Introduction to Statistical Analysis Using IBM SPSS Statistics

Student Guide

Course Code: 0G517

ERC 1.0

Introduction to Statistical Analysis Using IBM SPSS Statistics

0G517

Published October 2010

Table of Contents

Lesson 0: Course Introduction

0.1 *Introduction*

The focus of this two-day course is an introduction to the statistical component of IBM® SPSS® Statistics. This is an application-oriented course and the approach is practical. You'll take a look at several statistical techniques and discuss situations in which you would use each technique, the assumptions made by each method, how to set up the analysis using PASW® Statistics, as well as how to interpret the results. This includes a broad range of techniques for exploring and summarizing data, as well as investigating and testing underlying relationships. You will gain an understanding of when and why to use these various techniques as well as how to apply them with confidence, and interpret their output, and graphically display the results.

0.2 *Course Objectives*

After completing this course students will be able to:

- Perform basic statistical analysis using selected statistical techniques with PASW Statistics

To support the achievement of this primary objective, students will also be able to:

- Explain the basic elements of quantitative research and issues that should be considered in data analysis
- Determine the level of measurement of variables and obtain appropriate summary statistics based on the level of measurement
- Run the **Frequencies** procedure to obtain appropriate summary statistics for categorical variables
- Request and interpret appropriate summary statistics for scale variables
- Explain how to make inferences about populations from samples
- Perform crosstab analysis on categorical variables
- Perform a statistical test to determine whether there is a statistically significant relationship between categorical variables
- Perform a statistical test to determine whether there is a statistically significant difference between two groups on a scale variable
- Perform a statistical test to determine whether there is a statistically significant difference between the means of two scale variables
- Perform a statistical test to determine whether there is a statistically significant difference among three or more groups on a scale dependent variable
- Perform a statistical test to determine whether two scale variables are correlated (related)
- Perform linear regression to determine whether one or more variables can significantly predict or explain a dependent variable
- Perform non-parametric tests on data that don’t meet the assumptions for standard statistical tests

0.3 *About SPSS*

SPSS® Inc., an IBM® Company is a leading global provider of predictive analytics software and solutions. The Company’s complete portfolio of products - data collection, statistics, modeling and deployment - captures people's attitudes and opinions, predicts outcomes of future customer interactions, and then acts on these insights by embedding analytics into business processes. SPSS solutions address interconnected business objectives across an entire organization by focusing on the convergence of analytics, IT architecture and business process. Commercial, government and academic customers worldwide rely on SPSS technology as a competitive advantage in attracting,

retaining and growing customers, while reducing fraud and mitigating risk. SPSS was acquired by IBM® in October 2009. For more information, visit http://www.spss.com.

0.4 *Supporting Materials*

We use several datasets in the course because no one data file contains all the types of variables and relationships between them that are ideal for every technique we discuss. As much as possible, we try to minimize the need within one lesson to switch between datasets, but the first priority is to use appropriate data for each method.

The following data files are used in this course:

- Bank.sav
- Drinks.sav
- Census.sav
- Employee data.sav
- SPSS_CUST.sav

0.5 *Course Assumptions*

General computer literacy. Completion of the "Introduction to PASW Statistics" and/or "Data Management and Manipulation with PASW Statistics" courses or experience with PASW Statistics including familiarity with, opening, defining, and saving data files and manipulating and saving output. Basic statistical knowledge or at least one introductory level course in statistics is recommended.

Note about Default Startup Folder and Variable Display in Dialog Boxes

In this course, all of the files used for the demonstrations and exercises are located in the folder *c:\Train\Statistics_IntroAnalysis*.

Note: If the course files are stored in a different location, your instructor will give you instructions specific to that location.

Either variable names or longer variable labels will appear in list boxes in dialog boxes. Additionally, variables in list boxes can be ordered alphabetically or by their position in the file. In this course, we will display variable names in alphabetical order within list boxes.

