Introduction to statistics

Def: statistics is the science of collecting, organizing and analyzing the data.

Eg: Hight of student in the class

Types of statistics

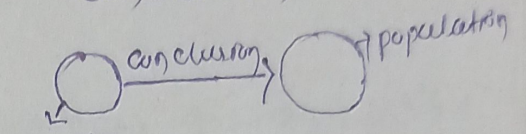
1. Descriptive stats

Def: It consists of organizing and summarizing the data.

- → Measure of central Tendency [Mean, Median, Mode]
- → Measure of dispersion [variance, std]
- → Histograms, Bar chart, piechart

2. Inferential stats

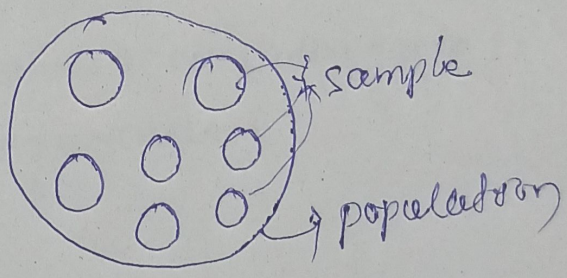
Def: It consists of using data you have measured to form conclusion.



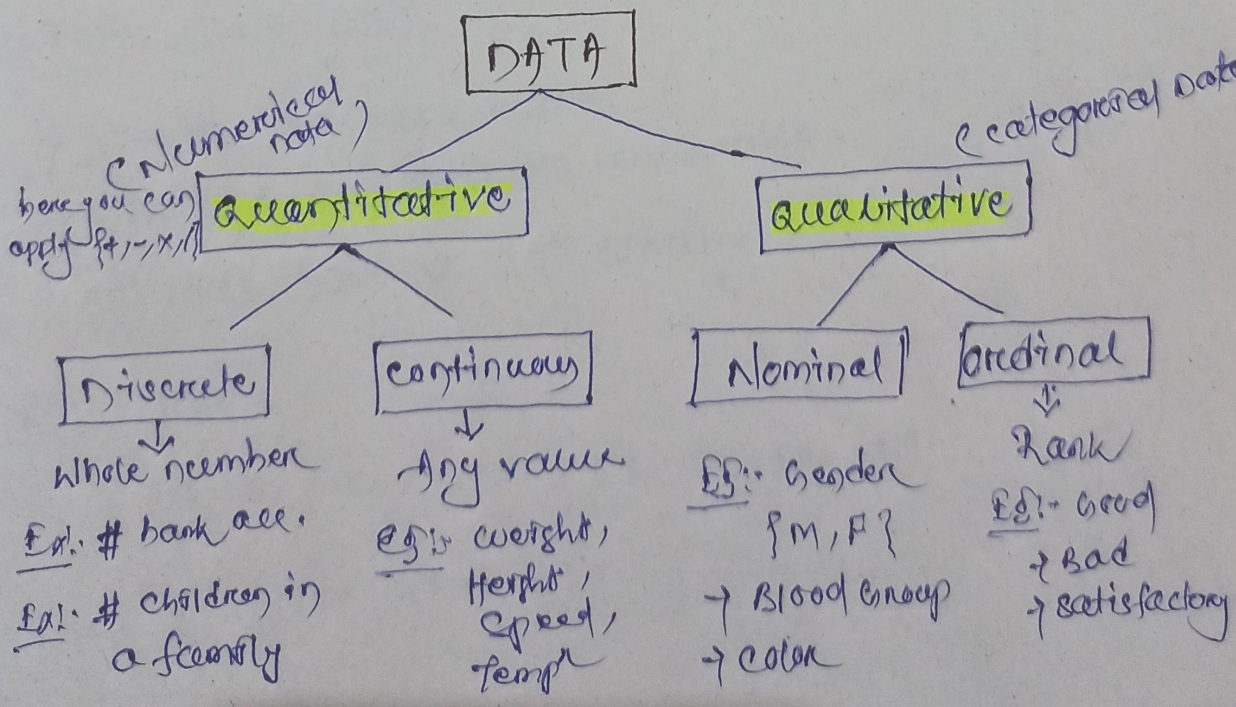
- → Z-test → t-test
- → Hypothesis testing, P value, Significance

sample data and population data

Eg: Exit poll



Types of data



Scale of measurement of data

1) Nominal scale data:

→ Qualitative / categorical data.

