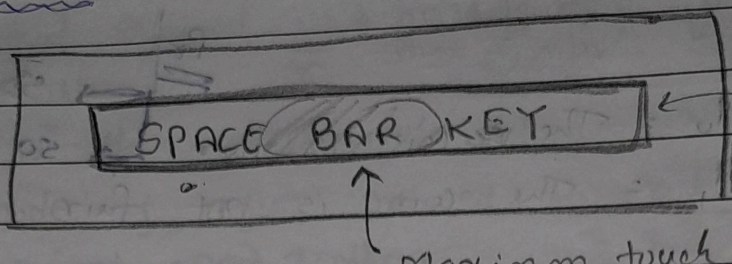
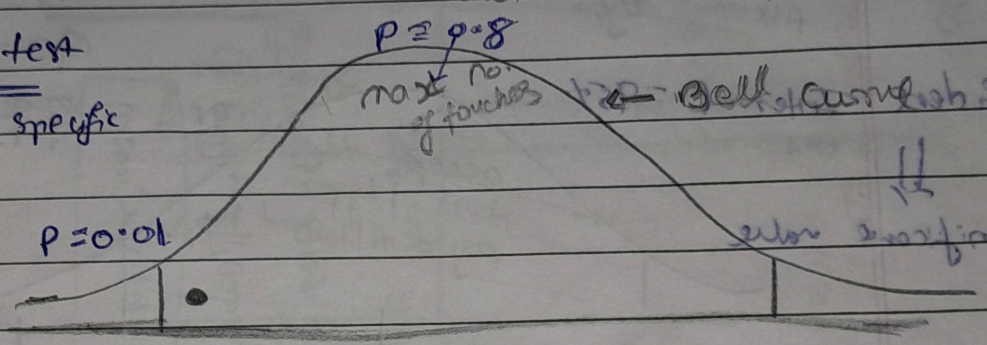
(12) * P- Value (Probability value)

eg:



Maximum touch point



2-Tailed test provides a specific ranges

P = 0.01

↳ This P = 0.01 represent that,

If we repeat this Experiment 100 times, then the number of time, ~~then the number of times~~ will be touching this area will be ⇒ 1 times

P = 0.8

Some Explanation ⇒ 80 times

Define ⇒ **P Value** ⇒ It is the Probability for the "Null Hypothesis" to be true.

Null Hypothesis ⇒ Treats Everything Same or Equal

★ P value we also called → Significant value.

Eg: Coin => fair coin -> Toss Experiment 100 times

Null Hyp (pr of H and T almost same)

