1 May

NOTES KASI

1 Population (N)

Collection of all items

eg: NewYork University Students

Sample (n)

Subset of Population

This is called Statistics.

Less costly, less time consuming

Sample

Randomness

Representativeness

Accurately represents the members of Entire population

NYU Students → Canteen, Leave, Class Students, Offline Class Students (Groups)

Sample: 50 students

Random :-

Representatives :- Considering all groups

2. Various types of Data :-

Based on type :-

1. Categorical :- Car brands BMW, Audi,... Yes/No Questions ; Climate Seasons

2. Numerical :- Discrete - No. of Children in Class - Integer

Continuous - Weight of a Person (depends on Scaling...)

$10/3 = 3.33... \bar{3}$ 63.4789 kgs

Based on Level of Measurement :-

1. Qualitative - Nominal - Car brands - No Order

Ordinal - Seasons - Summer → Rainy → Winter - Order

2. Quantitative (Represented by Num Quantity)

Interval - No True Zero - °C / °F (No True Zero)

Ratio - Real Scenarios - True Zero -

eg: Two Apples 6 apples

2 : 6

eg: Temp °K (Kelvin) (Has True Zero)

0°K

Today: 5°C or 41°F

Yesterday: 10°C or 50°F

In terms of °C → We feel More Temp → double

In terms of °F → We don't feel diff

NOTES

These scales are artificially created.

In terms of °K (Has True Zero).

Nothing can be colder than $-273°C \cong 0°K = -459°F$

$5°K : 10°K$ (We can)

$5°C : 10°C$ (We don't)

---

* Visualize Data :- Can be done by Knowing its Type & Measurement Level

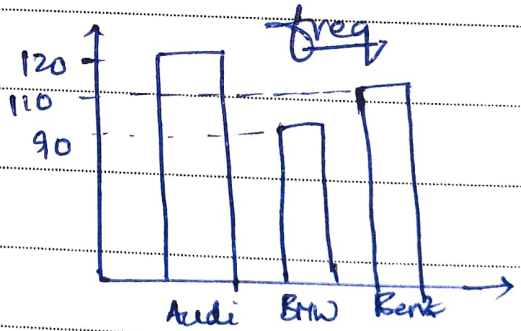
I. for Categorical Variables

1. Freq. Distribution Table
2. Bar Chart
3. Pie Chart
4. Pareto Diagram

1. Freq. Distribution Table:

| Car | Freq | Relative freq |
|---|---|---|
| Audi | 120 | $\frac{120}{320} \times 100 = 38\%$ |
| BMW | 90 | $\frac{90}{320} \times 100 = 28\%$ |
| Benz | 110 | $\frac{110}{320} \times 100 = 34\%$ |
| | 320 | 100% |

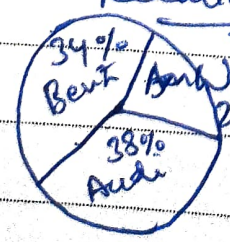
2. Bar Chart / Column Chart





3. Pie Chart


Relative freq




4. Pareto Diagram

Cumulative freq.

Sum of Relative freq.

Descending Order of freq.

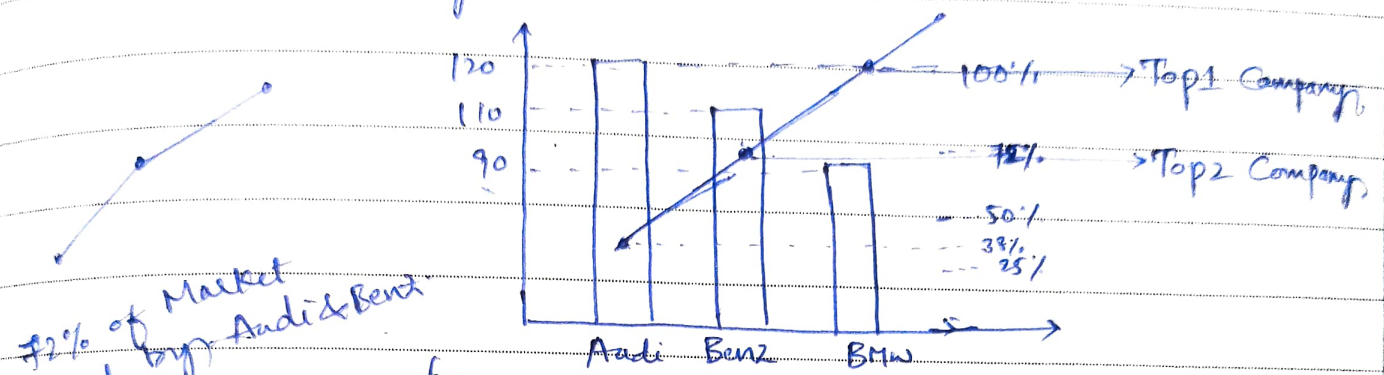
Bar Chart Data + Pie Chart data can be seen

NOTES

Reordering:- Descending Order of freq.

| Car | Freq. | Relative freq. | Cumulative freq. |
|---|---|---|---|
| Audi | 120 | 38% | → 38% |
| Benz | 110 | 34% | → 72% |
| BMW | 90 | 28% | → 100% |

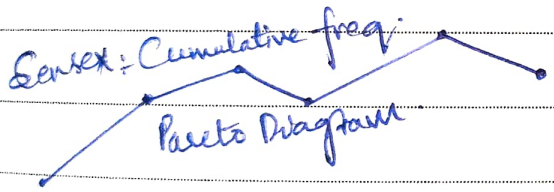
Pareto Diagram




72% of Market covered by Audi & Benz

~~Relative~~ Cumulative freq. difference, we get Relative freq.

Sensex:- Cumulative freq.