Eg:- Gender, colors, Blodgroup

→ order does not matter

2) ordinal scale data:

→ Ranking and order matter.

→ Difference cannot be measured.

Eg:- Qualification

PHd	I st
MS	2nd
B Tech	3rd
B com	4th

∴ here Ranking & order possible but difference cannot be measured.

3) Interval scale data

→ The rank and order matters.

→ Difference can be measured (excluding ratio).

→ Doesn't have "0" starting value.

Eg:- Temp^r

4) Ratio Scale Data

→ Order and rank matter.

→ Difference and ratio are measurable.

→ It does have a "0" starting

Eg:- mark

[0, 100, 90, 80, 75, 60, 50, 50]

100 : 50 ⇒ 2 : 1

Measures of central Tendency

1. Mean (μ) [Average]

Population (N)

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

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

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

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

Sample (n)

Sample mean $(s) = \sum_{i=1}^{n} \frac{x_i}{n}$

2. Median

Ex: 4, 5, 2, 3, 2, 1

here # element = even

sort → 1, 2, [2, 3], 4, 5

$\frac{(2+3)}{2} = 2.5$

median: 2.5

Ex: 1, 2, 2, [5], 4, 5, 7

element = odd

median = 3

3. Mode [maximum frequency]

{2, 1, 1, 1, 4, 5, 7, 8, 9, 10}

mode = 1

measures of central Tendency using Python

 # mean

```
age = [12, 13, 14, 21, 24]
import numpy as np
np.mean(age)
```

output

16.8

 # median

```
np.median(age)
```

O/P

14

 # mode

```
hight = [4.2, 5.0, 5.0, 5.2, 6.0]
np.mode(hight)
```

output

error

module 'numpy' has no attribute 'mode'

```
from scipy import stats
stats.mode(hight)
```

output

5.0

Measure of Dispersion

1. Variance : Spread of the data.

→ Population variance

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

where x_i = Data points
μ = population mean
N = population size

→ Sample variance

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

[Why does the sample variance have $n-1$ in the denominator?
Ans: to create unbiased estimator of the population variance.] → Bessil correction

where x_i = Data point
$\bar{x}$ = sample mean
n = sample size

Eg: $\{1, 2, 3, 4, 5\} \Rightarrow$ sample

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

$\bar{x}$ = sample mean = $\frac{1+2+3+4+5}{5} = 3$

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4

Now $s^2 = \frac{4+1+0+1+4}{5-1}$

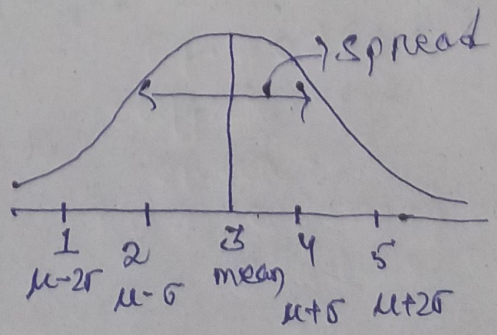
$= \frac{10}{4} = 2.5$

$\sqrt{s^2} = s$ = sample std deviation

Consider:

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

$\mu = 3$
$\sigma = 1$



measures of dispersion using python

variance & standard deviation

```
ages_list = [23, 43, 23, 56, 74, 82, 68, 98, 45, 32]
import numpy as np
mean = np.mean(ages_list)
mean :- o/p = 49.4
var = np.var(ages_list)
var :- o/p = 541.64
std = np.std(ages_list)
std :- 23.278169507331188
```

or

```
import pandas as pd
data = [[10, 12, 13], [34, 23, 45], [32, 34, 21]]
data :- o/p = [[10, 12, 13] [34, 23, 45], [32, 34, 21]]
df = pd.DataFrame(data, columns=["A", "B", "C"])
```

df :- o/p =

	A	B	C
0	10	12	13
1	34	23	45
2	32	34	21

column wise

df.var() :- o/p =

A	177.33333
B	121.0000
C	277.3333

Row wise

df.var(axis=1) :- o/p =

0	2.3333
1	121.0000
2	49.0000

Random Variable : Random variable is a process of mapping the output of a random process or experiment to a specific number.

