FORMULAE

1)

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

$$Range = Max(x_i) - Min(x_i)$$

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.

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

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.

$$Deviation = (x_i - \mu)$$

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 $x_1, x_2, \ldots, x_N$ 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

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

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, Qr 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)

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

IQR: Inter Quartile Range:

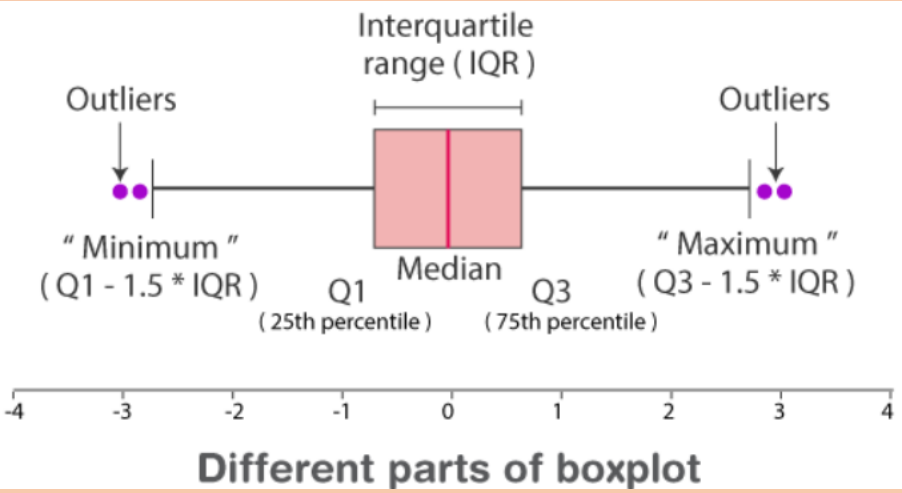
$$\textbf{IQR} = Q_3 - Q_1$$

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 Q1 and Q3 easier to find).
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- **Step 5:** *Find Q1 and Q3.* Q1 can be thought of as a median in the lower half of the data, and Q3 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, Q1 = 5, median = 9, Q3 = 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

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^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

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

20)

There are many methods for data normalization. We study *min-max normalization*, *z-score normalization*, and *normalization by decimal scaling*. For our discussion, let A be a numeric attribute with n observed values, $v_1, v_2, \ldots, v_n$.

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v'_i in the range $[new\_min_A, new\_max_A]$ by computing

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \quad (3.8)$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

In ***z*-score normalization** (or *zero-mean normalization*), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v'_i by computing

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}, \quad (3.9)$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. The mean and standard deviation were discussed in Section 2.2, where $\bar{A} = \frac{1}{n}(v_1 + v_2 + \cdots + v_n)$ and σ_A is computed as the square root of the variance of A (see Eq. (2.6)). This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.