Useful in Real Time

1. Histogram
~~2. Cross tables~~
~~3. Scatter Plots~~

for Numerical Variables :-

Freq. Distribution Table

Dataset - freq

| | |
|---|---|
| 1-1 | 50-1 |
| 9-1 | 51-1 |
| 22-1 | 59-1 |
| 24-1 | 60-1 |
| 27-1 | 74-1 |
| 29-1 | 79-1 |
| 30-1 | 80-1 |
| 31-1 | 89-1 |
| 41-1 | 90-1 |
| 42-1 | 92-1 |
| 46-1 | 94-1 |
| 48-1 | 100-1 |
| 49-1 | |

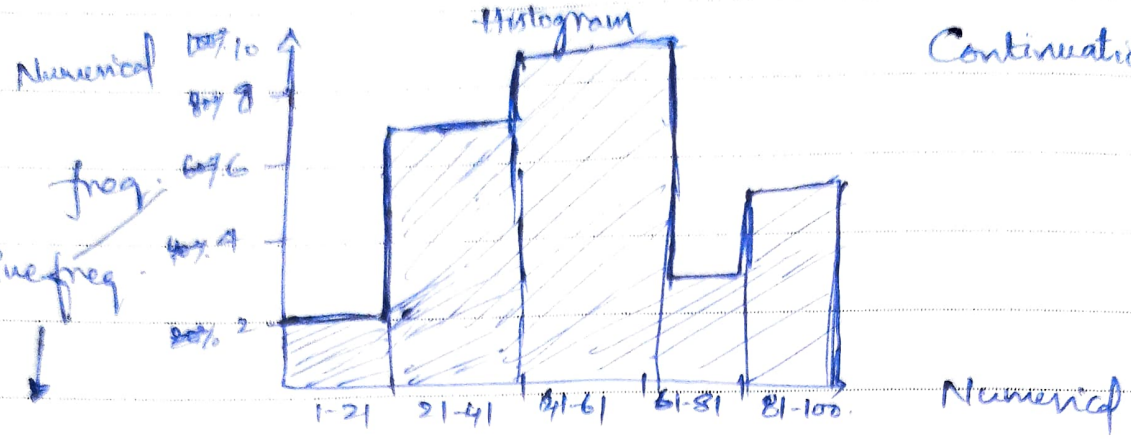
Desired intervals = 5

Interval width = $\frac{\text{Largest No.} - \text{Smallest No.}}{\text{Desired Interval}}$

$= \frac{100-1}{5} \cong 20$ (But have to use exact width)

| Interval start | Interval End | Freq | Relative freq. |
|---|---|---|---|
| 1 | 21 (Include) | 2 | $\frac{2}{26} \times 100 = 7.69\%$ |
| (Don't include) 21 | 41 | 7 | 26.9% |
| 41 | 61 | 9 | 34.6% |
| 61 | 81 | 3 | 11.5% |
| 81 | 100 | 5 | 19.2% |
| | | 26 | 100% |

NOTES




Continuation of Intervals

Graph looks same, bcz if we use Relative freq.

Histogram can be done with unequal Intervals

| Age | | diff Interval |
|---|---|---|
| 18-25 | → | 7 |
| 26-30 | → | 5 |
| 31-35 | → | 5 |
| 60+ | → | - |

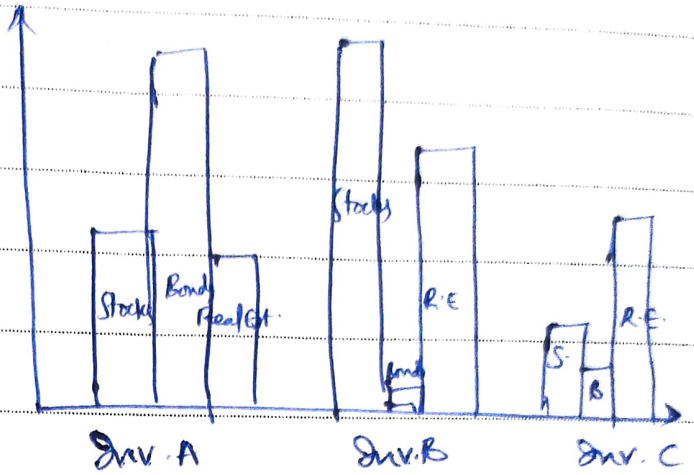
*** Relationship b/w 2 Variables:

1. Categorical - Cross Tables
2. Numerical - Scatter plots (Both Variables are Numerical)

Cross Tables

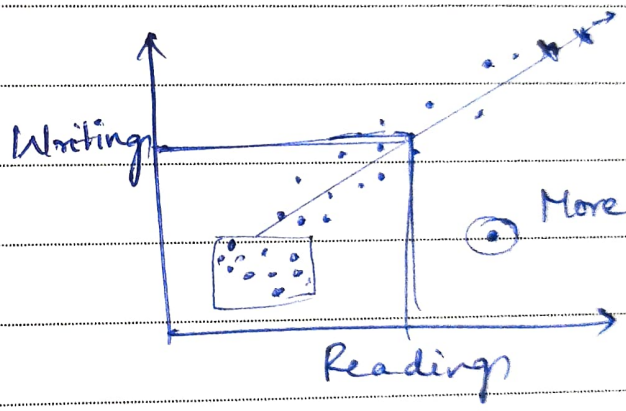
| Type of Investment \ Investor | Investor A | Inv B | Inv C | Total |
|---|---|---|---|---|
| Stocks | 96 | 185 | 39 | 320 |
| Bonds | 181 | 3 | 29 | 213 |
| Real Estate | 88 | 152 | 142 | 382 |
| Total | 365 | 340 | 210 | 915 |

## Cross Tables - Side by Side Bar Chart




## Scatter Plot:-

| Student ID | Scores: Reading | Writing |
|---|---|---|
| 1 | 273 | 216 |
| 2 | | |
| 3 | | |
| : | | |
| 10 | 390 | 206 |




from graph

Conclusion:-

Uptrend:-

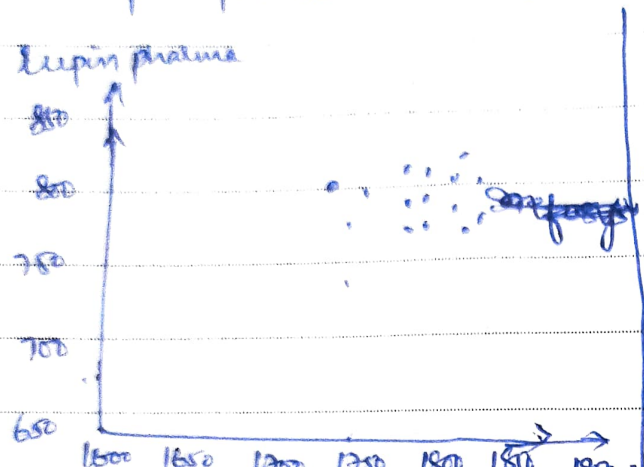
1. → Most students getting good scores in Writing, Scoring good in Reading.
2. → Average Most are concentrated here.
3. ** High Reading, High Writing scorer.
4. [ ] Bad/low scores people - Performance trends in deviation while performing different tasks
5. ⊙ Good at Reading, Bad @ Writing ⇒ Outliers.

