DOMS Page No.
Date

What is Statistics.

Statistics
- Descriptive
- Inferential

Ans Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.

Descriptive — data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Variable Type

- (Quantitative)
  - Numeric
    - Discrete
    - Continous
- (Qualitative)
  - Categoric
    - Binomial
    - Nominal
    - Ordinal

Hight	Height(mm)
tall	195
Short	156
very short	163
Tall	185
short	167

Size of T-shirt
XL
L
L
XL
XXL
XL
L
X
L

Color of car
red
silver
white

Shot on AWESOME A05s

XL - 3
L - 3
XXL - 1

Ordinal

red - 1
white - 3
Black - 1

Nominal

result
positive
Negative
positive

Binomial

No. of Participe
37
35
42
27

Discrete

Weight	Height
30.1kg	173.1cm
40.2kg	174.01
44.5kg	175.88
45.6kg	180.11

Continous

No. of Pen
7
5
2
6
7

~~Catergorical data~~

* Qualitative / Qualitative data
Describes the data that fits into the categories. Qualitative data are not numerical
* Nominal Data — Nominal Data is one of the types of qualitative information which helps to label the variables without providing numerical value. It cannot be ordered or measured.
* Ordinal Data - Data that follows a natural order. The ordical Data is commonly represented using a bar chart.
* Quantitative / Numerical Data
Quantitative data which represents the numerical value. Numerical data gives information about the quantities of a specific thing.
* Discrete Data - Discrete Data can take only discrete values. Discrete information contains only a finite no. of possible values.
* Continuous Data is data that can be calculated & has a infinite number of probable values that can be selected within a given specific range.

Shot on AWESOME A05s

DOMS Page No.
Date / /

Sampling

- Probablistic Sampling
  - Simple Random Sampling
  - Stratified Sampling
  - Systematic Sampling
  - Cluster Sampling
- Non-Probablistic Sampling
  - Convenience
  - Consecutive

Simple Random Sampling (SRS)

- Simplest and most common method.
- Sample is drawn unit by unit
- Each unit has equal probability of being selected.
- Two ways of selecting units:
  - without replacement: if the selected unit is not returned to the population
  - with replacement:
    if the selected unit is returned to the population.

Drawing samples with replacement and without replacement

```
import random

weight = [67, 76, 56, 47, 66, 57, 79, 88, 54, 63, 82,
          74, 95, 86, 94]
```

Shot on AWESOME A05s

# without replacement
random.sample(weight, k=10)

[79, 94, 56, 67, 86, 63, 66, 74, 54, 95]

# with replacement
random.choices(weight, k=10)

[63, 74, 74, 79, 82, 56, 66, 66, 47, 88]

Sampling

In Statistics, the sampling method or Sampling technique is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous.

Probability Sampling

The probability sampling method utilizes some form of random selection. In this method all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of population.

Shot on AWESOME A05s

Week 02

Simple Random Sampling

(i) simple and most common method
(ii) sample is drawn unit by unit
(iii) Each unit has equal probability of being selected.

In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as "Method of chance selection". Ex 1. lottery system.
2. throwing a dice.

SYSTEMATIC SAMPLING

The items are selected from the target population by selected the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size

Ex ① 2 groups 60% female and 40% male.

② 2 groups school students and University students.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30

5 employees.

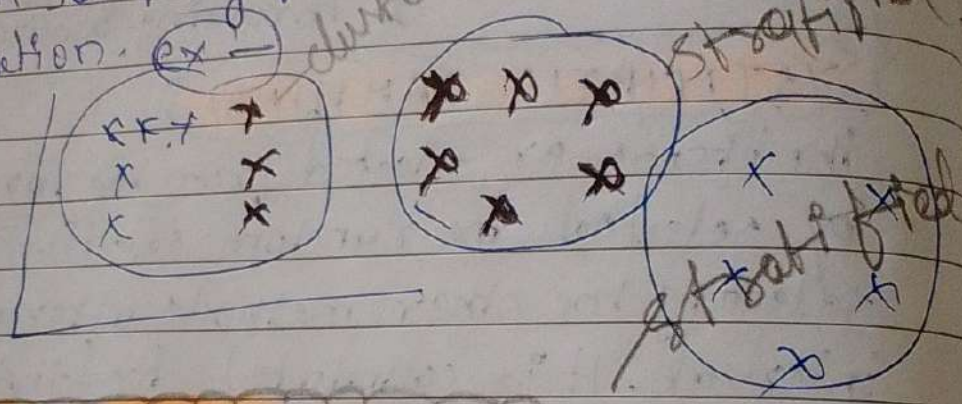
STRATIFIED SAMPLING

The total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group the statisticians randomly select the sample. Ex—

Shot on AWESOME A05s

CLUSTERED SAMPLING

The cluster of group of people are formed from the population set. The group has significatory characteristics. Also they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population. ex –



Non Probability Sampling

The non probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not at all the members of the population have a chance to participate in the study

Convenience Sampling

In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. The samples are easy to select and the researcher did not choose the sample that outlieries the entire population

Consecutive Sampling

consecutive sampling is similar to convenience sampling with a slight variation The researcher picks a single person or a group of people for sampling. Then the researcher researches for a period of time to analyze the result and move to another group if needed.

Shot on AWESOME A05s

Descriptive a

Date / /

Descriptive statistics

- Measure of Central Tendency
  - Mean:
  - Median:
  - Mode:
- Distribution of the Data
- Measures Dispersion

SL

(median)

12
12
12
13
22
28
29
29

$$\left(\frac{n+1}{2}\right) \Rightarrow \left(\frac{8+1}{2}\right) = 4.5$$

$$\frac{13+22}{2} = \frac{35}{2} = 17.5$$

$$\left(\frac{n+1}{2}\right) \Rightarrow \left(\frac{9+1}{2}\right) = 5$$

12	3
13	1
22	1
28	1
29	2
180	1

$\rightarrow$ mode = 12

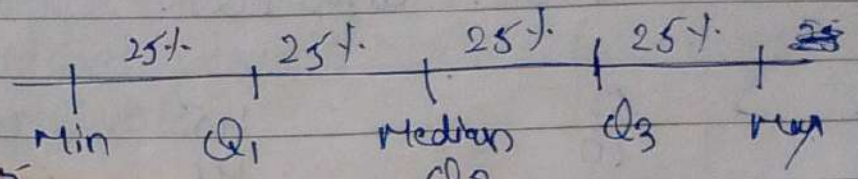
Shot on AWESOME A05s

Partition Values!!

Partition values are defined as the values that divide the data into equal parts. Quartiles divide the data into 4 equal parts, Deciles divide the data into 10 equal parts and percentiles divide the data into 100 equal parts

~~data~~

19
32
32
21 — 26.5
10
12
21 — 21
35
16
13
16 — 14.5
16
28



$$Q_1 = \left(\frac{n+1}{4}\right)^{th}$$

$$Q_2 = \left(\frac{n+1}{2}\right)^{th}$$

$$Q_3 = \frac{3}{4}(n+1)$$

$n = 13$

$Q_1 = \frac{14}{4} = 3.5^{th}$, $Q_2 = 14/2 = 7^{th}$, $Q_3 = 3/4 \times 14 = 10.5^{th}$

Q. A manager handles 12 branches of a Supermarket situated in the U.S.A. Consider one day sale (in dollars) of all the branches. Calculate the mean and median to find the average sale.

Sale = [165, 182, 140, 193, 172, 168, 174, 124, 187, 204, 148, 175]

mean = np.mean(Sale)

print("Mean", mean)

median = np.median(sale)

Mean: 169.33

Median: 173.[illegible]

Shot on AWESOME A05s

Measure of dispersion

A measure of dispersion the variability in the data. Some of the measures of dispersion are range, variance, standard deviation, coefficient of variation and IQR.

0	45
10	46
20	47
30	48
40	49
50	50
60	51
70	52
80	53
90	54
100	55
Mean = 50	50

range = max − min

IQR = $Q_3 - Q_1$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Range:

It is defined as the difference between the largest and smallest observation in the data. It is affected by the presence of extreme observations.

$$\boxed{\text{range} = \max - \min}$$

Variance:

It calculates the dispersion of the data from the mean. It is defined as the average of the sum of squares of the difference between the observation and the mean.

Shot on AWESOME A05s

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N}$$ variance

Standard Deviation:-

It is the positive square root of variance. The unit of standard deviation is the same as the unit of data points. The variable with near-zero standard deviation is least important for the analysis.

$$\text{Standard Deviation} = \sqrt{\text{variance}} = \sqrt{\sigma^2} = \sigma$$

Coefficient of variation

It is measure of the dispersion of data points around the mean. It is always expressed in percentage. We can compare the coefficient of variation of two or more groups to identify the group with more spread.

$$CV = \frac{S}{\bar{X}} \times 100\%$$ Sample

$$CV = \frac{\sigma}{\mu} \times 100\%$$ ← std deviation, ← mean, population

Interquartile Range (IQR)

It is defined as the difference between the third and first quartiles. It returns the range of the middle 50% of the data. IQR can

$$IQR = Q_3 - Q_1$$

Shot on AWESOME A05s

Degree of Freedom

35 — S_1 — S_2 (81) — S_3 (13) — S_4 — S_5 (?) ≈ 76

flexibility of changes

	population	Sample
Mean	μ	$\bar{x}$
st. deviation	σ	s
var.	σ^2	s^2
size	N	n

$[23, 41, 56]$ — 1, 2, 3

$$\sum_{i=1}^{3} x_i \Rightarrow x_1 + x_2 + x_3$$

* The sum of the deviations of the x values from the mean is always zero

Shot on AWESOME A05s

Q A manager handles 12 branches of a supermarket situated in the USA. Consider one day sale (in dollars) of all the branches calculate the standard deviation of the sale. Also, find the range in which the middle 50% of the sale would lie.

sale = [165, 182, 140, 193, 172, 168, 174, 124, 187, 204, 148, 175].

```
range = max(sale) - min(sale)
print('Range:', range)
var = np.var(sale)
print('Variance:', var)
```

```
q3, q1 = np.quantile(sale, [0.75, 0.25])
iqr = q3 - q1
stdev = np.std(sale)
print('Standard Dev:', stdev)
```

Range: 80.
Variance: 473.88888889.

Standard Dev: 21.768889563409 3.

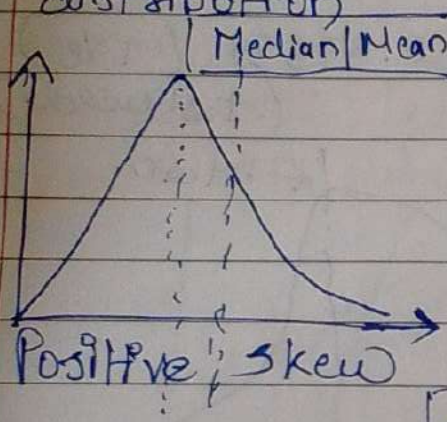
Measure of Shape

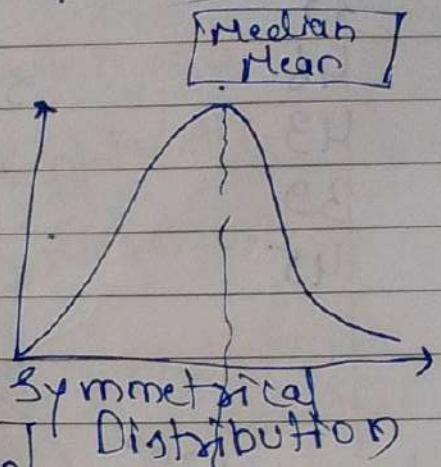
Skewness.

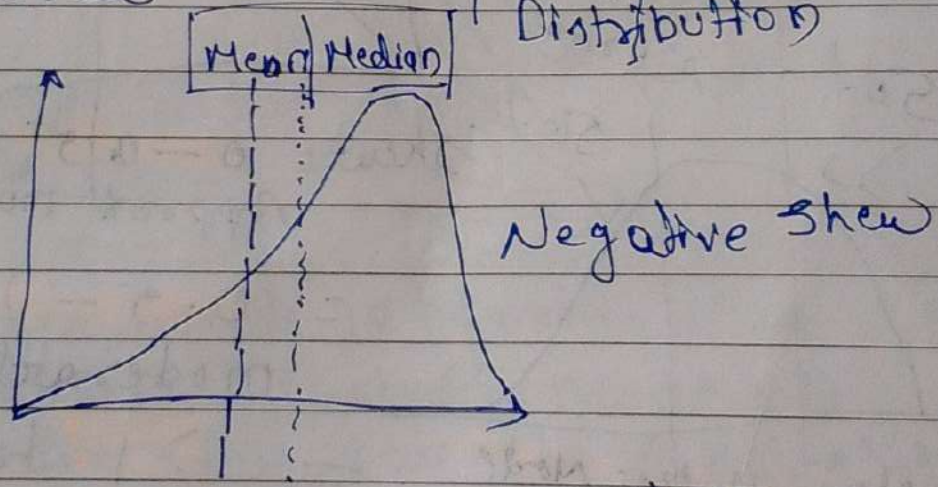
It measures of degree to which the distribution of the data differes from the normal distribution. The value of skewness can be positive, negative or zero.

