31/10/25

Measure of dispersion:

Variance:

Ex:

	person 1	person 2
Mon	7:00 AM	8:00Am
Tue	7:15 Am	9:00 Am
Wed	7:30 Am	11:00 Am
Thur	7:45 Am	7:00 Am
fri	8:00 Am	10:0m
Sat	? 7:00 Am - 8:00 Am.	? 9 - 10.
	7:15 - 15 mm 7:30 - 30 m 7:45 - 15m	7Am - 1 hr 11Am - 1 hr

Variance high prediction are low.

1. Variance
2. Standard Deviation.
3. Range

Variance:

1. population variance (σ^2)

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

2. Simple variance (s^2)

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

"n-1" We call it as Degree of freedom.

Ex:

Calculate variance on $\{1, 2, 2, 3, 4, 5\}$.

$$\mu = \frac{1+2+2+3+4+5}{6}$$

$$\mu = 2.8.$$

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

$$= \frac{\sum_{i=1}^{6} (x_i - \mu)^2}{6}$$

$$= \frac{(1-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (5-2.8)^2}{6}$$

$$= \frac{3.36 + 0.69 + 0.69 + 0.02 + 1.36 + 4.69}{6}$$

$$= \frac{10.8}{6} = 6.00$$

$$\sqrt{\sigma^2} = \sqrt{1.8\ km^2}$$

$\sigma^2 = 1.8\ km^2$ (variance)

$\sigma = 1.34$ km (Standard deviation)

Standard Deviation = $\sqrt{\text{variance}}$.

Sample SD

population SD

$\sigma = \sqrt{\sigma^2}$

$\sigma = \sqrt{1.8}$

$\sigma = 1.34$

$S = \sqrt{S^2}$

$S = \sqrt{2.16}$

$S = 1.46.$

Range = Maximum − Minimum

= 5 − 1

Range = 4.

percentaile and Quartile :

percentaile is the value below which certain percent of observation will ~~learn~~ lie.

Ex :

= { 1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8 }.

How much % of data will come below 6 ?

*) percentile rank of $x = \frac{\#\text{ no of value below}}{N} \times 100.$

$$= \frac{7}{11} \times 100$$

$$= 63.63$$

Observation of data value is <6.

Quartile :

Quartile helps to find the value which is present at the given percentile Rank.

Ex :

=

{ 1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8 }

Which value is present at 25% ?

$$\text{value} = \frac{\text{percentile}}{100} \times n+1$$

$$= \frac{25}{100} \times 12$$

$$= 3. \rightarrow \text{Index}$$

Value = 2

Calculating for 90%.

$= \frac{90}{100} \times 12$

$= 10.8 = 10. \rightarrow$ This value called Index.

$= 7$

Five number Summary:

1. Minimum
2. First Quartile (Q_1) 25%
3. Median(Q_2) 50%
4. Third Quartile (Q_3) 75%
5. Maximum.

Note: choose these 5 numbers after removing the outliers from the data by finding boundary values.

[Lower fence -4.5 upper fence 15.5]

Anything < -4.5 & > 15.5 is a

$LF = Q_1 - 1.5(IQR)$

$UF = Q_3 + 1.5(IQR)$

IQR (Inter Quartile Range)

$IQR = Q_3 - Q_1$

{1, 1, 2, 3, 4, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9, 28, 36} = 17

