

✓ Statistics is the Science of collecting, organizing and Analyzing data → {Facts or pieces of Information that can be measured}

① Descriptive Statistics

↳ consist of organizing and summarizing Data.

Ex: What is the Average age of the students in the class? {85, 68, 79, 92, 76}

{ ① Measure of Central Tendency.
② Measure of Dispersion }

Histogram, Pdf, cdf, Probability, Permutation, Mean, Median, Mode, Variance, Standard deviation

✓ ① Gaussian (Normal) Distribution

✓ ② Lognormal Distribution

✓ ③ Binomial Distribution

✓ ④ Bernoulli's Distribution

✓ ⑤ Pareto Dist {Power law}

✓ ⑥ Standard Normal Dist

✓ ⑦ Transformation and Standardization (Python)

✓ ⑧ Q-Q plot

✓ ⑨ Poisson Distribution (Pareto Dist)

② Inferential Stats

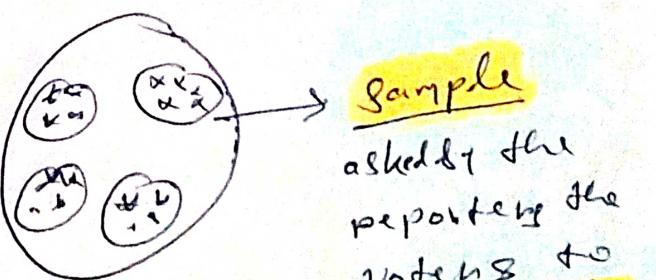
Z test, t test, ANova, CHISquare F Test
Hypothesis Testing (P value), test confidence Interval P

↳ Technique where we used the data that we have collected to form conclusions.

Ex: Are the Marks of the Students of this classroom similar to the Marks of the Maths classroom in the college?

Population(N) & Sample(n)

Elections → 60%, UP, Assam



Sample asked by the reporter the voters to collect data

Different Sample Techniques

① Simple Random Sampling

↳ Every member of the population (N) has an equal chance of being selected for sample (n)

Different

Sampling (i) Techniques

Depend on use
cases we need to
select sampling

combine also

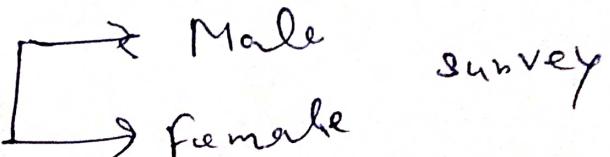
① Simple Random Sampling :

Every member of the Population (N) has an equal chance of being Selected for your Sample (n). eg: Exit Poll survey

② Stratified Sampling :

where the population (N) is split into non-overlapping groups (strata)

eg: Gender



Age group

(0-10) (10-20) (20-40) (40-100)

③ Systematic Sampling :

(N) → n^{th} individual

ex: - Mall → survey (covid) every individual

④ Convenience Sampling :

only those people basically interested or know particular topic ,

eg: ↗ Survey

↗ Data Science .

Statistics

Variable :

A variable is a property that can take on any value.

$$\text{eg: height} = \{ 182, 178, 168, 160 \}$$

$$\text{weight} = \{ 78, 99, 100, 60, 50 \}$$

Two kind of variables :

eg: Age, weight

① Quantitative Variable → Measured Numerically
 { Add, sub, Mul, divide }

② Qualitative / Categorical Variable.

eg: Gender [M, F] { Based on some characteristic & we can derive Categorical Variable }

eg: IQ

$\underline{\underline{0-10}}$	$\underline{\underline{10-50}}$	$\underline{\underline{50-100}}$
↓	↓	↓
<u>Less IQ</u>	<u>Medium IQ</u>	<u>Good IQ</u>

eg: Marital Status

eg: Blood Group.

Quantitative

① Discrete Variable

Eg: Whole Numbers

No. of Bank a/c's

② Continuous Variable

Eg: Height = 172.5, 162 cm

Weight = 100kg, 55kg

Total. of children in a family. eg: 1, 2, 3, 4, 5

(Q) Population of state

- ① Length of a river
- ② Amount of rain fall per year
- ③ Sari length.
- ④ Blood pressure

Statistics

613

Variable Measurement Scales :

4 types of Measurement Variable

- ① Nominal data - { Categorical data } → color, Gender, classes.
- ② Ordinal data - { Order of the data matters, but value does not } e.g. Rank - Marks
- ③ Interval
- ④ Ratio.

$$\begin{array}{rcl} 1 & = & 100 \\ 4 & = & 69 \\ 3 & = & 78 \\ 2 & = & 95 \end{array}$$

Interval data : Order Matters, Values also matter, here natural Zero is not present.

ex : Temperature

Fahrenheit

70-80 80-90 90-100



Frequency Distribution :

Sample dataset : { Rose, lilly, sunflower, Rose, lilly, sunflowers, Rose, lilly, lilly }

Based on flower type	Frequency
Rose	3
Lilly	4
Sunflower	2

(2) Cumulative Frequency

3) added

7) added

9) added

PDF = (Probability density Function)

① Bar Graph (Discrete)

② Histograms : - (Continuous) e.g. ages = { 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 }

① Arithmetic Mean for Population & Sample

In some range of data
Student Ages

① Mean (Average) :

Population (N)

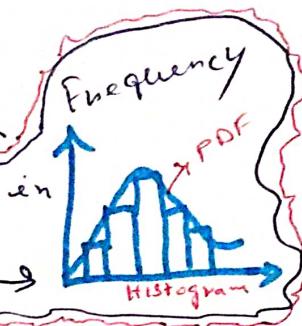
$$n = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$$

$$\mu = \sum_{i=1}^N \frac{n_i}{N}$$

$$= 1+1+2+2+3+3+4+5+5+6 \\ = 32$$

$$= \frac{32}{10} = 3.2$$

use → To remove outliers in the Distributions



Sample (n)

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \\ = 3.2$$

$$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$$

$$\text{Mean} = \frac{32 + 100}{11}$$

$$= \frac{132}{11} = 12 \\ M = 3.2 \\ + 100 \downarrow \text{huge difference}$$

$M = 12$

② Central Tendency :

- ① Mean ✓
- ② Median ✓
- ③ Mode ✓

Refers to the measure used to determine the

center of the distribution of data.

∴ To solve the problem of outliers in Mean we will use Median.

Most cases
to value =

② Median:

$$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112\}$$

C
central element

Two outliers

At in Median we sort the numbers of a dataset
2nd we will take the central element of the dataset.

→ if odd numbers = 11

$$\text{Median} = 3$$

→ if Even number = 12

$$\text{Median} = 3.5$$

$$\text{Avg} = \frac{3+4}{2} = 3.5$$

③ Categorical & Missing data
Mode : It takes the most frequent element

Mode = 6

{1, 2, 2, 3, 3, 4, 5, 6, 6, 6, 7, 8, 100} 100, 100, 100
if Mode = 100
X Problem

2) Magnitude of Dispersion : $\Rightarrow \{ \text{Spread} \}$

① Variance

② Standard Deviation

Q: -

$$N = \begin{cases} \{1, 1, 2, 2, 4\} \\ \{2, 2, 2, 2, 2\} \end{cases} \quad \frac{10}{5} = 2$$

Variance : $\{ \text{Spread} \}$

How to Identify these two distributions are different?

\Rightarrow At that time we will use variance & Standard Deviation

⊕ Population (N) Variance

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{10.84}{6} = 1.81$$

μ μ $\mu - \mu$ $(\mu - \mu)^2$

1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	+0.17	0.03
4	2.83	1.07	1.37
5	2.83	2.07	4.071
$\overline{ }$			10.84
$\mu = 2.83$			

② Sample (n) Variance

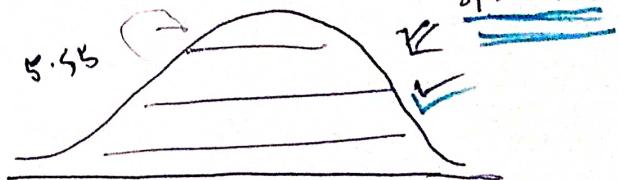
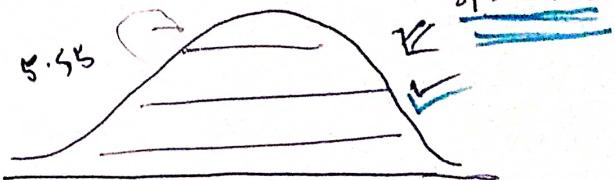
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

1.81

Q:

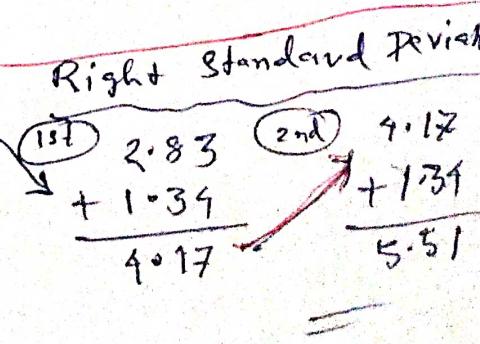
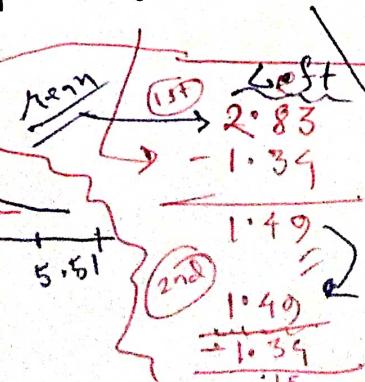
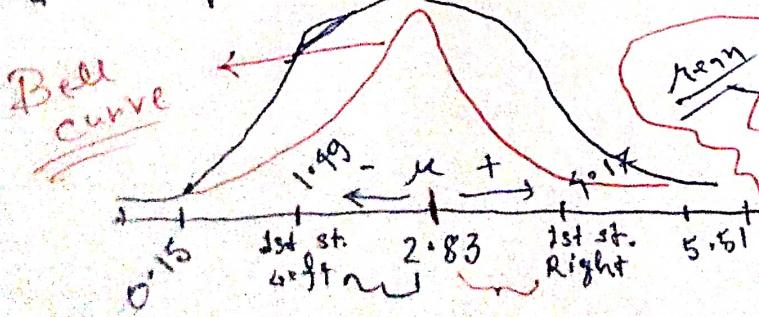
where Variance is more?

\Rightarrow more spread



Standard Deviation : $\{ \text{root of Variance} \}$

$$\sigma = \sqrt{\text{Variance}} = \sqrt{1.81} = 1.345$$



Percentiles And Quartiles {Find outliers?}

e.g:-

Percentage in Distribution : { 1, 2, 3, 4, 5 }

Q1 :- % of the numbers that are Odd ?

$$\Rightarrow \% = \frac{\text{number of numbers that are Odd}}{\text{total Numbers}}$$

$$= \frac{3}{5} = 0.6 = 60\%$$

Percentiles : { A percentile is a value below which a certain percentage of

e.g:-

observation lie. } ... 16, 17, ... 20

Dataset : { 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12 }

Q1 :- What is the percentile ranking of 10 ?

$$n = 10, n = 20$$

Percentile Rank of $n = \frac{\# \text{ of values below } X}{n} \times 100$

Q1 :- 11 ?

$$= \frac{17}{20} \times 100 = 85\%$$

$$= \frac{16}{20} \times 100 = 80\%$$

Q1 :-

$$= 80\% \checkmark$$

② What value exists at percentile ranking of 25% ?

$$\text{Value} = \frac{100}{100} \times (n+1)$$

$$= \frac{25}{100} \times (21)$$

$$= 5.25 \rightarrow \text{Index Position}$$

$$\text{Value} \rightarrow 5 \rightarrow 25\%$$

Q1 :- 75% ?

$$= \frac{75}{100} \times 21$$

$$= 15.75 \text{ (Index)}$$

$$\text{Value} \rightarrow 9 \downarrow$$



Five Number Summary :

1 = ① Minimum

3 = ② First Quartile (Q_1)

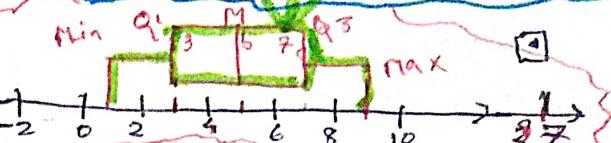
5 = ③ Median

7 = ④ Third Quartile (Q_3)

9 = ⑤ Maximum

Box plot

With the help of these → 5 will remove the outliers



e.g.: currently min

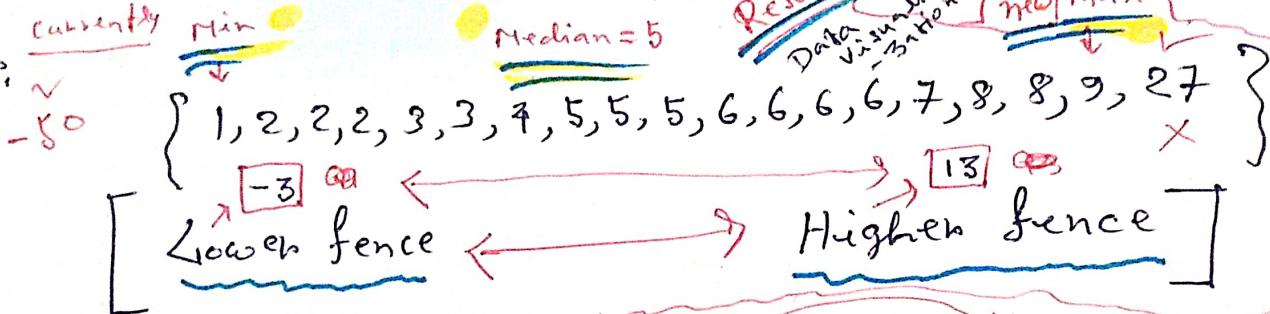
Median = 5

Result

Data visualization

new max

outlier
present state



$$\text{Lower fence} = Q_1 - 1.5 \text{ (IQR)}$$

$$\text{Upper fence} = Q_3 + 1.5 \text{ (IQR)}$$

$$Q_1 = (25\%)$$

$$Q_3 = (75\%)$$

$$\text{InterQuartile Range (IQR)} = Q_3 - Q_1$$

$$Q: - 25\% ? \text{ (Percentile)} / Q: - 75\% ? \text{ (Percentile)}$$

$$\Rightarrow \frac{25}{100} \times (19 + 1)$$

$$= \frac{25}{100} \times 20$$

$$= 5 \rightarrow \text{Index}$$

$$Q_1 = \underline{\underline{3}} \rightarrow \text{Value}$$

$$\therefore \text{InterQuartile Range is (IQR)} = Q_3 - Q_1$$

$$Q_3 \Rightarrow \underline{\underline{7}} \rightarrow \text{Value}$$

$$\textcircled{3} \text{ Higher fence} = Q_3 + 1.5 \text{ (IQR)}$$

$$= 7 + 1.5 \times 4$$

$$= 7 + 6$$

$$= \underline{\underline{13}}$$

$$\text{Now Lower fence} = Q_1 - 1.5 \text{ (IQR)}$$

$$= 3 - 1.5 \times 4$$

$$= 3 - 6$$

$$= \underline{\underline{-3}}$$

$$= 7 - 3$$

$$= \underline{\underline{4}}$$

$$\therefore \text{IQR} = \boxed{4}$$

Advance Statistics

Distribution :

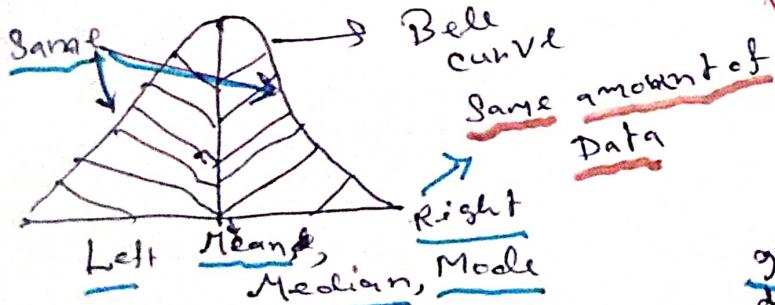
- ✓ Normal / Gaussian Distribution
- ✓ Standard Normal Distribution
- ✓ Z-Scores (/ Z-Table) \Rightarrow
- ✓ Log-Normal Distribution
- ✓ Bernoulli Distribution
- ✓ Binomial Distribution

Practical

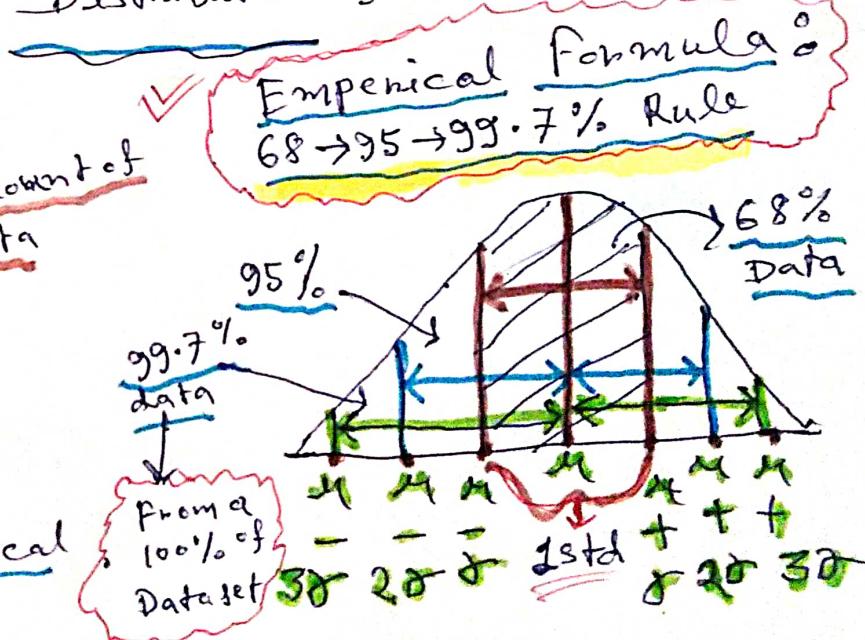
- ① Mean, Median, Mode
- ② Variance, Standard Dev.
- ③ Histogram, PDF, Bar plot, violin plot
- ④ IQR
- ⑤ Log-Normal Distribution

To visualize a set of data we use Distribution through various Graphs . e.g.: Histogram, pdf . to create report or for data Analysis etc.