$\therefore H_0$ = The coin is fair

$\therefore H_1$ = The coin is not fair (always opposite of H_0) ← Alternate Hyp

Pr → 50 time (Head) $Pr(H) = \frac{1}{2}$

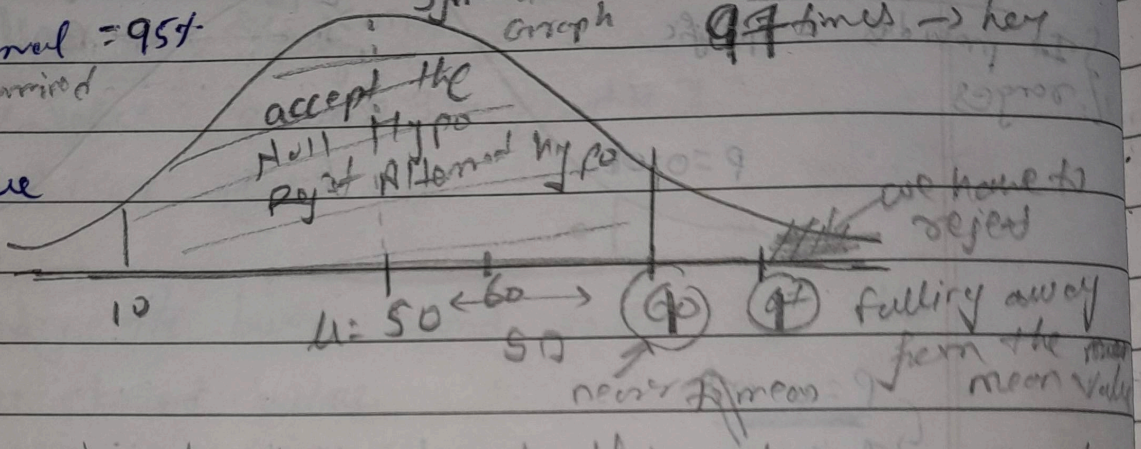
50 time (tail) $Pr(T) = \frac{1}{2}$

perform Experiment =

90 times → Head

97 times → her

Confidere Interval = 95% is determined by Significance value



we should always get this value nearer to the mean

* P - Value

⇒ P value is a measures of the probability that an observed difference could have occurred just by random chance. The lower the p-value, the greater the statistical significance of the observed difference. P value can be used as an alternative to or in addition to pre-selected confidence levels for hypothesis testing.

(14) * Confidence Interval

This is v. imp in order to find out CI

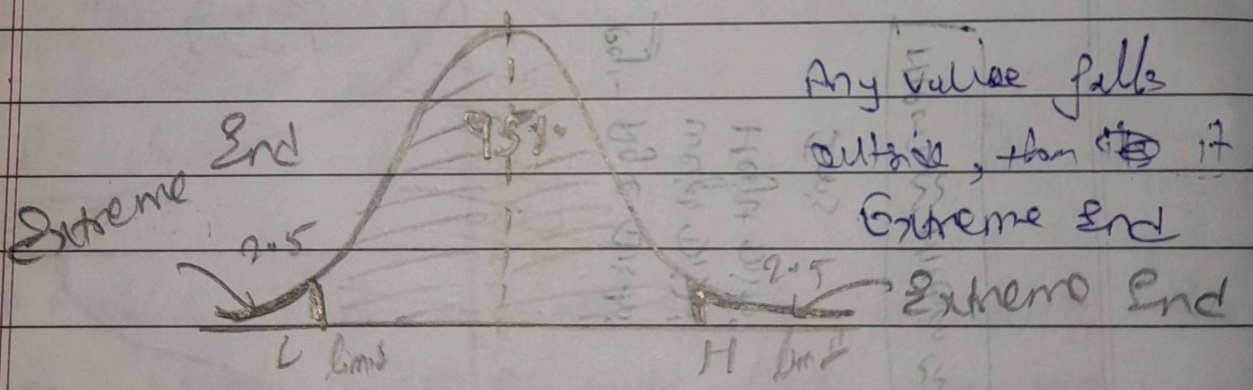
(1) Point Estimate :- $\bar{x}$ → basically calculate the parameter (μ) → μ

• with the help of $\bar{x}$ we will try to estimate the parameter of population mean (μ)

• with some CI = 95%, i will able to find out the range of the population mean (μ) from [L to H].

95% CI range { ↑ Lower Limit — ↑ Higher Limit

This is Basically my range that i can get w.r.t the Population mean (μ)



Any value falls between Lower Limit and Higher Limit, they are Basically CI.

(15) * Bernoulli distribution - Mean, Variance and Standard Deviation

~~In BD will have only~~ The outcome ~~would be~~ BD will always 2 outcome 2

** Eg:- Tossing a coin
up → H, T
P(H) = 0.5
P(T) = 0.5

** • Suppose if i say that whenever my random variable gets a tail,
X = 1 → (Success)

• If my random ~~variable~~ variable has zero this will basically be my fail,
X = 0 → (fail)

for this the P(X = x) = $p^x (1-p)^{1-x}$ → Probability Mass function

X (random variable) will have some value (x)

Note:
In Probability Density function outcome in → Continuous value
In Probability Mass function outcome in → Discrete value

Only two o/p → P
1 - P = q

x will usually have, ⊛ x = 1 & 1

P of getting success
↳ P(Success) = p

P of getting failure
↳ P(failure) = 1 − p = q

} based on this two probability we defined Bernoulli Distribution.

