Introduction to statistics

defn :⇒ Statistics is the science of collecting, organising, and analyzing the data. This will imply better decision making.

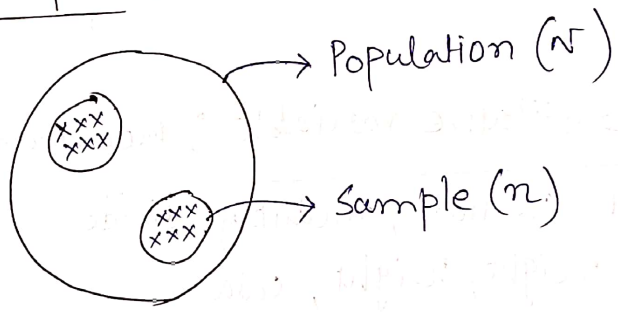
Types :

> Descriptive stats : Consist of organizing and summarizing of the data. ex: Mean, mode, median, S.D, var etc.

> Inferential Stats : It's a technique where in we use the data that we have measured to form conclusions.

ex: Are the age of students of this classroom is similar to the age of math dept students.

Population vs Sample :



Sampling Techniques :

i) Simple Random Sampling : Just go and randomly Pick a sample. Every member of the population has equal chance of being selected for your sample (n).

ii) Stratified Sampling : Population (N) is split into non overlapping groups.

ex: Gender → male, female } Male, Female different. Not overlapped.

ex2: Profession → .net, Python } Sometimes .net developer may know python then overlapping arises. Sometimes it's stratified, sometimes it's not.

iii) Systematic Sampling: Based on n^{th} individual.

Ex: For COVID survey we are checking for every 7th person. (No reason for 7th person. It's our choice)

iv) Convenience Sampling:

A domain expert will take part in specific surveys.

Ex: A ML engineer will survey ML & literacy in a population.

Note: Technique of sampling to be used depends on the use case.

Variables: It can take values.

Ex: weights = { 78, 75, 50, 52 .. }

Types:

i) Quantitative variable: Measured numerically. We can add, subtract, multiply, divide

Ex: Age, weight, height, salary.

ii) Qualitative/categorical variables:

Based on some characteristics we can have categorical variable.

Ex: IQ: (0-10) = Less IQ
(10-50) = (Medium IQ)
(50-100) = (Good IQ)

} No addition, subtraction, Mul, Div possible

Quantitative

Discrete variable

eg: total no of children in family:

eg: 2, 3, 4, 5, 6, 7

continuous variable

Any value it can have

eg: height: 170 cm, 180.5 cm

weight: 65 kg, 63.2 kg

Note:

variable	kind of variable
Grender	Categorical (Qualitative)
Marital status	categorical
River length	Continuous
Population of state	Discrete
Song length	Continuous
Blood pressure	Continuous

Variable Measurement Scales:

i) Nominal Data: categorical data (Qualitative)

ex: Grender, Type of flower

ii) Ordinal: Here order of the data matters but not the value.

ex:

Marks	Rank
100	1
96	2
90	3

This is ordinal data. Rank has order.

iii) Interval: Order and value both matters

Natural zero not present.

ex: Temperature: $(70-80)^\circ F$ $(81-90)^\circ F$ $(91-100)^\circ F$

but $0^\circ F$ is not here.

Ratio Data: It is measured along a numerical scale that has equal distances b/w adjacent values and a true Zero (or meaningful Zero)

ex: weight (kg) → 50, 70, 90

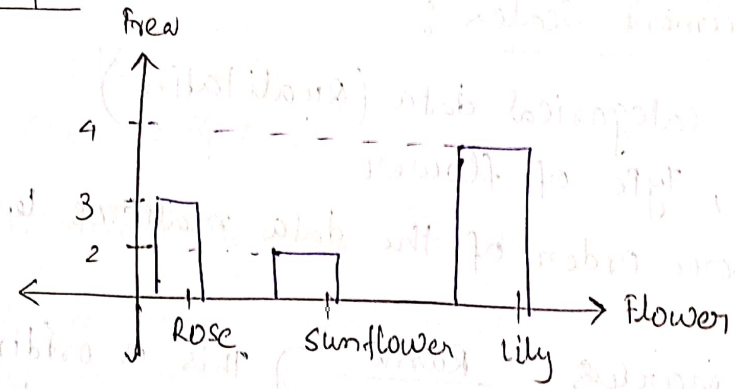
No of staff → 10, 20, 30

Frequency Distribution:

Dataset: Rose, Lilly, Sunflower, Rose, Rose, Lily, Sunflower, Lily, Lily

Flower	Frequency	Cummulative frequency
Rose	3	3
Sunflower	1	7
Lily	2	9

Bar Graph:

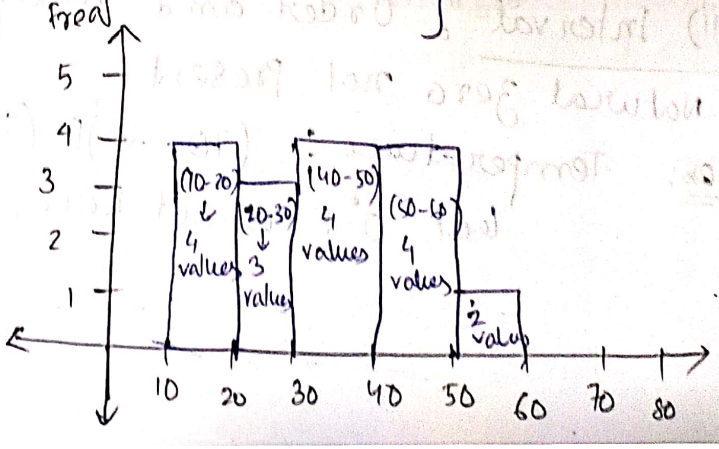


Histogram: Data should be continuous

ex: Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Bins = 10 (Grouping) ↓ Default size

> Pdf's can be built from here.



