24/07/19

Central Tendency: [to represent data in a single number]

- Mean: $\bar{x} = \frac{\Sigma x_i}{n}$ (or) $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$

- Median: even $= \text{avg}\left(\frac{n}{2}, \frac{n+1}{2}\right)$

  odd $= n/2$

- Mode : most frequent item

- Percentile: It tells the position of observation in dataset.

$$P_x = \frac{x(n+1)}{100}$$

- Quartile — Q_1 : contains 1st 25 % of data.

- Decile — divides data into 10 parts

Measure of Variation:

- Range $(\max - \min)$
- Inter Quartile Range $(Q_3 - Q_1)$
- Variance
- standard Deviation

Variance: Measure of variability from the mean

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

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

Standard Deviation:

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

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$$

Mostly $x_i - \mu > x_i - \bar{x}$

in order to adjust them we divide $\frac{(x_i - \mu)^2}{n}$ & $\frac{(x_i - \bar{x})^2}{n-1}$ ie, divides by $(n-1)$

This is to adjust the balance b/n s & σ

[downward bias problem]

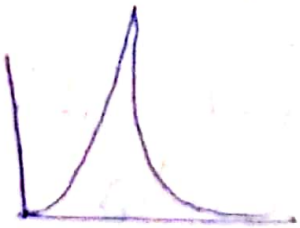
Skewness: it used to measure the symmetry of the distribution of data

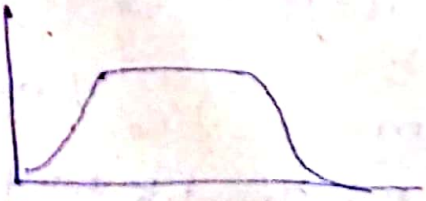
Skewness = 0 → Symmetric

-ve → left skewed

+ve → Right skewed

Kurtosis: at the top of the relational graph is sharp or blunt (flat)





solution

Random Variable: Function which assign a real number to each sample point in sample space.

Ex tossing 3 coins.

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

no. of tails(x) = f(n)

$\Rightarrow$ {0, 1, 1, 2, 1, 2, 2, 3}

Discrete Random Variable:—

A Random Variable which takes only finite values. in finite set of observations.

Continuous Random Variable:— A Random Variable which takes infinite values in finite observations.

Probability Mass functions:

Probability that the D.R.V 'X' takes on the value x_i given by $P(x_i) = P(X = x_i)$ for $i = 1, 2, 3, \cdots n$

(i) $f(x_i) \geq 0$

(ii) $\Sigma f(x_i) = 1$

Ex: If we flip a coin 2 times.

Random Variable (x) = no. of tails.

S = {HH, HT, TH, TT}

RV(x) = {0, 1, 1, 2}

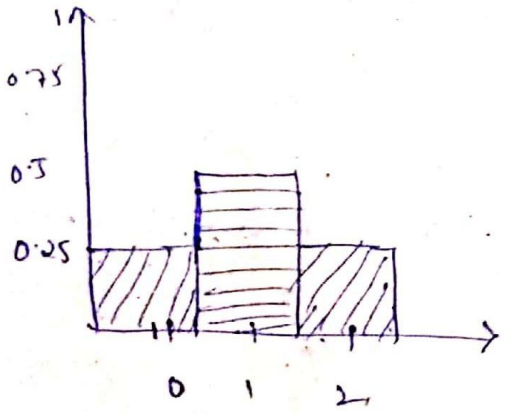
$f(1) = P(HT, TH) = 2/4 = 1/2 \geq 0$

$f(0) = P(HH) = 1/4 \geq 0$

$f(2) = P(TT) = 1/4 \geq 0$

$\Sigma f(x_i) = \frac{1}{4} + \frac{1}{4} + \frac{1}{2} = 1$

	0	1	2
x	0	1	2
f(x)	1/4	2/4	1/4



Probability thistogram.