## Scatter Plot data Analysis

NOTES

Each Share Value

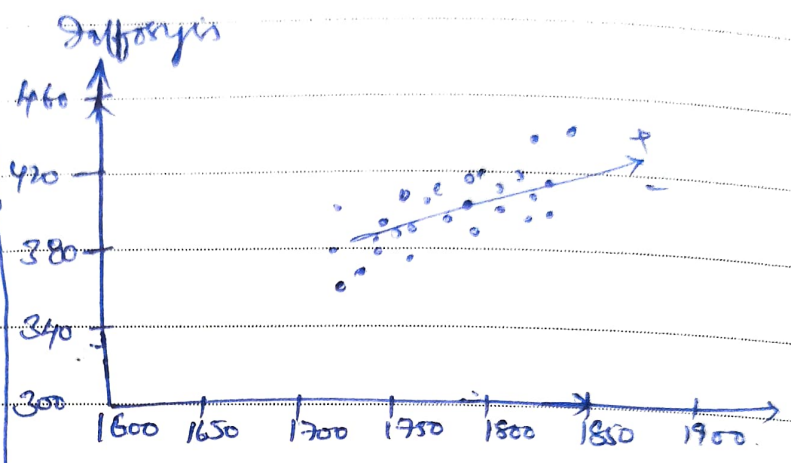
I.T. TCS Share +1800-
I.T. Infosys Share +400-
Lupin Pharma Share +800-




Yes

Independent

No Correlation bcz
When is ↑ing, the other not.
Vertical somewhat.




Yes

Dependent.
highly correlated bcz.
When One is higher, the other is & Vice versa

* Measures of Central Tendency:- Mean, Median, Mode
* Measuring Skewness
* 3-Measures of Variability
  (i) Standard Deviation & (ii) Coefficient of Variation
* Covariance
* Correlation Coefficient

NOTES

## * 3-Measures of Central Tendency:-

**Mean:-** $\mu$ (population) $\bar{x}$ (sample)

fairly effected by outliers

$$\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{or} \quad \frac{x_1 + x_2 + \cdots + x_N}{N}$$

| Student | Marks |
|---|---|
| 1 | 10 |
| 2 | 20 |
| 3 | 30 |
| 4 | 40 |
| 5 | 50 |
| 6 | 0 |

$$\text{Mean}(5) = \frac{10+20+30+40+50}{5} = 30 \quad \text{Marks}$$

$$\text{Mean}(6) = \frac{10+20+30+40+50+0}{6} = 25$$

**Median:-** Middle No. in Ordered Dataset.

No. at position $\frac{(n+1)}{2}$ in Ordered list

$$\text{Median}(5) = \frac{5+1}{2} = 3 \rightarrow \text{Marks } 30$$

$$\text{Median}(6) = \frac{6+1}{2} = 3.5 \rightarrow \frac{30+40}{2} = 35$$

**Mode:-** Used for both Categorical & Numerical

10, 20, 10, 30, 40, 50 → Mode = 10

"Repeated"

Can have single / multiple modes or no mode.

* Which is best? Depends on situation.

For eg:- Income of 1st year & 2nd years joiners in TCS → Mean is best

Income of all employees (including seniors) → Outliers → Median is best

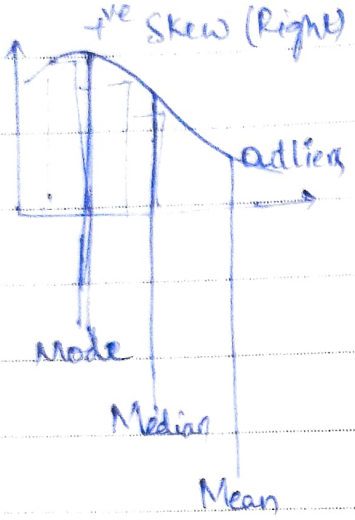
NOTES

To know Asymmetry of Distribution

**Skewness:**

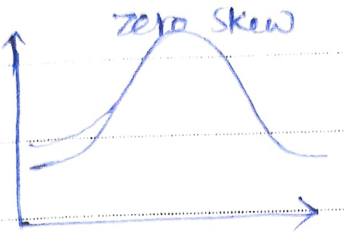
| Mean | Median | Mode |
|---|---|---|
| 2.79 | 2 | 2 |

Mean > Median




| Mean | Median | Mode |
|---|---|---|
| 4 | 4 | 4 |

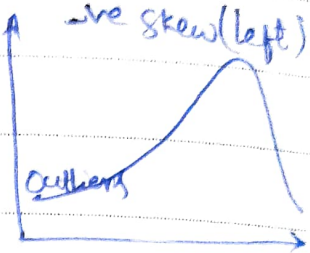
Mean = Median




Distribution is Symmetrical

| Mean | Median | Mode |
|---|---|---|
| 4 | 5 | 6 |

Mean < Median




Outliers are situated in left

How skewness is used?

It gives where most of the data is situated.

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^3}$$

3 Measures of Variability (i) Variance (ii) Std-Deviation.
(iii) Coef. of Variation.

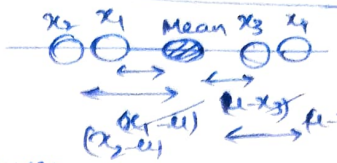
(i) **Variance** ($\sigma^2$)

Population (100) ← Sample 1 – 10 ... Sample 10 – 10 (100)

All 10 samples gives different measures.