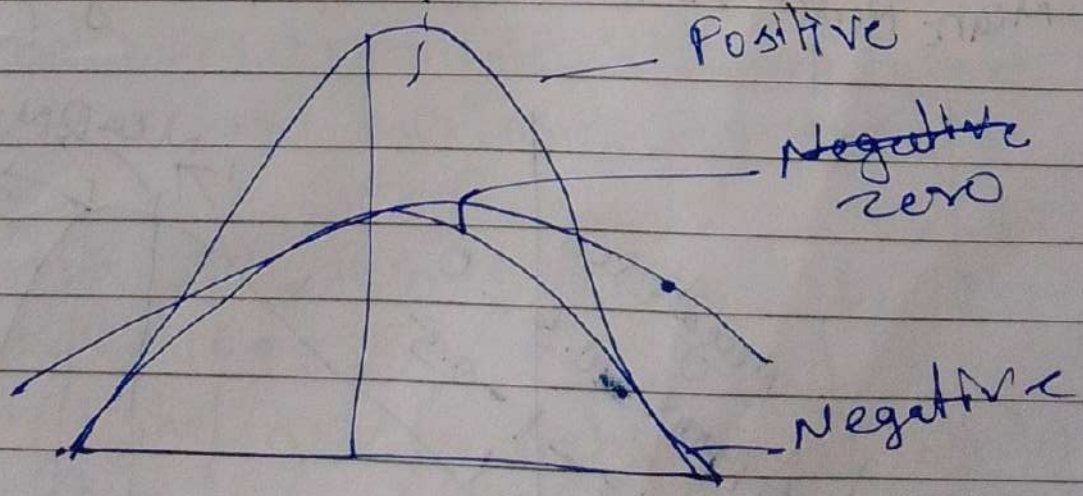
KURTOSIS

It identifies the peakeness of data distribution. The positive value of kurtosis represents the leptokurtic distribution, the negative value represents the playkurtic distribution and zero values represents the mesokurtic distribution









Covariance

It measures the degree to which two variables move together. The value of covariance can be between $-\infty$ to ∞. The magnitude of covariance is not easy to interpret

Correlation

It is the normalized value of covariance. The correlation value near to +1 indicates a strong positive correlation between the variables and value near to -1 indicates a strong negative correlation.
(-1, +1)

Data1	Data2
Weight	Height
67	156
45	144
82	176
77	163
63	152
37	133
89	182
	~~158~~
Mean 65.7143	158
Sd 17.1777	15.9732

cov 275.714

$$\boxed{\begin{aligned} &cov(x,y) \quad [-\infty \text{ to } \infty] \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n} \end{aligned}}$$

$$\boxed{\begin{aligned} &cor = \frac{cov(x,y)}{std(x)\,std(y)} \\ &[-1 \text{ to } 1] \end{aligned}}$$

Shot on AWESOME A05s

Quartiles	Percentiles	decile
$Q_1 = 1/4(n+)$	$\frac{30}{100}(n+1)$	$\frac{4}{10}(n+1)$
$Q_2 = 2/4(n+1)$	$\frac{60}{100}(n+1)$	$\frac{7}{10}(n+1)$
$Q_3 = 3/4(n+1)$		

PROBABILITY

- Probability is a branch of mathematics that deals with calculating the likelihood of a given events occurence.
- Probability is the measure of the likelinesss that an event will occur.
- Probability is quantified as a number between 0 and 1
- The higher the probability of a event the more certain that the event will occur.

$$P = \frac{\text{favourable Outcomes / Desired.}}{\text{Total Outcomes}}$$

Q From a deck of cards, 10 cards are picked at random and shuffled. The cards are as follows:
6, 5, 3, 9, 7, 6, 4, 2, 8, 2
Find the probability of picking a card having value more than 5 and find the probability of picking a card with an even number on it

total - possible = 10
desired outcomes = 5
prob = (desired outcomes / total possible)
print ('Probability: ', prob)
probability : 0.5

Addition Rule

जो events साथ में occur न हो

- If X and Y are two mutually exclusive events, then the probability of 'X Union Y' is the sum of the probability of X and the probability of Y and represented as,

ex - tossing a coin

$$P(X \cup Y) = P(X) + P(Y)$$

जो साथ में occur हो

- If X and Y are two mutually Non-Exclusive Events, then the probability of 'X union Y' is the difference between the sum of probability of X and the probability of Y and the probability of 'X intersection Y' and represented as,

ex King is a lord
heart is a lord

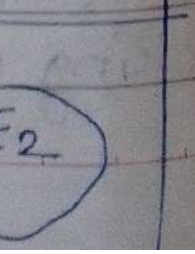
$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

जो कोई King Lord भी है और heart भी है

$S = \{1, 2, 3, 4, 5, 6\}$
$E_1 (n > 4) = \{5, 6\}$
$E_2 (3) = \{3\}$

Mutually Exclusive

$P(n > 4 \text{ or } 3) = P(n > 4) + P(3)$
$\frac{2}{6} + \frac{1}{6} = \frac{3}{6} = 0.5$



Shot on AWESOME A05s

A : 0.75

(B) prob_male_or_shoped = (1200/2000) + (800/2000) − (500/2000)

print('Ans B:', prob_male_or_shoped)

Ans B : 0.75

(C) prob_never_or_shoped = (1200/2000) + (800/2000)

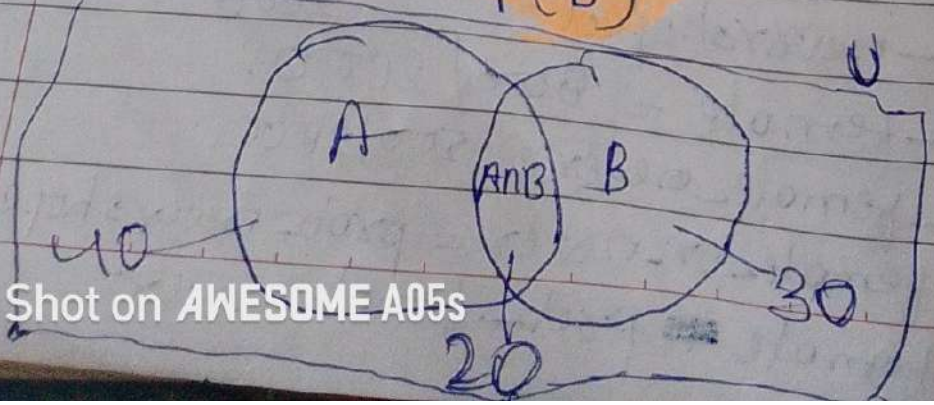
print('Ans C:', prob_never_or_shop

Ans C : 1.0

CONDITIONAL PROBABILITY

Conditional Probability is the probability th event will occur given that another event has already occured. If A and B are two events, then the conditional probability of A and B is written as P(A|B) and read as "the probability of A given that B has already occured." It can be used to determine a conditional probability from unconditional probabilities is

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



Q. From a two card is the p kings

P(K2/K1)

Q. Prob a bag blue

with replacement

Shot on AWESOME A05s

|| Multiplication Rule ||

According to the multiplication rule of probability of occurence of both the events A and B is equal to the product of the probability of B occuring and the conditional probability that event A occuring given that event B occurs.

⊙ If A and B are dependent events, then the probability of both events occuring simultaneously is given by

probability of B when event A has already occured.

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

जब दो events depend होते हैं

⊙ If A and B are two independent events in an experiment, then the probability of both events occuring simultaneously is given by

जब दो events independent हैं तो probability multiply होती है

$$\boxed{P(A \cap B) = P(A) \cdot P(B)}$$

Q1. What is probability of Red (first one) and Red (second one) also red?

$$\boxed{P(R_1 \cap R_2) = P(R_1) \cdot P(R_2/R_1)} \quad \frac{5}{12} \cdot \frac{4}{11} = \frac{20}{}$$

Q2. Blue (first one) and Red (second one)

$$\boxed{P(B_1 \cap R_2) = P(B_1) \cdot P(R_2/B_1)}$$

$$= 7/12 \cdot 5/11$$

Q3 Blue (first one) and Blue (second one)

$$P(B_1) \cdot P(B_2/B_1) = 7/12 \times 6/11$$

Shot on AWESOME A05s

BAYE'S THEOREM

Bayes theorem shows the relationship between one conditional probability and its inverse it provides a mathematical rule for revising an estimate of experience and observation

Likelihood
probability of 'B' being True, given 'A' is true

Prior
The probability 'A' being True
This is the knowledge

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

Posterior
The probability of "A" being True, given "B" is true

Marginalization
probability "B" being True.

Baye's theorem

$$P(\text{Absent}/\text{rain}) = \frac{P(\text{Absent} \cap \text{Rain})}{P(\text{rain})} \quad \text{—(1)}$$

$$P(\text{rain}/\text{Absent}) = \frac{P(\text{Absent} \cap \text{Rain})}{P(\text{Absent})} \quad \text{—(2)}$$

$$P(Ab/Rain)\,P(rain) = P(Ab \cap Rain)$$

$$P(rain/Ab)\,P(Ab) = P(Ab \cap Rain)$$

$$P(Ab/Rain)\,P(rain) = P(rain/Ab)\,P(Ab)$$

$$\boxed{P(rain/Ab) = \frac{P(Ab/rain)\,P(rain)}{P(Ab)}}$$

Shot on AWESOME A05s

Day-2 Session-3

Random Variable

Discrete RV	Continous RV
no. of cars	prob. of height
no. on dice	prob. of weight
no. of girls in class	prob of ant sugar
tossing a coin	
no. on tambola	

Discrete Prob. distribution	Continous Prob. distribution
→ Binomial distribution	→ Normal distribution
→ Hypergeometrical distribution	→ standardize Normal Distribution.
→ poisson distribution	

Random Variable = It is defined as a variable that takes all the possible values of a random variable

ex =

No cars	Families
0	30
1	70
2	50
3	40
4	10
	200

what is prob of selecting a family that own 2 cars = $\frac{50}{200}$

Probability

$P(x=2) = 50/200$

Random Variable

$P(x=0)$
$P(x=1)$
$P(x=2)$
$P(x=3)$
$P(x=4)$

Types

(1) Discrete Ran
A random varia
countably infinit
called a discr
let us study so
discrete randa
(1) Number of ou
the amusemen
(2) number of stu
(3) Number of hom
a restaurant
(4) Number of
an airport

(2) Continous R
A random
infinite numb
called a con
(1) weight of
(2) The diameter
(3) Mileage of

= Var v = Va

Shot on AWESOME A05s

Types

(1) Discrete Random variables

A random variable that takes finite or countably infinite number of values is called a discrete random variable.
let us study some of examples of a discrete random variable.

(1) Number of occupied car parking spots in the amusement park
(2) number of students selected in exam hall
(3) Number of home delivery orders taken by a restaurant in a day
(4) Number of passengers passed through an airports's security check

(2) Continous Random Variable

A random variable that takes an infinite number of values in a range is called a continous random variable.

(1) weight of people registered for fitness program
(2) The diameter of water pipe produced in a factory
(3) Mileage of car

Data Science

Shot on AWESOME A05s

||BINOMIAL DISTRIBUTION||

A discrete variable X taking values 0, 1, ... n follows a binomial distributions with parameters n and p, if the pmf of X is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, \ldots$$

① only two outcomes.

② p denotes the probability of success of experiment and q denotes the probability of failure (where, $p+q=1$)

③ If n independent bernoulli trials (each with a probability of success) are executed then the number of successes follows a binomial distribution.

② Both the outcomes are independent

The mean & variance of distribution is given as.

① complement लेना पड़ेगा

Mean = np.

Variance = npq.

S. No.

Date

Part 3 CONTINOUS PROBABILITY DISTRIBUTION

It is the probability distribution related to the continous random variable. The area under probability density function (pdf) gives the probability that a variable lies in a specific range. The value of pdf at a specific point is always 0.

Probability that X lies between [a, b] is given by

$$p(a \leq X \leq b) = \int_a^b f(x)dx$$

Area under the curve is 1 i.e $\int_{-\infty}^{\infty} f(x)dx$

For a continous random variable X the cummulative distribution function (cdf) is denoted by F(x) and defined as:

$$\boxed{F(x) = \int_{-\infty}^{x} f(u)\, du.}$$

The cdf of a random variable returns the probability that the variable takes all the values less than or equal to the specific value

Shot on AWESOME A05s

Normal Distribution

Questions Nainamesha diya hoga normal

A continuous variable X follows a distributes normal distribution with the parameters μ and σ if the pdf of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The normal distribution is a bell shape symmetric distribution with mean = median = mode

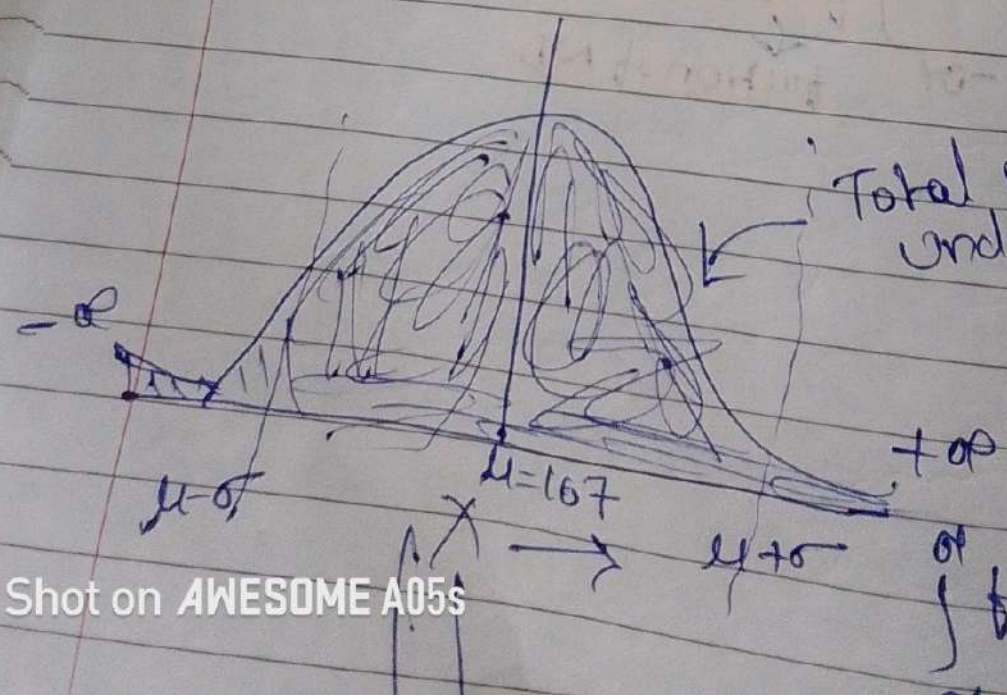
Mean = μ

~~Median~~ variance = σ^2

Z score

Note: The variable Z given as, $\boxed{Z = \frac{X-\mu}{\sigma}}$ follows a standard normal distribution with mean 0 and variance = 1

st. deviation = 1



$N(\mu, \sigma)$

$\int_{-\infty}^{\infty} f(x)dx = 1$

Shot on AWESOME A05s

According to Central Limit Theorem

1. whether the population is normally distributed or not, the sample distribution of that population will always be normally distributed (n>30)

