Chapter 200

Descriptive Statistics

Introduction

This procedure summarizes variables both statistically and graphically. Information about the location (center), spread (variability), and distribution is provided. The procedure provides a large variety of statistical information about a single variable.

Kinds of Research Questions

The use of this module for a single variable is generally appropriate for one of four purposes: numerical summary, data screening, outlier identification (which sometimes is incorporated into data screening), and distributional shape. We will briefly discuss each of these now.

Numerical Descriptors

The numerical descriptors of a sample are called statistics. These statistics may be categorized as location, spread, shape indicators, percentiles, and interval estimates.

Location or Central Tendency

One of the first impressions that we like to get from a variable is its general location. You might think of this as the center of the variable on the number line. The average (mean) is a common measure of location. When investigating the center of a variable, the main descriptors are the mean, median, mode, and the trimmed mean. Other averages, such as the geometric and harmonic mean, have specialized uses. We will now briefly compare these measures.

If the data come from the normal distribution, the mean, median, mode, and the trimmed mean are all equal. If the mean and median are very different, most likely there are outliers in the data, or the distribution is skewed. If this is the case, the median is probably a better measure of location. The mean is very sensitive to extreme values and can be seriously contaminated by just one observation.

A compromise between the mean and median is given by the trimmed mean (where a predetermined number of observations are trimmed from each end of the data distribution). This trimmed mean is more robust than the mean but more sensitive than the median. Comparison of the trimmed mean to the median should show the trimmed mean approaching the median as the degree of trimming increases. If the trimmed mean converges to the median for a small degree of trimming, say 5 or 10%, the number of outliers is relatively few.

Variability, Dispersion, or Spread

After establishing the center of a variable's values, the next question is how closely the data fall about this center. The pattern of the values around the center is called the *spread*, *dispersion*, or *variability*. There are numerous measures of variability: range, variance, standard deviation, interquartile range, and so on. All of these measures of dispersion are affected by outliers to some degree, but some do much better than others.

The *standard deviation* is one of the most popular measures of dispersion. Unfortunately, it is greatly influenced by outlying observations and by the overall shape of the distribution. Because of this, various substitutes for it have been developed. It will be up to you to decide which is best in a given situation.

Shape

The shape of the distribution describes the pattern of the values along the number line. Are there a few unique values that occur over and over, or is there a continuum? Is the pattern symmetric or asymmetric? Are the data bell shaped? Do they seem to have a single center or are there several areas of clumping? These are all aspects of the shape of the distribution of the data.

Two of the most popular measures of shape are skewness and kurtosis. *Skewness* measures the direction and lack of *symmetry*. The more skewed a distribution is, the greater the need for using robust estimators, such as the median and the interquartile range. Positive skewness indicates a longtailedness to the right while negative skewness indicates longtailedness to the left. *Kurtosis* measures the heaviness of the tails. A kurtosis value less than three indicates lighter tails than a normal distribution. Kurtosis values greater than three indicate heavier tails than a normal distribution.

The measures of shape require more data to be accurate. For example, a reasonable estimate of the mean may require only ten observations in a random sample. The standard deviation will require at least thirty. A reasonably detailed estimate of the shape (especially if the tails are important) will require several hundred observations.

Percentiles

Percentiles are extremely useful for certain applications as well as for cases when the distribution is very skewed or contaminated by outliers. If the distribution of the variable is skewed, you might want to use the exact interval estimates for the percentiles.

Percentile Type

The available percentile calculation options are

Ave X(p[n+1])

The $100p^{th}$ percentile is computed as

$$Z_p = (1 - g)X_{[k_1]} + gX_{[k_2]}$$

where k_1 equals the integer part of $p(n + 1)$, $k_2 = k_1 + 1$, g is the fractional part of $p(n + 1)$, and $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest. This gives the common value of the median.

Ave X(p[n])

The 100p^{th} percentile is computed as

$$Z_p = (1 - g)X_{[k_1]} + gX_{[k_2]}$$

where k_1 equals the integer part of np, $k_2 = k_1 + 1$, g is the fractional part of np, and $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest.

Closest to np

The 100p^{th} percentile is computed as

$$Z_p = X_{[k_1]}$$

where k_1 equals the integer that is closest to np and $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest.

EDF

The 100p^{th} percentile is computed as

$$Z_p = X_{[k_1]}$$

where k_1 equals the integer part of np if np is exactly an integer or the integer part of $np + 1$ if np is not exactly an integer. $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

EDF w/ Ave

The 100p^{th} percentile is computed as

$$Z_p = \frac{\left(X_{[k_1]} + X_{[k_2]}\right)}{2}$$

where k_1 and k_2 are defined as follows: If np is an integer, $k_1 = k_2 = np$. If np is not exactly an integer, k_1 equals the integer part of np and $k_2 = k_1 + 1$. $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

Custom Percentiles

Smallest Percentile

By default, the smallest percentile displayed is the 1st percentile. This option lets you change this value to any value between 0 and 100. For example, you might enter 2.5 to see the 2.5th percentile.

Largest Percentile

By default, the largest percentile displayed is the 99th percentile. This option lets you change this value to any value between 0 and 100. For example, you might enter 97.5 to see the 97.5th percentile.

Confidence Limits or Interval Estimates

An interval estimate of a statistic gives a range of its possible values. Confidence limits are a special type of interval estimate that have, under certain conditions, a level of confidence or probability attached to them.