1) Select **Edit...Options**
2) Select the **General** tab (if necessary)
3) Select **Display names** in the Variable Lists group on the General tab
4) Select **Alphabetical**
5) Select **OK** and **OK** in the information box to confirm the change

Lesson 1: Introduction to Statistical Analysis

1.1 *Objectives*

After completing this lesson students will be able to:

- Explain the basic elements of quantitative research and issues that should be considered in data analysis

To support the achievement of the primary objective, students will also be able to:

- Explain the basic steps of the research process
- Explain differences between populations and samples
- Explain differences between experimental and non-experimental research designs
- Explain differences between independent and dependent variables

1.2 *Introduction*

The goal of this course is to enable you to perform useful analyses on your data using PASW Statistics. Keeping this in mind, these lessons demonstrate how to perform descriptive and inferential statistical analyses and create charts to support these analyses. This course guide will focus on the elements necessary for you to answer questions from your data.

In this chapter, we begin by briefly reviewing the basic elements of quantitative research and issues that should be considered in data analysis. We will then discuss a number of statistical procedures that PASW Statistics performs. This is an application-oriented course and the approach will be practical. We will discuss:

1) The situations in which you would use each technique.
2) The assumptions made by the method.
3) How to set up the analysis using PASW Statistics.
4) Interpretation of the results.

We will not derive proofs, but rather focus on the practical matters of data analysis in support of answering research questions. For example, we will discuss what correlation coefficients are, when to use them, and how to produce and interpret them, but will not formally derive their properties. This course is not a substitute for a course in statistics. You will benefit if you have had such a course in the past, but even if not, you will understand the basics of each technique after completion of this course.

We will cover descriptive statistics and exploratory data analysis, and then examine relationships between categorical variables using crosstabulation tables and chi-square tests. Testing for mean differences between groups using T Tests and analysis of variance (ANOVA) will be considered. Correlation and regression will be used to investigate the relationships between interval/scale variables and we will also discuss some non-parametric techniques. Graphs comprise an integral part of the analyses and we will demonstrate how to create and interpret these as well.

1.3 *Basic Steps of the Research Process*

All research projects, whether analyzing a survey, doing program evaluations, assessing marketing campaigns, doing pharmaceutical research, etc., can be broken down into a number of discrete

components. These components can be categorized in a variety of ways. We might summarize the main steps as:

1) Specify exactly the aims and objectives of the research along with the main hypotheses.

2) Define the population and sample design.

3) Choose a method of data collection, design the research and decide upon an appropriate sampling strategy.

4) Collect the data.

5) Prepare the data for analysis.

6) Analyse the data.

7) Report the findings.

Some of these points may seem obvious, but it is surprising how often some of the most basic principles are overlooked, potentially resulting in data that is impossible to analyze with any confidence. Each step is crucial for a successful research project and it is never too early in the process to consider the methods that you intend to use for your data analysis.

In order to place the statistical techniques that we will discuss in this course in the broader framework of research design, we will briefly review some of the considerations of the first steps. Statistics and research design are highly interconnected disciplines and you should have a thorough grasp of both before embarking on a research project. This introductory chapter merely skims the surface of the issues involved in research design. If you are unfamiliar with these principles, we recommend that you refer to the research methodology literature for more thorough coverage of the issues.

Research Objectives

It is important that a research project begin with a set of well-defined objectives. Yet, this step is often overlooked or not well defined. The specific aims and objectives may not be addressed because those commissioning the research do not know exactly which questions they would like answered. This rather vague approach can be a recipe for disaster and may result in a completely wasted opportunity as the most interesting aspects of the subject matter under investigation could well be missed. If you do not identify the specific objectives, you will fail to collect the necessary information or ask the necessary question in the correct form. You can end up with a data file that does not contain the information that you need for your data analysis step.

For example, you may be asked to conduct a survey "to find out about alcohol consumption and driving". This general objective could lead to a number of possible survey questions. Rather than proceeding with this general objective, you need to uncover more specific hypotheses that are of interest to your organization. This example could lead to a number of very specific research questions, such as:

“What proportion of people admits to driving while above the legal alcohol limit?”

