WHAT IT IS

Return to Table of Contents

Descriptive statistics include the numbers, tables, charts, and graphs used to describe, organize, summarize, and present raw data. Descriptive statistics are most often used to examine:

- **central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **dispersion** (variability) of data, i.e. how spread out data are, as measured by the variance and its square root, the standard deviation.
- **skew** (symmetry) of data, i.e. how concentrated data are at the low or high end of the scale, as measured by the skew index.
- **kurtosis** (peakedness) of data, i.e. how concentrated data are around a single value, as measured by the kurtosis index.

Any description of a data set should include examination of the above. As a rule, looking at central tendency via the mean, median, and mode and dispersion via the variance or standard deviation is not sufficient. (See the definitions below for more details.)

DEFINITIONS

The following definitions are vital in understanding descriptive statistics:

- **Variables** are quantities or qualities that may assume any one of a set of values. Variables may be classified as nominal, ordinal, or interval.
  - **Nominal** variables use names, categories, or labels for qualitative values. Typical nominal variables include gender, ethnicity, job title, and so forth.
  - **Ordinal** variables, like nominal variables, are categorical variables. However, the order or rank of the categories is meaningful. For example, staff members may be asked to indicate their satisfaction with a training course on an ordinal scale ranging from "poor" to "excellent." Such categories could be converted to a numerical scale for further analysis.
  - **Interval** variables are purely numeric variables. The nominal and ordinal variables noted above are discrete since they do not permit making statements about degree, e.g., "Person A is three times more male than person B" or "Person A rated the course as five times more excellent than person B." Interval variables are continuous, and the difference between values is both meaningful and allows statements about extent or degree. Income and age are interval variables.
- **Frequency distributions** summarize and compress data by grouping them into classes and recording how many data points fall into each class. The frequency distribution is the foundation of descriptive statistics. It is a prerequisite for the various graphs used to display data and the basic statistics used to describe a data set, such as the mean, median, mode, variance, standard deviation, etc. (See the module on Frequency Distribution for more information.)
- **Measures of Central Tendency** indicate the middle and commonly occurring points in a data set. The three main measures of central tendency are discussed below.

- — **Mean** is the average, the most common measure of central tendency. The mean of a population is designated by the Greek letter mu (μ). The mean of a sample is designated by the symbol x-bar ($\bar{x}$). The mean may not always be the best measure of central tendency, especially if data are skewed. For example, average income is often misleading since those few individuals with extremely high incomes may raise the overall average.
- — **Median** is the value in the middle of the data set when the measurements are arranged in order of magnitude. For example, if 11 individuals were weighed and their weights arranged in ascending or descending order, the sixth value is the median since five values fall both above and below the sixth value. Median family income is often used in statistics because this value represents the exact middle of the data better than the mean. Fifty percent of families would have incomes above or below the median.
- — **Mode** is the value occurring most often in the data. If the largest group of people in a sample measuring age were 25 years old, then 25 would be the mode. The mode is the least commonly used measure of central tendency, particularly in large data sets. However, the mode is still important for describing a data set, especially when more than one value occurs frequently. In this instance, the data would be described as *bimodal* or *multimodal*, depending on whether two or more values occur frequently in the data set.

- **Measures of Dispersion** indicate how spread out the data are around the mean. Measures of dispersion are especially helpful when data are normally distributed, i.e. closely resemble the bell curve. The most common measures of dispersion follow.
  - — **Variance** is expressed as the sum of the squares of the differences between each observation and the mean, which quantity is then divided by the sample size. For populations, it is designated by the square of the Greek letter sigma (σ^2). For samples, it is designated by the square of the letter s (s^2). Since this is a quadratic expression, i.e. a number raised to the second power, variance is the second moment of statistics.

    Variance is used less frequently than standard deviation as a measure of dispersion. Variance can be used when we want to quickly compare the variability of two or more sets of interval data. In general, the higher the variance, the more spread out the data.

