Measures of Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

1.1 Frequency and Distribution

- A **histogram** is one way to depict a **frequency distribution**
- **Frequency** is the number of times a variable takes on a particular value
- **Relative frequencies** are particularly useful if you want to compare distributions drawn from two different sources (i.e. while the numbers of observations of each source may be different)
- We may summarize our data by constructing **histograms**, which are vertical bar graphs
- A **histogram** is used to **graphically** summarize the distribution of a data set
- A histogram divides the range of values in a data set into **intervals**
- Over each interval is placed a bar whose height represents the **frequency** of data values in the interval.

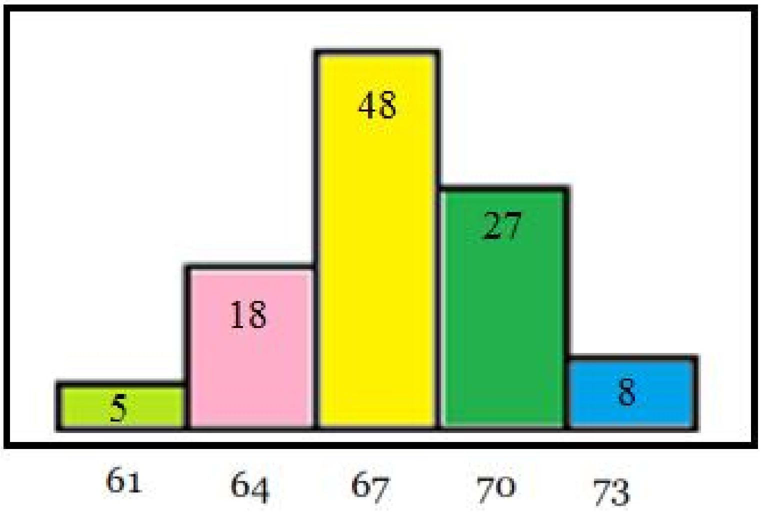
1.2 Building a Histogram

- To construct a **histogram**, the data are first **grouped** into categories
- The histogram contains one **vertical bar** for each category
- The **height** of the bar represents the number of observations in the category (i.e., **frequency**)
- It is common to note the **midpoint** of the category on the horizontal axis

<u>Example</u>

Here are grouped data for heights of 100 randomly selected male students

Height (inches)	Class Mark, x	Frequ-ency, f
59.5–62.5	61	5
62.5–65.5	64	18
65.5–68.5	67	42
68.5–71.5	70	27
71.5–74.5	73	8



The first thing you usually notice about a distribution's shape is whether it has one mode (peak) or more than one. If it's **unimodal** (has just one peak), like most data sets, the next thing you notice is whether it's **symmetric or skewed** to one side.

If the bulk of the data is at the left and the right tail is longer, we say that the distribution is **skewed right or positively skewed**; if the peak is toward the right and the left tail is longer, we say that the distribution is **skewed left or negatively skewed.**

1.3 Types of Distribution

1.3.1 Symmetric Distribution

a distribution is symmetric if it can be folded along the vertical axis so that the two sides coincide

if the distribution is symmetric, the mean, the median, and the mode are equal and are located at the same position along the horizontal axis.

1.3.2 Skewed Distribution

- if the two sides do not coincide, distribution is said to be asymmetric
- a distribution that is asymmetric with respect to a vertical axis is said to be skewed.

1.3.2.1 Two Types of Skewness

A- Positively Skewed or Skewed to the Right Distribution

* distribution tapers more to the right than to the left.

* has a longer tail to the right compared to a much shorter left tail.

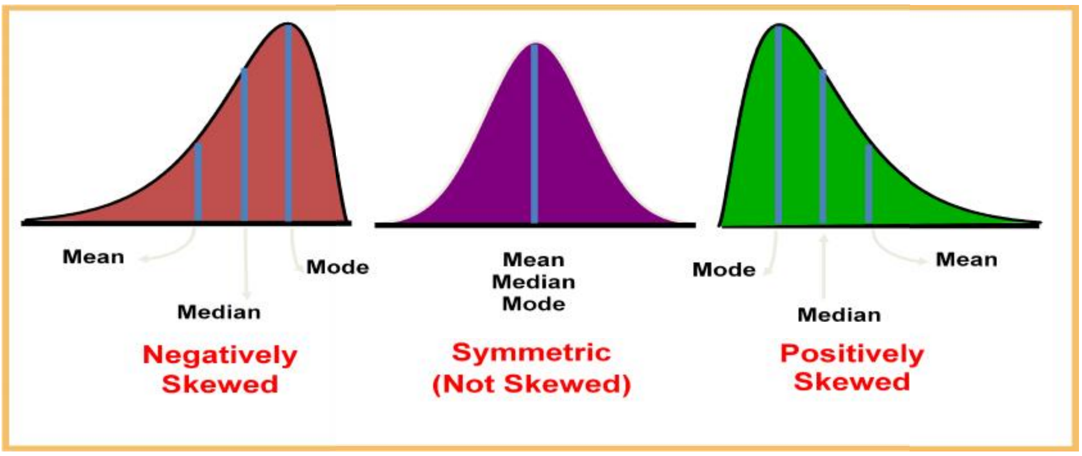
* values are more concentrated below than above the mean.

* large values in the right tail are not offset by correspondingly low values in the left tail and consequently the mean will be greater than the median.

