# Quantiles and Percentiles

Quantiles are statistical measure used to divide a set of numerical data into equal-size groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles are used in statistical analysis, including:

a. Quartiles: Divide the data into four equal parts, Q1 (25th percentile), Q2 (50th percentile), Q3 (75th percentile)

b. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile) --- D9 (90th percentile)

c. Percentile: Divide the data into 100 equal parts P1 (1st percentile), P2 (2nd percentile) ---

P99 (99th percentile).

d. Quintiles : Divide the data into 5 equal parts

Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation.
3. They are not actual values in the data.
4. All other tiles can be easily derived from percentile.

## Percentile

A Percentile is a statistical measure that represent the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observation in the dataset fall.

Formula to calculate the percentile value:

$$P_L = \frac{P}{100}(N+1)$$

Where:

- $P_L$ = the desired percentile value location
- $N$ = the total number of observations in the dataset.
- $P$ = the percentile rank (expressed as a percentage)

Example

Find the 75th percentile score from the below data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step 1 – sort the data (Asc)

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

$$PL = \frac{75}{100}(10+1) = \frac{3}{4} \times 11 = \frac{33}{4} = 8.25$$

96 — 98

8 — 9

Position

$$96 + 0.25(98-96) = 96.5$$

Percentile of a value

$$\text{Percentile rank} = \frac{x + 0.5y}{n}$$

x = number of values below the given value

y = number of values equal to the given values

N = total number of values in the dataset.

78, 82, 84, (88), 91, 93, 94, 96, 98, 99

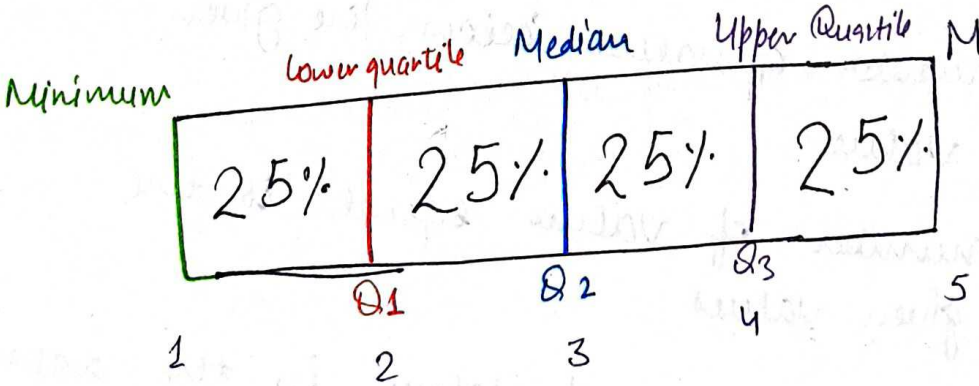
1 2 3

$$\frac{3 * 0.5(1)}{10} = \frac{3.5}{10} \times 100 = 35 \text{ percentile}$$

# 5 Number Summary

The five-number summary is a descriptive statistics that provides a summary of a dataset. It consist of five values that divide the dataset in four equal parts, also known as quartiles. The five-number summary includes the following values:

1. Minimum Value: The smallest value in the dataset
2. First quartile ($Q_1$): 25%.
3. Median ($Q_2$): 50%.
4. Third Quartile ($Q_3$): 75%.
5. Maximum Value: Largest values present in the dataset



