Exploratory Data Analysis
(EDA)

Introduction

A Need to Explore Your Data

The first step of data analysis should always be a detailed examination of the data. The examination of your data is called Exploratory Data Analysis (EDA).

Whether the problem you are solving is simple or complex, whether you're planning to do a t test or a multivariate repeated measures analysis of variance (ANOVA), you should first take a careful look at the data. In other words, you should do an EDA on the data.

Most researchers don't or never do an EDA on their data. They place too much trust in the confirmatory data analysis (statistical analysis). The data are analyzed straight away using specific statistical techniques or methods (such as frequency, mean, standard deviation, variance or the various confirmatory data analysis methods such as t test, ANOVA, multiple regression etc).

They relied on descriptive statistics and confirmatory data analysis exclusively. The data are not explored to see the assumptions of the selected test are met or violated and what other patterns might exist. Other modes of analysis might yield greater insight about your data or more appropriate.

The underlying assumption of EDA is that the more one knows about the data, the more effectively data can be used to develop, test and refine theory.

Thus a researcher should learn as much as possible about a variable or set of variables before using the data to test theories of social science relationships.

Reasons for using the Explore procedure

1. To detect or identify (scan your data set for) mistakes or errors.

   Data must make a hazardous journey before finding a final rest in a data file in your computer. First, a measurement is made or a response elicited, sometimes with a faulty instrument or by a careless enumerator (interviewer). Then it is coded and entered onto a data file at a later time. Errors can be introduced at any of these steps.

   Some errors are easy to spot. For example, forgetting to declare a value as missing, using invalid code/value labels, or entering the value 609 for age will be apparent from a frequency table (*it is highly recommended that you run a frequency procedure for all your variables first).* Other errors, such as entering an age of 54 instead of 45, may be difficult, if not impossible, to spot. Unless your first step is carefully check your data for mistakes, errors may contaminate all your analyses.

   If you put in a lot of junk in your data file then what you get out are also junk (Junk in, junk out). So scan your data set for unwanted junks/errors/mistakes.

2. Data screening. Data screening may show that you have unusual values, extreme values, gaps in the data, or other peculiarities.
   For example:

   If the distribution of data values reveals a gap--that is, a range where no value occur then you must ask why.

   If some values are extreme (far removed from the other values), you must look for reasons.

   If the pattern of numbers is strange, you must determine why.

   If you see unexpected variability in the data you must look for possible explanations; perhaps there are additional variables that may explain it.

3. Outlier identification,

4. Description,

5. Assumptions checking,
   The following assumptions must hold true to use parametric tests.

   A. The populations from which the samples are drawn are (approximately) normally distributed.

   B. The populations from which the samples are drawn have the same variance (or standard deviation).

   C. The samples drawn from different populations are random and independent.

6. Characterizing differences among subpopulations (groups of cases), and

7. Exploring the data can help to determine whether the statistical techniques you are considering for data analysis are appropriate. The exploration may indicate that you need to transform the data if the technique requires a normal distribution. Or, you may decide that you need nonparametric tests.

   In other words, it helps to prepare for hypothesis testing. Looking at the distribution of the values is also important for evaluating the appropriateness of the statistical techniques you are planning to use for hypotheses testing or model building. Perhaps the data must ***be transformed*** or ***reexpressed*** so that the distribution is approximately normal or so that the variances in the groups are similar (critical for parametric test); or perhaps nonparametric technique is needed or more appropriate.

   Data analysis has often been compared to detective work. Before the actual trial of a hypothesis, there is much evidence to be gathered and sifted/screened. Based on the clues, the hypotheses or models may be altered, or methods for testing may have to be changed or use a more resistant summary statistics such as median, trimmed mean, m-estimators for location; midspread for spread.

EDA is based on two principles: Skepticism and Openness

EDA seeks to maximize what is learned from the data. This requires adherence to two principles: ***skepticism*** and ***openness.***

Skepticism

You should be *skeptical* of measures that summarize your data since they can sometimes ***conceal*** or even ***misrepresent*** what may be the most informative aspects of the data. For instance, for skewed and multiple peaks distributions or if extremes values exist in the distribution then mean does not provide a good measure of central tendencies (MCT) because it misrepresent the data set.

Skepticism is an *awareness* that even widely used statistical techniques may have unreasonable hidden assumptions about the nature of the data at hand.

Openness

One should be *open to unanticipated patterns* in the data since they can be the most revealing outcomes of the analysis.

The researcher should remain *open to possibilities* that he or she does not expect to find.

Fundamental Concepts in EDA

Data = smooth + rough

All data analysis is basically the partitioning of data into the Smooth and the rough.

Statistical techniques for obtaining explained and unexplained variances, between-group and within-group sums of squares, observed and expected cell frequencies, and so on are examples of this basic process.

Smooth

The smooth is the underlying, ***simplified structure*** of a set of observations or data.

It may be represented by a straight line describing the relationship between two variables or by a curve describing the distribution of a simple variable. In either case the smooth is an important feature of the data.

It is the general *shape of a distribution* or the general *shape of a relationship.* It is the regularity or pattern in the data.

Since the data will almost never conform exactly to the smooth, the smooth must be ***extracted*** from the data. What is left behind is the rough, the *deviations from the smooth.*

Traditionally, social science data analysis has concentrated on the smooth. The rough is not treated either as *an aid in generating the smooth* or *as a component of the data* in its own right.

Rough

Rough is just as important as the smooth. Why?

1. Because smoothing is sometimes an iterative or repetitive process proceeding through the examination of successive roughs for additional smooth, and

2. Because points-that do not fit a model are often as instructive as those that do.

3. What is desirable is a rough that has no smooth; that is, the rough should contain no additional pattern or structure.

If the rough does contain additional structure not removed by the smooth, it is not rough enough and further smoothing should take place.

What distinguishes EDA from confirmatory data analysis is the willingness to examine the rough for additional smooth and the openness to *alternative models of relationships* that will remove additional smooth from the rough until it is rough enough.

The principle of openness takes two forms when extracting the smooth from the rough.

First, instead of imposing a hypothesized model of the smooth on the data, ***a model of the smooth is generated from the data.*** In other words, the data are explored to discover the smooth, and models generated from the data can then be compared to models specified by the theory. The more similar they are, the more the data confirm the theory.

For example, when looking at the relationship between two variables, the researcher should not simply fit a linear model to the data. Instead he should look at a summary statistics that compares the amount of rough to the amount of smooth and then test the statistical significance of the statistic to see if the ratio could have occurred by chance. The statistic might be significant, but the relationship in the data might not be linear, i.e., the smooth might not form a straight line or anything like it. Or the statistic might not be significance because the smooth is not a straight line. In either case, the researcher would have failed to discover something important about the data, namely the relationship between the two variables is not what he or she thought it was. Only by exploring data is it possible to discover what is not expected such as a nonlinear relationship in this case. And only by exploring the data is it possible to test fully a theory that specifies a relationship of a particular form. In short, one should be open to alternative models of relationships between variables.

The second form that openness takes is ***reexpression.*** The scale on which a variable was originally observed and recorded is not the only one on which it can be expressed.

In fact, reexpressing the original values into a different scale of measurement may prove to be more workable. For example by reexpressing the values in terms of log (natural logarithm), power transformations (square/cube/square root/reciprocal/reciprocal of the square root), it may be possible to extract additional smooth from the data or to explore data for unanticipated pattern.

The principle of skepticism also takes two forms when extracting the smooth from the data:

First, a reliance on visual representations of data, and second, the use of resistant statistics (median, midspread, m-estimators, etc).

Because of skepticism toward statistical summaries of data, major emphasis in the EDA is placed on visual representations of data.

The emphasis of EDA is upon using visual displays to reveal vital information about the data being examined. It thus makes *extensive* use of visual displays.

The reasons for this are:

1. The shape *(normal, skewed, multipeaks, outliers at the extremes or gap within the distribution of values)* of a distribution is at least as important as the location *(measured by mean, median or mode)* and spread/variability/dispersion of cases *(standard deviation, variance, sum of squared).* Excessive reliance on measures of location and spread can hide important differences in the way they are distributed.

2. Visual representations are superior to purely numeric representations for discovering the characteristic shape of a distribution. Shape is physical characteristic best communicated by visual techniques.

3. The choice of summary statistics to describe the data for a single variable should be dependent upon the appropriateness of the statistics for the shape of the distribution. When distributions depart markedly from the normal distribution, the more distribution-free measures ate to be preferred.

Data Requirements to Run EDA Procedure

The Explore procedure can be used for quantitative variables (interval- or ratio-level measurements). A factor variable (used to break the data into groups of cases) should have a reasonable number of distinct values (categories). These values may be short string or numeric. The case label variable, used to label outliers in boxplots, can be short string, long string (first 15 characters), or numeric.

Running EDA: A simple example

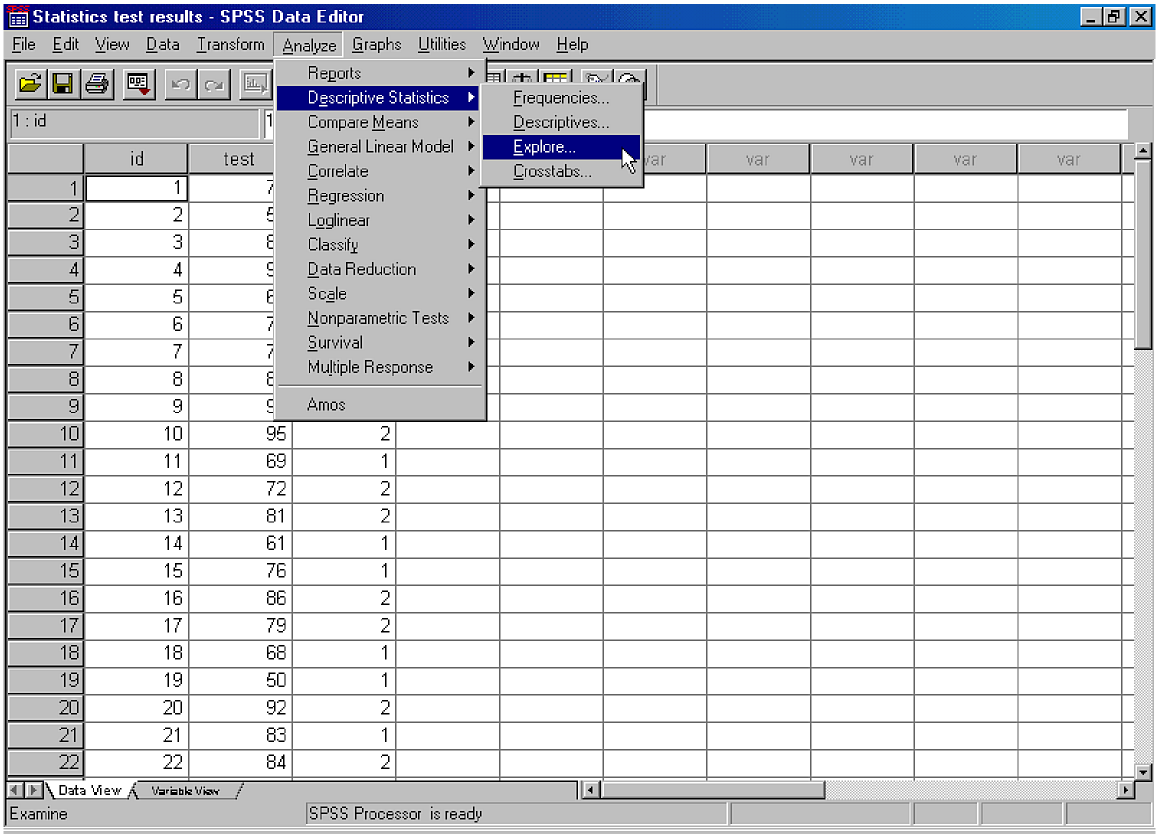
How to Explore Your Data

1. From the main menus choose:

File
Open
Data…
Open file: Statistics Test Results

2. From the main menus choose:

Analyze
Descriptive Statistics
Explore...

