Measurement Scales and Data Types

statsdirect.com/help/basics/measurement_scales.htm

It is important, in statistical analysis, to know about the different scales of measurement, these are:

- **INTERVAL**
  Scale with a fixed and defined interval e.g. temperature or time.

- **ORDINAL**
  Scale for ordering observations from low to high with any ties attributed to lack of measurement sensitivity e.g. score from a questionnaire.

- **NOMINAL with order**
  Scale for grouping into categories with order e.g. mild, moderate or severe. This can be difficult to separate from ordinal.

- **NOMINAL without order**
  Scale for grouping into unique categories e.g. eye colour.

- **DICHOTOMOUS**
  As for nominal but two categories only e.g. male/female.

In addition to the classification of measurement scales, other related terms are used to describe types of data:

- **CATEGORICAL vs. NUMERICAL (quantitative vs. qualitative)**
  Data that represent categories, such as dichotomous (two categories) and nominal (more than two categories) observations, are collectively called categorical (qualitative). Data that are counted or measured using a numerically defined method are called numerical (quantitative).

- **DISCRETE vs. ORDERED CATEGORICAL**
  Discrete data arise from observations that can only take certain numerical values, usually counts such as number of children or number of patients attending a clinic in a year. Ordered categorical data are sometimes treated as discrete data, this is wrong. For example, using the Registrar General's classification of social class, it would be wrong to say that class I is five times the socio-economic status as class V, as there is not a strict numerical relationship between these categories. It follows, therefore, that

average social class is a meaningless statistic. Thus, ordered categorical data should not be treated as discrete data for statistical analysis. Discrete data may be treated as ordered categorical data in statistical analysis, but some information is lost in doing so.

- **CONTINUOUS**
  Continuous data are numerical data that can theoretically be measured in infinitely small units. For example, blood pressure is usually measured to the nearest 2mm Hg, but could be measured with much greater resolution of difference. The interval measurement scale is intended for continuous data. Sometimes continuous data are given discrete values at certain thresholds, for example age a last birthday is a discrete value but age itself is a continuous quantity; in these situations it is reasonable to treat discrete values as continuous. Remember that information is lost when continuous data are recorded only in ranges (ordered categories), and the statistical analysis of continuous data is more powerful than that of categorical data.

- **PERCENTAGES and RATIOS**
  Percentages or ratios summarise two pieces of information, namely their constituent numerator and denominator values. Simple ratios (0 to 1, i.e. the denominator is the maximum possible value that the numerator can take) can be treated as continuous data. More difficult to analyse data arise when the ratio represents a change, and the value can be negative. Ratios of observations compared with reference values, e.g. height relative to the mean of a reference population for a given sex and age, are difficult to handle as values may fall either side of 1 (100%).

Many statistical methods are appropriate only for data of certain measurement scales. When selecting a statistical method, it is essential to understand how the data to be analysed were measured. The best stage of investigation for pondering measurement scales is the design stage, at which the statistical limitations imposed by certain measurement scales may influence your choice of observations and methods of measurement.

Types of Data & Measurement Scales: Nominal, Ordinal, Interval and Ratio

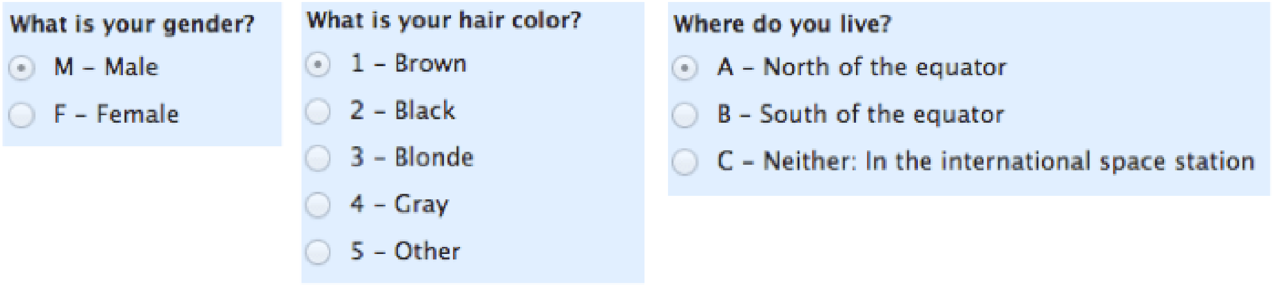
mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio

In statistics, there are four data measurement scales: nominal, ordinal, interval and ratio. These are simply ways to sub-categorize different types of data (here's an overview of statistical data types) . This topic is usually discussed in the context of academic teaching and less often in the "real world." If you are brushing up on this concept for a statistics test, thank a psychologist researcher named Stanley Stevens for coming up with these terms.