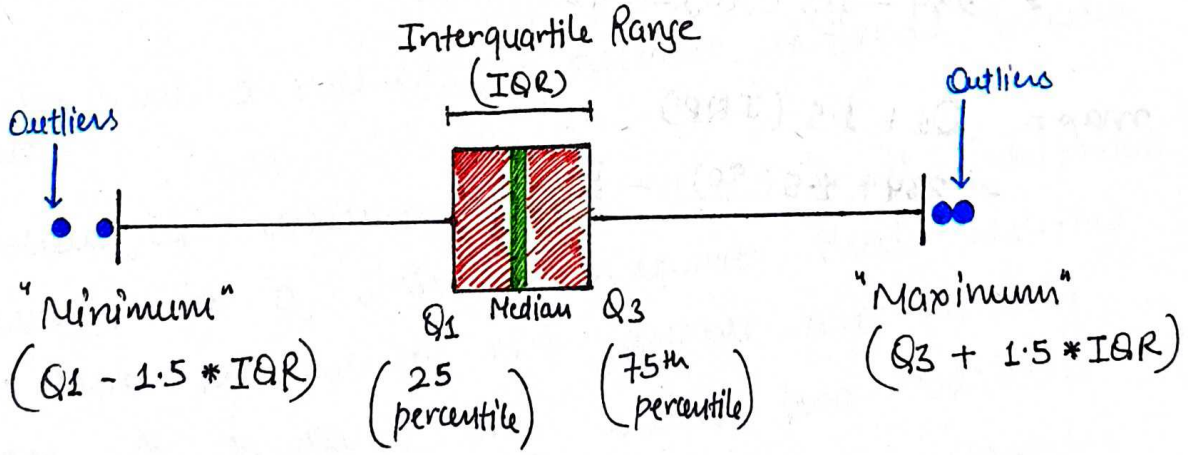

Interquartile Range is a measure of variability that is based on the five-number summary of a dataset. Specially, the IQR is defined

as the difference between the quartile ($Q_3$) and first quartile ($Q_1$) of a datasets.

# Boxplots

## 1. What is a boxplot

A boxplot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile ($Q_1$), the median ($Q_2$), and the third quartile ($Q_3$).




## 1. Benifit of a boxplot

→ Easy way to see the Distribution of data
→ Tells about skewness of data
→ can identify outliers
→ Compare 2 ~~assigned~~ catyon's of data.

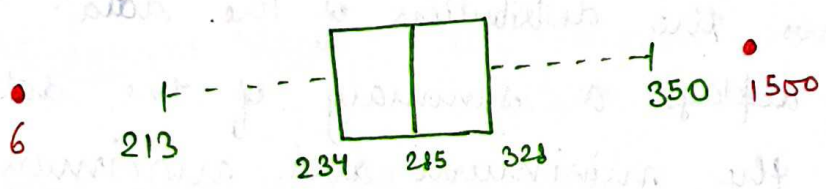
2. How to create boxplot with example.

| 6 | 213 | 241 | 260 | 281 | 290 | 314 | 321 | 350 | 1500 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

$Q_2 = \frac{50}{100} \times (11) = 5.5 = 285.5$

$Q1 = \frac{25}{100} \times (11) = 2.75 = 234$ $(213 + 0.75(241-213) = 234)$

$Q_3 = \frac{75}{100} \times 11 = 8.25 = 328.25$ $(321 + 0.25(350 - 321) = 328.25)$




min and max

$IQR = 328 - 234 = 90$

$min = Q1 - 1.5(IQR)$

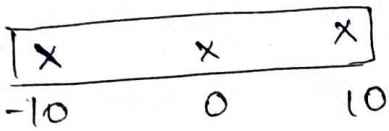
$= 234 - 1.5(90) = 93$

$max = Q1 + 1.5(IQR)$

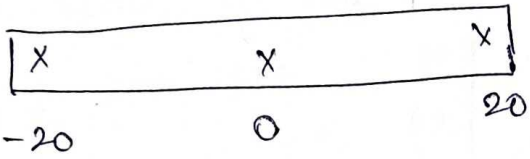
$= 234 + 1.5(90) = 469$

# Covariance

* what problem does covariance solve?



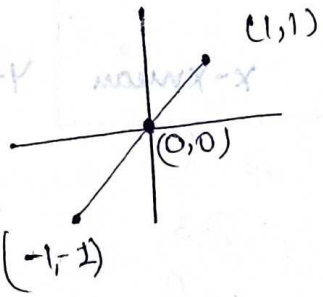

mean = 0



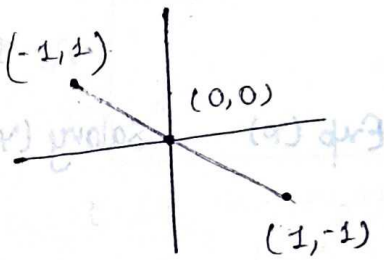

mean = 0

centre is same but spread is diff so we studied variance to solve this problem

Variance point of view data ~~spread~~ is same but this is wrong bcz points are diff.