“What demographic factors (e.g., age/sex/social class) are linked with a propensity to drunk-driving?”

“Does having a conviction for drunk-driving affect attitudes towards driving while over the legal limit?”

These specific research questions would then define the questionnaire items. Additionally, the research questions will affect the definition of the population and the sampling strategy. For example, the third question above requires that the responder have a drunk-driving conviction. Given that a

relatively small proportion of the general population has such a conviction, you would need to take that into consideration when defining the population and sampling design.

Therefore, it is essential to state formally the main aims and objectives at the outset of the research so the subsequent stages can be done with these specific questions in mind.

1.4 *Populations and Samples*

In studies involving statistical analysis it is important to be able to characterize accurately the population under investigation. The population is the group to which you wish to generalize your conclusions, while the sample is the group you directly study. In some instances the sample and population are identical or nearly identical; consider the Census of any country. In the majority of studies, the sample represents a small proportion of the population.

In the example above, the population might be defined as those people with registered drivers' licenses. We could select a sample from the drivers' license registration list for our survey. Other common examples are: membership surveys in which a small percentage of members are sent questionnaires, medical experiments in which samples of patients with a disease are given different treatments, marketing studies in which users and non users of a product are compared, and political polling.

The problem is to draw valid inferences from data summaries in the sample so that they apply to the larger population. In some sense you have complete information about the sample, but you want conclusions that are valid for the population. An important component of statistics and a large part of what we cover in the course involves statistical tests used in making such inferences. Because the findings can only be generalized to the population under investigation, you should give careful thought to defining the population of interest to you and making certain that the sample reflects this population. To state it in a simple way, statistical inference provides a method of drawing conclusions about a population of interest based on sample results.

1.5 *Research Design*

With specific research goals and a target population in mind, it is then possible to begin the design stage of the research. There are many things to consider at the design stage. We will consider a few issues that relate specifically to data analysis and statistical techniques. This is not meant as a complete list of issues to consider. For example, for survey projects, the mode of data collection, question selection and wording, and questionnaire design are all important considerations. Refer to the survey research literature as well as general research methodology literature for discussion of these and other research design issues.

First, you must consider the type of research that will be most appropriate to the research aims and objectives. Two main alternatives are experimental and non-experimental research. The data may be recorded using either **objective** or **subjective** techniques. The former includes items measured by an instrument and by computer such as physiological measures (e.g. heart-rate) while the latter includes observational techniques such as recordings of a specific behavior and responses to questionnaire surveys.

Most research goals lend themselves to one particular form of research, although there are cases where more than one technique may be used. For example, a **questionnaire survey** would be inappropriate if the aim of the research was to test the effectiveness of different levels of a new drug to relieve high blood pressure. This type of work would be more suited to a tightly controlled **experimental study** in which the levels of the drug administered could be carefully controlled and objective measures of blood pressure could be accurately recorded. On the other hand, this type of laboratory-based work would not be a suitable means of uncovering people’s voting intentions.

The classic experimental design consists of two groups: the **experimental group** and the **control group**. They should be equivalent in all respects other than that those in the former group are subjected to an effect or treatment and the latter is not. Therefore, any differences between the two groups can be directly attributed to the effect of this treatment. The treatment variables are usually referred to as **independent** variables, and the quantity being measured as the effect is the **dependent** variable. There are many other research designs, but most are more elaborate variations on this basic theme.

In non-experimental research, you rarely have the opportunity to implement such a rigorously controlled design. For example, we cannot randomly assign students to schools, however, the same general principles apply to many of the analyses you perform.

