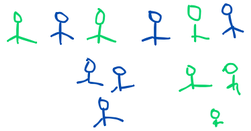
Day01stats

statistics:- is a branch of mathematics where we can,collect,organize, visualise , analyse the data for better decision making and future predictions
There 2 types of stats
1.Descriptive stats:- gives the summary of data
2.inferential stats:- is a process of data analysis where we can make the conclusion report about your data

#what is a population?
#population is a overall data that you want draw conclusion
#what is sample?
#sample is a part of population
#what is data?
#data is facts or piece of information that can be measured

Types of sampling techniques:-
1.simple random sampling:- is process of sampling where every member has equal chance to get selected

2.stratified sampling:- is process where population splits in to non overlapping groups.



3.systametic sampling:- is a probability sampling methods where researchers select population of nth intervel

4. **convinence sampling:**- is the process of taking the sample data from those who has knowledge on the reasearch data

#Variable:- are containers where we can store the data and reuse it

the variable are 2 types according to stats

1. Quantitative variable:- numerical data--> 2types --

a. continuous data:- a numeric value that has infinite number of values
ex:- weight of students in a class room
height of students in a class room

b. discrete data:- a fixed whole number
ex:- no of childrens in a family ---> 2,3,4,
Total population in a city

No of students in a class--> discrete data
House rent prices in a area---> continous
No of houses in a area-->discrete
count of sugar--->cont

1. Qualitative variable:- is categorical data based on some characterstics we derive some values
ex:- dog breeds, eye color, gender,level of education,marital status

#what are all variable measurment sclaes do we have?
#there are 4 types
#1.Nominal data:- categorical data and no order
#2.ordinal data:- order matters
#3.intervel:- order matters also values also matter
#4.ratio:- 1:2.3:2

revision:

np.where(condition,true,fals)
q)compute the indices of an array where the condition is true-->np.argwhere(condition)
pandas fillna-->
df['col'].fillna(methods='ffil',how,any)

Day02

Descriptive stats:-
1. The central limit theorem or central tendency theorem or central measure
2. Spread metrics

Q)what is the most common data point should i have to consider from my data set?
A)The central limit theorem or central tendency theorem
1.mean:- sum of observations / No of observation
l = [1,2,3,4,5,6,7,8,9]-->5
2.median:- The middel data point is know as median
l2 = [1,2,3,4,5,6]
  = 3+4/2---> 3.5
#Note:- when we apply median we need to sort it in a ascending order

3.mode :- most repeated elements in the data

#when we need to use mean,median and mode?
A) l = [1,2,3,4,5,6,7,8,9,100]
mean = 145/10--->14.5
median= 5+6/2---> 5.5
#Note:- when there is outliers in the data we use median
#when there no outliers we use mean
animal = ['dog', 'cow','cat','cat','elephant',]
#Note for the above we wont use the -- mean and median bcz it is a categorical data
Note:- for categorical data we use mode

2. spread metrics:- will give the distribution of the from center point
1.range:- (max - min)
2.IQR--> Inter quartile range
-- It defines 75th percentile of your data - 25th percentile of your data

#what is percentile and how can we calculate it?
-- percentile is the value below which certain percentage of values or observations exist

dataset = [2,2,3,4,5,6,6,7,8,8,8,8,9,9,10,11,11,12]
Q)what is the percentile of range of 10?

$$\text{percentile rank of x} = \frac{\text{No of values before x}}{n} \rightarrow 100$$

$$= \frac{16}{26} \times 100$$

$$= 80.1$$

Q)what is the value exist at 25th percentile?

$$\text{value} = \frac{\text{pernctile}}{100} * n+1 \qquad = \frac{25}{100} * 20+1$$

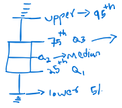
$$= 5.25 \rightarrow \text{index}$$

Q)what is the value exist at 75th percentile?

$$\frac{75}{100} \times 21 \longrightarrow 16 \longrightarrow 10$$

$$IQR = 10 - 5 = 5$$

$$10 + 1.5(5) \rightarrow 17.5$$



$$Q_3 + 1.5\,IQR \rightarrow UL$$

$$Q_1 - 1.5\,IQR \rightarrow LL$$

$$5 - 1.5(5) \rightarrow -2.5$$

3.standard deviation:- Shows the how far the elements are from the mean

$$\sigma = \sqrt{\text{variance}}$$

$$\text{varience} = \text{mean}(x^2) - (\text{Mean}(x))^2$$

$$x = [1,2,3,4,5]$$

$$x^2 = [1,4,9,16,25]$$

$$(M(x))^2 = (3)^2 = 9$$

$$M(x^2) = 11$$

$$\sigma = \sqrt{11-9} = \sqrt{2} \simeq 1.414$$

$$= \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

$$= \frac{10}{8} = \sqrt{2} = 1.414$$

q)what is the SD OF below data?
5, 5,9,9,9,10,5,10,10---->2.29

$$\bar{x} = 8 \qquad \begin{matrix} 9 \\ 9 \\ 1 \\ 1 \end{matrix} \quad = \frac{42}{9} = 4.67$$