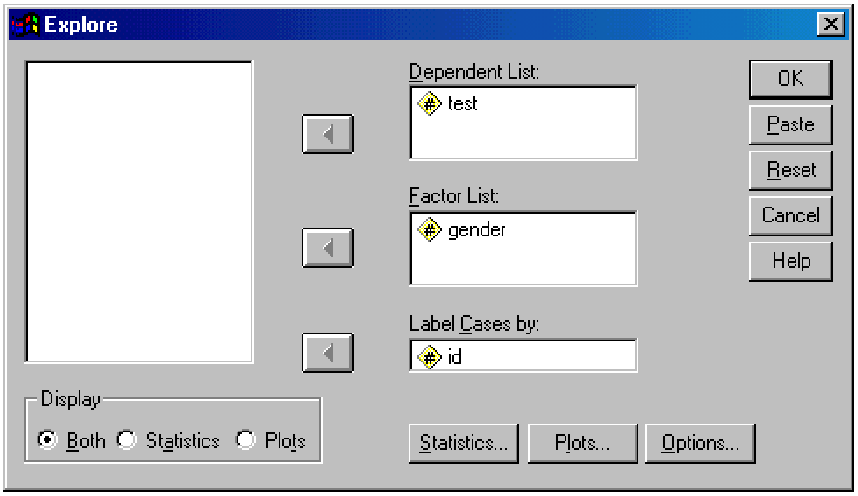


	id	test	
1	1	7	
2	2	5	
3	3	8	
4	4	9	
5	5	6	
6	6	7	
7	7	7	
8	8	8	
9	9	9	
10	10	95	2
11	11	69	1
12	12	72	2
13	13	81	2
14	14	61	1
15	15	76	1
16	16	86	2
17	17	79	2
18	18	68	1
19	19	50	1
20	20	92	2
21	21	83	1
22	22	84	2

Examine

SPSS Processor is ready

Than an explore dialogue box like this will appear



Dependent List: test

Factor List: gender

Label Cases by: id

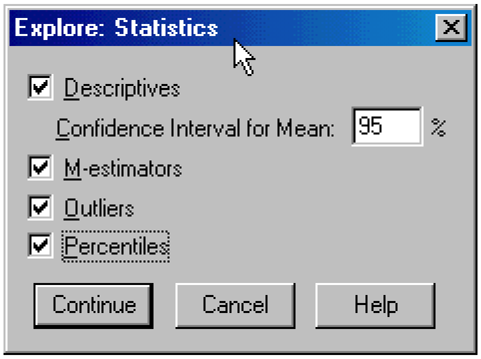
Display: ● Both ○ Statistics ○ Plots

Select the dependent variable (TEST) and place it in the Dependent List box.

Next select the factor variable whose values will define groups of cases (GENDER) and place it in the Factor List box.

Select an identification variable to label cases (ID) and place it in the "Label Cases by" box.

3. Click **Statistics** on the bottom of the main explore dialogue box and an Explore: Statistics dialogue box will appear.



☑ Descriptives

Confidence Interval for Mean: 95 %

☑ M-estimators

☑ Outliers

☑ Percentiles

Check the following Statistics: Descriptives, M-estimators, Outliers and Percentiles and,

Click Continue.

Descriptives: A descriptive statistics table will be displayed with the following statistics: arithmetic mean, median, 5% trimmed mean, standard error, variance, standard deviation, minimum, maximum, range, interquartile range, skewness and kurtosis and their standard errors, confidence interval for the mean (and specified confidence level),

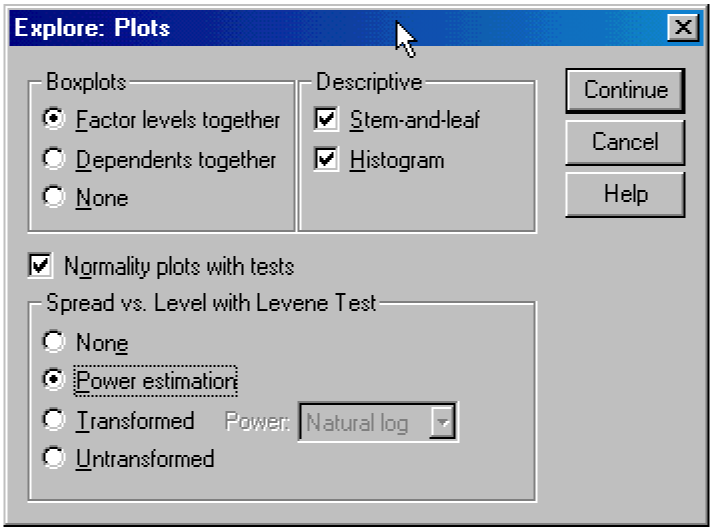
M-estimators: An M-estimators table will be given with several resistant statistics such as Huber's M-estimator, Andrew's wave estimator, Hampel's redescending M-estimator and Tukey's biweight estimator.

Outliers: A table of outlying values (outlier and extreme values) will be displayed. By default, it displays the five largest and five smallest values.

Percentiles: The Explore procedure in SPSS also displays a range of percentiles. By default, the percentiles table displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles.

4. Click **Plots** of the main explore dialogue box and this opens the Explore: Plots dialogue box shown below.

In the Explore: Plot dialogue box, a number of selections need to be carried out under the following headings: Boxplots, Descriptive, Normality plots with test and Spread vs. Level with Levene Test.



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
☑ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

Spread vs. Level with Levene Test
○ None
● Power estimation
○ Transformed   Power: Natural log
○ Untransformed

Boxplots

Choose "Factor levels together".

Two types of boxplot can generated from Explore procedure in SPSS namely factor levels together and dependents together.

Factor levels together generates a separate display for each dependent variable. Within a display, boxplots are shown for each of the groups defined by a factor variable.

Dependents together generates a separate display for each group defined by a factor variable. Within a display, boxplots are shown side by side for each dependent variable. This display is particularly useful when the different variables represent a single characteristic measured at different times.

Descriptive

Check the “Stem-and- leaf” and “Histogram” boxes.

Normality plots with tests

Check √ the “Normality plots with tests” box.

The Normality plots display both the normal probability and detrended normal probability plots.

The Normality statistics tests will display (a) the Kolmogorov-Smirnov statistic with a Lilliefors significance level for testing normality and (b) the Shapiro-Wilk statistic for samples with 50 or fewer observations.

Spread-versus level with Levene Test

Check the Spread-versus level with Levene Test box

The Spread-versus-level plot: For all spread-versus-level plot, the slope of the regression line is displayed. A spread-versus-level plot helps determine the power for a transformation to stabilize (make more equal) variances across groups.

Levene test for homogeneity of variance: If you select a transformation, Levene test is based on the transformed data. If no factor variable is selected, spread-versus-level plots are not produced. Power estimation produces a plot of the natural logs of the interquartile ranges against the natural logs of the medians for all cells, as well as an estimate of the power transformation for achieving equal variances in the cells.

Transformed allows you select one of the Power alternatives, perhaps following the recommendation from Power estimation, and produces plots of transformed data. The interquartile range and median of the transformed data are plotted. Untransformed produces plots of the raw data. This is equivalent to a transformation with a power of 1.

1. Descriptive Statistics

These measures of central tendency and dispersion are displayed by default. Measures of central tendency indicate the location of the distribution; they include the mean, median, and 5% trimmed mean. Measures of dispersion show the dissimilarity of the values; these include standard error, variance, standard deviation, minimum, maximum, range, and interquartile range. The descriptive statistics also include measures of the shape of the distribution, skewness and kurtosis are displayed with their standard errors. The 95% level confidence interval for the mean is also displayed; you can specify a different confidence level.

The data set that we are going to use in this EDA example is obtained from a statistic test taken by 30 college students. The scores breakdown by gender are listed in Table 1.