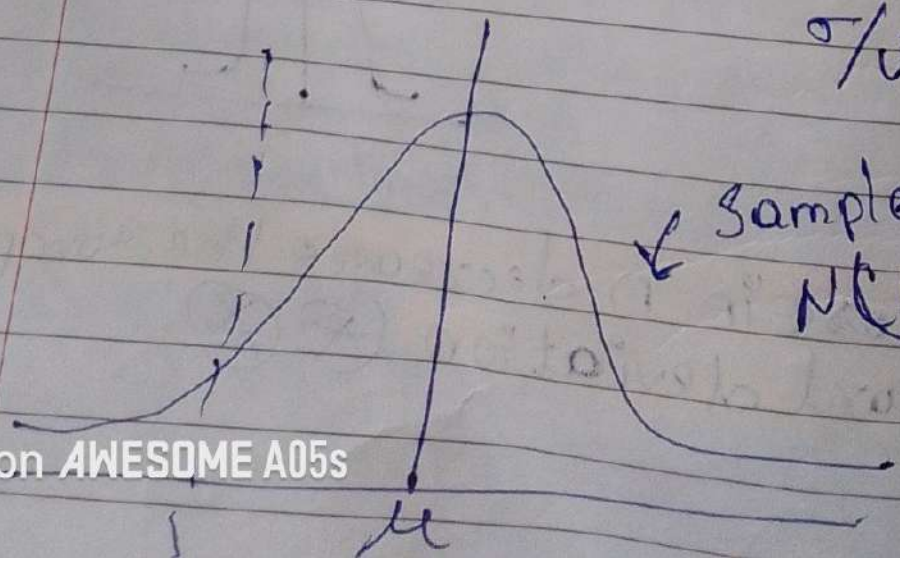
2. These sample means are distributed across the population mean

3. The SD of sample mean is always less than the pop mean and the SD of sample distribution is given by standard error (sigma / sqrt(n))

4. As the sample size increases the sample standard deviation decreases

population distribution $\qquad$ sample distribution

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



Sample dist. $N(\mu, \sigma/\sqrt{n})$

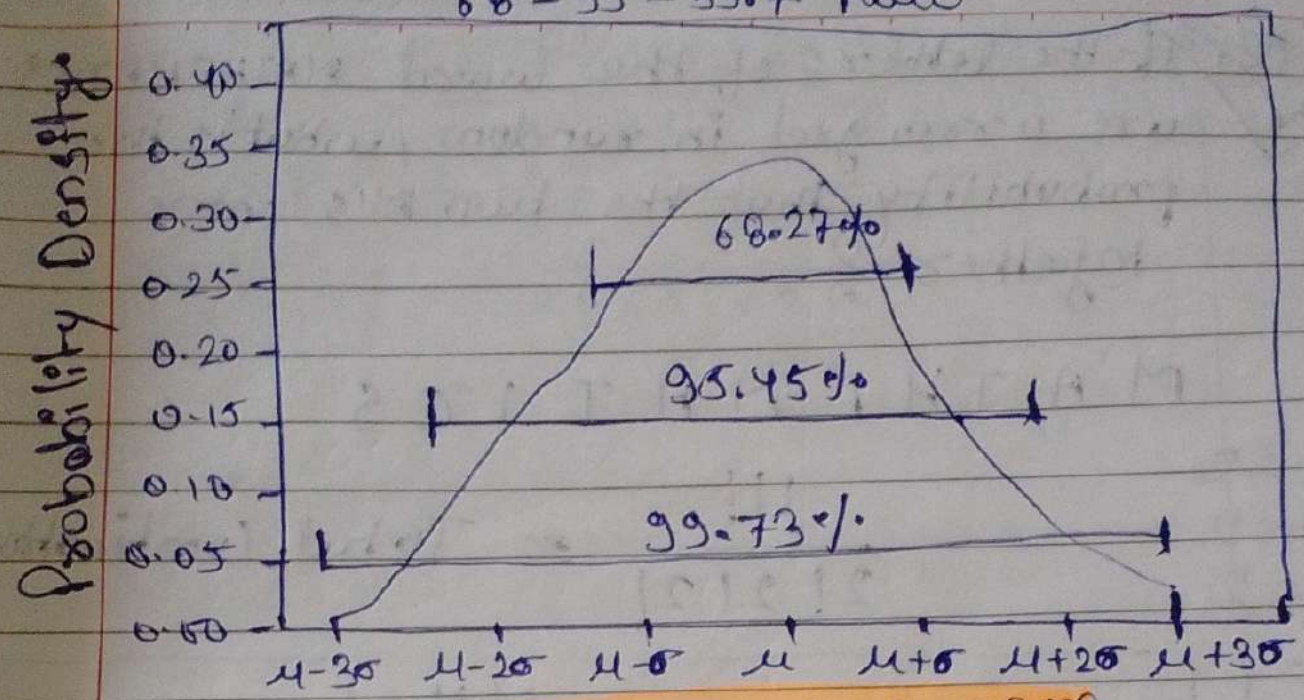
Q. In a emplo 2875 empl mean

Q2.

Shot on AWESOME A05s

EMPERICAL RULE

68 - 95 - 99.7 Rule



In 1st SD 68.26894921370858

(stats.norm.cdf(x=1) - stats.norm.cdf(x=-1))*100

In 2nd SD 95.44997361036415

(stats.norm.cdf(x=2) - stats.norm.cdf(x=-2))*100

In 3rd SD 99.73002039367398

(stats.norm.cdf(x=3) - stats.norm.cdf(x=-3))*100

Shot on AWESOME A05s

what is distribution is not Normal

chebyshev's Rule

This rule states if the data is not Normal then mu-k*sigma to mu+k*sigma will have atleast

(1-1/k^2)* 100 percent of data.

If data is not normal

mu-1.5*sigma to mu+1.5*sigma

```
k=1.5
per = (1- (1/(k**2)))* 100
per
print ('Your distribution will have atleast',
        round(per,2), '% of data')
```

Your distribution will have atleast 55.56% of data.

If data is not normal

mu-2*sigma to mu+2*sigma.

```
k=2
per = (1-(1/k**2)))* 100
per
print ("Your distribution will have atleast',
        round(per,2), '% of data')
```

Your distribution will have atleast 75.0% of data.

Shot on AWESOME A05s

Parameter Estimation

The value associated with the characteristic of the population is known as a parameter and the characteristic of the sample is denoted by a statistic. Usually, the population parameters are not known in most of the real life problems. Then we consider a subset of the population (sample) to estimate the population parameter using a sample statistic.

Point estimation and Interval estimation are two of methods to estimate the population parameter.

POINT ESTIMATION

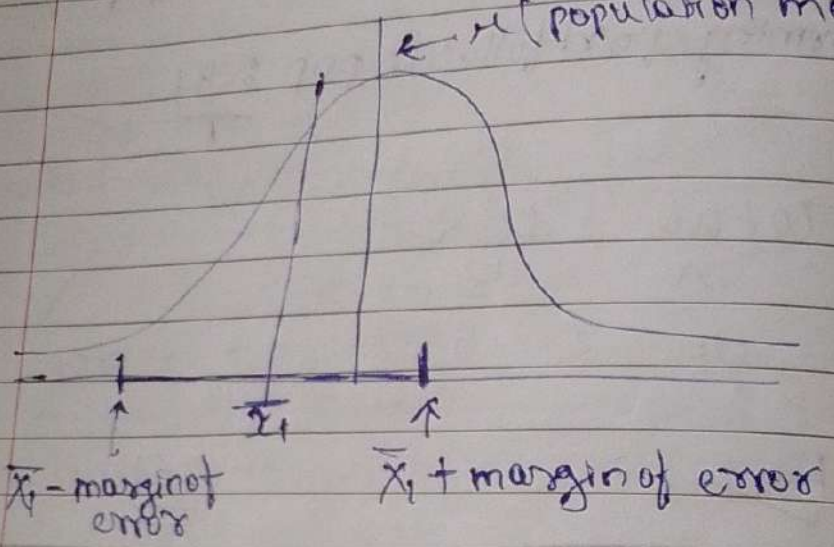
This method considers a single value (sample statistic) as estimate of the population parameter.

Let $X_1, X_2, X_3, \ldots X_n$ be the random sample drawn from a population with mean μ and standard deviation σ. The point estimation method estimates the population mean, $\mu = \bar{x}$, where $\bar{x}$ is the sample mean and population standard deviation, $\sigma = s$, where s is the standard deviation of the sample (standard

INTERVAL ESTIMATION

ex Avg of height of population is 168 cm ← Point estimate

Avg. of height of population is b/w 165 - 170 cm ← Interval estimate



$$\text{Interval} = \text{Sample mean} \pm \text{margin of error}$$

Each interval is constructed with a regard to a given confidence level, that's why we call it confidence ~~level~~ interval.

Confidence interval states that how much confidence we have that this interval contains the true population parameter.

↓

CONFIDENCE LEVEL.

This method considers the range of values in which the population parameter is likely to lie. The confidence ~~level~~ interval is an interval that describes the range of values in which the parameter lies with a specific probability. It is given by the formula

$$\boxed{\text{conf-interval} = \text{sample statistic} \pm \text{margin of error}}$$

The uncertainty of an estimate is described by the confidence level which is used to calculate the margin of error.

LARGE SAMPLE SIZE

Consider a population with mean μ and standard deviation σ. Let us take a sample of n observation from the population such that, $n \geq 30$. The central limit theorem states that the sampling distribution of mean follows a normal distribution with mean μ and standard deviation $\boxed{\frac{\sigma}{\sqrt{(n)}}}$.

The confidence interval for population mean with $100(1-\alpha)\%$ confidence level is given as

$$\boxed{\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

$\bar{X}$: Sample mean
α - Level of significance
σ: population st. deviation
n: Sample size.

Shot on AWESOME A05s

The above eq is sufficient ... sample size is ... σ by the SD of sample(s) if population SD is not known.

$Z_{\alpha/2}$ for different α values can be obtained using stat...

$$\text{interval estimation} = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

point estimate

margin of error → $Z_{\frac{\alpha}{2}}$

α — significance level

$$CL = (1-\alpha) * 100$$

confidence level

significance level

if let say we want interval with confidence level 90%

$$0.90 = (1-\alpha)$$

$$\alpha = 0.1$$ → significance level

If alpha is not provided we consider it to be

$$\alpha = 0.05 \text{ or } CL = 95\%$$

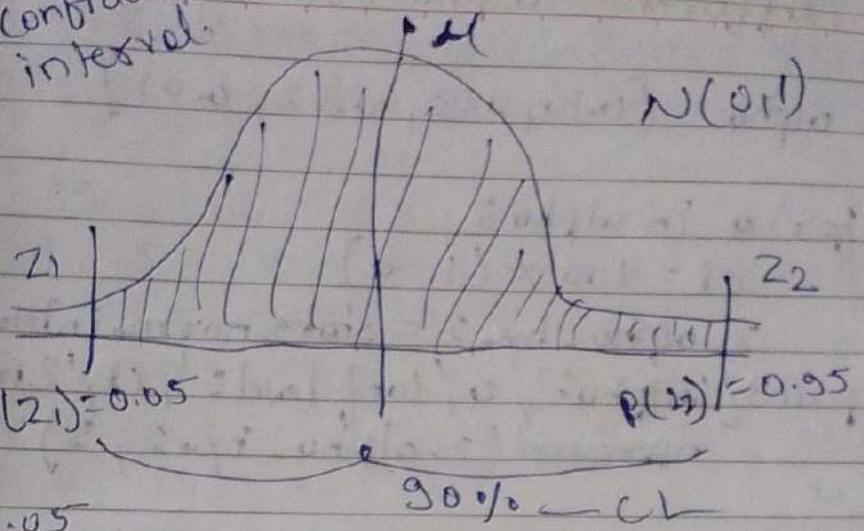
By default

CL = 90%	α = 0.1
CL = 95%	α = 0.05
CL = 97.5%	α = 0.025 → By default
CL = 99%	α = 0.01 → Some common use

Shot on AWESOME A05s

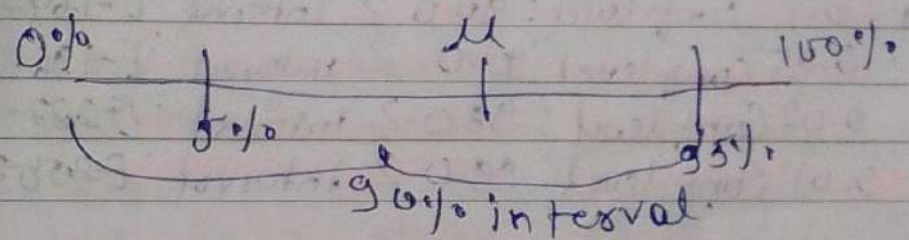
What do we mean by confidence level of 90%?