Derive

$P(x=0) = p^0 (1-p)^{1-0} = (1-p) \Rightarrow q$

$P(x=1) = p^1 (1-p)^0 = p$

$$PMF = \begin{cases} q = 1-p & \text{if } x=0 \\ p & \text{if } x=1 \end{cases}$$

$$\therefore PMF = p^x (1-p)^{1-x}$$

eg fair coin

(1)
- P(H) = p = 1/2
- P(T) = 1 − p = 1 − 1/2 ⇒ 1/2

(2)
- P(H) = 0.4 = p
- P(T) = 1 − 0.4 = 0.6

"i have one Experiment work"

⇒ In BD we have values like, p and q = 1 − p

Always Values between [0 & 1]

Mean, Variance & SD

- **Mean:**

$$E(x) = \sum_{i=1}^{x} x \cdot p(x)$$

Expected value of x

Eg: $x = 0$ or 1

$p(x=0) = 0.4 \Rightarrow 1-p$

$p(x=1) = 0.6 \Rightarrow p$

$$Ex = \sum_{i=1}^{x} x \cdot p(x)$$

$= 0(0.4) + 1 \times (0.6)$

$= \underline{0.6\ p}$

- **Variance (σ^2)**

$p(1-p) = pq$

σ (SD) $= \sqrt{pq}$

Mean	Variance	SD
p	pq	$\sqrt{pq}$

(16) 5 Number Summary And How to handle outliers using IQR [Removing the outlier]

1) Minimum → 1
2) 25% → 3 first Quartile (Q_1)
3) Median → 6 (central Element)
4) 75% → 9 Third Quartile (Q_3)
5) Maximum → 10

Percentiles

50% of this entire distribution is less than 6 value

value which is my 50% element

1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 10

↳ 25% value

↳ 75%

⇒ Calculate 25%

$$\text{Percentile} = \frac{25}{100} \times (n+1)$$ n = no. of elements

$$= \frac{25}{100}(13) = 3.25 \approx 3$$

Calculate 75%

$$\text{Percentile} = \frac{75}{100}(n+1)$$

$$= \frac{75}{100}(13) = 9.75 \approx 10$$

$$IQR = Q_3 - Q_1$$

$IQR = 9 - 3 \Rightarrow 6$

Dataset ⇒ 1, 2, 3, 4, 5, 5, 6, 7, 8, 9, 9, 10, 27 (outlier)

[Lower bracket ---------- Higher bracket]

↳ whichever values within this particular distribution follows between this lower bracket and higher bracket only those element will be kept.

↳ Remaining all the elements that are outside of this particular bracket will be treated as outlier.

Lower Bracket = $Q_1 - 1.5(IQR)$

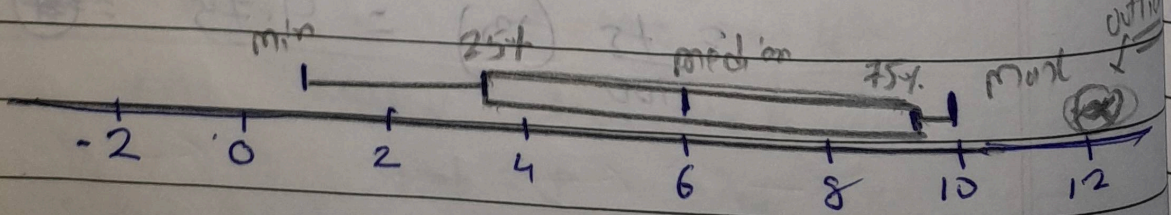
Higher Bracket = $Q_3 - 1.5(IQR)$

LB = $3 - 1.5(6)$ = -6

HB = $9 + 1.5(6)$ = 18

[LB (-6) ---------- HB 18]

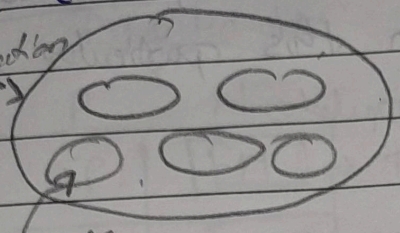
Elements not falling in between can be removed



(17) * Types of Sampling Techniques

① Random Sample Technique

eg: Exit Pole (election)



from Population we take Sample

Party A →
B →
C →

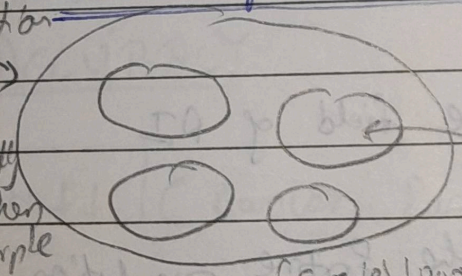
Here we are randomly select some kind of samples and we some kind of conclusions w.r.t the population

