Ex!- Suppose we have the following values for salary (in Thousands of dollars), shown in increasing order:

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110, we have

$$\text{Mean } \bar{x} = 30+36+47+50+52+52+56+60+63+70+70+110 / 12$$

$$\Rightarrow \frac{656}{12} = 58$$

Thus, the mean salary is $58,000.

* Suppose that we had only the first 11 value in the list. Given an odd number of values, the Median is the middlemost value. This is the sixth value in this list, which has a value of 52,000$

* There is an even number of observations (i.e 12); Therefore, the Median is not unique; it can be any value with in the Two Middlemost value of 52 and 56 (with in the sixth and seventh value in the list).

By convention, we assign the average of the Two middlemost values as the Median; i.e. $\frac{52+56}{2} = \frac{108}{2} = 5$

this the Median is 54000$

* Mode can be determined for qualitative and quantative attributes. It is possible for greatest frequency to correspond to several different values, which results in more than one mode. Datasets with one, Two, or Three modes are respectively called Unimodal, bimodal, and trimodal. In general, dataset with Two or more modes is multimodal. At the other extreme, if each data value occurs only once, then there is no mode.

* The data in this example are bimodal. The Two modes are $52,000 and $70,000.

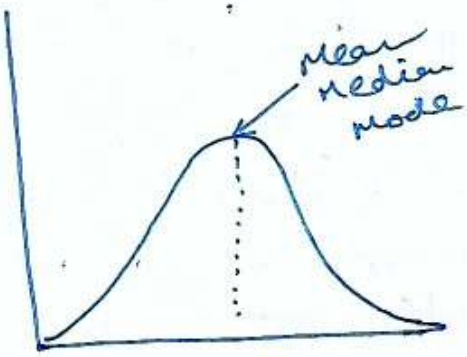
* Midrange of the data of same example is

$$\frac{30,000+110,000}{2}$$

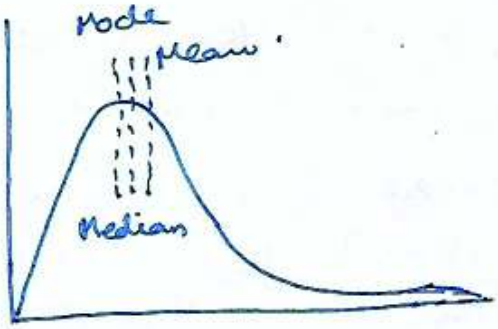
$$= \$70,000$$

in a unimodal frequency curve with Perfect symmetric data distribution, The mean, median, and mode are all at The same center value, As shown in fig (a)

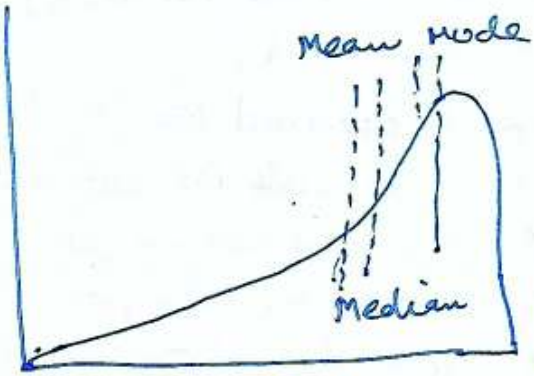
Data in most real applications are not symmetric. They may instead be either Positive skewed, where the mode occurs at a value that is ~~either~~ smaller than the median, or negatively skewed, where the mode occurs at a value greater than the median.



(a) Symmetric data



(b) Positively skewed data

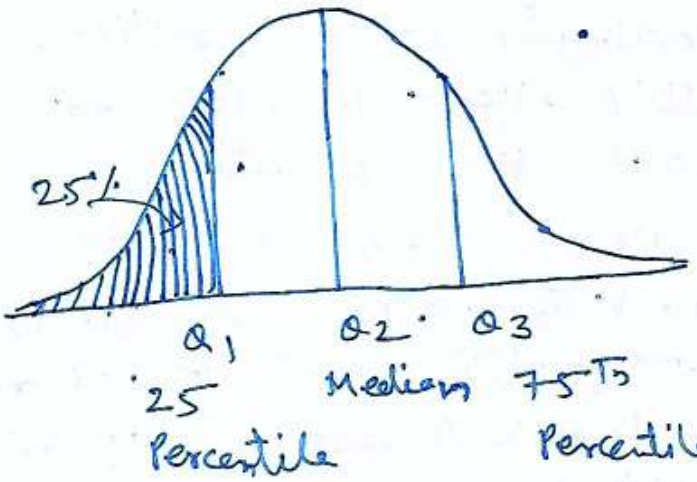


(c) Negatively skewed data

* Suppose that the data for attribute X are sorted in increasing numeric order. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive set, as in fig. These data points are called quantiles. Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.

