Statistics:- The science of collecting, organizing & analyzing data.

Data:- "facts of pieces of information"

2 types

Statistics

Descriptive stat
- It consist of organizing & summarizing of data

Inferential statistics
It consist of a technique to form some conclusions

Ex: lets say - 20 classrooms - Mixed age of a one classroom.
- [21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22]

1) Random Sampling
↑
100000

3] Systematic Sampling - Selecting n^{th} people from crowd.
Eg: Every 5th person come, he will go & ask for credit card.

2] Stratified Sampling
→ Gender - male/female ✓
→ Edu. Degree - BE, master, PHD, High school
these are overlapping

Strata
↑
Layers → clusters
↓
non overlapping

→ Blood group
✓ A+
B+

4] Convience Sampling - only those who are interested will be participating.

Data science
Art ×
Engg ✓ medical ×
Interested people.

Qualitative variable

categorical

Good / Bad

married / unmarried

Quantitive Variable

Discrete variable

eg: no. of children in a family [0, 1, 2, 3]

eg: whole number [1, 2, 3]

Continuous variable

eg. length, weight

3.2, 5.5

32.5

Histogram

Age: {10, 12, 13, 14, 16, 37, 25, 35, 40, 65, 55, 32}

1] Sort the no. [Ascending]

2] Bins — {groups} or gaps → no. of groups

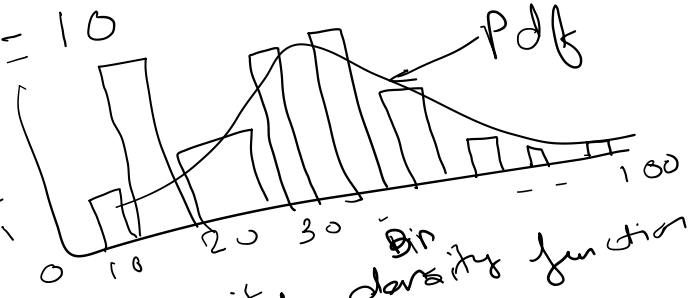
3] Bin size → Size of Bins

[0-100]

Bins = 10

Bin size = $\frac{100}{10} = 10$

Bins = 10



what is probability density function

→ Smoothing histogram

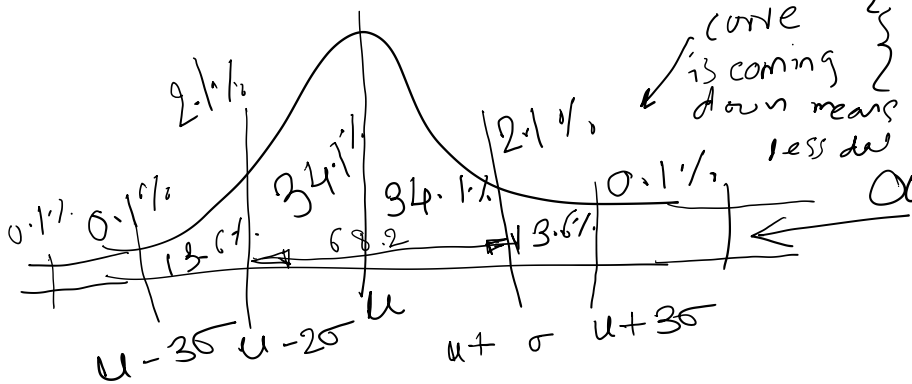
[through kernel density estimator]

$x = \{2, 4, 6, 10, 12 - - - - -\}$

{Emperical rule}

{Gaussian Distribution}

{Normal Distribution}



[68.2 — 95.4 — 99.7]

↓ Emperical Formula

How much % of data availble at center.

Z-Score $= \frac{x_i - \mu}{\sigma}$ — used for Standardization — means making all input units in one.

It is used in standard normal distribution.

