

FORMULAE

1)

Range is the given measure of how spread apart the values in a dataset are.

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$$

2)

Quartiles divide a rank-ordered data set into four equal parts, denoted by Q1, Q2, and Q3, respectively

The interquartile range is equal to Q3 minus Q1, i.e.. $IQR = Q3 - Q1$

3) Variance:

Variance describes how much a random variable differs from its expected value.

It entails computing squares of deviations.

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

x : Individual data points

n : Total number of data points

\bar{x} : Mean of data points

4) Standard Deviation:

Deviation is the difference between each element from the mean.

$$\text{Deviation} = (x_i - \mu)$$

Population Variance is the average of squared deviations.

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

Sample Variance is the average of squared differences from the mean.

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

Sta

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

5)

The most common and effective numeric measure of the “center” of a set of data is the (*arithmetic*) *mean*. Let x_1, x_2, \dots, x_N be a set of N values or *observations*, such as for some numeric attribute X , like *salary*. The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

6)

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

This is called the **weighted arithmetic mean** or the **weighted average**.

7)

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.$$

The mean of a random vector X is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}.$$

The **variance-covariance matrix** Σ is defined to be

$$\mathbb{V}(X) = \begin{bmatrix} \mathbb{V}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_k) \\ \text{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \text{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_k, X_1) & \text{Cov}(X_k, X_2) & \cdots & \mathbb{V}(X_k) \end{bmatrix}$$

8) Covariance of 2 random variables:

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

9)

Quartiles Formula

Suppose, Q_3 is the upper quartile is the median of the upper half of the data set. Whereas, Q_1 is the lower quartile and median of the lower half of the data set. Q_2 is the median. Consider, we have n number of items in a data set. Then the quartiles are given by;

$Q_1 = [(n+1)/4]$ th item

$Q_2 = [(n+1)/2]$ th item

$Q_3 = [3(n+1)/4]$ th item

Hence, the formula for quartile can be given by;

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f} (l_2 - l_1)$$

Where, Q_r is the r^{th} quartile

l_1 is the lower limit

l_2 is the upper limit

f is the frequency

c is the cumulative frequency of the class preceding the quartile class.

10)

$$\text{Quartile deviation} = (Q_3 - Q_1)/2$$

IQR: Inter Quartile Range:

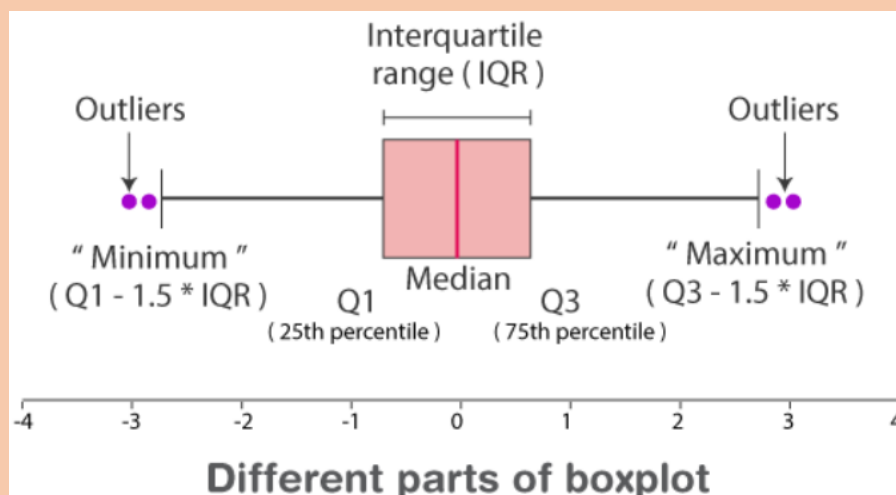
$$\text{IQR} = Q_3 - Q_1$$

11)

How to Find a Five-Number Summary: Steps

- **Step 1:** Put your numbers in ascending order (from smallest to largest). For this particular data set, the order is:
Example: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- **Step 2:** Find the minimum and maximum for your data set. Now that your numbers are in order, this should be easy to spot.
In the example in step 1, the minimum (the smallest number) is 1 and the maximum (the largest number) is 27.
- **Step 3:** Find the *median*. The median is the middle number. If you aren't sure how to find the median, see: [How to find the mean mode and median](#).
- **Step 4:** Place parentheses around the numbers *above and below* the median.
(This is not *technically* necessary, but it makes Q_1 and Q_3 easier to find).
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- **Step 5:** Find Q_1 and Q_3 . Q_1 can be thought of as a median in the lower half of the data, and Q_3 can be thought of as a median for the upper half of data.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- **Step 6:** Write down your summary found in the above steps.
minimum = 1, Q_1 = 5, median = 9, Q_3 = 18, and maximum = 27.

12)



13)

Correlation Formula

Correlation shows the relation between two variables. Correlation coefficient shows the measure of correlation. To compare two datasets, we use the correlation formulas.

Pearson Correlation Coefficient Formula

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data sets. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Where n = Quantity of Information

$\sum x$ = Total of the First Variable Value

$\sum y$ = Total of the Second Variable Value

$\sum xy$ = Sum of the Product of first & Second Value

$\sum x^2$ = Sum of the Squares of the First Value

$\sum y^2$ = Sum of the Squares of the Second Value

If given:

Market outcome	P(X,Y)	X	Y	X-E(X)	Y-E(Y)	(X-E(X))(Y-E(Y))
----------------	--------	---	---	--------	--------	------------------

13)

$$COV(x, y) = \sigma_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

14)

$$COV(x, y) = \sigma_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$VAR(x) = \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

15)

$$CORR(x, y) = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

16)

$$E(X) = \sum X[P(X)]$$

17)

$$COV(X, Y) = \sum (X - [E(X)])(Y - [E(Y)])P(X, Y)$$

18)

$$CORR(X, Y) = \frac{COV(X, Y)}{SD(X) \times SD(Y)}$$

19)

1. Calculate the correlation coefficient between the two variables x and y shown below:

X:	1	2	3	4	5	6
Y:	2	4	7	9	12	14

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

20)