Statistics is the science of collecting, organizing and analyzing data → {facts or pieces of information that can be measured}

① Descriptive Statistics

↳ It consist of organizing and summerizing Data.

Ex: What is the Average Marks of the students in the class ? {85, 68, 70, 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} (80% / 20%)
6. Standard Normal Dist
7. Transformation and standardization (Python)
8. Q-Q plot
9. Poison Distribution {follows Pareto Dist}

② Inferential Stats

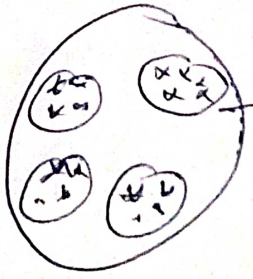
(Z test, t test, ANOVA, CHISquare, F Test, Hypothesis Testing (P value), test confidence Intervals)

→ Technique where in we used the Data that we have mapped to form Conclusions.

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

Population (N) & Sample (n)

Elections → Goa, UP, Assam



Sample asked by the reporters the voters to collect sample data

Different sample Techniques

1. Simple Random sampling
   ↳ Every member of the population (N) has an chance of being selected for sample (n)

Different Sampling(n) 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 (Strats)

eg: Gender → Male, female survey

Age group

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

(3) Systematic Sampling:

(N) → nth 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.

eg : height = { 182, 178, 168, 160, }

weight = { 78, 99, 100, 60, 50 }

Two kind of variables :

① Quantitative Variable → Measured Numerically { Add, sub, Mul, Divide } eg: Age, weight, height

② Qualitative / Categorical Variables.

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

eg: IQ

0 – 10 ⇓ Less IQ

10 – 50 ⇓ Medium IQ

50 – 100 ⇓ Good IQ

eg: Marital Status

eg: Blood Group.

Quantitative

① Discrete Variable

Eg: Whole Number

- No. of Bank a/c s
- ① Total. of children in a family, eg: 1, 2, 3, 4, 5
- ② Population of state

② Continuous Variable

Eg: Height = 172.5, 162 cm

weight = 100kg, 59kg

① Length of a river

② Amount of rain fall

③ Song length.

④ Blood Pressure

Statistics

613

Variable Measurement Scales :

4 types of Measurment Variable

1. Nominal data - { Categorical data } → colors, Gender classes.
2. Odinal data — { Order of the data matters, but value does not } eg. Rank - Marks

Rank	Marks
1	100
4	64
3	78
2	95

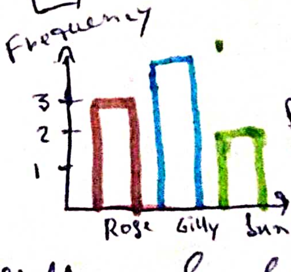
3. Interval
4. Ratio.

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

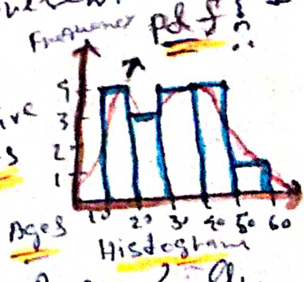
ex : Temprature
Fahrenhits

70-80   80-90   90-100   [0] Fahrenhit not matter present

Frequency Distribution :



Bargraph
Descriptive stats



smoothening of Histogram

Sample dataset : { Rose, lilly, Sunflower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly }

(1)

Based on flower type	Frequency
Rose	3
Lilly	4
Sunflower	2

(2) Cummelative Frequency

3) added
7) added
9

PDF = (Probability density Function)

(1) Bar Graph [: Discrete]
(2) Histograms : - (Continuous) eg: ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51 }

① Arithmetic Mean for population & Sample

① Mean (Average):

In some range of data
e.g. Student Ages

Population (N)

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

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

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{38}{10} = 3.8$$

Sample (n)

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

$$= 3.2$$

Used value here (100)

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

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

$= \frac{132}{11}$

$= 12$

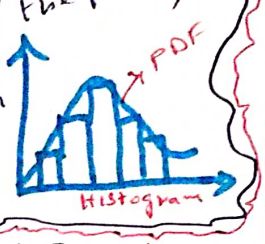
$M = 3.2$

$+100$ ↕ huge difference

$\bar{x} = 12$

X outlier

Use → To remove outliers in the Distributions



① Central Tendency:

① Mean ② Median ③ Mode

Refers to the measure used to determine the Centre of the distribution of data.

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

② Median:

$\{1, 1, 2, 2, 3, \underset{\text{Central element}}{\underset{\uparrow C}{3}}, 4, 5, 5, 6, 100, 112\}$ ← Too much outliers

Most cases to solve

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

→ if odd number = 11

Median = 3 ✓

→ if Even number = 12

Median = 3.5

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

③ Mode: It takes the most frequent element

Categorical & Missing data

$\{1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100\}$

2 3 1

Mode = 6

if Mode = 100 X Problem

2) Measure of Dispersion : => { Spread }

① Variance

② Standard Deviation

Q:-

$\mu \rightarrow \{1,1,2,3,4\}$ $\frac{10}{5} = 2$

Same Average

$\rightarrow \{2,2,2,2,2\}$ $\frac{10}{5} = 2$

How to Identify these two distribution are different ?

=> At that time we will use variance & Standard Deviation

Variance : { spread }

① Population (N) Variance

1st step

Sigma sq.

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

$= 1.81$

x	μ Mean	$x-\mu$	$(x-\mu)^2$
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	1.17	1.37
5	2.83	2.17	4.71
			10.84

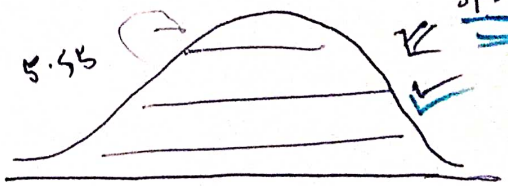
$\mu = 2.83$

② Sample (n) Variance

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

Q: where Variance is more ?