If the assumption of normality is valid, the confidence intervals for the mean, variance, and standard deviation are valid. However, the standard error of each of these intervals depends on the sample standard deviation and the sample size. If the sample standard deviation is inaccurate, these other measures will be also. The bottom line is that outliers not only affect the standard deviation but also all confidence limits that use the sample standard deviation. It should be obvious then that the standard deviation is a critical measure of dispersion in parametric methods.

Data Screening

Data screening involves missing data, data validity, and outliers. If these issues are not dealt with prior to the use of descriptive statistics, errors in interpretations are very likely.

Missing Data

Whenever data are missing, questions need to be asked.

1. Is the missingness due to incomplete data collection? If so, try to complete the data collection.
2. Is the missingness due to nonresponse from a survey? If so, attempt to collect data from the non-responders.
3. Are the missing data due to a censoring of data beyond or below certain values? If so, some different statistical tools will be needed.
4. Is the pattern of missingness random? If only a few data points are missing from a large data set and the pattern of missingness is random, there is little to be concerned with. However, if the data set is small or moderate in size, any degree of missingness could cause bias in interpretations.

Whenever missing values occur without answers to the above questions, there is little that can be done. If the distributional shape of the variable is known and there are missing data for certain percentiles, estimates could be made for the missing values. If there are other variables in the data set as well and the pattern of missingness is random, multiple regression and multivariate methods can be used to estimate the missing values.

Data Validity

Data validity needs to be confirmed prior to any statistical analysis, but it usually begins after a univariate descriptive analysis. Extremes or outliers for a variable could be due to a data entry error, to an incorrect or inappropriate specification of a missing code, to sampling from a population other than the intended one, or due to a natural abnormality that exists in this variable from time to time. The first two cases of invalid data are easily corrected. The latter two require information about the distribution form and necessitate the use of regression or multivariate methods to re-estimate the values.

Outliers

Outliers in a univariate data set are defined as observations that appear to be inconsistent with the rest of the data. An outlier is an observation that sticks out at either end of the data set.

The visualization of univariate outliers can be done in three ways: with the stem-and-leaf plot, with the box plot, and with the normal probability plot. In each of these informal methods, the outlier is far removed from the rest of the data. A word of caution: the box plot and the normal probability plot evaluate the potentiality of an outlier assuming the data are normally distributed. If the variable is not normally distributed, these plots may indicate many outliers. You must be careful about checking what distributional assumptions are behind the outliers you may be looking for.

Outliers can completely distort descriptive statistics. For instance, if one suspects outliers, a comparison of the mean, median, mode, and trimmed mean should be made. If the outliers are only to one side of the mean, the median is a better measure of location. On the other hand, if the outliers are equally divergent on each side of the center, the mean and median will be close together, but the standard deviation will be inflated. The interquartile range is the only measure of variation not greatly affected by outliers. Outliers may also contaminate measures of skewness and kurtosis as well as confidence limits.

This discussion has focused on univariate outliers, in a simplistic way. If the data set has several variables, multiple regression and multivariate methods must be used to identify these outliers.

Normality

A primary use of descriptive statistics is to determine whether the data are normally distributed. If the variable is normally distributed, you can use parametric statistics that are based on this assumption. If the variable is not normally distributed, you might try a transformation on the variable (such as, the natural log or square root) to make the data normal. If a transformation is not a viable alternative, nonparametric methods that do not require normality should be used.

NCSS provides seven tests to formally test for normality. If a variable fails a normality test, it is critical to look at the box plot and the normal probability plot to see if an outlier or a small subset of outliers has caused the nonnormality. A pragmatic approach is to omit the outliers and rerun the tests to see if the variable now passes the normality tests.

Always remember that a reasonably large sample size is necessary to detect normality. Only extreme types of nonnormality can be detected with samples less than fifty observations.

There is a common misconception that a histogram is always a valid graphical tool for assessing normality. Since there are many subjective choices that must be made in constructing a histogram, and since histograms generally need large sample sizes to display an accurate picture of normality, preference should be given to other graphical displays such as the box plot, the density trace, and the normal probability plot.

Data Structure

The data are contained in a single variable.

Height Dataset (Subset)

Height
64
63
67
.
.
.

Example 1 – Running Descriptive Statistics

This section presents a detailed example of how to run a descriptive statistics report on the *Height* variable in the Height dataset.

Setup

To run this example, complete the following steps:

1 Open the Height example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Height** and click **OK**.

2 Specify the Descriptive Statistics procedure options

- Find and open the **Descriptive Statistics** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Variables.....**Height**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Descriptive Statistics Reports

This report is rather large and complicated, so we will define each section separately. Usually, you will focus on only a few items from this report. Unfortunately, each user wants a different few items, so we had to include much more than any one user needs!

Several of the formulas involve both raw and central moments. The raw moments are defined as:

$$m'_r = \frac{\sum_{i=1}^{n} x_i^r}{n}$$

The central moments are defined as:

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n}$$

Large sample estimates of the standard errors are provided for several statistics. These are based on the following formula from Kendall and Stuart (1987):

$$Var(m_r) = \frac{m_{2r} - m_r^2 + 4m_2 m_{r-1}^2 - 2r m_{r-1} m_{r+1}}{n}$$

$$Var\big(g(x)\big) = \left(\frac{dg}{dx}\right)^2 Var(x)$$

Summary Statistics

Summary Statistics for Height

N	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
20	62.1	8.441128	1.887493	51	79	28

N

This is the number of nonmissing values. If no frequency variable was specified, this is the number of nonmissing rows.

Mean

This is the average of the data values. (See *Means Section* below.)