1.6 *Independent and Dependent Variables*

In general, the **dependent** (sometimes referred to as the **outcome**) variable is the one we wish to study as a function of other variables. Within an experiment, the dependent variable is the measure expected to change as a result of the experimental manipulation. For example, a drug experiment designed to test the effectiveness of different sleeping pills might employ the number of hours of sleep as the dependent variable. In surveys and other non-experimental studies, the dependent variable is also studied as a function of other variables. However, no direct experimental manipulation is performed; rather the dependent variable is hypothesized to vary as a result of changes in the other (independent) variables.
Correspondingly, **independent** (sometimes referred to as **predictor**) variables are those used to measure features manipulated by the experimenter in an experiment. In a non-experimental study, they represent variables believed to influence or predict a dependent measure.

Thus terms (dependent, independent) reasonably applied to experiments have taken on more general meanings within statistics. Whether such relations are viewed causally, or as merely predictive, is a matter of belief and reasoning. As such, it is not something that statistical analysis alone can resolve. To illustrate, we might investigate the relationship between starting salary (dependent) and years of education, based on survey data, and then develop an equation predicting starting salary from years of education. Here starting salary would be considered the dependent variable although no experimental manipulation of education has been performed. One way to think of the distinction is to ask yourself which variable is likely to influence the other? In summary, the dependent variable is believed to be influenced by, or be predicted by, the independent variable(s).

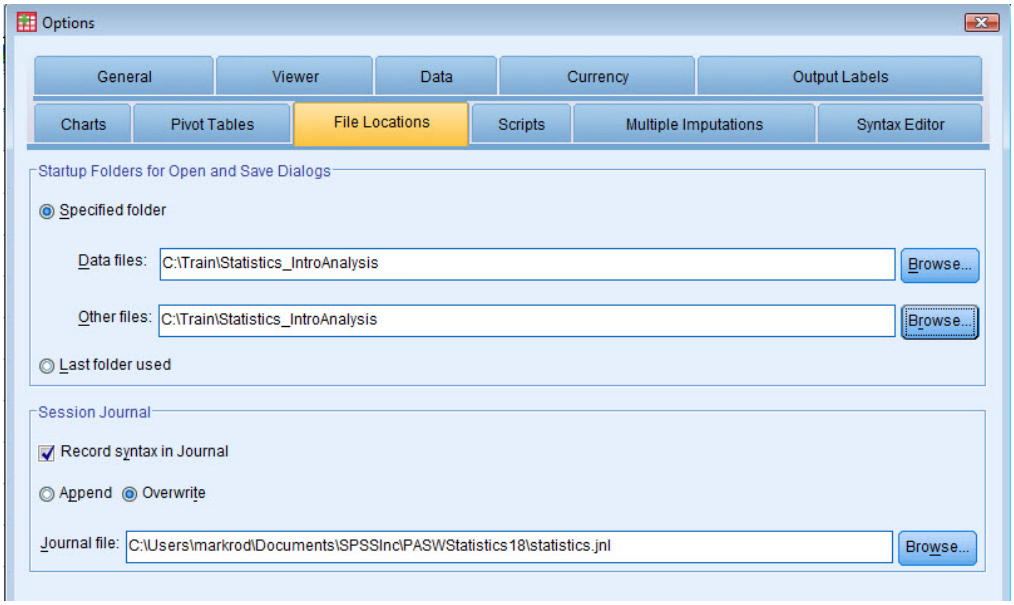
Finally, in some studies, or parts of studies, the emphasis is on exploring and characterizing relationships among variables with no causal view or focus on prediction. In such situations there is no designation of dependent and independent variables. For example, in crosstabulation tables and correlation matrices the distinction between dependent and independent variables is not necessary. It rather resides in the eye of the beholder (researcher).

1.7 *Note about Default Startup Folder and Variable Display in Dialog Boxes*

In this course, all of the files used for the demonstrations and exercises are located in the folder *c:\Train\Statistics_IntroAnalysis*.You can set the startup folder that will appear in all Open and Save dialog boxes. We will use this option to set the startup folder.