90% confidence interval

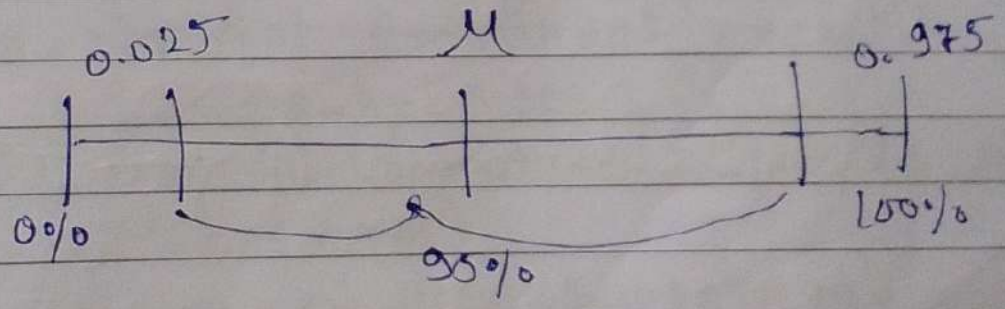


$Z_{\alpha/2}$

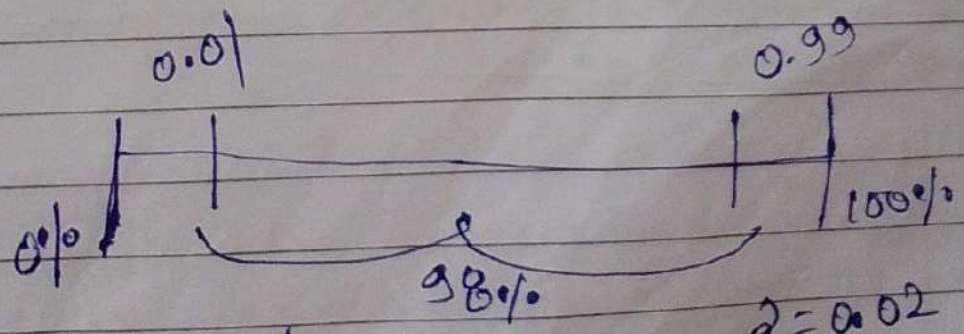
$Z_{0.05}$



95% $\alpha = 0.05$



$Z_{0.025}$



$CL = 98\%$

$\alpha = 0.02$

$0.98 = (1-\alpha)$

$\alpha = 1 - 0.98 = 0.02$

Shot on AWESOME A05s

confidence interval for the production population mean.

cl=0.9
sigma=8
x_bar=24

alpha

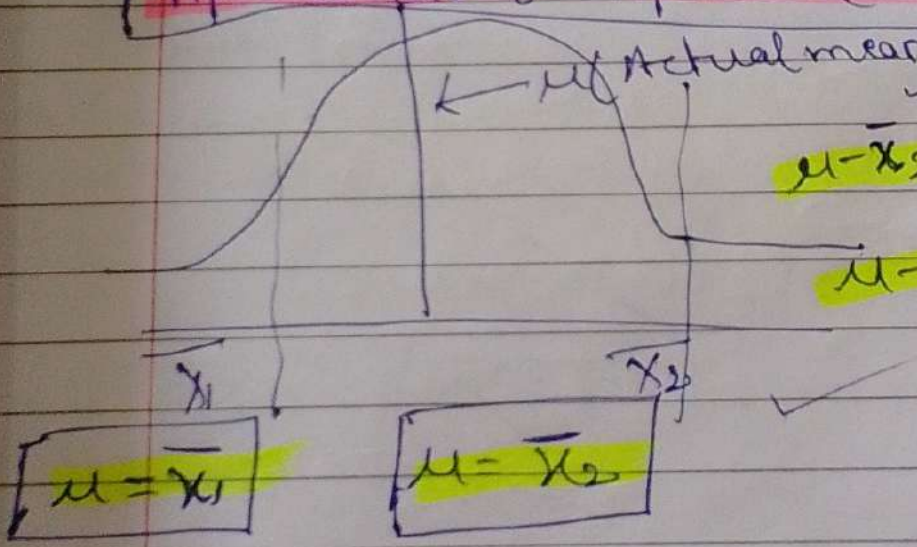
stats.norm.interval(confidence=0.9, loc=x_bar, scale=sigma/(97**0.5))

(21.836701835709 07, 26.163298164290 93)

Sample error

print('Sample error:', np.mean(data)-np.mean(sample))

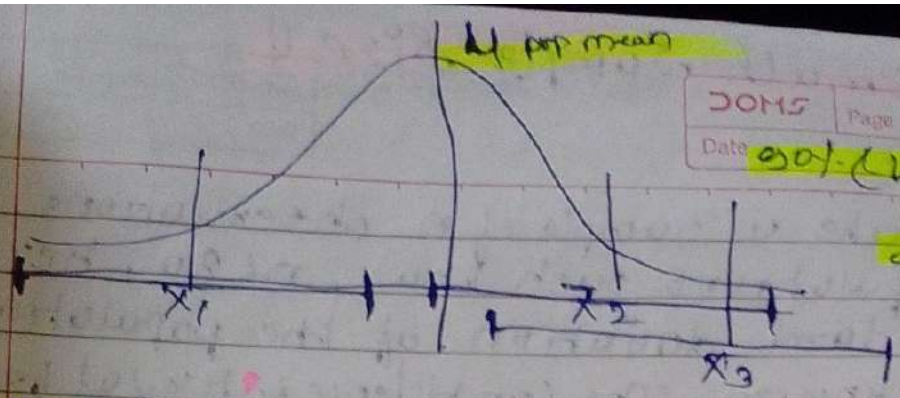
np.mean(data) - np.mean(sample)



$\mu - \bar{x}_2 \Rightarrow$ Sample error

$\mu - \bar{x}_1 \Rightarrow$ Sample error

Shot on AWESOME A05s



DOMS Page No.
Date 90% CL

$\alpha = 0.95$

$\bar{x}_1 \pm$ moe → no pop mean
$\bar{x}_2 \pm$ moe → pop mean
$\bar{x}_3 \pm$ moe → pop mean

By 90% confidence level we mean there are 90% samples which are like x_2 an x_3 and remaining 10% samples means are like x_1

$$\text{moE} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

fix

Shot on AWESOME A05s

If small Sample size

Let us take a sample of n observations from the populations such that, $n < 30$, here standard deviation of the population is unknown. The confidence interval for population mean with $100(1-\alpha)\%$ confidence level is given as $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

where,

$\bar{x}$ = Sample mean

α: level of significance

s = Sample standard deviation

$n-1$ = degrees of freedom = df

The ratio $\frac{s}{\sqrt{n}}$ is the estimate of the standard error of the mean. And $t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$ is the margin of error of the estimate

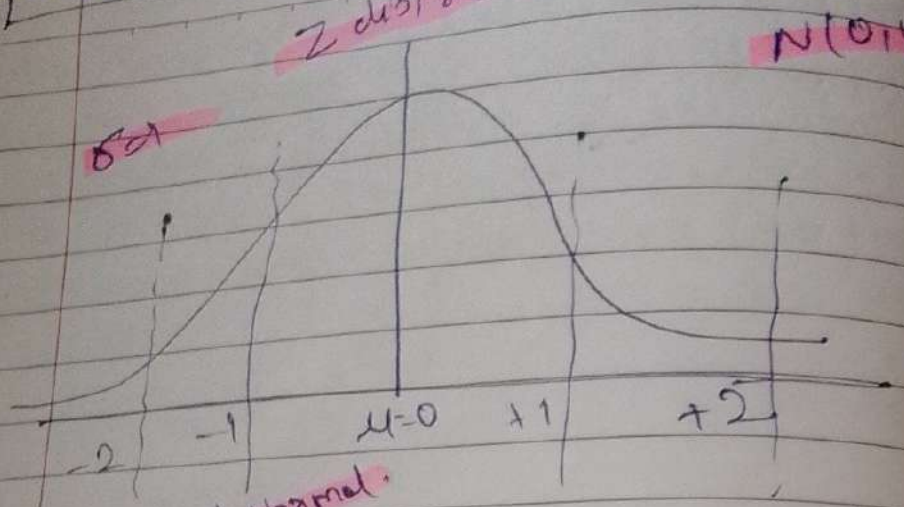
The value of $t_{\alpha/2, n-1}$ for different α values can be obtained using the stats.t.isf() from scipy library

Shot on AWESOME A05s

Z-Distribution V/S T-Distribution

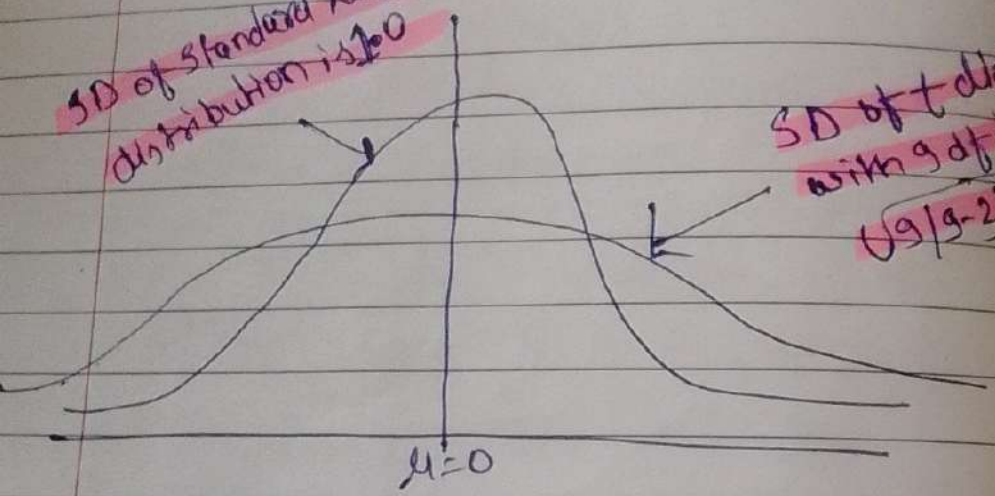
Z distribution

N(0,1)



SD of Standard Normal distribution is 1.0

SD of t distribution with 9 df is $\sqrt{9/(9-2)} = 1.13$



t distribution for df=9 and standard normal distribution.

T-Distribution
(sample size < 30)

T-distribution

$T(0, \sqrt{df/(df-2)})$

$\boxed{df = n-1}$

if $n = 10$

$df = 9$

$T_{df=19}(0, 1.057)$

$T_{df=29}(0, 1.0363)$

lets us find the t-
alpha values (0.1,

alpha = [0.10, 0.05,

for a in alpha:

cl = 100*(1-a)

t_alpha_by_2 =

print ('Alpha:', a
np.round

Alpha: 0.1 (conf Level
Alpha 0.05 (conf Level
Alpha 0.02 (conf Level:
Alpha 0.01 (conf Level

Q There are 1
randomly ch
that the ove
grams with
grams. Fi
for the po

n=17

df=n-1

x_bar=78

s=23

cl=0.9

t_alpha_by_
se = s/(n
moe = t_a
print ('In

Shot on AWESOME A05s

0.02, 0.01) at df = 10

```
alpha = [0.10, 0.05, 0.02, 0.01]
for a in alpha:
    cl = 100*(1-a)
    t_alpha_by_2 = stats.t.interval((1-a), df=10)
    print('Alpha:', a, 'Conf Level:', cl, 't interval:',
          np.round(t_alpha_by_2, 2))
```

```
Alpha: 0.1 Conf Level: 90.0 t interval: [-1.81 1.81]
Alpha 0.05 Conf Level: 95.0 t interval: [-2.23 2.23]
Alpha 0.02 Conf Level: 98.0 t interval: [-2.76 2.76]
Alpha 0.01 Conf Level: 99.0 t interval: [-3.17 3.17]
```

Q There are 150 apples on a tree. You randomly choose 17 apples and found that the average weight of apples is 78 grams with a standard deviation of 23 grams. Find the 90% confidence interval for the population mean.

```
n = 17
df = n-1
x_bar = 78
s = 23
cl = 0.9
t_alpha_by_2 = 1.75
se = s/(n**0.5)
moe = t_alpha_by_2 * se
print('interval estimation:', x_bar-moe, 'to',
      x_bar+moe)
```

Shot on AWESOME A05s

INTERVAL ESTIMATION FOR PROPORTION

Consider a population in which each observation is either a success or a failure. The population proportion is denoted by P which the ratio of the number of successes to the size of the population.

The confidence interval for the population proportion with $100(1-\alpha)\%$ confidence level is given as :

$$\boxed{p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}}$$

where,

p : Sample proportion

α : Level of significance

n : Sample size.

The quantity $Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ is margin of error

Ex- proportions

$P \rightarrow$ 60% people prefer flights.

$n = 150$

$p = 80/150 = 0.53$ = Sample proportion

$$\boxed{p \pm \left(Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)} \rightarrow \text{moe}$$

CL = 90%

TESTING OF HYPOTHESIS

Two types of Hypothesis
⇒ Null Hypothesis H_0, H_0
⇒ Alternate Hypothesis H_a, H_1

What is a hypothesis?
A Hypothesis is a claim about the population parameter

Examples:
- The average diameter of bolts manufactured is 200 cm
- The new drug is more effective.

Hypothesis
- NULL HYPOTHESIS
- ALTERNATIVE HYPOTHESIS.