Standard Deviation

This is the standard deviation of the data values. (See *Variation Section* below.)

Standard Error

This is the standard error of the mean. (See *Means Section* below.)

Minimum

The smallest value in this variable.

Maximum

The largest value in this variable.

Range

The difference between the largest and smallest values for a variable. If the data for a given variable is normally distributed, a quick estimate of the standard deviation can be made by dividing the range by six.

Data Summary

Data Summary for Height

Rows	Sum of Frequencies	Data Values			Sum of Squares	
		Missing	Unique	Sum	Total	Adjusted
20	20	0	14	1242	78482	1353.8

Rows

This is the total number of rows available in this variable.

Sum of Frequencies

This is the number of nonmissing values. If no frequency variable was specified, this is the number of nonmissing rows.

Data Values: Missing

The number of missing (empty) rows.

Data Values: Unique

This is the number of unique values in this variable. This value is useful for finding data entry errors and for determining if a variable is continuous or discrete.

Data Values: Sum

This is the sum of the data values.

Sum of Squares: Total

This is the sum of the squared values of the variable. It is sometimes referred to as the *unadjusted sum of squares.* It is reported for its usefulness in calculating other statistics and is not interpreted directly.

$$\text{Sum of Squares} = \sum_{i=1}^{n} x_i^2$$

Sum of Squares: Adjusted

This is the sum of the squared differences from the mean.

$$\text{Adjusted Sum of Squares} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Mean and Location Statistics

Mean and Location Statistics for Height

				95% Confidence Interval Limits*		Test of H0: Value = 0	
Statistic	**N**	**Value**	**Standard Error**	**Lower**	**Upper**	**T-Value**	**P-Value**
Mean	20	62.1	1.887493	58.14943	66.05057	32.9008	0.00000
Median	20	59.5		56	67		
Geometric Mean	20	61.57052		57.84089	65.54064		
Harmonic Mean	20	61.05865		57.53493	65.04214		
Sum	20	1242	37.74987	1162.989	1321.011		
Mode**	20	52 (3)					

* The geometric mean confidence interval assumes that ln(Height) values are normally distributed.
The harmonic mean confidence interval assumes that 1/(Height) values are normally distributed.
** The value for the mode is displayed as "Mode Value (Mode Value Frequency)".

Statistic

The mean or location statistic reported.

N

The number of data values used to calculate the associated statistic.

Mean: Value

This is the average of the data values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean: Standard Error

This is the standard error of the mean. This is the estimated standard deviation for the distribution of sample means for an infinite population.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Mean: Lower and Upper 95% Confidence Interval Limits

This is the upper and lower values of a $100(1-\alpha)\%$ confidence interval estimate for the mean based on a *t* distribution with *n* – 1 degrees of freedom. This interval estimate assumes that the population standard deviation is not known and that the data for this variable are normally distributed.

$$\bar{x} \pm t_{a/2,n-1} s_{\bar{x}}$$

Mean: T-Value

This is the *t*-test value for testing that the sample mean is equal to zero versus the alternative that it is not. The degrees of freedom for this t-test are *n* – 1. The variable that is being tested must be approximately normally distributed for this test to be valid.

$$t_{\alpha/2,n-1} = \frac{\bar{x}}{s_{\bar{x}}}$$

Mean: P-Value

This is the *p*-value of the above *t*-test, assuming a two-tailed test. Generally, this *p*-value is compared to the level of significance, 0.05 or 0.01, chosen by the researcher. If the *p*-value is less than the pre-determined level of significance, the sample mean is different from zero.

Median: Value

The value of the median. The median is the 50th percentile of the data set. It is the point that splits the dataset in half. The value of the percentile depends upon the percentile method that was selected.

Median: Lower and Upper 95% Confidence Interval Limits

These are the values of an exact $100(1-\alpha)\%$ confidence interval for the median. These exact confidence intervals are discussed in the *Percentile Section*.

Geometric Mean: N

The number of positive numbers used in computing the geometric mean.

Geometric Mean: Value

The geometric mean (GM) is an alternative type of mean that is used for business, economic, and biological applications. Only non-negative and non-zero values are used in the computation. If one of the values is zero, the geometric mean is defined to be zero. Instead of returning a zero, **NCSS** omits values equal to zero.

One example of when the GM is appropriate is when a variable is the product of many small effects combined by multiplication instead of addition.

$$GM = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

An alternative form, showing the GM's relationship to the arithmetic mean, is:

$$GM = \exp\left(\frac{1}{n}\sum \ln(x_i)\right)$$

Geometric Mean: Lower and Upper 95% Confidence Interval Limits

These are the values for the limits of a $100(1-\alpha)\%$ confidence interval for the geometric mean.

Harmonic Mean: N

The number of nonzero numbers used in computing the harmonic mean.

Harmonic Mean: Value

The harmonic mean is used to average rates. For example, suppose we want the average speed of a bus that travels a fixed distance every day at speeds s_1, s_2, and s_3. The average speed, found by dividing the total distance by the total time, is equal to the harmonic mean of the three speeds. The harmonic mean is appropriate when the distance is constant from trial to trial and the time required was variable. However, if the times were constant and the distances were variable, the arithmetic mean would have been appropriate.

Only nonzero values may be used in its calculation.

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Harmonic Mean: Lower and Upper 95% Confidence Interval Limits

These are the values for the limits of a $100(1-\alpha)\%$ confidence interval for the harmonic mean.

Sum: Value

This is the sum of the data values. The standard error and confidence limits are found by multiplying the corresponding values for the mean by the sample size, *n*.