Items = {1, 1, 2, 3, 4, 5, 5, 8, 7, 5, 9, 10, 10, 20} = 14 items

95% percentile $= \frac{95}{100}(n+1) = \frac{95}{100} \times 15 = 14.25$

14.25 is index

Find out average delivery time of items = 20 days

So we choose 20 = 20 days

Log-normal distribution

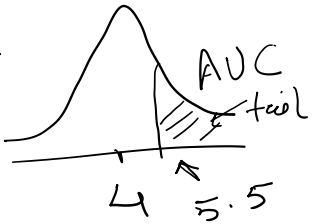
always right skewed

$x =$ { } $\Rightarrow$ apply $\log_e(x) \Rightarrow$ { } $= x$ $\Leftarrow e^x$ $\Leftarrow$ [2]

$\mu = 4$, $\sigma = 1$

% of scores less than 4.5?

e.g. 1



$\mu = 4$, $\sigma = 1$

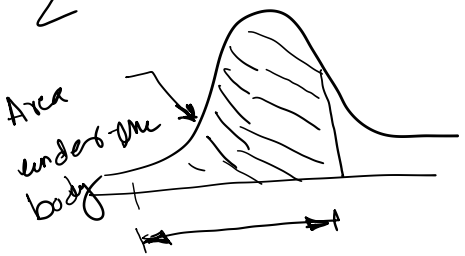
Z-score $= \frac{5.5 - 4}{1} = 1.5$

From Z table 1.5 → 0 → 0.93319

Area of tail $= 1 - 0.93319$

AUC $= 0.6681$

Z Table -



In india, the average IQ IS 100, with a standard deviation of 15. what is the percentage of population would you expect to have and IQ lower than 85?

1] lower than 85

2] higher than 85

3] between 85 to 100

$\mu = 100 \quad \sigma = 15$

$Z\text{-score} = \frac{85-100}{15}$

$= \frac{-15}{15} = -1$

% of score between 100 to 125 ∪ 45.15%

Area of Body $= 0.1587$

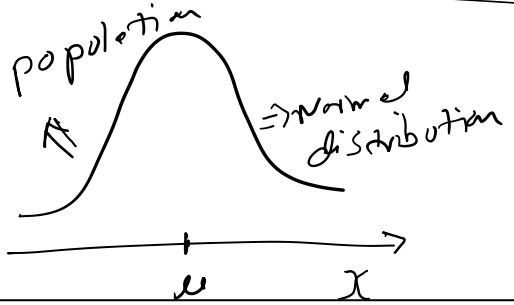
$= 1 - 0.1587 = 84.13\%$

$\frac{125-100}{15} = \frac{5}{3} = 0.9515 - 0.5$

$= 0.4515$

$= 45.15\%$

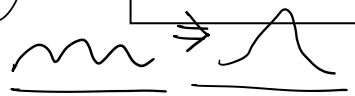
Central limit Theorem:-



If data follows, normal or non- normal distribution,

If multiple samples i.e n >= 30 , and if we take its sample mean, and if we plot , it follows normal or gaussian distribution.

The moe the value of n, more the data follows normal distribution.



n is greater than 30

$\{x_1, \ldots x_n\} \rightarrow \bar{x}_1$

$\{x_1, \ldots x_n\} \rightarrow \bar{x}_2$

$\{x_1, \ldots x_n\} \rightarrow \bar{x}$

Probability : measure of the likelihood of an even.

E.g tossing a fair coin , p(H) = 0.5 p(T)= 0.5 Q=1-P

{BERNOULLI'S DISTRIBUTION} +> TWO OUTCOMES

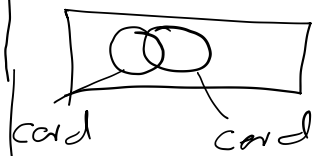
E.G ROLLING A DICE? {1,2,3,4,5,6}

P(I) = 1/6 P(2)= 1/6

1] Mutual exclusive events: two events are mutually exclusive if they can not occur at the same time.