(ex) A pharmaceutical company has claimed to have produced a new drug that is effective against cancer treatment

To test the pharmaceuticals claim, first write the test hypotheses - Null Hypothesis and the Alternative Hypothesis

Shot on AWESOME A05s

Week (2)
Day (1) Session (1)

Revision whole concept from week (1)

(1)

Stats

- Desc stats → Desc the data
- Inf. Statistics

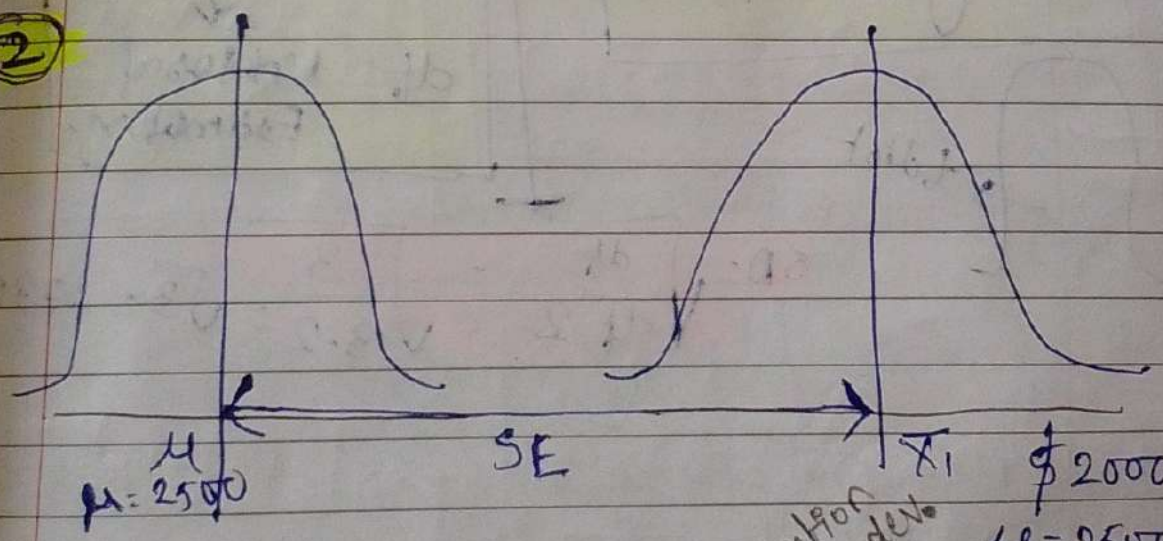
Data → Pop, Sample

Pop: μ, σ → Parameter

Sample: $\bar{x}$, s → Sample statistic

$$z\text{score} = \frac{x_i - \mu}{\sigma}$$

(2)



$\mu = \{\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_n\}$

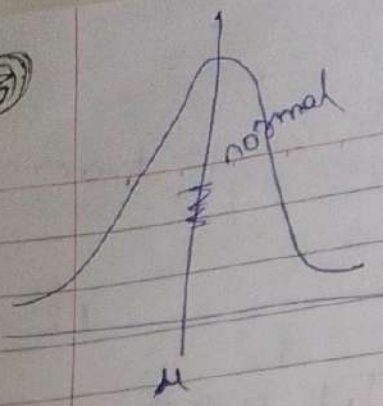
$\mu = \bar{x} \pm 500$

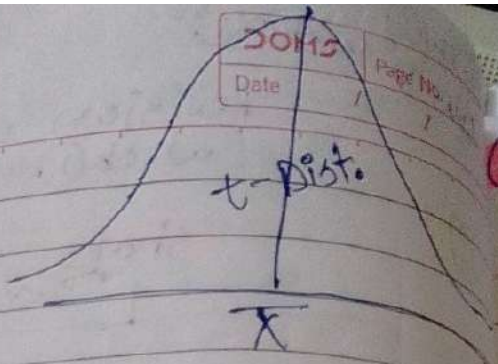
$SE = \frac{\sigma}{\sqrt{n}}$ (population st. dev.) | $\frac{s}{\sqrt{n}}$ (sample st. dev.) | t dist.

standard error वो होता जो μ population mean से दूसरी distance पे होता जैसे ऊपर show किया (SE)

Shot on AWESOME A05s

③





Gaussian Curve

$$Z = \frac{x_i - \mu}{\sigma}$$

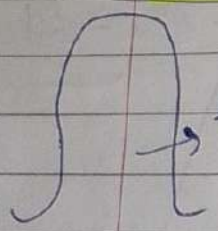
$$[df = n-1]$$

(A) $df = n-1$

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}$$

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

df = Unbiased Estimator.



-3σ $+3\sigma$

Dought $+3\sigma = 5\sigma$

$$SD = \sqrt{\frac{df}{df-2}} = \sqrt{\frac{3}{3-2}} = \sqrt{3} = 1.732$$

EDA Mein हमने देखा था की -3σ, $+3\sigma$ के बाहर Outlier होते हैं। इसलिए हमने $SD = \sqrt{\frac{df}{df-2}}$ लगाकर एक σ sigma value निकाली and एक sigma की value 1.732 आई, इसलिए -5σ, $+5\sigma$ को bad Outlier मानते हैं।

t distribution में

Zee distribution में -3σ, $+3\sigma$

⑤

confidence Interval

α = error

$CL = 1$

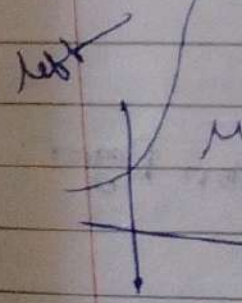
$CI = \bar{X}$

जहाँ $\bar{x}$ हो

जहाँ S हो

तो pendi

⑥



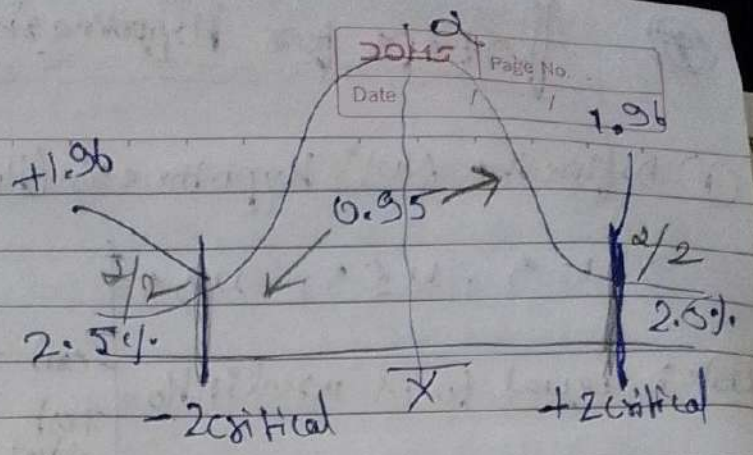
Null Hyp H_0 :

Alt Hyp H_a :

σ = Var

Shot on AWESOME A05s

(5)



confidence Interval

α = error = 0.05

$\boxed{CL = 1-\alpha} = 1-0.05$
$= 95\%$

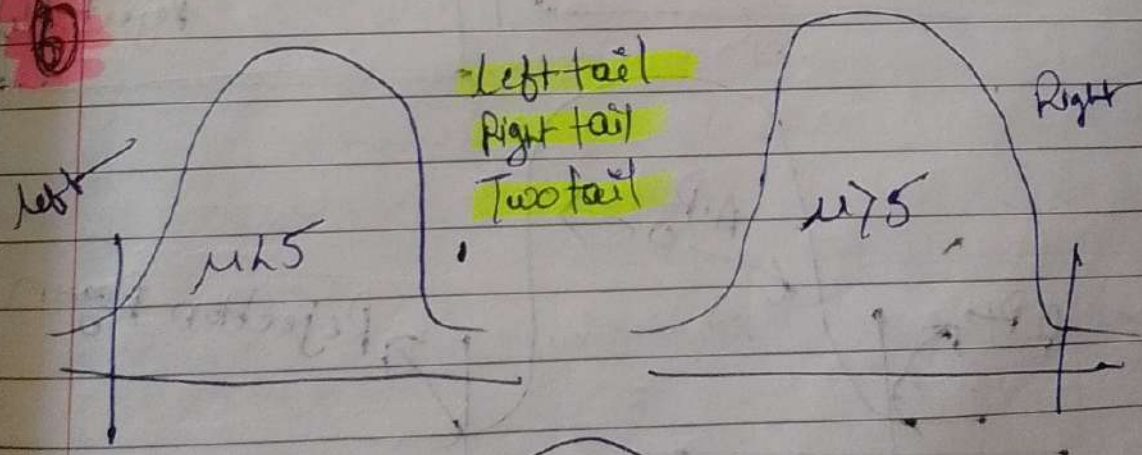
σ - Zcritical/Zscore

s - tcritical

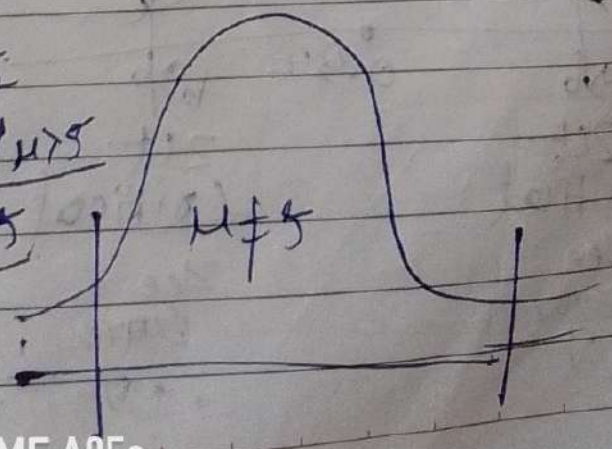
$\boxed{CI = \bar{X} \pm Z \cdot Se}$ → Margin of error

जहाँ σ दिया हो वहाँ Zcritical/Zscore निकलेगा।
जहाँ s tcritical निकलेगा।
तो pencil/pen की line को ही Zcritical/Zvalue बोलते हैं।

(6)



Null Hyp Ho : μ=5
Alt Hyp Ha : μ<5 / μ>5
μ≠5



Shot on AWESOME A05s

Steps for Hypothesis testing

(1) Define the Null Hypothesis (H_0)

$\mu = 5$, $\mu \leq 5$, $\mu \geq 5$

Note: equal to is used in H_0

	(case)
Zcrit	1.96
test stat	3.09

(2) Define Alternate Hypo (H_a)

$\mu \neq 5$, $\mu > 5$, $\mu < 5$

(3) Calculate Critical value

(4) Calculate Test Statistic

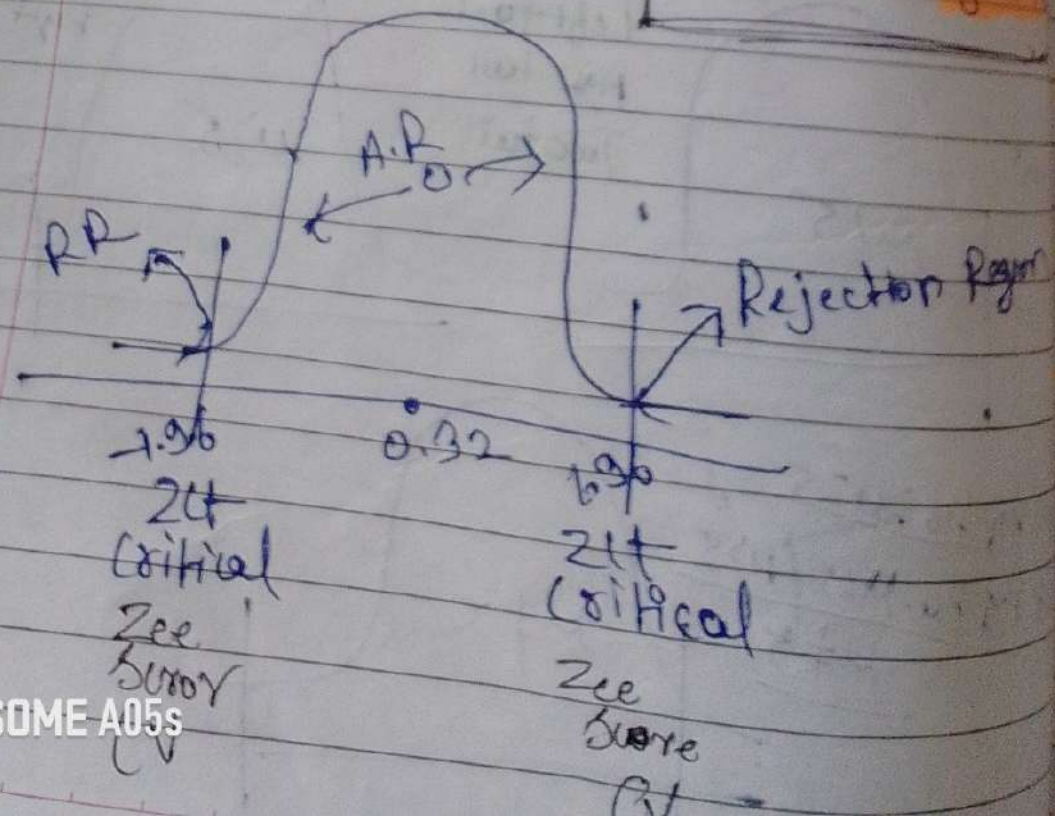
$$Z_{test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = 3.09$$

(5) Draw Inferen

test st > CV → Reject

test st < CV → Fail

Reject



Shot on AWESOME A05s

12

$$\text{t-test formula} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

(Sample SD)

Here,

$$\text{Sample SD } s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Day ② Session ① week ②

(13)

(P value)

$H_0 = =, \geq, \leq$

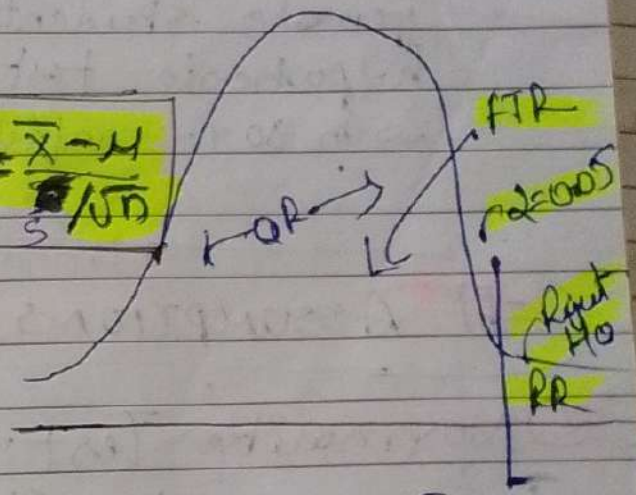
$H_a = >, <, \neq$

$$\text{test stats} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

1 tail test

$CL = 1 - \alpha$

$= 1 - 0.05 = 95\%$



2 tail test

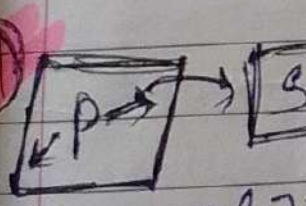
$CL = 1 - (\alpha/2)$

$= 1 - 0.025$

$= 97.5\%$ (AR in b/w)

Doubt

(14)



$\mu \neq 300ml$] Two tail

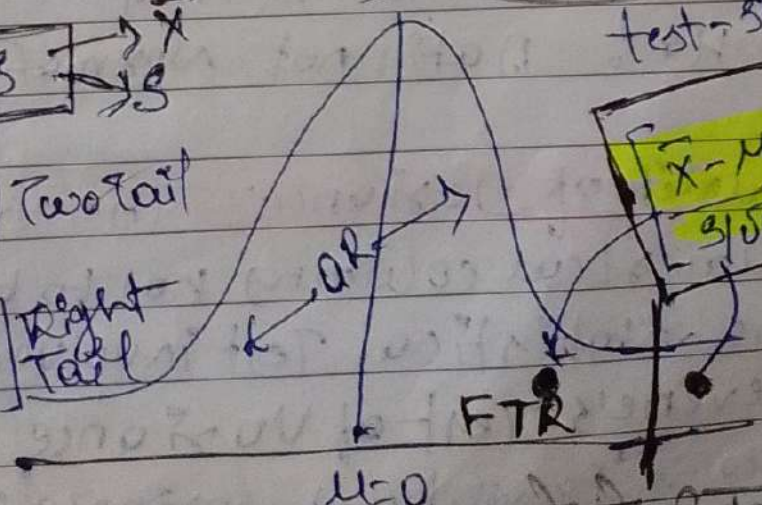
$\mu = 300ml$

$\mu > 300ml$] Right Tail

$\mu < 300ml$

test-statistic

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \text{t-test}$$



test-statistic mean k pass ayega to to utna hi Null hypo. satisfy karega and jitna dur reject Null hypothesis.