Sum: Standard Error

This is the standard deviation of the distribution of sums. With this standard error, confidence intervals and hypothesis testing can be done for the sum. The assumptions for the interval estimate of the mean must also hold here.

$$s_{sum} = n s_{\bar{x}}$$

Mode: Value

This is the most frequently occurring value in the data. The value for the mode is displayed as "Mode Value (Mode Value Frequency)". For example, a mode of "50 (4)" indicates that the most frequently occurring value is 50, and it occurs 4 times in the dataset. The mode is not given if all of the data values are unique or if the number of modes is greater than one.

Variation Statistics

Variation Statistics for Height

Statistic	N	Value	Standard Error	95% Confidence Interval Limits*	
				Lower	Upper
Variance	20	71.25263	17.01612	41.20865	152.0011
Standard Deviation	20	8.441128	1.425427	6.419396	12.32887
Standard Deviation (Unbiased)	20	8.552877			
Standard Error of the Mean	20	8.441128	0.3187352	1.435421	2.756819
Range	20	28			
Interquartile Range	20	14			

Statistic

The variation statistic reported.

N

The number of data values used to calculate the associated statistic.

Variance: Value

The sample variance, s^2, is a popular measure of dispersion. It is an average of the squared deviations from the mean.

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

Variance: Standard Error

This is a large-sample estimate of the standard error of s^2 for an infinite population.

Variance: Lower 95% Confidence Interval Limit

This is the lower limit value of a $100(1-\alpha)\%$ confidence interval estimate for the variance based on the chi-squared distribution with *n - 1* degrees of freedom. This interval estimate assumes that the variable is normally distributed.

$$LCL = \frac{s^2(n-1)}{\chi^2_{\alpha/2,n-1}}$$

Variance: Upper 95% Confidence Interval Limit

This is the upper limit value of a $100(1-\alpha)\%$ confidence interval estimate for the variance based on the chi-squared distribution with *n - 1* degrees of freedom. This interval estimate assumes that the variable is normally distributed.

$$UCL = \frac{s^2(n-1)}{\chi^2_{1-\alpha/2,n-1}}$$

Standard Deviation: Value

The sample standard deviation, *s*, is a popular measure of dispersion. It measures the average distance between a single observation and its mean. The use of *n - 1* in the denominator instead of the more natural *n* is often of concern. It turns out that if *n* (instead of *n - 1*) were used, a biased estimate of the population standard deviation would result. The use of *n - 1* corrects for this bias.

Unfortunately, *s* is inordinately influenced by outliers. For this reason, you must always check for outliers in your data before you use this statistic. Also, *s* is a biased estimator of the population standard deviation. An unbiased estimate, calculated by adjusting *s*, is given under the heading *Unbiased Std Dev*.

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

Another form of the above formula shows that the standard deviation is proportional to the difference between each pair of observations. Notice that the sample mean does not enter into this second formulation.

$$s = \sqrt{\frac{\sum_{all\ i,j\ where\ i<j}\left(x_i - x_j\right)^2}{n(n-1)}}$$

Standard Deviation: Standard Error

This is a large sample estimate of the standard error of *s* for an infinite population.

Standard Deviation: Lower 95% Confidence Interval Limit

This is the lower limit value of a $100(1-\alpha)\%$ confidence interval estimate for the standard deviation based on the chi-squared distribution with *n* - 1 degrees of freedom. This interval estimate assumes that the variable is normally distributed.

$$LCL = \sqrt{\frac{s^2(n-1)}{\chi^2_{\alpha/2,n-1}}}$$

Standard Deviation: Upper 95% Confidence Interval Limit

This is the upper limit value of a $100(1-\alpha)\%$ confidence interval estimate for the standard deviation based on the chi-squared distribution with *n* - 1 degrees of freedom. This interval estimate assumes that the variable is normally distributed.

$$UCL = \sqrt{\frac{s^2(n-1)}{\chi^2_{1-\alpha/2,n-1}}}$$

Standard Deviation (Unbiased): Value

This is an unbiased estimate of the standard deviation. If the data come from a normal distribution, the sample variance, s^2, is an unbiased estimate of the population variance. Unfortunately, the sample standard deviation, s, is a biased estimate of the population standard deviation. This bias is usually overlooked, but division of s by a correction factor, c_4, will correct for this bias. This is frequently done in quality control applications. The formula for c_4 is:

$$c_4 = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma\big((n-1)/2\big)}$$

where

$$\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt$$

Standard Error of the Mean: Value

This is an estimate of the standard error of the mean. This is an estimate of the precision of the sample mean. Its standard error and confidence limits are calculated by dividing the corresponding Standard Deviation value by the square root of n.

Standard Error of the Mean: Standard Error

This is an estimate of the standard error of the standard error of the mean.

Standard Error of the Mean: Lower and Upper 95% Confidence Interval Limits

These are the values for the limits of a $100(1-\alpha)\%$ confidence interval for the standard error of the mean.

Range: Value

The difference between the largest and smallest values for a variable. If the data for a given variable is normally distributed, a quick estimate of the standard deviation can be made by dividing the range by six.

Interquartile Range: Value

This is the interquartile range (IQR). It is the difference between the third quartile and the first quartile (between the 75th percentile and the 25th percentile). This represents the range of the middle 50 percent of the distribution. It is a very robust (not affected by outliers) measure of dispersion. In fact, if the data are normally distributed, a robust estimate of the sample standard deviation is IQR/1.35. If a distribution is very concentrated around its mean, the IQR will be small. On the other hand, if the data are widely dispersed, the IQR will be much larger.