Layers or Groups
Partition
They Non overlapping groups

② Stratified Sampling technique

⇒ Stratified sampling technique, we know in a technique where we basically train our model we do the train-test-split in stratified manner

⋆⋆⋆ → Here we provide ration in [ ] : 1



(if it is Equally represented then Stratified sample)

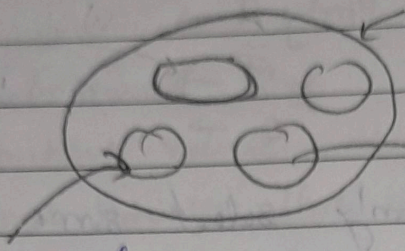
Ratio of Gender in [7:3] [There should be Equal Proportion of Gender]

If we try to pick up the sample with this particular ratio this may become Completely Biased

Because, of this we may not get proper output or proper inferences based on this entire population

③ Systematic Sampling Techniques

Here we are just selecting the n^{th} person and every n^{th} person will try to select 10^{th} person i see in this particular location



Selecting this particular Equation and Every 10^{th} person that ~~the~~ i see i will try to record the statement, like which party this particular guy has actually voted.

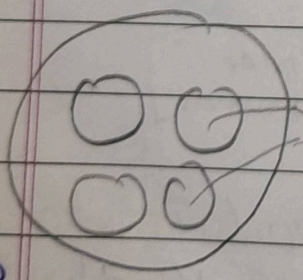
④ Cluster Sampling Technique → Domain

Taking some in the AI Domain

AI → field

↑

Expert in the field of AI



from the Entire population, i just pick up people who are familiar with this domain knowledge

Live class

⑤ Convenience Sampling (or) {Voluntary Response Sampling

Survey → sended → 1000 people → reply → No → Interested → All

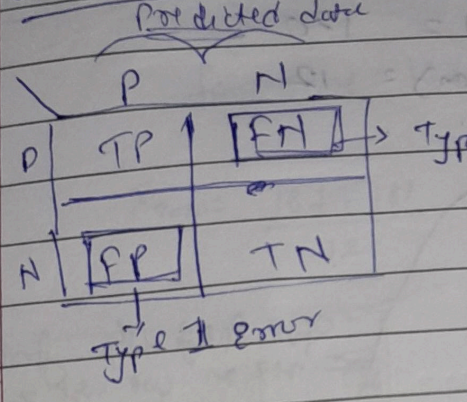
generated at Survey and mailed to 1000 people

only interested people will fill the form

eg:- A survey on Privacy

(18) Reduce false Positive (or) Negative (confusion Matrix)

Confusion Matrix



we should always try to reduce FN↓↓

3 use cases

① whether the person has a disease or not
② whether the market will crash or Not
③ Vaccination side effect or Not

* Most of time the Matrix we are going to use, is ACCURACY

FN↓↓ (Cancer Early Stage treatment, covID)

use case ① Person has a disease or Not
→ we should always try to reduce FN↓↓
In this use case (cancer)

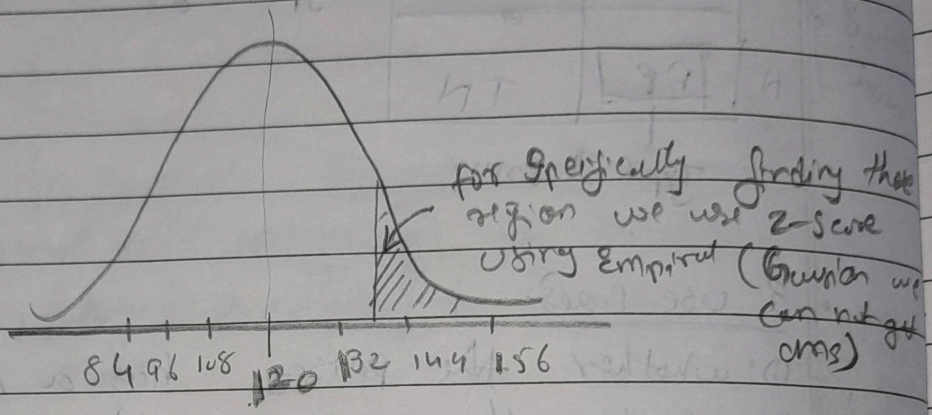
TP↑ TN↑↑ FN↓ FP↓

use ② Market will crash or Not
⇒ Aim: Person should be able to save his money. FN↓↓ we should reduce,