$Q_1 = \frac{25}{100} \times 18$

$= 4.5 = \text{Index.}$

$Q_1 = 3$

$Q_3 = \frac{75}{100} \times 18$

$= 13.5 = \text{Index}$

$Q_3 = 8$

$IQR = Q_3 - Q_1$

$= 8 - 3$

$IQR = 5$

$LF = 3 - 1.5(5)$

$= -4.5$

$UF = 8 + 1.5(5)$

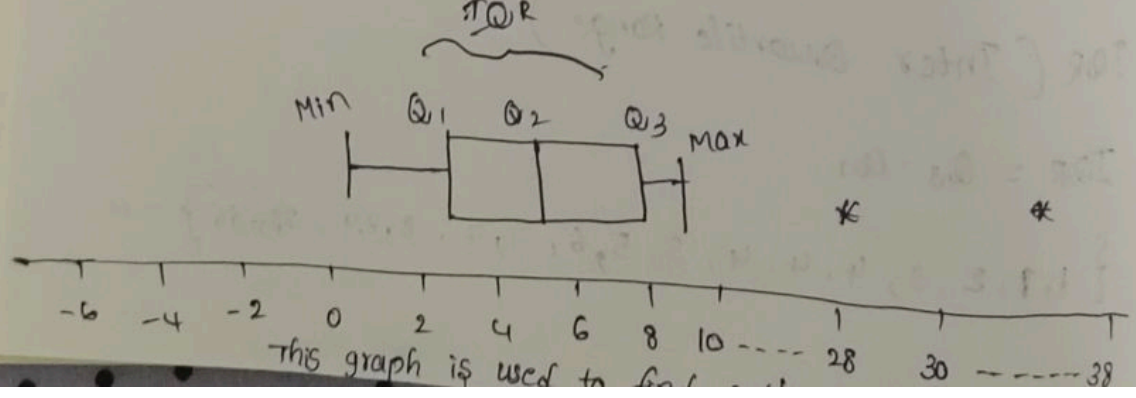
$= 15.5$

Minimum = 1, Median = 5, Maximum = 9

$Q_1 = 3$

$Q_3 = 8$

Box plot :



This graph is used to

3/11/25

Different types of Distributions:

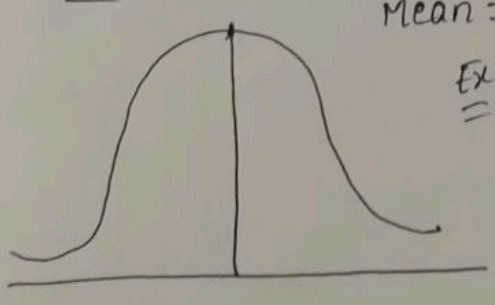
1. To understand Data patterns.
2. To Summarize the data easily.
3. To calculate the probabilities.
4. To make prediction and Decision.
5. To choose right statistical test.

There are two category of distribution.

1. Continuous Distribution. (Numerical Distribution)
2. Descrete Distribution. (Categorical Distribution).

1. Normal Distribution. } Continous Distribution
2. Standard normal Distribution. }
3. Bernoulli Distribution. } Categorical Distribution
4. Binomial Distribution. }
5. poision Distribution. }

Normal Distribution:

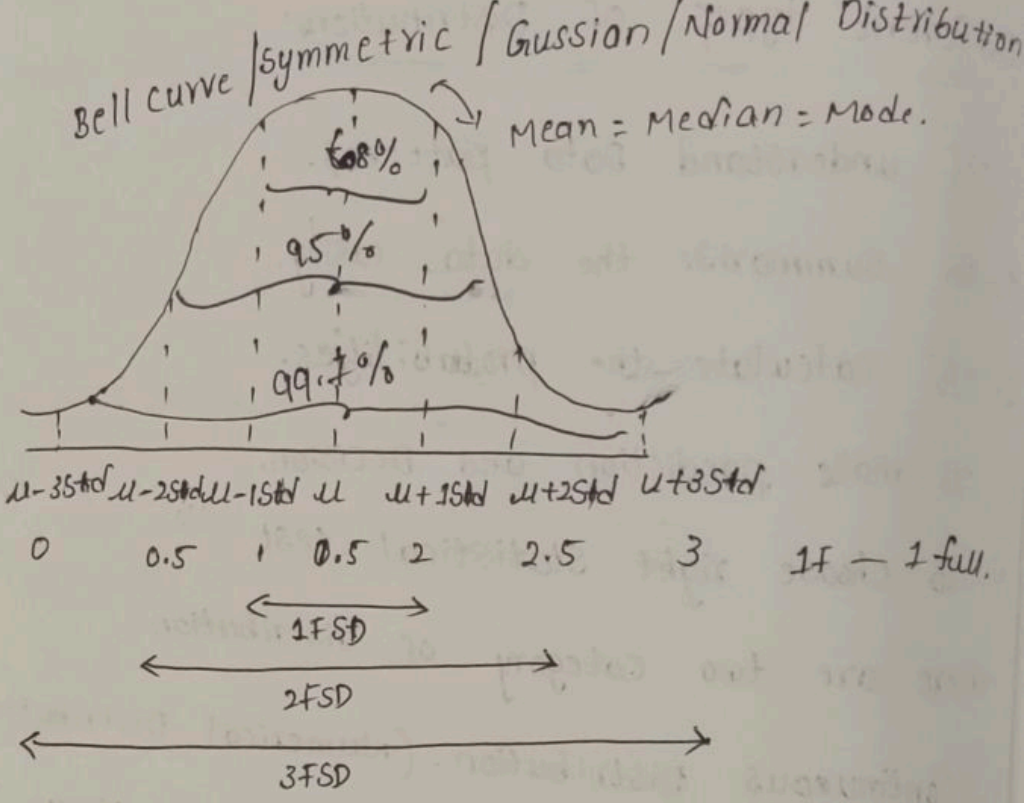


Mean = Median = Mode.

Ex!

1. Temp
2. Speed
3. Strength level

The Normal Distribution follows Rules of emphenical.



→ 68% - 95% - 99.7% called as Confidence Intervals.

Emphenical Rules:

68% of data will present in 1SD.

95% of data will present in 2SD.

99.7% of data will present in 3SD.

Standard Normal Distribution (SND):

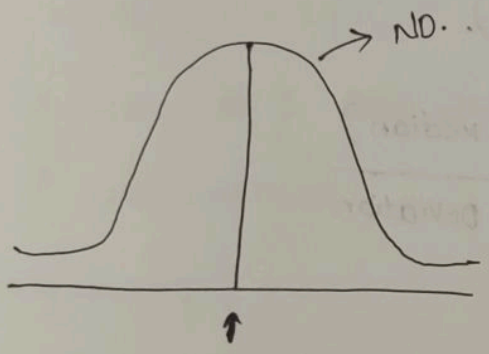
$\mu = 0$

SD = 1

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

Normal Distribution data	Standard Normal Distribution data.
2	-0.9
7	1.57
5	0.57
4	0.07
1	-1.43
3	-0.43
5	0.57
$\mu = 3.86$	$\mu = 0.02$
$\sigma = 2.$	$\sigma = 0.93 = 1$

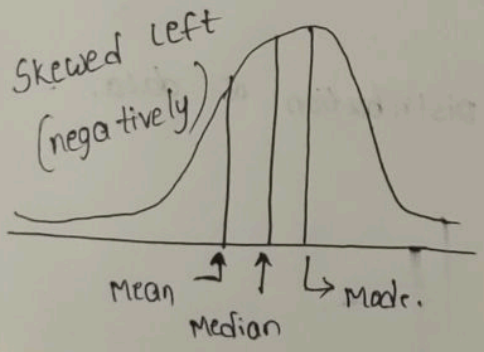
Normal Distribution:

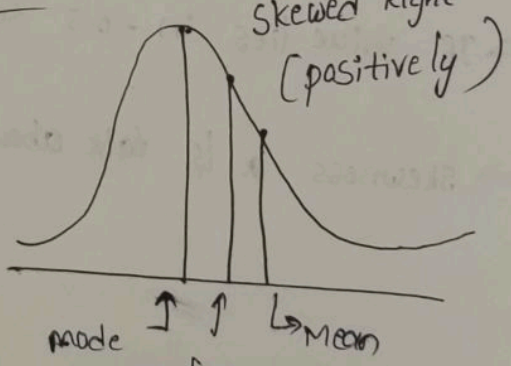


Mean = Median = Mode.

Symmetric

Skewed Data





4/11/25

Different types of Skews:

1. positive Skew (Right side):

Tail on the right side is longer most data are on the left.

2. Negative Skew (Left side):

Tail on the left side is longer most data are on the right.

3. Zero Skew (Symmetric):

The data is evenly distributed around the mean (like a normal distribution).

$$\text{Skewness} = \frac{3(\text{mean} - \text{Median})}{\text{Standard Deviation}}$$

It is only for Individual columns

→ If value near to -1 then it is -ve. Skew.

→ If value near to +1 then it is +ve Skew.

→ If value lies in -0.5 to 0.5 then it is Zero Skew.

→ Skewness is talk about Distribution of data.