Shot on AWESOME A05s

S. No.

totalscore training institute
206 Nature learning
199 Nature learning

$\phi = \{\}$

P value is the probability getting test statistics or extreme or possible such that null hypothesis is being true.

Day 2 week 2 Session 2 (*) data दिया हुआ तो 100% chance H0 को Test लगाना।

Q A survey claims that in a math test female students tend to score fewer marks than the average marks of 75 out of 100. Consider a sample of 24 female students and perform a hypothesis test to check the claim with 90% confidence

ASSUMPTIONS OF HYPOTHESIS TESTING

(1) Normality Test: The should be normal. Here we apply Shapiro walk Test of Normality to check if the Data is Normal

H0 : Data is Normal

Ha : Data is not Normal

(2) Equality of Variance: The variance of Numerical columns has to be equal. The statistical Test involved is called Levene's Test of Variance

H0 : Equality of variance condition is met

Ha : Equality of variance condition is not met

Shot on AWESOME A05s

The test statistic for proportion z-test is given as

$$\boxed{Z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}}$$

where,

p - Sample proportion

n - Sample size.

Under H_0, the test statistic follows a standard normal distribution.

(15) Zprop Test

$\bar{x}$ | Sample prop = $\hat{p}$ $\rightarrow \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$

Population mean $(\mu) \neq$ Hyp. prop = P_0 → Population mean

$\mu = 300ml$

Hyp = 0.10.

t-test = $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ | $\frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$

Shot on AWESOME A05s

← Parametric Tests →

Assumptions — Shapiro, Y, Levene

Data → Y → σ → Z test $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Data → Y → s → t-test $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

Data → A → No

Non Parametric Tests

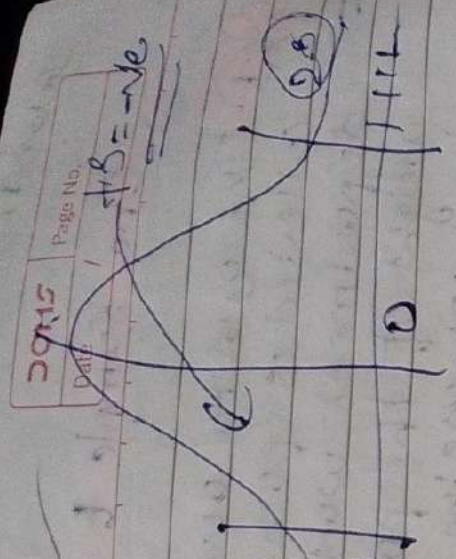
T-Test equivalent → Mann whitney U

Anova equivalent → Kruskal wallis Test (KS Test)

Shapiro N Normality

Levene N EQ of variance

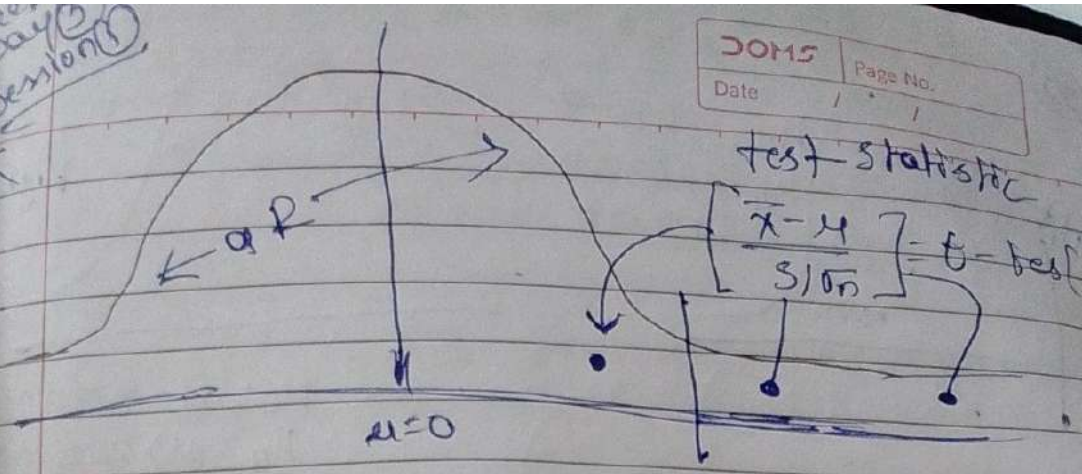
chisQ → chisQ Test



test v/s (critical value

statistic

Pvalue v/s alpha



test statistic

$$\left[\frac{\bar{x}-\mu}{s/\sqrt{n}}\right] = t\text{-test}$$

ONE SAMPLE TEST

σ is given > 30 → Z test → $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

OR

σ is given < 30 → t-test → $\frac{\bar{x}-\mu}{s/\sqrt{n}}$

ttest_1samp (sample data, popmean)

Pvalue < α → Reject Ho

α → Level of significance

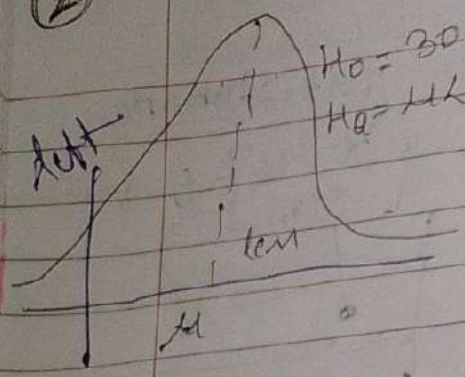
$$CI \Rightarrow \bar{x} \pm \underbrace{Z/t_{critical} \times \text{Standard Error}}_{\text{Margin of error (MOE)}}$$

$$CI \Rightarrow \bar{x} \pm Z/t_{critical} \times \frac{s}{\sqrt{n}}$$

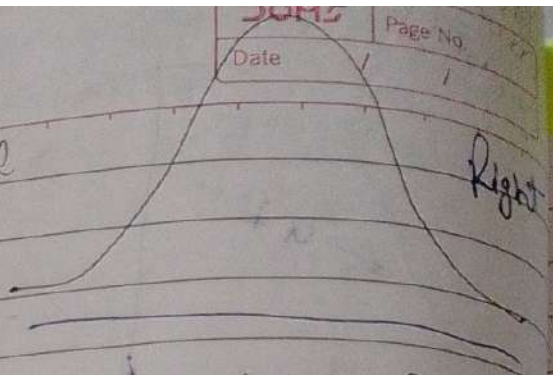
confidence interval का width α पे and Z critical/t critical पे depend करेगा।

Shot on AWESOME A05s

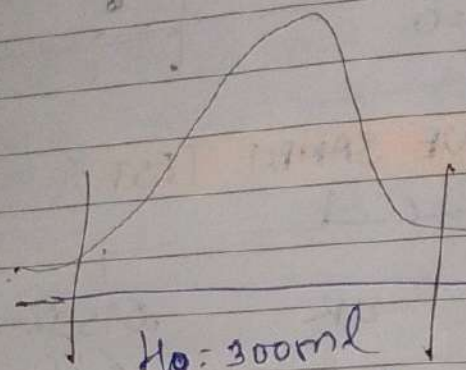
(21)



$H_0 = 300ml$
$H_a = \mu < 300ml$



$H_0 : \mu = 300ml$
$H_a : \mu > 300ml$

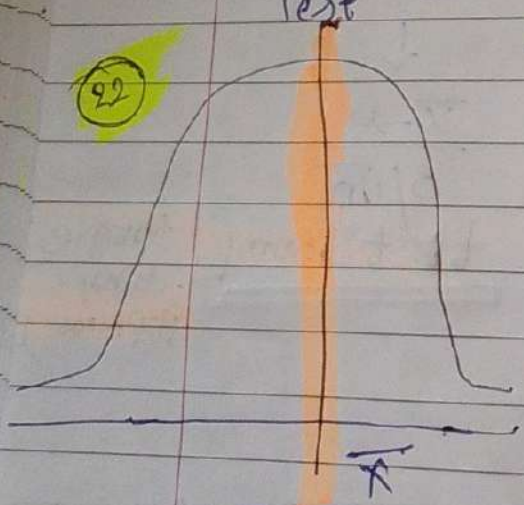


Two tail

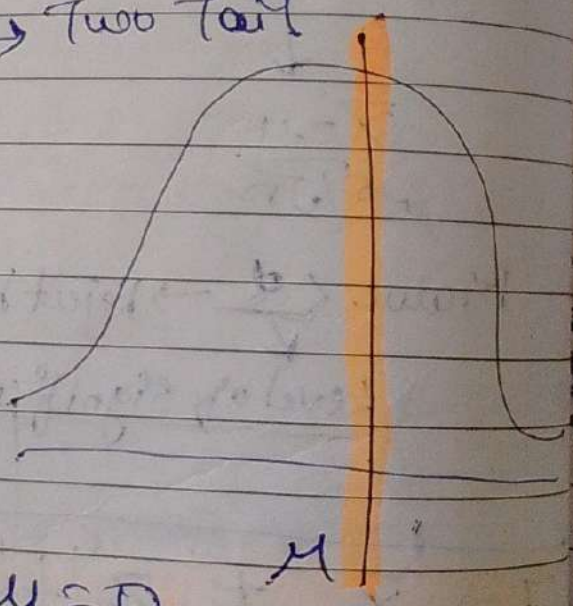
$H_a : 300ml$
$H_a \neq 300ml$

Types of Test → One Tail → Left, Right
Types of Test → Two Tail

(22)



$\bar{x} = \mu$



$\bar{x} - \mu = 0$

अगर $\bar{x}$ & μ दोनों equal हो गए तो $\bar{x} - \mu = 0$ हो जाएगा and t statics भी 0 हो जाएगा and probability 1 होती है

$300 - 300$

$\frac{\bar{x} - \mu}{s/\sqrt{n}} = 0$ ✓

t statics = 0 और 0 पे probability 1 होती है

Two Sample 2 Tes

(24)

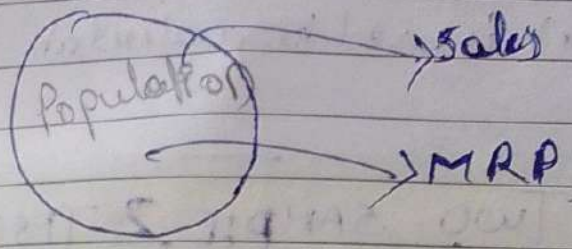
Shot on AWESOME A05s

Two Sample Z test / t test

Big Mart

MRP S_1 v/s Sales $N\mu$ S_2 $N\ 300$

$H_0: \mu_1 = \mu_2 = 300$



$H_0: \mu_1 = \mu_2$

$\hookrightarrow \mu_1 - \mu_2 = 0$

$H_a: \mu_1 < \mu_2$ (left Tail)

$\mu_1 > \mu_2$ (Right Tail)

$\mu_1 \neq \mu_2$ (Two Tail)

Two Sample Z Test

$$\frac{\bar{x}_1 - \bar{x}_2 - \mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{N\ \bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

One Sample Test	Two Sample Z Tests
$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$	$\frac{\bar{x}_1 - \bar{x}_2 - \mu_1 - \mu_2}{\frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}}$

Shot on AWESOME A05s

from statsmodels.stats import weightstats as wstats

wstats.ztest()

direct library for python two sample Ztest.

Week 02 Day 03 Session 02

||LARGE SAMPLE TEST||

If the sample size is sufficiently large (usually $n > 30$) then we use the Z-test. If population standard deviation (σ) is unknown, then the sample standard deviation(s) is used to calculate the test statistic

||TWO SAMPLE Z TEST||

Let us perform a two sample Z-test for the population mean. We compare the means of the two independent populations. The samples are assumed to be taken from populations such that they follow a normal distribution. Also, the sample must have equal variance

The Shapiro-Wilk-Test is used to check the normality of the data. The assumption of equal variances of the population is tested using the Levene's Test. The hypothesis of the levene's test is given as

H_0: The variances are equal.

H_1: The variances are not equal.

The levene() from scipy library performs levene's test

Shot on AWESOME A05s

sample Z-test is given as

$H_0: \mu_1 - \mu_2 = \mu_0$ or $\mu_1 - \mu_2 \geq$ or $\mu_1 - \mu_2 \leq \mu_0$

$H_1: \mu_1 - \mu_2 \neq \mu_0$ or $\mu_1 - \mu_2 < \mu_0$ or $\mu_1 - \mu_2 > \mu_0$

Consider two normally distributed independent populations. Let us take a sample of size (n_1) from the first population with standard deviation (σ_1) and samples ~~are assumed to~~ of size (n_2) from a second population with standard deviation (σ_2) such that $n_1, n_2 > 30$

The test statistic for two sample Z-test is given as

$$\boxed{Z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}}}$$

where

$\bar{x}_1, \bar{x}_2$: Mean of both the samples.