=> more spread

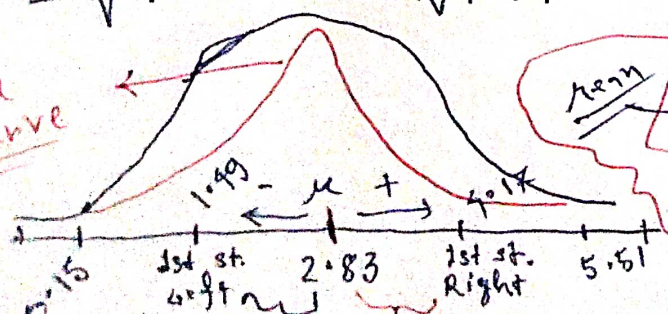


Standard Deviation : { root of Variance }

2nd step

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

Bell curve



Mean

1st → 2.83 − 1.34 = 1.49

2nd → 1.49 = 1.34 = .15

Left Standard Deviation

Right Standard Deviation

1st 2.83 + 1.34 = 4.17

2nd 4.17 + 1.34 = 5.51

Percentiles And Quartiles {Find outliers}

eg:-

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

Q:- % of the numbers that are Odd ?

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

$$= 3/5 = 0.6 = 60\%$$

Percentiles : { A percentile is a value below which a certain percentage of observation lie. }

eg:

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

(positions 1 ... 16 17 ... 20; 25% and 75% marked)

Q:- what is the percentile ranking of 10 ?

$x = 10$, $n = 20$

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

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

(16 ← position)

$= 80\%$

Q:- 11 ?

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

Q:-

(2) what value exists at percentile ranking of 25% ?

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

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

$= 5.25 \longrightarrow$ Index Position

Value 5 → 25%

Q:- 75% ?

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

$= 15.75$ (Index)

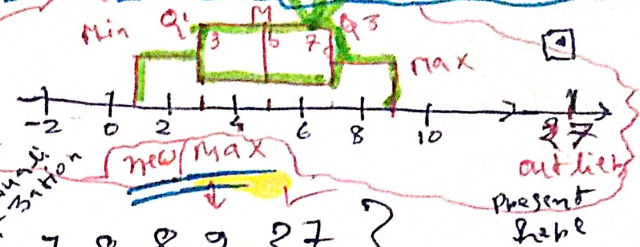
Value 9

Five Number Summery :

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 Removing the outliers



eg: currently Min ✓ -50 Median = 5

{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 } ✗

[Lower fence (-3) ⟷ Higher fence (13)]

Lower fence = $Q_1 - 1.5\,(IQR)$ Fixed

Upper fence = $Q_3 + 1.5\,(IQR)$

Inter Quantile Range (IQR) = $Q_3 - Q_1$

$Q_1 = (25\%)$

$Q_3 = (75\%)$

Q:- 25% ? (Percentile)

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

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

$= 5 \rightarrow$ Index

$Q_1 = \boxed{3} \rightarrow$ value

Q:- 75% ? (Percentile)

$Q_3 \Rightarrow \boxed{7} \rightarrow$ value

(3) Higher fence = $Q_3 + 1.5\,(IQR)$
$= 7 + 1.5 \times 4$
$= 7 + 6$
$= \boxed{13}$

(1) ∴ Inter Quantile Range is (IQR) = $Q_3 - Q_1$
$= 7 - 3$
∴ IQR = 4

(2) Now Lower fence = $Q_1 - 1.5\,(IQR)$
$= 3 - 1.5 \times 4$
$= 3 - 6$
$= \boxed{-3}$

Advance Statistics

Distributions :

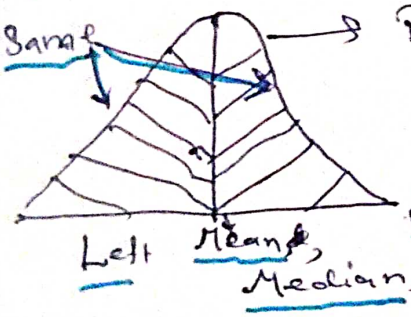
- Normal / Gaussian Distribution
- Standard Normal Distribution
- Z - Scores (/ Z-Table) ⇒
- Log Normal Distribution
- Bernullis Distribution
- Binomial Distribution

Practical

1. Mean, Median, Mode
2. Variance, Standard D.
3. Histogram, PDF, Bar plot, violen plot
4. IQR
5. Long Normal Distribution

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

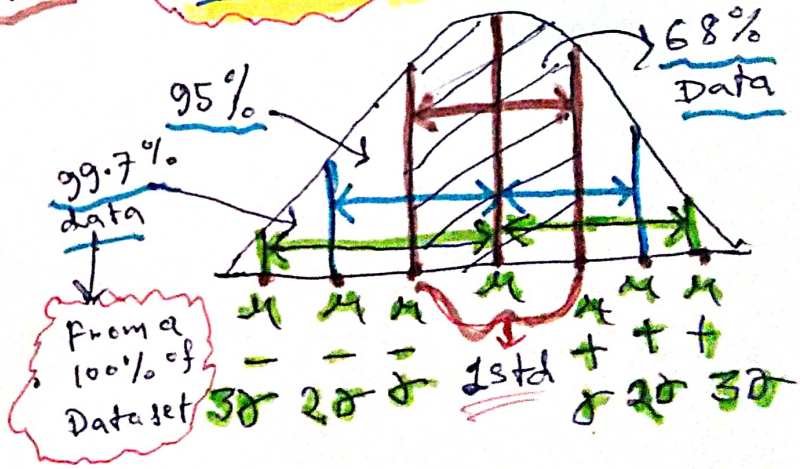
① Gaussian / Normal Distribution :



Right part to the Left part of the distribution exactly Similar / Similtical

Empenical formula :

68 → 95 → 99.7 % Rule



eg : Height (Dataset) → Normally Distributed

↳ According to Domain Expert .

② Weight , ③ IRIS DATASET .

⇒ Z-Score

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

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

$$= \frac{4.75 - 4}{1}$$

$$= .75 \text{ Std.}$$

① Z score $= \frac{3.75 - 4}{}$

$= -0.25$ Std.

Advance Statistics

② Standard Normal Distribution :

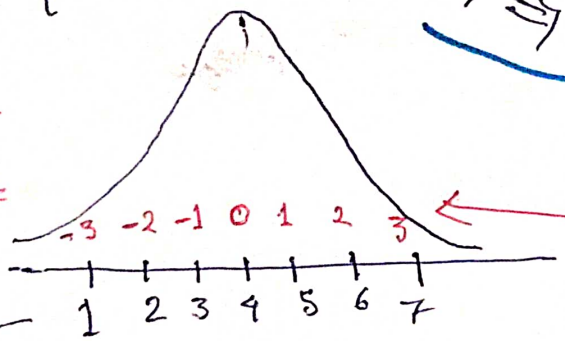
It is a Distribution that a set of normally Distributed Data converts by the Z-Score to a Standard Normal Distribution set of Data.