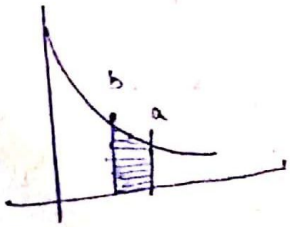
Probability Density function (continuous Rv) :-

If x is a continuous Random Variable the function f(x) is Probability Density function

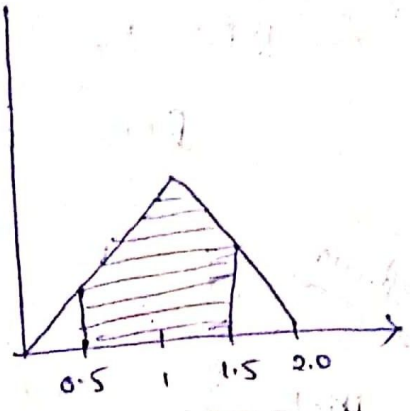
$$f(x) = P(a \leq x \leq b) = \int_a^b f(x) \cdot dx$$

if (i) $f(x) \geq 0$

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



Ex $p(x) = \begin{cases} x & 0 \le x < 1 \\ 2-x & 1 \le x < 2 \\ 0 & \text{other} \end{cases}$



$$\Rightarrow \int_{0.5}^{1.5} f(x)\,dx$$

$$\Rightarrow \int_{0.5}^{1} f(x)\,dx + \int_{1}^{1.5} f(x)\,dx$$

$$\Rightarrow \int_{0.5}^{1} x\,dx + \int_{1}^{1.5} 2-x\,dx$$

$$\Rightarrow \left[\frac{x^2}{2}\right]_{0.5}^{1} + 2[x]_{1}^{1.5} - \left[\frac{x^2}{2}\right]_{1}^{1.5}$$

$$\Rightarrow \frac{1}{2}[1^2 - 0.5^2] + 2[1.5-1] - \frac{1}{2}[(1.5)^2 - 1^2]$$

$$\Rightarrow \frac{1}{2}[0.75] + 2[0.5] - \frac{1}{2}[1.25]$$

$$\Rightarrow 1 + \left(\frac{0.75 - 1.25}{2}\right)$$

$$\Rightarrow 1 + \frac{0.50}{2}$$

$$\Rightarrow 1 + 0.25 = \underline{1.25}$$

Cumulative Distribution function (CDF) :-

[irrespective of continuous & discrete function, we can use Cumulative distributive function (CDF)]

C.D.F of R.V 'x' is defined as the probability that R.V 'x' takes a value "less than & equal to x"

C.D.F = $F_x(x) = P(X \leq x)$

$$f(a) = P(X \leq a) = \sum_{x \leq a} f(x)$$

31/7/19

Histogram:

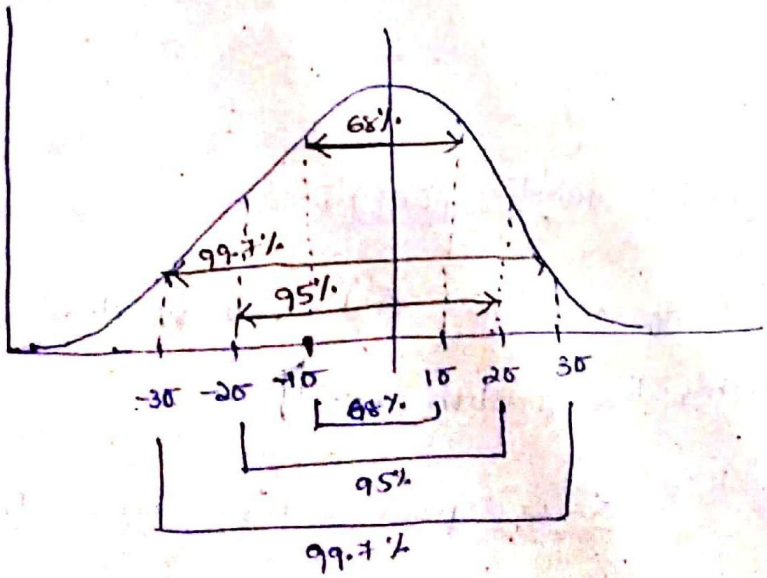
Pictorial representation of frequency distribution.
which groups the data into groups.