These four measurement scales (nominal, ordinal, interval, and ratio) are best understood with example, as you'll see below.

Nominal

Let's start with the easiest one to understand. Nominal scales are used for labeling variables, without any quantitative value. "Nominal" scales could simply be called "labels." Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. A good way to remember all of this is that "nominal" sounds a lot like "name" and nominal scales are kind of like "names" or labels.



What is your gender?

● M – Male
○ F – Female

What is your hair color?

● 1 – Brown
○ 2 – Black
○ 3 – Blonde
○ 4 – Gray
○ 5 – Other

Where do you live?

● A – North of the equator
○ B – South of the equator
○ C – Neither: In the international space station

Examples of Nominal Scales

Note: a sub-type of nominal scale with only two categories (e.g. male/female) is called "**dichotomous**." If you are a student, you can use that to impress your teacher.

Bonus Note #2: Other sub-types of nominal data are "nominal with order" (like "cold, warm, hot, very hot") and nominal without order (like "male/female").

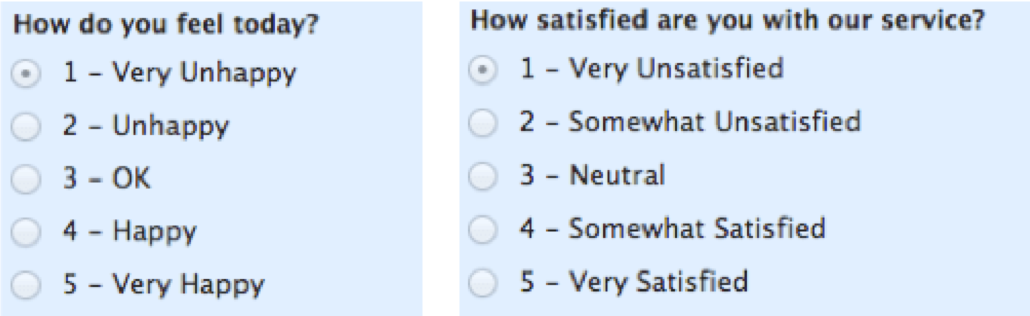
Ordinal

With ordinal scales, the *order* of the values is what's important and significant, but the differences between each one is not really known. Take a look at the example below. In each case, we know that a #4 is better than a #3 or #2, but we don't know–and cannot quantify–how *much* better it is. For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.

Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

“Ordinal” is easy to remember because is sounds like “order” and that’s the key to remember with “ordinal scales”–it is the *order* that matters, but that’s all you really get from these.

Advanced note: The best way to determine *central tendency* on a set of ordinal data is to use the mode or median; a purist will tell you that the mean cannot be defined from an ordinal set.



How do you feel today?

● 1 – Very Unhappy
○ 2 – Unhappy
○ 3 – OK
○ 4 – Happy
○ 5 – Very Happy

How satisfied are you with our service?

● 1 – Very Unsatisfied
○ 2 – Somewhat Unsatisfied
○ 3 – Neutral
○ 4 – Somewhat Satisfied
○ 5 – Very Satisfied

Example of Ordinal Scales

Interval

Interval scales are numeric scales in which we know both the order and the exact differences between the values. The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

Interval scales are nice because the realm of statistical analysis on these data sets opens up. For example, *central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated.

Like the others, you can remember the key points of an “interval scale” pretty easily. “Interval” itself means “space in between,” which is the important thing to remember–interval scales not only tell us about order, but also about the value between each item.

Here’s the problem with interval scales: they don’t have a “true zero.” For example, there is no such thing as “no temperature,” at least not with celsius. In the case of interval scales, zero doesn’t mean the absence of value, but is actually another number used on the scale, like 0 degrees celsius. Negative numbers also have meaning. Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide.