Select **Edit...Options**, and then select the **File Locations** tab
Select the **Browse** button to the right of the **Data Files** text box
Select **Train** from the Look In: drop down list, then select **Statistics_IntroAnalysis** from the list of folders and select **Set** button
Click the **Browse** button to the right of the **Other Files** text box and repeat the process to set this folder to **Train\Statistics_IntroAnalysis**

Figure 1.1 Set Default File Location in the Edit Options Dialog Box



Note: If the course files are stored in a different location, your instructor will give you instructions specific to that location.

Either variable names or longer variable labels will appear in list boxes in dialog boxes. Additionally, variables in list boxes can be ordered alphabetically or by their position in the file. In this course, we will display variable names in alphabetical order within list boxes.

Select **General** tab
Select **Display names** in the Variable Lists group on the General tab
Select **Alphabetical** (Not shown)
Select **OK** and then **OK** in the information box to confirm the change

1.8 *Lesson Summary*

In this lesson, we reviewed the basic elements of quantitative research and issues that should be considered in data analysis.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Explain the basic elements of quantitative research and issues that should be considered in data analysis

To support the achievement of the primary objective, students should now also be able to:

- Explain the basic steps of research process
- Explain differences between populations and samples
- Explain differences between experimental and non-experimental research designs
- Explain differences between independent and dependent variables

1.9 *Learning Activity*

In this set of learning activities you won't need any supporting material.

1. In each of the following scenarios, state the possible goals of the research, the type of design you can use, and the independent and dependent variables:
    a. The relationship between gender and whether a product was purchased.
    b. The difference between income categories (e.g., low, medium, and high) and number of years of education.
    c. The effect of two different marketing campaigns on number of items purchased.

2. In your own organization/field, are experimental studies ever done? If not, can you imagine how an experiment might be done to study a topic of interest to you or your organization? Describe that and the challenges such an experimental design would encounter.

Lesson 2: Understanding Data Distributions – Theory

2.1 *Objectives*

After completing this lesson students will be able to:

- Determine the level of measurement of variables and obtain appropriate summary statistics based on the level of measurement

To support the achievement of this primary objective, students will also be able to:

- Describe the levels of measurement used in PASW Statistics
- Use measures of central tendency and dispersion
- Use normal distributions and z-scores

Introduction

Ideally, we would like to obtain as much information as possible from our data. In practice however, given the measurement level of our variables, only some information is meaningful. In this lesson we will discuss level of measurement and see how this determines the summary statistics we can request.

Business Context

Understanding how level of measurement impacts the kind of information we can obtain is an important step before we collect our data. In addition, level of measurement also determines the kind of research questions we can answer, and so this is a critical step in the research process.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

2.2 *Levels of Measurement and Statistical Methods*

The term **levels of measurement** refers to the properties and meaning of numbers assigned to observations for each item. Many statistical techniques are only appropriate for data measured at particular levels or combinations of levels. Therefore, when possible, you should determine the analyses you will be using before deciding upon the level of measurement to use for each of your variables. For example, if you want to report and test the mean age of your customers, you will need to ask their age in years (or year of birth) rather than asking them to choose an age group into which their age falls.

Because measurement type is important when choosing test statistics, we briefly review the common taxonomy of level of measurement.

The four major classifications that follow are found in many introductory statistics texts. They are presented beginning with the weakest and ending with those having the strongest measurement properties. Each successive level can be said to contain the properties of the preceding types and to record information at a higher level.

- **Nominal** — In nominal measurement each numeric value represents a category or group identifier, only. The categories cannot be ranked and have no underlying numeric value. An example would be marital status, coded 1 (Married), 2 (Widowed), 3 (Divorced), 4 (Separated) and 5 (Never Married); each number represents a category and the matching of specific numbers to categories is arbitrary. Counts and percentages of observations falling into each category are appropriate summary statistics. Such statistics as the mean (the average marital status?) would not be appropriate, but the mode would be appropriate (the most frequent category).