Normal Distribution:

Normal distribution of data usually allow the usuase of parametric test to analyse the data

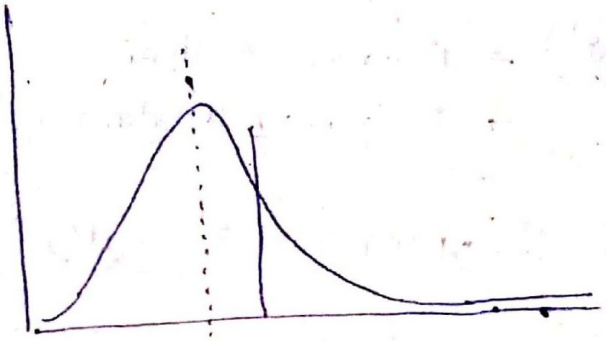
Note In normal distribution mean, median, mode are located at same place (ie,, middle)

- Parametric test has lots of tools, it is well developed system



Positively skewed data

The mean shift right side



mean.

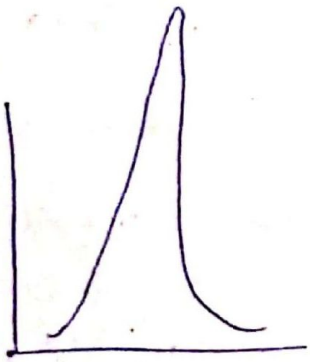
(+ve skewed data)

To measure normality — measure of distribution.

— Skewness [Symmetry of the distribution]

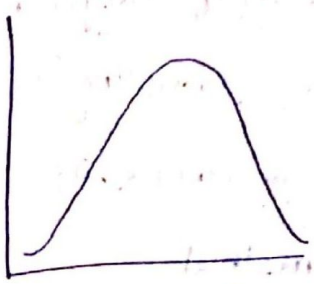
— Kurtosis [peakedness] — measure of peakedness.

for normal distribution kurtosis = 0



Lepto kurtic

(a)



meso kurtic

(b)

platy kurtic

(c)

Standard error = $\frac{\sigma}{\sqrt{n}}$ ⇒ $\frac{\text{standard deviation}}{\sqrt{\text{no. of observations}}}$

If ≠ 3 times of standard error. then we can't say it is in normal distribution.

$K < 3\sigma_x$ (standard error)

$K < 3\left[\frac{\sigma}{\sqrt{n}}\right]$.

Why

skewness $< 3\left(\frac{\sigma}{\sqrt{n}}\right)$

11/8/19

→ Remove outliers

→ increasing the data size

Sampling and Estimation:

Sampling is a process of selecting subgroup from population to make inferences about population parameters such as mean, median, proportion, standard deviation etc

Steps used in Sampling process:

1. Identification of target population
2. Decide the sampling frame
3. Determine the sample size
4. Sampling method
   - probablistic sampling
   - Non-probablistic Sampling

- **Probablistic Sampling**
  - Random Sampling → If population is homogeneous, it is used
  - Stratified Sampling
  - clustered Sampling
  - Bootstraping (Bagging)

→ Stratified Sampling:

- divide population with equal classes.
- Then the ratio of proportions should be equal in each class

→ Clustered Sampling:

- only important clames only considered.
- neglecting the negligible clam.

→ Bootstrapping:

- Non-probablistic Sampling:

1. Convenience Sampling
2. Voluntary Sampling.

- Sampling Distribution:

Probability distribution of static such as mean, standard deviation computed from random sample.

Population

1	2	3	4	5	6	7	8	9	10
5	10	15	20	25	30	35	40	45	50

Sample of size 2:

Sample	5,5	5,15	10,5	10,15	10,45	20,15	45,15
mean	5	10	7.5	12.5	27.5	17.5	30

50,20 25,10
35 17.5

If we take sample sufficiently large sample size from population. The mean of all samples follow approximately normal distribution with mean equal to population mean and std. deviation equal to $\frac{\sigma}{\sqrt{n}}$

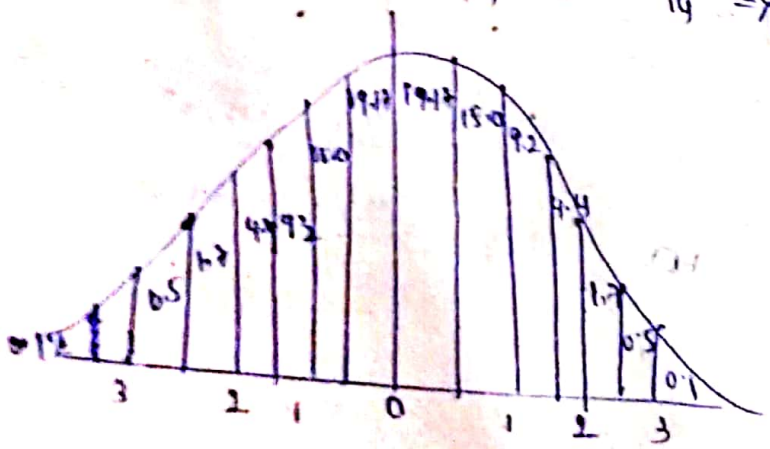
Standard Normal Distribution:

It is a normal distribution, it has mean '0' & stand. deviation '1'

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

$$z = \frac{x-\mu}{\sigma}$$

$$P(x<72) \Rightarrow P\left(x < \frac{72-86}{14}\right) = \frac{-14}{14} = p(x<-1)$$



Confidence Intervals:

CI is the range in which the value of population parameter is likely to be with certain probability.

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Prb If we measure the height of 40 randomly choosen men and get a mean height 175cm. We also know that standard deviation is 20cm

Soln $s = 40$

$s = 20$

$\bar{x} = 175$

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

$$= 175 \pm 1.96 \times \frac{20}{\sqrt{40}}$$

$$= 175 \pm 6.20$$

$\Rightarrow$ 168.8 cm to 181.2 cm

6/8/19

Hypothesis testing:

- Null hypothesis — When comparing any two, there is no difference
- Alternate hypothesis
- Rejection of Null hypothesis

let sample $- n = 10$

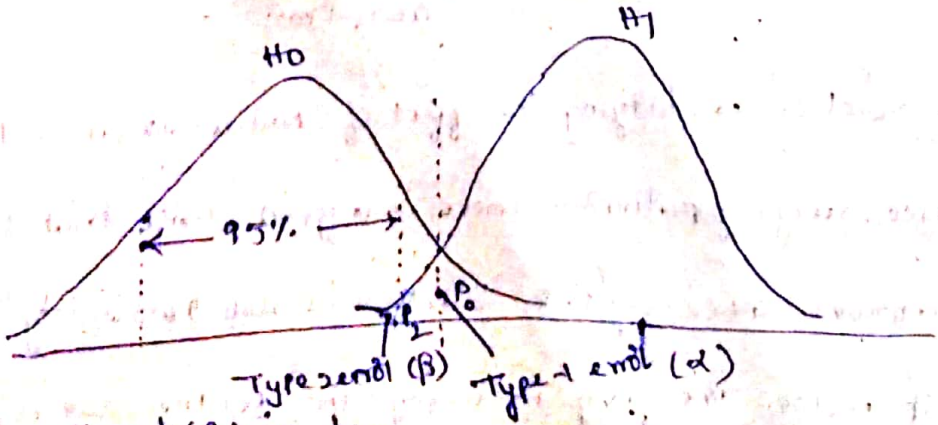
$\bar{x} = 44.5$

$\sigma = 11.41$

Hypothesis is an assumption about population parameter

$H_0 = \mu \geq 45$ Null hypothesis

$H_1 = \mu < 45$ alternate hypothesis



Here though (P_0) is from Ho. but we are considering it as not $\in$ Ho, Since it is

not in the 95% confidence of the H_0, it is considered false.

ie, type -1 error (α).

The t_2 belongs to H_1, but it is in the range of 95% confidence interval, though it is belongs to H_1, it is accepted as belongs to H_0. This is of type 2 error (β)

7/8/19

α – false rejection of H_0

β – false acceptance of H_0

Ex: A Researcher thinks that a knee surgery patients go to physical therapy twice a week. (instead of 3 times) their recovery period will be longer, Average recovery time for knee surgery patients is 8.2 weeks.

Soln: $\mu_0 = 8.2$ weeks

$H_0 : \mu \leq 8.2$

$H_1 : \mu > 8.2$

Always test on null hypothesis, but not on alternate hypo

→ Always alternate hypothesis is accepted.

Ex: A Researchor is studying the effect of Radical exercise program on knee surgery patients. there is a good chance that the therapy will improve recovery time, but there is also possibility it will make it worse. The avg. Recovery time for knee surgery Patients is 8.2 weeks.