(19) ~~Vaccination side effect~~

Z-Score And its Applications

μ (population mean) = 120
σ (Standard Deviation) = 12



Z-Score will help us to find out whether a value like how far this value is from the mean in the terms of Standard Deviation.

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

$$= \frac{144 - 120}{12} = 2$$

How much SD away from the mean

Application ① Standardization

② Helps us to compare scores between different distribution

Date :
Page :

India perform good in 2021

eg:- India Cricket Team

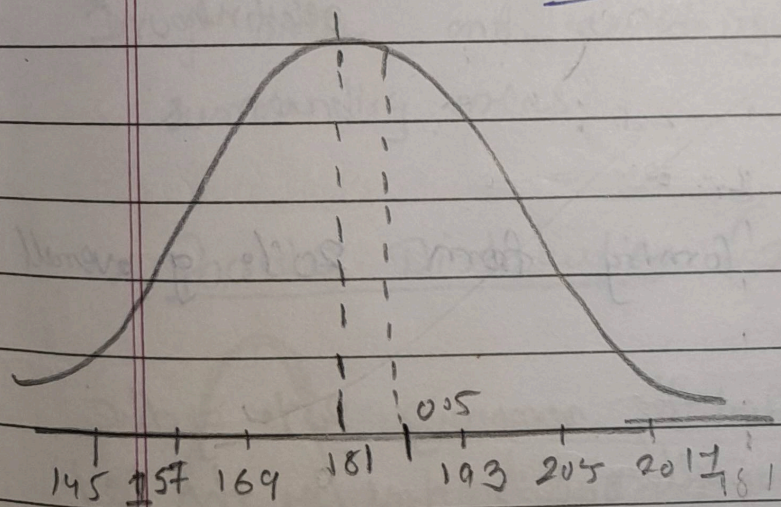
	2020 (cricket)	2021 (cricket)
Assume	Avg = 180	Avg = 182
	$\sigma = 12$	$\sigma = 5$
	final score = 187	final score = 185