① Gaussian / Normal Distribution :



Right part to the left part of the distribution exactly similar / symmetrical

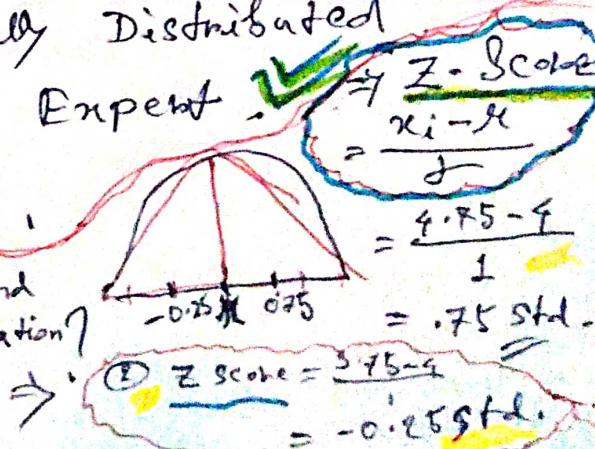


e.g.: Height (Dataset) \rightarrow Normally Distributed

\hookrightarrow According to Domain Expert

② weight, ③ Iris DATASET .

e.g.: Mean = 4 Std = 1 where the value 4.75 and 3.75 will fall in terms of standard deviation?



Advance Statistics

② Standard

Normal Distribution

It is a Distribution that a set of normally Distributed Data converts by the Z-Score to a Standard Normal Distribution set of Data.

e.g:

$$\{1, 2, 3, 4, 5, 6, 7\}$$

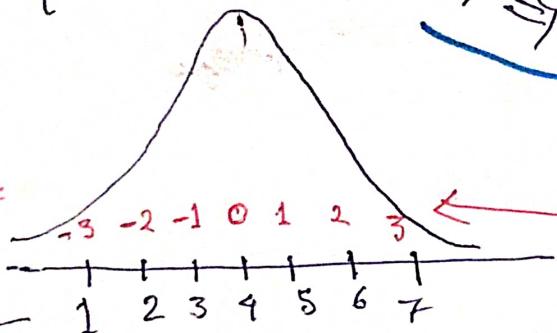
Normal / Gaussian

Distribution

$$x_i - \mu$$

$$\mu = 4$$

$$\sigma = 1$$



$$Z(1) = \frac{1-4}{1} = -3, \quad Z(3) = \frac{3-4}{1} = -1$$

$$Z(2) = \frac{2-4}{1} = -2, \quad \dots$$

Practical Application

Standardization

Converted to
 $\mu = 0, \sigma = 1$

(Value - mean) / std.

Z score works internally

Dataset:

Age → (year)
Salary → (RS)
Weight → (kg)

Standard Normal Distribution

Age	Salary	Weight
24	40k	70
25	80k	80
26	60k	55
27	70k	95

Z score

what percentage of scores fall above

$$4.25?$$

H-Score

$$Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

$$1 - 0.5987$$

from some Z-table

$$= 0.4013$$

$$= 40\%$$

Standard Normal Distribution

$y \sim SND(\mu=0, \sigma=1)$

Normalization

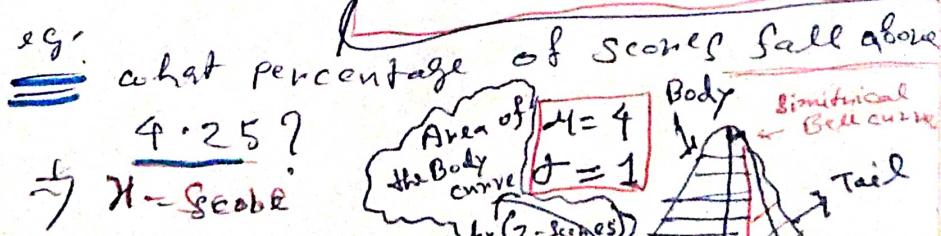
(Value - min)

(max - min)

upper

Min Max scales bound

(values 0 to 1)



3 Types of Z-Scores
we get.

Coding

```
⇒ import pandas as pd  
import seaborn as sns  
import numpy as np  
import matplotlib.pyplot as plt  
%matplotlib inline
```

```
import statistics
```

mean, median, mode

```
df = sns.load_dataset('tips') ← # Define Datasets  
# Load Database
```

```
df.head()
```

=
np.mean(df['total_bill'])

=
np.median(df)

=
statistics.mode(df['total_bill'])

=
sns.boxplot(df['total_bill'])

=
sns.histplot(df['total_bill'])

=
sns.histplot(df['total_bill'], kde=True) # Probability density function (PDF)

```
df1 = sns.load_dataset('iris')
```

```
df1.head()
```

=
sns.countplot(df1['species']) # Barplot/Bargraph

=
np.percentile(df1['sepal_length'], [25, 75])

=
df2 = pd.read_csv("dat.csv")

```
df2.head()
```

=

foot:

Google Colab Pro
or
Jupyter Notebook

Collaboratory



IQR / outliers | Z Score

Code

| lower fence

upper fence

→ import seaborn as sns
 import numpy as np
 import matplotlib.pyplot as plt
 %matplotlib inline

define Dataset

dataset = [11, 10, 12, 14, 12, 15, 14, 13, 15, 102, 12, 14, 17, 19, 107, 10, 13, 12, 108, 12, 11, 14, 13, 15, 10, 15, 12, 10, 14, 13, 15, 10]

plt.hist(dataset)

outliers = []

def detect_outliers(data):

threshold = 3 # 3 std deviation

mean = np.mean(data)

std = np.std(data)

for i in data:

z-score = (i - mean) / std

if np.abs(z-score) > threshold:

outliers.append(i)

return outliers

detect_outliers(dataset)

→ [102, 107, 108]

dataset = sorted(dataset)

dataset

q1, q3 = np.percentile(dataset, [25, 75])

print(q1, q3)

iqr = q3 - q1

print(iqr)

lower_fence = q1 - (1.5 * iqr)

upper_fence = q3 + (1.5 * iqr)

print(lower_fence, upper_fence)

sns.boxplot(dataset),

google Colab
or
Jupyter Notebook

Probability

Probability is a measure of the likelihood of an Event.

Eg: Roll a dice $\{1, 2, 3, 4, 5, 6\}$

$$\Pr(6) = \frac{\# \text{ of way an event can occur}}{\# \text{ of possible outcome}}$$

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

" Toss a coin $\{H, T\}$

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

Q2: Roll a Dice if 1, 3, or 6?
 $\Pr(1 \text{ or } 3 \text{ or } 6) = \Pr(1) + \Pr(3) + \Pr(6)$
 $= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$
 $= \frac{3}{6} = \frac{1}{2} = 0.5$

① Addition Rule (Probability, "OR")

① Mutual Exclusive Event:

Two Events are mutual exclusive if they cannot occur at the same time.

Eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

② Non Mutual Exclusive:

Multiple events can occur at the same time. $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$

Eg: Deck of Cards $\{Q, \heartsuit\}$

combination

Q1: If I Toss a coin, what is the probability of the coin landing on heads or tails?

Mutual Exclusive (Additional Rule)

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) \quad \text{or} \quad \Pr(A \text{ or } B) = 1$$
$$= \frac{1}{2} + \frac{1}{2} = \frac{1}{2}$$

Probability

Non Mutual Exclusive / Additional Probability

You are picking a card randomly from a deck. What is the probability of choosing a card that is queen or a heart?

$$\Rightarrow P(Q) = \frac{4}{52} \quad P(H) = \frac{13}{52} \quad P(Q \text{ and } H) = \frac{1}{52}$$

Addition Rule for non mutual Exclusive events

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

$$P(Q \text{ or } H) = P(Q) + P(H) - P(Q \text{ and } H)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{16}{52}$$

~~≈ 0.30769~~ something like this

② Multiplication Rule

Independent Event

Eg: Rolling a dice {1, 2, 3, 4, 5, 6}

Dependent Event

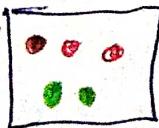
Eg: 1, 1, 2, 3, 5, 5, 6, 9. Each and every one is independent

Eg:

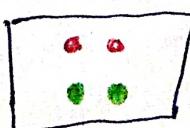
$$P(Red) = \frac{3}{5} \xrightarrow{\text{Impacted after removing}} P(Green) = \frac{2}{4}$$

From A Box

has Total 5 Marbles 3 Red 2 Green



removed 1 Red then



Naive Bayes
(conditional Probability)

Probability

Independent Event / Multiplication

Q :- what is the probability of rolling a "5" and then a "4" in a dice?