B- Negatively Skewed or Skewed to the Left Distribution

* distribution tapers more to the left than to the right

* has longer tail to the left compared to a much shorter right tail
* values are more concentrated above than below the mean
* small values in the left tail will make the mean less than the median



When the skewness is presented in absolute term i.e, in units, it is absolute skewness. If the value of skewness is obtained in ratios or percentages, it is called relative or coefficient of skewness. When skewness is measured in absolute terms, we can compare one distribution with the other if the units of measurement are same. When it is presented in ratios or percentages, comparison become easy. Relative measures of skewness is also called coefficient of skewness.

1.4 Measure of skewness

- Shows the degree of asymmetry, or departure from symmetry of a distribution.
- Indicates also the direction of the distribution.

1.4.1 Nature of Skewness

Skewness can be **positive** or **negative** or **zero**.

When the values of mean, median and mode are equal, there is no skewness.

- When mean > median > mode, skewness will be positive.
- When mean < median < mode, skewness will be negative.

Mathematical measures of skewness can be calculated by :

1.4.2 Pearson's First and Second Coefficient of Skewness

1- **Karl Pearson's** coefficient of skewness is defined in terms of mode as

$$SK = \frac{\bar{X} - Mo}{s} \quad Where\ \ X = mean, Mo = mode,\ s = standard\ deviation$$

2- **Karl Pearson's** coefficient of skewness is defined in terms of median as

$$SK = \frac{3(\bar{X} - Md)}{s} \quad where\ \bar{X} = mean\ , Md = median, s = standard\ deviation$$

Since the mode is frequently only an approximation, formula 2 is preferred.

Interpretation of the measure of skewness

$$SK > 0 : positively\ skewed\ since\ \ \bar{X} > Md > Mo$$

$$SK < 0 : negatively\ skewed\ since\ \ \bar{X} < Md < Mo$$

Example: Given the following statistics on weights of court justices in kilograms:

$$\bar{X} = 74.1\ ,\ Md = 75\ \ ,\ Mo = 84\ \ , s = 11.25$$

Using the first formula

$$Sk = \frac{74.1 - 84}{11.25} = -0.88 \quad (skewed\ to\ the\ left)$$

Using the second formula,

$$Sk = \frac{3(74.1 - 75)}{11.25} = -0.24 \quad (skewed\ to\ the\ left)$$

The skewness of a random variable , x_2, ...,x_N, the formula for skewness is:

$$Sk = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 / N}{s^3}$$

where ***x*** is the mean, ***s*** is the standard deviation, and ***N*** is the number of data points.

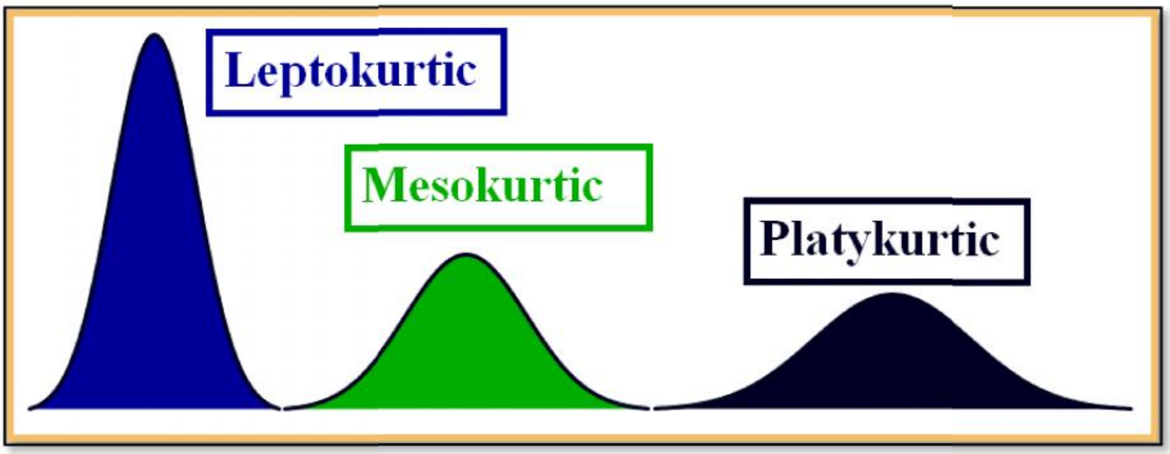
The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness. Many software programs actually compute the adjusted Fisher-Pearson coefficient of skewness.

1.5 Measure of kurtosis

If a distribution is symmetric, the next question is about the central peak: is it high and sharp, or short and broad? You can get some idea of this from the histogram, but a numerical measure is more precise. The height and sharpness of the peak relative to the rest of the data are measured by a number called kurtosis. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the excess kurtosis is presented: excess kurtosis is simply kurtosis−3.

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called **mesokurtic**.
- A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**.
- A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**.
    - **Leptokurtic**: high and thin
    - **Mesokurtic**: normal in shape
    - **Platykurtic**: flat and spread out



Mathematical measures of kurtosis can be calculated by :

$$kurtosis = \frac{\sum_{i}^{n} (x_i - \bar{x})^4}{ns^4}$$

Or

$$kurtosis = \frac{\sum_{i}^{n} (x_i - \bar{x})^4}{ns^4} - 3$$

Interpretation of *Kur*

***Kurtosis* < 0, platykurtic**

***Kurtosis* = 0, mesokurtic**

***Kurtosis* > 0, leptokurtic**