Confused? Ok, consider this: 10 degrees C + 10 degrees C = 20 degrees C. No problem there. 20 degrees C is not twice as hot as 10 degrees C, however, because there is no such thing as “no temperature” when it comes to the Celsius scale. When converted to Fahrenheit, it’s clear: 10C=50F and 20C=68F, which is clearly not twice as hot. I hope that makes sense. Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale…

Example of Interval Scale

Ratio

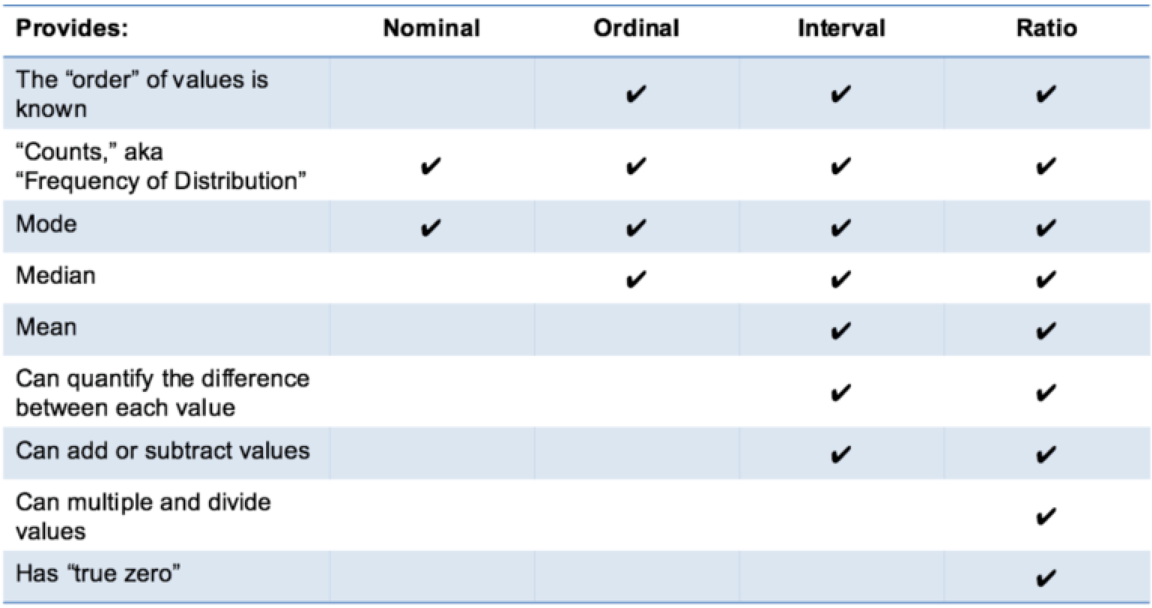
Ratio scales are the ultimate nirvana when it comes to measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied. At the risk of repeating myself, everything above about interval data applies to ratio scales, plus ratio scales have a clear definition of zero. Good examples of ratio variables include height, weight, and duration.

Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

This Device Provides Two Examples of Ratio Scales (height and weight)

Summary

In summary, **nominal** variables are used to "*name*," or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.



Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Summary of data types and scale measures

Understanding Descriptive Statistics

towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291

Statistics is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.

Initially, when we get the data, instead of applying fancy algorithms and making some predictions, we first try to read and understand the data by applying statistical techniques. By doing this, we are able to understand what type of distribution data has.

This blog aims to answer following questions:

1. What is Descriptive Statistics?

2. Types of Descriptive Statistics?

3. Measure of Central Tendency (Mean, Median, Mode)

4. Measure of Spread / Dispersion (Standard Deviation, Mean Deviation, Variance, Percentile, Quartiles, Interquartile Range)

5. What is Skewness?

6. What is Kurtosis?

7. What is Correlation?

Today, let's understand descriptive statistics once and for all. Let's start,

What is Descriptive S**tatistics?**

Descriptive statistics involves summarizing and organizing the data so they can be easily understood. Descriptive statistics, unlike inferential statistics, seeks to describe the data, but do not attempt to make inferences from the sample to the whole population. Here, we typically describe the data in a sample. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory.

Types of Descriptive Statistics?

Descriptive statistics are broken down into two categories. Measures of central tendency and measures of variability (spread).

Measure of Central Tendency

Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way "central" to the set.

Mean / Average

Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out. In a way, it is a single number which can estimate the value of whole data set.

Let’s calculate mean of the data set having 8 integers.

$$x = \frac{12+24+41+51+67+67+85+99}{8} = 55.75$$

Image 2

Median

Median is the value which divides the data in 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either **ascending or descending order**.

Note: If you sort data in descending order, it won’t affect median but IQR will be negative. We will talk about IQR later in this blog.

