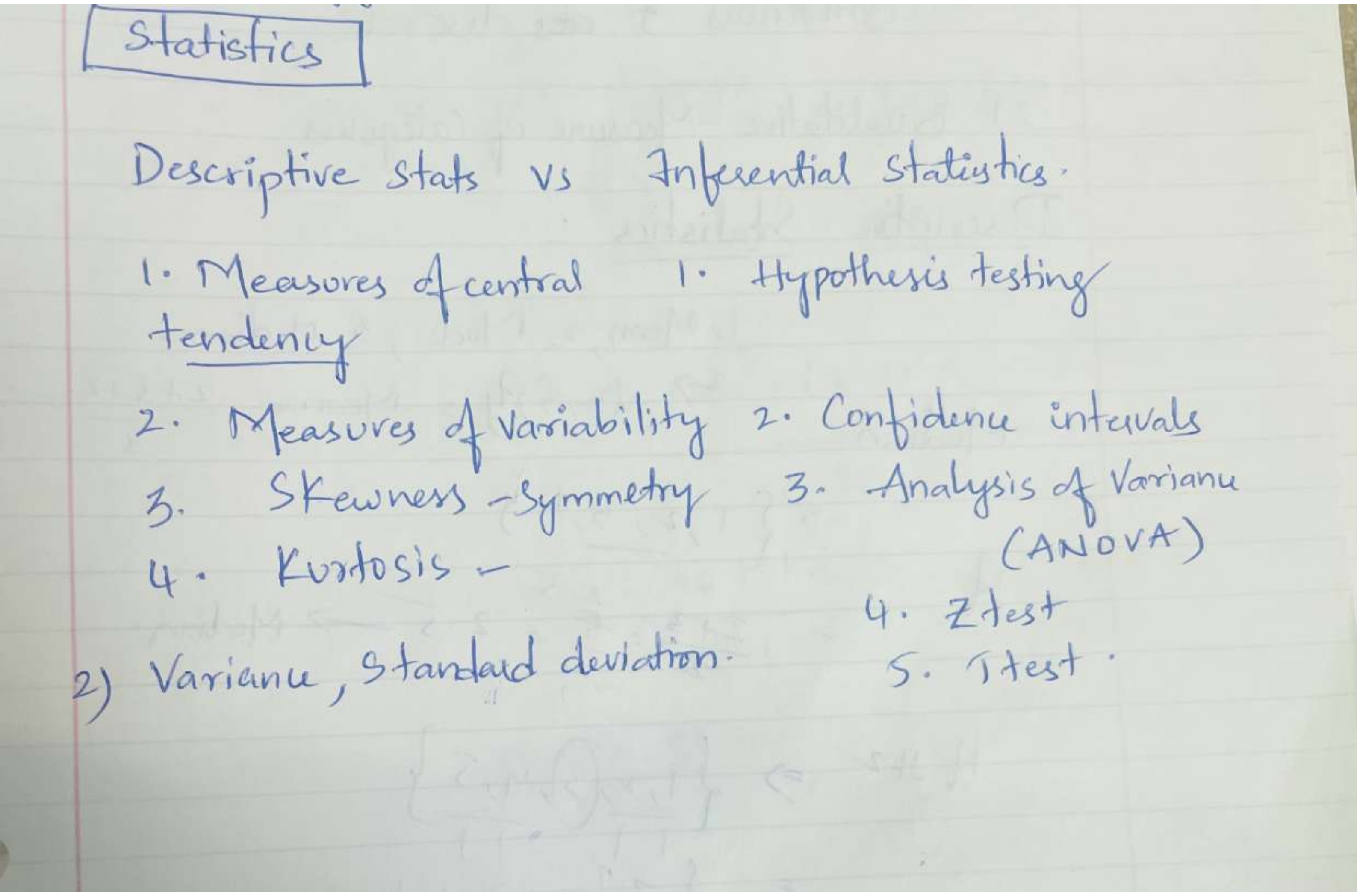


Statistics

Descriptive stats vs Inferential statistics.

1. Measures of central tendency
2. Measures of variability
3. Skewness - Symmetry
4. Kurtosis -

2) Variance, standard deviation.