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

$$P(5 \text{ and } 4) = \frac{1}{6} * \frac{1}{6}$$

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

Dependent Event

Q :- what is the probability of drawing a Queen and then a Ace from a deck of cards?

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

conditional probability

$$\frac{3}{5} \quad \frac{2}{4}$$

Bay's Theorem

$$P(G \text{ and } R) \downarrow$$

$$= P(G) * P(R/G)$$

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

$$= \frac{4}{52} * \frac{1}{51}$$



Probability

Permutation & Combination

eg: School trip } chocolate factory } → Dairy, strawberry, Kedgeree, Milky bar, Silk, Mix fruit.

Student can pick Any 3 of 6 types of different chocolate for free at one time.

$$\Rightarrow \{ \text{Dairy, Silk, Milky} \} \quad \text{Total chocolates} = n = 6 \\ \{ \text{Milky, Silk, Dairy} \} \quad \rightarrow \begin{array}{l} n=6 \\ r=3 \end{array}$$

can repeat same elements

↓ Permutation formula: $n_{P_r} = \frac{n!}{(n-r)!}$

$$= \frac{6!}{(6-3)!}$$

$$= \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 120$$

Combination

Unique combination Each Time.

Dairy Silk Milky

$$n_{Cr} = \frac{n!}{r!(n-r)!} = \frac{6!}{3!(6-3)!}$$

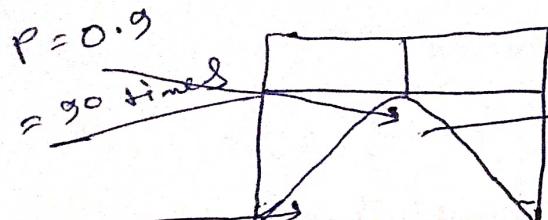
$$= \frac{6 \times 5 \times 4 \times 3!}{3! \times 2 \times 1 \times 3!}$$

$$= 20$$

Influential Statistics

① P Value

Every 100 time I touch the mouse pad 80 times
I touched this specific region.



= only one time

P = Probability
 $P = 0.8$
= 80 times

Alpha

CI

Li

② Hypothesis testing, Confidence Interval, Significance value

Coin → Test whether this coin is a fair coin or

not by performance.

In 100 times tossed

$$P(H) = 0.5, P(T) = 0.5$$

if
eg: $P(H) = 100\%$.
unfair

gloomy Slim
coin
50% 50%

But,

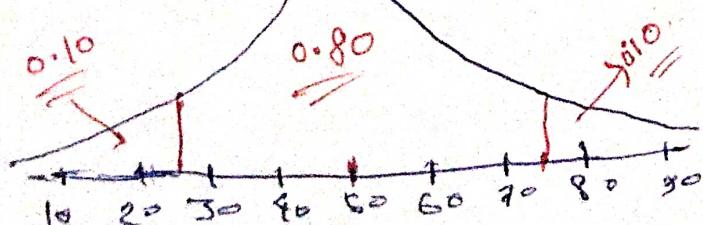
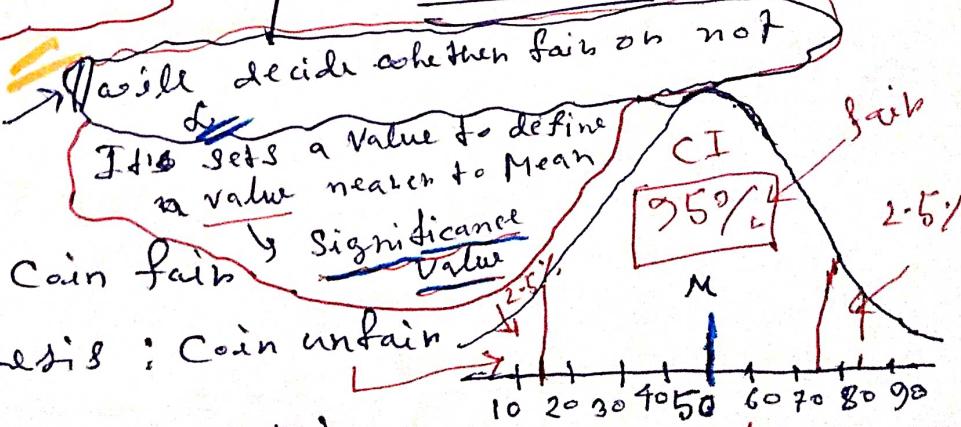
Hypothesis Testing

① Null Hypothesis: Coin fair

② Alternate Hypothesis: Coin unfair

③ Experiment (eg: t-test, z-test, chi test)
↓ (based on it)

④ Reject or Accept the Null Hypothesis



Value decide when
the Fair point will
fall.
Alpha $\alpha = 0.20 \rightarrow 20\%$.
then $CI = 80\% \rightarrow 0.80$
Confidential Interval

Influential Statistics

① Type I and Type II Errors:

Null Hypothesis (H_0) = coin is fair

Alternative Hypothesis (H_1) = coin is not fair

Reality Check

Null hypothesis is true or Null hypothesis is false.

Decision: (Always 4 Types of Outcomes)

Null hypothesis is true or Null hypothesis is false

outcome 1: we reject the Null Hypothesis, when in reality it is false → Yes

outcome 2: we reject the Null Hypothesis, when in reality it is true → Type I Error

outcome 3: we Accept the Null Hypothesis, when in reality it is false → Type II Error

outcome 4: we Accept the Null Hypothesis, when in reality it is true → Yes

e.g.: Confusion Matrix:

P N

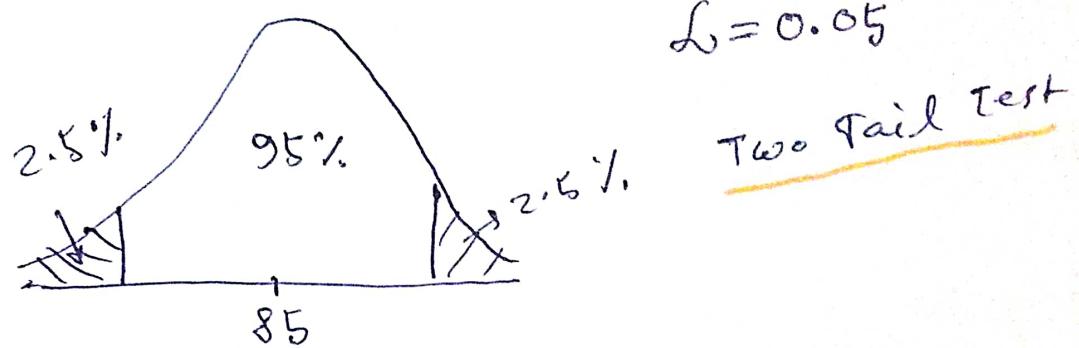
T	TP	TN
F	FP	FN

1. Type I Error 2. Type II Error

Inferential Statistics

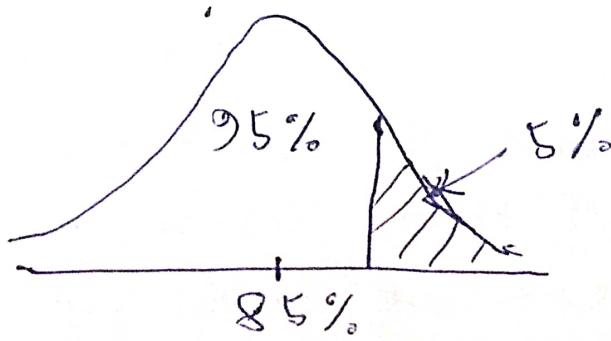
② 1 Tail and 2 tail Test :

Eg: Colleges in Karnataka have an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88% with a standard deviation 4%. Does this college has a different placement rate?



Does this college have a placement rate greater than 85%?