Rolling a dice

2] Non mutual exclusive events: Picking randomly from the deck of card, two are " hearts and king" are not mutual exclusive



eg Bag of marble:-

Red, Green, Red & Green

Red or green

what is the probability of coin landing on heads or tails

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

$= \frac{1}{2} + \frac{1}{2} =$

what is probbbilty of getting 1 or 6 or 3 while rolling a dice?

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

Addition Rule $= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$

$= 0.5 = \frac{1}{2}$

Bag of marbles : 10 red, 6 green 3(red and green)

When picking randomly from a bag of marbles, what is the probability of choosing a marble that is red or green?

$P(R) = \frac{10}{19}$ $P(G) = \frac{6}{19}$ $P(R \text{ \& } G) = \frac{3}{19}$

Addition Rule for non-mutual ex event

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

$= \frac{10}{19} + \frac{6}{19} - \frac{3}{19} = \frac{13}{19}$

* multiplication Rule →

1] Independant event : Two events are independant if they do not effect one another

e.g Tosing a coin $P(H) = 0.5$ $P(T) = 0.5$

here one event not impacting other event.

2] Dependant events

↳ Two events are dependant if they affect one another.

e.g prob of rolling "5" & then "3" with normal 6k sided dice

- multiplication Rule here

$P(A \text{ \& } B) = P(A) \times P(B)$

$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

e.g

Bag of color marbles. $\{ \text{oooo} \; \text{xxx} \} \rightarrow 7$ (one removed)

$P(o) = \frac{4}{7} \rightarrow P(x) = \frac{3}{6} \rightarrow P(o) = \frac{3}{5} \rightarrow P(x) = \frac{2}{4}$

→ dependent → removing one by one

e.g) orange marble & 3 yellow marble

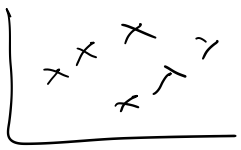
→ Prob. of orange & then yellow → multiplicative Rule of conditional events

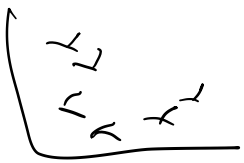
$P(o \text{ \& } y) = P(o) \times P(y/o)$ ← "Bayes"

$= \frac{4}{7} \times \frac{3}{6} = \frac{12}{42} = \frac{2}{7}$

* Covariance

$\{ x\uparrow \; y\uparrow \;;\; x\downarrow \; y\downarrow \}$ $\{ x\uparrow \; y\downarrow \;;\; x\downarrow \; y\uparrow \}$





Age	wt
12	20
15	30
19	50
15	55
x_i	x_i

$\bar{x}$ $\bar{y}$

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) \times (y_i - \bar{y})}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

x	y
1	3
2	4
3	6

$\bar{x} = 2$ $\bar{y} = 4.334$

$$cov(x,y) = \frac{[(1-2)(3-4.334) + (2-2)(4-4.334) + (3-2)(6-4.334)]}{2} = 1.497$$

can be written as $Cov(x,x)$

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

Pearson & Spearman Rank correlation

* Spearman coefficient :- & Pearson coefficient

Inferential stat.

1] P value :- Out of all

It is the probability of the null hypothesis is true?

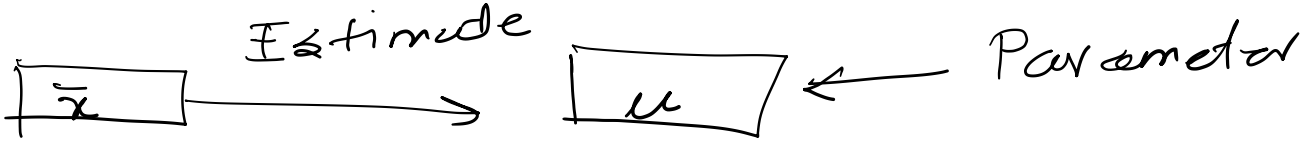
① Null Hypothesis - coin is fair - (H_0)

② Alternate Hypothesis - coin is not fair - (H_1)

③ Perform Experiments

Significant value - 0.05 ⇒ 95% CI

Point Estimate :- The value of any statistics that estimates the value of a parameter is called Point Estimate.