μ_0 : Mean difference given in the null hypothesis

σ_1, σ_2 : Standard deviation of both populations

n_1, n_2 : Size of samples from both the population

Under H_0 the test statistic follows a standard normal distribution

If standard deviations of populations are unknown, use the standard deviations of samples (s_1, s_2) instead of σ_1 and σ_2 to calculate the test statistic

$$\boxed{Z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}}$$

Shot on AWESOME A05s

Shapiro and levene are always applied on data

Day (3) session (3) Day 8

(27) If $n > 30$ or more ~ Z test ~ is given

If $n < 30$ ~ T-Test ~ Sample SD is given

Two Sample t-test small — [8]

$$\text{test statistic} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p \times \sqrt{1/n_1 + 1/n_2}}$$

pooled standard deviation

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 - S_2^2(n_2-1)}{(n_1+n_2-2)}}$$

(28)

$$S_1 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

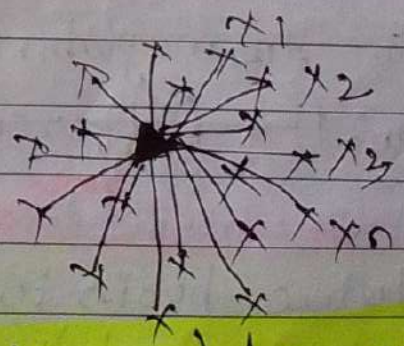
$$S_1^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

pooled std. dev.

$$\sum (x_i + \bar{x})^2 = S_1^2 (n-1)$$

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



$(n_1 + n_2 - 2)$ degrees of freedom

Two sample Z test

T test

let solve using python Function.

from statsmodels.stats import weightstats as ws

ws.ztest(x1 = n1, x2 = sg1, alternative = "larger")

(0.15125511253494914, 0.4309887298408498025)

Paired t-Test

A paired t-test is used to compare the mean of the population for two dependent samples. The dependant samples can be the scores before and after a specific treatment.

let X_i be the sample before the treatment and Y_i be the sample after the treatment. Let μ_x, μ_y be the mean of the data X and Y respectively. The mean difference $\mu_d = \mu_y - \mu_x$

The null and alternative hypothesis is given as

$H_0 : \mu_d = \mu_0$ or $\mu_d \geq \mu_0$ or $\mu_d \leq \mu_0$

$H_1 : \mu_d \neq \mu_0$ or $\mu_d < \mu_0$ or $\mu_d > \mu_0$

The test statistic for paired t-test is given as

$$t = \frac{\bar{X}_D - \mu_0}{\frac{S_D}{\sqrt{n}}}$$

where,

$\bar{X}_D$ = Mean difference b/w the pairs

μ_0 = Mean difference given in the null hypothesis

S_D = Standard Deviation of differences between the pairs.

n : Sample size

Under H_0, the test statistic follows a distribution with $(n-1)$ degrees of freedom

Shot on AWESOME A05s

F-Critical

(category) dfn $(n-1)$ — dfd (data) $(N-k)$

n = no. of categories
dfn
$3-1 = 2$

N = Total no. of data points
k = no. of category
$N-k$
$9-3 = 6$

F distribution table

dfd \ dfn	0.01	0.025	0.05
6	0.05	5.14	

F critical
www.sjsu.edu

cat ①	cat ②	cat ③
1	2	2
2	4	3
5	2	4

(4) calulate Test Statistic

(i) calulate the Mean of data.

$$= \frac{1+2+5+2+4+2+2+3+4}{9}$$

$$\bar{X} = \boxed{2.77}$$

(ii) calulate the $\bar{X}$ of all of categories

$$\bar{X}_{cat1} = \frac{1+2+5}{3} = 8/3 = 2.66$$

$$\bar{X}_{cat2} = \frac{2+4+2}{3} = 8/3 = 2.66$$

$$\bar{X}_{cat3} = \frac{2+3+4}{3} = 9/3 = 3.00$$

Shot on AWESOME A05s

(iii) calulate sum of squares total

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \Rightarrow$$

$[(1-2.77)^2 + (2-2.77)^2 + (5-2.77)^2 +]$ cat ①

$[(2-2.77)^2 + (4-2.77)^2 + (2-2.77)^2 +]$ cat ②

$[(2-2.77)^2 + (3-2.77)^2 + (4-2.77)^2]$ cat ③

$\boxed{13.56}$

(iv) Calulate the sum of squares category wise

cat① $\sum (x_i - \bar{x}_1)^2 = (1-2.66)^2 + (2-2.66)^2 + (5-2.66)^2$
$\boxed{8.67}$

cat② $\sum (x_i - \bar{x})^2 = (2-2.66)^2 + (4-2.66)^2 + (2-2.66)^2$
$\boxed{2.67}$

cat③ $\sum (x_i - \bar{x})^2 = (2-3)^2 + (3-3)^2 + (4-3)^2$
$= \boxed{2.00}$

$8.67 + 2.67 + 2.00 = \boxed{13.34}$ ✓

SS within category.

$\boxed{13.56}$ ✓

SS within Total data points

(v) Sum Square between the Groups.

SS Total (datapoints) − SS within (category)

$13.56 - 13.34$

$SS_b = 0.22$ ✓

sum Square between the groups

Shot on AWESOME A05s

(vi) F Test statistic

Page :
Date :

$$\boxed{\text{F Test Statistic} = \frac{\text{Mean Square between}}{\text{Mean square within}}}$$

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

$$\boxed{\text{Mean Square between } MS_{bet} = \frac{\text{Sum of Square between}}{dfn}}$$

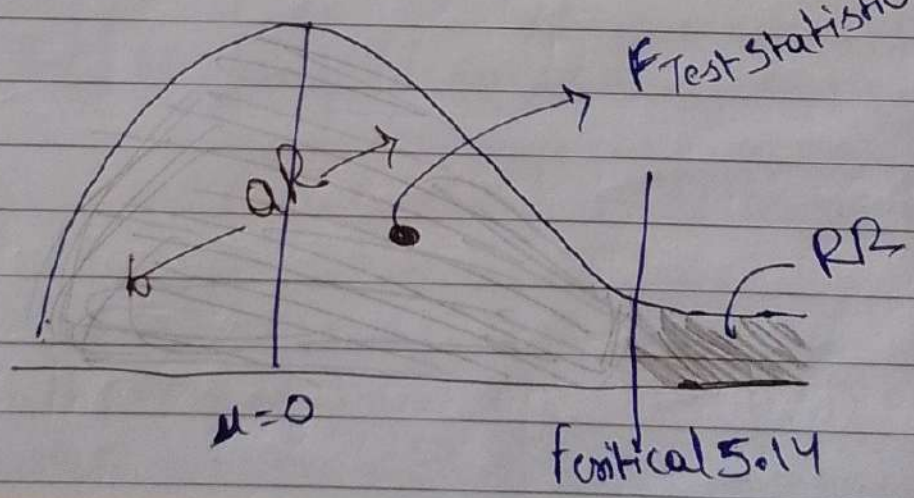
$$\boxed{\frac{SS_{betw}}{dfn}}$$

Here, dfn = degree of freedom for categories.

$$MS_b = \frac{0.22}{2} = \boxed{0.11}$$

$$MS_w = \frac{SSW}{dfd} = \boxed{\frac{13.34}{6}} = \cancel{0.049} \quad 2.22$$

$$\boxed{F_{test}} = \frac{0.11}{2.22} = \boxed{0.049}$$



F test statistic < F critical
→ Fail to Reject Ho

Note As the Sample size increases F - distribution becomes Normal distribution

Shot on AWESOME A05s

Anova more than 2 categories

cat v/s Num

ANOVA

one category with 3 levels cat - cat1 cat2 cat3

One way Anova — 1 cat var. — 1 Num var.

Two way Anova — 2 cat — Type ① Type ② Outlet Type — 1 Num — Sales

cat	No
cat1	2
cat2	4
cat3	3

1 cat Var with 3 Levels [cat1, ..., cat3]

Q① Total marks in aptitude exam are ~~conducted~~ recorded for students with different race/ethnicity. Test whether all the races/ethnicities have an equal average score with 0.05 level of significance. (one-way anova)

```
# read the students performance data

df_student = pd.read_csv('studentPerformance.cs

# display the first two observations
df_student.head(2)
```

	gender	race/ethnicity	lunch	testPrep.course	mathscore	readin
0	female	groupB	standard	none	89	55
1	female	groupC	standard	completed	55	63

~~gender~~	writingscore	Totalscore	training institute
	56	200	Nature Learning
	72	190	Nature Learning

Shot on AWESOME A05s

One-way ANOVA

It is used to check the equality of population means for more than two independent samples. Each group is considered as a treatment. It assumes that the samples are taken from normally disturbed populations. To check this assumptions. we can use the shapiro-Wilk-Test Also, the population variances should be equal, this can be tested using using the heveneis Test

The null and alternative Hypothesis is given as

H_0: The averages of all treatments are the same
H_1: At least one treatment has a different average

Consider there are t treatments and N number of total observations. The test statistic is given as

$$F = \frac{MTrSS}{MESS}$$

where,

$$MTrSS = \frac{TrSS}{dfTr}$$

$$TrSS = \sum_i^t \sum_j^m n_i (\bar{x}_{i.} - \bar{x}_{..})$$

n_i is the nuber of observations in i^{th} treatment
$\bar{x}_{i.}$ is the mean over i^{th} treatment
$\bar{x}_{..}$ is the grand mean (i.e. mean of all the observatio

dfTr is the degrees of freedom for treatments (= t-1

$$MESS = \frac{ESS}{dfe}$$

Shot on AWESOME A05s

$ESS = \sum_i \sum_j (a_{ij} - \bar{a}_{i.})$

dfe is the degrees of freedom of error (= N-t)

Under H_0, the test statistic follows F-distribution with (t-1, N-t) degrees of freedom.

Difference between One way Anova and Two way Anova

1. One way Anova is when you introduce One categorical variable and A Numerical Variable

2. On the other hand, if we want to analysis multiple categorical variables with a numerical ~~value~~ variable, then it is called Two way Anova

3. For example, Pclass Vs Fare is an example of One way Annova with One categorical variable with 03 Levels (Class 1, Class 2 & 3) and a Numerical variable

4. whereas Two-way Anova is two or more categorical variables with a numerical variables.

Shot on AWESOME A05s

Kruskall Wallis Test

- If one of the assumptions of one-way ANOVA is not stisfied, then we can perform the krushal-Wallis H-test which is a non parametric equivalent test for ANOVA

- Wilcoxon Signed-Ranks test is the non-parametric version of Two Sample Related T-Test

- If the Data is not following Normal Distribution, then one can use Mann whitney U test inplace of Independent Two Sample T-Test. It is also known as Mann whitney wilcoxon Test or the wilcoxon Rank Sum Test.

Q1. A gym trainer wants to provide an energy bar to all his customers to increase muscle strength. Three different companies approached the gym trainer with their high-quality energy bars. The trainer collects an amount of calcium(ing) in the energy bar from three companies and he wants to study whether all the bars have an equal amount of calcium on average. Test at 5% level of significance.
given data:

- alpha_bar = [24.4, 20.7, 56.9, 19.5]
- beta_bar = [53.2, 54.7, 20.5, 15.8, 56.6]
- gamma_bar = [54, 31, 22.8, 24.7]
- Ho: That All the energy Bars have Same Quantity of calcium
- H1: That All the Energy Bars Do not have Same Quantity of Calcium.

Shot on AWESOME A05s

* Non Parametric are performed on Median
* Parametric Tests are performed on Mean.

Date :

Post-hoc Analysis

If one way ANOVA rejects the null hypothesis; we conclude that at least one treatment has a different mean. The test does not distinguish a treatment with the different average ~~rate~~ value. The post-hoc test or multi comparison test is used to identify such treatment(s).

In this section, we study the Tukey's HSD test. The test calculate the mean difference for each pair of treatments and returns a pair(s) with different average

The test statistic of Tukey's HSD test is given as.

$$T_\alpha = q_\alpha(t, f)\sqrt{\frac{MSE}{n}}$$

The value of $q_\alpha(t, f)$ is obtained from the tukey table.

t: Number of treatments
f: degrees of freedom for error (dfe)
MSE: Mean error sum of squares $\left(= \frac{ESS}{dfe} = \sum_i^t \sum_j^{n_i} (x_{ij} - \bar{x}_{i.})^2\right)$
n: Number of observations in a treatment

This test is efficient when the sample size for each treatment is equal. If the sample size is not equal to each treatment then we can use the Scheffe test. The scikit_posthocs. posthoc_scheffe()

Shot on AWESOME A05s

Day 5 Session (1)

Z proportion Test Two Sample

Page :
Date :

Q1.

Perform Two sample z test for the population proportion. we check the equality of population proportion P_1 and P_2

The null and alternative hypothesis given as.

$H_0 : P_1 - P_2 = P_0$ or $P_1 - P_2 \geq P_0$ or $P_1 - P_2 \leq P_0$

$H_1 : P_1 - P_2 \neq P_0$ or $P_1 - P_2 < P_0$ or $P_1 - P_2 > P_0$

The test statistic for two sample proportion z-test is given as

$$Z = \frac{(P_1 - P_2) - P_0}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\bar{P} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

where,

P_1, P_2 = Sample proportions
P_0 = Hypothesized Proportions
$\bar{P}$ = Proportion of pooled sample
n_1, n_2 = Samples sizes.

$$\frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \Longrightarrow \frac{(\tilde{P}_1 - \tilde{P}_2) - P_0}{SE} \rightarrow \mu_1 - \mu_2 = 0$$

$$SE \rightarrow \sqrt{P_0(1-P_0) \times \frac{1}{n_1} + \frac{1}{n_2}}$$

$n_1 \{ \hat{P}_1 =$

$n_2 \{ \hat{P}_2 =$

$P_0 = ?$

Shot on AWESOME A05s

Chi-Square Test

It is a non-parametric test. Non-parametric tests do not require any assumptions on the parameter of the populations from which the sample is taken. These tests can be applied to the ordinal/nominal data. A non-parametric test can be performed on the data containing outliers.

