## Correlation

- It is a measure used to describe how strongly the given two random variables are related to each other.
- It is the estimated measure of covariance and is dimensionless.
- The value of correlation lies betwee -1 and +1
- It measures the direction and strength of the linear relatioship between the given two variables.
- Not sensitive to scale of the data.

Correlation, $P(x, Y) = cov(x, y) \sigma x \, \sigma y$

where:

- $P(x, y)$ = Correlation between the Variables $x$ and $y$
- $cov(x, y)$ = covariance between the variables $X$ and $Y$
- $\sigma X$ = Standard deviation of the $x$ varible
- $\sigma Y$ = Standard deviation of the $Y$ Varible

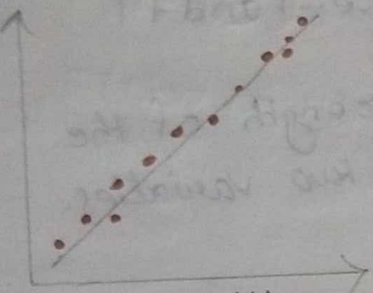
## Types of Correlation

- postive and Negative correlation
- Linear and nonlinear correlation
- Simple multiple correlation

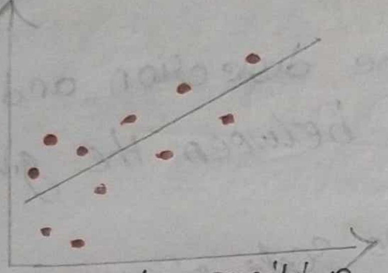
# Positive and negative correlation

- when two variables move in same direction i.e, when one increases other also increases or one decreases other also increases - postive correlation

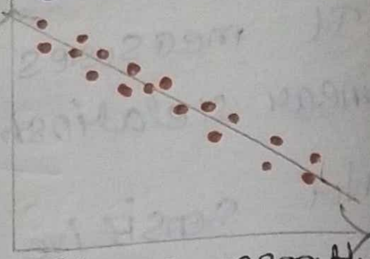
- when one increases, other decreases - Negative correlati



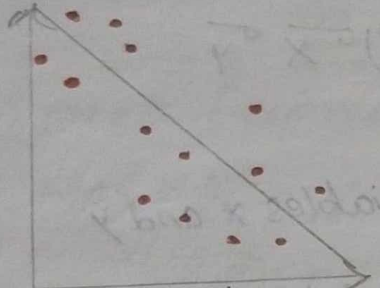
strong positive correlation



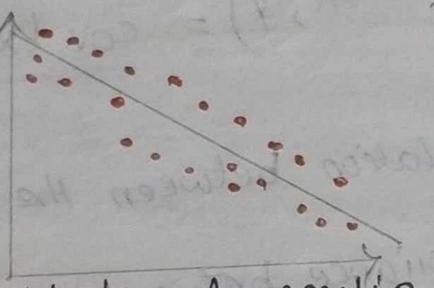
weak positive correlation



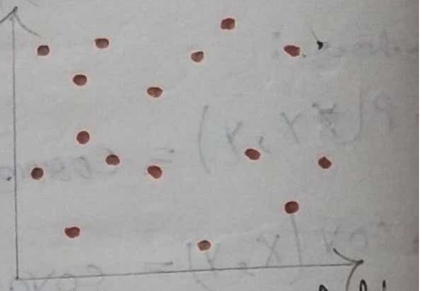
Strong negative correlatio



weak negative correlation



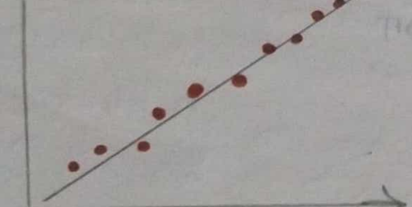
Moderate negative correlation



No correlation

## Linear and nonlinear Correlation

- Linear correlation - Amount of change in one varia tends to bear a constant ration to the amount of ch change in anothervarible. Then the graph will be a straight line