Variance measures the dispersion of a set of data points around their mean. (Spread)

Mean ($\mu$)



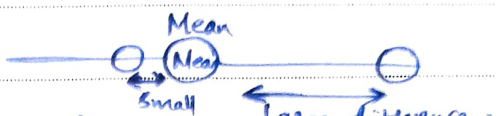

Error :- We square, ∆ ⇒ Otherwise

NOTES

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

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

* Squaring :- Error Sum +0.2 & -0.2 May Cancel out
To avoide, We square.




* ~~Error amplified by Squaring~~.

eg

Population:
1
2
3
4
5

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

pop. Variance $\sigma^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5} = 2$

Consider as Sample, Sample Variance $s^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1} = 2.5$

Why $s^2_{\text{Sam Variance}} > \sigma^2_{\text{Variance}}$ ?

Imaginary population
1
1
2
3
4
5
5
5
5

Mean = 3.2
$\sigma^2 = 2.96$

sample popul
1
2
3
4
5

eg:- Salaries of 5 Team Members on our project.

30000/-
20000/-
1,20,000/-
45,000/-
32,000/-

Variance ($s^2$) = 1636800000

3000
3100
3200
3300
3400

$s^2 = 25000$

(ii) Standard Deviation = ($\sigma$) ($s$)

$\sqrt{\sigma^2}$ $\sqrt{s^2}$

NOTES

(iii) Coef. of Variation :- / Relative Std. Deviation.

$$C.V = \frac{\text{Std. Deviation}}{\text{Mean}} = \frac{\sigma}{\mu} \text{ (or) } \frac{s}{\bar{x}}$$

population: $C_v = \frac{\sigma}{\mu}$ Sample: $\hat{C}_v = \frac{s}{\bar{x}}$

$\sigma$ :- Std. Deviation is the most common measure of variability for a Single Dataset.

$C_v$ :- Coef. of Variation is comparing Two or More Datasets

eg:-

| | Dollars | Rupees |
|---|---|---|
| 1 | | 73/- |
| 2 | | 146/- |
| 3 | | |
| 4 | | |
| : | | |
| 10 | | 730/- |
| Mean $\bar{x}$ | 5.5 | 401.5 |
| Var $s^2$ | 9.16 | 48849.17 |
| $s$ | 3.02 | 221.01 |
| $\hat{C}_v = \frac{s}{\bar{x}}$ | 0.581 | 0.551 |

Dollar Rupees
$\hat{C}_v \simeq \hat{C}_v$

Conclusion:-
Two Datasets have same variability

eg:- Income of People in U.S. Vs. Denmark Sample.
Currency is different. $ Kr
We can compare $\hat{C}_v$.

NOTES

Measures of Relationship b/w Variables 1. Covariance
2. Linear Correlation Coef.

* Covariance:

| Land Size (x) | Price (y) |
|---|---|
| 650 | ₹ 72,000 |
| 780 | 998000 |
| 1200 | 1200000 |
| 720 | 800000 |
| 975 | 895000 |

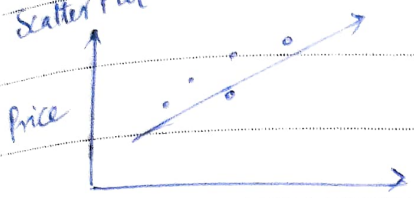
Sample Covariance

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Pop Covariance N

$$\sigma_{xy} = \frac{\sum_{i=1} (x_i - \mu_x) * (y_i - \mu_y)}{N}$$

Scatter Plot




During $\Sigma$ (Summation) + - also summed up sign also.

Cov. Sample = 33,491,250

Covariance >0, the two variables moving together
<0, " " " opposite
=0, " " " are independent.

Covariance
30000.
50
0.23

How Can We Know, Covariance More or just more?
Sol:- We go for Correlation Coef.

* Correlation Coef:- -1 to 1

$$\frac{Cov(x,y)}{Std.dev(x) * Stdev(y)}$$

Cor. Coef = 0.87
Highly correlated
+ve value indicate
If Size of Land ↑ ⇒ Price ↑

1 → Perfectly Correlated ↑↑ or ↓↓
0 → Independent
-1 → Perfectly Correlated ↑ ↓
Icecream & Umbrella in Season (Rainy)

* Causality:-
Size causes Price, but not Vice Versa
Correlation doesn't imply causation.

# Descriptive Statistics

NOTES

Descriptive Statistics Real Example

Product Table

| ID | Type of Property | Area | Price | Status | Cust. ID | Name | Age | Gender | Country | State |
|---|---|---|---|---|---|---|---|---|---|---|
| ↳ Categorical (Qualitative (Nominal)) | | ↳ Numerical | ↳ Numerical | → Sold / Pending → Empty - No one has bought | → Categorical | ↳ Categorical | ↳ Quantitative (Ratio) | Categorical (Nominal) | | Categorical |

Customer Table

Age
Mean 46.16
Median 45
Mode 48
Skew 0.24
(Right skew)

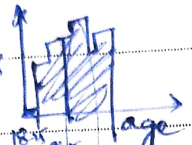
Sample Dataset

Sample formulas

Variance 164
St. dev 12.84

③ * Age → Histogram → (Numerical)

@ What age more customers are buying Hocere?




* Gender → Pie Chart. (Categorical)

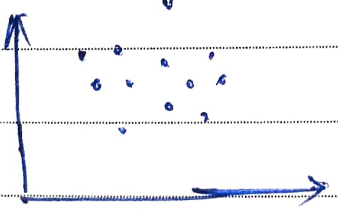
① Individuals, Firms, Unknown, Corporates, Male, Female, Individuals

* State → Pareto diagram

② N.C. NY

@ Age Vs Price. (Scatter Plot)

④



Covariance -17618.87
Correlation coef. -0.17

Inference:
Price Not related to age.

1. Females are buying more. (Gender)
2. Most sales on N.C. state (State)
3. ~~Age~~ 36-46 are buying more with Avg mean & Std dev of
Right skewness → Younger people buy more than older people
4. No relation b/w Age & Price they are buying.