Median will be a middle term, if number of terms is odd

Median will be average of middle 2 terms, if number of terms is even.

Image 3

$$12+24+41+51+67+67+85+99 = 59$$

The median is 59 which will divide set of numbers into equal two parts. Since there are even numbers in the set, the answer is average of middle numbers 51 and 67.

Note: When values are in arithmetic progression (difference between the consecutive terms is constant. Here it is 2.), **median is always equal to mean**.

Image 4

$$2, 4, 6, 8, 10$$

An mean of these 5 numbers is 6 and so median.

Mode

Mode is the term appearing maximum time in data set i.e. term that has highest frequency.

Image 5

$$12, 24, 41, 51, 67, 67, 85, 99$$

In this data set, mode is 67 because it has more than rest of the values, i.e. twice.

But there could be a data set where there is no mode at all as all values appears same number of times. If two values appeared same time and more than the rest of the values then the data set is **bimodal**. If three values appeared same time and more than the rest of the values then the data set is **trimodal** and for n modes, that data set is **multimodal**.

Measure of Spread / Dispersion

Measure of Spread refers to the idea of variability within your data.

Standard deviation

Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

There are situations when we have to choose between sample or population Standard Deviation.

When we are asked to find SD of some part of a population, a segment of population; then we use sample Standard Deviation.

$$\textbf{S.D.} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x-\bar{x})^2}$$

Image 6

where x̅ is mean of a sample.

But when we have to deal with a whole population, then we use population Standard Deviation.

$$\textbf{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x-\mu)^2}$$

Image 7

where μ is mean of a population.

Though sample is a part of a population, their SD formulas should have been same, but it is not. To find out more about it, refer this link

As you know, in descriptive statistics, we generally deal with a data available in a sample, not in a population. So if we use previous data set, and substitute the values in sample formula,

$$\frac{\sqrt{(12-55.75)^2+(24-55.75)^2+(41-55.75)^2+(51-55.75)^2+(67-55.75)^2+(67-55.75)^2+(85-55.75)^2+(99-55.75)^2}}{\sqrt{7}}$$

Image 8

And answer is 29.62.

Mean Deviation / Mean Absolute Deviation

It is an average of absolute differences between each value in a set of values, and the average of all values of that set.

Mean Deviation [Image 9] (Image courtesy: My Photoshopped Collection)

So if we use previous data set, and substitute the values,

$$\textbf{M.D.} = \frac{1}{n}\sum_{i=0}^{n}|x_i - \bar{x}|$$

$$\frac{(|12-55.75|)+(|24-55.75|)+(|41-55.75|)+(|51-55.75|)+(|67-55.75|)+(|67-55.75|)+(|85-55.75|)+|99-55.75|)}{8}$$

Image 10

And answer is 23.75.

Variance

Variance is a square of average distance between each quantity and mean. That is it is square of standard deviation.

Image 11

$$\textbf{Variance} = (S.D.)^2$$

And answer is 877.34.

Range

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

Image 12

12, 24, 41, 51, 67, 67, 85, 99

Range is 99–12 = 87

Percentile

Percentile is a way to represent position of a values in data set. To calculate percentile, values in data set should always be in ascending order.

Image 13

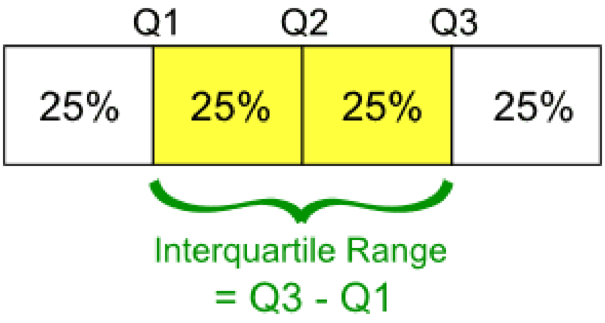
12, 24, 41, 51, 67, 67, 85, 99

The median 59 has 4 values less than itself out of 8. It can also be said as: In data set, 59 is 50th percentile because 50% of the total terms are less than 59. In general, if **k** is **nth** percentile, it implies that **n%** of the total terms are less than **k**.