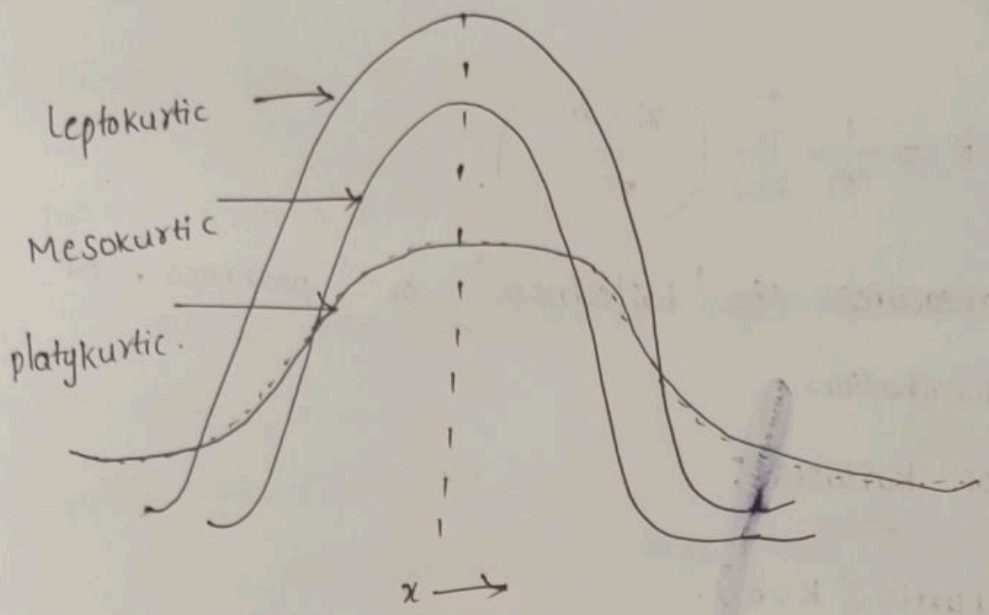
Kurtosis :

$$K = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right).$$

It measures the "tailedness" or "peakness" of data distribution.

Types of kurtosis :

1. Meso kurtic ($K = 3$).
   * Normal Distribution.
   * No outliers.
   * Moderate tail and peak.
2. Lepo kurtic ($K > 3$).
   * Heavy tails and sharp peak.
   * More outliers.
3. platy kurtic ($K < 3$).
   * Light tails and flat peak.
   * Fewer outliers.



Discrete Distribution:

Bernouli Distribution :

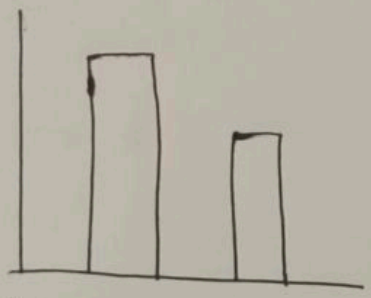
It is simplest form of discrete probability distribution. and models a random experiment with exactly 2 outcomes.

Success denoted by = p

failure denoted by = 1-p.

→ Total probability is 1.

→ The range of probability is 0-1.



$P(T) = \frac{1}{2} = 0.5$

$P(H) = \frac{1}{2} = 0.5$

Binomial Distribution :

It generalizes the Bernoulli Distribution to multiple trails. It models the number of success in a fixed number of independent and identical Bernoulli trails.

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

k = no. of success

n = no. of trails.

p = probability of success.

poision Distribution :

It is used to model the number of events that occur in a fixed time interval or space and occur independently the parameter λ represents the avg number of event in the Interval.

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

k! = factorial of k.

λ = Avg no. of events.

k = No. of (Arguments.) occurences.

e = Euler's number.

Inferential Statistics :

1. probability :

It Measures likelyhood of an event.

Eg : Dice = { 1, 2, 3, 4, 5, 6 }.

$$P(x) = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes.}}$$

$$P(3) = \frac{1}{6}$$

$$P(2,4,5) = P(2) + P(4) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$= 0.5$$

Toss the 2 coins.

{ HH, HT, TH, TT }.

1) What is the probability of getting only 1 H ?

A) $= \frac{2}{4_2} = \frac{1}{2}$.