Normal / Gaussian

eg: {1, 2, 3, 4, 5, 6, 7} → Distribution

$\mu = 4$

$\sigma = 1$



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

{−3, −2, −1, 0, 1, 2, 3} → Standard Normal Distribution

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

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

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

Practical Application :

Standardization → converted to $\mu = 0, \sigma = 1$

(Value − mean) / Std. ← Z score works internally

DATASET :

Age (year)	Salary (RS)	Weight (kg)
24	40 kg	70
25	80 k	80
26	60 k	55
27	70 k	45

Z score

3 Types of Z-scores we get.

⇓

Standard Normal Distribution

Normalization

$$\frac{(Value - min)}{(max - min)}$$

upper bound, lower bound

Minmax scaler

(value 0 to 1)

eg: what percentage of scores fall above 4.25?

⇒ μ-Score

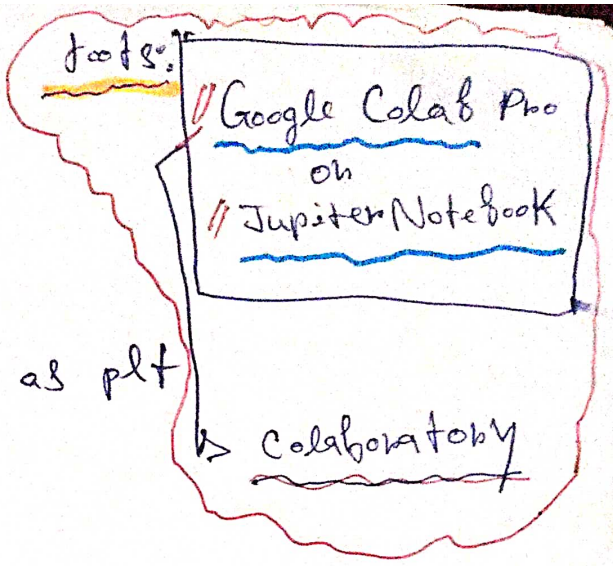
$$Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

1 − 0.5987 (from some Z table)

= 0.4013

= 40%

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.media(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 / barchart
=
np.percentile(df1['sepal_length'], [25, 75])
=
df2 = pd.read_csv("dat.csv")
df2.head()
=
```

Code

IQR / outliers / Z score / lower fence / upper fence

google colab or Jupiter Note book

```
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, 10, 107, 10, 13, 12, 14,
           12, 108, 12, 11, 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]

IQR
1. Sort the Data
2. calculate Q1 and Q3
3. IQR (Q3 - Q1)
4. Find the Lower fence (Q1 - 1.5(IQR))
5. Find the upper fence (Q3 + 1.5(IQR))

```
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)
```

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?

$$Pr(1 \text{ or } 3 \text{ or } 6) = Pr(1) + Pr(3) + Pr(6)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$
$$= \frac{3}{6_2} = \frac{1}{2} = 0.5$$

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

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

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

⇒ Mutual Exclusive (Additional Rule)

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

$$Pr(A \text{ or } B) = 1$$

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 ?

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

Addition Rule for non mutual Exclusive events

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

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

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

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

$$\approx 0.30769$$

Something like this

(2) Multiplication Rule :

- Independent Event
- Dependent Event

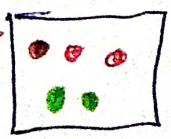
Eg: Rolling a dice {1,2,3,4,5,6}
1,1,2,3,5,5,6,4 Each and every are independent

Eg:

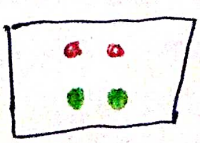
$$Pr(\text{Red}) = \frac{3}{5} \xrightarrow{\text{Impacted after removing}} Pr(\text{Green}) = \frac{2}{4}$$

after

From A Box has Total 5 Marbells 3 Red 2 Green



removed 1 Red then



⇒ Naive Bays (conditional Probability

Probability

Independent Event / Multiplication

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

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

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

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

Dependent Event

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

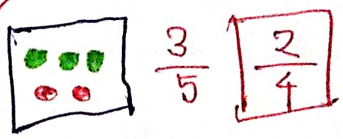
conditional probability

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

Bay's Theorem

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

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



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

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

Probability

Permutation & Combination

eg: School trip { chocolate factory } → Dairy, 5Star, Kedberry, Milky bar, Silk, Mix fruit.

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

⇒ { Dairy, Silk, Milky
Milky, Silk, Dairy } → can swap sample elements

Total chocolates = $n = 6$

$n = 6$

$r = 3$

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

$$ {}^{n}C_{r} = \frac{n!}{r!\,(n-r)!} = \frac{6!}{3!\,(6-3)!} $$

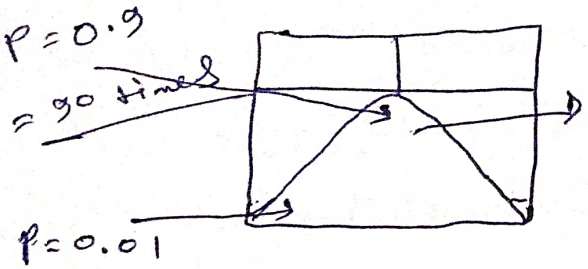
$$ = \frac{{}^{2}6 \times 5 \times {}^{1}4 \times 3!}{3 \times 2 \times 1 \times 3!} $$

$$ = 20 $$

Influential Statistics

① P Value :

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



P = Probability

CI → Confidence Interval

Alpha → α

Hypothesis testing, Confidence Interval, Significance Value,

Coin → Test whether this coin is a fair coin or not by performance.