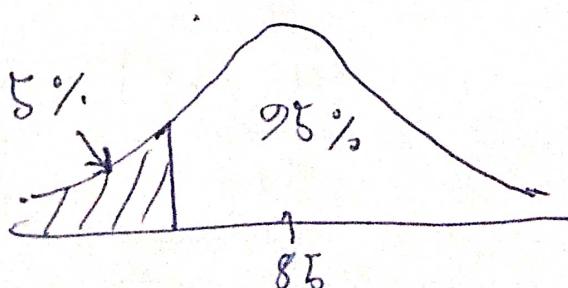
$$L = 0.05$$



(ob)

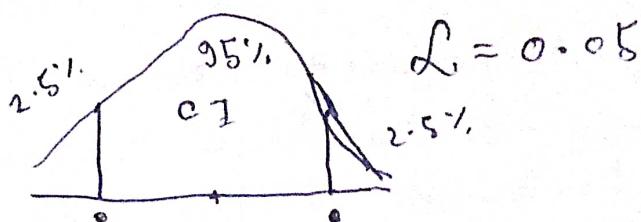
Does this college have a placement rate lesser than 85%?

$$L = 0.05$$



Influential Statistics

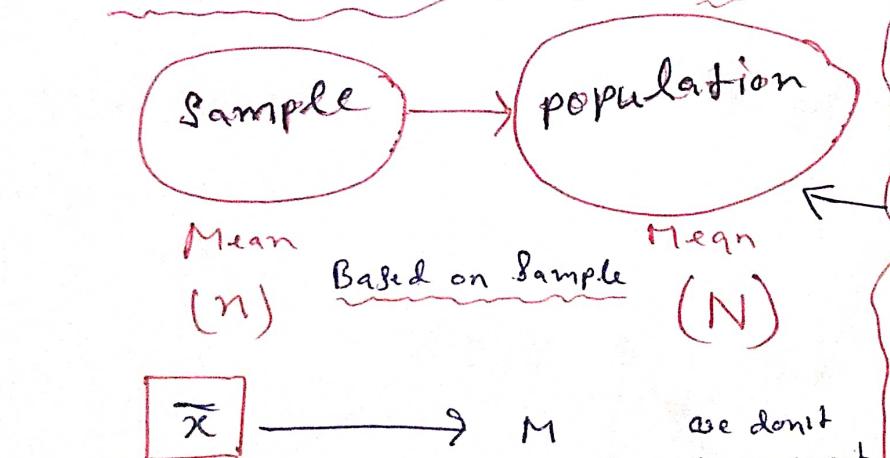
③ Confidence Interval (How to find) ?



Point Estimate

The value of any statistics that estimate the value of a parameter is called Point Estimate.

In Influential stats



$$\bar{x} = 2.9 \quad \text{Approximately } M = 3 \text{ members}$$

Confidence Interval

Point Estimate \pm Margin of Error

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

Standard Error

$$\text{Upper Bound} = \bar{x} + Z_{0.05} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\text{Lower Bound} = \bar{x} - Z_{0.05} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$$

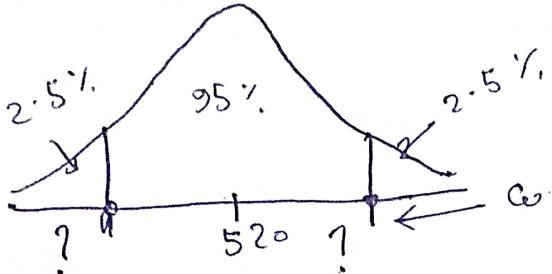
In Influential stats
Based on sample data are try to estimate the all population data.
eg: reporter collecting about the vote report from sample of public.

Inferential Statistics

Z-Test, Z-Table:

Q:- On the quant test of CAT Exam, the population standard deviation is known to be 100. A sample of 25 test taken has a mean of 520 score. Consider for simple calculation (3σ) a 95% CI about the mean?

$$\Rightarrow \sigma = 100, n = 25, \alpha = 0.05, \bar{x} = 520$$



what will be at those position?

2 of them
when fullfill

① whenever Population Std. is given, Then Z Test
usually ② $n \geq 30$

Point Estimate \pm Margin of Error

$$\bar{x} \pm Z_{\alpha/2} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) \rightarrow \text{Standard Error}$$

$$Z_{0.05} = Z_{0.025}$$

$$1 - 0.025 = 0.975$$

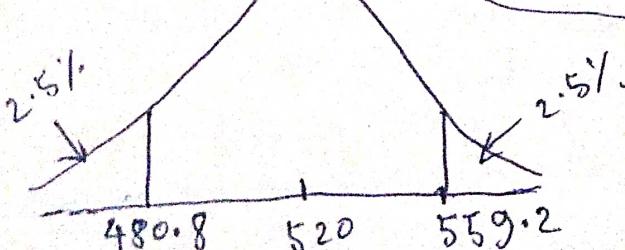
According to Z-table
Value is position
 $\rightarrow 1.96 (a6)$

$$\text{Upper Bound of CI} = \bar{x} + Z_{0.05} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = \bar{x} + 1.96 \cdot \left(\frac{100}{\sqrt{25}} \right)$$

$$\text{Lower Bound of CI} = \bar{x} - Z_{0.05} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = \bar{x} - 1.96 \cdot \left(\frac{100}{\sqrt{25}} \right)$$

$$\text{Upper} = 520 + 1.96 (20) = 559.2$$

$$\text{Lower} = 520 - 1.96 (20) = 480.8$$



1.96
one tail

0.975% 0.025%

Influential Statistics

T-Test: (to find values of CI Range)

Q) On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a standard deviation of 80. Construct 95% Confidence Interval about the Mean?

⇒ Condition

Since population std is not given → t-test

$n = 25$

$$\bar{x} = 520 \quad s = 80$$

$$(\text{Alpha}) \rightarrow \alpha = 0.05$$

↑ Significance value

Sample std

Point Estimate \pm Margin of Error

$$\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right) \rightarrow \text{Standard Error}$$

$$\text{Upper Bound} = \bar{x} + t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

$$\text{Degree of Freedom} = n - 1 = 25 - 1 = 24$$

$$\begin{aligned} &= 520 + 2.064 \cdot \left(\frac{80}{\sqrt{25}} \right)^{16} \\ &= 553.024 \end{aligned}$$

0.05	2	0.025
2	2	0.025
2	2	0.025
2	2	0.025
2	2	0.025

→ t-Table

	0.025	2
2	2	0.025
2	2	0.025
2	2	0.025
2	2	0.025

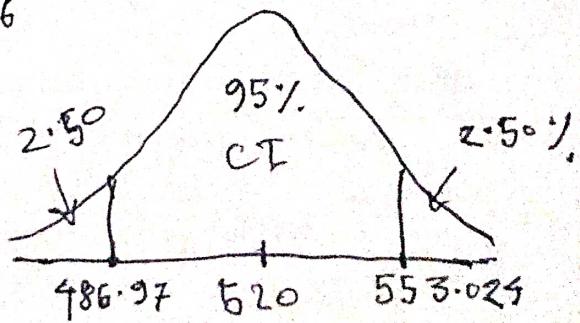
24 = 2.064

$$\text{Lower Bound} = \bar{x} - t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right)$$

$$= 520 - 2.064 \cdot \left(\frac{80}{\sqrt{25}} \right)^{16}$$

$$= 486.97$$

$$[486.97 \longleftrightarrow 553.024]$$



Influential Statistics

① One Sample Z - Test (Hypothesis Testing) :

Rules

- ① Population std is given
- ② Sample size $n \geq 30$

Q:- In the population, the average IQ is 100 with a std. of 15. Researchers wants to test a new medication to see if there is a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean IQ of 140. Did the medication effect the intelligence? (Increase or Decrease) $\alpha = 0.05$, C.I = 95%

⇒ 1) Define Null Hypothesis is $H_0: M = 100$ $S = 140$

$$\bar{x} = 140$$

2) Alternate Hypothesis when $H_1: M \neq 100$

$\alpha/2 = 0.025$

3) State Alpha

$$\alpha = 0.05$$

4) State Decision Rule

2 tail

5) Calculate Z Test Statistics

-1.96

95%

2.5%

+1.96

2 Tail Test

$$1 - 0.025 = 0.975$$

From Z Table

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

when huge Data

$$\left\{ \frac{\sigma}{\sqrt{n}} \right\} \rightarrow \text{Standard Error}$$

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}}$$

$$= \frac{40}{15} \times \sqrt{30}$$

$$= 14.60$$

Ans State out Decision	
1	$14.60 > 1.96$ $Z = 14.60$
if Z is less than -1.96 or greater than 1.96, Reject the null hypothesis	
Medication improve the Intelligence Or Decrease? (Improve Very Much)	

Influential Statistics

② One Sample T-test (Hypothesis testing)

\Leftarrow Z-test \Rightarrow population std.

\Leftarrow t-test \Rightarrow unknown population std.

Q:- Population the Average IQ = 100 (Same Question)

$$n = 30, \bar{x} = 140, S = 20$$

Did the medication affect intelligence?

$$\alpha = 0.05$$

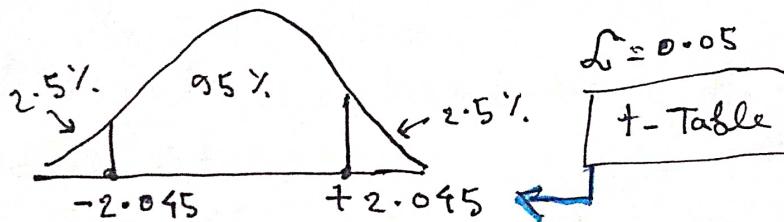
\Rightarrow ① $H_0 = M = 100$

\Leftarrow ② $H_1 = M \neq 100$

\Leftarrow ③ calculate the degree of freedom

$$n-1 = 30-1 = 29$$

④ State Decision Rule



\Rightarrow ⑤ T-Test

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

$$= 10.96$$

$$\begin{aligned} \bar{x} &= 140 \\ \mu &= 100 \\ s &= 20 \\ n &= 30 \end{aligned}$$

\Rightarrow ⑥ State out decision

$$T = 10.96 > 2.045$$

Increased the Intelligence.