2) What is the probability of getting both Tails ?

$\frac{1}{4}$

There are 2 rules in probability :

1) Addition Rule (OR)

2) Multiplication Rule (AND)

Addition Rule :

1. Mutual Exclusive Events.

2. Non Mutual Exclusive Event.

5/11/25

Mutual Exclusive Event :

The different events won't occur at the same time is called Mutual Exclusive Event.

Ex: If you toss the coin what is the probability of landing on heads or tails.

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(H \text{ or } T) = P(H) + P(T)$$

$$P(H \text{ or } T) = \frac{1}{2} + \frac{1}{2} = 1.$$

Non Mutual Exclusive Event :

Here multiple events can occur at the same time is called Non Mutual Exclusive Event

Ex: picking the cards from deck of cards

What is the probability of getting Jack or heart?

$$P(J \text{ or } H) = P(J) + P(H) - P(J \cap H)$$

$$P(J \text{ or } H) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$P(J \text{ or } H) = 0.07 + 0.25 - 0.01$$

$$P(J \text{ or } H) = 0.33 = 0.31.$$

$$P(J \text{ or } H)$$

Multiplication Rule:

1. Independent Event:

Here all the values have the same priority after n. number trails also (or) 1 Event don't depend on another event.

1st toss the coin

$$P(H) = \frac{1}{2}$$

7th toss the coin

$$P(H) = \frac{1}{2}$$

Ex: What is the probability of Dice rolling and getting a 5 and then 4?

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

$p(5 \text{ and } 4) = p(5) * p(4)$

$= \frac{1}{6} * \frac{1}{6}$

$= \frac{1}{36} = 0.027$

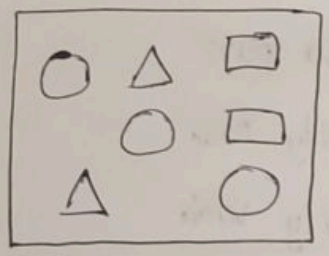
2.7 % is getting.

2) Dependent Event :

present event depend on the previous event.

1st time

$p(\bigcirc) = \frac{3}{7} = 0.43.$



2nd time

$p(\triangle) = \frac{2}{6} = 0.3$

3rd time

$p(\square) = \frac{2}{5} = 0.4$

4th time

$p(\bigcirc) = \frac{2}{4} = 0.5$

Ex: From a deck of cards what is the probability of getting a king and then 8 ?

$p(A \text{ and } B) = p(A) * p(B/A)$

$p(k \text{ and } 8) = \frac{4}{52} * \frac{4}{51} = 0.07 * 0.07$

$p(k \text{ and } 8) = 0.07 = 7\%$ chance.

permutation and Combination :

permutation means the value should be repeated but order shouldn't be repeated.

Ex: Dosa, Idly, vada, puri

1. Dosa Idly
2. Dosa vada
3. Dosa puri
4. Idly Dosa
5. Idly vada
6. Idly puri
7. vada Dosa
8. vada Idly
9. vada puri
10. puri Dosa
11. puri Idly
12. puri vada.

$$n_{P_r} = \frac{n!}{(n-1)!} = \frac{4!}{(4-2)!} = \frac{4!}{2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 12$$

n = no. of items r = no. of ways.

→ permutation refers to the different way in which a set of items can be arranged in order and in permutation the orders of the items matters but not items.

Combination :

{ Dosa, Idly, vada, puri }

1. Dosa Idly
2. Dosa Vada
3. Dosa puri
4. Idly Vada
5. Idly puri
6. Vada puri

$$nC_r = \frac{n!}{(n-r)!r!} = \frac{4!}{(4-2)!2!}$$

$$= \frac{4!}{2!2!}$$

$$= \frac{4 \times 3 \times 2!}{2!2!}$$

$$= \frac{12}{2} = 6$$

→ It refers to the different way of selecting item from a set where the order of selection doesn't matters but items should not repeat.

6/11/25

Hypothesis testing :

H_0/H_N → Null hypothesis (True Statements)

H_1/H_A → Alternative hypothesis.

Ex:

H_0 : The Avg height of Indian man is 5.7.

H_A : No, The Avg height of Indian man is not 5.7.

H_0 : True Statement.