- **Ordinal** — For ordinal measures the data values represent ranking or ordering information. However, the difference between the data values along the scale is not equal. An example would be specifying how happy you are with your life, coded 1 (Very Happy), 2 (Happy), and 3 (Not Happy). There are specific statistics associated with ranks; PASW Statistics provides a number of them mostly within the Crosstabs, Nonparametric and Ordinal Regression procedures. The mode and median can be used as summary statistics.

- **Interval** — In interval measurement, a unit increase in numeric value represents the same change in quantity regardless of where it occurs on the scale. For interval scale variables such summaries as means and standard deviations are appropriate. Statistical techniques such as regression and analysis of variance assume that the dependent (or outcome) variable is measured on an interval scale. Examples might be temperature in degrees Fahrenheit or SAT score.

- **Ratio** — Ratio measures have interval scale properties with the addition of a meaningful zero point; that is, zero indicates complete absence of the characteristic measured. For statistics such as ANOVA and regression only interval scale properties are assumed, so ratio scales have stronger properties than necessary for most statistical analyses. Health care researchers often use ratio scale variables (number of deaths, admissions, discharges) to calculate rates. The ratio of two variables with ratio scale properties can thus be directly interpreted. Money is an example of a ratio scale, so someone with $10,000 has ten times the amount as someone with $1,000.

The distinction between the four types is summarized below.

Table 2.1 Level of Measurement Properties

Level of Measurement	*Property*			
	Categories	**Ranks**	**Equal Intervals**	**True Zero Point**
Nominal	✓			
Ordinal	✓	✓		
Interval	✓	✓	✓	
Ratio	✓	✓	✓	✓

These four levels of measurement are often combined into two main types, **categorical** consisting of nominal and ordinal measurement levels and **scale** (or continuous) consisting of interval and ratio measurement levels.

The measurement level variable attribute in PASW Statistics recognizes three measurement levels: Nominal, Ordinal and Scale. The icon indicating the measurement level is displayed preceding the variable name or label in the variable lists of all dialog boxes. The following table shows the most common icons used for the measurement levels. Special data types, such as Date and Time variables have distinct icons not shown in this table.

Table 2.2 Variable List Icons

Measurement Level	Data Type	
	Numeric	String
Nominal		
Ordinal		Not Applicable
Scale		Not Applicable

Rating Scales and Dichotomous Variables

A common scale used in surveys and market research is an ordered rating scale usually consisting of five- or seven-point scales. Such ordered scales are also called Likert scales and might be coded 1 (Strongly Agree, or Very Satisfied), 2 (Agree, or Satisfied), 3 (Neither agree nor disagree, or Neutral), 4 (Disagree, or Dissatisfied), and 5 (Strongly Disagree, or Very Dissatisfied). There is an ongoing debate among researchers as to whether such scales should be considered ordinal or interval. PASW Statistics contains procedures capable of handling such variables under either assumption. When in doubt about the measurement scale, some researchers run their analyses using two separate methods, since each make different assumptions about the nature of the measurement. If the results agree, the researcher has greater confidence in the conclusion.

Dichotomous (binary) variables containing two possible responses (often coded 0 and 1) are often considered to fall into all of the measurement levels except ratio (at least as independent variables). As we will see, this flexibility allows them to be used in a wide range of statistical procedures

Implications of Measurement Level

As we have discussed, the level of measurement of a variable is important because it determines the appropriate summary statistics, tables, and graphs to describe the data. The following table summarizes the most common summary measures and graphs for each of the measurement levels and PASW Statistics procedures that can produce them.