Bar VS Histogram

> Bar graph is discrete. Histogram is continuous

Measure of Central Tendency:

Mean :

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

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} \quad (\text{Population } (N))$$

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

Now for Sample (n) mean is denoted by $\bar{x}$.

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

Now if in the data set (x) a point '100' is added.

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

$\mu' = 12$

Previously $\mu = 3.2$ Now it's 12. So huge difference.
So, '100' is an outlier. So, Median is better.

Median :

$x = \{1, 1, 2, 2, 3, (3), 4, 5, 5, 6, 100\}$

always sort the data in case of median.

Mode = 3

Now if 112 also added.

$x' = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, (112)\}$ → outlier

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

** Adding outlier also the median change is not high. But in case of avg the difference is much higher.
So, in case of outlier median works better.

Mode : It's defined as most freaquent element.

$x = \{1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 200\}$

Mode $(x) = 6$ (most freaquent)

Significance: Suppose in a dataset there are many missing values. We can use Mode to fill those missing values. It is very useful for Categorical data.

Measure of dispersion :

'Dispersion' means spread.

ex: $x_1 = \{1, 1, 2, 2, 4\}$ $\mu_1 = 2$

$x_2 = \{2, 2, 2, 2, 2\}$ $\mu_2 = 2$

$\mu_1 = \mu_2$. But x_1 and x_2 are differently distributed.

> Variance

Population variance :

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

Sample variance.

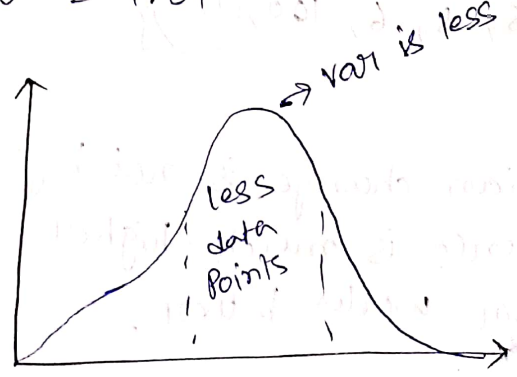
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

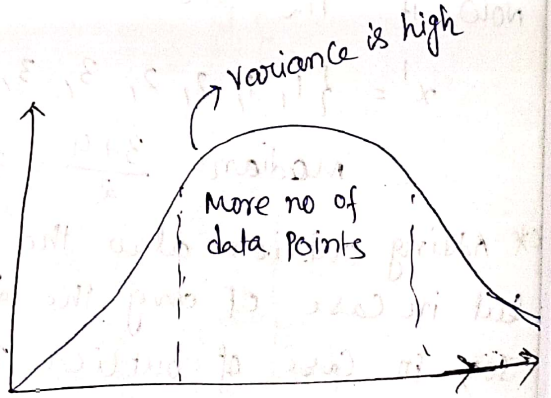
$x = \{1, 2, 2, 3, 4, 5\}$

$\mu = 2.83$

$$\sigma^2 = \frac{(1-2.83)^2 + (2-2.83)^2 + \dots + (5-2.83)^2}{6}$$

$\sigma^2 = 1.81$



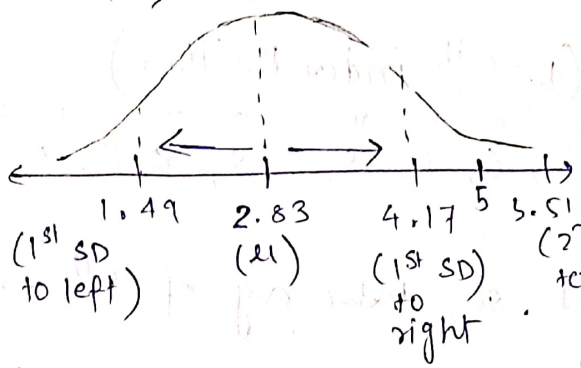


> Standard deviation (S.D)(σ):

$$SD(\sigma) = \sqrt{var(x)}$$

Previous data set $S.D = \sqrt{1.81} = 1.34$

$\mu = 2.83$, $SD = 1.34$



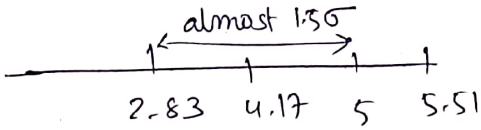
$2.83 + 1.34 = 4.17$, $4.17 + 1.34 = 5.51$

$2.83 - 1.34 = 1.49$

> In b/w 2.83 and 4.17 the values of the dataset falls in 1st SD to the right of mean

> In b/w 1.49 to 2.83 the values of the data set falls in 1st SD to the left of the mean

> 5 falls almost 1.5σ to the right of the mean.