1. Hypothesis testing
2. Confidence intervals
3. Analysis of Varianu (ANOVA)
4. Z test
5. T test.

Sampling techniques

- Simple Random Sampling.
- Stratified Sampling.
- cluster sampliug
- Systematic sampling.
- Convienience Sampling

Types of Variables :-

1) Quantities - Measure of Numeric

- Conti~~ti~~naus & ~~ats~~ discrete

2) Qualitative - Measure of Categories

Descriptive Statistics :-

1, Mean, 2, Mode, 3, Median,

$\hookrightarrow \{2,3,2\} = \text{Mean} = \frac{2+3+2}{3} = 2.8.$

Median

if $\Rightarrow \{1, \underline{2, 3}, 4\}$

$= \frac{2+3}{2} = \frac{\overset{2.5}{\cancel{5}}}{\underset{1}{\cancel{2}}} = 2.5 \rightarrow$ Median.

if #2 $\Rightarrow \{\underset{1\ 1}{\underline{1, 2}}, \boxed{3}, \underset{1\ 1}{\underline{4, 5}}\}$

$= \{3\} \rightarrow$ Median

Mode → Categorical data

Mode - Most repeated element

$\{a, a, b, b, b, b, c\}$

$= b$

* wheneves i have Categorical data we will use Mode.

* → wheneva we have out lies in data then we will go with Median.

- *Variance

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

↓ Mean of Sample

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

$\bar{x} = \frac{(1+2+3+4+5)}{5}$

$\bar{x} = 3$

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

$$= \frac{(1-3)^2 + (2-3)^2 + (3-0)^2 + (4-3)^2 + (5-3)^2}{5-1}$$

$$= \frac{4+1+0+1+4}{5-1} = \frac{10}{4} = 2.5$$

$V = 2.5$

Standard deviation !

$= \sqrt{V}$

Skewness :-

Kurtosis :- Measures the peakedness or flatness of a distribution. It provides info about the presence of outliers or Extreme Values in dataset.

* Normal distribution :-

- When the data is symmetric ~~to~~ on both sides then that distribution is called Normal distribution.

Standard Normal distribution !

if $m=0$; $SD=1$ then we can say that our
ta is Standard Normal distribution

Day - 16 - python:

1) why numpy?
2) why pandas
3)

* Measure of central tendencies
  ↳ Mean, Median, Mode.

+

Percentile

- At what particular percentile my value is present at?

$$\text{percentile} = \frac{(\#)\ \text{Number of Values below } x}{n} \times 100$$

Ex :- $n = \{1, 2, 5, \overset{\downarrow x}{6}, 10\}$

$$= \frac{\cancel{4}}{\cancel{0}} = \frac{\cancel{3}^{0.6}}{\cancel{5}_{1}} \times 100 = \frac{\cancel{6}}{1\cancel{0}} \times 10\cancel{0} = \underline{0.6}$$

percentile = 0.6%

- what if he gives ~~p~~percentile and we have to know the element.

* Ex :- 75% ?

$\{2, 5, 5, 7, 8, 9, 10\}$

$$= \frac{\cancel{75}^{3}}{\cancel{100}_{4}} \times \cancel{(8)}^{2} = \boxed{6}$$

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

6 is index of given element.

Q_3

$$\frac{\cancel{75}^{\,3}}{\cancel{100}_{\,\cancel{4}\,2}} \times \cancel{6}^{\,3} = \frac{9}{2}$$

$= 4.5$

$= 5$

* 5-Number Summary:

$\{1\ \underset{25\%}{\underline{5}}\ \underset{Med}{\underline{7}}\ 8\ 9\}$

1)- Mine - (1) minimum Value

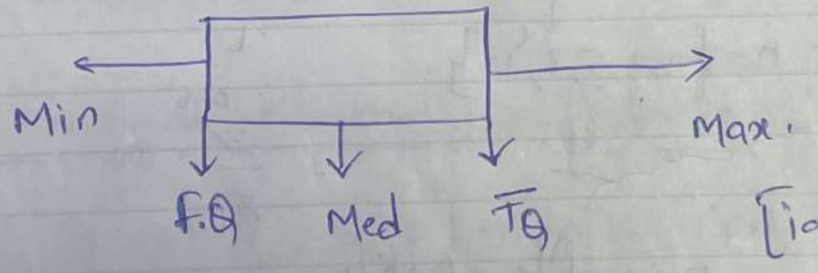
2) first quartile → 25% - $\textcircled{2}$ ⇒ 5

3) Median - 7

4) thirdquartile - 8.

5) Max - 9.

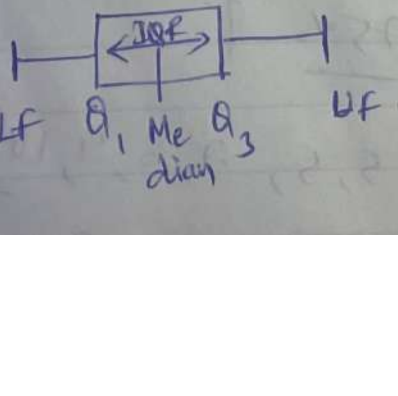
we will do our Box plot with this.