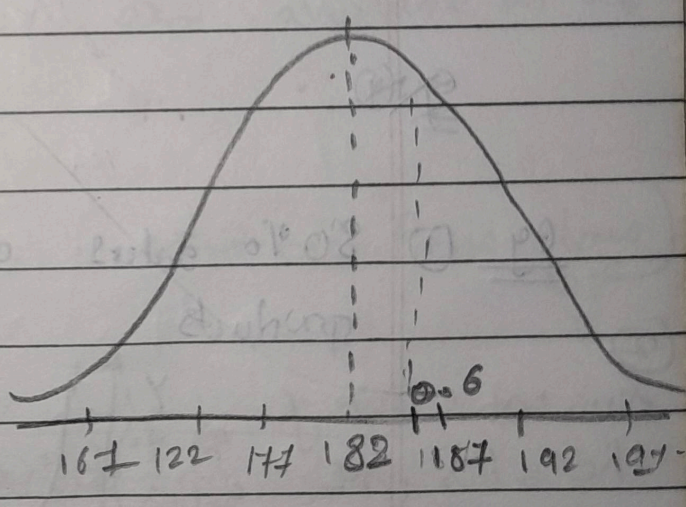
$$Z_{2020} = \frac{187 - 181}{12}$$

$$= 0.5$$

$$Z_{2021} = \frac{185 - 182}{5}$$

$$= 0.6$$

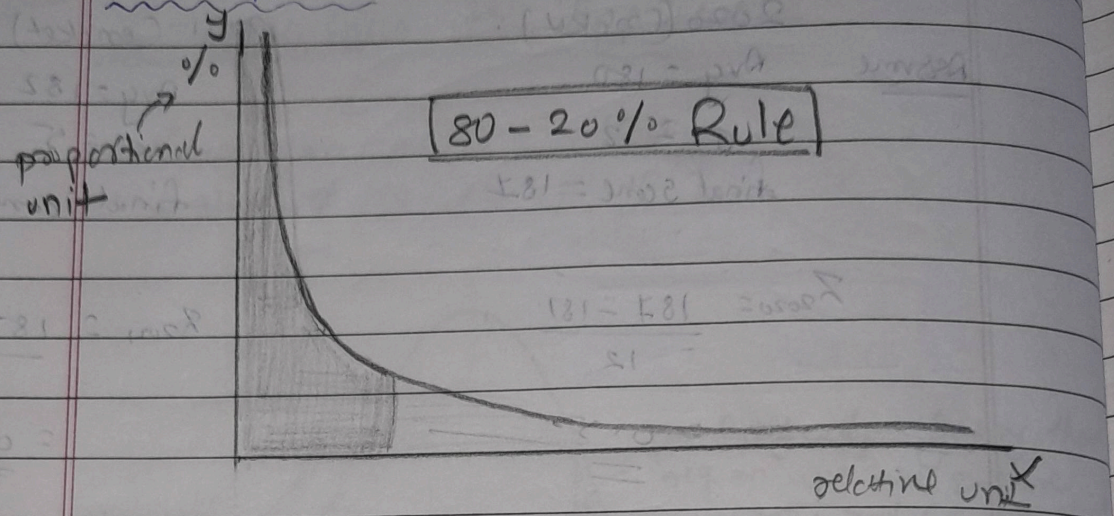




A power law → is the relationship between two quantities where in a relative change in one quantity results in a proportional change in the other quantity

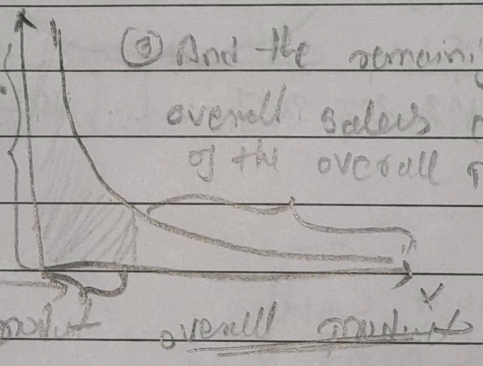
Date :
Page :

(20) Power law Distribution And its Examples and Applications



eg (1) 80% sales are coming from 20% of overall products

(2) This entire 80% of the sales are done by the 20% of the overall products



(3) And the remaining 20% of the overall sales done by the 80% of the overall products

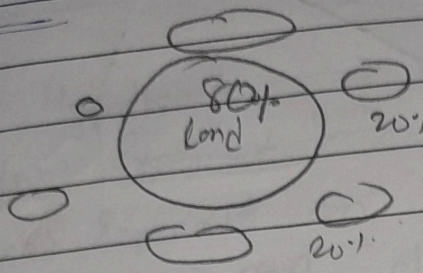
(1) This 20% of the overall product is actually doing the 80% of the Sales.

(2) 80% of windows crash ~~are~~ because of 20% of all the overall bugs.