  - — **Standard deviation** is expressed as the positive square root of the variance, i.e. σ for populations and s for samples. It is the average difference between observed values and the mean. The standard deviation is used when expressing dispersion in the same units as the original measurements. It is used more commonly than the variance in expressing the degree to which data are spread out.

- — **Coefficient of variation** measures relative dispersion by dividing the standard deviation by the mean and then multiplying by 100 to render a percent. This number is designated as V for populations and v for samples and describes the variance of two data sets better than the standard deviation. For example, one data set has a standard deviation of 10 and a mean of 5. Thus, values vary by two times the mean. Another data set has the same standard deviation of 10 but a mean of 5,000. In this case, the variance and, hence, the standard deviation are insignificant.
- — **Range** measures the distance between the lowest and highest values in the data set and generally describes how spread out data are. For example, after an exam, an instructor may tell the class that the lowest score was 65 and the highest was 95. The range would then be 30. Note that a good approximation of the standard deviation can be obtained by dividing the range by 4.
- — **Percentiles** measure the percentage of data points which lie below a certain value when the values are ordered. For example, a student scores 1280 on the Scholastic Aptitude Test (SAT). Her scorecard informs her she is in the 90th percentile of students taking the exam. Thus, 90 percent of the students scored lower than she did.
- — **Quartiles** group observations such that 25 percent are arranged together according to their values. The top 25 percent of values are referred to as the upper quartile. The lowest 25 percent of the values are referred to as the lower quartile. Often the two quartiles on either side of the median are reported together as the interquartile range. Examining how data fall within quartile groups describes how deviant certain observations may be from others.

- **Measures of skew** describe how concentrated data points are at the high or low end of the scale of measurement. Skew is designated by the symbols Sk for populations and sk for samples. Skew indicates the degree of symmetry in a data set. The more skewed the distribution, the higher the variability of the measures, and the higher the variability, the less reliable are the data.

  Skew is calculated by either multiplying the difference between the mean and the median by three and then dividing by the standard deviation or by summing the cubes of the differences between each observation and the mean and then dividing by the cube of the standard deviation. Note that the use of cubic quantities helps explain why skew is called the *third* moment.

  More conceptually, skew defines the relative positions of the mean, median, and mode. If a distribution is skewed to the right (positive skew), the mean lies to the right of both the mode (most frequent value and hump in the curve) and median (middle value). That is, mode > median > mean. But, if the distribution is skewed left (negative skew), the mean lies to the left of the median and the mode. That is, mean < median < mode.

In a perfect distribution, mean = median = mode, and skew is 0. The values of the equations noted above will indicate left skew with a negative number and right skew with a positive number.

- **Measures of kurtosis** describe how concentrated data are around a single value, usually the mean. Thus, kurtosis assesses how peaked or flat is the data distribution. The more peaked or flat the distribution, the less normally distributed the data. And the less normal the distribution, the less reliable the data.

  Kurtosis is designated by the letter K for populations and k for samples and is calculated by raising the sum of the squares of the differences between each observation and the mean to the fourth power and then dividing by the fourth power of the standard deviation. Note that the use of the fourth power explains why kurtosis is called the *fourth* moment. Three degrees of kurtosis are noted:
  - — **Mesokurtic** distributions are, like the normal bell curve, neither peaked nor flat.
  - — **Platykurtic** distributions are flatter than the normal bell curve.
  - — **Leptokurtic** distributions are more peaked than the normal bell curve.

  The ideal value rendered by the equation for kurtosis is 3, the kurtosis of the normal bell curve. The higher the number above 3, the more leptokurtic (peaked) is the distribution. The lower the number below 3, the more platykurtic (flat) is the distribution.

WHEN TO USE IT

Descriptive statistics are recommended when the objective is to describe and discuss a data set more generally and conveniently than would be possible using raw data alone. They are routinely used in reports which contain a significant amount of qualitative or quantitative data. Descriptive statistics help summarize and support assertions of fact.

Note that a thorough understanding of descriptive statistics is essential for the appropriate and effective use of all normative and cause-and-effect statistical techniques, including hypothesis testing, correlation, and regression analysis. Unless descriptive statistics are fully grasped, data can be easily misunderstood and, thereby, misrepresented.