Skewness and Kurtosis Statistics

Skewness and Kurtosis Statistics for Height

Statistic	Value	Standard Error
Skewness	0.471155	0.3343679
Skewness (Fisher's g1)	0.5102501	
Kurtosis	2.140641	0.5338696
Kurtosis (Fisher's g2)	-0.7479873	
Coefficient of Variation (COV)	0.135928	0.0148992
Coefficient of Dispersion (COD)	0.1142857	

Skewness: Value

This statistic measures the direction and degree of asymmetry. A value of zero indicates a symmetrical distribution. A positive value indicates skewness (longtailedness) to the right while a negative value indicates skewness to the left. Values between -3 and +3 are typical values of samples from a normal distribution. For an alternative measure of skewness, see *Fisher's g1*, below.

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}}$$

Skewness: Standard Error

This is a large sample estimate of the standard error of skewness for an infinite population.

Fisher's g1: Value

Fisher's g_1 is an alternative measure of skewness.

$$g_1 = \frac{\sqrt{n(n-1)}b_1}{n-2}$$

Kurtosis: Value

This statistic measures the heaviness of the tails of a distribution. The usual reference point in kurtosis is the normal distribution. If this kurtosis statistic equals three and the skewness is zero, the distribution is normal. Unimodal distributions that have kurtosis greater than three have heavier or thicker tails than the normal. These same distributions also tend to have higher peaks in the center of the distribution (leptokurtic). Unimodal distributions whose tails are lighter than the normal distribution tend to have a kurtosis that is less than three. In this case, the peak of the distribution tends to be broader than the normal (platykurtic). Be forewarned that this statistic is an unreliable estimator of kurtosis for small sample sizes. For an alternative measure of skewness, see *Fisher's g2*, below.

$$b_2 = \frac{m_4}{m_2^2}$$

Kurtosis: Standard Error

This is a large sample estimate of the standard error of kurtosis for an infinite population.

Fisher's g2: Value

Fisher's g_2 is an alternative measure of kurtosis.

$$g_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left[b_2 - \frac{3(n-1)}{n+1}\right]$$

Coefficient of Variation (COV): Value

The *coefficient of variation* is a relative measure of dispersion. It is most often used to compare the amount of variation in two samples. It can be used for the same data over two time periods or for the same time period but two different places. It is the standard deviation divided by the mean:

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

Coefficient of Variation (COV): Standard Error

This is a large sample estimate of the standard error of the estimated coefficient of variation.

Coefficient of Dispersion (COD): Value

The *coefficient of dispersion* is a robust, relative measure of dispersion. It is frequently used in real estate or tax assessment applications.

$$COD = \frac{\left(\frac{\sum_{i=1}^{n}|x_i - \text{Median}|}{n}\right)}{\text{Median}}$$

Alpha-Trimmed Data Statistics

Alpha-Trimmed Data Statistics for Height

Percentage of Data Trimmed	N	Mean	Standard Deviation
5%	18	61.77778	7.448297
10%	16	61.5	6.552353
15%	14	61.35714	5.692196
25%	10	60.9	3.60401
35%	6	60.5	2.428992
45%	2	59.5	0.7071068

Percentage of Data Trimmed

We call *100g* the trimming percentage, the percent of data that is trimmed from each side of the sorted data. Thus, if *g* = 5%, for a sample size of 200, 10 observations are ignored from each side of the sorted array of data values. Note that our formulation allows fractional data values. Different trimming percentages are available, but 5% and 10% are the most common in practice.

N

This is the number of observations remaining after the trimming operation.

Mean

These are the alpha-trimmed means discussed by Hoaglin (1983, page 311). These are useful for quickly assessing the impact of outliers. You would like to see stability in these trimmed means after a small degree of trimming. The formula for the trimmed mean for 100g% trimming is

$$\bar{x}_{(\alpha)} = \frac{1}{n(1-2\alpha)}\left\{(1-r)\left[X_{(g+1)} + X_{(n-g)}\right] + \sum_{i=g+2}^{n-g-1} X_{(i)}\right\}$$

where $g = [\alpha n]$ and $r = \alpha n - g$.

Standard Deviation

This is the standard deviation of the observations that remain after the trimming. It can be used to evaluate changes in the standard deviation for different degrees of trimming. The formula for the trimmed standard deviation for *100g%* trimming is the standard formula for a weighted average using the weights given below.

$$a_i = \begin{cases} 0 & \text{if } i \le g \text{ or } i \ge n-g+1 \\ \dfrac{1-r}{n-2\alpha n} & \text{if } i = g+1 \text{ or } i = n-g \\ \dfrac{1}{n-2\alpha n} & \text{if } g+2 \le i \le n-g-1 \end{cases}$$

Mean Deviation Statistics

Mean Deviation Statistics for Height

Statistic	Base Function	Value	Standard Error
Mean Absolute Deviation from the Mean (MAD)	|X - Mean|	7.01	1.134273
Mean Absolute Deviation from the Median (MADM)	|X - Median|	6.8	
Mean Squared Deviation from the Mean (m2, 2nd Moment)	$(X - Mean)^2$	67.69	16.16531
Mean Cubed Deviation from the Mean (m3, 3rd Moment)	$(X - Mean)^3$	262.392	181.2807
Mean 4th-Power Deviation from the Mean (m4, 4th Moment)	$(X - Mean)^4$	9808.281	3522.41