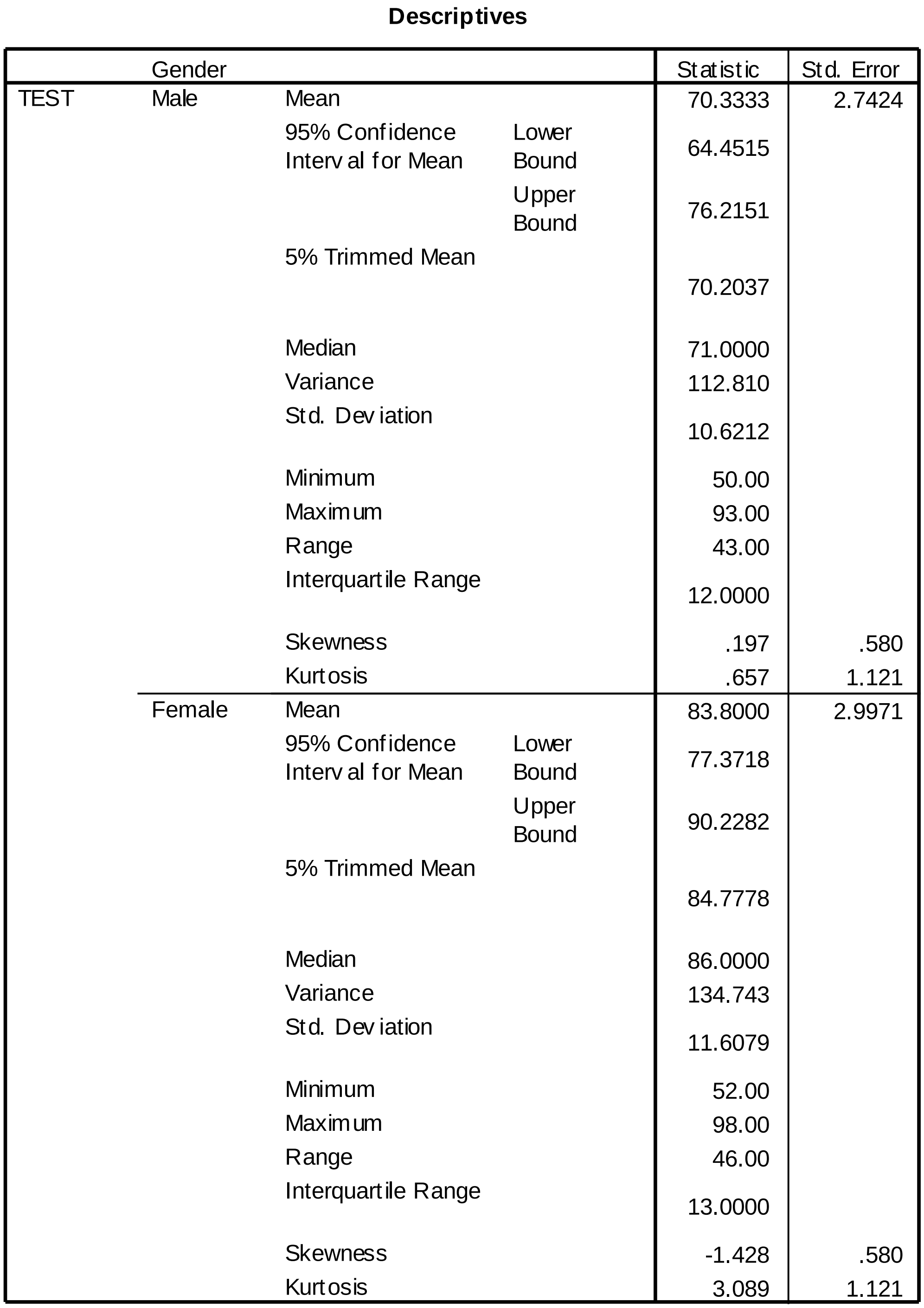
Table 1: Statistics Test Result (TEST) by Gender

Statistics Test Result

		TEST	Gender
1		75.00	Male
2		52.00	Female
3		80.00	Male
4		96.00	Female
5		65.00	Male
6		79.00	Female
7		71.00	Male
8		87.00	Female
9		93.00	Male
10		95.00	Female
11		69.00	Male
12		72.00	Female
13		81.00	Female
14		61.00	Male
15		76.00	Male
16		86.00	Female
17		79.00	Female
18		68.00	Male
19		50.00	Male
20		92.00	Female
21		83.00	Male
22		84.00	Female
23		77.00	Female
24		64.00	Male
25		71.00	Male
26		87.00	Female
27		72.00	Male
28		92.00	Female
29		57.00	Male
30		98.00	Female
Total	N	30	30

Table 2 gives a summary of some of the descriptive statistics for the statistic test results by gender

Table 2: Descriptive Statistics for the Statistic Test Results by Gender



Descriptives

	Gender			Statistic	Std. Error
TEST	Male	Mean		70.3333	2.7424
		95% Confidence Interval for Mean	Lower Bound	64.4515	
			Upper Bound	76.2151	
		5% Trimmed Mean		70.2037	
		Median		71.0000	
		Variance		112.810	
		Std. Deviation		10.6212	
		Minimum		50.00	
		Maximum		93.00	
		Range		43.00	
		Interquartile Range		12.0000	
		Skewness		.197	.580
		Kurtosis		.657	1.121
	Female	Mean		83.8000	2.9971
		95% Confidence Interval for Mean	Lower Bound	77.3718	
			Upper Bound	90.2282	
		5% Trimmed Mean		84.7778	
		Median		86.0000	
		Variance		134.743	
		Std. Deviation		11.6079	
		Minimum		52.00	
		Maximum		98.00	
		Range		46.00	
		Interquartile Range		13.0000	
		Skewness		-1.428	.580
		Kurtosis		3.089	1.121

How to Calculate CI_{95}?

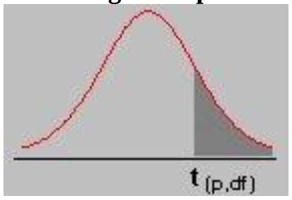
Formula:
CI_{95} = Mean ± ($t_{critical}$) (s_{Mean})

Where:
Mean = the sample mean (x-bar).

($t_{critical}$) = the appropriate t-value associated with the CI_{95} and is dependent upon the number of degree of freedom (Read the *t*-table for α/2, df = n-1).

s_{Mean} = the standard error of the mean

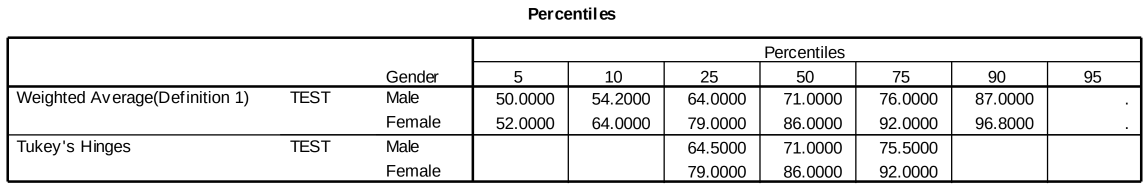
t table with right tail probabilities



df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405

15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Table 3: Percentiles for the Statistic Test Results by gender

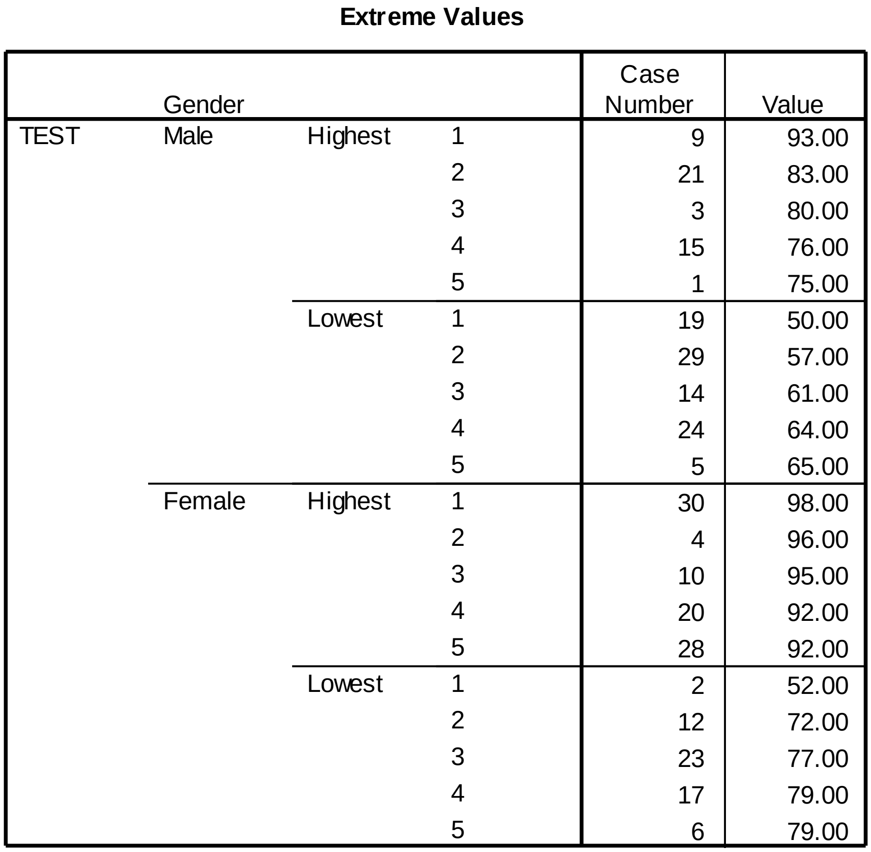


Percentiles

			Percentiles						
		Gender	5	10	25	50	75	90	95
Weighted Average(Definition 1)	TEST	Male	50.0000	54.2000	64.0000	71.0000	76.0000	87.0000	.
		Female	52.0000	64.0000	79.0000	86.0000	92.0000	96.8000	.
Tukey's Hinges	TEST	Male			64.5000	71.0000	75.5000		
		Female			79.0000	86.0000	92.0000		

In addition Explore procedure of SPSS also displays the five largest and five smallest values (extreme values), with case labels (see Table 4).

Table 4: Extreme Values for the Statistic Test Results by gender



Extreme Values

	Gender			Case Number	Value
TEST	Male	Highest	1	9	93.00
			2	21	83.00
			3	3	80.00
			4	15	76.00
			5	1	75.00
		Lowest	1	19	50.00
			2	29	57.00
			3	14	61.00
			4	24	64.00
			5	5	65.00
	Female	Highest	1	30	98.00
			2	4	96.00
			3	10	95.00
			4	20	92.00
			5	28	92.00
		Lowest	1	2	52.00
			2	12	72.00
			3	23	77.00
			4	17	79.00
			5	6	79.00

2. Resistant Statistics

We often use the *arithmetic mean* to estimate central tendency or location.

This mean is heavily influenced by outliers (very large or very small value can change the mean dramatically). Thus it is a nonresistant measure.

The *median*, on the other hand, is insensitive to outliers; addition or removal of extreme values has little effect on it. Thus the median is called a resistant measure, since its value depends on the main body of the data and not on outliers. The median value is obtained in Table 2.

Other better estimators of location are called robust estimators. Robust estimators depend on simple, fairly nonrestrictive assumptions about the underlying distribution and are not sensitive to these assumptions.

Two Robust Estimators of Central Tendency are:

The Trimmed Mean, and

M-Estimators.

The Trimmed Mean
The trimmed mean is a simple robust estimator of location which is obtained by "trimming" the data to *exclude values that are far removed from the others.*

For example a 5% trimmed mean disregards the smallest 5% and the largest 5% of all observations. The estimate is based on only 90% of data values that are in the middle. Trimmed mean is provided in Table 2.