H_A : It will helps to find either we have to accept (or) reject H_0.

P-Value :

If $p <= \alpha$ we can reject Ho.

If $p > \alpha$ we have to accept the Ho.

→ "α" is called as Significance value.

α - Significance value :

$$\alpha = 1 - CI$$

→ If CI is 68% what is the α is "0.32".

→ If CI is 95% then α is "0.05".

→ If CI is 99.7% then α is "0.03".

CI = Confidence Interval.

P-value formula :

$$z = \frac{\hat{p} - p_0}{\frac{p_0(1-p_0)}{n}}$$

$\hat{p}$ = Sample proportion.

p_0 = Assumed population proportion in the Ho.

n = Sample size.

$p = 0.03$

$\alpha = 0.05$

$p < \alpha$ reject Ho.

→ Hypothesis testing is a frame work for making inference about data and models in Machine learning.

→ It helps in model evaluation feature selection assumption validation and ensuring the robustness and

reliability of conclusions drawn from the models.

Type I and Type II Error :

R → Reality.

D → Decision.

R Ho is True D Ho is True.

R Ho is True D Ho is false → Type I Error.

R Ho is false D Ho is True → Type II Error.

R Ho is false D Ho is false.

If you failed to accept Ho is type I error.

If you failed to Reject ~~to~~ Ho is type II Error.

and also most dangerous.

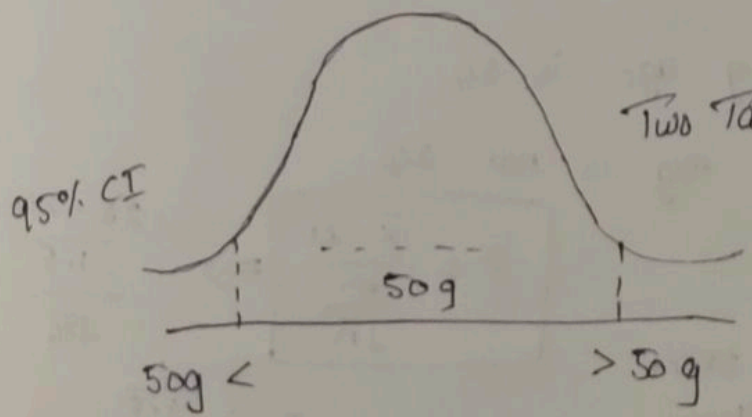
One Tail test and Two Tail test :

Ho :- The chips packet weight is 50g.

HA :- The chips packet weight is not 50g.

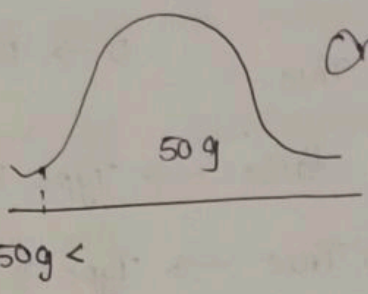
Ho = 50g

HA ≠ 50g



→ Ho :- The chips packet weight is more than 50g.

HA :- No The chips packet weight is not more than 50g.



Z Test and T Test :

→ The Average age of college student is 24 years with the standard deviation 1.5 Sample of 36 Students the mean is 25 years with 95% confidence interval do the age will vary or not.

Z-Test : If they given population standard deviation go with z test.

T-test : If they give Sample standard deviation go with T test.

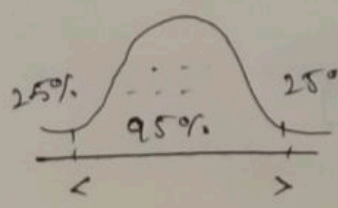
$\mu = 25, \sigma = 1.5, n = 36, \bar{x} = 25, CI = 95\%, \alpha = 0.05$

Ho : The avg age is 24

HA : No The AVg is not 24

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow \frac{25 - 24}{\frac{1.5}{\sqrt{36}}}$$

$$z \Rightarrow \frac{1}{\frac{1.5}{6}} = 4$$



Z table check value for 4 with α of 0.05.

0.99997.

Area under curve $= 1 \Rightarrow 1 - 0.99997$

$\Rightarrow \frac{0.00003}{2} = 0.000015.$

This is p value $= 0.000015.$

p is $0.000015 < \alpha\ 0.05$ So we can reject Ho.