Mean Absolute Deviation from the Mean (MAD): Value

This is a measure of dispersion, called the *mean absolute deviation* or the *mean deviation*. It is not affected by outliers as much as the standard deviation, since the differences from the mean are not squared. If the distribution for the variable of interest is normal, the mean deviation is approximately equal to 0.8 standard deviations.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Mean Absolute Deviation from the Mean (MAD): Standard Error

This is an estimate of the standard error of the *mean absolute deviation*.

$$SE_{MAD} = \sqrt{\frac{2s^2(n-1)}{\pi n^2}\left[\frac{\pi}{2} + (n^2 - 2n)^2 - n + \arcsin\left(\frac{1}{n-1}\right)\right]}$$

Mean Absolute Deviation from the Median (MADM): Value

This is an alternate formulation of the *mean deviation* above that is more robust to outliers since the median is used as the center point of the distribution.

$$MAD_{Robust} = \frac{\sum_{i=1}^{n} |x_i - \text{Median}|}{n}$$

Mean Squared Deviation from the Mean (m2, 2nd Moment): Value

This is the second moment about the mean, m_2.

$$m_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Mean Squared Deviation from the Mean (m2, 2nd Moment): Standard Error

This is the estimated standard error of the second moment.

Mean Cubed Deviation from the Mean (m3, 3rd Moment): Value

This is the third moment about the mean, m_3.

$$m_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$$

Mean Cubed Deviation from the Mean (m3, 3rd Moment): Standard Error

This is the estimated standard error of the third moment.

Mean 4th-Power Deviation from the Mean (m4, 4th Moment): Value

This is the fourth moment about the mean, m_4.

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}$$

Mean 4th-Power Deviation from the Mean (m4, 4th Moment): Standard Error

This is the estimated standard error of the fourth moment.

Percentiles

Percentiles of Height

		95% Confidence Interval Limits	
Percentile	**Value**	**Lower**	**Upper**
10th	52		
25th	56	51	59
50th	59.5	56	67
75th	70	60	76
90th	75.7		

Percentile Formula: Ave X(p[n+1])

This report gives the values for the 10th, 25th, 50th, 75th, and 90th percentiles. Of course, the 25th percentile is called the *first (lower) quartile*, the 50th percentile is the *median*, and the 75th percentile is called the *third (upper) quartile*. Confidence intervals are also given for the quartiles (25th, 50th, and 75th percentiles).

Percentile

The percentile reported on the row.

Value

These are the values of the specified percentiles. Note that the definition of a percentile depends on the type of percentile that was specified and is listed at the bottom of the report.

Lower and Upper 95% Confidence Interval Limits

These are the $100(1-\alpha)\%$ exact confidence interval limits for the population percentile. This confidence interval does not assume normality. Instead, it only assumes a random sample of *n* items from a continuous distribution. The interval is based on the equation:

$$1-\alpha = I_p(r, n-r+1) - I_p(n-r+1, r)$$

Here $I_p(a,b)$ is the integral of the incomplete beta function:

$$I_q(n-r+1, r) = \sum_{k=0}^{r-1} \binom{n}{k} p^k (1-p)^{n-k}$$

and $q = 1 - p$ and $I_p(a, b) = 1 - I_{1-p}(b, a)$.

Additional Percentiles

Additional Percentiles of Height

Percentile	Value	95% Confidence Interval Limits		Exact Confidence Level
		Lower	Upper	
99	79			
95	78.85			
90	75.7			
85	72.7	64	79	95.5319%
80	71	64	79	95.6328%
75	70	60	76	96.1823%
70	66.4	59	76	97.5218%
65	64.65	59	73	96.8303%
60	63.6	58	71	96.3010%
55	61.65	58	71	95.9722%
50	59.5	56	67	95.8611%
45	59	56	65	95.9722%
40	58.4	52	64	96.3010%
35	58	52	63	96.8303%
30	56.6	52	60	97.5218%
25	56	51	59	95.5904%
20	52.8	51	58	95.6328%
15	52	51	58	95.5319%
10	52			
5	51.05			
1	51			

Percentile Formula: Ave X(p[n+1])

This section gives a larger set of percentiles than was included in the Percentiles report. Use it when you need less-common percentiles.

Percentile

The percentile reported on the row.

Value

These are the values of the specified percentiles. Note that the definition of a percentile depends on the type of percentile that was specified and is listed at the bottom of the report.

Lower and Upper 95% Confidence Interval Limits

These are the $100(1-\alpha)\%$ exact confidence interval limits for the population percentile. This confidence interval does not assume normality. Instead, it only assumes a random sample of *n* items from a continuous distribution. The interval is based on the equation:

$$1-\alpha = I_p(r, n-r+1) - I_p(n-r+1, r)$$

Here $I_p(a,b)$ is the integral of the incomplete beta function:

$$I_q(n-r+1, r) = \sum_{k=0}^{r-1} \binom{n}{k} p^k (1-p)^{n-k}$$

and $q = 1 - p$ and $I_p(a, b) = 1 - I_{1-p}(b, a)$.

Exact Confidence Level

Because of the discrete nature of the confidence interval constructed above, **NCSS** finds an interval that is less than the specified alpha level. This column gives the actual confidence coefficient of the interval.