In 100 times tosses

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

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

Sholey Sholey coin 50-50%

will decide whether fair or not

& It sets a value to define a value nearer to Mean → Significance Value

But,

Hypothesis Testing

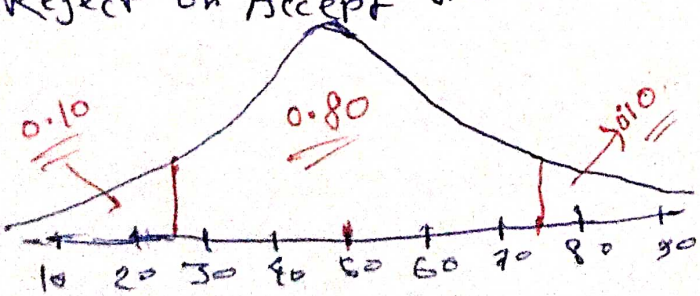
① Null Hypothesis : Coin fair

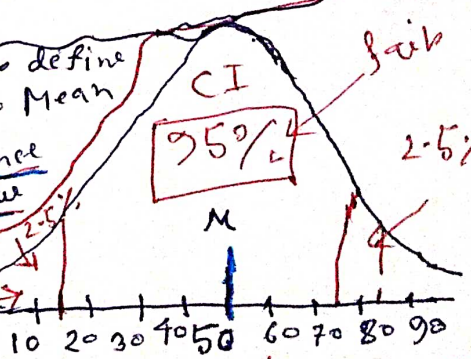
② Alternate Hypothesis : Coin unfair

③ Experiment (eg: t-test, z-test, chi test)

⇓ (based on it)

④ Reject or Accept the Null Hypothesis





eg: $\alpha = 0.05$ / 100-5% = 95% CI

Domain Expert → the significance value decide when the fair point will fall.

Alfa $\alpha = 0.20 = 20\%$

then CI = 80% = 0.80

Confidential Interval

Influential Statistics

① Type 1 and Type 2 Error :

Null Hypothesis (H_0) = Coin is fair

Alternative Hypothesis (H_1) = Coin is not fair

Reality check

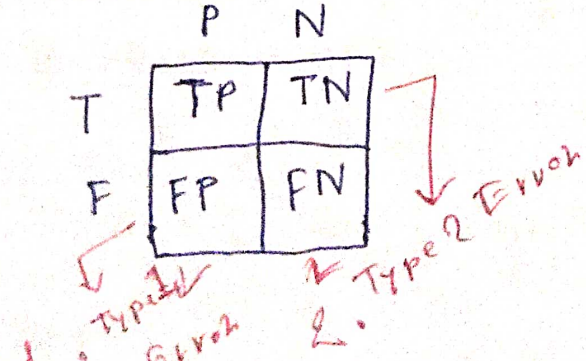
Null hypothesis is true or Null hypothesis is false.

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 1 Error
- **outcome 3 :** we Accept the Null hypothesis, when in reality it is false → Type 2 Error
- **outcome 4 :** we Accept the Null Hypothesis, when in reality it is true → Yes

eg : Confusion Matrics :

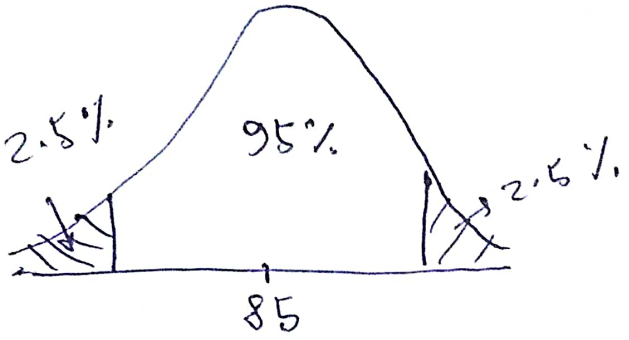


Inferential Statistics

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 ?

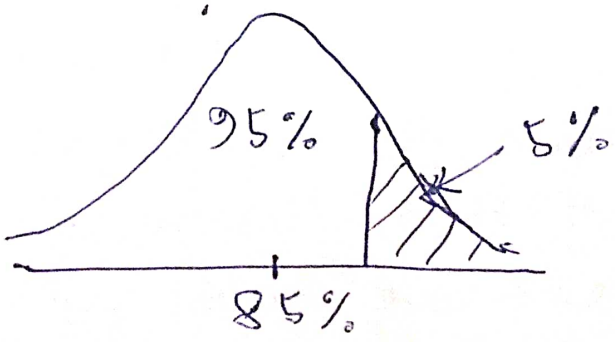
$\alpha = 0.05$



Two Tail Test

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

$\alpha = 0.05$

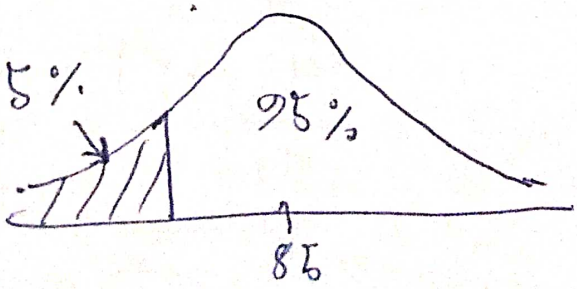


one Tail Test

(or)

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

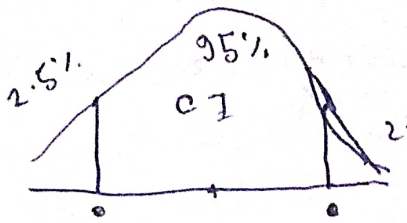
$\alpha = 0.05$



one Tail Test

Influential Statistics

(3) Confidence Interval (How to find) ?



$\alpha = 0.05$

Point Estimate

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

In Influential stats real data

Sample → population

Mean (n) Based on Sample — Mean (N)

$\bar{x}$ ⟶ M

$\bar{x} = 2.9$ Approximately nearest $M = 3$

we don't know exact Mean of population so these calculate

In Influential stats Based on Sample data we try to estimate the all population data.

eg: reporter collecting about the vote report from sample of public.

Confidence Interval

Point Estimate ± Margin of Error

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

Upper Bound $= \bar{x} + Z_{\frac{0.05}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$

Lower Bound $= \bar{x} - Z_{\frac{0.05}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$

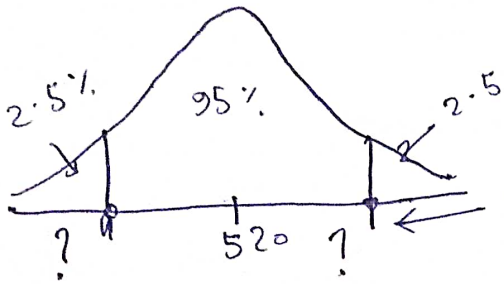
Inferential Statistics

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 (30) test takers has a mean of 520 score. Construct a 95% CI about the mean?

For simple calculation

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



What will be at those position?