Point estimate $\pm$ Margin of error = Parameter

Confidence Interval

Lower fence = Point estimate − margin of error
Higher fence = Point Estimate + margin of error } C.I

Lower Fence = $\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ Higher Fence = $\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

t test :-

1] Degree of freedom :- $n-1 = 25-1 = 24$

* One Tail & 2 Tail

Lower fence =

Higher fence =

1) Anova

The problem statement : The factory has a machine that fills 80ml of baby medicine in a bottle . An employee believs the average amunt of baby medicine is not 80ml using 40 samples, he measures the average amount dispered by the machine to be 78ml with a std. dev of 2.5

a] state null and alt hyptotheis b] at 95% C.I, is ther enough evidence to support whether machine is working properly or not.[]

1) $H_0 : \mu = 80$ — null Hypothesis

$H_1 : \mu \neq 80$ — Alternative Hypothesis

Ztest

$SV = 1 - CI$.

$n = 40$

$\bar{x} = 78$

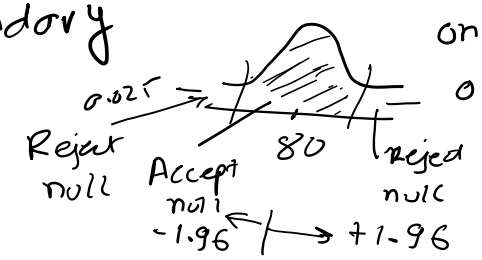
$s = 2.5$

Ztest & t-test

$\mu = 80ml$

2] step $\alpha = 0.05$ $CI = 95\%$ $\alpha = 1 - 0.95 = 0.05$

3] Decision Boundary one tail or two tail



From Z table

4] calculate

Test statistics

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

$n = 40$

sample size

std. error

$$\frac{78 - 80}{2.5/\sqrt{40}} = \frac{2 \times \sqrt{40}}{2.5} = -5.05$$

5] Decision Rule

If $Z = -5.05$ is less than -1.96 or greater than 1.96, then we reject null hypothesis with 95% CI.

Conclusion. Reject the null hypothesis.

↓

Fault in m/c or m/c is not working properly.

. A complain was registered, the boys in a govement school are underfed. average weight of the boys of age 10 is 32 kgs with s.d = 9 kg. A sample of 25 boys were selected from the government school and the average weight was found to be 29.5 kgs? with CI=95% . Choose whether it is true or false?

Conditions for Z-TEST

1[WE KNOW THE POPULATION S.D OR

2] WE DO NOT KNOW THE POPULATION S.D BUT OUR SAMPLE IS LARGE N>= 30.

CONDITIONS FOR T-TEST

1] WE DO NOT KNOE THE POPULATION VARIANCE OR S.D

2] OUR SAMPLE SIZES IS SMALL I.E N<=30

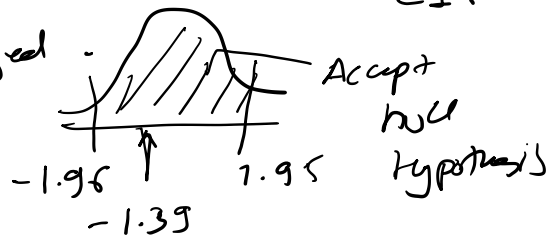
$H_0: \mu = 32$

$H_1: \mu \neq 32$

$\sqrt{} \longrightarrow$ Z-score $= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{29.5 - 32}{9/\sqrt{25}} = -1.39$

conclusion $-1.39 > -1.96$ so Accept the null hypothesis 95% CI.

Students are not underfed.