So, ① Reject the Null Hypothesis

② $P \leq$ Significance Value

③ Accept the Alternative Hypothesis.

Influential Statistics

Chi Square Test :

chi square Test claims about population proportions
It is a non parametric test that is performed on categorical data (Nominal or Ordinal data).

Q :- In the 2000 Indian Census, the age of the individual in a small town were found to be the following :

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

In 2010, age of $n=500$ individuals were sampled.
Below are the results

< 18	$18 - 35$	> 35
121	288	91

Using $\alpha = 0.05$, could you conclude the population distribution of ages has changed in the last 10 years?



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

{ Population } 2000

Expected

< 18	$18 - 35$	> 35
121	288	91
500×0.2 $= 100$	500×0.3 $= 150$	500×0.5 $= 250$

$n = 500$ { Sample }

\Rightarrow Observed = f_0

\Rightarrow Expected = f_e

(continue)

(continued)

Influential Statistics

< 18	$18 - 35$	> 35
121	288	91
100	150	250

$\Rightarrow n=3$ categories

\Rightarrow observed (f_o)

\Rightarrow Expected (f_e)

① H_0 = The data meets distribution 2000 census

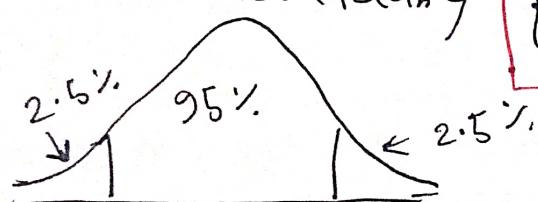
H_1 = The data does not meet in 2000 census

② $\alpha = 0.05$ (95% CI)

③ Degree of freedom = $n-1 = 3-1 = 2$

$\rightarrow df=2, \alpha=0.05$

④ Decision Boundary



{ check in the Chi Square Table }

2 tail Test

χ^2

denoted by

from χ^2 Table

If χ^2 is greater than 5.99 reject H_0

⑤ calculate Chi Square Test statistic.

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

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

$$= 232.494$$

$$\therefore \chi^2 = 232.494 > 5.99 \quad \left\{ \begin{array}{l} \text{Reject the Null Hypothesis} \\ H_0 \end{array} \right.$$

Coding Z-test

Influential Statistic 8

Q:- Suppose the IQ in a certain population is normally distributed with a mean of $\mu = 100$ and standard deviation of $\sigma = 15$. A researcher wants to know if a new drug affects IQ levels, so he recruits 20 patients to try it and records their IQ levels. The following code shows how to perform a one sample Z-test in python to determine if the new drug causes a significant difference in IQ levels.

=> from statsmodels.stats.weightstats import ztest
 => as ztest

Enter IQ levels for 20 patients

data = [88, 92, 94, 99, 96, 97, 97, 99, 99, 105, 109, 109, 110, 112, 112, 113, 114, 115]

ztest(data, value = 100)
 # if 110 = 0.002 (Reject)

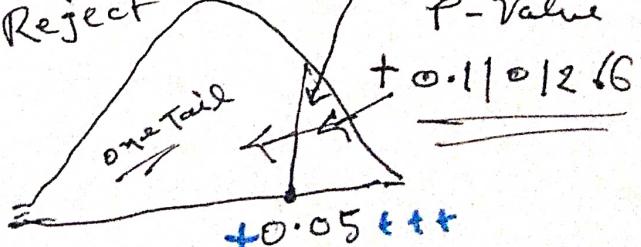
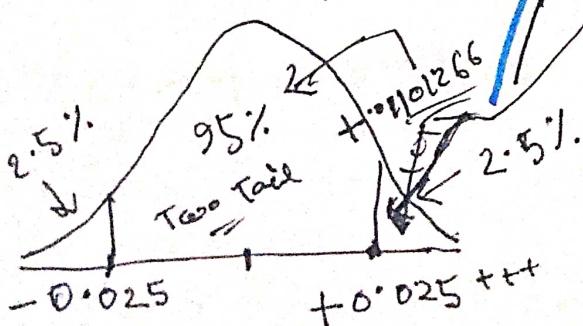
=> (1.5976240527147705, 0.1101266701438426)
 Z-test value , P-value

Assume

Significant value, $\alpha = 0.05$

① $0.11 > 0.05$
 ↓ (not less than 0.05)
 Accept Reject the Null Hypothesis

② $0.002 \leq 0.05$ Tail Reason
 ↓
 Accept the Null hypothesis
 Reject



Fail to Reject (or)
 Accept H_0

If P-Value \leq Significance Value, $P\text{-Value} \leq 0.05 = \alpha$
 Reject the Null Hypothesis (OR) { Accept the Null Hypothesis

Influential Statistics

Covariance :

X weight	Y Height	X↑ Y↑	X↓ Y↓
50	160		
60	170		
70	180		
75	181		

Increasing Both
Decreasing Both

(2) // No. of hours
study

2	6
3	4
4	3

play

X↑ Y↓
X↓ Y↑

one Increasing (ob) Decreasing
then another doing
opposite of each other

Quantity Relationship between \bar{X} & \bar{Y} $\frac{\text{mean of } X}{\text{mean of } Y}$

Covariance :

$$\text{cov}(X, Y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

(ob) $n-1$ If working with sample

{ After calculation }
3 Types of Results

$$= +ve \text{ (ob)} -ve \text{ (ob)}$$

OR number

① **+ve** Indicates

X↑ Y↑
X↓ Y↓

+ve Covariance

② **-ve** Indicates

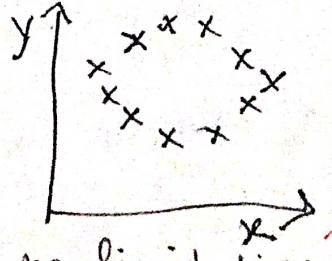
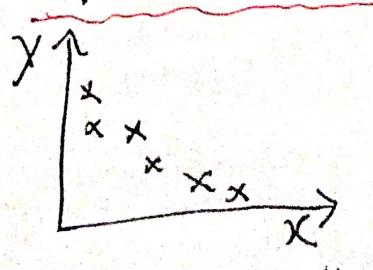
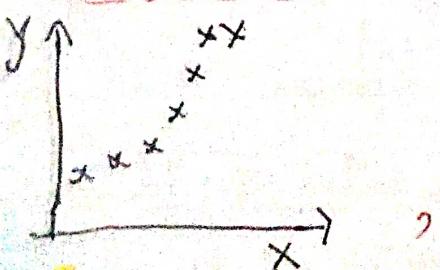
X↓ Y↑
X↑ Y↓

-ve Covariance

③ **0** Indicates, there is no relation between

X & Y
↓

0 = Covariance



x^{10^0}, x^{10^0}
 x^{20^0}, x^{20^0}

No fixed value in Co-Variance

Abt.

// Disadvantage of Covariance is there is no limitation / boundary

Influential Statistics

To solve the co-variance boundary problem we use —

② Pearson Co-relation Co-efficient :

(-1 to +1) The more towards (+1) (~~(abs)~~-1), it more positively correlated.

And The more towards (-1), it more negatively correlated.

$$f(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} \quad \{-1 \text{ to } +1\}$$

③ Spearman's ^{Rank} Co-relation coefficient :

$$\text{Spear}(x, y) = \frac{\text{cov}(R(x), R(y))}{R_{fx} \cdot R_{fy}}$$

x	y	R(x)	R(y)
Height	weight		
170	75	2	2
160	X	3	3
150	ignored	4	4
145	55	5	5
180	85	1	1

we use because,
↓
It captures the
Non Linear
Properties

Coding

Influential Statistics

Jupyter Notebook

t-test :

ages = [10, 20, 35, 50, 28, 40, 55, 18, 16, 55, 30, 25, 43, 18, 30, 28, 14, 24, 16, 17, 32, 35, 26, 27, 65, 18, 43, 23, 21, 20, 19, 70]

import numpy as np

ages_mean = np.mean(ages)

ages_mean =

⇒ 30.34375

sample_size = 10

age_sample = np.random.choice(ages, sample_size)

age_sample =

⇒ array([19, 35, 30, 17, 25, 25, 65, 50, 28])

from scipy.stats import ttest_1samp

ttest_1samp(age_sample, 30) =

⇒ Ttest_1sampResult(statistic=0.31410216574189925, pvalue=0.2606021861911046)

np.mean(age_sample) =

⇒ 31.5

ttest_1samp(age_sample, 31) =

⇒ Ttest_1sampResult(statistic=0.10470072191396641, pvalue=0.918909520333335)

ttest_1samp(age_sample, 28) =

⇒ Ttest_1sampResult(statistic=0.7329050533977649, pvalue=0.2791164328097811)

ttest_1samp(age_sample, 26) =

⇒ Ttest_1sampResult(statistic=1.1517079910536306, pvalue=0.2791164328097811)