What is the advantage of a trimmed mean?
Like the median, it results in an estimate that is not influenced by extreme values.

However, unlike the median it is not based solely on a single value, or two values, that are in the middle.

It is based on a much larger number of middle values. In general, a trimmed mean makes better use of the data than does the median.

M-Estimators

In calculating the trimmed mean, we treated observations that are far from most of the others by excluding them altogether.

A less extreme alternative is to include them but give them smaller weights than cases closer to the center, that is by using M-estimator, or generalized *maximum-likelihood estimator.*
Robust provides alternatives to the sample mean and median for estimating the center of location.

Examples of M-estimators provided by the Explore procedure in SPSS are Huber's, Hampel's, Tukey's, and Andrew's M-estimators. If there are no outliers, then no m-estimators are provided by SPSS output (see Table 5).

Table 5: M-Estimators for the Statistic Test Results by gender

M-Estimators

	Gender	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
TEST	Male	70.1518	69.9863	70.1233	69.9591
	Female	85.3664	86.1350	85.5799	86.1340

[a] The weighting constant is 1.339.

[b] The weighting constant is 4.685.

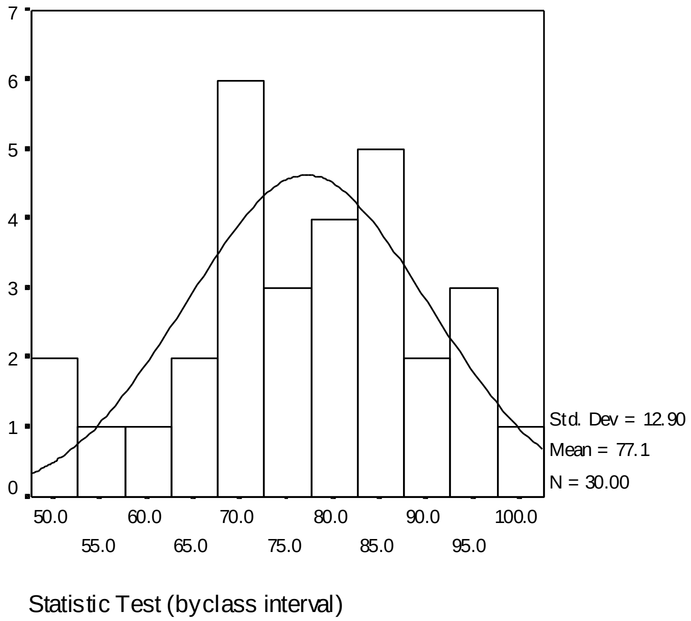
[c] The weighting constants are 1.700, 3.400, and 8.500

[d] The weighting constant is 1.340*pi.

3. The Histogram

The histogram is commonly used to represent data graphically. The range of observed values is subdivided into equal interval, and number of cases in each interval is obtained. Each bar in the histogram represents the number of cases (frequencies) with values within the interval (see Figure 1).

Figure1: Histogram with Normal Curve Display for the Statistic Test Results by gender



4. The Stem-and-Leaf Plot

In a stem-and-leaf display of quantitative data, each variable value is divided into two portions/parts--a stem and a leaf. Then the leaves for each stem are shown separately in a display. An advantage of this display over a frequency distribution or histogram is that, we do not lose information on individual observations. It is constructed only for quantitative data.

Figure 2: The stem-and leaf plot for the Statistic Test Results by gender

```
Frequency     Stem &  Leaf     (Class interval with a width of 5)

    2.00         5 .  02       (50-54)
```

```
     1.00        5 .  7        (55-59)
     2.00        6 .  14       (60-64)
     3.00        6 .  589      (65-69)
     4.00        7 .  1122     (70-74)
     5.00        7 .  56799    (75-79)
     4.00        8 .  0134     (80-84)
     3.00        8 .  677      (85-89)
     3.00        9 .  223      (90-94)
     3.00        9 .  568      (95-100)

 Stem width:      10.00
 Each leaf:        1 case(s)
```

Both the histogram and the stem-and-leaf plot provide valuable or useful information about the shape of a distribution for univariate variables. We can see how tightly cases cluster together. We can see if there is a single peak or several peaks. We can determine if there are extreme values.

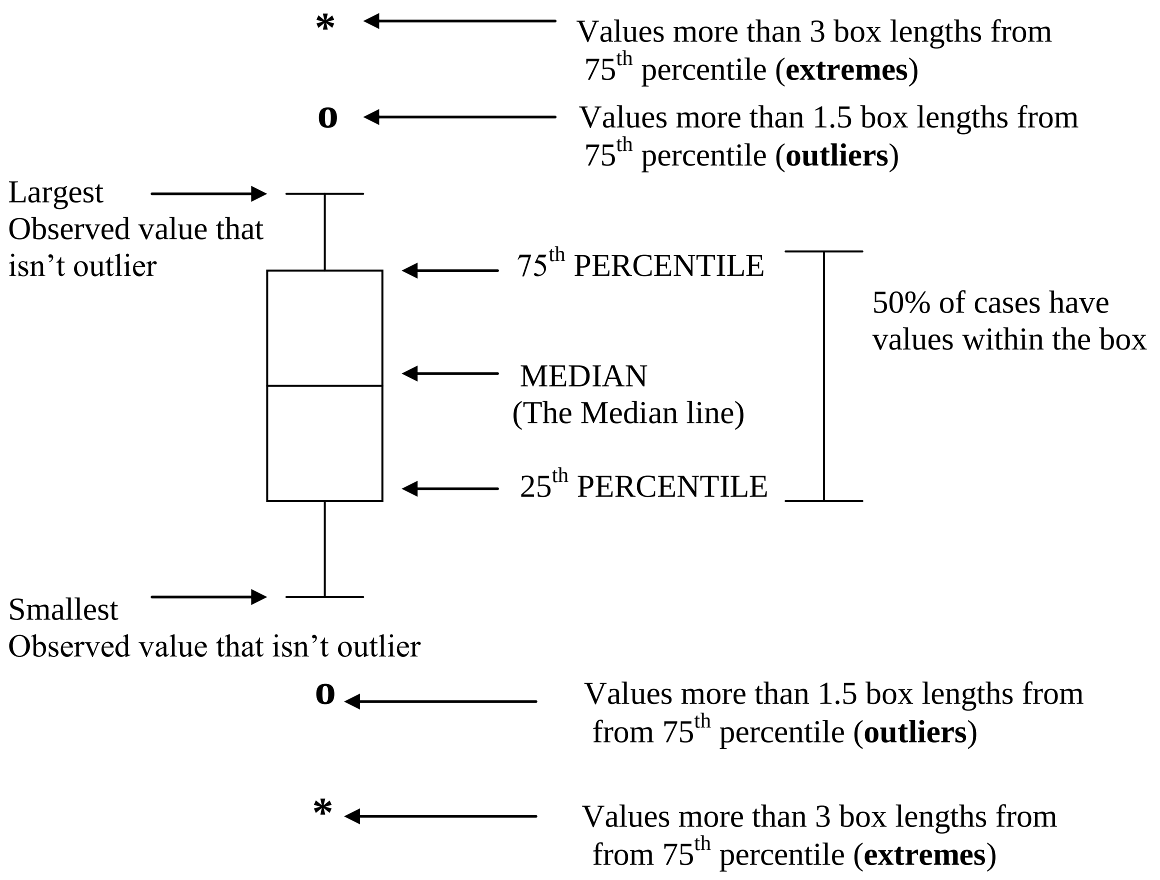
5. The Box-and-Whisker Plot (Boxplot)

A display that further summarizes information about the distribution of the a data set is the box-and-whisker plot (a boxplot). Instead of plotting the actual values, a boxplot displays summary statistics for the distribution. It plots the median, 25th percentile, the 75th percentile, and values that are far removed from the rest (extreme values).

The Figure 3 shows an annotated sketch of a boxplot. The lower boundary of the box is the 25th percentile and the upper boundary is the 75th percentile. (These percentiles, sometimes called Tukey's hinges). The horizontal line inside the box represents the median. **Fifty percent** of the cases have values within the box. The length of the box corresponds to the interquartile range, which is the difference between the 75th and 25th percentile.

The boxplot includes two categories of cases with outlying values. Cases with values that are more than 3 box-length from the upper or lower edge of the box are called **extreme** values. On the boxplot, these are designated an asterisk (*). Cases with values that are between 1.5 and 3 box-length from the upper and lower edge of the box are called **outliers** and are designated with a circle. The largest and smallest observed values that aren't **outliers** are also shown. Lines are drawn from the ends of the box to these values. (These lines are sometimes called whiskers and the plot is called a box-and-whiskers plot).

Figure 3: An annotated sketch of a box-and-whisker plot



* Values more than 3 box lengths from 75th percentile (**extremes**)

0 Values more than 1.5 box lengths from 75th percentile (**outliers**)

Largest Observed value that isn't outlier

75th PERCENTILE

50% of cases have values within the box

MEDIAN (The Median line)

25th PERCENTILE

Smallest Observed value that isn't outlier

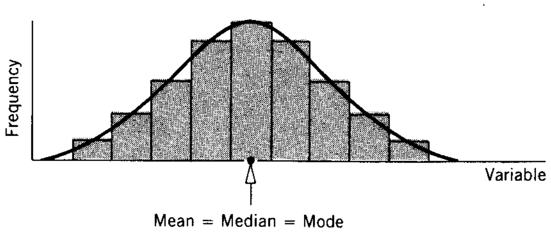
0 Values more than 1.5 box lengths from from 75th percentile (**outliers**)

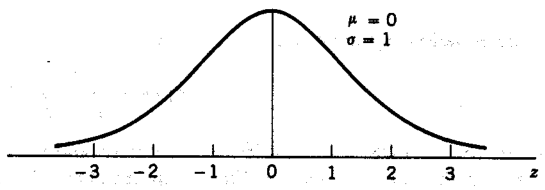
* Values more than 3 box lengths from from 75th percentile (**extremes**)

What can you tell about your data from the box-and-whisker plot?

1. From the median line, you can determine the central tendency, or location.

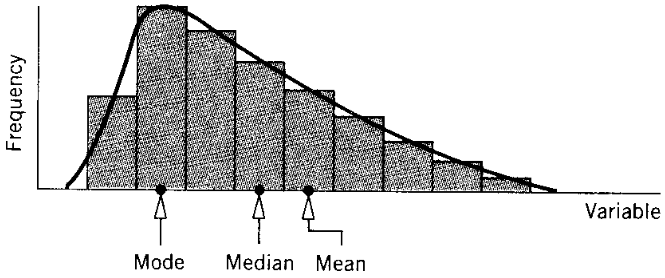
- If the median line is at or near the center of the box, then the observed values are normally distributed.