(3) 80% of Data Scientists uses 20% of overall most common Software products

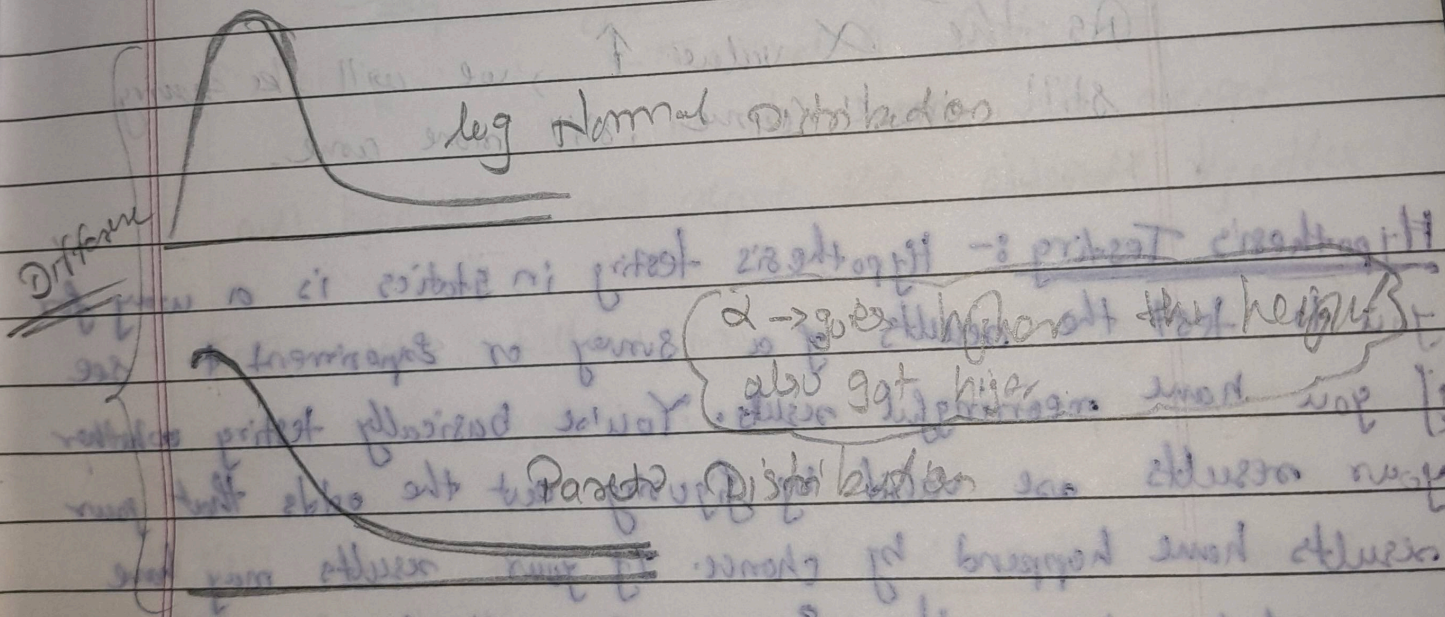
④ 80% of the IPL match are won by 20% of the the team

⑤ Oil fields



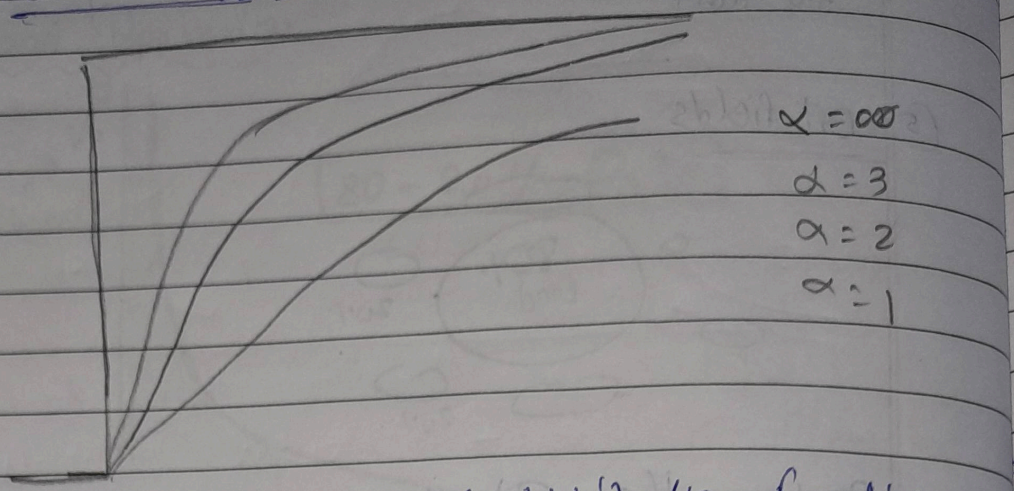
80% of the oils are available in this 20% of area and remaining 20% are available on the surrounding areas

* Pareto Distribution (It is a kind of Power law Dist)



Pareto Type I probability density function for variable x with $x_m = 1$. As $\alpha \to \infty$, then distribution approaches $\delta(x - x_m)$, where δ is the Dirac delta function.

