```
'''
What is data?
Data is a set of information. It's a collection of values that convey some information.Now, d
alotof data together is also not information.data is a set of values that gives information,a
which is of importance.


Now, data values can be some kind of quantity,like weight,no.of students,sales,proit,etc.  is
Or quality- thereare set standards to measure quality.,likeuality of drug,vaccine,etc.
or Facts like your date of birth
or statistics-like average salary of people in different states.

Types of data-
now,when we work on data,weneed to organize the data so that we can easily analyze our data.W
Such kind of data is structured data. for eg, data stored in Excel sheets, SQl table, pandas
But it'snot always possible to organize our data in tabular fromat.Such kind of data is unstr
in the same way,there might be sme other user following us,This kind of info cannot be easily
social media people,store there data in graphs and that is Unstructured data.like videos,audi

Semi-structured data-The data not completely unstructured like videos and audios,alsonot comp


Now, example of data can be-Giving link for online class, amazon products,stock markets,block

Now,how to collect data? giving reviews,through measurements,survey forms,or collected by not

Now,before analysing your data,umust clean the data,like removing null values, removing outli
After cleaning our data,wemove to analysis of data.

Now,difference between data,information,intelligence and knowledge is-
Data regarding same theme or topic gives information like  we have data of name,ages,salary.N
Now, we draw some insights basedon this information, it becomes intelligence.Now we use this
Because if we want to analyse our data,likemean medain mode,then that will give wrong results


Andrew NG,

AI is research based industry.There is no exact aanswer.

Data Analytics - data analyst: identify, collect, clean, analyze, and interpret data. needs M
(i.e. understand the problem from domain point of view by studying the data.i.e. understandin
Tools required- Excel,Advance Excel, SQL, Tableau, PowerBi, and a little bit of programming.D

Data Analyst-
Data Analyst analyzes numeric data and uses it to help companies make better decisions.

Data analysts are one of the data consumers. A data analyst answers questions about the prese
What should we do to avoid/achieve ABC? What is the trend in the past 3 years? Is our product

A data analyst’s job includes 3 main parts:
```

Understand the metrics/business problem, i.e ask the right questions.
Find out the answers or more insights from the data.
Communication. This includes creating dashboards with appropriate visualisations and explaini

Skills requirements

SQL: This is essential for all data-related roles to interact with databases.

Data visualisation: The more important thing is knowing how to visualise the data in a proper
Most companies have licensed Business Intelligence tools like Power BI, Tableau, Looker, Qlik
You don’t need to know how to use all of them. If you understand the core concepts of data an

Domain knowledge: I’d say domain knowledge is much more critical for a data analyst than othe
These kinds of domain knowledge are necessary to ask the right questions, to be able to find

Data Scientist - Do Data Diagnostic,Data Prediction
He is next to ceo.They take business decisions. Evrything Data Analyst knows + alot of progra
should of strong hold on algorithms for searching, sorting, finding also knows big data.

Data scientists are another data consumer. Instead of answering questions about the present,
they try to find patterns in the data and answer the questions about the future, i.e predicti
This technique has actually existed for a long time. You must have heard of it, it’s called s
Machine learning and deep learning are the 2 most popular ways to utilise the power of comput
Data scientists also build products based on those predictions. For instance, a recommendatio
a ranking system predicts the order of popularity, NLP predicts what a sentence means.
Data scientists build these products not to help make business decisions, but to solve busine

Skills requirements:

SQL: This is essential for all data-related roles to interact with databases.

Statistics/Mathematics: You have to master statistics knowledge such as theories behind each
This part is quite academic and theoretical, that’s why most of the data scientist roles woul

Programming skills: To apply statistics knowledge to solve real-world problems, you have to e
Training models, writing algorithms, building next-generation products are all done on a lapt
Currently, Python and R are the most popular programming language.
Software development: Just like any other engineer, software development skills are essential
Git workflow, CI/CD, DevOps, etc are all basic in a data scientist’s arsenal.

Data Engineer - Previously,data eng was referred as SQL developer.He sould know,mathematics a
There job role is to collect the data and give to data analyst or data scientist.

everything that happens to the data before reaching the database is taken care of by data eng

A data engineer cares most about:

How to ingest data from disparate sources to one single destination for analysts and scientis

Make sure the data pipeline, storage, data structure are optimised and most cost-efficient fo
Make sure the data that analysts and scientists use is the most updated, validated and accoun

Skills requirements

SQL: In addition, a data engineer should understand the ins and outs of each different databa

Sometimes need to know DBA (database administration) commands like monitoring accesses of tea
maintaining schemas to optimise database performance.

Cloud computing: As now nearly all of the data is on the cloud, from storage to database to w
technology. AWS (Amazon), Azure (Microsoft), and GCP (Google) are the 3 most popular cloud se

Software development: Same as the abovementioned.