A FACTORY MFG. CARS WITH A WARRENTY OF 5 YEARS ON THE ENGINE AND TRANSMISSION. AN ENGINEER BELIEVES THAT THE ENGINE OR TRANSMISSION WILL MALFUNCTION IN LESS THAN 5 YEARS. HE TESTS A SMPLE OF 40 CARS AND FINDS THE AVERAGE TIME TO BE 4.8 YEARS WITH A STD. DEV OF 0.50 .

1] STATE THE NULL AND ALTERNATIVE HYPOTHEISIS

2] AT 2% SIGNIFICANT LEVEL. IS THERE ENOUGH EVIDANCE TO SUPPORT THE IDEA THAT THE WARRENTY SHOULD BE REVISED.

$H_0: \mu \geq 5$

$H_1: \mu < 5$

$n = 40$

$\bar{x} = 4.8$ yrs

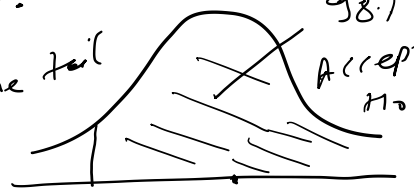
s.d $= s = 0.50$

$\alpha = 0.02$ CI = 98 %.

From Z table $= -2.05$

Z-score is $= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -2.52$

$\alpha = 0.02 \rightarrow$ Reject H_0



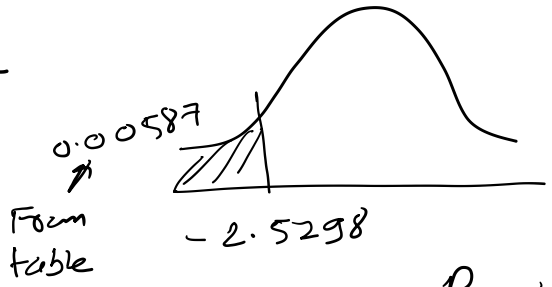
So, $-2.52 < -2.05$

$\rightarrow$ conclusion $\rightarrow$ Reject null

$\rightarrow$ warranty should be Revised.

So, company may consider warranty of 4 years, as most of the part will fail within 4.8 years.

using P-value - from table - Z-score = -2.52



$1 - 0.0587$

$= 0.99413$

$\alpha = 002$

P value $= 0.00587$

Reject null Hypothesis $\leftarrow$ P value $< \alpha$

EX; THE AVERAGE WEIGHT OF ALL RESIDENT IN A TOWN XYZ IS 168 POUNDS . A NUTRISIOANIST BELIEVS THE MEAN TO BE DIFFERENT. SHE MEASURED THE WEIGHT OF 36 INDIVIDUALS AND FOUND THE MEAN TO BE 169.5 POUNDS WITH A STANDARD DEVIATION OF 3.6.

A] NULL AND ALTERNATE HYPOTHESIS

2] 95% IS THERE ENOUGH EVIDENCE TO DISCARD THE NULL HYPOTHSIS?

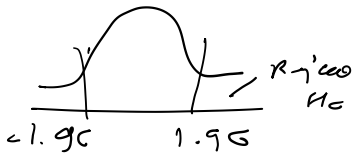
$\bar{x} = 169.5$ $S = 3.9$ $n = 36$ $\mu = 168$

1] $H_0 : \mu = 168$

$H_1 : \mu \neq 168$

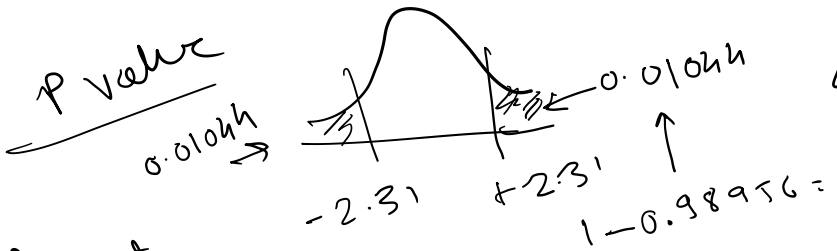
2] $\alpha = 0.05$

3) Decision Boundary



4] $Z\text{score} = \dfrac{168 - 169.5}{3.5/\sqrt{36}} = \underline{\underline{2.31}}$