$$\frac{1^2+0^2+1^2}{3} = \frac{2}{3}$$

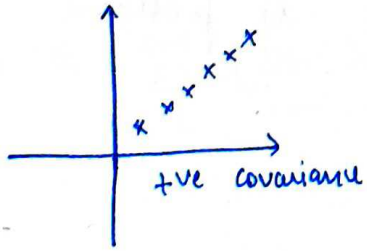


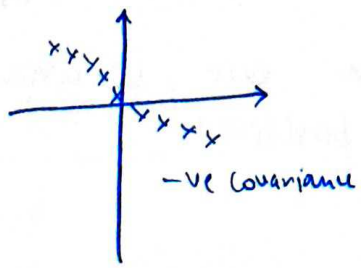

$$\frac{1^2+0^2+1^2}{3} = \frac{2}{3}$$

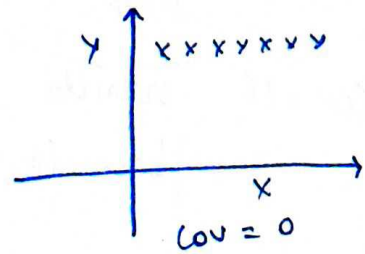
So we want a method to describe different between data.

* what is covariance and how is it interpret?

Covariance is a statistics measure that describes the degree to which two variable are linearly related. It measures how much two variables change together, such that when one variable increase, does the other variable also increase, or does it decrease?



* How is it calculated?

## Population

$$\sigma_{xy} = \frac{\sum (X - \mu_x)(Y - \mu_y)}{N}$$

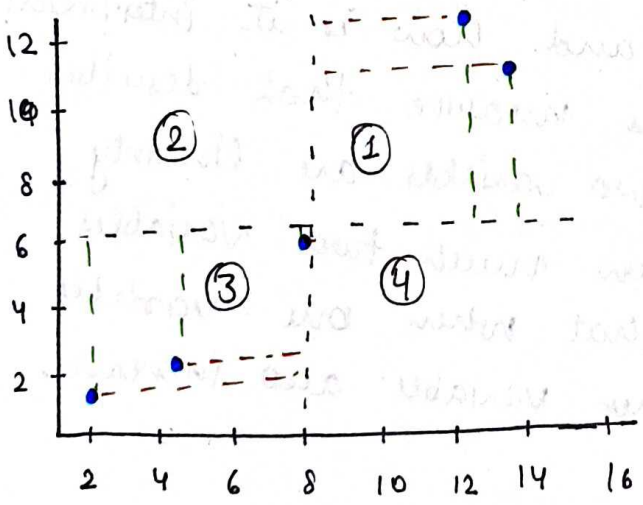
## Sample

$$S_{xy} = \frac{\sum (X - \bar{x})(Y - \bar{Y})}{n - 1}$$

| Exp (x) | Salary (Y) | x - x mean | Y - Y mean | (x-mean) * (Y-mean) |
|---|---|---|---|---|
| 2 | 1 | 6 | 5 | 30 |
| 5 | 2 | -3 | -4 | 12 |
| 8 | 5 | 0 | -1 | 0 |
| 12 | 12 | 4 | 6 | 24 |
| 13 | 10 | 5 | 4 | + 20 |
| | | | | 85 |

$\bar{X} = 8 \quad \bar{Y} = 6$

$\frac{80}{N=5} = 16$




① and ③ contain the positive points

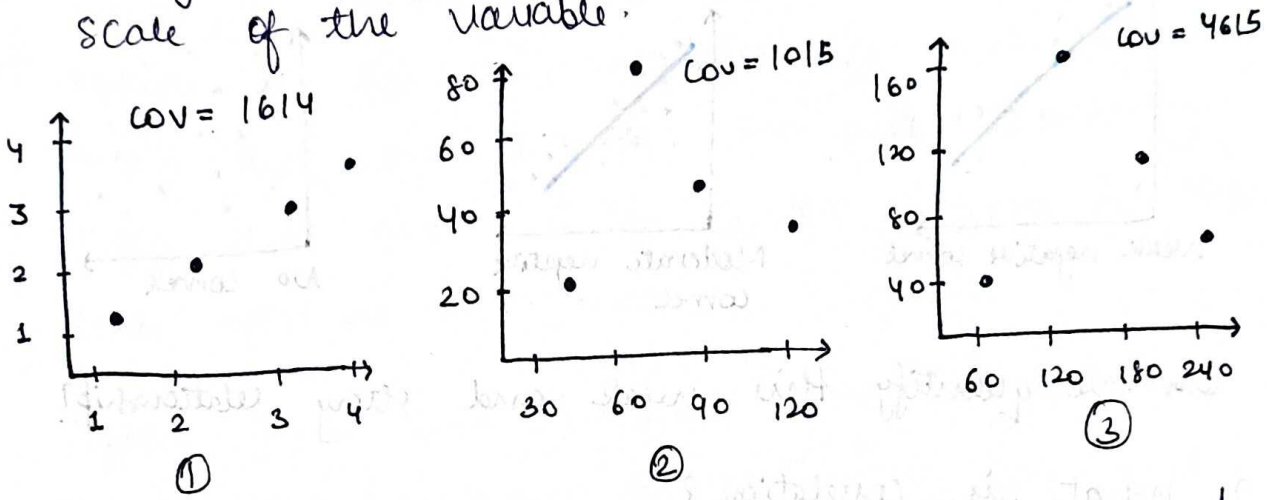
In ① box, both x and $y^{mean}$ are positive but