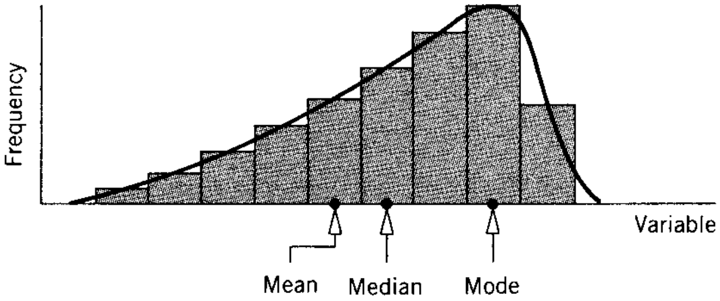


- If the median line is not in the center of the box, you know that the observed values are skewed.

- If the median line is closer to the bottom of the box than to the top, the data are positively skewed (Skewed to the right). Mean > median > mode.



- If the median line is closer to the top of the box than to the bottom, the opposite is true: the distribution is negatively skewed (Skewed to the left). Mean < median < mode.



2. From the length of the box, you can determine the spread, or variability, of your observation.

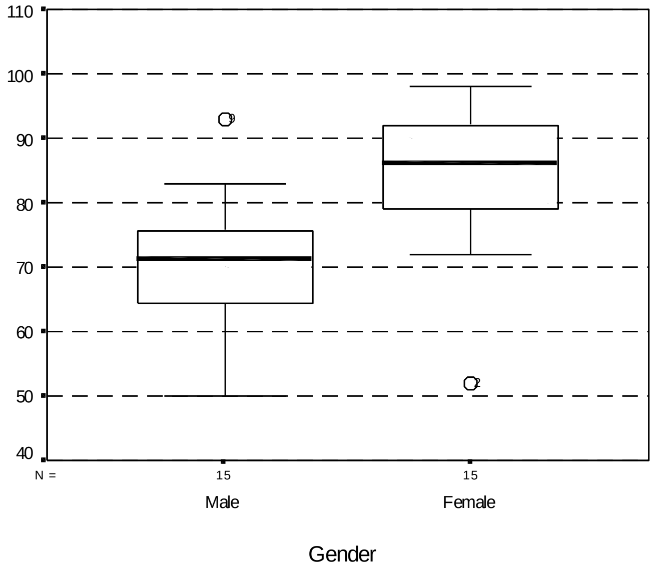
- A tall box indicates a high variability among the values observed.

- A short or compressed box shows a low spread or little variability among the values observed.

3. Boxplots are particularly useful for comparing the distribution of values in several groups. (e.g. see Figure 4 below).

It contains boxplots of the statistic test result data by gender. From these two boxplots, you can see that both the male and female students have similar distributions for their statistic test results. They are both slightly negatively skewed (median line is closer to the top of the box). The statistic test results of the female students have higher variability/larger spread (slightly taller box) compared to the male students. The female students have much higher median statistic test results than the male (the female students median line is higher than the male students). Both groups have one outlier each. The male student with an outlier (the outlier value is 92) is case no. 9 while the female student with an outlier (the outlier value is 52) is case no. 2.

Figure 4: Boxplots for the Statistic Test Result by Gender



6. Normality Plots

We often want to examine the assumption that our data come from a normal distribution. Two ways to do this are with a normal probability plot and a detrended normal plot.

A Normal Probability Plot

In a normal probability plot, each observed value is paired with its expected value from the normal distribution. (The expected value from the normal distribution is based on the number of cases in the sample and the rank order of the case in the sample)

If the sample is from a normal distribution, we expect that the points will fall more or less on a straight line. (The points cluster around a straight line).

Figure 5: Normal probability plot for the Statistic Test Result for Female Students

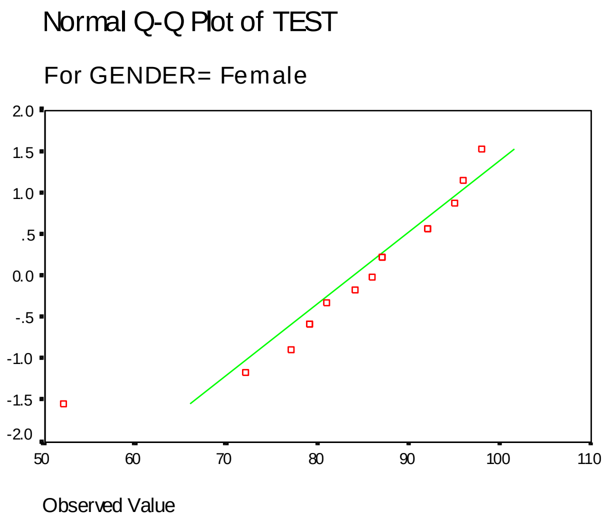
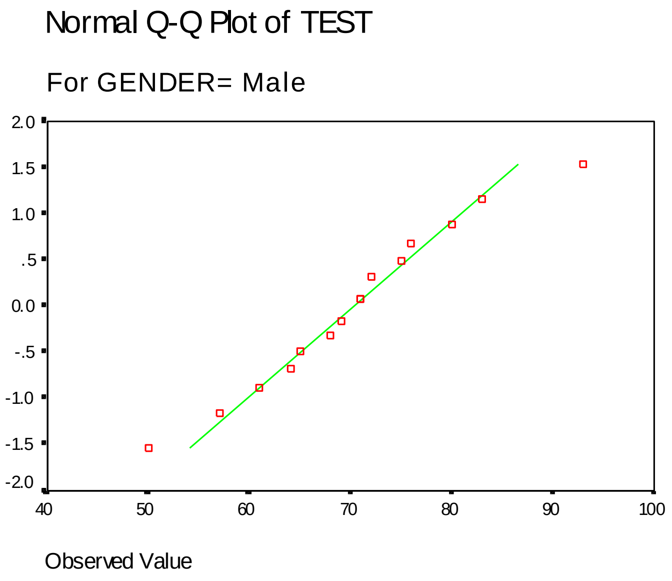


Figure 6: Normal probability plot for the Statistic Test Result for Male Students



A Detrended Normal Plot

This is a plot of the actual deviation of the points from a straight line. Use detrended normal probability plot as an aid to characterize how the values depart from a normal distribution. In this display the difference between the usual 2 scores for each case and its expected score under normality are plotted against the data values.

If the sample is from a normal population, the points should cluster around a horizontal line through 0, and there should be no pattern.

A striking pattern suggests departure from normality.

Figure 7: Detrended Normal plot for the Statistic Test Result for Female Students

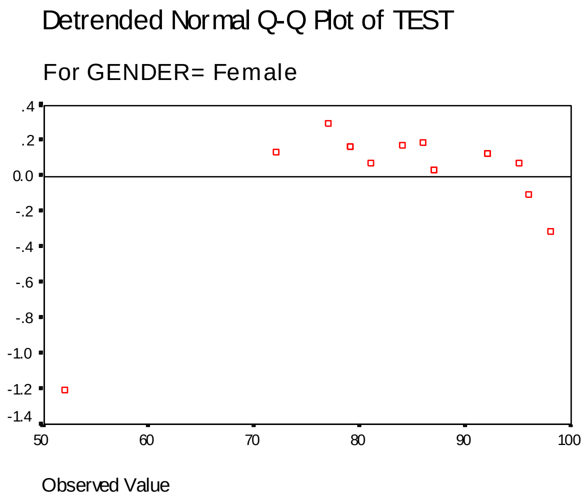
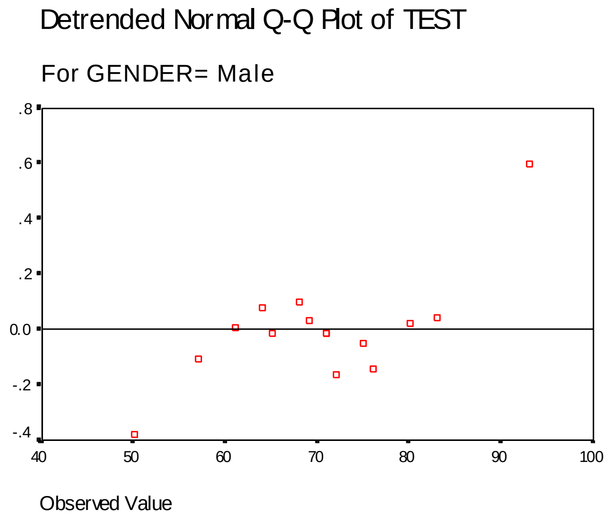


Figure 8: Detrended Normal plot for the Statistic Test Result for Male Students



7. Normality Test

Although normal probability plots provide a visual basis for checking normality, it is often desirable to compute a statistical test of the hypothesis that the data are form a normal distribution.

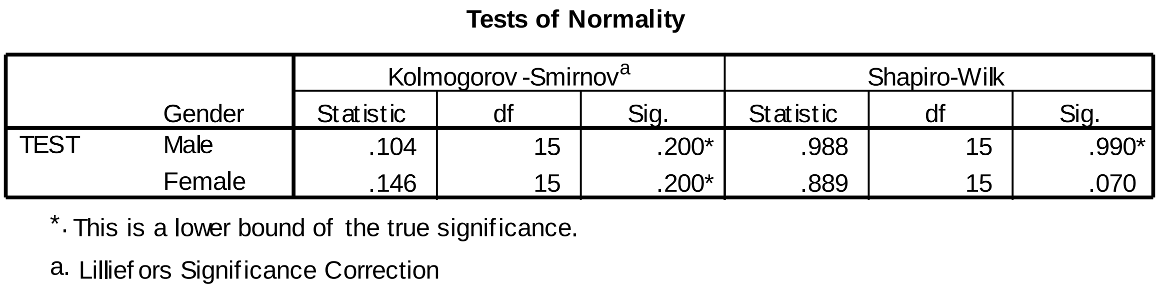
Two commonly used tests are:

1. The Shapiro-Wilk's test
2. The Lilliefors' test (based on a modification of the Kolmogorov-Smirnov test).

The hypotheses for normality test are as follows:

H_O: Groups or sample come from normally distributed population.
H_A: The samples are not normally distributed.

Table 6: Normality Tests for the Statistic Test Result for Female Students



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender	Statistic	df	Sig.	Statistic	df	Sig.
TEST	Male	.104	15	.200*	.988	15	.990*
	Female	.146	15	.200*	.889	15	.070

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

From the large observed significance levels (larger than 0.05), you see that the hypothesis of normality cannot be rejected, or the sample comes from a normally distributed population (The populations from which the samples are drawn are approximately normally distributed).