2 of them

Often fulfill

usually (2) $n \geq 30$

① Whenever Population Std. is given, Then Z Test

Point Estimate ± Margin of Error

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

Upper Bound of CI $= \bar{x} + Z_{\frac{0.05}{2}} \cdot \left(\frac{100}{\sqrt{25}}\right)$

Lower Bound of CI $= \bar{x} - Z_{\frac{0.05}{2}} \cdot \left(\frac{100}{\sqrt{25}}\right)$

Upper $= 520 + 1.96\,(20) = 559.2$

Lower $= 520 - 1.96\,(20) = 480.8$

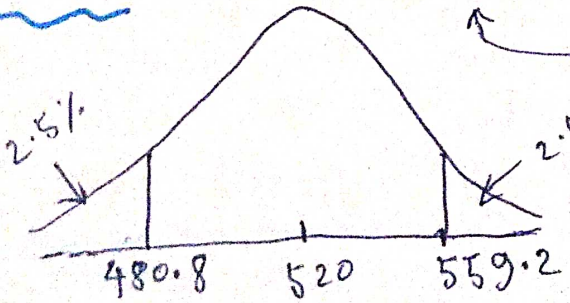
$Z_{0.05} = Z_{0.025}$

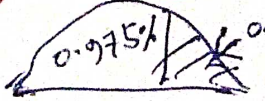
$1 - 0.025 = 0.975$

According to Z table Value is Position

$Z = 1.96\,(06)$

$\frac{1.96}{20} = \frac{5}{}$





↑ 1.96

One Tail

Influential Statistics

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

⇒ (2) 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

Here population std is not given ⟶ t-test

$n = 25$

$\bar{x} = 520$ $\quad s = 80$ ↳ Sample std

(Alpha) ↳ $\alpha = 0.05$ ↳ Significance value

Point Estimate ± Margin of Error

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

$$\text{Upper Bound} = \bar{x} + t_{\frac{0.05}{2}} \left(\frac{s}{\sqrt{n}}\right)$$

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

$$= 520 + 2.064 \cdot \left(\frac{80}{5}\right)^{16}$$

$$= 553.024$$

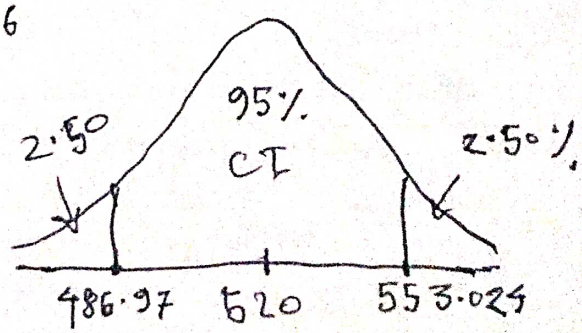
$$\text{Lower Bound} = \bar{x} - t_{\frac{0.05}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

$$= 520 - 2.064 \cdot \left(\frac{80}{5}\right)^{16}$$

$$= 486.97$$

[486.97 ⟷ 553.024]

$\frac{0.05}{2} \Rightarrow 0.025$	
→ t-Table	
	0.025 to 1 Tail
24 = T	0.05 2 Tail ↓
24 =	2.064



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%

$S = 140$

$\bar{x} = 140$

⇒ 1) Define Null Hypothesis is

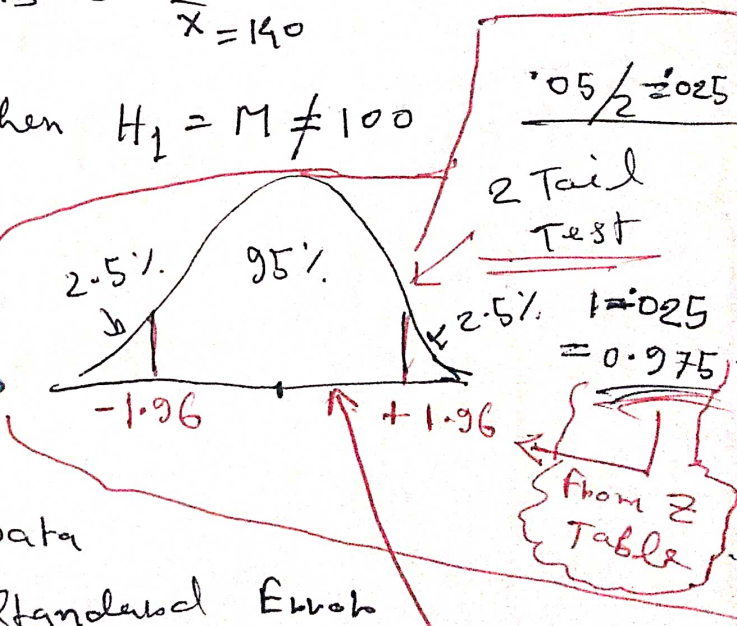
$H_0 = M = 100$

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

3) State Alpha

$\alpha = 0.05$

4) State Decision Rule



5) Calculate Z Test statistics

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

when huge Data $\frac{\sigma}{\sqrt{n}}$ → Standard Error

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

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

$$= 14.60$$

Ans State out Decision

$14.60 > 1.96$, $Z = 14.60$

① if Z is less than −1.96 or greater than 1.96, Reject the null hypothesis

② Medication improve the Intelligence or Decrease? (Improve Very Much)

Influential Statistics

(2) One Sample T-test (Hypothesis testing)

- Z-Test ⇒ Population Std.
- t-test ⇒ Unknown Population Std.

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

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

Did the Medication affect intelligence?

$\alpha = 0.05$

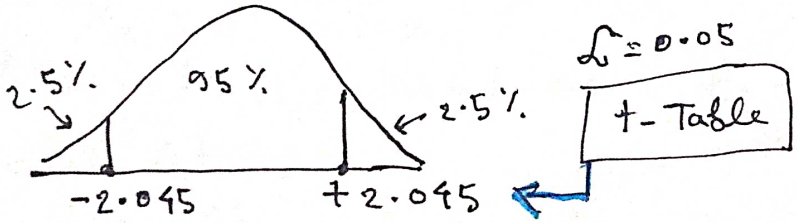
⇒ ① $H_0 = M = 100$

② $H_1 = M \neq 100$

③ Calculate the degree of freedom

$n - 1 = 30 - 1 = 29$

④ State Decision Rule



⑤ T-Test

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

$\bar{x} = 140$
$\mu = 100$
$S = 20$
$n = 30$

$= 10.96$

⇒ ⑥ State out decision

$T = 10.96 > 2.045$

So, ① Reject the Null Hypothesis → Increased the Intelligence.