All four moments should be explored whenever possible. Skew and kurtosis should be examined any time you deal with interval data since they jointly help determine whether the variable underlying a frequency distribution is normally distributed. Since normal distribution is a key assumption behind most statistical techniques, the skew and kurtosis of any interval data set must be analyzed. Data that show significant variation, skew, or kurtosis should not be used in making inferences, drawing conclusions, or espousing recommendations.

HOW TO PREPARE IT

Since statistical analysis software and most spreadsheets generate all required descriptive statistics, computer applications offer the best means of preparing such information. Nonetheless, for reference purposes, formulas for the various measures follow in the same order in which previously discussed. Note that in all cases, the use of Greek or upper case Latin letters refer to population parameters and that the use of lower case Latin letters refers to sample statistics.

Measures of Central Tendency

Formulas for assessing the central tendency of data sets follow:

- **Mean**
  - — Population data:

$$\mu = \frac{\Sigma FX_i}{N}$$

μ = population mean
N = number of items

$\sum FX_i$ = sum of frequency of each class times the class midpoint X

Note that F = 1 for raw data since each datum is a class unto itself.

  - — Sample data:

$$\bar{x} = \frac{\Sigma fx_i}{n}$$

$\bar{x}$ = sample mean
n = number of items

$\sum fx_i$ = sum of frequency of each class times the class midpoint x

Note that f = 1 for raw data since each datum is a class unto itself.

- **Median**
  - — Population raw data in ascending or descending order:

$$\text{Median} = \text{the } \left(\frac{N+1}{2}\right) \text{th item of data}$$

N = number of items

— Sample raw data in ascending or descending order:

$$\text{median} = \text{the } \left(\frac{n+1}{2}\right)\text{th item of data}$$

n = number of items

— Population grouped data in ascending or descending order:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{F_M}\right) \times C$$

L = lower limit of median class
N = number of items
F_M= frequency of median class
C = width of class interval
F = sum of frequencies up to but not including median class

— Sample grouped data in ascending or descending order:

$$\text{median} = l + \left(\frac{\frac{n}{2} - f}{f_m}\right) \times c$$

l = lower limit of median class
n = number of items
f_m= frequency of median class
c = width of class interval
f = sum of frequencies up to but not including median class

- **Mode**
  — Population grouped data:

$$\text{Mode} = L + \left(\frac{D_1}{D_1 + D_2}\right) \times C$$

L = lower limit of modal class
C = width of class interval
D_1 = modal class frequency minus frequency of previous class
D_2 = modal class frequency minus frequency of following class

— Sample grouped data:

$$\text{Mode} = l + \left(\frac{d_1}{d_1 + d_2} \right) \times c$$

l = lower limit of modal class
c = class interval width
d_1 = modal class frequency minus frequency of previous class
d_2 = modal class frequency minus frequency of following class

Measures of Dispersion

Formulas for assessing the dispersion of data sets:

- **Variance**

— Population data:

$$\sigma^2 = \frac{\Sigma F(X_i - \mu)^2}{N}$$

σ^2 = population variance
N = number of items
F = frequency of each class
$X_i - \mu$ = difference of each item and population mean

Note that F = 1 for raw data since each datum is a class unto itself.

— Sample data:

$$s^2 = \frac{\Sigma f(x_i - \bar{x})^2}{n - 1}$$

s^2 = sample variance
n = number of items
f = frequency of each class
$x_i - \bar{x}$ = difference of each item and sample mean

Note that f = 1 for raw data since each datum is a class unto itself.

- **Standard deviation**
  - Population data:

$$\sigma = \sqrt{\frac{\Sigma F(X_i - \mu)^2}{N}}$$

σ = population standard deviation
N = number of items
F = frequency of each class
$X_i - \mu$ = difference of each item and population mean

Note that F = 1 for raw data since each datum is a class unto itself.