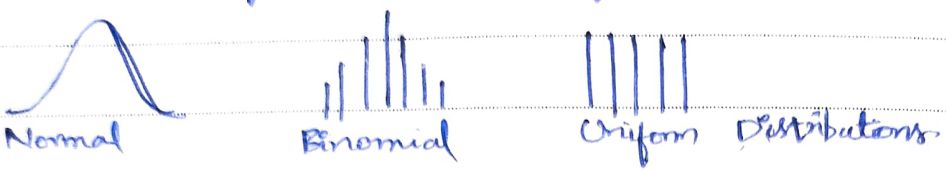
# Inferential Statistics :- Infer → To give some proper conclusions.

Probability Theory & Distributions based on sample data.

NOTES

- Distribution
- Point Estimates
- Confidence Intervals

**Distribution** :- Usually Means Probability distribution.



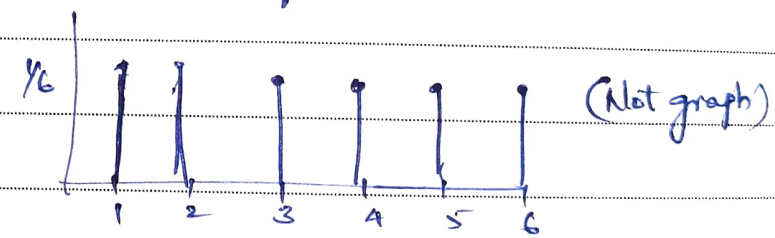

A distribution is a function that shows the possible values for a variable and how often they occur.

Numerical - Discrete Distribution :

Die & prob of getting of 1

| Outcome | Probability |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |
| 7 | 0 |

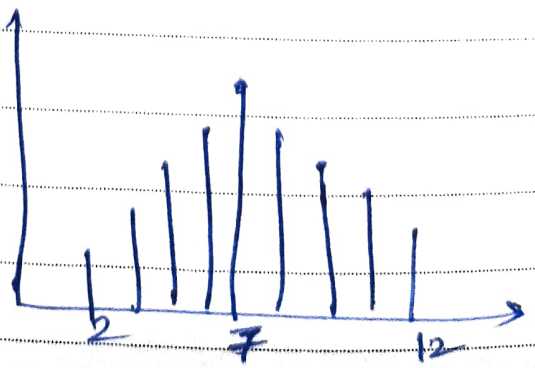
$\Rightarrow$ Discrete uniform Distribution.




Die :- Rolling Two Dice.

(1,1) (1,2) ... (1,6)
(2,1)
⋮
(6,1) ..... (6,6)

| Outcome Sum | Prob. |
|---|---|
| 2 | 1/6 |
| 3 | 2/36 |
| 4 | |
| ⋮ | ⋮ |
| 12 | 1/6 |




Binomial

T. Kani Pradeep

NOTES

## Continuous Distribution

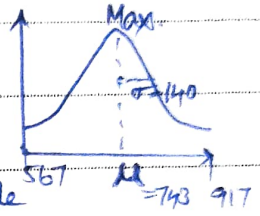
### Normal Distribution :-

* They approximate a wide variety of Random Variables
* Distribution of sample means with large enough sample sizes could be approximated to Normal
* ~~Computable~~ Decisions based on Normal Distribution is good.

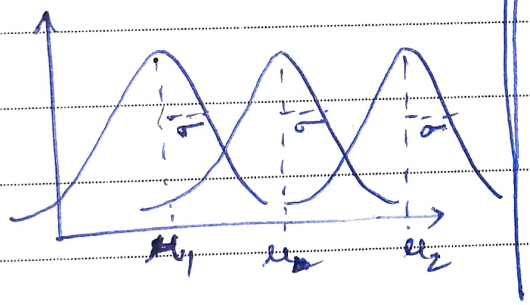
$N \sim (\mu, \sigma^2)$

Distribution, Mean, Variance

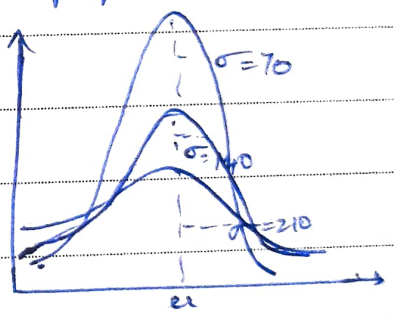
Bell Curve
Normal Distribution
Gaussian Distribution
Symmetrical ⇒ Skew = 0
⇒ Mean = Median = Mode




Controlling for Std. Deviation:




Controlling for Mean :- By Changing Std. dev.




More data on Middle
less in thicker tails

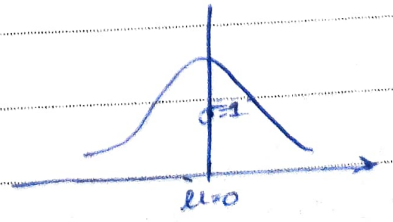
Higher Std. deviation ⇒ Makes graph flat ⇒ fatter tails

### Standard Normal Distribution

$N \sim (\mu, \sigma^2) \sim (0, 1)$

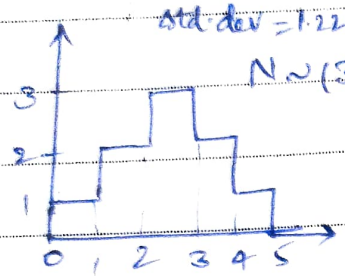
***

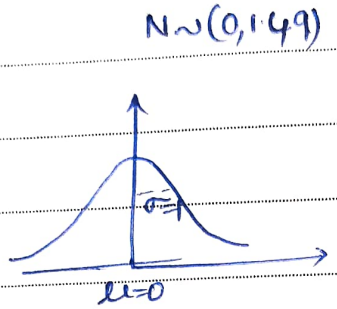
$Z = \frac{x - \mu}{\sigma}$ $\quad Z \sim N(0,1)$