Cooling

Influential Statistics

Jupyter Notebook

t-test

```
ages = [10, 20, 35, 50, 28, 40, 55, 18, 16, 55, 30, 25, 43, 18, 30, 28, 14, 24, 16,
        17, 32, 35, 26, 27, 65, 19, 18, 23, 21, 20, 19, 70]
```

```
import numpy as np
```

```
ages_mean = np.mean(ages)
```

```
ages_mean
```

```
⇒ 30.34375
```

```
sample_size = 30
```

```
age_sample = np.random.choice(ages, sample_size)
```

```
age_sample
```

```
⇒ array([19, 35, 30, 17, 25, 25, 65, 50, 28])
```

```
from scipy.stats import ttest_1samp
```

```
ttest_1samp(age_sample, 30)
```

```
⇒ Ttest_1sampResult(statistic=0.31910216574189925,  
                      pvalue=0.2606081761911046)
```

```
⇒ np.mean(age_sample)
```

```
⇒ 31.5
```

```
ttest_1samp(age_sample, 31)
```

```
⇒ Ttest_1sampResult(statistic=0.10970072191396691,  
                      pvalue=0.9189095203333335)
```

```
ttest_1samp(age_sample, 28)
```

```
⇒ Ttest_1sampResult(statistic=0.7329050533977649,  
                      pvalue=0.2791164328097816)  
                      pvalue=0.4822676675935259C)
```

```
ttest_1samp(age_sample, 26)
```

```
⇒ Ttest_1sampResult(statistic=1.1517079910536366,  
                      pvalue=0.2791164328097811)
```

Coding

e.g.: T-Test

Influential Statistics

Jupyter Notebook

consider another Example

ages of the college students (population)

1 class student mean of all the ages.

```
import numpy as np
```

```
import pandas as pd
```

```
import scipy.stats as stats
```

```
import math
```

```
np.random.seed(6)
```

```
school_ages = stats.poisson.rvs(loc=18, mu=35, size=1500)
```

```
classA_ages = stats.poisson.rvs(loc=18, mu=10, size=60)
```

```
school_ages
```

```
=> array([62, 59, 44, ..., 45, 52, 50])
```

```
classA_ages
```

```
=> array([52, 46, 40, 40, 47, 50, 51, 45, 44, 52, 46, 53, 43, 44, 51, 50, 54, 42, 54, 45, 61, 53, 49, 46, 47, 41, 45, 51, 43, 45, 48, 50, 40, 52, 44, 55, 54, 40, 45, 46, 54, 42, 46, 35, 51, 51, 46, 48, 47, 35, 52, 52, 39, 44, 48, 40, 42, 46, 47, 45])
```

```
classA_ages.mean()
```

```
=> 46.9
```

```
ttest_1sample(classA_ages, popmean = school_ages.mean())
```

```
=> Ttest_1sampResult(statistic = -9.604796510704091, pvalue = 1.139027071016194e-13)
```

```
school_ages.mean()
```

```
=> 53.30333333333335
```

```
if p-value < 0.05:
```

```
    print("Reject H0")
```

```
else:
```

```
    print("Accept H0)")
```

```
=> Reject H0
```

```
import seaborn as sns
```

```
df = sns.load_dataset('iris')
```

```
=> df.head()
```

Table

```
df.corr()
```

Table

```
sns.pairplot(df)
```

```
=> <seaborn.axisgrid.PairGrid  
at 0x16fd93c14c0>
```

different graphs

Visualization of corr

Influential Statistics

% P-Value & % Significance Value

P-Value is A: Area
(or) % of tested points - e.g. Z-test, t-test etc.

↳ Derive the P-Value % from Test

Q: - The average weight of all residents in Bangalore city is 168 pound with a standard deviation 3.9 we take a sample of 36 individuals and the mean is 169.5 pounds. CI = 95%.

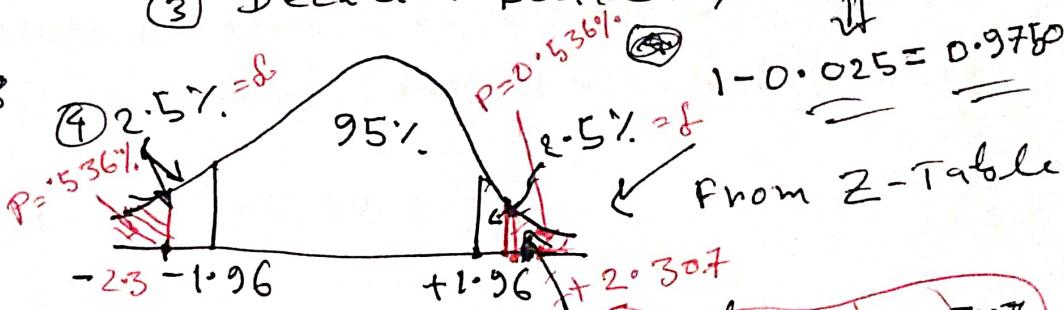
⇒ Z-Test: $M = 168$, $\sigma = 3.9$, $\bar{X} = 169.5$, $n = 36$, $\alpha = 0.05$

$$\textcircled{1} \quad H_0: M = 168$$

$$H_1: M \neq 168$$

$$\textcircled{2} \quad \alpha = 0.05$$

③ Decision Boundary



④ Z-Test

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{169.5 - 168}{3.9 / \sqrt{36}} \\ &= \frac{1.5}{3.9} \times 6 \\ &= 2.307 \end{aligned}$$

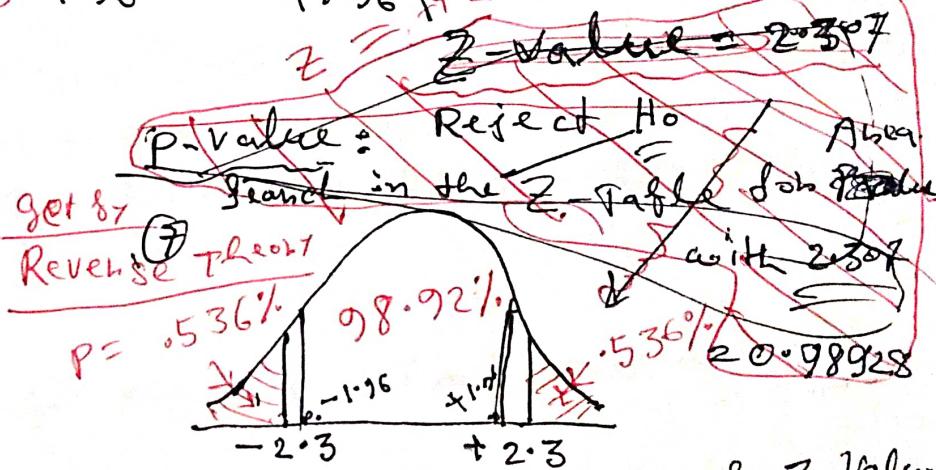
$$\textcircled{5} \quad Z = 2.307 > 1.96$$

↓
Reject the Null Hypothesis

(H_0) \Leftrightarrow

∴ P-value = 0.01072 < Significance value = 0.05

∴ Reject the Null Hypothesis for P-value also. (H_0)



Search the Area of Z-value

2.307 is Z-Table again

for p-value $2.3 = 0.98928$

$$P = 1 - 0.98928$$

$$= 0.01072 \Rightarrow 1.072\%$$

$$= 0.01072 / 2$$

$$= 0.00536 \Rightarrow 0.536\%$$

Reverse Theory \leftarrow One Tail

We get p-value area from Z-Test point (or) $\frac{1 - 0.98928}{2}$

Coding

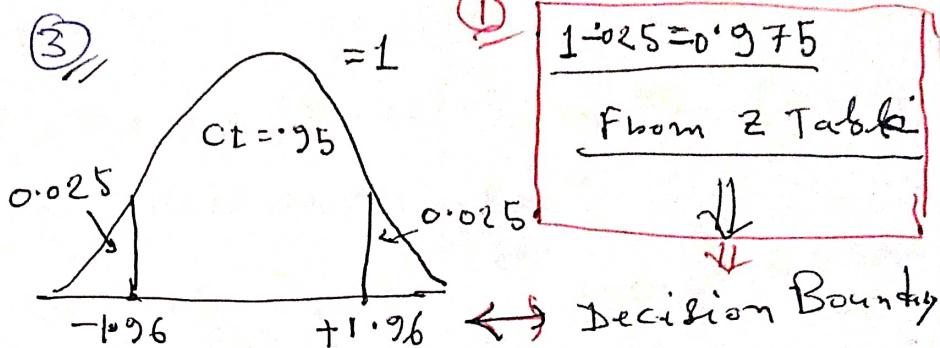
Inferential Statistics

Q Average age of a college is 24 years with a standard deviation 1.5. Sample of 36 student. Student's mean is 25 years. with $\alpha = 0.05$ CI = 95%, do the age vary?