Eg:- Tossing a coin
Rolling a dice

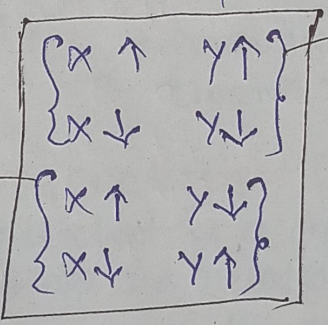
Covariance and correlation: covariance indicates the relationship of two variables whenever one variable changes, if an increase in one variable results in an increase in the other variable, both ~~variance~~ variables are said to have a positive covariance. Decrease in one variable also cause a decrease in the other, both variable said to have a +ve covariance. But in case of increase in one variable other is decrease then both said to have negative covariance.

Eg:

X	Y
2	3
4	5
6	7
8	9

∴ here X & Y are two variable

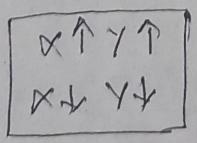
{Posible Relationship between X and Y can made?}



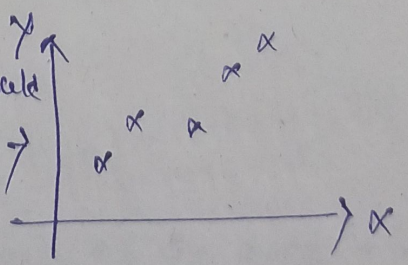
When X & Y both are increase or decrease then both variable said to have +ve covariance.

When among X and Y variable one of them increase other is decrease then both variable said to have -ve covariance.

When



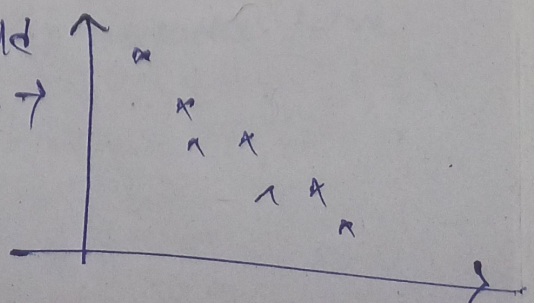
graph should look like →



When

X ↑ Y ↓
X ↓ Y ↑

graph should look like →



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

where x_i = data point
$\bar{x}$ = sample mean
n = sample size

Also we know that,

$var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \Rightarrow cov(x,x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$

[Spread of data]

Eg:-

x	y
2	3
4	5
6	7

$\bar{x} = 4$ $\bar{y} = 5$

$cov(x,y) = \sum_{i=1}^{n} \frac{(x - \bar{x})(y - \bar{y})}{n-1}$

$= \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{2}$

$= \frac{4 + 0 + 4}{2} = 4$ (+ve)

#Advantage

→ relationship betw x & y +ve or -ve value.

#Disadvantage

→ covariance does not specific limited value.

Pearson correlation coefficient [-1 to 1]

The pearson coefficient is a type of correlation coefficient that represents the relationship betw two variables that are measured on the same interval or ratio scale. The Pearson co-efficient is a measure of the strength of the association betw two continuous variables.

$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = [-1 \text{ to } 1]$

→ The more the value towards +1 the more (+ve correlated it is (x,y).

→ The more the value towards -1 the more (-ve correlated it is (x,y).

Spearman Rank correlation

$$r_s = \frac{cov(R(x), R(y))}{\sigma R(x) \times \sigma R(y)}$$

Ex:

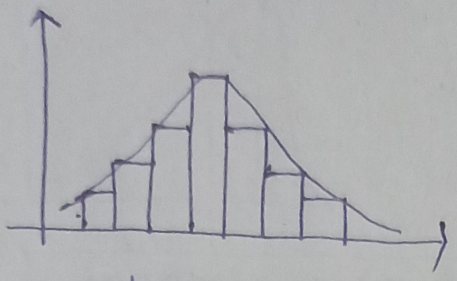
X	Y	R(x)	R(y)
1	2	4	5
3	4	3	4
5	6	2	3
7	8	1	1
0	7	5	2
		$\overline{R(x)}=3$	$\overline{R(y)}=3$

covariance And correlation with python

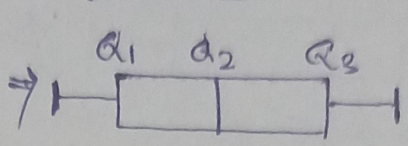
```
import seaborn as sns
df = sns.load_dataset("xyz")
# covariance
df.cov()
# or using numpy
import numpy as np
np.cov()
# correlation
df.corr(method = 'spearman')
df.corr(method = 'pearson')
```

Skewness: Skewness is a measure of how asymmetrical a distribution is about it's mean.