Data Analytics->level1- MIS(Management Information System),level2->DescriptiveAnalysis,level3
level4->predictive modelling predicts what is likely to happen.This is by
ML engineer, Level5-> big data : gives ans of what can be done using this data. For which, we

----

How to see data?-
The 2 tyes of data are qualitative data and quantitative data.
1.Qualitative data is divided into 2 -Nominal and Ordinal
And qunatitative data is divided into discrete and continuous data.

So,qualitative data is the data wich cannot be countedlike, your name,cities name,like sum of
and,quantitative data is the data that can be counted. Data like sales,profit,etc.This can be

Under qualitative data we have nominal data-i.e data which doenot follow any order.Que is whi
is a Nominal data.Ordinal data has specific order.Like colors,they have significance but they

Ordinal data follow some kind of ranking or order. Like class 1,2 3 or grades.

Then comes qunatitative data is divided into discrete and continuous data.
Continuous data is like temperature,height,weight,etc.It can take any value from-ve to +ve.
Discrete data is like-cannot be subdivided,i.e. cannot have dicmalpoints,like number of emplo

Statistician-In the domain of statistics, data is costly,U r paid to collect the data, and do

ML Engineer-Here,data is cheap.U r paid to aska right ques and draw insights from data.

Now, calculating mean median variance stddev comes under Descriptive statistics.

```
But what is statistics-Statistics is To collect, Analyse,Summarize,interpret and to draw conc

Population is a universalset. Sample is subset. Statistics is study of sample.Parameter is mu
std div is sigma in parameter. S is std dev in sample

population    sample
universal set   sub set
parameter   statistics
mu -mean    Xbar
sigma -std dev    S
Nbar - population (N-1)bar
var-
  (x-xhat)^2/N    (x-xbar)^2/N-1


Descriptive statistics is how well we can describe our data by concepts like- Measure of cent

For population-
1+2+3+4+5=15=total
15/5=3=mean
(3-1)^2+(3-2)^2+(3-3)^2+(3-4)^2+(3-5)^2=4+1+0+1+4=10/5=2=var
sqrt(2)=1.414=std dev








'''


#to see continuous data,we can useline plot,histogram plots,boxplot,scatterplot or dot plot,e
#plots to see categorical data- countplot,bar plot,pie plot


#Primary data is data collected by a researcher.For eg. I gave my students task to collect da
#Then comes secondry data-which is working on primary data and finding inferences out of it.
#If there is data then only data can be analysed.
#Like outof 20 classmates- 8 likes playing games in mobile.3 likes to read book,4like towatch


'''
#Correlation between 2 variables given by correlation coefficient.- gives relationship betwee
# for eg. there is a variable x->y
#change in x changes y.Here x is an independent variable and y is a dependent variable.i.e. y
#For eg.
#Temp and icecream sale
 12         100
 15         110
 20         200
 23         230
 30         400
 #i.e. with the incrase in temp, ice cream sales increase.
 So, the type of correlation are-
 1. perfect +ve corr
 2.weak +ve corr
 3.perfect -ve corr
 4.weak -ve corr
 5. No corr

 Now, corr is always between -1,0,1

 -1 = negative corr or near to -1
 0=no corr
 1= positive corr or near to 1

 We can use scatter plot to see corr.When data is scattered all over, then there is no corr.


 '''
```

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

```
x=[22,15,11,8,5,-2] #temp
y=[2500,1500,800,300,200,90]#ice-cream sales
plt.figure(figsize=(6,6))
sns.lineplot(x,y)
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass tl
  FutureWarning
```

```
# formula for corr func is-
#corr between x,y i.e r= summation of(x-xmean)(y-ymean)/root(summation of(x-xmean)^2*summatio
```

```
meanx=sum(x)/len(x)
meany=sum(y)/len(y)
print("mean x",meanx)
print("mean y",meany)
```

```
mean x 9.833333333333334
mean y 898.3333333333334
```

```
xdiff=[(i-meanx) for i in x]
ydiff=[(i-meany) for i in y]
```

```
xdiffsq=[(i-meanx)**2 for i in x]
ydiffsq=[(i-meany)**2 for i in y]
```

```
import math
mulxydiff=[xdiff[i]*ydiff[i] for i in range(len(xdiff))]
```

```
sum(mulxydiff)
```

```
36518.33333333333
```

```
mulxydiffsq=[xdiffsq[i]*ydiffsq[i] for i in range(len(xdiffsq))]
```

```
sum(mulxydiffsq)
```

```
493507954.1666666
```

```
corr=sum(mulxydiff)/math.sqrt(sum(mulxydiffsq))
corr
```