In ③ box both x and y mean may negative or not.

Cov = 16 which mean our covariance is positive bcz 16 is positive.

## * Disadvantage of using covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variable, since the magnitude of covariance is affected by the scale of the variable.




Graph ① is highly correlated and cov is 1614 and graph ② is correlated but not highly and cov is 1015 after increase scale (multiply by 2 x and y) of x and y of graph(3) and we can see graph ③ is same as graph ② (not highly correlated) but the cov is greater than graph ①. So covariance does not tell about strength of the relationship.

## * covariance of a variable of itself

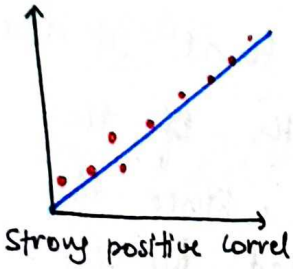
$$= \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

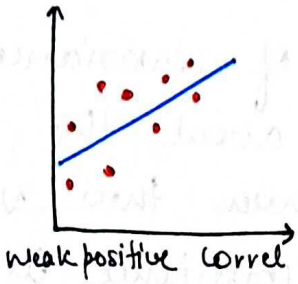
$$= \sum_{i=1}^{n} \frac{(x-\bar{x})(x-\bar{x})}{n-1}$$

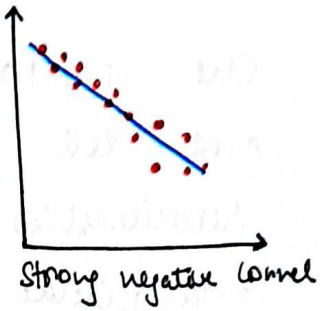
$$= \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$$

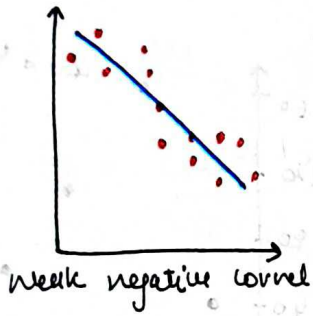
Variance formula

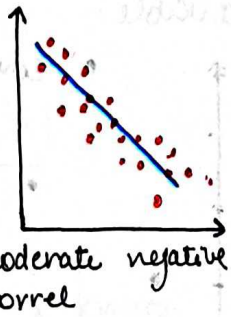
1. What problem does correlation solve?

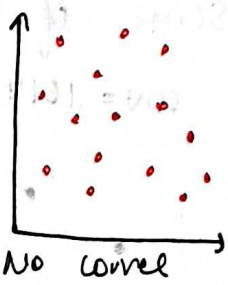



Can we quantify this weak and strong relationship?

2. What is correlation?

Correlation refers to a statistical relationship between two or more variables specifically, it measure the degree to which two variable are related and how they tend to change together.

Correlation is often measured using a statistical tool ^ (called) the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive correlation.

$$\text{correlation} = \frac{Cov(x, y)}{\sigma x * \sigma y}$$

# Correlation and Causation

The phrase "correlation does not imply causation" means that just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation betn two variable is the reason for the other variables does not necessarily imply that one variable does not neccesarily imply that one variable is the reason for the other variable's behaviour.

Suppose there is a positive correlation betn the number of firefighters present at a fire and the amount of damage caused by the fire. One might be tempted to conclude that the presence of firefighters cause more damage. However this correlation could be explained by a third variable - the severity of the fire. More severe fires might require more firefighter to be present, and also cause more damage.