Quartiles

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an **ascending order.**

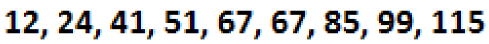
Quartiles [Image 14] (Image courtesy:



https://statsmethods.wordpress.com/2013/05/09/iqr/)

There are three quartile values. First quartile value is at 25 percentile. Second quartile is 50 percentile and third quartile is 75 percentile. Second quartile (Q2) is median of the whole data. First quartile (Q1) is median of upper half of the data. And Third Quartile (Q3) is median of lower half of the data.

So here, by analogy,


12, 24, 41, 51, 67, 67, 85, 99, 115

Q2 = 67: is 50 percentile of the whole data and is median.

Q1 = 41: is 25 percentile of the data.

Q3 = 85: is 75 percentile of the date.

Interquartile range (IQR) = Q3 - Q1 = 85 - 41 = 44

Note: If you sort data in descending order, IQR will be **-44**. The magnitude will be same, just sign will differ. Negative IQR is fine, if your data is in descending order. It just we negate smaller values from larger values, we prefer ascending order (Q3 - Q1).

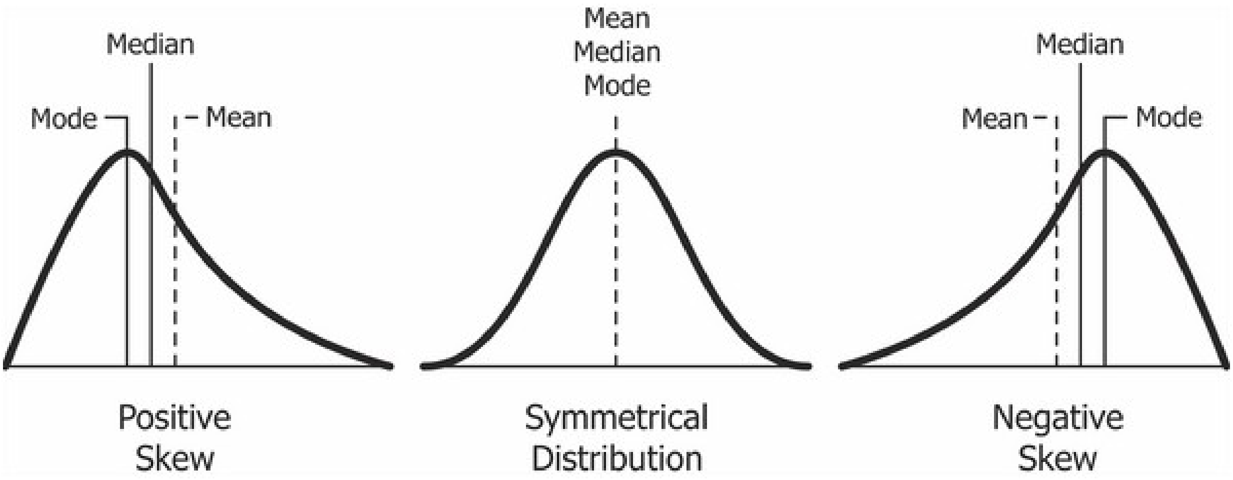
Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness.



Skewness [Image 16] (Image courtesy: https://www.safaribooksonline.com/library/view/clojure-for-data/9781784397180/ch01s13.html)

How to the skewness coefficient?

To calculate skewness coefficient of the sample, there are two methods:

1] Pearson First Coefficient of Skewness (Mode skewness)

Image 17

$$\frac{\textbf{Mean - Mode}}{\textbf{Standard Deviation}}$$

2] Pearson Second Coefficient of Skewness (Median skewness)

Image 18

$$\frac{\textbf{3 (Mean - Median)}}{\textbf{Standard Deviation}}$$

Interpretations

- The direction of skewness is given by the sign. A zero means no skewness at all.
- A negative value means the distribution is negatively skewed. A positive value means the distribution is positively skewed.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.

Sample problem: Use Pearson's Coefficient #1 and #2 to find the skewness for data with the following characteristics:

- Mean = 50.
- Median = 56.
- Mode = 60.
- Standard deviation = 8.5.

Pearson's First Coefficient of Skewness: -1.17.

Pearson's Second Coefficient of Skewness: -2.117.

Note: Pearson's first coefficient of skewness uses the mode. Therefore, if frequency of values is very low then it will not give a stable measure of central tendency. For example, the mode in both these sets of data is 9:

1, 2, 3, 4, 4, 5, 6, 7, 8, 9.