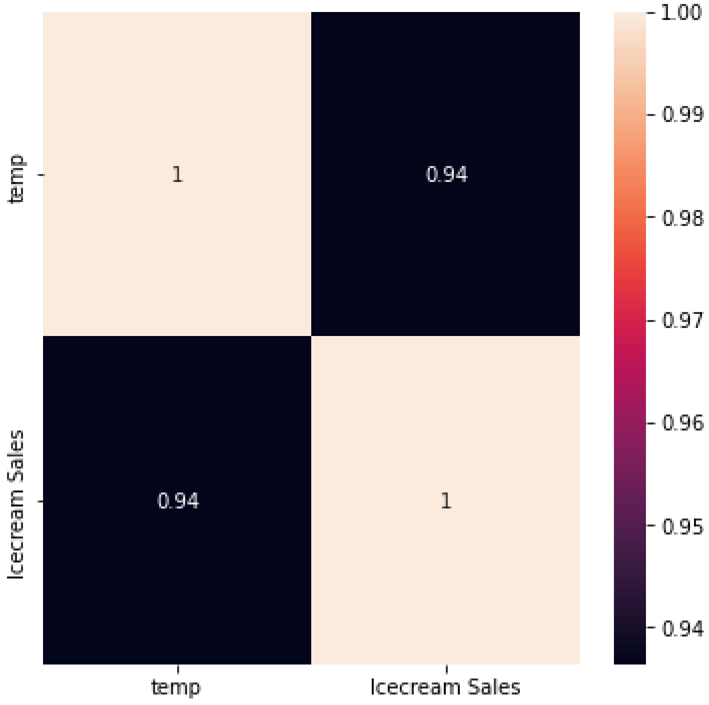
```
1.6438563743556045
```

```
data={"temp":[22,15,11,8,5,-2],"Icecream Sales":[2500,1500,800,300,200,90]}
df=pd.DataFrame(data,columns=["temp","Icecream Sales"])
df
```

	temp	Icecream Sales
0	22	2500
1	15	1500
2	11	800
3	8	300

```
corr=df.corr()
```

```
plt.figure(figsize=(6,6))
sns.heatmap(corr,annot=True)
plt.show()
```



```
#DWD_Probability
'''
1. Introduction to Probability
2.Set theory
3.Dependent and independent probability
4.Conditional Probability
5.Bayes Theorem
6.Random VAriable
7.Probability distribution.
'''
```

```
# Introduction to prob-Probability is simply how likely something is to happen. Whenever we'r
#we can talk about the probabilities of certain outcomes—how likely they are. The analysis of
'''
```

```
for eg, i have 1 coin , it has 2 sides- H and Tails. The prob of getting h is 1/2 and prob of
chance of getting T.
lly, prob og geeting 3 in dice=1/6 =0.16% chance of getting any side.
So formula for prob=no. of ways it can happen/total no. of outcomes'''


'''
prob line is between 0----0.5-----1
0 means impossible, 0.5 means even chance,1 means certain
0----0.5 is unlikely
0.5----1 is likely
prob of getting even no. from dice=3/6=0.5 and odd=3/6=0.5
prob is just the guideline to find facts from the event.
experiment=repeateble procedure like tossing coin
sample space=possible outcome toss=(head,tail)
event=outcome based on experiment for head (1/2) and tail=(1/2)
'''



# set theory
'''
In probability sample set is set of data,now to understand sets, we need to dive into set the
What is set? Set is a unique collection of objects(character,alphabets,symbols,numbers)
eg, numset={all the numbers}
even set={all even numbers}
odd set={all odd numbers}

if a={1,2,3,4}
b={3,4,5,6}
then,aUb={1,2,3,4,5,6}
a intersection b={3,4}
now, how set theory relates to prob?=
experiment={rolling the dice}
s={1,2,3,4,5,6}
chance of getting no. divisible by 2={2,4,6}=3/6
chance of getting prime nos.={2,3,5}=3/6
prob of getting even >3={4,6}=2/6
E1 intersection E2=2/6

Proof of prob-

toss coin getting H=1/2
toss coin getting T=1/2
sum of H and T=0.5+0.5=1


'''


'''
Dependent and Independent Prob or Conditional Probability-
```

When 2 events outcome doesnot change its prob values, its called independent prob
Tossing a coion getting H or T are the eg of Independent prob

Note: i.e. events are disconnected from each other.

Dependent prob-when the outcome of the first event influences the outcome of the second event
Q. prob of getting 3 aces in a row-
P(getting 3 aces in a row)=4/52*3/51*2/50=0.0181%-------This is called Conditional Prob.Depen

Baye's Theorem-

When two events, A and B are dependent, the probability of occurrence of A and B is:

P(A and B) = P(A) · P(B|A), or,
The probability of simultaneous happening of two events A and B is equal to the probability o
the conditional probability of B with respect to A.
P(A∩B) = P(B).P(A/B)
The probability of simultaneous happening of two events A and B is equal to the probability o
the conditional probability of A with respect to B.

Q.Shareen has to select two students from a class of 23 girls and 25 boys. What is the probab

Solution: Total number of students = 23 + 25 = 48

Probability of choosing the first boy, say Boy 1 = 25/48

Probability of choosing the second boy, say Boy 2 = 24/47

Now,

P(Boy 1 and Boy 2) = P(Boy 1) and P(Boy 2|Boy 1)

= (25/48) × (24/47)

= 600/2256

Two cards are drawn one by one from a pack of 52 cards without replacement. What is the proba
second is queen?