② $P \leq$ Significance Value

③ Accept the Alternate Hypothesis.

Inferential Statistics

Chi Square Test :

Chi square Test claims about Population proportions

It is a non parametric test that is performed on catagorical 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 :

Less than 18	18 - 35	> 35
20 %	30 %	50 %

In 2020, age of $n = 500$ individuals were sampled. Below are the results

< 18	18 - 35	> 35
121	288	91

using $\alpha = 0.05$, would 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

$n = 500$ { Sample }

⇒ observed = f_o

500 X 0.2 = 100	500 X 0.3 = 150	500 X .5 = 250

⇒ Expected = f_e

(continue)

(continued)

Influential Statistics

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

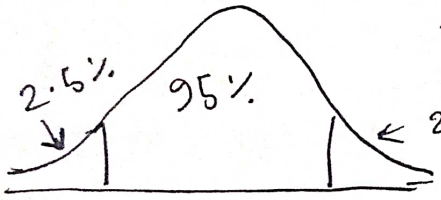
→ n=3 catagories
→ observed (f_o)
→ Expected (f_e)

① H_o = The data meets distribution 2000 census
H_1 = The data does not meet n 2000 census

② $\alpha = 0.05$ (95%=CI)

③ Degree of freedom = $n-1 = 3-1 = 2$ → df = 2, $\alpha = 0.05$

④ Decision Boundary { check in the Chi Square Table } → χ^2 ← denoted by



2 Tail Test

from χ^2 Table

If χ^2 is greater than 5.99 reject H_o

⑤ calculate Chi Square Test statistics.

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

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

$$= 232.494$$

$\therefore \chi^2 = 232.494 > 5.99$ { Reject the Null Hypothesis H_o

Coding Z-test — Influential Statistics

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, 94, 96, 97, 97, 99, 99, 105, 109, 109, 110, 112,
        112, 113, 114, 115]

ztest(data, value = 100)   # if 110   =.002 (Reject)
=> (1.5976240527147705, 0.11012667014384926)
```

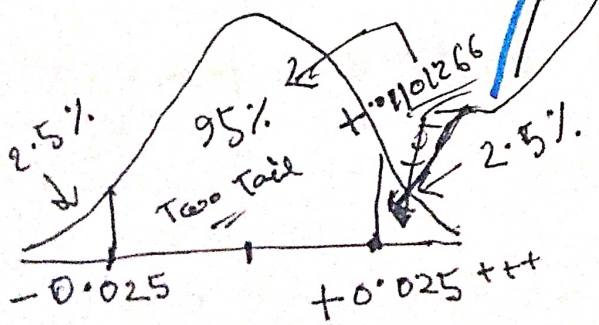
1.5976240527147705 ⇑ Z-test value, 0.11012667014384926 ⇑ P-value

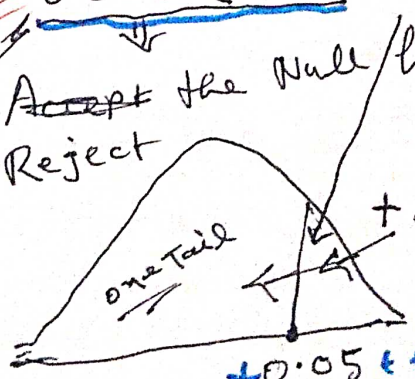
Assume Significant value, $\alpha = 0.05$

① $0.11 \geq 0.05$ (not less than 0.05) ⇓ Accept ~~Reject~~ the Null Hypothesis

② $0.002 \leq 0.05$ ⇓ ~~Accept~~ Reject the Null Hypothesis

Tail Reason





Fail to Reject (or) Accept Ho

Rule: If P-Value ≤ Significance Value, ⇓ Reject the Null Hypothesis (OR) { P-value ⇓ ≥ 0.05 = α, Accept the Null Hypothesis

Influential Statistics

Covariance :

① //

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

- X↑ Y↑ — Increasing Both
- X↓ Y↓ — Decreasing Both

② //

No. of hours study	play
2	6
3	4
4	3

- X↑ Y↓
- X↓ Y↑

one Increasing (or) Decreasing then another doing opposit of each other

Quantify Relationship between X & Y

Co-Variance :

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

($\bar{x}$ = mean of X, $\bar{y}$ = mean of Y)

→ If working with Sample

→ population

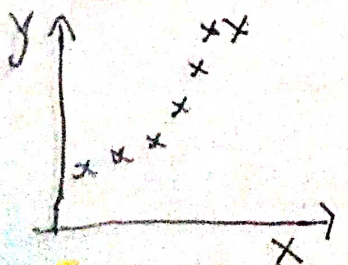
After calculation 3 Types of Results:

= +ve (or) −ve (or) 0 number

① +ve Indicates

X↑ Y↑
X↓ Y↓

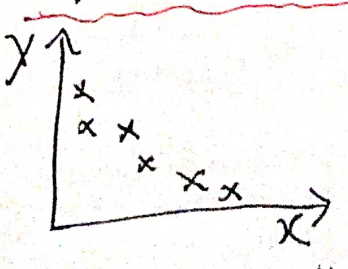
⇓ +ve covariance



② −ve Indicates

X↓ Y↑
X↑ Y↓

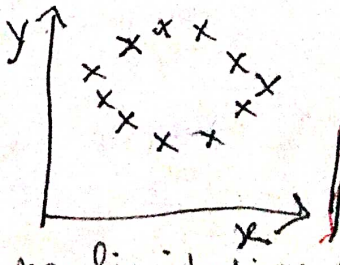
⇓ −ve covariance



③ 0 Indicates, there is no Relation between

X & Y

⇓ 0 = Covariance



X: 100, +100, −200, +200

No fixed value in Co-variance

// Disadvantage of Co-variance is there is no limitation / boundary

Influential Statistics

To Solve the co-variance Boundary problem we use —

② Pearson Co-relation Co-efficient :

$(-1 \text{ to } +1)$ The more towards $(+1)$, it more positively co-related.

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

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

③ Speabmanis Rank co-relation coefficent :

$$Spear(x,y) = \frac{Cov(R(x), R(y))}{R_{\sigma_x} \cdot R_{\sigma_y}}$$

we use because,

It captures the Non Linear Propertyes

x Height	y weight	R(x)	R(Y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

Rank

X ignored

Accept

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, 21, 25, 65, 50, 28])