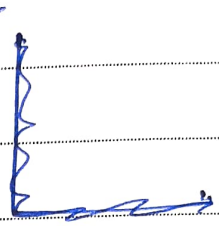


NOTES

| Original dataset | Subtract Mean ($x - \mu$) → divide by $\sigma^2$ |
|---|---|
| 1 | -2 → -1.63 |
| 2 | -1 → -0.82 |
| 2 | -1 → -0.82 |
| 3 | 0 → 0 |
| 3 | 0 → 0 |
| 3 | 0 → 0 |
| 4 | 1 → 0.82 |
| 4 | 1 → 0.82 |
| 5 | 2 → 1.63 |
| Mean = 3 | Mean = 0 → Mean = 0 |
| Std dev = 1.22 | $N \sim (0, 1)$ |



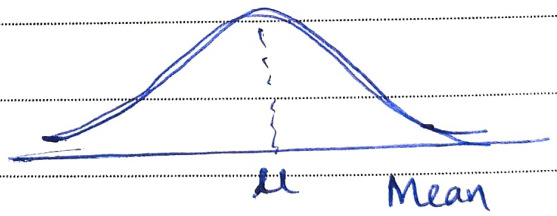
## * Central Limit Theorem:-

| | Mean |
|---|---|
| Sample 1 | 2600/- |
| Sample 2 | 3201/- |
| Sample 3 | 2844/- |

~~Mean~~ ~~see~~ different values we get for different samples.

Sampling Distribution of the Mean

We take some more samples and calculate means

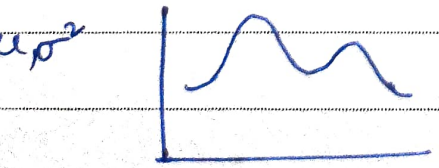


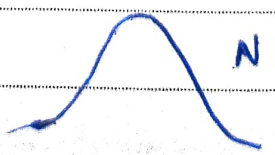

We are not calculating Population Mean.

We calculate Means of different samples in population. (Average of sample Means)

→ All together gives good result precisely.

Original Distribution → Sampling Distribution



$N\left(\mu, \frac{\sigma^2}{n}\right)$ Population Variance / sample size

The bigger the samples we get less variance

$n \geq 30$

My Baby NOTES

Central Limit ~~att~~ Theorem allows us to perform tests, solve problems and make inferences using Normal Distribution, even when the population is not Normally Distributed

Standard Error:

$$S.E. = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Stand. Deviation

As sample size ↑, Error ↓

— x — x — x — x —

↳ Kanpradeep

Estimators and Estimates

* Point Estimate :- Exact point of Estimate · Middle of Confidence Interval
* Confidence Estimate :- ☐ Range · Good for Inferences

* Point Estimators and Estimates

Estimator of Parameter

$\bar{x}$ of $\mu$

$s^2$ of $\sigma^2$

* Unbiased Estimator

$\bar{x}$ of $\bar{\mu}$

* Bias Estimator

$\bar{x} + 1$ ft of $\bar{\mu}$

* Efficiency :-

Most Efficient Estimator is Unbiased Estimator with Smallest Variance

* Estimators :- A type of Statistic

eg:- Avg. pizza cost

160/- 200/- 250/-

* 100% Confidence → We have to Consider Entire population But not possible

* $\alpha$ Confidence level

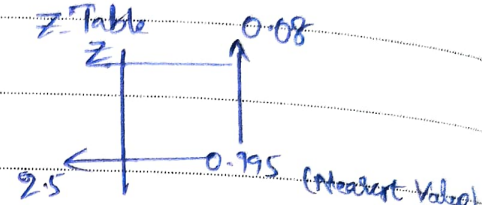
5% ← 95%

NOTES


[Point Estimate ± Reliability Factor * Standard Error , P.E = R.F * SE]

$$\left[\bar{x} - R.F * \frac{\sigma}{\sqrt{n}} \;,\; + \right]$$

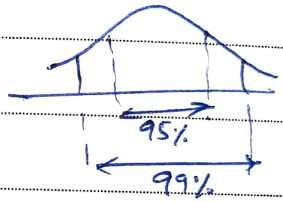
| 90% Confidence Interval | 99% |
|---|---|
| $\alpha = 0.1$ | $\alpha = 0.01$ |
| $\alpha/2 = 0.05$ | $1 - \alpha/2 = 1 - 0.005 = 0.995$ |
| $1 - \alpha/2 = 0.95$ | |




$$\left[\bar{x} - Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}\right] \qquad R.F = 2.5 + 0.08 = 2.58$$

[Sample Mean ± R.F * Deviation ; + ]
(R.F: From Z-Table.)

Confidence 95%, Narrower
[Less Value, Less Value]




eg:- Average of Salaries

Min.

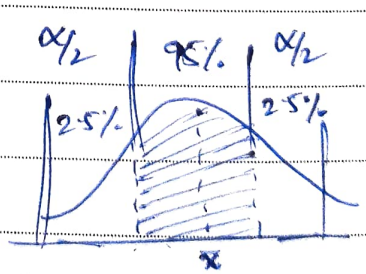
Max.

100% confidence [Min, Max]

Confidence Intervals:-

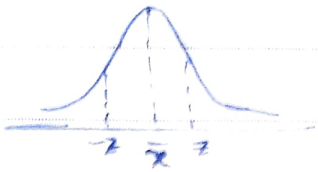
95% of Confidence.

We are 95% confident the true population Means falls within this Interval.




Majority of observations are around Mean.

Std Normal Distribution Mean = 0
Limit [-z, +z]

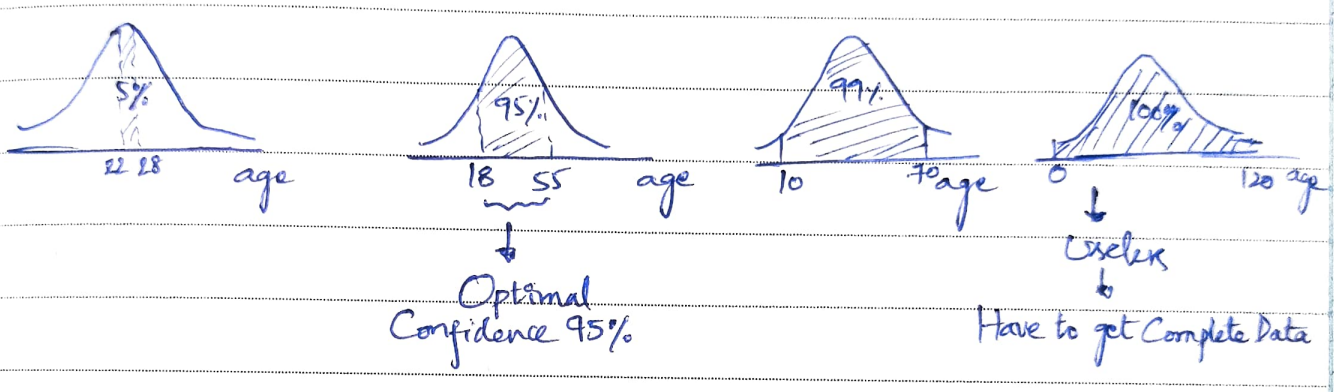



Confidence Interval

90% is smaller
$\Rightarrow (1-\alpha)$ smaller
$\Rightarrow$ Confidence Interval is Narrower

