Introduction to Descriptive Statistics

1 Measures of dispersion

Knowing the "typical value" of the data alone is not enough. We also need to know how "concentrated" or "spread out" it is. That is, we need to know something about the "variability" of the data. Measures of dispersion are a way of quantifying this idea numerically. it is degree of scatter or variation of individual value of a variable about the central value such as the median or the mean. These include range, mean absolute deviation, interquartile range, variance, standard deviation, coefficient of variation and etc.

1.1 The range

This is the simplest method of measuring dispersion. It is the difference between the largest and the smallest value in a set of data. It is commonly used in statistical quality control. However, the range may fail to discriminate if the distributions are of different types. So, for our ranked data, we have

Range $= x_{(n)} - x_{(1)}$.

1.2 Quartiles and the interquartile range

Whereas the median has half of the data less than it, the *lower(or first) quartile* has a *quarter* of the data less than it, and the *upper(or third) quartile* has a quarter of the data above it. So the lower quartile is calculated as the $(n+1)/4^{th}$ smallest observation, and the upper quartile is calculated as the $3(n+1)/4^{th}$ smallest observation. Again, if this is not an integer, *linearly interpolate* between adjacent observations as necessary (examples below). There is no particularly compelling reason why $(n+1)/4$ is used to define the position of the lower quartile — $(n+2)/4$ and $(n+3)/4$ seem just as reasonable.

Example. Calculating lower quartiles

n	LQ at	LQ is
15	$(15+1)/4 = 4$	$x_{(4)}$
16	$(16+1)/4 = 4\frac{1}{4}$	$\frac{3}{4}x_{(4)} + \frac{1}{4}x_{(5)}$
17	$(17+1)/4 = 4\frac{1}{2}$	$\frac{1}{2}x_{(4)} + \frac{1}{2}x_{(5)}$
18	$(18+1)/4 = 4\frac{3}{4}$	$\frac{1}{4}x_{(4)} + \frac{3}{4}x_{(5)}$
19	$(19+1)/4 = 5$	$x_{(4)}$

The *inter-quartile range* is the difference between the upper and lower quartiles, that is

$$\text{IQR} = \text{UQ - LQ}.$$

It measures the range of the middle 50% of the data. It is an alternative measure of spread to the standard deviation. It is of interest because it is much more robust than the standard deviation, and thus is often used to describe asymmetric distributions.

1.3 Mean absolute deviation (M.A.D.)

This is the average absolute deviation from the mean $\overline{x}$.

M.A.D. $= \frac{|x_1-\overline{x}|+\cdots+|x_n-\overline{x}|}{n} = \frac{\sum||x_i-\overline{x}||}{n}$

1.4 Variance and standard deviation

The population variance,σ^2 is given by

$$\sigma^2 = \frac{1}{n}\sum (x_i\text{-}\overline{x})^2 .$$

It is the average squared distance of the observations from their mean value. The *population standard deviation*,σ , is just the square root of the variance. It is preferred as a summary measure as it is in the units of the original data. However, it is often easier from a theoretical perspective to work with variances. Thus the two measures are complimentary.

Advantage

1. It is well defined and uses all observations in the distribution.
2. It has wider application in other statistical technique like skewness, correlation, and quality control e.t.c

Disadvantage

1. It cannot be used for computing the dispersion of two or more distributions given in different unit.

1.5 Coefficient of variation

This is a dimension less quantity that measures the relative variation between two servers observed in different units. A measure of spread that can be of interest is known as the *coefficient of variation*. This is the ratio of the standard deviation to the mean,

$CV = \frac{\sigma}{\bar{x}}$

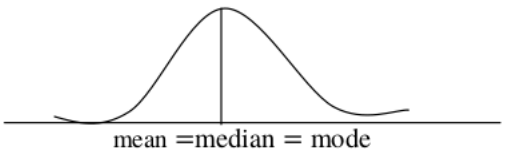
and thus has no units. The coefficient of variation does not change if the (linear) *scale*, but not the *location* of the data is changed. That is, if you take data $x_1, ..., x_n$ and transform it to new data, $y_1, ..., y_n$ using the mapping $y_i = \alpha x_i + \beta$, the coefficient of variation of $y_1, ..., y_n$ will be the same as the coefficient of variation of $x_1, ..., xn$ if $\beta = 0$ and $\alpha > 0$, but not otherwise. So, the coefficient of variation would be the same for a set of length measurements whether they were measured in centimeters or inches (zero is the same on both scales). However, the coefficient of variation would be different for a set of temperature measurements made in Celsius and Fahrenheit (as the zero of the two scales is different).The distribution with smaller C.V is said to be better.

2 Skewness and Kurtosis

Skewness means "lack of symmetry". We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. If in a distribution mean in median = mode, then that distribution is known as symmetrical distribution. If in a distribution, mean≠median≠mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negative skewed.