$$\Rightarrow H_0: \mu = 24, \sigma = 1.5, n = 36, \bar{x} = 25, \alpha = 0.05$$

$$H_1: \mu \neq 24$$

$$\textcircled{2} \quad \alpha = 0.05 / 2 = 0.025$$

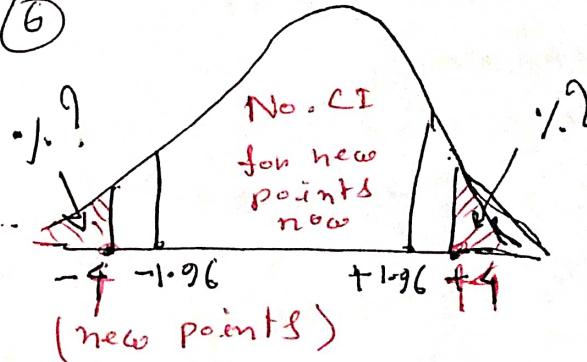


$$\textcircled{4} \quad Z\text{-Score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{25 - 24}{1.5} \times 6 \quad \textcircled{6}$$

$$= \frac{1 \times 6}{1.5}$$

$$= 4$$



$$\textcircled{5} \quad 4 > 1.96$$

Reject H_0

$$\textcircled{7} \quad \text{From Z Table} \quad \textcircled{2}$$

$$\text{again } q = 0.99997$$

Reverse $1 - q = ?$

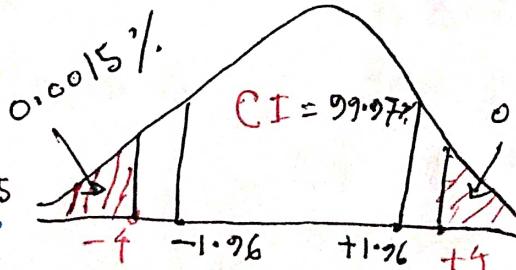
$$1 - 0.99997 = 0.00003$$

$$P = \frac{0.00003}{2}$$

$$\begin{aligned} \text{Two Tail} &= 0.000015 \\ &= 0.0015\% \end{aligned}$$

$$\textcircled{8} \quad \therefore P = 0.00003 < \alpha = 0.05$$

(P-value is less than significant value)



$$CI = 1 - 0.00003$$

$$= 0.99997$$

$$= 99.97\%$$

Z-Table

Search on Internet

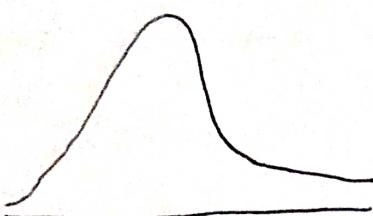
ztable.net/api-content/uploads/2018/11/positivetable.png

∴ Reject the Null Hypothesis too
(Reject H_0)

Advance Statistics (Descriptive statistics)

(3)

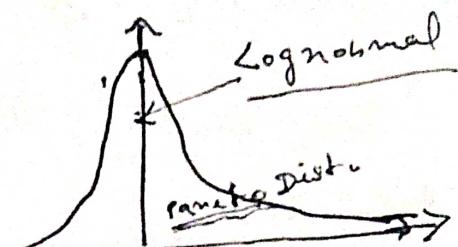
Log Normal Distribution :



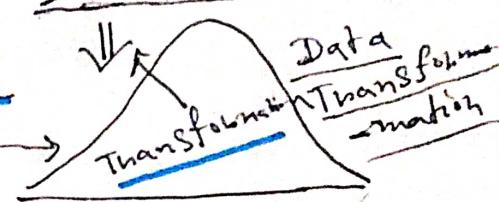
Eg: ① wealth distribution

② people writing big comments

$\{ Y \xrightarrow{\text{(random)}} \text{Log Normal} \}$ \leftarrow $\text{Normal distribution}$



$\log(Y) \rightarrow \text{Normal Distribution}$



Bernoulli Distribution :

in this only 2 outcomes

0 or 1

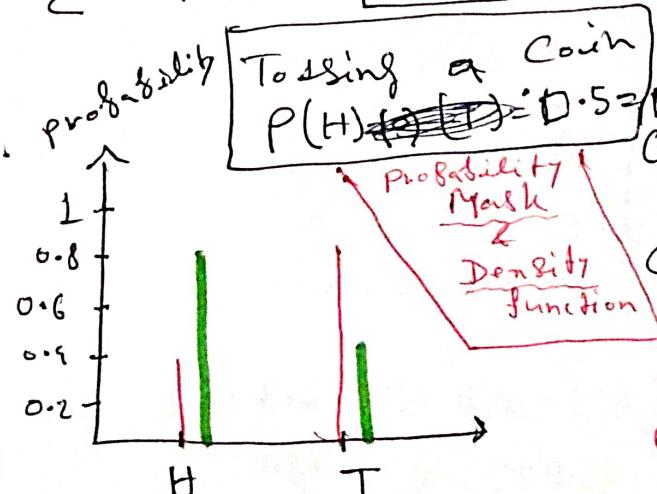
Single Trial
Distribution

$$\textcircled{1} \quad p = 0.5$$

$$q = 1 - p = 0.5$$

$$\textcircled{2} \quad p = 0.3$$

$$q = 1 - p = 0.7$$



$$\textcircled{1} \quad p(n=0) = 0.52 \quad \text{and} \quad p(n=1) = 0.48$$

$$\textcircled{1} \quad p(n=0) = 0.8 \quad \text{and} \quad p(n=1) = 0.2$$

$$\textcircled{3} \quad p(H) = 0.3$$

$$p(T) = 0.7$$

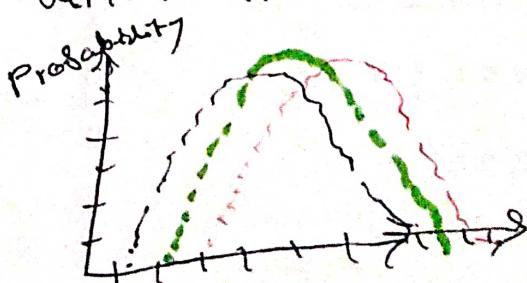
Binomial Distribution :

Multiple Trial

Every time \rightarrow Bernoulli distribution with Multiple Trial

$$\downarrow \quad p(H) = 0.5, p(H) = 0.6, \dots$$

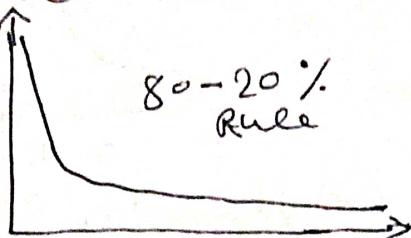
$$p(T) = 0.5, p(T) = 0.4, \dots$$



Advance Statistics (Descriptive Statistics)

⑥ Pareto Distribution (80% & 20%)

This is a non Gaussian/normal Distribution. This is a Power Law Distribution. Here 80-20 Rule exist.



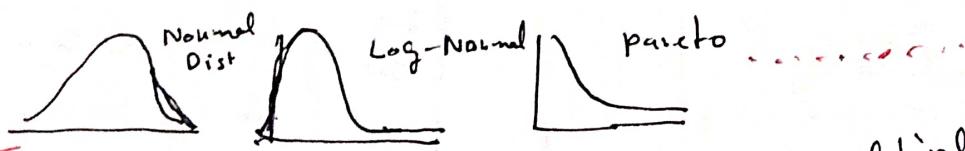
e.g.: 80% of the wealth is distributed 20% of the people.

- ③ 80% of sales is done by the 20% of famous product.
- ④ 80% of the match is won by 20% of the team.

- ② 80% of the company project is done by the 20% of the people in a team.

⑦ Central Limit Theorem

When we have any kind of distribution



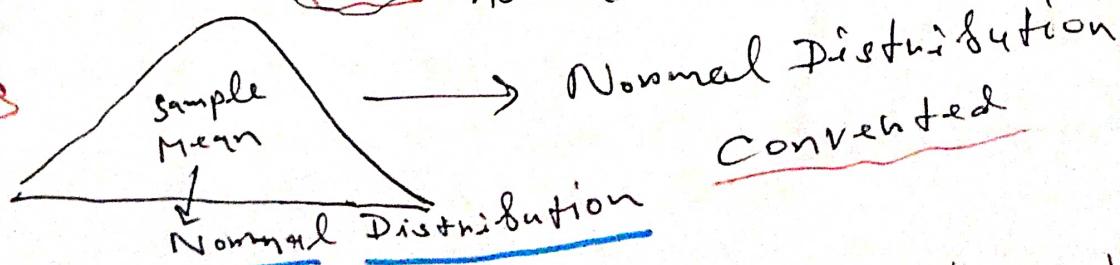
$$\begin{aligned} S_1 &\rightarrow \bar{x}_1 \\ S_2 &\rightarrow \bar{x}_2 \\ S_3 &\rightarrow \bar{x}_3 \\ S_4 &\rightarrow \bar{x}_4 \\ \vdots & \\ S_m &\rightarrow \bar{x}_m \end{aligned}$$

Start finding mean \downarrow Start taking Multiple Samples

$n \geq 30$
Sample Size

more Central Limit
Theorem Holds

Sample Mean and Population in the form of PDF
can be Anything
then, it will converted into a normal Distribution



- ⑧ Poisson distribution follows the Pareto Distribution