The chi-square test statistic follows a Chi-Square (χ^2) distribution under the null hypothesis. It can be used to check the relationship between the categorical variable

Chi Square Test for Goodness of Fit

This test is used to compare the distribution of the categorical data with the expected distribution.

H0: There is no significant difference between the observed and expected frequencies from the expected distribution

H1: There is a significant difference between the observed and expected frequencies from the expected distribution

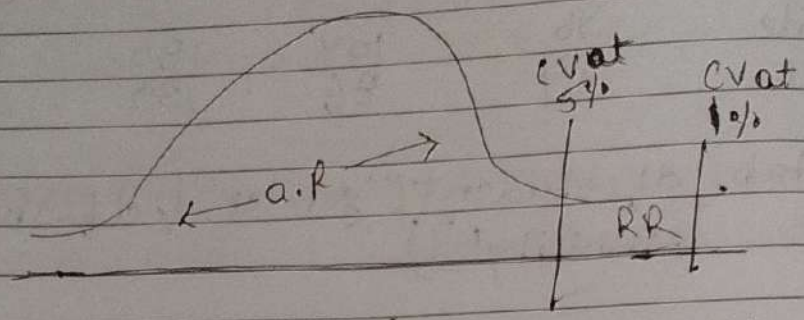
The test statistic is given as:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - N$$

Shot on AWESOME A05s

Type① Error
Type② error

Errors in Hypothesis testing

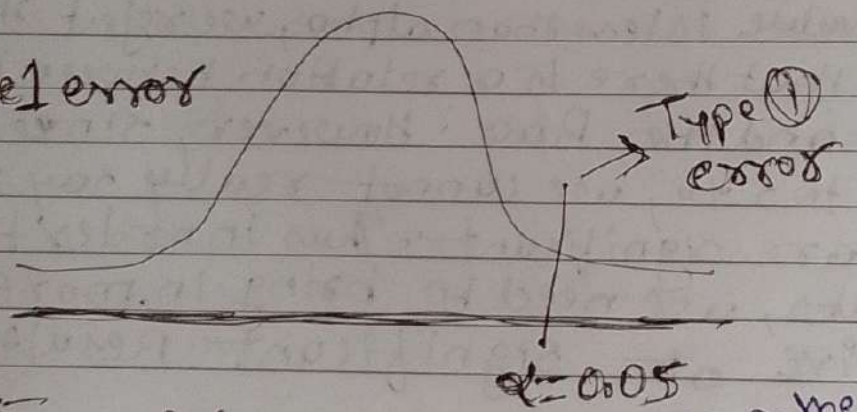


test statistics > Critical Value ⇒ Reject Ho

test statistics < Critical Value ⇒ Fail to Reject

$\alpha = 0.05 - 0.01$ } FTR
Pvalue = 0.04

$\alpha \sim$ Type1 error



$\alpha = 0.05$
Pvalue = 0.04 Rej.

$\alpha = 0.01$
Pvalue = 0.04
FTR

अगर Null hypothesis सही है, अगर result में आए की Null hypothesis reject Ho जो = Type① error

Rejecting Null hypothesis when it is True is called Type① error

Ex Ho = Fire, alarm,
Ha = No Fire, No alarm
Error = No Fire, alarm

Type⑪

Failing

Shot on AWESOME A05s

error (1) - Fire, No

Q. which error is more harmful error (1) and error (2)

Ans. Both errors are harmful

error (1) and error (2) examples

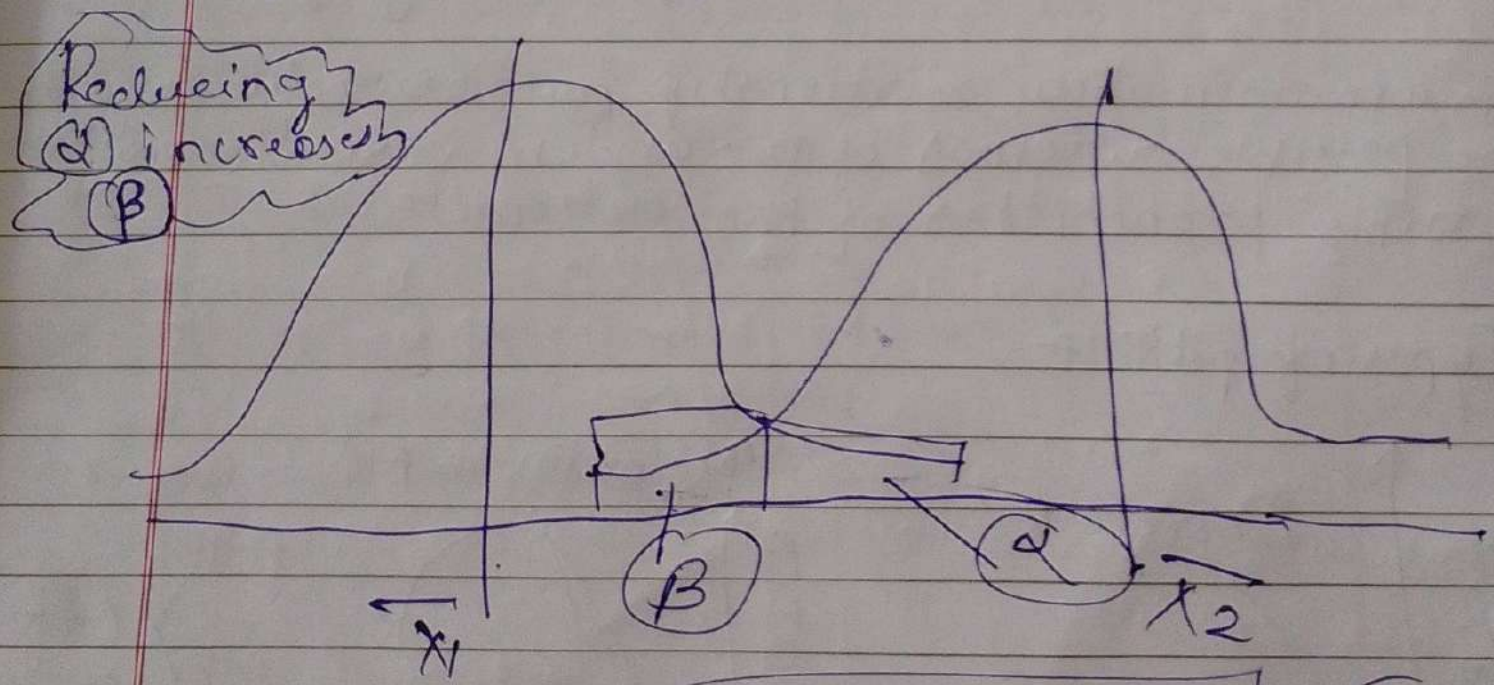
error (2)

Failing / Fail to reject an incorrect hypothesis..

Type (1) error False positive (α)

Type (2) error False Negative (β)

Type (1) error = confidence interval/range बढ़ा दो तो, Type (1) का nikal jaega.



$1-\alpha$ = confidence

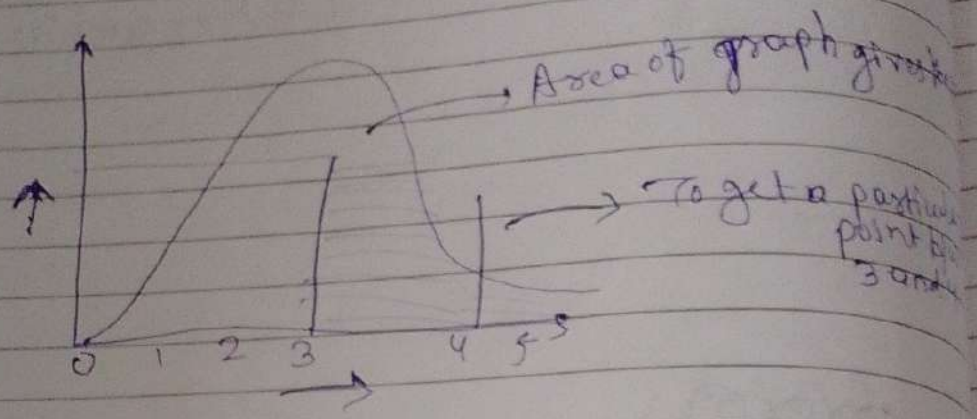
power of Test $= 1-\beta$

Shot on AWESOME A05s

P-Density function (PDf)

It is a mathematical func^n that describes the prob. of continuous Random variable.

(eg) of pdf



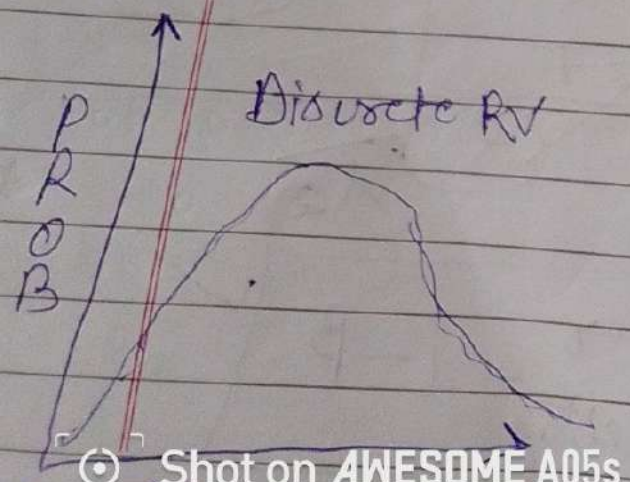
Probability distribution

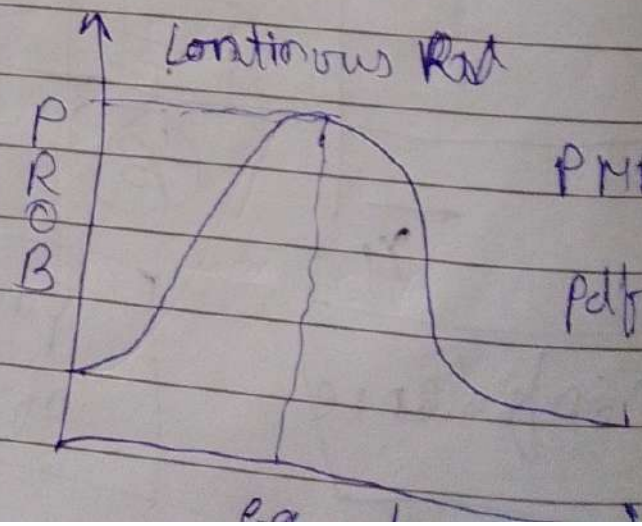
A probability distribution is a list of all of the possible outcomes of a random variable along with that corresponding probability values.

Probability distribution function = $y = f(x)$

It is a mathematical relationship between the possible outcomes of a random variable with possible probabilities of the outcomes

Types of pdf -





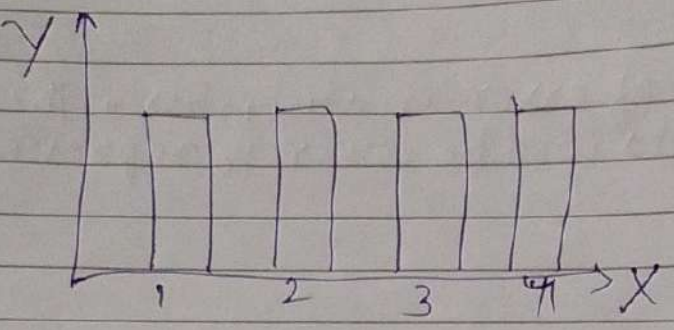
eg. - Die

eg.

Shot on AWESOME A05s

pmf

It is a mathematical func^n that describes ~~discrete Random variable~~. The pmf of a DRV assign a probability to each possible value of a random variable.



Two Conditions of PMF

① The probability assigned to each value must be non-negative ($>=0$)

② The sum of probabilities assigned to all variables values must equal to 1 | $f(x=6) = 1/6$

eg. Dice

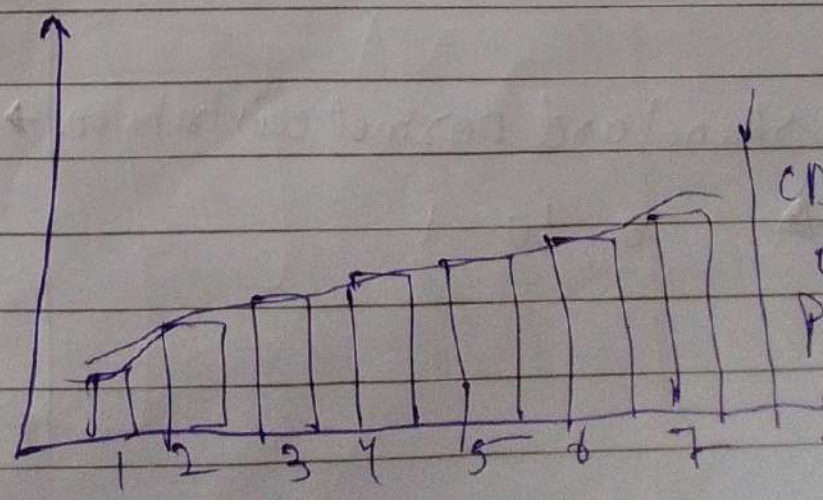
~~Cond~~. CDF of PMF = describes the probabilities that random variable x with a probability dist. will be found of a value less or equal to x

eg. $f(x \leq 4) = f(x=4) + f(x=3) + f(x=2) \ldots$

It will go to 1 max

pmf = gives prob. of a particular point

cdf = gives a prob. of all the points up to x.



Shot on AWESOME A05s