypothesis

If p-value > alpha, results are not statistically significant and fail to reject the null hypothesis

The popular parametric tests are the following:

1-sample z-test and t-test:

These tests are applied when you want to compare a sample's mean to the population's mean. You use the z-test if you have the population variance, else use the T-test. The test will give you a p value, which you will compare to the significance level (as described above) to decide whether to reject the null hypothesis or not.

For example: You would calculate a z-test score (the p-value) to find out if the average spend of your customers is over $500. There is only 1 sample you are dealing with here (your customer spend)

2- sample z-test and t-test

This test is applied when you want to compare the mean of 2 samples. You use the z-test if you have the population variance, else use the T-test.

For example: You would use this to find out if customers of Product A spend 10% more than customers of Product B. This test will help you prove or disprove this hypothesis. There are 2 samples here - Product A customer spend and Product B customer spend.

2. Non-Parametric Test

A non-parametric test is done when your dataset doesn't follow any assumption about whether the sample has a normal distribution or not.

You would choose a non-parametric test under one or more of the following circumstances:

- When your data is very skewed, and a median value is a better representation than a mean.
- When your sample size is small (less than 30 values)
- Dataset has many outliers
- Ordinal data (data is ordered by hierarchy)

There are many non-parametric tests but the most commonly used one is the Mann Whitney Test. The Mann Whitney test (Wilcoxon-Mann-Whitney Test) tests the part if the two samples are drawn from the same distribution irrespective of whether the distribution is normal or not. This test is the non-parametric equivalent of the 2 sample t-test.

QUIZ 10

1. What test would you choose if you had a sample size of 57 and does not have a normal distribution ?
   a) z-test
   b) t-test
   c) Mann-Whiteny test
2. If your p-value is less than the significance level, what do you do
   a) Reject the null hypothesis
   b) Reject the alternative hypothesis

Mann Whitney Test

In [6]:
```
from scipy.stats import mannwhitneyu
import numpy as np


a = [1,3,5,6,12,2,4,4]
b = [99,104,33,12,33,4,4,5]
np.random.shuffle(b)
np.corrcoef(a,b)
```

Out[6]:
```
array([[ 1.        , -0.26018152],
       [-0.26018152,  1.        ]])
```

In [7]:
```
mannwhitneyu(a, b)
```

Out[7]:
```
MannwhitneyuResult(statistic=11.0, pvalue=0.014859704830565621)
```

In [19]:
```
# T Test Code

## Import the packages
import numpy as np
from scipy import stats

## Define 2 random distributions
#Sample Size
N = 20
#Gaussian distributed data with mean = 2 and var = 1
a = np.random.randn(N) + 4
#Gaussian distributed data with with mean = 0 and var = 1
b = np.random.randn(N)
#Method 1
#For unbiased max likelihood estimate we have to divide the var by N-1,
 and therefore the parameter ddof = 1
var_a = a.var(ddof=1)
var_b = b.var(ddof=1)
#std deviation
s = np.sqrt((var_a + var_b)/2)
## Calculate the t-statistics
t = (a.mean() - b.mean())/(s*np.sqrt(2/N))
## Compare with the critical t-value
#Degrees of freedom
df = 2*N - 2
#p-value after comparison with the t
p = 1 - stats.t.cdf(t,df=df)
print("t = " + str(t))
print("p = " + str(2*p))

#we get a good p value of 0.0005 and thus we reject the null hypothesis
 and thus it proves that the

#mean of the two distributions are different and statistically significa
nt.
```

```
t = 10.14064599497438
p = 2.3125945602942O1e-12
```

In [ ]:
```
#Method 2
## Cross Checking with the internal scipy function
t2, p2 = stats.ttest_ind(a,b)
print("t = " + str(t2))
print("p = " + str(p2))
```

Hypothesis Testing

A hypothesis is an assumption.
Hypothesis testing proves or disproves the assumption with data.

Steps in Hypothesis Testing

01 — **Define the hypothesis**
The null hypothesis.

02 — **Define the decision criteria**
The significance level is the criteria using which you accept or fail to accept the null hypothes is (usually 5%).

03 — **Calculate the Hypothesis test**
The p score (0 to 1).

04 — **Make the decision**
Compare the p-score to the significance level. Decide whether to accept the null hypothesis or the alternative hypothesis.

Applications in Data Science:

1. To make claims about data distribution.
2. To compare 2 sets of results.

ANSWERS

Quiz 1

1. c

2. (a) 217,000/412,000

   (b) 198,000/412,000

   (c) 30,000/195,000

   (d) 34,000/412,000

Quiz 2

1. b

2. d

Quiz 3

1. c

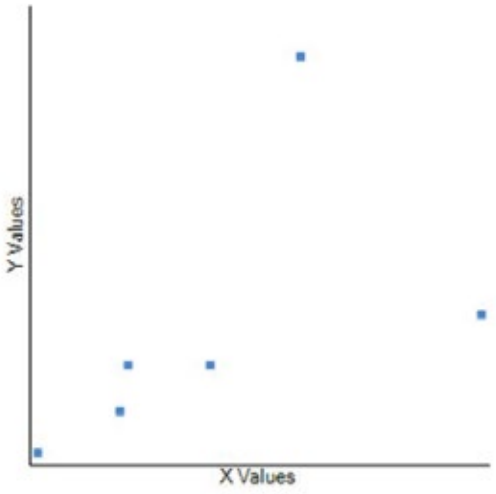
2. b

3. a

Quiz 4

1. Age and BMR show negative correlation

2.



3. 0.53

Quiz 5

1. Age and BMR show negative correlation. Weight and Age show no correlation. It also means that to find out if Weight can be used to predict BMR, there should be some causal relationship between them (and there isn't)

2. Age and Weight are positively correlated and BMR is negatively correlated

3. 67

Quiz 6

1. a

2. b

3. Mean = 18; Variance= 24

Quiz 7

1. c

2. c

3. a

4. b

Quiz 8

1. 278

2. 384.15

Quiz 9

1. 0.167

2. 3

3. 0.87

Quiz 10

1. c

2. a

You can read about data science programmes offered by the best Institutions here:

FIND YOUR PROGRAMME

ERUDITUS

EXECUTIVE EDUCATION