```
from scipy.stats import ttest_1samp
ttest_1samp(age_sample, 30)
```

⇒ Ttest_1sampResult(statistic = 0.3141021657418992 5,
pvalue = 0.2606021961911046)

```
np.mean(age_sample)
```

⇒ 31.5

```
ttest_1sample(age_sample, 31)
```

⇒ Ttest_1sampResult(statistic = 0.10470072191396641,
pvalue = 0.9189095203333335)

Ttest_1sampResult

```
ttest_1sample(age_sample, 28)
```

⇒ Ttest_1sampResult(statistic = 0.7329050533977649,
~~pvalue = 0.27911643280978 11~~
pvalue = 0.4822676675435254)

```
ttest_1sample(ag_sample, 26)
```

⇒ Ttest_1sampleResult(statistic = 1.1517079410536306,
pvalue = 0.27911643280978 11)

Influential Statistics

Jupyter Notebook

Coding

eg: T-Test

```
#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_1samp(classA_ages, popmean = school_ages.mean())
=> Ttest_1sampResult(statistic=-9.604796510704091, pvalue=1.13902
                     707 1016194e-13)
school_ages.mean()
=> 53.303333333333335
if p_value < 0.05:
    print("Reject Ho")
else:
    print("Accept Ho")
=> Reject Ho
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
=> #Table
df.corr()
=> #Table
sns.pairplot(df)
=> <seaborn.axisgrid.PairGrid
   at 0x16fd53c14c0>
# different graphs
```

Visualization of corr

Influential Statistics

P-value is A. Area (or) % of tested points - eg. Z-test, T-test etc.

% P-Value & % Significance Value:

↳ 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%

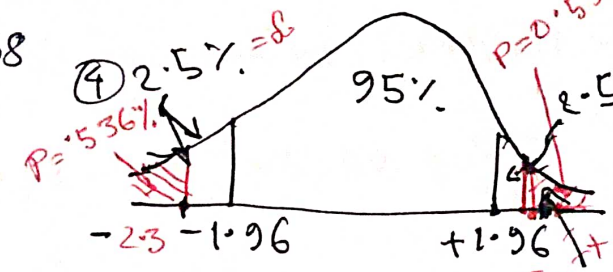
⇒ Z-Test: $\mu = 168$, $\sigma = 3.9$, $\bar{x} = 169.5$, $n = 36$, $\alpha = 0.05$

① $H_0 = \mu = 168$

$H_1 = \mu \neq 168$

② $\alpha = 0.05$

③ Decision Boundary



⑤ 1 − 0.025 = 0.9750

From Z-Table

⑧ Z-Test

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}}$$

$$= \frac{1.5}{3.9} \times 6$$

$$= 2.307$$

Z-value = 2.307

P-value: Reject Ho

Area found in the Z-Table for P-value with 2.307

get by Reverse Theory

P = .536%, 98.92%, .536%

-2.3, -1.96, +1.96, +2.3

0.98928

⑥ $Z = 2.307 > 1.96$

⇓

Reject the Null Hypothesis (Ho)

Search the Area of Z-value 2.307 is Z-Table again for P-value

2.3 = 0.98928 = 98.92%

⇓

P = 1 − 0.98928

= 0.01072 ⇒ 1.072%

= 0.01072/2

= 0.00536 ⇒ 0.536%

one Tail

Reverse Theory To Table we get P-value area from Z-Test point (or)

⑧

∴ P-value = 0.01072 < Significance value = 0.05

we Reject the Null Hypothesis for P-value also. (Ho)

Coding

Inferential Statistics

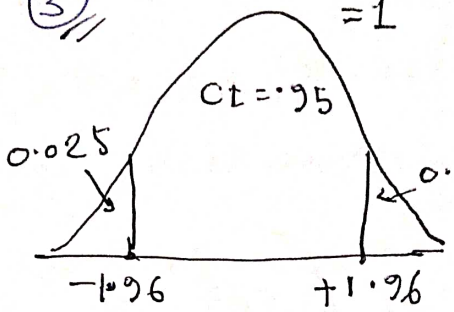
② Average age of a college is 24 years with a standard deviation 1.5. Sample of 36 student. Students mean is 25 years. with $\alpha = 0.05$ CI = 95%, do the age vary?

⇒ ① $H_0 = \mu = 24$, $\sigma = 1.5$, $n = 36$, $\bar{x} = 25$, $\alpha = 0.05$

$H_1 = \mu \neq 24$

② $\alpha = 0.05/2 = .025$

③



① $1 - .025 = 0.975$