Percentile and Quartiles (Outlier):

Percentile: A Percentile is a value below which a certian Percentage of observation lie.

Ex: $x = \{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, \boxed{10}, 11, 12\}$

12 values ← | → 16 values, $n = 20$

Percentile value of 10 ?

$$Percentile(x) = \frac{\#\ values\ below\ x}{N} * 100\ \%$$

$$Percentile(10) = \frac{16}{20} * 100\ \% = 80\ Percentile.$$

So, 80 % of the dataset less than 10.

$$Percentile(11) = \frac{17}{20} * 100\ \% = 85\ Percentile.$$

∴ what value exists at percentile rating of 25%?

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

$$\Rightarrow \text{value} = \frac{25}{100} * (20+1)$$

$$\Rightarrow \text{value} = 5.25 \quad (5.25^{th} \text{ index position})$$

i = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

$x = \{2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12\}$

17 18 19 20

5.25 index not present. so, take avg of 5th and 6th index value

$$\frac{x[5] + x[6]}{2} = \frac{5+5}{2} = \frac{10}{2} = 5 \text{ (ans)}$$

value of 75 percentile?

$$\text{value} = \frac{75}{100} * (21)$$

$$= 15.75$$

$$\frac{x[15] + x[16]}{2} = \frac{9+9}{2} = 9$$

Five no summary: Five no summary includes

① minimum ② first quartile (Q_1) ③ Median

④ Third quartile (Q_3) ⑤ Maximum

ex: $\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

$$\boxed{IQR = Q_3 - Q_1}$$

Q_1 (25%) $= \frac{25}{100} * (19+1) = 5^{th}$ index $= 3$

Q_3 (75%) $= \frac{75}{100} * (19+1) = 15^{th}$ index $= 7$

$$IQR = Q_3 - Q_1 = 7 - 3 = 4$$

Lower fence (LF) $Q_1 - 1.5\,(IQR)$
High fence (HF) $Q_3 + 1.5\,(IQR)$

$LF = 3 - 1.5\,(4) = -3$

$HF = 7 + 1.5\,(4) = 13$

So, anything less than -3 and greater than 27 is outlier. Remove them.

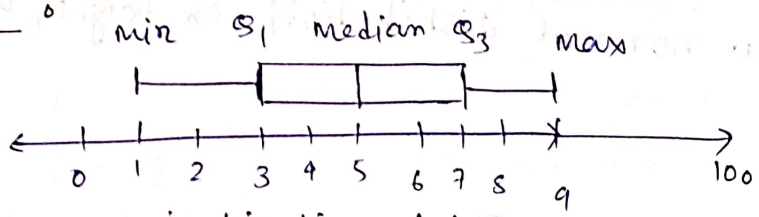
$$\text{outlier} := \begin{cases} x < LF \\ x > HF \end{cases}$$

'27' is removed from the data set.

min = 1
Q_1 = 3
Median = 5
Q_3 = 7
max = 9

} 5 Number Summery

Box - Plot :



Box. plot is a visulization technique to see if there is any outlier.

Population variance vs Sample variance :

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$ This is Population variance

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

Here $(n-1)$ is called as degree of freedom. In Sample data there is lack of information about majority of the data. So, $(n-1)$ is used instead of N. which acts as an automatic ~~shield~~ Shield for the analyst and decision makers.

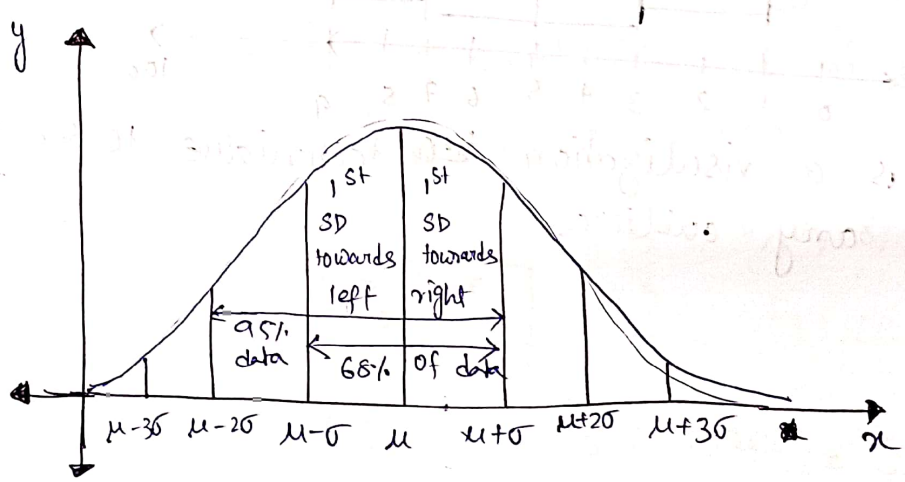
$$\text{value}(s^2) > \text{value}(\sigma^2)$$

So, Sample variance is higher than population variance.