Solution:

Let A be the event of drawing a king and B be the event of drawing a queen. Since, the first

Total number of balls = 52

Number of kings = 4

Therefore,

Probability of drawing a king, P(A) = 4/52

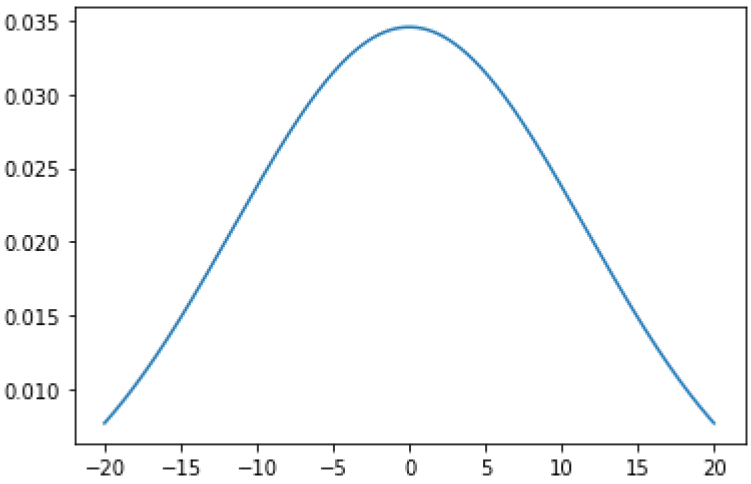
```
The number of cards in the deck now is 52 - 1 = 51

Number of queen = 4

A queen is drawn given that a king is drawn. Therefore, conditional probability of B given th

P(B/A) = 4/51

Now, the probability that events A and B occur simultaneously is given by,

P(A∩B) = P(A).P(B/A)

Substituting the respective values,

P(A∩B) = 4/52
 × 4/51
 = 4/663

Therefore, the probability that the first card drawn is a king and second is queen is 4/663.
'''
```

```
'''
Random Variable
In probability, a random variable is a real valued function whose domain is the sample space
For example, let us consider an experiment for tossing a coin two times.

Hence, the sample space for this experiment is S = {HH, HT, TH, TT}

If X is a random variable and it denotes the number of heads obtained, then the values are re

X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.

Similarly, we can define the number of tails obtained using another variable, say Y.

(i.e) Y(HH) = 0, Y(HT) = 1, Y(TH) = 1, Y(TT)= 2.
'''
```

```
# Probability Distribution-
'''
What is binomial distribution?which kind of graph it produces?Some eg.
What is Random variable?
What is Probability Distribution?-Outcomes of an experiment when we plot in a bar graph,etc.
```