8. Spread vs. Level with Levene Test and Transformation

Controls data transformation for spread-versus-level plots. For all spread-versus-level plots, the slope of the regression line and Levene test for homogeneity of variance are displayed. If you select a transformation, Levene test is based on the transformed data. If no factor variable is selected, spread-versus-level plots are not produced. Power estimation produces a plot of the natural logs of the interquartile ranges against the natural logs of the medians for all cells, as well as an estimate of the power transformation for achieving equal variances in the cells. A spread-versus-level plot helps determine the power for a transformation to stabilize (make more equal) variances across groups. Transformed allows you select one of the Power alternatives, perhaps following the recommendation from Power estimation, and produces plots of transformed data. The interquartile range and median of the transformed data are plotted. Untransformed produces plots of the raw data. This is equivalent to a transformation with a power of 1.

Many statistical procedures, such as ANOVA and so forth require that *all groups* come from normal populations with the same variance.

Therefore, before choosing a statistical hypothesis, we need to test that all the group variances are equal or that the samples come from normal populations.

If it appears that the assumptions are violated, we may want to determine appropriate transformations.

Re-expression and Transformation of Data

Steps in Transforming a Data

1. Obtain a spread-versus-level plot

It is a plot of the values of spread and level of each group. Very often there is a relationship between the average value, or level, of a variable and the variability, or spread, associated with it.

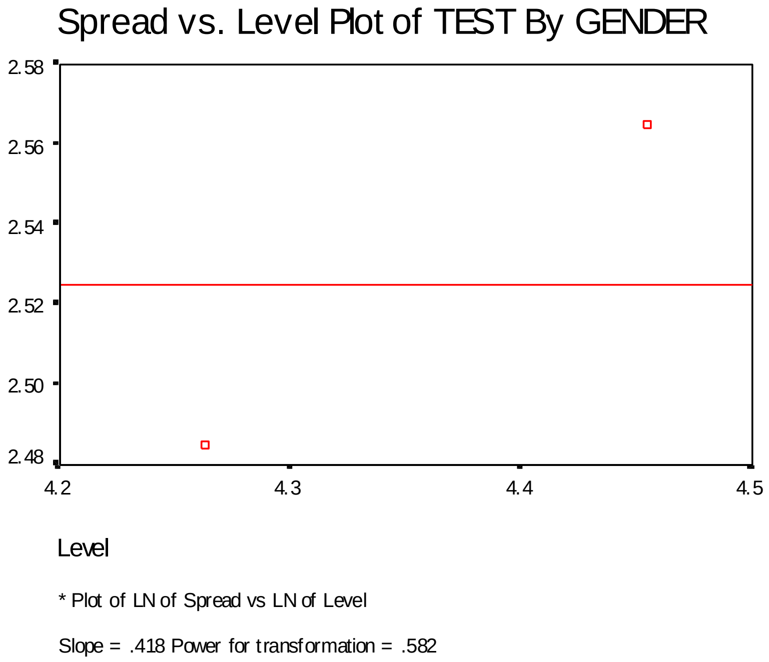
You can see that there is a fairly strong linear relationship between spread and level.

2. Determine or estimate the power value the power value that will eliminate or lessen this relationship.

If there is no relationship, the points should cluster around a horizontal line.

If this is not the case, we can used the observed relationship between the variables to choose an appropriate transformation.

Figure 9: Spread-versus-level plot for the Statistic Test Result by Gender



* Plot of LN of Spread vs LN of Level

Slope = .418 Power for transformation = .582

Take note of the slope value.
The power is obtained by subtracting the slope from 1.

Power = 1 - slope

= 1- 0.418,

= 0.582 (for the statistic test result example)

≈ 1.0

Then choose the closest power that is a multiple of 1.0.
To transform data, you must select a power for the transformation. You can choose one of the following most *commonly used transformations*:

Power	Transformation	Description
3	Cube	Each data value is cubed
2	Square	Each data value is squared
1	No change	No transformation or re-expression is required
½	Square root	The square root of each data value is calculated
0	Logarithm/ Natural log	Natural log transformation
- ½	Reciprocal of the square root	For each data value, the reciprocal of the square root is calculated
- 1	Reciprocal	The reciprocal of each data value is calculated

Power transformation is frequently used to stabilize variances. A power transformation raises each data value to specified power.

As shown be Figure 9 above, the slope of the least-square line for the bank data is 0.418, so the power for the transformation is 0.582. Rounding 0.582 to the nearest whole number gives you 1.0. Therefore, no change or no transformation or re-expression is required.

3. After applying the power transformation, it is wise to obtain a spread-versus-level plot for the transformed data. From this plot you can judge the success of the transformation.

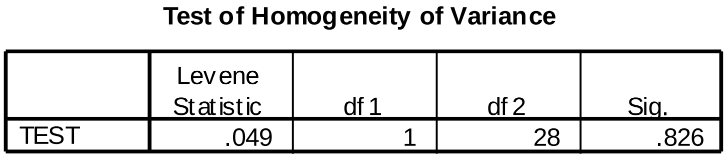
The Levene Test

The Levene test is a homogeneity-of-variance test that is to test the null hypothesis that all the group variances are equal.

The hypotheses for Levene test are as follows:

H_O: Groups or samples come from populations with the same variance or equal variance
H_A: The variances are unequal

Table 7: Levene Test of Homogeneity of Variance for the Statistic Test Result for Female Students



Test of Homogeneity of Variance

	Levene Statistic	df1	df2	Sig.
TEST	.049	1	28	.826

If the observed significant level is equal or larger than 0.05 level, then the null hypothesis that all group variances are equal is not rejected (true), or the groups variances are equal. Since the observed significant level is 0.826 (which is greater than alpha value of 0.05). So you cannot reject the H_O, the H_O is true. This implies that you don't have significant evidence to suspect the variances are unequal or variances are the same.

Conversely, if the observed significant level is smaller than the 0.05 level, then the null hypothesis that all group variances are equal is rejected, or the groups variance are unequal.

For example if the observed significance level is 0.04 (which is <0.05) than H_O is rejected or the H_A is accepted (the variances are unequal).

In this case we should consider transforming the data if we plan to use a statistical test which requires equality of variance.

Supplementary Notes

Table 1: The Boxplot Display & Computation of 5% Trimmed Mean for Male & Female Students Test Scores

Male Test Score	Boxplot	Female Test Score	Boxplot	5% Trimmed Mean
93	Outlier	98	Upper whisker	▨
83	Upper whisker	96		
80		95		
76	75th Percentile	92	75th Percentile	
75		92		
72		87		
71		87		
71	Median	86	Median	
69		84		
68		81		
65		79		
64	25th Percentile	79	25th Percentile	
61		77		
57		72	Lower whisker	
50	Lower whisker	52	Outlier	▨

Mean (arithmetic mean):
Male = 70.33, Female = 83.800

5% Trimmed Mean:
Male 70.20, Female = 84.78

Legends:

⇔

Only these cases (13) are used for the computation of the 5% trimmed mean.

↔

50% of the cases have test scores are within the **25th Percentile** and **75th Percentile or** within the box.

Upper whisker: Largest test score (observed value) that is not an outlier.

Lower whisker: Smallest test score (observed value) that is not an outlier.

Outliers: Values more than 1.5 box length for the **25th Percentile** and/or **75th Percentile**.

Extremes: Values more than 3 box length for the **25th Percentile** and/or **75th Percentile**.

Percentiles: 5, 10, 25, 50, 75,90 & 95th

A percentile point is defined as the point on the distribution below which a given % of the scores is found.

Example: 25th percentile is the point below which 25% of the score fall. For male students the 25th percentile is 64%. In other word 25% of the female students' score fall below 64%. By comparison, the 25th percentile score for the female students is 79% or 25% of the female students obtained scores smaller than 79.

Normality: Degree to which the distribution of the sample data corresponds to a normal distribution.

Normal Probability Plot: A graphical comparison of the form of the distribution to the normal distribution.

In the graph (Normal Probability Plot), the normal distribution is represented by a straight line angled at 45 degrees.

The actual distribution is plotted against this straight line. Any differences are shown as deviation from straight line, making identification of differences quite simple.

Detrended Q-Q Plot: If your sample data are normally distributed, then 95% of the data should fall between –2 and +2.

Or, then 99% of the data (only 1 in 1000) should fall outside –3 and +3.

Null Hypothesis (H_O): Is a hypothesis of no difference (same, equal, similar).

Rule of thumb: If the *observed significance level* is **smaller** than *the alpha level* (**0.05**) then you reject the null hypothesis (H_O) (or you accept the alternative hypothesis/H_A). Conversely, if the *observed significance level* is **larger** than *the alpha level* (**0.05**) then you accept (you cannot reject or fail to reject) the null hypothesis (H_O).

Exploratory data analysis (EDA) for checking assumptions

Exploratory data analysis procedures could assist us is checking the normality and equality of variance assumptions.

How EDA help us make decision to choose appropriate confirmatory data analysis procedures?

1. If the two assumptions are met, then you could proceed with your CONFIRMATORY DATA ANALYSES using the parametric test you have proposed.
2. If either one or both assumptions is (are) not met then you need to carry out DATA TRANSFORMATION.