#which spread metrics is more sensitive to the the outliers?
a = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,40,50]

#cumulative frequency:-
#[rose,lilly,sunflower,rose,lilly,sunflower, rose,lilly,lilly]

Flower	Freq	CF
R	3	3
L	4	7
S	2	9

Day03-Stats
Probability:- A particular event happens or not is known as probability
Q)How much percentage you are confident to clear the exam?
Q)Today is sunday what is probability that tomorrow is monday?
Q)what is the probability that tomorrow is going to rain?

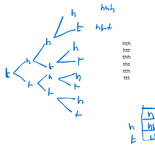
#what is sample space?
#all the possible outcomes
week = (sun,mon,tues,wed,thu,fri,sat)
ss = (1,2,3,4,5,6)
dec of cards = 52 => 13cards each
toss a coin ss--> (H,T)
#What is the ss when we toss 2 coins?
#ss(H,H)(H,T)
#toss 3 coins How many sample spaces ?--> 2n--> 2**3-->8**
#what is probability of getting tail when you toss a single coin--->
ss = (H,T)
N = 2
fav = 1

p(t) = $\dfrac{\text{No of fav outcomes}}{\text{Total No of possibilities}} = \frac{1}{2}$

q)what is probability of 6 will appear on dice -->1/6
Q)when you toss 2 coins what is probability of getting (HH)---> 1/4
q)what is probability of appearing even numbers on dice
ss = (1,2,3,4,5,6)
fav = (2,4,6)
= 3/6 => 1/2

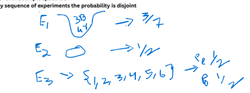
a)There are 6 pillows in a bed

#what is the probability of picking the yellow pillow---> 1/3

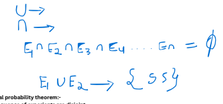


	h	t
h	hh	ht
t	th	tt

Axioms of probability:-
#Prob can be lays b/w 0 and 1
#prob of entire sample space is 1
#for any sequence of experiments the probability is disjoint



Q)what is the total probability of p(e) and p(o)---> p(e) + p(o)---> 1



#Law of total probability theorem:-
when the sequence of experients are disjoint
if you combine all the events it gives over all sample spaces

E1--> getting the even number when you roll a dice ---> (2,4,6)
E2-->probability of getting 5 on dice--> (5)
E3--> probability of getting 1 or 3 ---> (1,3)
e1,e2,e3 is known as disjoint event/exhustive event

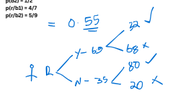
#what is the probability of A in above 3 events?

P(A) = $P_{E_1} \times P_{A/E_1} + P_{E_2} \times P_{A/E_2} + P_{E_3} \times P_{A/E_3}$

a)

#what is the probability of getting the red ball from 2 bags?

p(b1) = 1/2
p(b2) = 1/2
p(r/b1) = 4/7
p(r/b2) = 5/9



Q)find the probability that the person can able to complete the work?

= 0.65*0.32 + 0.35*0.80 => 0.488

Q)Find the total probability that the person cannot able to finish the work?

#2.Conditional probability theorem:- it is opposite to the total probability theorem

#what is the probability that the red ball is choosen--> Total probability theorem
#Given that red ball is choosen find the probability that is from bag1?
- this can be solve by using the conditional probability theorem or the bayes theorem

p(r/b1) = $\dfrac{p(e).p(r/e1)}{p(e1).p(r/e1) + p(e2).p(r/e2)} = \dfrac{\frac{1}{2} \times \frac{3}{5}}{\left(\frac{1}{2} \times \frac{3}{5}\right) + \left(\frac{1}{2} \times \frac{4}{7}\right)}$

$= \dfrac{0.3}{0.3 + 0.28}$

$= 0.517$

#Bernoulli's Trail:- it is an experiement with 2 outcomes
success (1) or failure (0)
when we need to apply the bernoulis trail?
1.The outcomes should be two
ex:- yes or n0, TRUE OR FALSE ...etc
2. all the trails should be independent on each other
3.The probability of all the tarils should remain same
4.The trails should be finite.