Distributions :

Gaussian Distribution :

It is also normal distribution. ex: Height, weight



Bell curve

$(\mu - \sigma)$ to $(\mu + \sigma)$ = 68% of data

$(\mu - 2\sigma)$ to $(\mu + 2\sigma)$ = 95% of data

$(\mu - 3\sigma)$ to $(\mu + 3\sigma)$ = 99.7% of data.

suppose for a dataset

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

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

$$Z(4.75) = \frac{4.75 - 4}{1} = 0.75 \Rightarrow$$ 0.75 SD towards right of mean (μ).

$$Z(3.75) = \frac{3.75 - 4}{1} = -0.25 \Rightarrow$$ 0.25 SD towards left to mean (μ).

Ex: Dataset $x = \{1, 2, 3, 4, 5, 6, 7\}$ $\mu = 4, \quad \sigma = 1$

$\Rightarrow z(1) = \frac{1-4}{1} = -3$ $\quad z(4) = \frac{4-4}{1} = 0$

$z(2) = \frac{2-4}{1} = -2$ $\quad z(5) = \frac{5-4}{1} = 1$

$z(3) = \frac{3-4}{1} = -1$ $\quad z(6) = \frac{6-4}{1} = 2$

$z(7) = \frac{7-4}{1} = 3$

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

$\Downarrow$

z - score

$z(x) = \{-3, -2, -1, 0, 1, 2, 3\}$

$z(x) \Rightarrow \mu = 0, \ \sigma = 1$ } This is a property of standard normal distribution

$z(x)$ = Standard normal distribution.

Standardization: Performing z-score to all the datapoints of the dataset making $(\mu = 0, \sigma = 1)$ is called standardization.

Basically the process to get standard normal distribution is called standardization.

Q: 2020 ODI series :-

Avg score = 260

S.D = 12

Pant score = 245

2021 ODI series

Avg score = 250

S.D = 10

Pant = 240

Which year Pant played better?

$Z_{2020}(245) = \frac{245-260}{12} = -1.25$

$Z_{2021}(240) = \frac{240-250}{10} = -1$

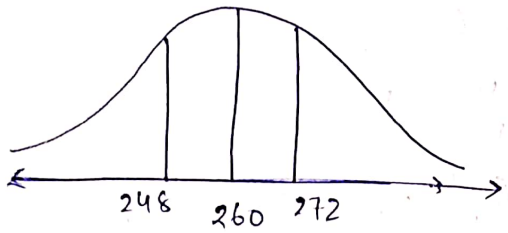


fig: 2020

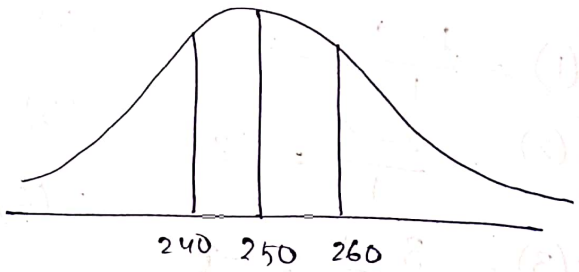
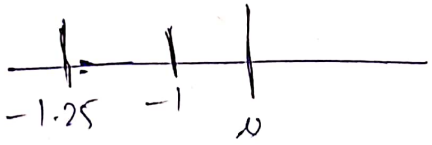


fig: 2021



-1 is close to avg than -1.25. In 2021 Pant played better compared to 2020

Q: In maths you get 83 and in Physics = 68 :-

$\mu_{math} = 83$, $SD_{math} = 10$

$\mu_{Physics} = 62$, $SD_{phy} = 6$.

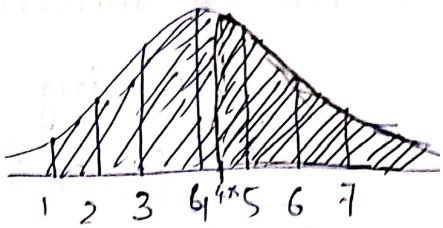
which subject you did better?

$Z(Math) = \frac{80-83}{10} = -0.3$

$Z(Phy) = \frac{68-62}{6} = 1$

In Physics the marks is 1 SD more than avg. You did physics well.

(4) $\mu = 4$, $\sigma = 1$, what % of scores falls above 4.25?



∴ Total area = 1

$$z(4.25) = \frac{4.25 - 4}{1} = 0.25$$

▨ ⇒ Required area

▨ = 1 − ▨ (▨ → from z-table left)

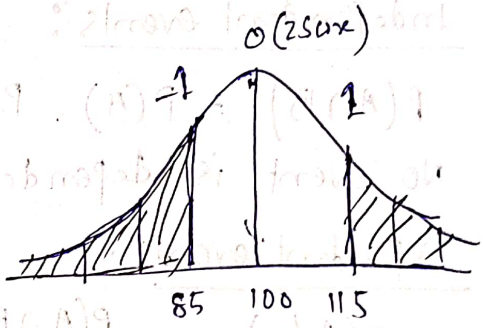
$= 1 - z(0.25)$

$= 1 - 0.5987$

$= 0.4013$

= 40.13 % of the data above 4.25 (ans)

(5) Avg I.Q = 100. SD = 15. What % of the population would you expect to have an IQ lower than 85?