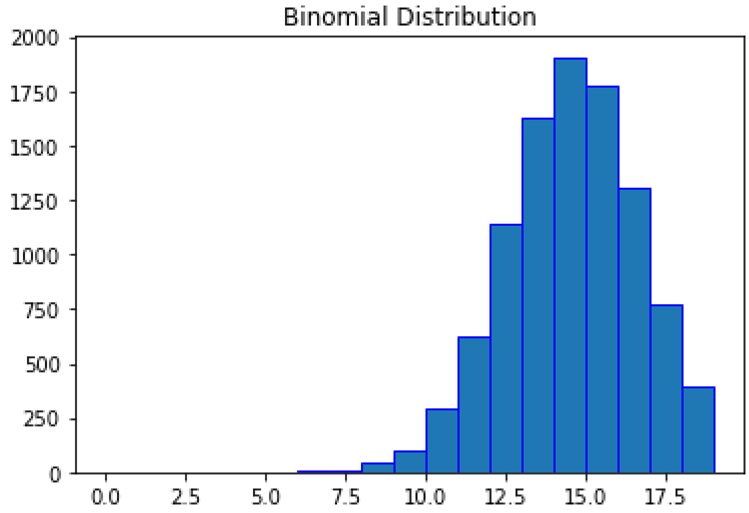
```
'''
Homework
Random Variable and Distribution (ProbabilityDistribution)
Normal Distribution
```

```
Binomial Distribution
Cumulative Distribution
Uniform Distribution
Multinomial Distribution
Continuous Random Variable Distribution


'''



'''
A probability distribution is a function under probability theory and statistics- one that gi
are in an experiment.

It describes events in terms of their probabilities; this is out of all possible outcomes.
Let’s take the probability distribution of a fair coin toss.
Here, heads take a value of X=0.5 and tails gets X=0.5 too.
Two classes of such a distribution are discrete and continuous.
The former represented by a probability mass function and the latter by a probability density

'''


# Normal Distribution

import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
import statistics

# Plot between -10 and 10 with .001 steps.
x_axis = np.arange(-20, 20, 0.01)

# Calculating mean and standard deviation
mean = statistics.mean(x_axis)
sd = statistics.stdev(x_axis)

plt.plot(x_axis, norm.pdf(x_axis, mean, sd))
plt.show()

'''
The empirical rule in statistics allows researchers to determine the proportion of values tha
The empirical rule is often referred to as the three-sigma rule or the 68-95-99.7 rule.

The Empirical Rule(also called the 68-95-99.7 Rule or the Three Sigma Rule) states that for a

68% of the observed values lie between 1 standard deviation around the mean : (mu-sigma) to (
95% of the observed values lie between 2 standard deviations around the mean : (mu-2*sigma) t
99.7% of the observed values lie between 3 standard deviation around the mean : (mu-3*sigma)
'''
```



```
'''
Binomial Distribution in Python
Python binomial distribution tells us the probability of how often there will be a success in
Such experiments are yes-no questions.

There must be only 2 possible outcomes.
Each outcome has a fixed probability of occurring. A success has the probability of p, and a
Each trial is completely independent of all others.
The binomial random variable represents the number of successes(r) in n successive independen

Probability of achieving r success and n-r failure is :

n*p^r * (1-p)^{n-r}

Example 1: If a coin is tossed 5 times, find the probability of:

(a) Exactly 2 heads

Number of trials: n=5

Probability of head: p= 1/2 and hence the probability of tail, q =1/2

For exactly two heads:

x=2

P(x=2) = 5C2 p2 q5-2 = 5! / 2! 3! × (½)2× (½)3

P(x=2) = 5/16

Calculating distribution table :

Approach :

Define n and p. n= Required. Specify number of trials, must be >= 0. Floats are also accepted
p=Required. Specify probability of success in each trial, must be in range [0, 1]. float or a
Define a list of values of r from 0 to n.
```

```
Get mean and variance.
For each r, calculate the pmf and store in a list.

'''


import matplotlib.pyplot as plt
import numpy as np

#fixing the seed for reproducibility
#of the result
np.random.seed(10)

size = 10000
#drawing 10000 sample from
#binomial distribution
sample = np.random.binomial(20, 0.7, size)
bin = np.arange(0,20,1)

plt.hist(sample, bins=bin, edgecolor='blue')
plt.title("Binomial Distribution")
plt.show()
```



```
'''
Poisson Distribution in Python
Python Poisson distribution tells us about how probable it is that a certain number of events
This assumes that these events happen at a constant rate and also independent of the last eve

For example, a Poisson distribution could be used to explain or predict:

Text messages per hour
Machine malfunctions per year
Website visitors per month
Influenza cases per year

'''
```

```
'''
In statistics, uniform distribution refers to a type of probability distribution in which all
A deck of cards has within it uniform distributions because the likelihood of drawing a heart
A coin also has a uniform distribution because the probability of getting either heads or tai
'''
```

```
'''
Inferential statistics use laws of probability to make inferences about a population based on
'''
```

```
#Inferential Statistics
'''
Why we need Inferential Statistics?
Pre-requisites
Sampling Distribution and Central Limit Theorem
Hypothesis Testing
Types of Error in Hypothesis Testing
T-tests
Different types of t-test
ANOVA
Chi-Square




1. Why do we need Inferential Statistics?
Suppose, you want to know the average salary of Data Science professionals in India. Which of

Meet every Data Science professional in India. Note down their salaries and then calculate th
Or hand pick a number of professionals in a city like Gurgaon. Note down their salaries and u
Well, the first method is not impossible but it would require an enormous amount of resources
But today, companies want to make decisions swiftly and in a cost-effective way, so the first

On the other hand, second method seems feasible. But, there is a caveat.
 What if the population of Gurgaon is not reflective of the entire population of India?
 There are then good chances of you making a very wrong estimate of the salary of Indian Data

Now, what method can be used to estimate the average salary of all data scientists across Ind

Enter Inferential Statistics
In simple language, Inferential Statistics is used to draw inferences beyond the immediate da

With the help of inferential statistics, we can answer the following questions:

Making inferences about the population from the sample.
Concluding whether a sample is significantly different from the population. For example,
let’s say you collected the salary details of Data Science professionals in Bangalore.
And you observed that the average salary of Bangalore’s data scientists is more than the aver
Now, we can conclude if the difference is statistically significant.
```

```
If adding or removing a feature from a model will really help to improve the model.
If one model is significantly better than the other?


Now, u should must know-

Statistic – A Single measure of some attribute of a sample. For eg: Mean/Median/Mode of a sam
Population Statistic – The statistic of the entire population in context. For eg: Population
Sample Statistic – The statistic of a group taken from a population. For eg: Mean of salaries
Standard Deviation – It is the amount of variation in the population data. It is given by σ.
Standard Error – It is the amount of variation in the sample data. It is related to Standard

'''

'''
Sampling------
What is sampling?
Imgine there is a population of shark-10,000
whose avg weight is 250kg

This is called as a population avg

Now,if I take a random sample from population and size of sample is 50 and we get avg of thos

So population mean=250
and sample mean=280
Now, we took onemore sample,and we calculate one more sample of 50 size and we get avg of tho

So,this way if we take samples and its avg weight, then we must get the data distribution nor
Then we can conclude that the avg weight of shark is 250 to 275.

i.e in Inferential Statistics , from sample dataset wetry to infer about the population.

1.e. based on sample we make some assumption on population.

Vaccination drive staarted on sample,and after getting valid results, population vaccination

'''


'''
Sampling Distributions-Sampling Distribution is the graph obtained by plotting sample means.

Sampling Distribution helps to estimate the population statistic.
But how ?

This will be explained using a very important theorem in statistics – The Central Limit Theor

Central Limit Theorem
It states that when plotting a sampling distribution of means, the mean of sample means will
```