The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.

The 100-quantiles are most commonly referred to as percentiles; they divide the data distribution into 100 equal-sized consecutive sets. The median, quartiles, and percentiles are the most widely used forms of quantiles



A quartiles give an indication of a distribution's center, spread and ~~center~~ shape. The first quartile, denoted by Q_1, is the 25 percentile. it cuts off the lowest 25% of the data. The third quartile, denoted by 75Th Percentile - it cuts off the lowest 75% (or highest 25%) of the data The second quartile is the 50Th percentile. As the median, it

* The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the IQR and is defined as

$$IQR = Q_3 - Q_1$$

* is our data values; $Q_1 = \$47000$ and $Q_3 = \$63000$
Thus the interquartiles range is $IQR = 63-47 = 16000\$$ and median is $52000\$$.

Five number summary, Box Plots, and outliers :→

In the symmetric distribution, the median (and other M/S of central) splits the data into equal-size halves. This does not occur for skewed distributions.

Therefore, it is more informative to also provide the two quartiles Q_1 and Q_3, along with median. A common rule of thumb for identifying suspected outliers is to single out values falling atleast $1.5 \times IQR$ above the third quartile or below the first quartile.

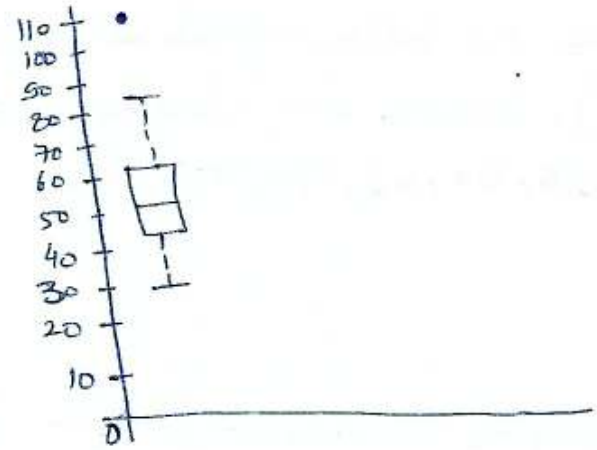
Because Q_1, the median (Q_2), and Q_3 together contain no information about the end points (eg tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the five number summary. The five number summary of a distribution consists of the median (Q_2), the quartiles Q_1 and Q_3, and the smallest and largest individual observations, written in the order of Minimum, Q_1, median, Q_3, maximum.

Box plot are a popular way of visualizing a distribution. A box plot incorporates the five number summa as follows:

* typically the ends of the box are the quartiles so that the box length is the interquartile range.

* The median is marked by a line within the box.

* Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.



Variance and standard deviation: → Variance and standard deviation are measure of data dispersion.

They indicate how spread out a data distribution.

A low standard deviation means that the data observations tend to be very close to mean, while a high standard deviation indicates that the data are spread out over a large ~~number~~ range of values.

The variance of N observations, $x_1, x_2 \cdots x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \left(\frac{1}{N-1} \sum_{i=1}^{N} x_i^2 \right) - \bar{x}^2$$

The standard deviation, σ, of the observations is the square root of the variance, σ^2.

Properties of σ :-

σ measure spread about the mean and should be considered only when the mean is chosen as the measure of center.

$\sigma = 0$ only when there is no spread, that is, when all observations have the same value, otherwise $\sigma > 0$

Example:- Suppose we have the following value for salary (in thousands of dollars), shown in increasing order

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

Sol:- we have $\bar{x} = 58$

$$\sigma^2 = \frac{1}{12-1}(30^2 + 36^2 + 47^2 + \cdots + 110^2) - 58^2$$

$$= \frac{1}{11}(44918) - 3364$$

$$\Rightarrow 4083.45 - 3364$$

$$\Rightarrow 719.454$$

$$\boxed{\sigma = 26.82}$$ Ans.