Soln: $\mu_0 = 8.2$ ω

H_0 : $\mu = 8.2$

H_1 : $\mu \neq 8.2$

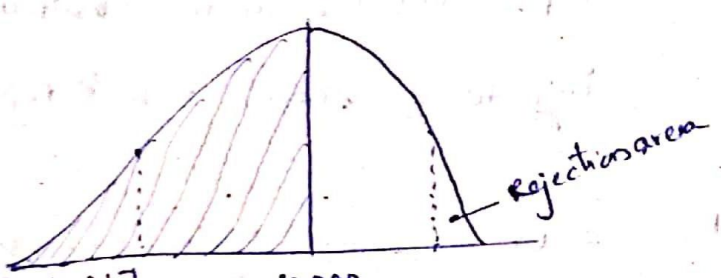
Ex: Salary of Machinelearning expert on average is atleast $ 10,000

$\mu_0 = 10,000$

H_0 = $\mu \leq 10,000$

H_1 = $\mu > 10,000$

Ex:



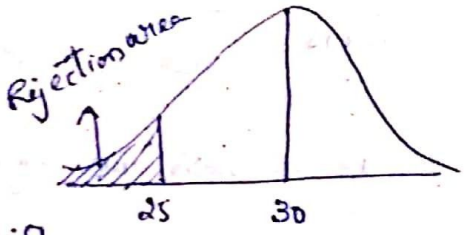
One tail test [Right-tail]

8/8/19

Ex: Average waiting time at London Heethrow airport security check is less than 30min.

H_0 : $\mu \geq 30$

H_1 : $\mu < 30$

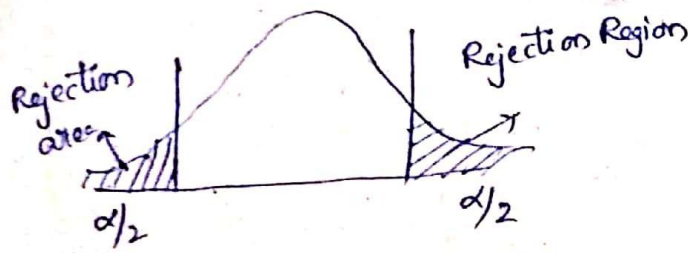


one tail test [left-tail]

Ex Avg Salary of male and female MBA students at graduation is different.

H_0 : $\mu_m = \mu_f$

H_1 : $\mu_m \neq \mu_f$



two tailed test

Steps for hypothesis testing:

- State the null hypothesis
- State the alternate hypothesis
- Determine the value of α
- find z-score associated with alpha level
- Find the test statistic using formulae
- If the value of the static is less than z-score of alpha level [p-value is less than α-value] reject the null hypothesis otherwise don't reject the null hypothesis.

Pro $\bar{x} = 44.5$, $S = 11.41$, $n = 10$ CI = 95%. $\alpha = 5\%$. (type I error)

$\bar{x} - z\frac{\sigma}{\sqrt{n}}$

$H_0: \mu \geq 45$

$H_1: \mu < 45$

$44.5 \pm 1.96 \cdot \frac{11.41}{\sqrt{10}}$ $44.5 \pm \frac{11.41}{3.162}$

$\Rightarrow 44.5 \pm \frac{11.41}{3.162}$ $\Rightarrow$ 37.42 to 51.57

if < 37.42 & 51.57 are rejection areas.

if $\mu_0 = 45$, then. > 37.42 & < 51.57 there is no statistic difference.

14/8/19

25, 60, 43, 56, 32, 43, 47, 59, 39, 41

$n = 10$ $\quad S = 11.41$

$\bar{x} = 44.5$ $\quad \mu_0 = 45$

Perfolmance initiatives taken by the HR Team are not upto mark [one tail test]

H_0 : $\mu \geq 45$ (null hypothesis)

H_1 : $\mu < 45$ (Alternative)

commonly $\alpha = 5\%$ it may be 1%, 2% or any thing.

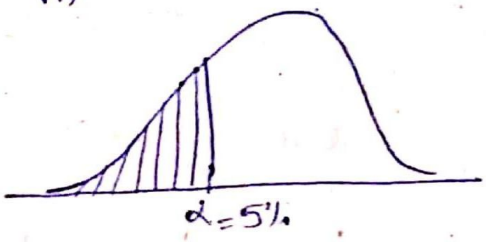
So, CI → confedence Interval : 95%.