Day 04:-

#Binomial distribution:- if a event follows bernoulis trail then the probability can be determined by using the binomial distribution.

ex:- toss a coin

#how many times i need to repeat the experiment to get the 1st tail?
success ---> p(t) = 1/2
failure --q = 1-p --1/2

#tomorrow rains = 70
p(r)---> 0.7
q(not rain) = 1-p
= 0.3

dice = s {1,2,3,4,5,6}
find the probability of getting the numbers greater then 4
{1,2,3,4} {5,6}

toss a coin--->{H,T}
2-->trails independent
3-->probability remains same
4-->i am tossing only 2 times so it is a finite trail

what is probabilty of appearing 1 tail when you toss 2 coins at a time
s = {HH,HT,TH,TT}
P(T) = 2/4
Q(NOT TAIL) = 1/2

= q**2, qp, pq, p**2

$$\boxed{P_r = {}^nC_r\, p^r q^{n-r}}$$

C ---> stands for combinations
n ---> No of onsite scenes
r ---> No of outcomes

$$p(t) = \frac{n!}{r!(n-r)!} \times p^r \times q^{n-r}$$

$$= \frac{2\times 1}{1(2-1)}\, p^1 \times q^{2-1}$$

$$= 2pq$$

#if i toss 3 coing what is probability of getting 2 heads

$$\frac{3\times 2\times 1}{2(3-2)} \times p^2 \times q^{3-2}$$

$$3p^2q$$

#bernoulis trail--> Binomial distribution:- How many time a particular data point appeared or repeated in combinations in overall data we use Binomial distribution.

$$\boxed{P_x = {}^nC_x\, p^x q^{n-x}}$$

#geometrical distribution:- it tells you what is probability till you get 1st success.
ex:- rolling a dice and appears the 1st 6?

GD = p --> success
q --> failure

$$P_{r+1} = q^{\text{no of trials}-1}\, p$$

q)A user is rolling a dice untill and unless get even number on dice what is probability that the user can succeed in the 4th trail?
ss = {1,2,3,4,5,6}
={2,4,6} , {1,3,5}
p = 1/2
q = 1/2
p(x=4) = 1/2*1/2 4-1

$$= \frac{1}{16} = 0.06 \qquad \frac{1}{32} = 0.03$$

q)In a compitation wining prob is 30% what is the probabilty the user can win in 5th trail?

$$P_{n=5} = 0.3 \times (0.7)^4 = 0.072$$

-- The probability where we deal with constant numbers is known as "discrete probability" -- assign a specific value for particular event
Eg:- no of students in a class room.
when it is discrete values we find the probability by using the Binomial distribution and geometrical distribution.

when the values are non-discrete then they are called "continuous probability"
Ex:- water in a glass
area of a circle

##Expectation:- is the average value of repitations of the experiment

#what is the expected number coin flips for getting a head?
x - number of flips
a. the first flip head - probability of flipping -->0.5
b . the first flip is tail--> 1 trail is wasted

x = 0.5(1) + 0.5(1+x)
x = 2

what is the expected number of coin flip s fro getting 2 consecutive heads?

Ex = $\sum x_i P_i + x_2 P_2 + \cdots + x_n P_n$

Candidates are appearing for the interview on after the other.the probability of each candidate selected is 0.16. what is the expected number of candidates that you will need to interview to make sure that you select somebody

let 'x' be the no of canditates to be interviewed
probability of selecting 1st candidate is 0.16

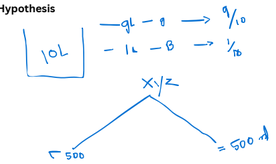
(1-0.16)*(x+1)

x = 0.16+(1-0.16)*(x+1)--> 6.25

#continuous distribution:- has a range of values that are infinite.

1.Uniform distribution:-
2.Normal distribution:-

Day05 :- Hypothesis



when there is 2 derivations this is called hypothesis
#what is hypothesis?
#Hypothesis is nothing but quantitative statments of population
There are 2 types of hypothesis

1. H0-->Null hypothesis-->it is a claim about population that is assumed to be true until it declared as false
2. H1--> Alternate hypothesis--> opposite to the null hypothesis

ex:- h0 :- says weight of the milk is equal to 500ml
h1 : says weight of milk is != 500ml
b)A beer manufacturing company adds 5% alcohol to the beer then a person claims the company is adding more then 5%
H0:- OH % is 5%
H1:- OH is > 5%
c)delhi govt says the pollution is <0.07ppm how ever people says it is more then 0.07ppm
H0:- pollution is < 0.07ppm
H1:- pollution is >0.07ppm

when we analyse either left tail or right tail we call it as a "one tail test"
when we analyse both the tails it is know as two tails test
we use to tests to find the hypothesis

1. T-test
2. Z-test

Z-score test:- the sample size is more then or equals to 30

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

x--> sample mean
u--> population mean
sigma -- SD
N -- no of samples
T - test:- when the samples less then 30

q)A sample size of 400 was drawn and the sample mean has found to be 99 , test weather a sample would have come from normal population mean is 100 and SD is 8 at 5% level of significance?
n = 400
x = 99
u = 100
SD = 8

$$Z = \frac{99 - 100}{8} \sqrt{400}$$
$$z = -2.5$$
$$\alpha = 0.05\%$$
$$z_T = 0.0054$$
$$|z| = 2.5$$

if the |z| is greater then the z alpha reject the null hypothesis and accept the alternate hypothesis.