$\Rightarrow z(85) = \frac{85 - 100}{15} = -1$

~~It's left part. Directly~~ check left Part of 85 = right Part 115. bcz of symmetry.

Area = $1 - z(1)$

$= 1 - 0.84134$

$= 0.15866$

= 15.86 % of the data (ans)

Probability

It's a measure of likelyhood of an event.

eg: Rolling of dice.

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

$P(\text{even}) = \frac{3}{6} = \frac{1}{2} = 50\%$

Mutual exclusive event:

Two events can't happen at the same time.

eg: Tossing a coin. We can get either head or tail. Not both at the same time.

Non mutual exclusive event:

Multiple events can happen at the same time.

ex: Deck of Cards.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independent events:

$$P(A \cap B) = P(A) \cdot P(B)$$

No event is dependent on other event.

Dependent events:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$\Rightarrow P(A \cap B) = P(B) \cdot P(A/B)$$

P. value :

Here hypothesis testing, Confidence interval, Significance value all are included.

Problem: Test whether the given coin is a fair coin or not.

we know $P(H) = 0.5 = P(T)$

Hypothesis testing:

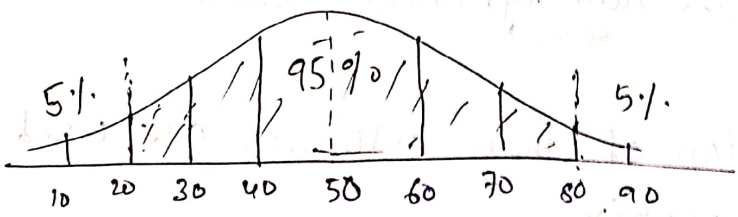
(1) Null hypothesis : Coin is fair.

(2) Alternate hypothesis : Coin is un-fair.

(3) Experiment (Z test, T test or anything)

(4) Reject or Accept the null hypothesis.

Procedure : Tossing of 100 times suppose we get 30 heads.



No of heads : $\mu = 50$, $\sigma = 10$ } suppose.

Significance value (α) = 0.05

So, $\alpha = 0.05 = 5\%$. (α is set by domain expert)

Confidence interval = $1 - \alpha$

$= 100 - 5$

$= 95\%$

Now Suppose 20 to 80 covers 95% of the area Then if no of heads is in b/w 20 to 80 then it's a fair coin. Otherwise not.

Decision :

Null hypothesis is true or Null hypothesis is false.

Outcome - 1 :

We reject the null hypothesis, when in reality it's false.

Outcome - 2 :

We reject the null hypothesis, when in reality it is true ⇒ Type 1 error.

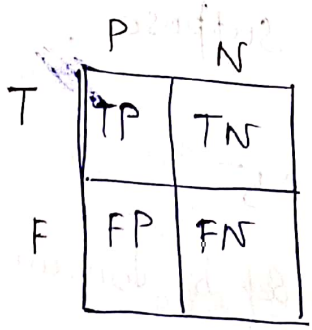
outcome - 3 :

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.

ex: Application of these outcomes are used in confusion matrix



Actual →

Pred ↓	1	0
1	TP	FP
0	FN	TN

TP

	1	0
1	TP	FP
0	FN	TN

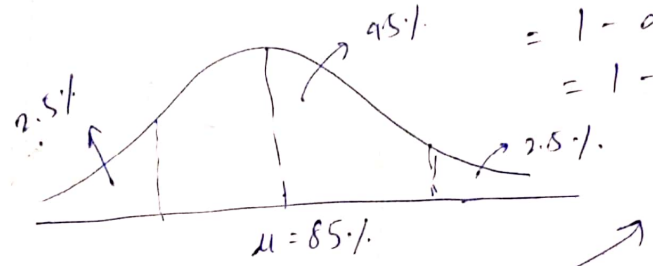
0 1
F T
T F
1 0

(T)(T)
(T)(T)

1 tail and 2 tail test :

Suppose a college has 85% placement record. Now you are given a new college. It is found that at the new college a sample of 150 students had a placement rate of 88% with a standard deviation of 4%. Does this new college has a different placement rate? significance value (α) = 0.05.

Confidence interval (CI)
$= 1 - \alpha$
$= 1 - 0.05 = 95\%$
(3 part area divided)



2 tail test

Now if the problem statement was: does this new college has better placement than that old college.

1 tail test.

Here better is told so, $> 85\%$ R.H.S of μ. 1 tail test because area is divided in two parts

Point estimate :

The value of any statistic that ~~estimate~~ estimate the value of a parameter.

Inferential stat :

Sample → estimate → Population

$\bar{x}$ given estimate Population mean (μ)

Confidence Interval:

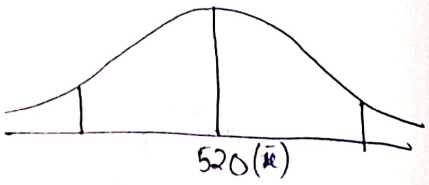
Point estimate ± Margin of error.