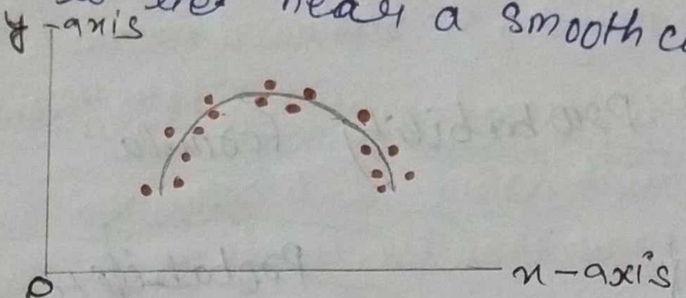


X:Y = constant ratio (throughout the data)

| X | Y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |
| 4 | 8 |

Here, X:Y = 1:2

- Non linear correlation- Amount of change in one varible does not bear a constant ration to the amount of change in another Variable.
- diagram tend to lies near a smooth curve



Simple multiple correlation

- Relationship between two variable-simple
- Relationship between three or more ~~varible~~ variable ~~multip~~ multiple.

```
import numpy as np
# Using seed function to generate the same random number every time with the same seed value
 np.random.seed(4)
# Create a random array of 500 integers between 0 and 50.
 x = np.random.randint(0, 50, 500)
# Create the second array using first array by adding some noise
 y 2x + np.random.normal(0, 10, 500)
 correlation = np.corrcoef(x, y)
#print the result.
 print ("The correlation between x and y is:
          \n", correlation)
```

Output

```
The correlation between x and y is;
[[1.          0.82477049]
  [0.82477049  1.        ]]
```

Condition Probability

- Condition probability is the probability of an event occuring given that another event has already occurred. The concept is one of the quintessential concept in probability theory. Note that conditional probability does not state that there is always a causal relationship between the two event, as well as it does not indicate that both event occur simultaneously. The concept of conditional probability is primarily related to the Bayes theorem which is one of the most influential theory in statistics

Conditional Probability formula

Probability of A and B

$$P(A|B) = \frac{P(A \cap B)}{\text{probability of } B}$$

probability of A given B

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B) \rightarrow \text{Probability of A and B}}{P(B) \rightarrow \text{Probability of B}}$$

where:

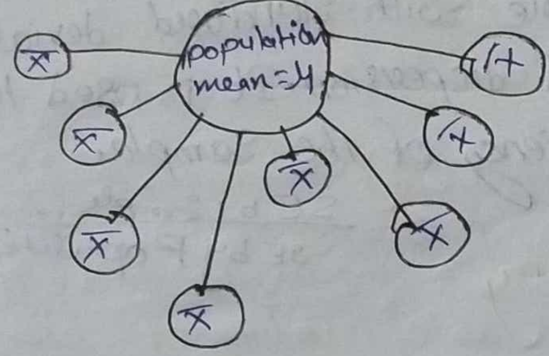
- P(A/B) – the conditional probability; the probability of even A occuring given that event B has already occurred.
- P(A∩B) – the joint probability of events A and B; the probability that both events A and B occur
- P(B) – the probability of event B

Sampling Distribution

- A researcher wished to estimate the birth weight of babies in a developing nation, a single sample of babies would provide the average birth weight of babies (sample mean)
- The estimate would likely to get more ~~as~~ accurate if a second sample was drawn and the mean calculated again
- Theoretically as more samples are drawn the estimate of the man for the population becomes more and more accurate as
- the total number of observations increases

# Sampling Distribution Of Mean

- If all the mean values calculated based on different samples were aggregated into a single data set the distribution of this
- set mean values would from a sampling distribution of mean.



- The sampling distribution can be developed for any specific statistic like mean, a proportion, the standard deviation etc.
- These distributions can be used to determin how likely it is that any specific measurement in the sample also appears in the population with the same relative frequency.
- If we knew the standard deviation of the sampling distribution, we could interpret the accuracy of the specific statistic used.

## Standard Error of Mean

- The standard error of the mean is a method used to evaluate the standard deviation of a sampling distribution. It is also called the standard deviation of the mean and is abbreviated as SEM
- The smaller the SE of the mean implies the lesser variability of sample mean from the population value
- he formula for standard error of the mean is equal to the ratio of the standard deviation to the root of ~~sample~~ population size.

$SEM = SD/\sqrt{N}$

where 'SD' is the standard deviation and N is the number of observation.

SEM for ~~Sample~~ = SD

## Standard Error of the proportion

- The standard error of the proportion is defined as the spread of the sample proportion about the population proportion
- More specifically, the standard error is the estimate of the standard deviation of a statistic.
- It has a similar nature with standard deviation, as both are the measures of dispersion. It is used to find the accuracy and efficiency of the sample.

$$\frac{\text{SE by Sample}}{\text{SE by Population}}$$