* Cumulative Distribution function



Above Type 1 cumulative distribution functions for various α with $x_m = 1$

Explain: It basically is cumulative summation we just doing cumulative summation over the probability.
As the α value ↑, we will be having still more curve, still more curve.

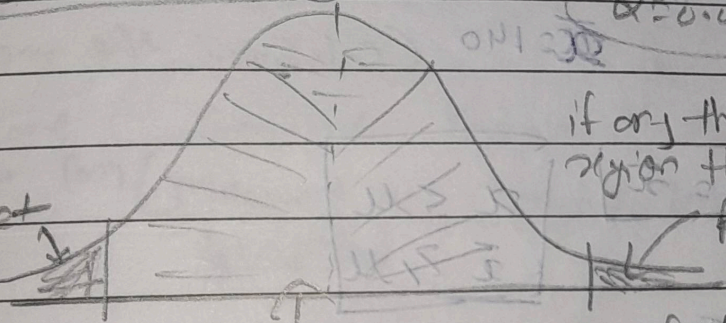
Hypothesis Testing :- Hypothesis testing in statics is a way for you to test the results of a survey or Experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happend by chance. If your results may have happened by chance, the Experiment won't be repeatable and so has little use.

- **Null Hypothesis** = is often an initial claim that is based on previous analyses or specialized knowledge.
- **Alternative Hypothesis** = States that a population parameter is smaller, greater or different than the hypothesized values in the null hypothesis.

How to perform Hypothesis Testing - CI, Z-Test Statistics, Derive Conclusion

- Hypothesis testing

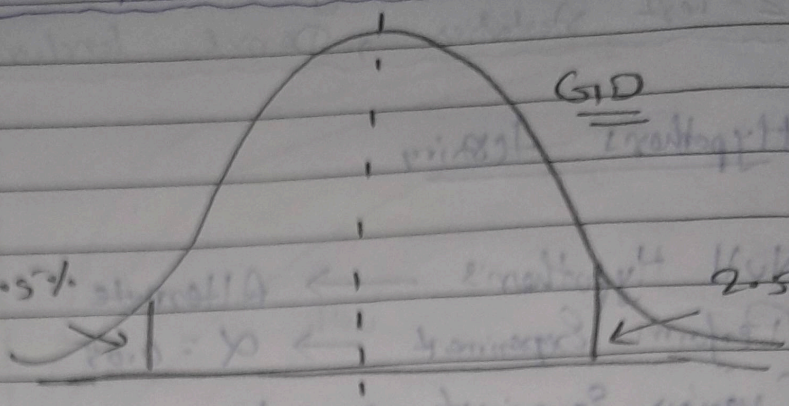
① Null Hypothesis → Alternate Hypothesis
② Perform Experiment with → $\alpha = 0.05$
(various Experiment can be performed like, T-test, Z-Test, chi-square TOA, Anova)
③ C.I based on the (α) value. ← it is depend on the Domain Expertise



→ Test value falls in Extreme END → Reject the Null Hypothesis, Accept the Alternate Hypothesis
Ci → is defined based on the Confidence Interval → (α)

→ Test value fall in
- Accept the Null Hypothesis and
- Reject the Alternate Hypothesis

1 tail test & 2 tail Test



eg: $N = 100$ $\mu = 120$ $\sigma = 5$ {Population}

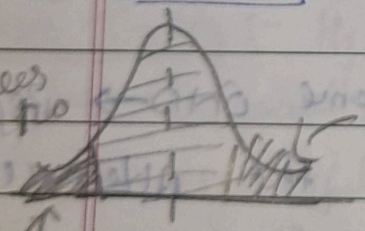
Sample => 30

$\bar{x} = 140$ (calculated the mean)

$\therefore \boxed{\mu = \bar{x}}$ $\boxed{\begin{array}{l} \bar{x} \leq \mu \\ \bar{x} \geq \mu \end{array}}$

$\therefore$ These kind of test is called as 2 tail test

* 1 Tail test



$\bar{x} < 120$

or $\bar{x} < 140$

(can be SD, mean etc)

Hypothesis testing

* Population → with some (Information)
  - use theory
  - Sample from population
* Sample → perform some Experiment {Statistical testing}
  - Conclusion about the Population