P value



In 2 tail test

→ P value $= 0.01044 + 0.01044 = 0.02088$

P value $= 1 - 0.02088 = 0.97952$

$\alpha = 0.05 \rightarrow 0.02 < 0.05$ - Reject null Hypothesis

Z test with proportions

(1) A tech company believes that the percentage of residents in town XYZ that owns a cell phone is 70%. A marketing manager believes that this value to be different. He conducts a survey of 200 individuals and found that 130 responded yes to owning a cell phone

(a) State the Null and Alternate Hypothesis?

(b) At a 95% C.I, is there enough evidence to reject the Null Hypothesis?

→ Z test with proportions

- null → $H_0 : p_0 = 70\%$

$H_1 : p_0 \neq 70\%$. $q = 1 - p_0$

2] $\alpha = 0.05$

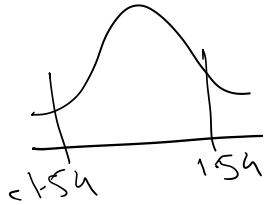
3]

Z test with proportion

$$Z\,test = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$-1.54 > -1.96$

- Accept null



Pvalue =

$0.06178 + 0.06178$

$= 0.12356$

$0.12356 > 0.05$ - Accept null

Chi square test:

It is a non parametrical test that is performed when data is categorical and ordinal data.

q] In the 2000 U.S cencus the ages of individual in a small town were found to be the following.

<18	18-35	>35
20%	30%	50%

Expected

In 2010, ages of n=500 individuals were sampled. below are the results.

Using alpha= 0.05 , would you conclude that the population distribution of ages has changed in the last 10 years?

	<18	18-35	>35
Observed →	121	288	91
Expected →	100	150	250

Ans H_0 :- The data meet expected distribution

H_1 : The data does not —11— —11—

2] $\alpha = 0.05$ C.I = 95% $n = 500$

chi sq Table

df / α — 0.05

2 — 5.991

3] DOF = df = k - 1 = 3 - 1 = 2

4] Desicion Bundary 5.991

{ $\chi^2 < 5.991$ - Accept null

5] Chi square Test Statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_e = expected

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$\chi^2 = 232.494$ → This is greater than 5.991

so we reject null Hypothesis.

QUE] 500 ELEMENTARY SCHOOL BOYS AND GIRLS ARE ASKED WHICH IS THEIR FAVORITE COLOR,: BLUE , GREEN OR PINK . RESULT ARE SHOWN BELOW.

USING APLHA = 0.05, WOULD YOU CONCLUDE THAT THERE IS A RELATIONASHIP BETWEEN GENDER AND FAVORITE COLOR?

	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	230
	120	180	200	500

DOF = (rows-1) (col-1)
= (2-1)(3-1)
= 1 × 2 = 2

NULL HYPTOTHSIS H0: - GENDER AND FAVORITE COLOR ARE RELATED

H1: GENDER AND FAVORITE COLOR ARE NOT RELATED

$\alpha = 0.05$

From Table - 5.991

DOF = 2

DOF	α — 0.05
2	5.991

* Chi Sq. Test Stat

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} =$$

→ Expected

B

G

	Blue	Green	Pink	
B	64.8	97.2	108	270
G	55.2	82.8	92	230
	120	180	200	500

$$= \frac{(100-64.8)^2}{64.8} + \frac{(150-97.2)^2}{97.2} + \frac{(20-108)^2}{108} + \frac{(20-55.2)^2}{55.2} + \frac{(30-82.8)^2}{82.8} + \frac{(180-92)^2}{92}$$

$\chi^2 = 259.7$

It's greater than 5.991.

→ Reject H_0 - null hypo → so they are not related.