Normality Tests

Normality Tests for Height

Test Name	Test Statistic	P-Value	Critical Values α = 0.1	Critical Values α = 0.05	Reject Normality at α = 0.05?
Shapiro-Wilk	0.9374	0.21373			No
Anderson-Darling	0.4434	0.28629			No
Martinez-Iglewicz	1.0259		1.2162	1.3573	No
Kolmogorov-Smirnov	0.1482		0.1760	0.1920	No
D'Agostino Skewness	1.0367	0.29986	1.6450	1.9600	No
D'Agostino Kurtosis	-0.7855	0.43216	1.6450	1.9600	No
D'Agostino Omnibus	1.6918	0.42916	4.6050	5.9910	No

This section displays the results of seven tests of the hypothesis that the data come from the normal distribution. The Shapiro-Wilk and Anderson-Darling tests are usually considered as the best. The Kolmogorov-Smirnov test is included because of its historical popularity but is bettered in almost every way by the other tests.

Unfortunately, these tests have small statistical power (probability of detecting nonnormal data) unless the sample sizes are large, say over 100. Hence, if the decision is to reject, you can be reasonably certain that the data are not normal. However, if the decision is to accept, the situation is not as clear. If you have a sample size of 100 or more, you can reasonably assume that the actual distribution is closely approximated by the normal distribution. If your sample size is less than 100, all you know is that there was not enough evidence in your data to reject the normality assumption. In other words, the data might be nonnormal, you

just could not prove it. In this case, you must rely on the graphics and past experience to justify the normality assumption.

Shapiro-Wilk Test

This test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of n observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of squares of the observations about the sample mean. The test statistic W may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. Since these weights are asymptotically proportional to the corresponding expected normal order statistics, Shapiro-Wilk is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer Shapiro-Wilk is to one, the more normal the sample is.

The test was developed by Shapiro and Wilk (1965) for samples up to 20. **NCSS** uses the approximations suggested by Royston (1992) and Royston (1995) which allow unlimited sample sizes. Note that Royston only checked the results for sample sizes up to 5000 but indicated that he saw no reason larger sample sizes should not work.

The *p*-values for Shapiro-Wilk are valid for sample sizes greater than 3.

Shapiro-Wilk may not be as powerful as other tests when ties occur in your data.

The test is not calculated when a frequency variable is specified.

Anderson-Darling Test

This test, developed by Anderson and Darling (1954), is the most popular normality test that is based on EDF statistics. In some situations, it has been found to be as powerful as the Shapiro-Wilk test.

The test is not calculated when a frequency variable is specified.

Martinez-Iglewicz Test

This test for normality, developed by Martinez and Iglewicz (1981), is based on the median and a robust estimator of dispersion. They have shown that this test is very powerful for heavy-tailed symmetric distributions as well as a variety of other situations. A value of the test statistic that is close to one indicates that the distribution is normal. This test is recommended for exploratory data analysis by Hoaglin (1983). The formula for this test is:

$$I = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)s_{bi}^2}$$

where s_{bi}^2 is a biweight estimator of scale.

Martinez-Iglewicz Test Critical Values (α = 0.1 and α = 0.5)

The 10% and 5% critical values are given here. If the value of the test statistic is greater than this value, reject normality at that level of significance.

Kolmogorov-Smirnov Test

This test for normality is based on the maximum difference between the observed distribution and expected cumulative-normal distribution. Since it uses the sample mean and standard deviation to calculate

the expected normal distribution, the Lilliefors' adjustment is used. The smaller the maximum difference the more likely that the distribution is normal.

This test has been shown to be less powerful than the other tests in most situations. It is included because of its historical popularity.

Kolmogorov-Smirnov Test Critical Values (α = 0.1 and α = 0.5)

The 10% and 5% critical values are given here. If the value of the test statistic is greater than this value, reject normality at that level of significance. The critical values are the Lilliefors' adjusted values as given by Dallal (1986). If the test value is greater than the reject critical value, normality is rejected at that level of significance.

D'Agostino Skewness Test

D'Agostino (1990) describes a normality test based on the skewness coefficient, $\sqrt{b_1}$. Recall that because the normal distribution is symmetrical, $\sqrt{b_1}$ is equal to zero for normal data. Hence, a test can be developed to determine if the value of $\sqrt{b_1}$ is significantly different from zero. If it is, the data are obviously nonnormal. The statistic, z_s, is, under the null hypothesis of normality, approximately normally distributed. The computation of this statistic, which is restricted to sample sizes $n > 8$, is

$$z_s = d \ln\left(\frac{T}{a} + \sqrt{\left(\frac{T}{a}\right)^2 + 1}\right)$$

where

$$b_1 = \frac{m_3^2}{m_2^3}$$

$$T = \sqrt{b_1\left(\frac{(n+1)(n+3)}{6(n-2)}\right)}$$

$$C = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$W^2 = -1 + \sqrt{2(C-1)}$$

$$a = \sqrt{\frac{2}{W^2-1}}$$

$$d = \frac{1}{\sqrt{\ln(W)}}$$

D'Agostino Skewness Test: P-Value

This is the two-tail, significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

D'Agostino Kurtosis Test

D'Agostino (1990) describes a normality test based on the kurtosis coefficient, b_2. Recall that for the normal distribution, the theoretical value of b_2 is 3. Hence, a test can be developed to determine if the value of b_2 is significantly different from 3. If it is, the data are obviously nonnormal. The statistic, z_k, is, under the null hypothesis of normality, approximately normally distributed for sample sizes n>20. The calculation of this test proceeds as follows:

$$z_k = \frac{\left(1-\frac{2}{9A}\right) - \left(\frac{1-\frac{2}{A}}{1+G\sqrt{\frac{2}{A-4}}}\right)^{1/3}}{\sqrt{\frac{2}{9A}}}$$

where

$$b_2 = \frac{m_4}{m_2^2}$$

$$G = \frac{b_2 - \left(\frac{3n-3}{n+1}\right)}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}}$$

$$E = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$A = 6 + \frac{8}{E}\left(\frac{2}{E} + \sqrt{1+\frac{4}{E^2}}\right)$$