2) In the population the Avg IQ is 100 with a SD of 1.5 Researchers want to test a new medication to see if there is +ve or -ve effect on intelligence or no effect at all a sample of 30 participants who have taken the medication has a Mean of 140. did the Medication offers the intelligence or not with a C.I 95% ?

A) Ho : $\mu = 100$, $n = 30$, $\sigma = 15$, $\bar{x} = 140$. C.I = 95% = 0.05

Ho : The Avg IQ of the population is 100.

Since population SD (σ) is known - Z-test.

HA : No The Avg IQ of

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

$$Z = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = \frac{4}{\frac{15}{\sqrt{5.477}}} = \frac{40}{2.738} = 14.61.$$

At 95% Confidence (two-tailed test):

$\alpha = 1 - 0.95$

$= 0.05$

$Z_{critical} = \pm 1.96$.

$Z_{calculated} = 14.61 > 1.96$.

$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

$1 - 0.025 = 0.975$

Reject H_0.

$p < \alpha$ So we can reject

∴ At 95% Confidence level, the Sample mean is significantly higher than the population mean. So, the medication has a significant positive effect on intelligence.

7/11/25

$\mu = 100$, $\sigma = 15$, $n = 30$, $\bar{x} = 140$, C.I = 95% = 0.05

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 3.66$$

$= 0.00013$

$p = 1 - 0.00013$

$= \frac{0.99987}{2}$

$= 0.499$

$p > 0.05$ So accept the value.

1) Some above question the $\mu = 100$, $n = 30$, $\bar{x} = 140$, $S = 20$, $\alpha = 0.05$.

2) Here Sample SD is given so we have to do t-test.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{140 - 100}{\frac{20}{\sqrt{30}}} = 10.95$$

DOF = n - 1
= 30 - 1
DOF = 29

t > table value with DOF = 29 ~~accept Ho~~ & $\alpha = 0.05$

then reject Ho else Accept Ho.

10.95 > 2.045 so We can reject.

2) Credit Card Launch :

$n = 140$, $\bar{x} = 1990$, $\sigma = 2500$, $S = 2833$, C.I = 95%, $\alpha = 0.05$

We have to calculate the Internal ~~storage~~. Range.

Default always the CI is 95%

$$CI = \bar{x} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Lower Bound

$$1990 - z_{\frac{0.05}{2}} \frac{2500}{\sqrt{140}}$$

Upper Bound.

$$= 1990 - z_{0.025}(211.29)$$

$$= 1990 - (1.96)(211.29)$$

$$= 1990 - 414.1284$$

$$= 1576.$$

Upper Bound

$$\bar{x} + z\alpha \frac{\sigma}{\sqrt{n}}$$

$$= 1990 + \frac{z_{0.05}}{2}\left(\frac{2500}{\sqrt{140}}\right)$$

$$= 1990 + (1.96)(211.29)$$

$$= 1990 + 414.12$$

$$= 2404.$$

∴ The Avg balance they are going to maintain after full fledge launch is [1576 2404]

3) An a quantitive test of a cat exam of a Sample of 25 test takers as a sample mean of 520 ~~which~~ with Sample Standard deviation of 80 Construct 95% Confidence Interval about the mean?

A) $n = 25$, $\bar{x} = 520$, $s = 80$, CI = 95% $\alpha = 0.05$.

$$t = \bar{x} \pm t\alpha \frac{s}{\sqrt{n}}$$

Lower Bound

$$\bar{x} - t\alpha \frac{s}{\sqrt{n}}$$

$$= 520 - \frac{t_{0.05}}{2} \frac{80}{\sqrt{25}}$$

$$= 520 - (2.06)(16) = 520 - 31.36.$$

$$= 488.64$$

upper Bound

$$\bar{x} + t\alpha \frac{s}{\sqrt{n}}$$

$$= 520 + \frac{t_{0.05}}{2} \frac{80}{\sqrt{25}}$$

$$= 520 + (2.06)(16)$$

$$= 520 + 31.36$$

$$= 552.$$

∴ the Avg balance they are going to maintain after fullfledged [487, 552].

Statistical tests :