95% is Larger
Larger.
Broader

↳ Age of persons taking Statistics Course.




Optimal Confidence 95%

Useless

Have to get Complete Data

T-table:-

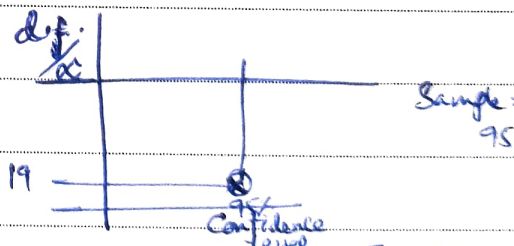
After 30 df (degree freedom) t-statistics ≃ z-statistics

* Student T-distribution:-

Inference through small samples.

Unknown population Variance

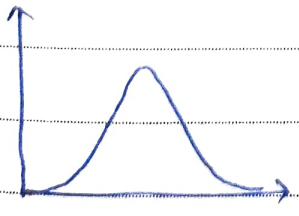
Huge Real Life Application.




$$t_{n-1,\alpha} = \frac{\bar{x} - \mu}{(S/\sqrt{n})}$$

n - sample size
n-1 freedom degree.
$\alpha$ significance level

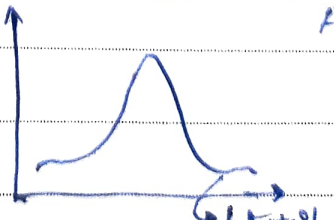
Normal distribution



z - statistic

Student T distribution

App. to Normal




t - statistic.

NOTES

Population Variance

Unknown → $\bar{x} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$

Known → $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

(Generally we don't Know Pop. Variance → So T-distribution is useful)

[Student T-table easy to analyse]

Eg: Data Scientist Sample

| S No | Salaries (Dataset) |
|---|---|
| 1 | : |
| 2 | : |
| 3 | → |
| 4 | → |
| 5 | → |
| 6 | → |

~~CI 95%~~ (

* $C.I_{95\%\ unknown} \cong t \equiv (LL, UL) \equiv$ Width More

* $C.I_{95\%\ Known} \cong Z \equiv (LL, UL) \equiv$ less

* population variance is unknown, our boundaries are more little less accurate.

Margin of Error:

M.E

$$\bar{x} \pm \left[ Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \right]$$

M.E

$$\bar{x} \pm \left[ t_{n-1,\alpha/2} * \frac{S}{\sqrt{n}} \right]$$

Conf. Int. = $\bar{x} \pm$ M.E

* M.E ↓ es ⇒ ~~Confidence~~ Level is less

* M.E ↓ es ⇒ $\sigma$ (Stdev.) ↓ es.

* M.E ↓ es ⇒ n ↑ es

↳ Higher no. of samples, we get near to True Mean.

## * Dependent

* Before & After situation
Cause and Effect

## Independent

* Population Variance Known
* Population Variance, Unknown, but assumed to be equal.
* Population Variance is Unknown, but assumed to be different

NOTES

## ① Calculating Confidential Intervals for 2 Means with Dependent Samples

eg: Weight loss - Blood samples of a Person are Dependent

* Person is same here

eg:- Sugar Test:

Before B Fast 90

After B Fast 160 → Why? Increasing?

Reason - Cause :- Had breakfast

What happened - Effect - Increased Sugar level in the body

Here We do testing on same Person.

Testing Approaches:-

1. Confidence Intervals for Dependent Samples
2. Statistical Methods like Regressions

| Patient | Before Drug taken | After Drug taken | Difference |
|---|---|---|---|
| 1 | 2 | 1.7 | -0.3 |
| 2 | | | |
| 3 | | | |
| . | | | |
| . | | | |
| 10 | 1.5 | 2.4 | 0.9 |

Mean diff =

Std. Dev = diff

Confidence interval

$$\bar{d} \pm t_{n-1,\alpha/2} * \frac{s_d}{\sqrt{n}}$$

for eg

Result : (0.01, 0.65)

* True Mean will fall in this interval
* Whole interval is +ve ⇒
* Level of drug is ↑ed (because interval)
⇒ Drug is Effective

* Population Var is Unknown
* <30 samples ⇒ We go 't' distribution.

## ② Confidence Interval for the difference of 2 Means Independent Samples, variance known:

eg:

| | Engg $x$ | M.B.A $y$ | Diff $x-y$ |
|---|---|---|---|
| Size | 100 | 70 | |
| Sample Mean | 58 pts | 65 pts | -7 |
| Pop. Std. dev | 10 pts | 5 pts | |

Considerations

1. Populations are Normally Distributed.
2. Population Variances are known
3. Sample sizes are different.

NOTES

* The grades getting in Engg. doesnot effect the grades of MBA Students
* the samples are truely Independent.

Considerations

↳ ~~Popula~~ samples are big. ⇒ Z Statistics
↳ Population Variance Known
↳ Populations are assumed to follow Normal Dist.

Variance of difference: $\sigma^2_{diff} = \frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}$

$= \frac{10^2}{100} + \frac{5^2}{70} = 1.36$

Mean diff = -7.