⑧ ** In CAT exam std = 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean? ($\alpha = 0.05$)

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

CI = Point estimate ± margin of error

$$CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



$$\text{Standard error} = \frac{\sigma}{\sqrt{n}}$$

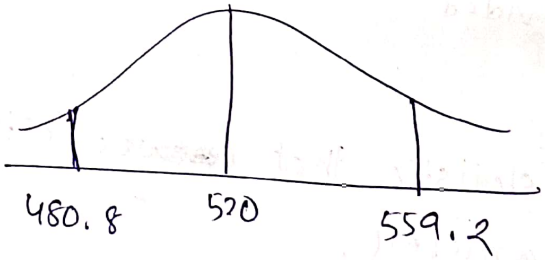
$$CI = \begin{cases} \text{upperbound} = \bar{x} + Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}} \\ \text{lowerbound} = \bar{x} - Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}} \end{cases}$$

$$CI = \bar{x} + Z_{0.025} \frac{100}{5}$$

($Z_{0.025} = 1.96$ from Z table)

$CI_{upper} = 520 + 1.96 * 20 = 559.2$

$CI_{lower} = 520 - 1.96 * 20 = 480.8$



(8) In CAT exam a sample of 25 test takers has a mean of 520 with a Std = 80. Construct 95% CI about the mean?

⇒ Here population std not given. Sample std is given. So, Perform T-test. ($\alpha = 1 - 95\% = 0.05$)

$\alpha = 0.05$, $\bar{x} = 520$, $\sigma = 80$

$$CI = \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\text{Standard error} = \frac{s}{\sqrt{n}}$$

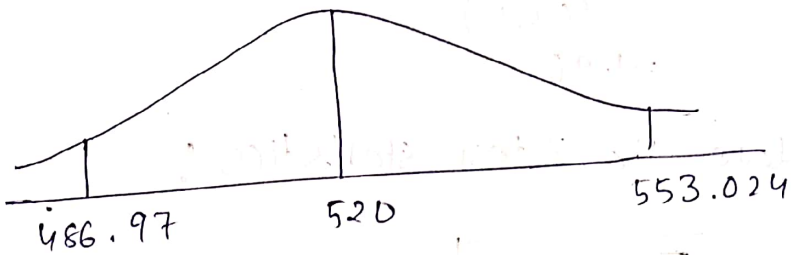
$$\text{Degree of freedom} = n - 1 = 25 - 1 = 24$$

** $t_{\alpha/2}$ is dependant on degree of freedom.

$$\text{upper bound} = \bar{x} + t_{\frac{0.05}{2}}\left(\frac{s}{\sqrt{n}}\right) = 520 + 2.064\left(\frac{80}{\sqrt{25}}\right) = 553.024$$

$$\text{lower bound} = \bar{x} - t_{\frac{0.05}{2}}\left(\frac{s}{\sqrt{n}}\right) = 486.97$$

($t_{0.025} = 2.064$; from t table with degree of freedom 24)



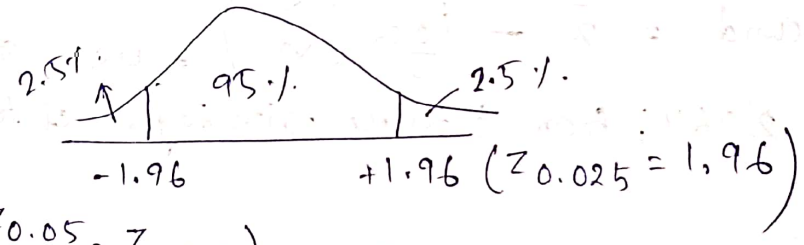
(9) In the population avg IQ = 100. with std = 15. Researchers to test new medicine to see if there is positive or negetive effect on intelligence or not effect. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication effect the intelligence ? ($\alpha = 0.05$)

⇒ ① Null hypothesis (H_0) : $\mu = 100$

② Alternate hypothesis (H_1) : $\mu \neq 100$

③ $\alpha = 0.05$

④ State decision rule : here it is said 'did the medication affect the intelligence ?' It means it may increase or decrease. So, perform 2 tail test



($Z_{\frac{0.05}{2}} = Z_{0.025}$)

$= -1.96$

⑤ Calculate the Z test statistics :

$$Z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} = \frac{140 - 100}{15/\sqrt{30}} = 14.60$$

⑥ State our decision :

Value has to be in b/w -1.96 to $+1.96$ but $14.60 > 1.96$.

So, Null hypothesis (H_0) is rejected. That means avg IQ $\neq 100$.

$14.60 > 1.96 \Rightarrow$ improved intelligence.

Note :

i) Z-test ⇒ Population Std

ii) T-test ⇒ Unknown Population Std.

⑧ Population avg IQ = 100, $n = 30$, $\bar{x} = 140$, $S = 20$

$\alpha = 0.05$. Did the medication affect?

① $H_0 : \mu = 100$ (Population Std not given. So, T-test)

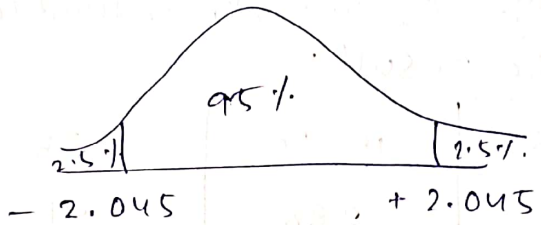
② $H_1 : \mu \neq 100$

③ Degree of freedom $= n - 1 = 30 - 1 = 29$

④ Decision Rule:

$t_{\frac{0.05}{2}} = 2.045$

(Degree of freedom = 29)



⑤ T-test:

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

$$\Rightarrow T = \frac{140 - 100}{20/\sqrt{30}} = 10.96$$

⑥ State our decision:

$10.96 > 2.045$.