From Z Table

↓

↔ Decision Boundry

④ $Z\text{-score} = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

$= \dfrac{25 - 24}{1.5} \times 6$

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

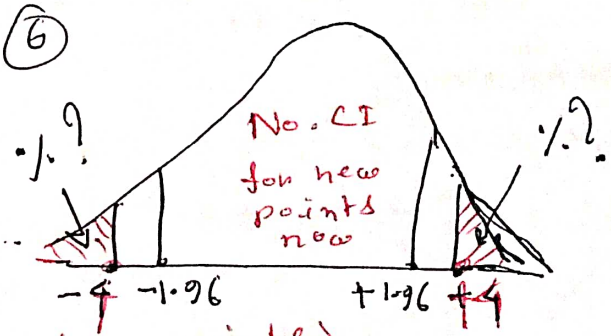
$= 4$

⑤ $4 > 1.96$

↓

Reject Ho

⑥



Reverse %?

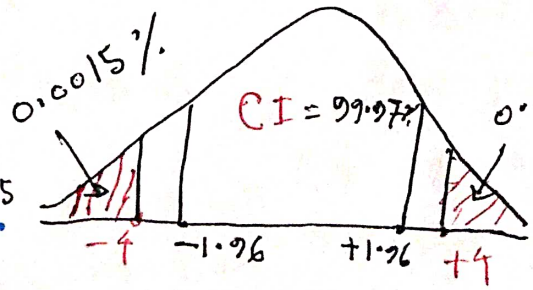
⑦ From Z Table ② again $4 = .99997$

$1 - .9997 = 0.00003$

$P = \dfrac{.00003}{2}$

Two Tail = 0.000015

= 0.0015%



⑧ $\therefore P = .00003 < \alpha = 0.05$

(P-value is Less than significant value)

↓

$\therefore$ Reject the Null Hypothesis too

(Reject Ho)

$CI = 1 - 0.00003$

$= .9997$

$= 99.97\%$

Z-Table

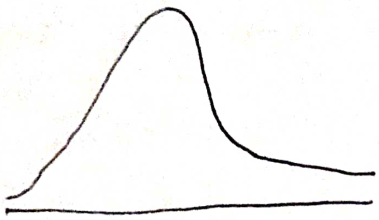
Search on Internet

↓

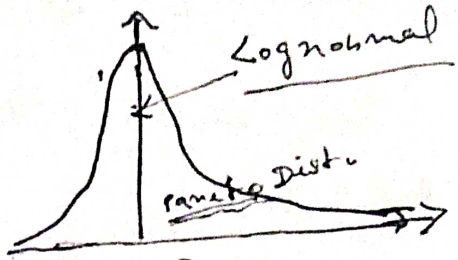
ztable.net/wp-content/uploads/2018/11/positiveztable.png

③ Advance Statistics (Discriptive statistics)

Log Normal Distribution :



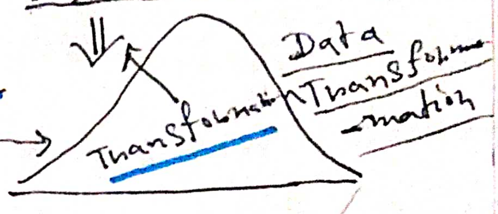
Eg: ① wealth Distribution
② people writing big comments



{ Y (random) ~ LogNormal Distribution }

⇓

log (Y) ⟶ Normal Distribution



④ Burnoullis Distribution :

Single Trial Distribution

in this only 2 outcomes [0 or 1]

① p = 0.5
q = 1 − p = 0.5

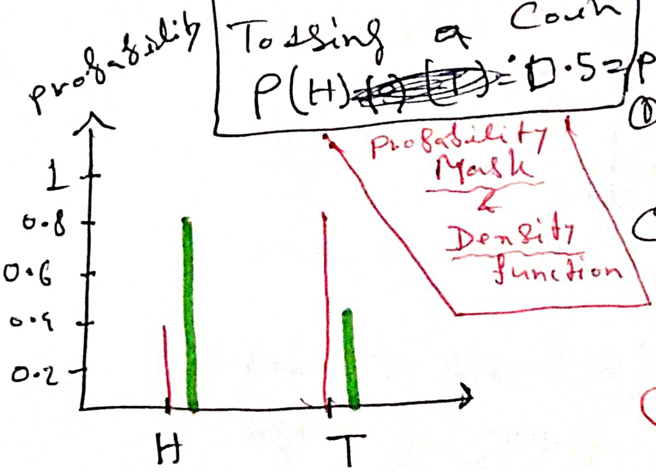
② p = 0.3
q = 1 − p = 0.7

Tossing a Coin
P(H) = 0.5 = P(T) = 0.5

① P(x=0) = 0.02 and P(x=1) = 0.8

① P(x=0) = 0.8 and P(x=1) = 0.2

③ P(H) = 0.3
P(T) = 0.7



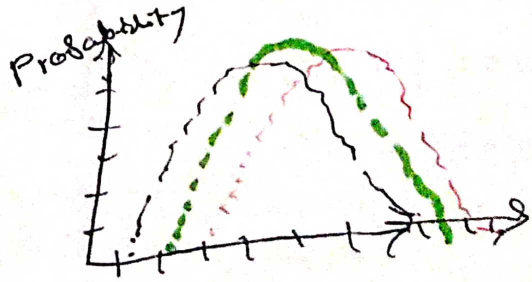
① catagonical (PMF)
② continuous val. (PDF)

⑤ Binomial Distribution : Multiple Trial

Every time ⟶ Burnoulli distribution with Multiple Trial

↓

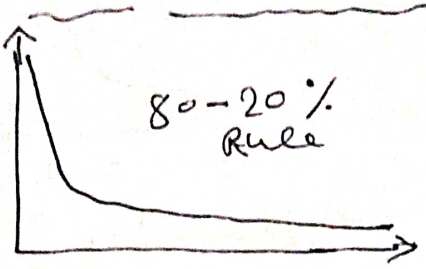
P(H) = 0.5 , P(H) = 0.6 , ...
P(T) = 0.5 , P(T) = 0.4, ...



Advance Statistics (Discriptive Statistics)

(6) Pareto Distribution (80% & 20%) :

This is a non Gausian/normal Distribution. This is a Power law Distribution. Here 80-20 Rule exist.



Eg : ① 80% of the wealth is Distributed 20% of the people.

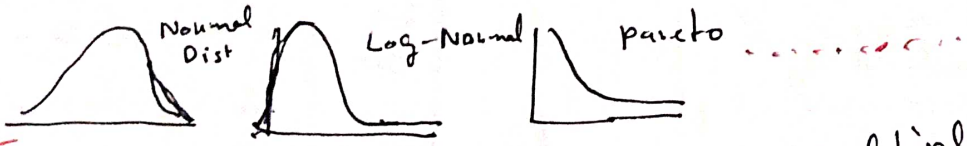
② 80% of the company project is done by the 20% of the people in a Team.

③ 80% of Sales is done by the 20% of famous product.

④ 80% of the match is won by 20% of the team.

(7) Central Limit Theorem :

When we have any kind of distribution



Start finding mean ⇐ start taking Multiple Samples

$S_1 \rightarrow \bar{x}_1$
$S_2 \rightarrow \bar{x}_2$
$S_3 \rightarrow \bar{x}_3$
$S_4 \rightarrow \bar{x}_4$
$S_m \rightarrow \bar{x}_m$

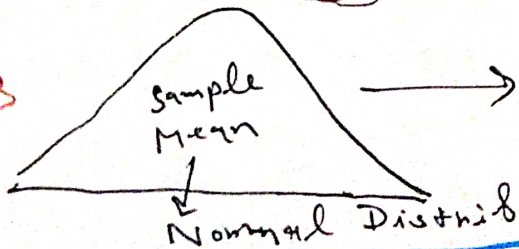
$n \geq 30$ Sample size

more Centre Limit theorem Holds

Sample Mean

can be Anything

and Populate in the form of PDF then, it will converted into a normal Distribution



→ Normal Distribution Converted

(8) Poison distribution follows the Pareto Distribution :