**descriptive statistics**

**descriptive statistics**

**Q1: define the standard deviation and its importance :**

**Solution:**

Standard deviation is the measure of dispersion of a set of data from its mean. It measures the absolute variability of a distribution; the higher the dispersion or variability, the greater is the standard deviation and greater will be the magnitude of the deviation of the value from their mean.

Standard deviation is important because it tells us how spread out the values are in a given dataset.

**Q2:**

**The monthly salaries (in thousands of dollars) of a sample of 5 sales representatives are as follows: 10, 8, 9, 13, 10.**

**Find the standard deviation of this sample of monthly salaries**.

**Solution:**

Mean=(10+8+9+13+10)/5 = 10

S=√Σ((xi)\*\*2 – (n\*xbar)\*\*2) / (n-1)

S=√((514)-(500))/(4)

S=3.7416

**Q3:**

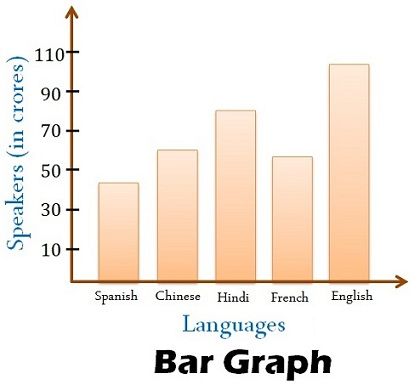
**explain this plots type [Bar charts, Scatter plots, histogram].**

**Solution:**

**Bar graph or bar chart:**

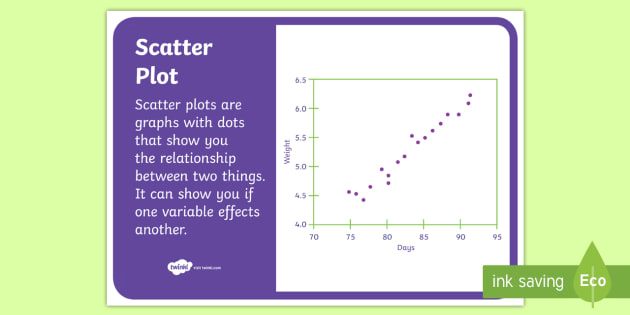
A bar graph is a chart that graphically represents the comparison between categories of data. It displays grouped data by way of parallel rectangular bars of equal width but varying the length. Each rectangular block indicates specific category and the length of the bars depends on the values they hold. The bars in a bar graph are presented in such a way that they do not touch each other, to indicate elements as separate entities.

Bar diagram can be horizontal or vertical, where a horizontal bar graph is used to display data varying over space whereas the vertical bar graph represents time series data. It contains two axis, where one axis represents the categories and the other axis shows the discrete values of the data. See figure given below:



**scatter plot:**

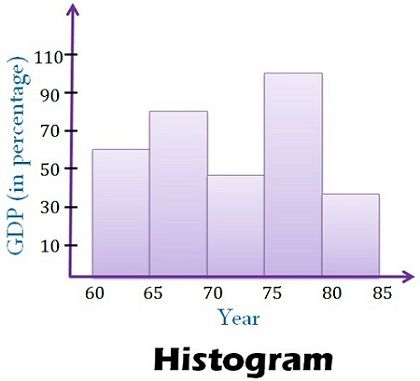
A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



**Histogram:**

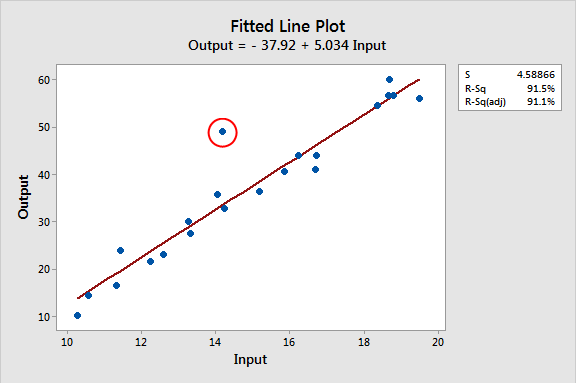
In statistics, Histogram is defined as a type of bar chart that is used to represent statistical information by way of bars to show the frequency distribution of continuous data. It indicates the number of observations which lie in-between the range of values, known as class or bin.

The first step, in the construction of histogram, is to take the observations and split them into logical series of intervals called bins. X-axis indicates, independent variables i.e. classes while the y-axis represents dependent variables i.e. occurrences. Rectangle blocks i.e. bars are depicted on the x-axis, whose area depends on the classes. See figure given below:



**Q4:**

**What does the red circle around the point mean? explain its theory.**

****

**Solution:**

**Outlier**

**Note**:

What are Outliers?

An outlier is an extreme element that is outside the distinctive pattern of a particular group or combination and affects the result of the analysis, but sometimes its presence is useful for some applications that depend on knowledge of different behavior such as fraud detection. From the image below, the green dots are next to each other, while the two red dots are far from them and are considered anomalies.

**Another note**:

The following are the most common reasons for outliers

1- Errors due to human reasons, such as entering wrong data

2- Errors due to technical reasons such as taking data from sensors

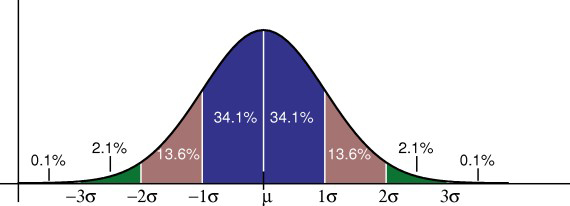
3. Intentional errors such as fictitious outliers for testing fraud detection methods

4- Errors as a result of sampling, such as extracting or mixing data from wrong or various sources

5- Natural errors such as novelties in the data (mutations).

**Q5:**

**Explain this plot** µ

****

**Class:** -3 σ -<-2 σ,-2 σ -<-1 σ, -1 σ -< µ , µ -<2 µ , 2µ -<4 µ

**Frequenc**y: 0.1%,2.1%,13.6%,34.1%,13.6%,2.1%,0.1%

***Shape*** : symmentrical

***Cennter:***

* Mean : (0.1+2.1+13.6+34.1+13.6+2.1+0.1)/7 =9.3857%
* Q2 =median =34.1%

**Spread**:

* range = max-min=34.1%-0.1%=34%
* Q1 =0.25(n+1)=0.25(7+1)=2 , Q1 = -2.1%
* Q3 =0.75(n+1)=0.75(7+1)=6 , Q3 = 2.1%
* IQR =Q3-Q1= 2.1-(-2.1)=4.2%
* V=Σ((xi)\*\*2 – (n\*xbar)\*\*2) / (n-1)

V=((1541.57)-(8139.67)/(6) = -1099.68333

* STANDARD DEVIATION =√variance =√1099.6833 =33.161