→ Symmetrical Distribution



Box plot



$Q_3 - Q_2 \approx Q_2 - Q_1$

⇒
- → Normal / Gaussian Distribution
- → Symmetric Distribution
- → No Skewness
- → The mean, median, mode all are perfectly at center.

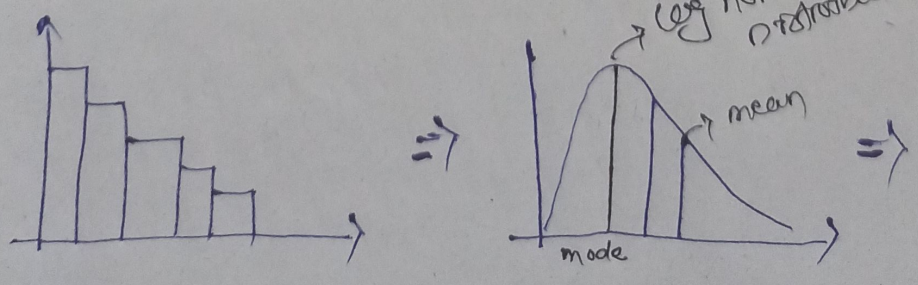
[mean = median = mode]

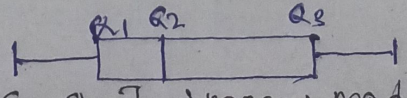
Def of Skewness: A distribution is said to be skewed when the mean and the median fall at different points in the distribution, and the balance is shifted to one side or the other - to left or right side

measures of Skewness tell us the distribution and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness.

When data points on a bell curve are not distributed symmetrically to the left and right sides of the median, the bell curve is skewed. distributions can be positive and right skewed, or negative and left skewed.

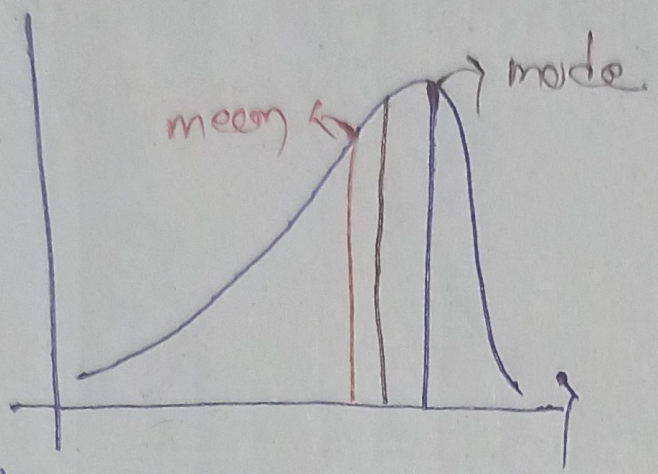
Right Skewed Distribution [positive skewed]

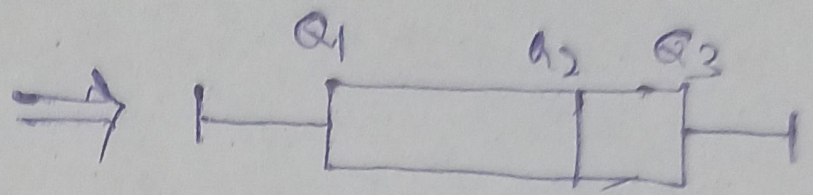




[$Q_3 - Q_2 > Q_2 - Q_1$] ⇒ mean > median > mode

Left skewed distribution





$Q_3 - Q_2 < Q_2 - Q_1$

* mean < median < mode

Histogram [Frequency]

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

Bins = 10

Bin size = $\frac{50 - 0}{10} = 5$