Confidence interval $\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$

$\sqrt{10} = 3.162$

$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$

$\Rightarrow 44.5 \pm (1.64)\left(\frac{11.41}{\sqrt{10}}\right)$



$\Rightarrow \mu = 38.5$

allowed region, > 37.42 & < 51.57

$\therefore \mu = 38.5$ is in the allowed region.

So, H_0 accepted, H_1 Rejected.

1) $H_1 < 45$

$H_0 \geq 45$

$\alpha = 5\%$

$Z = 1.64$. [One tail test]

$$Z \Rightarrow \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} \Rightarrow \frac{44.5 - 56}{\left(\frac{11.41}{\sqrt{10}}\right)} \Rightarrow \frac{-8.5}{\left(\frac{11.41}{3.162}\right)}$$

$$\Rightarrow -3.12$$

Pvalue = -3.12

16/8/19

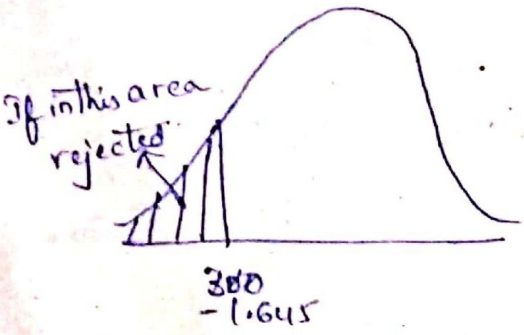
A Manufacturer purchare bulbs. that are suppose to burn for a mean life time of atleast 3000 hrs. with a standard deviation of 500 hrs. If a sample of 100 bulbs is taken with mean $\bar{x}$ = 2800 hrs

$n = 100$

$\bar{x} = 2800$ hrs

H_0 : $\mu \geq 3000$ hrs.

H_1 : $\mu < 3000$ hrs.



so, single tail test.

— decide α = 5%

confidence : 95%

then $Z = 1.64$.

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \Rightarrow \frac{2800 - 3000}{\frac{500}{\sqrt{100}}}$$

$$\Rightarrow \frac{-200}{50}$$

$$\Rightarrow -4$$

Z is in rejection area so, it is rejected.

reject null hypothesis.

Ex: A company claims that its weight reducing drug will cause atleast 10 pounds with in 1 month. A Random sample of 64 subjects is taken and the avg. weight loss is 7 pounds. with standard deviation $s = 4$ pounds & $\alpha = 0.01$

→ $n = 64$

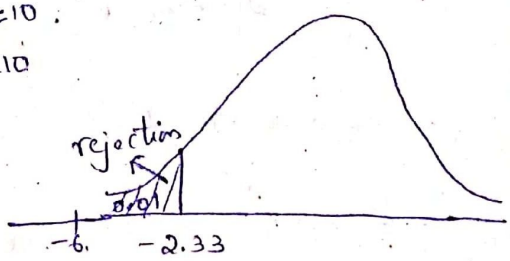
$\bar{x} = 7$

$s = 4$

$\alpha = 0.01$

$H_0 = \mu \geq 10$.

$H_1 = \mu < 10$



rejected.

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$\Rightarrow \frac{7 - 10}{\frac{4}{\sqrt{64}}} \Rightarrow \frac{-3}{\frac{4}{8}} = -6$$

reject null hypothesis.

So, given alternate hypothesis is true.

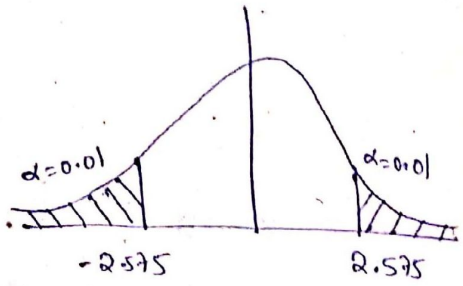
A researcher claims that 10 years old watch 6.6 hrs of TV daily. You try to verify this with following sample data.