INTERVIEW QUESTIONS

Asked in Amazon

1] size of all the sharks

$n=40 \quad \bar{x}=5\text{ m} \quad S=0.5\text{m}$

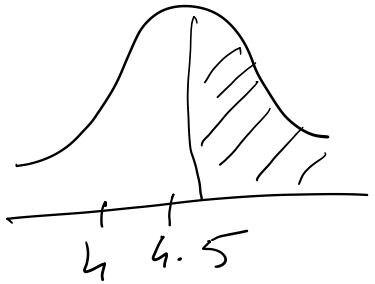
$\alpha=0.05 \quad n>30 \quad$ - Z test

confidence interval → $\bar{x} \pm$ margin of error

$\bar{x} \pm Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$

To calculate size of sharks.

2]



$\sigma = 0.5$

non linear data

Diff betn spearman Rank & pearson correlation

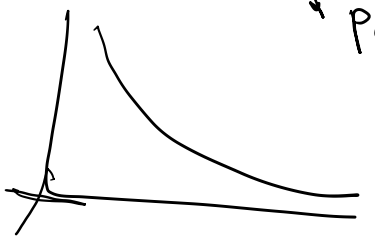
↑ in linear data

---

* oil Distribution in a place follows which distribution.

Ans- pareto distribution

↓ power laws

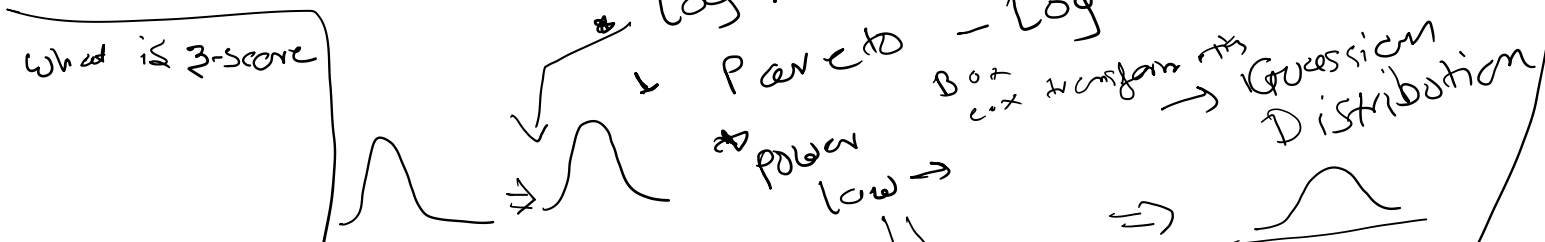


Distributions may be asked

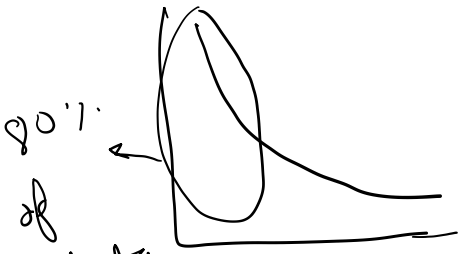
* Log normal $\xrightarrow{\log(x)}$ normal

↓ Pareto - Log normal

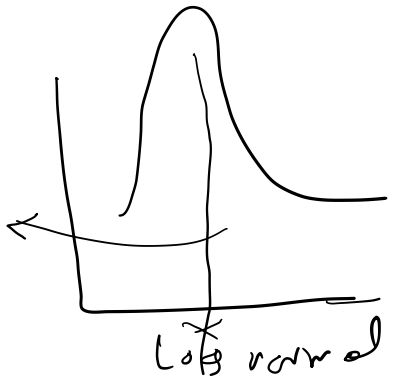
* power low → $\xrightarrow{\text{Box cox transform}}$ Gaussian Distribution



Pareto Distribution



Pareto or Power law



* Disadvantage of replacing NAN value by mean.

↓ variance ↑es - most values will be at