C.I. $(\bar{x} - \bar{y}) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$

$(\bar{x} - \bar{y})$ ↑ different point estimator

95% Confidence level
$(-9.28, -4.72)$

Inferences :- 1. True Mean falls in this interval we are 95% confident.
2. Whole interval is -ve ⇒ Engineers consistently gets lower grades.
3. We Calculate diff b/w Management − Engineers ⇒ $(4.72, 9.28)$

## ③ Confidence Interval for difference of 2 Means, Independent samples, Variance Unknown but assumed to be Equal

| Newyork Apples | | L.A. Apples |
|---|---|---|
| 1. | 3.8/- | 3.1/- |
| 2 | 3.4/- | 8/- |
| | 1/- | 6.5/- |
| | : | 3.2/- |
| | : | |
| 10. | 3.62/- | |

| | NY | L.A. |
|---|---|---|
| Mean | 3.9 | 3.25 |
| Std. Dev. | 0.18 | 0.27 |
| Sample Size | 10 | 8 |

We assume variance is same for both

Pop Variance Unknown
Small Samples → T-dist

Pool Variance $S_p^2 = \dfrac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{(n_x-1)+(n_y-1)}$

Pooled Variance $= \dfrac{(10-1)(0.18)^2 + (8-1)(0.27)^2}{(10-1)+(8-1)} = 0.05$

~~Pooled Std. Dev = $\sqrt{0.05}$ = 0.22~~

C.I. $(\bar{x}-\bar{y}) \pm t_{n_x+n_y-2,\,\alpha/2}\sqrt{\dfrac{S_p^2}{n_x} + \dfrac{S_p^2}{n_y}}$

↳ pooled std. dev.

C.I. (0.47, 0.92)

Inferences: Apples Cost N.Y. > Apples Cost in L.A. bcz diff C.I. Value is +ve

## ④ C.I. for diff. of 2 Means, Indep. Samples, Vari. Unknown but assumed to be different.

Apples | Oranges

C.I. $(\bar{x}-\bar{y}) \pm t_{v,\alpha/2}\sqrt{\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}}$

Here Main problem is Calculating Degree of Freedom

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1) + \left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

# Inferential Statistics

NOTES

Al Bundy Shoe Shop in US: Operating in UK, Canada, Germany.

1. Some shoes are no one buying.
2. Some shoes are more buying and failed to supply as per demand.

Here Comes Confidential Intervals:

Shop Last 20 yrs Data, We are about to Consider Last 3 yrs

Size Conversion Table

| Invoice No | Date | Country | Product ID | Shop | Gender (M/F) | Shoe Size (US) | Shoe size (UK) | Shoe size (Europe) | Unit Price | Discount |
|---|---|---|---|---|---|---|---|---|---|---|

Shoe Size Conversion Table

| Men | | | & | Women |
|---|---|---|---|---|
| UK | US | Euro | Inches | |

Inferences Regarding Dataset

1. Sample Data
2. Two Main groups Men group & Women (Almost Independent) - ~~different Sizes~~ ~~different~~ Models

Segment the data by 1. Shoe Size
2. Country
3. Gender

Men Group

| US Men Size | Canada | U.S | UK | Germany | Total |
|---|---|---|---|---|---|
| 6. | 15 | 59 | 6 | 30 | 105 |
| | | | | | |
| | | | | | |
| Total | | | | | |

Women Group

| US Women Group Size | Canada | US | UK | Germany | Total |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | | | | | |
| Total | | | | | |

Sum or Average of No.'s ⇒ We assume Normality ⇒ Center Limit Theorem

Game Plan:

95% of Confidence Interval
1. Last 12 months of Sales. (Last 5 yrs, Last 3 yrs)
2. Only for Men Shoes (Size 6, Size 6.5 ...)
3. Only for USA (All ~~Countries~~ Canada, U.K. ...)

U.S. 2016

| US Size | 1 Jan | 2 Feb | 3 Mar | ... | 12 Dec | Mean | Std Error | M.E | 95% CI 2016 | No of Pairs |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 4 | 1 | 3 | ... | 0 | Avg → 2.92 | | 12 | 1.8 to 4.04 | 4 |
| 6.5 | | | | | | | | | 0.46 to 2.88 | 3 |
| 9 | | | | | | | | | | |
| 9.5 | | | | | | → 30.33 | | t-dist | 25.00 to 35.67 | → 36 |
| 16 | | | | | | 0 | | 95% | 0.00 to 0.00 | 0 |
| Total | | | | | | | | | | |

$n = 12$

$t_{11, 0.025} = 2.2$

Shoe size of "6", We require 4 pairs Max.

Shoe Size of 9.5", We require 36 pairs

Shoe size of 16", Nothing to Manufacture.

## Another Application of Confidential Intervals

Two shops are selling the same no. of Shoes

95% Confidentiality of 2 Shops Sales in Germany:

| US Size | GER1 1 | 2 | 3 | ... | 12 months |
|---|---|---|---|---|---|
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |
| 11.5 | | | | | |

| US Size | GER2 1 | 2 | 3 | ... | 12 |
|---|---|---|---|---|---|
| 4 | | | | | |
| 5 | | | | | |
| 11.5 | | | | | |

Logically:
Same Years, Same People, won't buy same brand shoes in different shops
So, Two samples are Independent

NOTES

Same country, same People.

We assume population Variance of Sample 1 ≤ Pop Variance of Sample 2.

Here The 2 samples are Independent, Population Variance Unknown, but assumed to be equal.

| | Mean | | Sample Variance | | Pooled Variance | ME | 95% CI | |
|---|---|---|---|---|---|---|---|---|
| | Ger1 | Ger2 | Ger1 | Ger2 | | | | |
| | | | | | | | 0.00 | 0.00 |
| 4 | | | | | | | -0.23 | 0.9 |
| 4.5 | | | | | | | - | + |
| ⋮ | | | | | | | | |
| 11.5 | | | | | | | | |

95% confidence

$t_{12+12-2,\,0.025} = 2.07$

We can't Conclude
One shop has more no. of Pair shoes than Other on all sizes.
Reason Below

All confidence Intervals start on Negatives and finish on +ve

[L.L, U.L]

↓

'0' is included

(-0.23, 0.90)
Mostly +ve

(-1.11, 0.27)
Mostly Negative

In terms of Sales, Both shops are almost same.