And the sampling distribution is equal to normal distribution with variance equal to σ/√n wh
σ is the std dev of population and n is the sample size.

1.The shape of the Sampling Distribution will remain the same (remember the normal curve- bel
2.The number of samples have to be sufficient (generally more than 50) to satisfactorily achi
Also, care has to be taken to keep the sample size fixed since any change in sample size will
it will no longer be bell shaped.
3.As we increase the sample size, the sampling distribution squeezes from both sides giving u
it lies somewhere in the middle of the sampling distribution (generally).

Now,it is important to know where the population mean lies with respect to a particular sampl
This brings us to our next topic – Confidence Interval.

Confidence Interval
The confidence interval is a type of interval estimate from the sampling distribution which g
population statistic may lie

Foreg.wewant toestimate mean weight of certain species of turtlein florida.
Now,there are more than 1000,so it will be time consuming togoand checkweight of individual t
of 50turtles,and calc sample mean.Now,here itsnot guaranteed that the mean ofthis sample will
So 1 way can be,takking very heavy weight turtle and least weight turtle and then takeout mea

Now,in order to capture uncertaininty here we can create confidence interval in a range of v
It can be calculated as-
CI=point estimation+-(critical value)*standard error.

This formula creates interval between lower bound and upper bound
so that,its likely to contain population parameter with certain level of confidence.

confidence inteval=[lower bound,upper bound]

Here steps are-
1. Take randomsamples with extreme values and apply the formula,
where criticalvalue is z value, and std error is std deviation.

So our final formula is like-
CI=sample mean+-z*(sampleStddev/sqrt(sample size))

Now, this z value we will use will depend on CI we will choose.
Now,tis zvalue is not constant becuse we are only forecasting . It is not necessarythat forec

eg
sample size=25
sample mean=300
sampe sd=13.5

Now,to calc 90% confidence inerval (this is decide by researcher)for true population mean wei

The researchers would then utilize the following table to determine their Z value:

Confidence Interval	Z value
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

90%CI = 300 +- 1.645*(13.5/sqrt(25)) ==[293.91,306.09]

i.e.there is a 90% chance that ,CI of [293.91,306.09] contains population mean weight of tur

Hypothesis Estimation-

ConfidenceIntervalis estimation of population fromthe sample data.
but there is noproof that population will be under given interval only.
So we need to provewith scientific method.Which is clled Hypothesis.

eg.

Class 8th has a mean score of 40 marks out of 100.
The principal of the school decided that extra classes are necessary in order to improve the
The class scored an average of 45 marks out of 100 after taking extra classes.
Can we be sure whether the increase in marks is a result of extra classes or is it just rando

Hypothesis testing lets us identify that.
It lets a sample statistic to be checked against a population statistic or statistic of anoth
Extra classes being the intervention in the above example.

Hypothesis testing is defined in two terms – Null Hypothesis and Alternate Hypothesis.

Null Hypothesis being the sample statistic to be equal to the population statistic.
For eg: The Null Hypothesis for the above example would be that the average marks after extra
Alternate Hypothesis for this example would be that the marks after extra class are significa
Hypothesis Testing is done on different levels of confidence and makes use of z-score to calc
So for a 95% Confidence Interval, anything above the z-threshold for 95% would reject the nul

Points to be noted:

We cannot accept the Null hypothesis, only reject it or fail to reject it.
As a practical tip, Null hypothesis is generally kept which we want to disprove.
For eg: You want to prove that students performed better after taking extra classes on their
The Null Hypothesis, in this case, would be that the marks obtained after the classes are sam