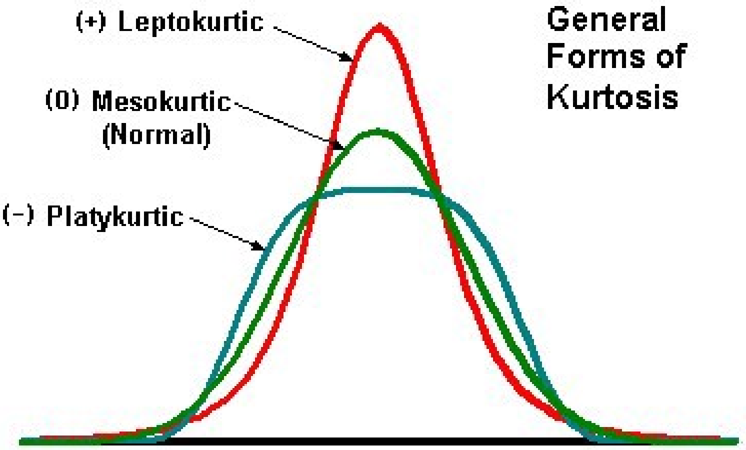
In the first set of data, the mode only appears twice. So it is not a good idea to use Pearson's First Coefficient of Skewness. But in second set,

1, 2, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 6, 7, 8, 9, 10, 12, 12, 13.

the mode 4 appears 8 times. Therefore, Pearson's Second Coefficient of Skewness will likely give you a reasonable result.

Kurtosis

The exact interpretation of the measure of Kurtosis used to be disputed, but is now settled. Its about existence of outliers. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution.



Kurtosis [Image 19] (Image courtesy: https://mvpprograms.com/help/mvpstats/distributions/SkewnessKurtosis)

There are three types of Kurtosis

Mesokurtic

Mesokurtic is the distribution which has similar kurtosis as normal distribution kurtosis, which is zero.

Leptokurtic

Distribution is the distribution which has kurtosis greater than a Mesokurtic distribution. Tails of such distributions are thick and heavy. If the curve of a distribution is more peaked than Mesokurtic curve, it is referred to as a Leptokurtic curve.

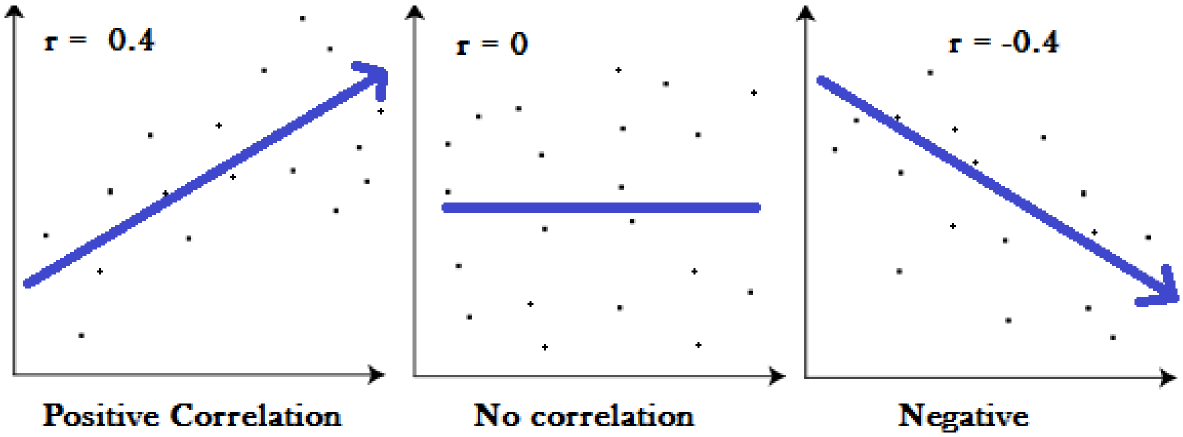
Platykurtic

Distribution is the distribution which has kurtosis lesser than a Mesokurtic distribution. Tails of such distributions thinner. If a curve of a distribution is less peaked than a Mesokurtic curve, it is referred to as a Platykurtic curve.

> The main **difference between skewness** and **kurtosis** is that the skewness refers to the degree of symmetry, whereas the kurtosis refers to the degree of presence of outliers **in the** distribution.

Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.



Correlation [Image 20] (Image courtesy: http://www.statisticshowto.com/what-is-correlation/)

The main result of a correlation is called the **correlation coefficient** (or “r”). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an “inverse” correlation).