D'Agostino Kurtosis Test: P-Value

This is the two-tail significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

D'Agostino Omnibus Test

D'Agostino (1990) describes a normality test that combines the tests for skewness and kurtosis. The statistic, K^2, is approximately distributed as a chi-square with two degrees of freedom. After calculated z_s and z_k, calculate K^2 as follows:

$$K^2 = z_s^2 + z_k^2$$

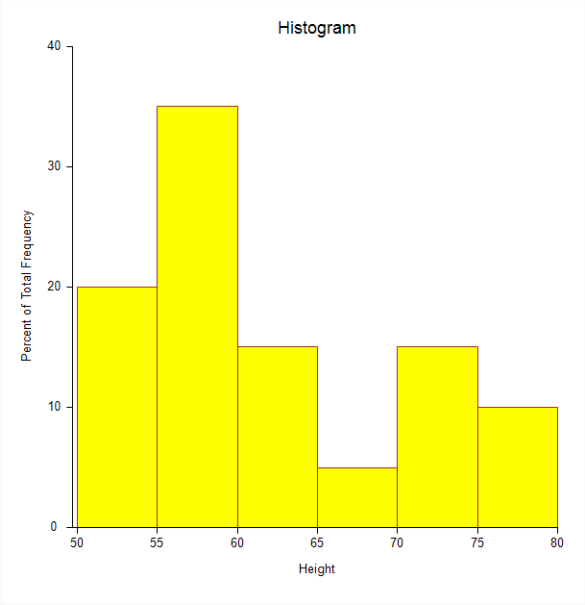
D'Agostino Omnibus Test: P-Value

This is the significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

Histogram

The following plot shows a histogram of the data.

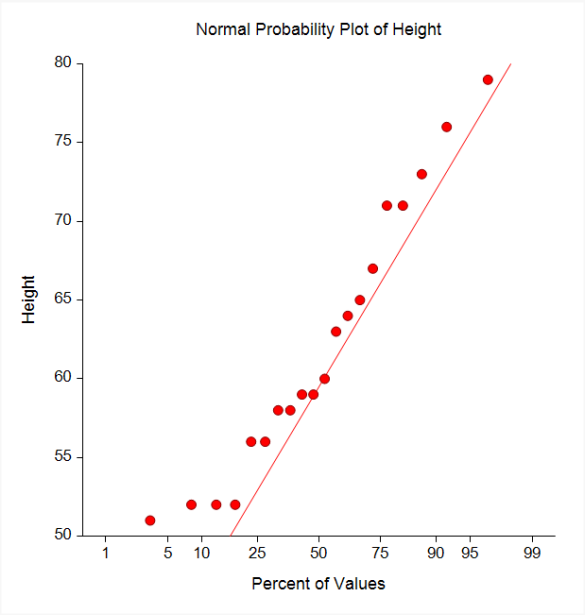
Plots of Height



The histogram is a traditional way of displaying the shape of a group of data. It is constructed from a frequency distribution, where choices on the number of bins and bin width have been made. These choices can drastically affect the shape of the histogram. The ideal shape to look for in the case of normality is a bell-shaped distribution.

Normal Probability Plot

Plots of Height



This is a plot of the inverse of the standard normal cumulative versus the ordered observations. If the underlying distribution of the data is normal, the points will fall along a straight line. Deviations from this line correspond to various types of nonnormality. Stragglers at either end of the normal probability plot indicate outliers. Curvature at both ends of the plot indicates long or short distribution tails. Convex, or concave, curvature indicates a lack of symmetry. Gaps, plateaus, or segmentation in the plot indicate certain phenomenon that need closer scrutiny.

Confidence bands may be added to serve as a visual reference for departures from normality. If any of the observations fall outside the confidence bands, the data are not normal. The numerical normality tests will usually confirm this fact statistically. If only one observation falls outside the confidence limits, it may be an outlier. Note that these confidence bands are based on large sample formulas. They may not be accurate for small samples (less than 30).

Stem and Leaf Plot

Stem and Leaf Plot of Height

Depth	Stem		Leaf
4	5*	|	1222
10	.	|	668899
10	6*	|	034
7	.	|	57
5	7*	|	113
2	.	|	69

Unit = 1 Example: 1 | 2 Represents 12

The stem and leaf plot is a type of histogram which retains much of the identity of the original data. It is useful for finding data-entry errors as well as for studying the distribution of a variable.

Depth

This is the cumulative number of leaves, counting in from the nearest end.

Stem

The stem is the first digit of the actual number. For example, the stem of the number 523 is 5 and the stem of 0.0325 is 3. This is modified appropriately if the batch contains numbers of different orders of magnitude. The largest order of magnitude is used in determining the stem. Depending upon the number of leaves, a stem may be divided into two or more sub-stems. A special set of symbols is then used to mark the stems.

The star (*) represents numbers in the range of zero to four, while the period (.) represents numbers in the range of five to nine.

Leaf

The leaf is the second digit of the actual number. For example, the leaf of the number 523 is 2 and the leaf of 0.0325 is 2. This is modified appropriately if the batch contains numbers of different orders of magnitude. The largest order of magnitude is used in determining the leaf.

Unit

This line at the bottom indicates how the data were scaled to make the plot.