So, $\mu \neq 100$, H_0 is rejected.

IQ has increased.

Chi Saluare Test :

> Chi saluare test claims about Population Proportion

> It is a non parametric test that is Performed on categorical (nominal or ordinal) data.

(3) In the 2000 indian Census, the age of the individual in a small ~~tea~~ town were found to be the following:

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

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

< 18	18 - 35	> 35
121	288	91

$\alpha = 0.05$ would you conclude the Population distribution of ages has changed in the last 10 years?

⇒ From Census of 2000 expected Population

< 18 : 500 * 0.2 = 100

18 - 35 : 500 * 0.3 = 150

> 35 : 500 * 0.5 = 250

~~Observed : 100 150 250~~

Observed : 121 288 91

Expected : 100 150 250

Step - 1 : H_0 = The data meets the distribution of 2000

H_1 = The data does not meet the distribution.

Step - 2 : $\alpha = 0.05$

Step - 3 : Degree of freedom = $n - 1 = 3 - 1 = 2$

step 4 : Find chi saluare (χ^2) value at α and degree of freedom.

Here $\alpha = 0.05$, degree of freedom = 2

$$\chi^2_{\alpha=0.05} = 5.99$$

dof = 2

so, if χ^2 is greater than 5.99 then reject H_0

Step-5 : Calculate the χ^2 value for this problem.

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

(f_o = observed, f_e = expected)

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

$$= 232.94$$

$\chi^2 = 232.94 > 5.99$ (Reject the null hypothesis (H_0))

So, H_1 is true.

Covariance :

Weight	Height
50	160
60	170
70	180
75	181

$x\uparrow$	$y\uparrow$
$x\downarrow$	$y\downarrow$

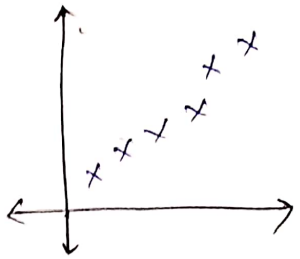
Study Time	Play Time
2	6
3	4
5	2

$x\uparrow$	$y\downarrow$
$x\downarrow$	$y\uparrow$

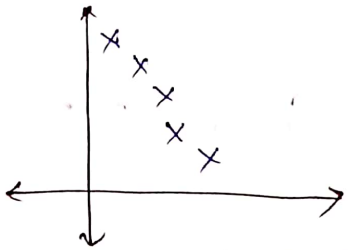
$$Cov(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

if $Cov(x,y) = +ve \Rightarrow$ $x\uparrow$ $y\uparrow$ $\Rightarrow x \propto y$; $x\downarrow$ $y\downarrow$

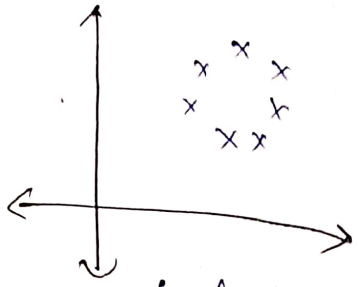
$Cov(x,y) = -ve \Rightarrow$ $x\downarrow$ $y\uparrow$ $\Rightarrow x \propto \frac{1}{y}$; $x\uparrow$ $y\downarrow$



+ve Cov(x,y)



- ve Cov(x,y)



Cov(x,y) = 0
(No relation)

Disadvantage of Cov(x,y):

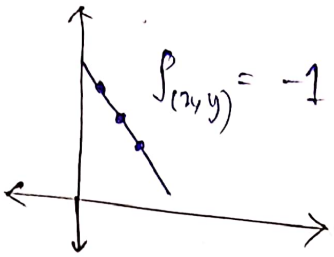
We can tell that it's +ve Cov(x,y) or -ve Cov(x,y). The cov But the value can't be told.

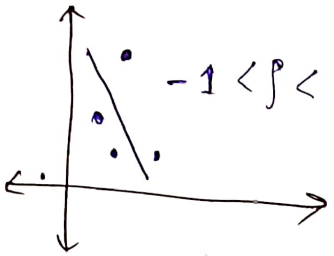
Soln ⇒

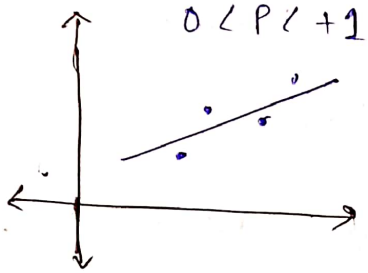
Pearson Correlation Co-efficient:

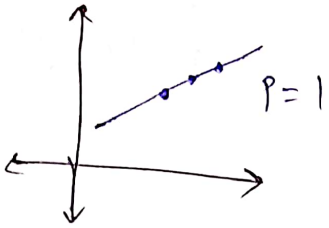
It ranges b/w -1 to +1. The more towards +1 more positive co-relation. The more towards -1 more negative co-related.