[iqr = Inter Quartile Range]

Lower fence ⇒ $Q_1 - 1.5 * iqr$

Upper fence ⇒ $Q_3 + 1.5 * iq\sigma$. [∵ $iqr = Q_3 - Q_1$]



Standard Normal Distribution.

$$Z = \frac{x_i - \mu\,(\text{Mean})}{\sigma\,(\text{S.D})}$$

↓

N.D → S.N.D

Day - 17 - python

Questions :- How to choose Datastructures

→ Lambda function

→ oops, class, object, Inheritance, polymorphism, Encapsulation

→ why numpy!

→ why panda?

* probability.

Emperical Rule

68	95	99.7
68% of data lies in 1st S.D Ive ,,	95% data lies in Second S.D	99.7 data lies in 3rd S.D.

Z Score

Probability - Measure of likelihood of an event

Ex :- Roll of Dice $\{1, 2, 3, 4, 5, 6\}$

$P(4) = 1/6$.

$P(1) = 1/6$.

Toss of a Coin $\{H, T\}$

$P(H) = 1/2$; $P(T) = \frac{1}{2}$

Additive Rule $P(A \text{ or } B)$

$\{H, T\} \Rightarrow \overset{\text{As we know}}{P(H)} \rightarrow 1/2$

$P(T) = 1/2$

what is $P(H \text{ or } T) \Rightarrow P(H) + P(T)$

$= \frac{1}{2} + \frac{1}{2}$

$P(H \text{ or } T) = 1$

Mutually Exclusive Events:

In a Dice.

-> Is it possible to get 1 at the same time when we get 6?

-> Answer to this question is $\overset{\text{No}}{\text{wrong}}$ ×

for Example we Can take Coin

- So What if we want to get H & T at Same time?

- Answer would be $\overset{\text{No}}{\text{wrong}}$ × again

Non-Mutually Exclusive Events

for Example

~~To~~ Deck of cards [52]

P(Q) & P(♡) i.e; P(Q & ♡) ?

is it possible to get both at Same time?

-> Answer to this question is yes ✓ we can

this type of events are called Non Mutual Exclus

events.

for mutual Exclusive .

what is $P(Q) = \frac{4}{52}$

and $P(♡) = \frac{13}{52}$

- $P(Q \text{ or } ♡) = \frac{4}{52} + \frac{13}{52}$

$= \frac{17}{52}$

and $P(Q \text{ and } ♡) = \frac{1}{52}$

Non-Mutual Exclusive :

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

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$

Conditional probability :-

Dependent event :-

permutation & combination

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

$n=6$; $r=3$
(creating a pair of 3 elements)

= {a, b, c, d, e, f}

↑

$$^{*}\,{}^{n}P_{r} = \frac{6\times5\times4\times\cancel{3\times2\times1}}{\cancel{3\times2\times1}}$$

$$= 120.$$

So we can have 120 permutations ~~from~~

Ex :- → abc,
cba
deb
acb
⋮

Combination :- The main difference b/w permutation & Combination is there is no chance of repetion like what we see in permutations.

= {a,b,c,d,e,f}

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

$n=6;\ r=3$

$$= \frac{\cancel{6}\times5\times4\times\cancel{3\times2\times1}}{\cancel{3\times2\times1}\,(\cancel{3}\times\cancel{2}\times1)}$$

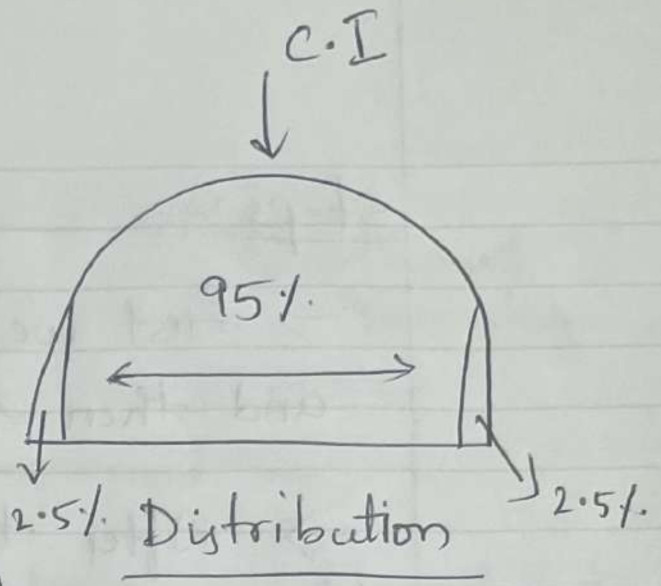
$$= 5\times4$$

$$= 20.$$

So we can get 20 Combinations

*** Infrential Statistics :-

- Drawing Conclusions or making predictions about larger dataset based on Sample data.