  - Sample data:

$$s = \sqrt{\frac{\Sigma f(x_i - \bar{x})^2}{n - 1}}$$

s^2 = sample variance
n = number of items
f = frequency of each class
$x_i - \bar{x}$ = difference of each item and sample mean

Note that f = 1 for raw data since each datum is a class unto itself.

- **Coefficient of variation**
  - Populations:

$$V = \frac{\sigma}{\mu}$$

V = coefficient of variation
σ = population standard deviation
μ = population mean

  - Samples:

$$v = \frac{s}{\bar{x}}$$

v = coefficient of variation
s = sample standard deviation
$\bar{x}$ = sample mean

- **Range**
  - Populations or samples:

    Range = Largest Value - Smallest Value

    An approximation of the standard deviation is the range divided by 4.

- **Percentiles**
  - Populations or samples:

    The *p*th percentile is the value for which at most *p*% of the values are than that value and at most (100 - *p*%) of the values are greater than that value. The median is the 50th percentile.

- **Quartiles**
  - Populations or samples:

    First quartile = Q1 = 25th percentile
    Second quartile = Q2 = 50th percentile (median)
    Third quartile = Q3 = 75th percentile

Measures of Skew

Formulas for assessing the skew of a data set follow below. The ideal value of these equations is 0, the skew of the perfectly distributed normal bell curve. Positive values indicate a distribution skewed to the right (positive skew). Negative values indicate a distribution is skewed to the left (negative skew).

- **Skew**
  - Populations using median:

$$Sk = \frac{3(\mu - \text{Median})}{\sigma}$$

Sk = population skew μ = population mean
σ = population standard deviation

  - Samples using median:

$$sk = \frac{3(\bar{x} - \text{median})}{s}$$

sk = sample skew $\bar{x}$ = sample mean
s = sample standard deviation

— Population data:

$$\mathbf{Sk} = \frac{\Sigma F(X_i - \mu)^3}{\sigma^3}$$

Sk = population skew
σ = population standard deviation
F = frequency of each class
$X_i - \mu$ = difference of each item and population mean

Note that F = 1 for raw data since each datum is a class unto itself.

— Sample data:

$$\mathbf{sk} = \frac{\Sigma f(x_i - \bar{x})^3}{s^3}$$

sk = sample skew
s = sample standard deviation
f = frequency of each class
$x_i - \bar{x}$ = difference of each item and sample mean

Note that f = 1 for raw data since each datum is a class unto itself.

Measures of Kurtosis

Formulas for assessing the kurtosis of a data set follow below. The ideal value of these equations is 3, the kurtosis of the perfectively distributed normal bell curve. The higher the value above 3, the more peaked is distribution. The lower the value below 3, the more flat is distribution.

- **Kurtosis**

— Population data:

$$\mathbf{K} = \frac{\Sigma F(X_i - \mu)^4}{\sigma^4}$$

K = population kurtosis
σ = population standard deviation
F = frequency of each class
$X_i - \mu$ = difference of each item and population mean

Note that F = 1 for raw data since each datum is a class unto itself.

— Sample data:

$$\mathbf{k} = \frac{\Sigma f(x_i - \bar{x})^4}{s^4}$$

k = sample kurtosis
s = sample standard deviation
f = frequency of each class
$x_i - \bar{x}$ = difference of each item and sample mean

Note that f = 1 for raw data since each datum is a class unto itself.

ADVANTAGES

Descriptive statistics can:

- be essential for arranging and displaying data
- form the basis of rigorous data analysis
- be much easier to work with, interpret, and discuss than raw data
- help examine the tendencies, spread, normality, and reliability of a data set
- be rendered both graphically and numerically
- include useful techniques for summarizing data in visual form
- form the basis for more advanced statistical methods

DISADVANTAGES

Descriptive statistics can:

- be misused, misinterpreted, and incomplete
- be of limited use when samples and populations are small
- demand a fair amount of calculation and explanation
- fail to fully specify the extent to which non-normal data are a problem
- offer little information about causes and effects
- be dangerous if not analyzed completely