3. Proceed with your confirmatory data analyses if both assumptions are met after data transformation.

4. If one or both assumptions are not met after transformation then you to use the NONPARAMETRIC EQUIVALENT TESTS.

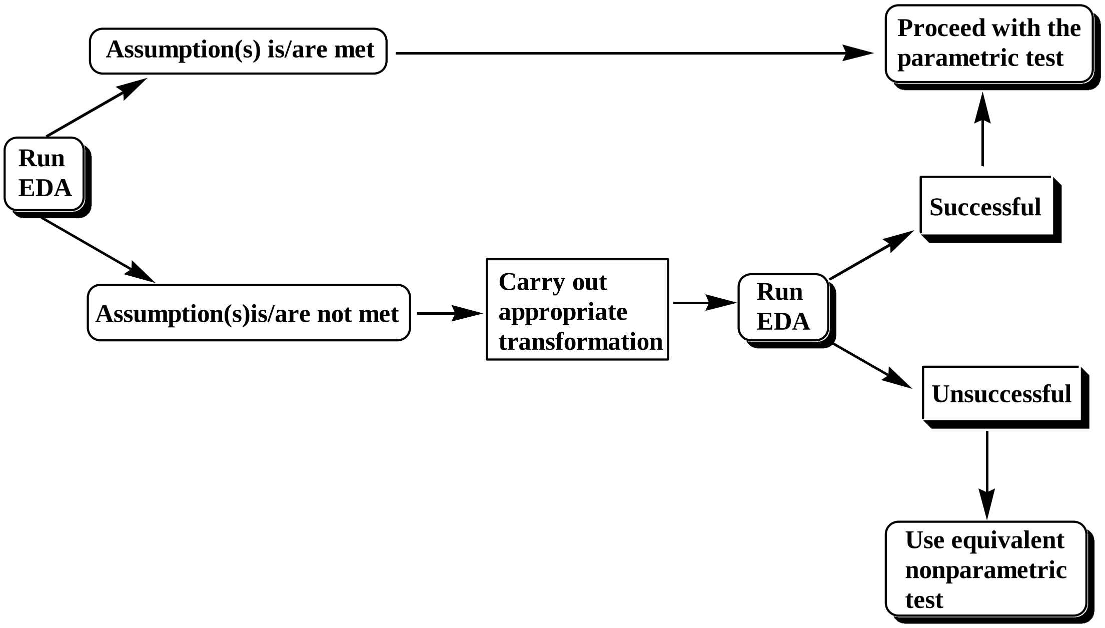

Run EDA
Assumption(s) is/are met
Proceed with the parametric test
Assumption(s)is/are not met
Carry out appropriate transformation
Run EDA
Successful
Unsuccessful
Use equivalent nonparametric test

Figure 1: EDA Decision Tree

Table 2: Nonparametric (NPAR) Equivalent Test of the Parametric Test

Parametric Tests	Nonparametric Equivalent Test
Independent-samples t-test	Mann-Whitney U-test
Paired-samples t-test	McNemar, Sign and Wilcoxon Test
One-way ANOVA	Kruskal-Wallis
Two-way ANOVA	Friedman and Cochran Test
Pearson product-moment correlation	Chi-square test of independent (Spearman Rho, Phi and Cramer’s V etc)

Assumption Testing in Parametric Tests

Statistical Procedure	*Assumption(s) Must Be Met*		
	Population Normality	Homogeneity of Variance	Linearity
One-sample t-test	Yes	-	-
Paired-samples t-test	Yes	-	-
Independent-samples t-test	Yes	Yes	
One-way ANOVA	Yes	Yes	
Two-way ANOVA (Simple Factorial)	Yes	Yes	
Bi-variate Correlation	Yes	-	Yes
Simple/Multiple Linear Regression*	Yes	Yes	Yes
One-way Analysis of Covariance (ANCOVA)*	Yes	-	Yes
Factor Analysis*	Yes	-	Yes
Multivariate Analysis of Variance (MANOVA)*	Yes		Yes

* These statistical procedures need additional assumption testing. For detaiks refer to Coakes, S. and Steed, L (2003). SPSS, Analysis Without Anguish, Version 11.0 for Windows. Milton, Queenland: John Wiley & Sons Australia, Ltd.

Checking Normality with Skewness

Skewness (Asymmetry)

Skewness is a **MEASURE of SYMMETRY**. It provides an indication of departure from normality. A distribution displaying no skewness (i.e., Sk = 0) would indicate coincidence (i.e., mean = median = mode). A negative skewness (i.e., Sk < 0) would display a distribution tailing off to the left while a positive skewness (i.e., Sk > 0) would display a distribution tailing off to the right. It should be noted that the measures of mean and standard deviation are useful only if the data are not significantly skewed.

The ***skewness*** of a distribution is an indicator of its asymmetry. A traditional way to compute skewness is to average the 3rd power of the quotient of each value's deviation from the mean divided by the standard deviation. Wow! That's a mouthful. Can we make it more bite-sized? Since to say deviations from the mean divided by the standard deviation is cumbersome, let's introduce a shorthand method for expressing that idea more economically and see where it leads:

$$z_i = \frac{(X_i - X)}{s}$$

You will find the term z_i throughout both measurement and statistical literature. It expresses the powerful idea of standardizing values in a distribution do that the standardized form has a mean of zero and unit standard deviation. After each value in the distribution has been so transformed, it is a simple matter to discover how many standard deviations a value is away from its mean. The value of z embodies the answer.

For example, if a distribution has a mean of 50 and a standard deviation of 10, then a value of 60 in the distribution can be transformed to *z* = (60 - 50) / 10 or *z* = 1. We immediately know that 60 is one standard deviation above its mean. This is an extremely convenient way to compare test scores from two different exams to get an idea of comparability of results. Suppose you take a test and receive a 60 score where the mean and standard deviation are 50 and 10 as above. Your friend takes a test from another course and receives an 80. Did she do better? It sounds like it, doesn't it? After all, 80 is better than 60. Right? What if you discovered that the mean on your friend's exam is 90 and the standard deviation is 20? After transforming her score to z-form, you would see that the result is -0.5 indicating that she scored one-half a standard deviation below the mean. Now do you think she performed better? Obviously, other factors need be considered in making comparisons like this but the example should make the point that *z* is valuable in its own right.

We now can write an expression for skewness that is easily read:

$$skewness = \frac{\sum z_i^3}{n}$$

Even though this formula is technically accurate, we are going to substitute a simpler formula for skewness that makes interpretation a little easier. It has the advantage that its sign always gives the direction of the mean from the median. That formula is:

$$skewness = \frac{3(Mean - Median)}{s}$$

In the example dataset, this formula for skewness returns values of -.8, -.7, -.4, and +.4 for age, height, weight, and fitness respectively. Notice, in the formula, if the mean is smaller than the median the coefficient will be negative. On the other hand, if the mean is greater than the median, the coefficient is positive. Either way, the sign of the coefficient indicates the direction of skew since the mean is pulled in the direction of skew.

The interpretation of this skewness statistic is very straightforward. If its value is more negative than -.5, then the distribution is ***negatively*** skewed (meaning that the left-hand tail of the distribution is pulled out more than would be the case if the distribution were symmetric). If the coefficient's value exceeds +.5, then the distribution is ***positively*** skewed (meaning that the right hand tail of the distribution is pulled out more than would be the case if the distribution were symmetric). If the coefficient's value falls between -.5 and +.5 inclusive then the distribution appears to be relatively symmetric.

Measure of skewness

- Worry if skewness > 1 or skewness < 1
- Do something if
    - Skew / Std Err(Skew) < -2

- Skew / Std Err(Skew) > 2

Rule of thumb for checking normality base on measure of skewness

According to George & Mallery (2003)[1], a skewness value between ±1.0 is considered excellent for most psychometric purposes, but a value between ±2.0 is on many cases also acceptable, depending on your particular application (page 99).

What do we do if the distribution is very skewed?

Carry out mathematical transformation[2]

In the example dataset, age and height appear to have a slight negative skew, while weight and fitness are more nearly symmetric. Consequently, an analyst would likely consider a mathematical transformation, such as a logarithmic transformation, of both age and height prior to further analysis. You are encouraged to do a log transform and investigate its effect on skew. You might also try a square root transformation. These are just a few out of many possibilities. The goal will always be to bring the distribution into a more symmetric shape before pursuing advanced statistical techniques. If you have questions about how to do mathematical transformations, contact your instructor.

If the skewness is between 0 and 2, a square root transformation is probably appropriate. To perform this transformation, type 'SQRT(variable)' in the box labeled 'Numerical expression', where the term 'variable' is replaced with the name of that variable. For example, you might type SQRT (confid)
If the skewness is between 2 and 5, a log transformation might be appropriate. Type 'LG10(variable) in the box labeled 'Numerical expression'.
If the skewness exceeds 5, an inverse transformation might be suitable. Type '1/variable' in the box labeled 'Numerical expression'. In this instance, note that higher scores on the transformed variable correspond to lower scores on the original variable.

Kurtosis (Peakedness)

Kurtosis is a MEASURE of PEAKEDNESS. A strongly platykurtic distribution (i.e., $k < 3$) would indicate either a very flat distribution or one that is bimodal (for this reason, it is prudent to construct frequency distribution graphs for the purpose of visual assessment). Kurtosis values of $k = 3$ and $k > 3$ indicate mesokurtic and leptokurtic distributions, respectively. Both skewness and kurtosis show how far a distribution deviates from normality (this is an important consideration when choosing a statistical procedure for hypothesis testing, many of which assume the data to be normally distributed).

[1] George, Darren & Mallery, Paul (2003). SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update. Third Edition. Allyn & Bacon, USA.

[2] Transformations are recommended as a remedy for outliers, breaches in normality, non-linearity, and lack of homoscedasticity.

The ***kurtosis*** of a distribution is an indicator of how peaked a distribution is in the vicinity of its mode compared to the peakedness of a normal distribution. One traditional measure of kurtosis is to average the 4th power of the quotient of each value's deviation from the mean divided by the standard deviation. The formula is:

$$kurtosis = \frac{\sum z_i^4}{n}$$

A kurtosis returned by this formula that is greater than +3 indicates a distribution more peaked than that of a normal curve while a kurtosis below +3 indicates a flatter than normal distribution. The result of this calculation can not be a negative number. For the working dataset, the estimates of kurtosis are 1.5, 2.6, 1.3, and 1.4 respectively for age, height, weight and fitness based upon this method. Thus, all four variables appear flatter than the normal distribution but height is very close to having the same peakedness as a normal distribution. Since age, weight and fitness are not unimodal, their kurtosis measures are of questionable value.

Measure of kurtosis

- Do something if
  - Kur / Std Err(Kur) < -2 (1)
  - Kur / Std Err(Kur) > 2 (5)

According to George & Mallery (2003), a kurtosis value between ±1.0 is considered excellent for most psychometric purposes, but a value between ±2.0 is on many cases also acceptable, depending on your particular application (page 98).

Nonparametric Tests

Distribution-free or nonparametric statistics/tests make few assumptions[3] about the nature of data. They may be particularly suitable where

- there are relatively few cases in the population we wish to examine, or
- where a frequency distribution is badly skewed (rather than symmetrical).

Nonparametric tests can handle data at nominal (classificatory) and ordinal (rank-ordered) levels of measurement. For some nonparametric tests, social class (code, perhap 1 to 5) and sex (coded 1 and 2) could be used as dependent variables.

[3] In parametric tests, you have to assume that each group is an independent random sample from a normal population, and that the groups variances are equal.

The disadvantage of nonparametric statistics/tests is that they are less likely to find a true difference when it exists than the test based on the assumption of normality.