Hypothesis testing :-

→ Confidence Interval. (C.I)

→ Significance Value

→ Null hypothesis ⇒ H_0

→ Alternate Hypothesis ⇒ H_A

→ P-Value.

C.I = 95%

$= 1 - \frac{CI}{100}$

Singificance Value ⇒ 1-0.95

S.V (α) value = 0.05

Next checking the P value

if probability of touching at 2.5% is 0.01

and the α Value = 0.05

we will check $(P < \alpha)$ in our Case.

$0.01 < 0.05$

it Satisfies the Condition.

then we will reject our Null hypothesis

Steps :-

→ first we Create our Null Hypothesis and then Alternate Hypothesis.

→ After that we will define our Significance Value. ~~anp~~ and p value (after Conducting Some tests)

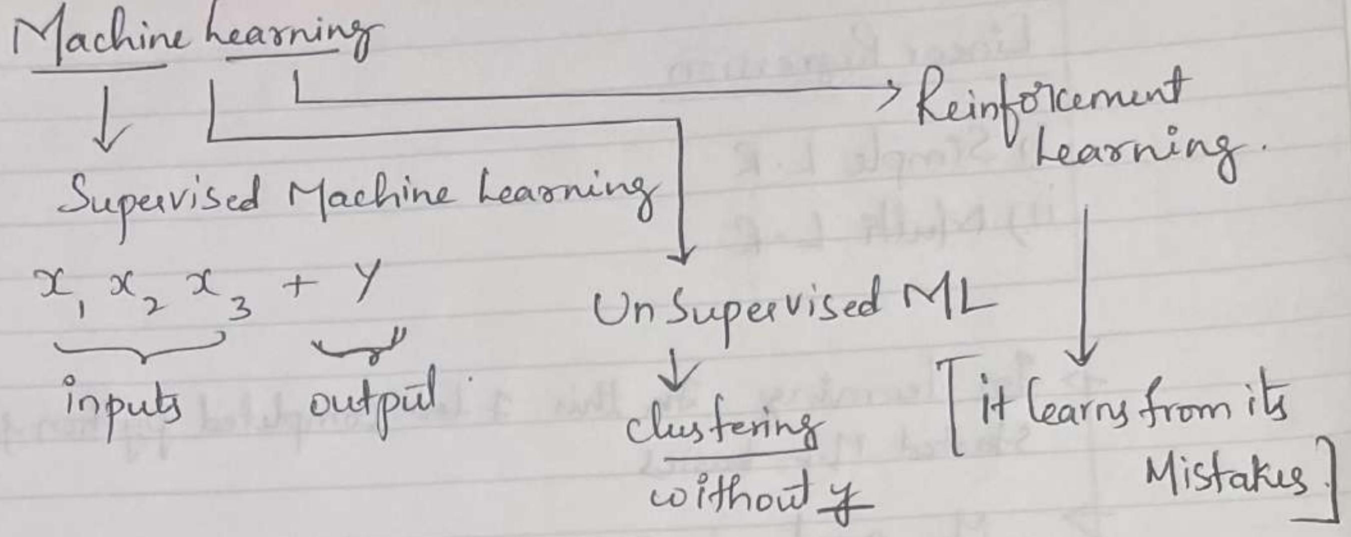
→ And Based on that p value we will decide which one to choose.

Different Kind of tests!

1) One Sample T-test
2) One Sample Z-test
3) One Sample proportion test.
4) Two Sample T-test
5) Two Sample Z-test
6) Two Sample proportion test.
7) paired-T test
8) Anova-test
9) Chi-Square test.
10) 1 tail & 2 tail test.

ML
Deeplearnin
NLP

Day-18 :-



Machine Learning

Supervised Machine Learning

$\underbrace{x_1\ x_2\ x_3}_{\text{inputs}} + \underbrace{y}_{\text{output}}$

Un Supervised ML

clustering

without y

Reinforcement Learning.

[it Learns from its Mistakes]

Supervised M.L :-

-> If my output ~~data~~ Variable is Continous then it is Regression. problem.

-> If my O/P Variable is discrete then it is ~~Continous~~ classificatio problem.

- we have two different Models i) Parametric models ii) Non parametric Mode

i) P.M's

a) Linear Regression -> R

b) Logistic Regression. -> C

ii) Non P.M's

a) Decesion tree (DT)

b) Random forest (RF)

c) Support Vector Machine (SVM)

d) Ada Boost, Extra Variant boost