H0 is rejected and H1 is accepted -- Type-1 Error
H0 is accepted even H1 is true ---> Type -II error
Q)A medicine is tested on a person
let H0 :- medicine is curing the diesease
correct:- no medicine is not curing the diesease

#what type of error -->TYPE 1
q) Delhi govt claims ppm in air is less then 70
#H0 is incorrect but our anlysis says more then 70--> 1
#H0 is incorrect but our analysis it is == 70

#How the null hypothesis is accepted ?
#Null hypothesis tested under 2 bais
#1.parametric testing:- we compute mean median mode
ex:- Z-Test and T-test
#--> Parametric testing applied when my data is normally distributed
#2.Non-parametric testing:- when the data is not normally distributed we non-parametric testing
ex:- Chi-square test

Day06:- Parametric test

$$t = \frac{\bar{x} - \mu}{SD} \sqrt{n}$$

	X_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	63	63−67	16
2	63		16
3	64		9
4	68		4
5	66		1
6	69		4
7	69		
8	70		
9	70		
10	71		
$\bar{x}$	67		

$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{48}{9}}$$

$\sigma = 3.06$

$n = 10$

$\alpha = 0.05$

$\mu = 65$

$$\frac{67 - 65}{3.126} \sqrt{10} =$$

$t = 2.02$

$t_\alpha = 2.262$

$t < t_\alpha$

h_0 ✓

h_1 ✗

Non-parametric test:- when the data is not normally distributed we use non parametric test

1.Chi - square test:-

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

df = 1

df = 2

df = 3

M	T	W	Th	Fr
28	22	18	20	32

	ob − ex	$(ob - ex)^2$	$(ob - ex)^2 / ex$
28	28−24	4²	16/24 = 0.66
22		2²	0.16
18		6²	1.5
20		4²	0.8
32		8²	2.66
24			

$5.66 / 5$

$\chi_c = 1.12$

$\chi_T = 9.4488$

$\chi < \chi_T$

```
for i range (n, m+1):
    for j in (n, m+1):
        if i % j = 0:
            break

else:
```

Day07:- stats

T-test:- 1.sample size less then 30
2. The population and SD is unknown
3. The population from which samples are taken are normally distributed

$$t = \frac{\bar{x} - \mu}{SD} \sqrt{n}$$

q)10 students in a class room avg is u = 65 level of significance is 0.05?

Q) given the mean of the population u = 140, n = 26, x= 147 SD = 16, alpha = 1%

$$t = \frac{\bar{x} - \mu}{SD} \sqrt{n}$$

$$= \frac{7}{16} \sqrt{26}$$

$$= 2.23$$

$df = 25$

$t_\alpha = 2.787$

$t < t_\alpha$

H_0

#Non-parametric test:- Chi-square test

Condition to apply Chi-square-test.
1.when the data is not normally distributed
2.when i take df = 1 --> the no of samples = 2
3. always the total area under the curve is 1
#q)when a data is normally distributes?--> mean is 0 and SD = 1

q) THERE is employees on company

	M	T	W	Th	Fri
Ab	5	4	2	3	6

from the above what si the degree of freedom and avg?

avg = 4

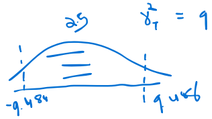
df = 4

$$M = \frac{(ob - Ex)^2}{Ex\ avg} = \frac{(5-4)^2}{4}$$

$= 0.25$

$\chi^2 = 2.5$

$\chi^2_T = 9.484$



when the calculated chi-square value present in between table values then we need to accept the null hypothesis

Q)a healthy human from age (19 -28) the weight avergae for this group 70kg i would like to know 5 persons are belongs healthy candidate population or not?
age (19-28)---->u → 168

q)In americal it has been observed that the dolls purchased by the different race is different we have to verfy it?

	B	W
ob	61	28
ob-Ex	169	-169
	272.25	272.29

$df = 1$

$\chi^2 = 12.236$

$\alpha = 0.05$

3.841

-3.841

Anova testing
Kurtosisi