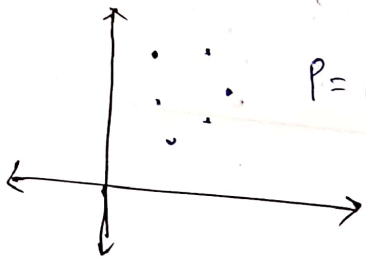
$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$











Pearson Co-relation co-efficient is of linear type.

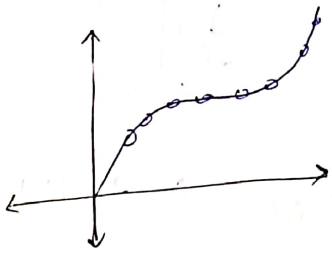
For non linear co-efficient we can use spearman co-relation co-efficient.

Spearman Corelation : $\frac{Cov(R(x), R(y))}{R\sigma_x * R\sigma_y}$

$p(i)$: Rank of i

Ex:

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



$p(x) = 1$ for height = 180 (highest)

Rank 2 = 170 height.

spearman Co-relation Captures the non-Linear Properties.

(2) The avg weight of all residants in Bangalore city is 168 Pounds with a std = 3.9. we take a sample of 36 Persons and mean = 169.5 Pounds. CI = 95%.

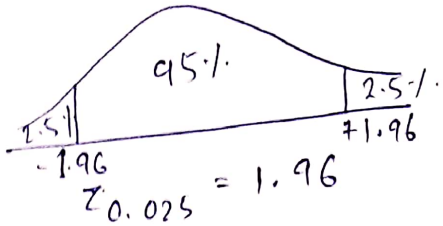
$\Rightarrow \mu = 168, \sigma = 3.9, \bar{x} = 169.5, n = 36, \alpha = 0.05$

(1) H_0 : $\mu = 168$

H_1 : $\mu \neq 168$.

(2) $\alpha = 0.05$

(3)



$Z_{0.025} = 1.96$

(4) ~~α = ... $Z_{0.025}$~~

$$Z_{test} = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$$

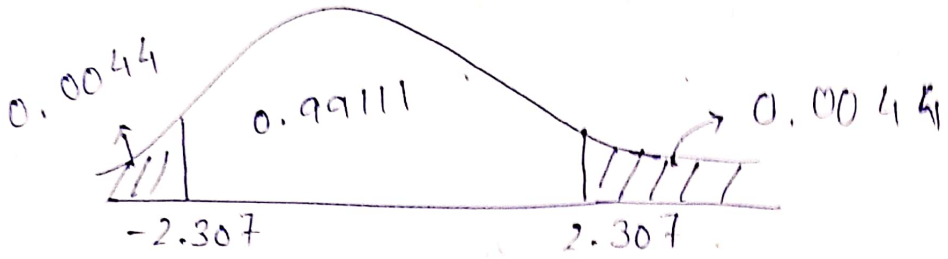
$$Z_{test} = \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.307$$

(5) $Z_{test} > 1.96$, so, H_0 is rejected.

P-value :

Previous Problem $Z_{0.025} = 1.96$

$Z_{test} = 2.307$



$Z_{2.307} = 0.99111$

⇒ Area $= 1 - Z_{2.307}$

$= 1 - 0.99111$

$= 0.00889$

Area (▨) $= \frac{0.00889}{2} = 0.0044$

$P_{value} = 0.0044 + 0.0044$

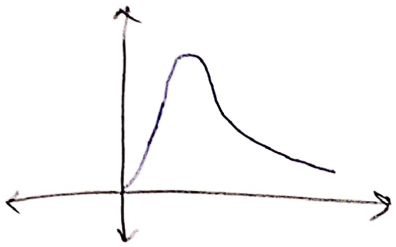
$P_{value} = 0.0088$

if $P_{value} < \alpha$ = Reject H_0

$P_{value} > \alpha$ = Accept H_0

Here $0.0088 < 0.05 \Rightarrow$ Reject H_0

Log normal distribution



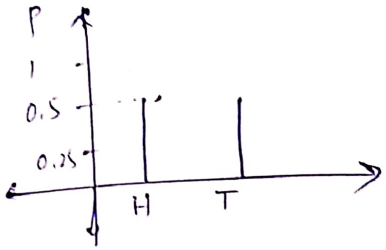
eg: wealth distribution

y is random variable and it belongs to log normal distribution if log(y) follows normal distribution.

Bernoulli Distribution :

It has two outcomes: (P and Q; Q = 1 - P)

P(H) = 0.5 = P(T)



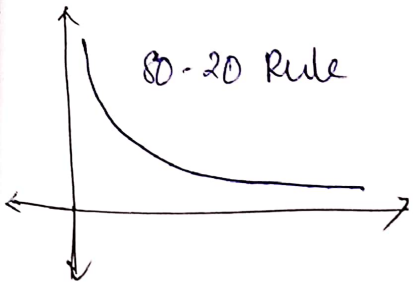
(Prob mass function)

It is for single trial.

Binomial Distribution :

It is for multiple trails. Each trail follows Bernoulli Distribution.

Power-law distribution :



eg: 80% of the wealth is distributed 20% of people.