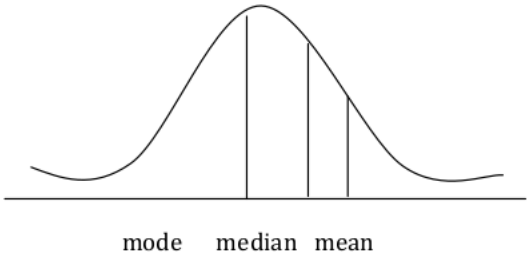
2.1 Skewness

(a) Symmetrical distribution:



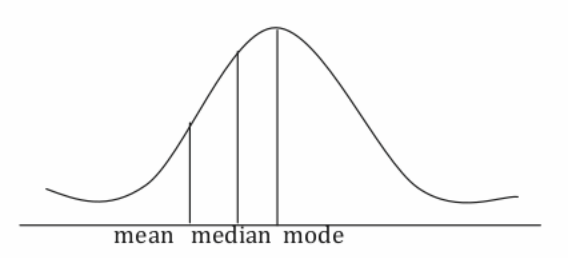
Mean, median and mode coincide and the spread of the frequencies is the same on both sides of the center point of the curve.

(b) Positively Skewed distribution:



In a positive skewed distribution, the value of the mean is the maximum and that of the mode is the least, and the median lies in between the two. The frequencies are spread out over a grater range of values on the right hand side than they are on the left hand side.

(c) Negatively Skewed distribution:



In a negatively skewed distribution, the value of the mode is the maximum and that of the mean is the least. The median lies in between the two. The frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

The important measures of skewness are:

1. Karl-Pearson's coefficient of skewness
2. Bowley's coefficient of skewness
3. Measures of skewness based on moments.

2.1.1 Karl-Pearson's coefficient of skewness

This is given by:
Karl-Pearson's Coefficient Skewness= $\frac{Mean-Mode}{standard-deviation}$.

In case of mode is ill-defined, the coefficient can be determined by the formula:

$$\text{Coefficient of Skewness}= \frac{3(Mean-Median)}{standard-deviation} .$$

2.1.2 Bowley's coefficient of skewness.

This is given by:
Bowley's Coefficient of Skewness= $\frac{Q_3+Q_1-2Median}{Q_3-Q_1}$.
Where Q_1, Q_3 are first and third quartiles respectively.

2.1.3 Measure of skewness based on moment

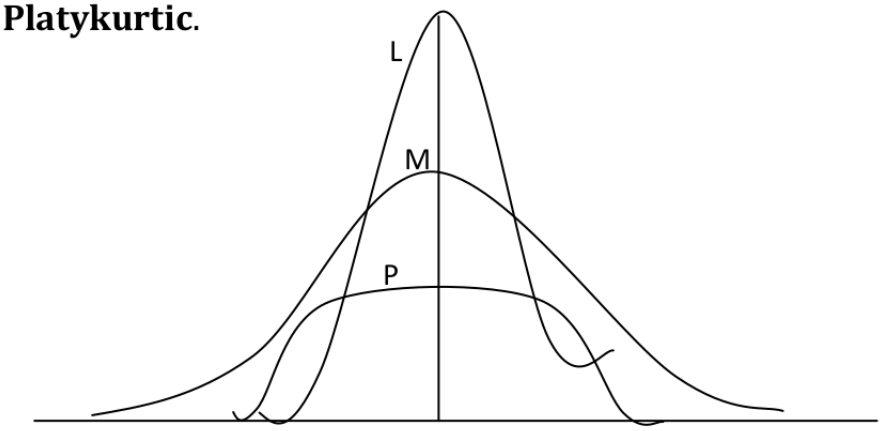
First, we note the following moments:
1^{st} moment about mean: $\mu_1 = \frac{\sum(x-\overline{x})}{n} = 0$
2^{nd} moment about mean: $\mu_2 = \frac{\sum(x-\overline{x})^2}{n} = \sigma^2$
3^{rd} moment about mean: $\mu_3 = \frac{\sum(x-\overline{x})^3}{n}$
4^{th} moment about mean: $\mu_4 = \frac{\sum(x-\overline{x})^4}{n}$
Then the measure of skewness based on moments denoted by$\beta_1$is given by:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

2.2 Kurtosis

The expression 'Kurtosis' is used to describe the peakedness of a normal curve. The three measures – central tendency, dispersion and skewness, describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution. Measures of kurtosis tell us the extent to which a distribution is more picked or more flat topped than the normal curve, which is symmetrical end bell – shaped, is designated as *Mesokurtic*. If a curve is relatively more narrow and peaked at the top,

it is designated as *Leptokurtic*. If the frequency curve is more flat than normal curve, it is designated as *Platykurtic*.



Measures of Kurtosis

The measure of kurtosis of a frequency distribution based on moment is denoted by$\beta_2$and is given by:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

- If$\beta_2 = 3$, the distribution is said to be normal and the curve is Mesokurtic.
- If$\beta_2 > 3$, the distribution is said to be more peaked and the curve is Leptokurtic.
- If$\beta_2 < 3$, the distribution is said to be flat peaked and the curve is Platykurtic.