At least 15 different nonparametric tests are available from SPSS

Table 1: Nonparametric tests/statistics available in SPSS

	Test	**N (nominal), O (ordinal), D (dichotomous)**	Purpose	**Number of samples**	**Independent (I) or Related (R)**
1	Chi-square	N	Tests fit between observed and expected frequency	1	-
2	Binomial	N (D)	Tests for biased proportion	1	-
3	Runs	N (D)	Test whether run sequence is random	1	-
4	Kolmogorov-Smirnov	O	Tests match between frequencies	1 or 2	I
5	Wald-Wolfowitz	O	Tests for differences in ranks of independent groups	2	I
6	Moses	O	Compares ranges of two independent groups	2	I
7	Mann-Whitney	O	Tests of differences between two independent groups **(equivalent to independent sample *t*-test)**	2	I
8	McNemar	N (D)	Tests before/after change in D variables	2	R
9	Sign	O	Tests paired data for + or - bias	2	R
10	Wilcoxon	O	Test for differences between related groups **(equivalent to paired sample *t*-test)**	2	R
11	Median	O	2-way frequency association	2-*k*	I
12	**Kendall**	O	Agreement between judges	2-*k*	R

13	Kruskal-Wallis	O	**One-way ANOVA**	*k*	I
14	Cochran	N (D)	Tests before/after change in *k* groups	*k*	R
15	Friedman	O	Two-way ANOVA	*k*	R

Choosing a nonparametric test

The choice of a nonparametric test depends on a number of factors:

1. The nature (level of measurement) of data (e.g. nominal or ordinal)

2. The number of sample or groups (1, 2 or *k*)

3. Whether the two (or more) samples are related or independent

4. What we want to do with the data (e.g. assess the level of association between two variable, determine differences in the central tendency or span of two or more groups, examine sequence patterns, or examine proportions within a single group)

5. The preference of the user (i.e. sometimes a number of alternative test will be suitable for assessing some aspect of the data.

As mentioned above, the statistics available on the nonparametric test submenu in SPSS do not rely on the assumption of normality. Many of the tests use the rank order of each observation rather than the data value as recorded. When this is the case, SPSS automatically transforms the data to ranks for you.

The nonparametric statistics fall under several categories, comparing:

A. Two or more independent groups

When you want to compare the center of location for two distributions (assuming their shapes are the same), use the 2 Independent Samples submenu option to request the Mann-Whitney U test--it is the nonparametric analog of the two-sample t test. The Kolmogorov-Smirnov Z, Moses extreme reactions, and the Wold-Wolfowitz runs tests are also available.

When you have more than two independent groups, use the K Independent Samples option to select the Kruskal-Wallis H test. It, like the Mann-Whitney test, uses ranks.

B. Paired or related variables

When, for each subject (case), you want to compare two variables, use the 2 Related Samples submenu option to request the Wilcoxon test--it is the nonparametric analog of the paired or dependent t test. The sign test and the McNemar test are also available. The latter is appropriate for comparing two categorical variables that are coded with the same two values.

When you want to compare more than two measures for each subject, use the K Related Samples option to request the Friedman test. Kendall’s W and Cochran’s Q reside on the same dialog box with the Friedman test.

C. A sample versus a chi-square, binomial, normal, or other distribution

If you want to compare the observed frequencies of a categorical variable with expected frequencies, use the Chi-Square submenu option and specify whether the expected values should be equal for all categories or equal to proportions you specify.

To compare the observed frequency of a dichotomous variable with expected frequencies from the binomial distribution, use the Binomial option.

To compare the observed cumulative distribution for a variable with a theoretical distribution that is normal, uniform, or Poisson, use the 1-Sample K-S option.

The Runs option provides an additional one-sample test that tests whether the order of the observations fluctuates randomly. For example, do the residuals from a regression model fluctuate randomly around 0?

Chi-Square Test

Chi-Square tests hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if you want to test the hypotheses that people are equally likely to buy six different brands of cereals, you can count the number buying each of the six brands. Based on the six observed counts, you can use the Chi-Square procedure to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories do not have to be equal. You can specify whatever hypothetical proportions you want to test.

The Chi-Square Test procedure tabulates a variable into categories and computes a chi-square statistic. This goodness-of-fit test compares the observed and expected frequencies in each category to test either that all categories contain the same proportion of values or that each category contains a user-specified proportion of values.

Examples: The chi-square test could be used to determine if a bag of jelly beans contains equal proportions of blue, brown, green, orange, red, and yellow candies. You could also test to see if a bag of jelly beans contains 5% blue, 30% brown, 10% green, 20% orange, 15% red, and 15% yellow candies.

Statistics: Mean, standard deviation, minimum, maximum, and quartiles. The number and the percentage of nonmissing and missing cases, the number of cases observed and expected for each category, residuals, and the chi-square statistic.

Binomial Test

Binomial tests the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, if you want to test that the probability that an item on the assembly line is defective is one out of ten (p=0.1), you can take a sample of 300 items and record whether each is defective or not. You can then use the Binomial procedure to test the hypothesis of interest.

The Binomial Test procedure compares the observed frequencies of the two categories of a dichotomous variable to the frequencies expected under a binomial distribution with a specified probability parameter. By default, the probability parameter for both groups is 0.5. To change the probabilities, you can enter a test proportion for the first group. The probability for the second group will be 1 minus the specified probability for the first group.

Example: When you toss a dime, the probability of a head equals 1/2. Based on this hypothesis, a dime is tossed 40 times, and the outcomes are recorded (heads or tails). From the binomial test, you might find that 3/4 of the tosses were heads and that the observed significance level is small (0.0027). These results indicate that it is not likely that the probability of a head equals 1/2; the coin is probably biased.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

Runs Test

Runs tests whether the two values of a dichotomous variable occur in a random sequence.

The Runs test is appropriate only when the order of cases in the data file is meaningful. For example, if you are monitoring items coming off a production line to see whether they are defective or not, you can use the Runs test to see if defective items cluster together.

The Runs Test procedure tests whether the order of occurrence of two values of a variable is random. A run is a sequence of like observations. A sample with too many or too few runs suggests that the sample is not random.

Examples: Suppose that 20 people are polled to find out if they would purchase a product. The assumed randomness of the sample would be seriously questioned if all 20 people were of the same gender. The runs test can be used to determine if the sample was drawn at random.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

One-Sample Kolmogorov-Smirnov (K-S) Test

1-Sample K-S tests whether a sample comes from a specified distribution. You can test against either the uniform, normal, or Poisson distributions.

For example, you can test that lottery numbers are uniformly distributed.

Alternative tests for normality are available in the Explore procedure from the Summarize submenu. The Normal P-P and Normal Q-Q plots available from the Graphs menu can also be used to examine the assumption of normality.

The One-Sample Kolmogorov-Smirnov Test procedure compares the observed cumulative distribution function for a variable with a specified theoretical distribution, which may be normal, uniform, or Poisson. The Kolmogorov-Smirnov Z is computed from the largest difference (in absolute value) between the observed and theoretical cumulative distribution functions. This goodness-of-fit test tests whether the observations could reasonably have come from the specified distribution.

Example: Many parametric tests require normally distributed variables. The one-sample Kolmogorov-Smirnov test can be used to test that a variable, say INCOME, is normally distributed.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

Two-Independent-Samples Tests

The Two-Independent-Samples Tests procedure compares two groups of cases on one variable.

2 Independent Samples is used to compare the distribution of a variable between two nonrelated groups.

Only limited assumptions are needed about the distributions from which the samples are selected.

The Mann-Whitney U test is an alternative to the Independent-Samples T Test.

The actual values of the data are replaced by ranks.

The Kolmogorov-Smirnov Z test is based on the differences between the observed cumulative distributions of the two groups.

The Wald-Wolfowitz runs test sorts the data values from smallest to largest and then performs a runs test on the groups' numbers.

The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

Example: New dental braces have been developed that are intended to be more comfortable, to look better, and to provide more rapid progress in realigning teeth. To find out if the new braces have to be worn as long as the old braces, 10 children are randomly chosen to wear the old braces, and another 10 are chosen to wear the new braces. From the Mann-Whitney U test, you might find that, on average, those with the new braces did not have to wear the braces as long as those with the old braces.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Mann-Whitney U, Moses extreme reactions, Kolmogorov-Smirnov Z, Wald-Wolfowitz runs.

Tests for Several (K) Independent Samples

The Tests for Several Independent Samples procedure compares two or more groups of cases on one variable.

K Independent Samples compares the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected.

The Kruskal-Wallis test is an alternative to One-Way ANOVA, with the actual values of the data replaced by ranks.

The Median test counts the number of cases in each group that are above and below the combined median and then performs a chi-square test.

Example: Do three brands of 100-watt lightbulbs differ in the average time the bulbs will burn? From the Kruskal-Wallis one-way analysis of variance, you might learn that the three brands do differ in average lifetime.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Kruskal-Wallis H, median.

Two-Related-Samples Tests

The Two-Related-Samples Tests procedure compares the distributions of two variables.

2 Related Samples compares the distribution of two related variables.

Only limited assumptions are needed about the distributions from which the samples are selected.

The Wilcoxon and Sign tests are nonparametric alternative to the Paired-Samples T Test.

The Wilcoxon test is more powerful than the Sign test.

The McNemar test is used to determine changes in proportions for related samples. It is often used for "before-and-after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analyzing the number of people whose preference for a candidate changed based on the speech. Using the McNemar test, you analyze the changes to see if change in both directions is equally likely.

Example: In general, do families receive the asking price when they sell their homes? By applying the Wilcoxon signed-rank test to data for 10 homes, you might learn that seven families receive less than the asking price, one family receives more than the asking price, and two families receive the asking price.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Wilcoxon signed rank, sign, McNemar.

Tests for Several (K) Related Samples

The Tests for Several Related Samples procedure compares the distributions of two or more variables.

K Related Samples compares the distribution of two or more related variables.

Only limited assumptions are needed about the distributions from which the samples are selected.

The Friedman test is a nonparametric alternative to a single-factor repeated measures analysis of variance.

You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of five products when each person is asked to rate each of the products on a scale.

Cochran's Q test can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users.

Kendall's W measures the agreement among raters. Each of your cases corresponds to a rater, each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank seven ice-cream flavors from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.

Example: Does the public associate different amounts of prestige with a doctor, a lawyer, a police officer, and a teacher? Ten people are asked to rank these four occupations in order of prestige. Friedman’s test indicates that the public does in fact associate different amounts of prestige with these four professions.

Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Friedman, Kendall’s W, and Cochran’s Q.