Now, for hypothesis,
first state the hypo
2 Determine the significance level to use for hypo

```
3 find test statistics
4 accept or reject uing p value
5 interpret the result from process above

Significance level-is the prob that the event could haveoccured by chance.

Now,after this 2 types of decision errors ae possible-
In statistics, a Type I error is a false positive conclusion, while a Type II error is a fals


Null hypo is  True        False
true          accept      type1 error-reject a true null hypo
false         type 2      accept
              accept a
              false null
              hypo


You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could
Type I error (false positive): the test result says you have coronavirus, but you actually do
Type II error (false negative): the test result says you don’t have coronavirus, but you actu


------------
z-test is a statistical method for the comparison of mean in a sample from the normally distr

z-test is used when:

Population variance is unknown
Sample size is greater than 30

2 types of z-test are:
1 tailed z-test: The region of rejection is located either extreme left or extreme right of t
2 tailed z-test : A two-sample test is used when we have to compare the mean of two samples.
The region of rejection is located on both the extreme (left and right) of the distribution


There are 3 steps in Hypothesis Testing:

1 State Null and Alternate Hypothesis
2 Perform Statistical Test
3 Accept or reject the Null Hypothesis

Z-scores are measured in standard deviation units.

For example, a Z-score of 1.2 shows that your observed value is 1.2 standard deviations from
A Z-score of 2.5 means your observed value is 2.5 standard deviations from the mean and so on

The closer your Z-score is to zero, the closer your value is to the mean. The further away yo
the further away your value is from the mean. Typically, you will not see Z-scores that are m
This is because most data points lie within 3 standard deviations of the mean.
```

x. The value for which you want to calculate the Z-score. We sometimes call this the raw scor

$\boldsymbol{\mu}$. The population mean.

σ. The population standard deviation.

Z-Score= $x-\mu/\sigma$

eg.Your niece has just been born weighing 6.9 pounds, and you want to know if this is a norma

Niece’s birth weight (x): 6.9 pounds or 3130 grams
Mean birth weight (**μ**): 7.224 pounds or 3276 grams
Standard deviation (**σ**): 1.25 pounds or 567 grams

Z-score=(3130 - 3276)/567 = -146/567 = -0.26

Her birth weight is 0.26 standard deviations below the mean.

A z-statistic, or z-score, is a number representing the result from the z-test

1 tailed Z-test: The formula is:

Z-Score= $x-\mu/(\sigma/\text{sqrt}(n))$

Q is A gym trainer claimed that all the new boys in the gym are above average weight.

A random sample of thirty boys weight have a mean score of 112.5 kg and the population mean w

Is there a sufficient evidence to support the claim of gym trainer.

H0= Equal to avg weight i.e.μ= 100 kg
H1=$\mu>100$

Z-Score= $x-\mu/(\sigma/\text{sqrt}(n))$= 4.56

Now, for accepting or rejecting the null hypothesis, we have significance value=0.05, for whi

Now, z-score>significance value, we need to reject null hypo.

Z-test is a statistically significant test for the hypothesis testing (null and alternative h
when the sample size is large, and the population parameter (mean and variance) is known.

T-tests
T-tests are similar to the z-scores, the only difference being that instead of the Population
The rest is same as before, calculating probabilities on basis of t-values.

A t-test is a statistical test that is used to compare the means of two groups.

You want to know whether the mean petal length of iris flowers differs according to their spe
You find two different species of irises growing in a garden and measure 25 petals of each sp
You can test the difference between these two groups using a t-test and null and alterative h

The null hypothesis (H0) is that the true difference between these group means is zero.
The alternate hypothesis (Ha) is that the true difference is different from zero.

The different type of t-tests are- One-sample, two-sample, or paired t-test

If the groups come from a single population (e.g. measuring before and after an experimental
If the groups come from two different populations (e.g. two different species, or people from
perform a two-sample t-test (a.k.a. independent t-test).
If there is one group being compared against a standard value (e.g. comparing the acidity of
perform a one-sample t-test.

One-tailed or two-tailed t-test?
If you only care whether the two populations are different from one another, perform a two-ta
If you want to know whether one population mean is greater than or less than the other, perfo

In your test of whether petal length differs by species:

Your observations come from two separate populations (separate species), so you perform a two
You don’t care about the direction of the difference, only whether there is a difference, so

t-test for 1 sample-
t-test=(meansample - mean population)/(samplesd/sqrt(samplesize))

p-value=degree of freedom

whatever sample size we will take,it will be always -1. i.e if we have 40 sample then it's 40

degree of freedom value we can choose 0.10,0.05 and 0.01

for null hypothesis-

eg.we have taken turtle sample size of 40,310 pound is the weight of turtle in florida .Thisi

step 1
samplesize=40
sample value=300
sample mean=310
samplestd dev=18.5

step 2

H0:population mean=310
H1: population not equalto 310

```
calc t-test:
t=300-310/18.5/sqrt(40)
t=-3.4187

step-4:calc the p-value

degree of freedom=40-1=39

p-value=0.000745

note: p value is always between 0-1

p-value<0.05 then, reject nullhypothesis.

i.e.as per sample mean,300 pound is notthe weight of the population.

-----------------------
'''

data=[14,14,16,13,12,17,15,14,15,13,15,14]#plant height


import scipy.stats as stats
stats.ttest_1samp(a=data,popmean=15)
```

```
     Ttest_1sampResult(statistic=-1.6848470783484626, pvalue=0.12014460742498101)
```