$n = 100$

$\bar{x} = 6.1$ hrs

$S = 2.5$ hrs

$\alpha = 0.01$



$H_0 \Rightarrow \mu = 6.6$ hrs

$H_1 : \mu \neq 6.6$ hrs

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \Rightarrow \frac{6.1 - 6.6}{\frac{2.5}{\sqrt{100}}} \Rightarrow \frac{-0.5}{\frac{2.5}{10}} \rightarrow \frac{-5}{25} \times \frac{10}{10}$$

$\Rightarrow -\frac{1}{5} \times \frac{2}{10}$

$\Rightarrow 0.2 \times 2 \Rightarrow -2$

it is not in -2.575 ie, not in rejection region.

$\Rightarrow$ then, H_1 is rejected.

Confidence interval $\bar{x} \pm Z \frac{S}{\sqrt{n}}$

$\Rightarrow 6.1 \pm 2.575 \times \frac{2.5}{\sqrt{100}}$

$= 6.1 \pm (2.575 \times 0.25)$

$\Rightarrow 6.1 \pm 0.64375 \cong 0.64$

$\Rightarrow 6.1 \pm 0.64$

$\Rightarrow$ 5.36 to 6.74

2.575 × 0.25

0.64375

6.10
.64
5.36

A manufacturer produces bolts with a thickness of exactly 1 inch. A customer takes a random sample of 100 bolts, and find $\bar{x} = 1.2$ inches, $s = 0.40$ inches. should the manufacturers claim that the bolts be rejected. ?

$n = 100$

$\bar{x} = 1.2$ inches

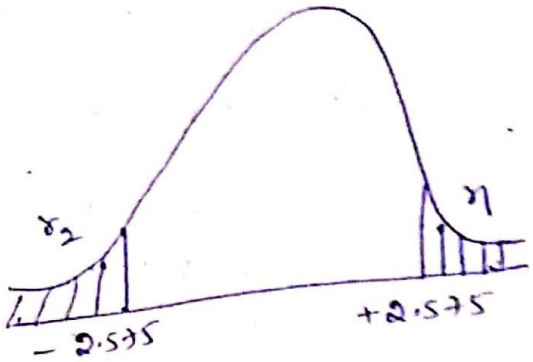
$s = 0.40$ inches.

$\alpha = 0.01$

$Z = 2.575$

$H_0 : \mu = 1$ inch.

$H_1 : \mu \neq 1$ inch.



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

$$\Rightarrow \frac{1.2 - 1}{\frac{0.40}{\sqrt{100}}} \Rightarrow \frac{0.2}{\left(\frac{0.40}{10}\right)} \Rightarrow \frac{0.2 \times 10}{0.40} \Rightarrow \frac{2}{0.40} \Rightarrow \frac{2}{\left(\frac{40}{100}\right)}$$

$$\Rightarrow \frac{200}{40} \Rightarrow 5$$

$$\Rightarrow \pm 5$$

$+5$ is in r_1 region so, it is rejected.

using confidence intervals.

$$\bar{x} \pm Z \frac{s}{\sqrt{n}}$$

$$\Rightarrow 1.2 \pm 2.575 \times \frac{0.40}{\sqrt{100}}$$

$$= 1.2 \pm \left(2.575 \times \frac{0.40}{10}\right)$$

$$\frac{2.575 \times 0.40}{10}$$

$$= 1.07 \text{ to } 1.303$$

A company claims that their battery have a life of atleast 100 hrs. The following data.

$n = 121$

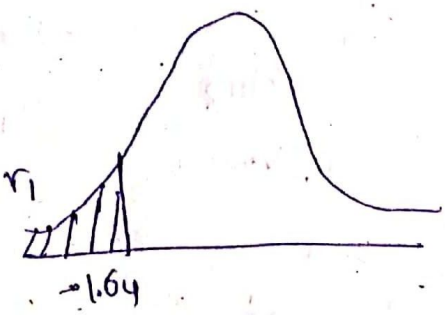
$s = 3$ hrs

$\bar{x} = 97$ hrs

$\alpha = 0.05 = 5\%$

$H_0 : \mu \geq 100$ hrs

$H_1 : \mu < 100$ hrs



$$Z = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \Rightarrow \frac{97 - 100}{\frac{3}{\sqrt{121}}} \Rightarrow \frac{-3}{\left(\frac{3}{11}\right)} \Rightarrow -3 \times \frac{11}{3} \Rightarrow -11$$

Hence it is in rejection region. So, reject null hypothesis

→ Sample size should be ≥ 30.

→ It should follow standard normal distribution.