p_value>0.05 Accept nullhypo

```
'''When should we use t-test than z-test?

If the population standard deviation is known and the sample size is greater than 30, Z-test
If the population standard deviation is known, and the size of the sample is less than or equu
T-test is recommended. If the population standard deviation is unknown, T-test is recommended
'''

'''
t-test and z-test are for continuous data.

'''



'''
One-Way ANOVA in Python: One-way ANOVA (also known as “analysis of variance”) is a test that
significant difference between the mean values of more than one group.

Now, the question arises – Why do we need another test for checking the difference of means b
Why can we not use multiple t-tests to check for the difference in means?

The answer is simple. Multiple t-tests will have a compound effect on the error rate of the r
```

```
Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA kee

ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Squa

Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between g

Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)


Steps to perform ANOVA
1 Hypothesis Generation
2 Null Hypothesis : Means of all the groups are same
  Alternate Hypothesis : Mean of at least one group is different
3 Calculate within group and between groups variability
4 Calculate F-Ratio
5 Calculate probability using F-table
6 Reject/fail to Reject Null Hypothesis
There are various other forms of ANOVA too like Two-way ANOVA, MANOVA, ANCOVA etc. but One-Wa

Hypothesis involved:
A one-way ANOVA has the below given null and alternative hypotheses:

H0 (null hypothesis): μ1 = μ2 = μ3 = … = μk (It implies that the means of all the population
H1 (null hypothesis): It states that there will be at least one population mean that differs
Statement:

Researchers took 20 cars of the same to take part in a study. These cars are randomly doped w
 allowed to run freely for 100 kilometers each. At the end of the journey, the performance of

'''

'''

Python provides us f_oneway() function from SciPy library using which we can conduct the One-
'''

# Importing library
from scipy.stats import f_oneway

# Performance when each of the engine
# oil is applied
performance1 = [89, 89, 88, 78, 79]
performance2 = [93, 92, 94, 89, 88]
performance3 = [89, 88, 89, 93, 90]
performance4 = [81, 78, 81, 92, 82]

# Conduct the one-way ANOVA
f_oneway(performance1, performance2, performance3, performance4)
```

```
     F_onewayResult(statistic=4.625000000000002, pvalue=0.016336459839780215)
```

```
'''
Analyse the result:

The F statistic and p-value turn out to be equal to 4.625 and 0.016336498 respectively. Since
 we would reject the null hypothesis.
This implies that we have sufficient proof to say that there exists a difference in the perfo

'''
```

```
'''
chi2 hypothesis testing-it is useful for categorical data when we want tocheck two category d
eg.voting vs color of person in US

It is also known as test of independence.

H0:nullhypothesis
H1:alternate hypothesis.

x^2=E(observation-estimation)^2/total_estimation

E=summation
          republic      democrat    independent    total
male        120            90          40            250
female      110            95          45            250
total       230           185          85            500


H0: there is link between gender and politicalparty preference

'''
```

```
data=[[120,90,40],
      [110,95,45]]
```

```
import scipy.stats as stats
stats.chi2_contingency(data)
```

```
(0.8640353908896108, 0.6491978887380976, 2, array([[115. ,  92.5,  42.5],
        [115. ,  92.5,  42.5]]))
```

```
x=stats.chi2_contingency(data)
print(x[0])
print(x[1])
```

```
0.8640353908896108
0.6491978887380976
```

```
'''
p-value>0.05-we fail to reject null hypothesis
i.e. there is an associaton between gender and political party.

'''
```

```
'''


Sampling as we feed data in to machine learning model,

Data Sampling forms the essential part of the majority of research, scientific and data exper
There are many sampling techniques that can be used to gather a data sample depending upon th

random

Random Sampling
The simplest data sampling technique that creates a random sample from the original populatio
In this approach, every sampled observation has the same probability of getting selected duri
Random Sampling is usually used when we don’t have any kind of prior information about the ta

For example random selection of 3 individuals from a population of 10 individuals.
Here, each individual has an equal chance of getting selected to the sample with a probabilit


systematic
Systematic sampling is defined as a probability sampling approach where the elements from a t
and after a fixed sampling interval.
We calculate the sampling interval by dividing the entire population size by the desired samp


cluster
Cluster sampling is a probability sampling method in which you divide a population into clust
such as districts or schools, and then randomly select some of these clusters as your sample.

stratified
Stratified Sampling is a data sampling approach, where we divide a population into homogeneou
(e.g., age, race, gender identity, location, event type etc.).

Every member of the population studied should be in exactly one stratum.
 Each stratum is then sampled using Cluster Sampling, allowing data scientists to estimate st
 We rely on Stratified Sampling when the populations’ characteristics are diverse and we want
 characteristic is properly represented in the sample.



'''
```

```
#https://www.analyticsvidhya.com/blog/2017/01/comprehensive-practical-guide-inferential-stati
```

```
l=list()
l.append([1,2,[3,4]])
l.extend([5,6,7])
l
```

```
[[1, 2, [3, 4]], 5, 6, 7]
```

```
from random import randint
#import random
for i in range(5):
  print(random.randint(1,5))
```

```
3
5
1
3
5
```

Colab paid products - Cancel contracts here