Table 2.3 Summary of Descriptive Statistics and Graphs

	NOMINAL	ORDINAL	SCALE
Definition	Unordered Categories	Ordered Categories	Metric/Numeric Values
Examples	Labor force status, gender, marital status	Satisfaction ratings, degree of education	Income, height, weight
Measures of Central Tendency	Mode	Mode Median	Mode Median Mean
Measures of Dispersion	N/A	Min/Max/Range, InterQuartile Range (IQR)	Min/Max/Range, IQR, Standard Deviation/Variance
Graph	Pie or Bar	Pie or Bar	Histogram, Box & Whisker, Stem & Leaf
Procedures	Frequencies	Frequencies	Frequencies, Descriptives, Explore

Measurement Level and Statistical Methods

Statistics are available for variables at all levels of measurement for more advanced analysis. In practice, your choice of method depends on the questions you are interested in asking of the data and the nature of the measurements you make. The table below suggests which statistical techniques are most appropriate, based on the measurement level of the dependent and independent variable. Much more extensive diagrams and discussion are found in Andrews et al. (1981), or other standard statistical texts.

Table 2.4 Level of Measurement and Appropriate Statistical Methods

Dependent Variable	Independent Variables		
	Nominal	**Ordinal**	**Interval/Ratio**
Nominal	Crosstabs	Crosstabs	Discriminant, Logistic Regression
Ordinal	Nonparametric tests, Ordinal Regression	Nonparametric correlation, Optimal Scaling Regression	Ordinal Regression
Interval/ Ratio	T Test, ANOVA	Nonparametric Correlation	Correlation, Linear Regression

Best Practice

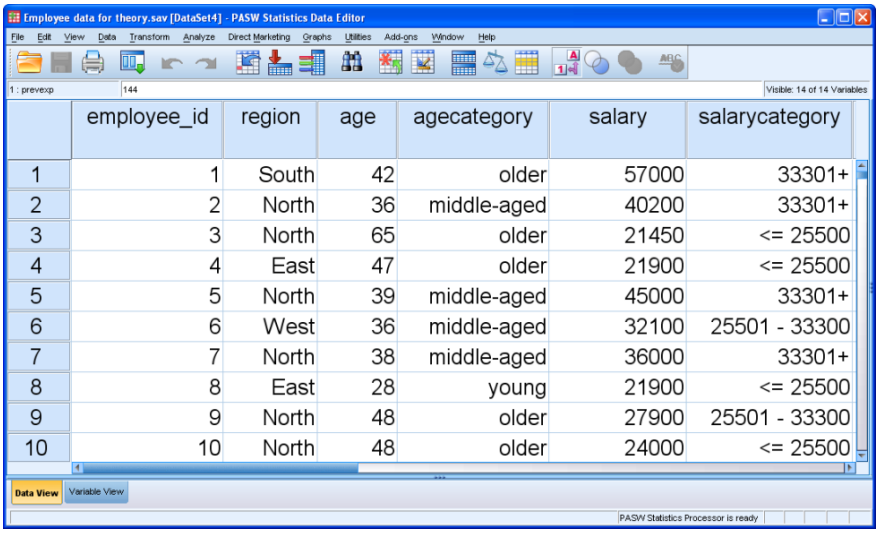
If in doubt about the measurement properties of your variables, you can apply a statistical technique that assumes weaker measurement properties and compare the results to methods making stronger assumptions. A consistent answer provides greater confidence in the conclusions.

Apply Your Knowledge

1. PASW Statistics distinguishes three levels of measurement. Which of these is *not* one of those levels?
   a. Categorical
   b. Scale
   c. Nominal
   d. Ordinal

2. True or false? An ordinal variable has all properties of a nominal variable?

3. Consider the dataset depicted below. Which statements are correct?
   a. The variable *region* is an ordinal variable
   b. The variable *age* is a scale variable
   c. The variable *agecategory* is an ordinal variable
   d. The variable *salarycategory* is a scale variable



	employee_id	region	age	agecategory	salary	salarycategory
1	1	South	42	older	57000	33301+
2	2	North	36	middle-aged	40200	33301+
3	3	North	65	older	21450	<= 25500
4	4	East	47	older	21900	<= 25500
5	5	North	39	middle-aged	45000	33301+
6	6	West	36	middle-aged	32100	25501 - 33300
7	7	North	38	middle-aged	36000	33301+
8	8	East	28	young	21900	<= 25500
9	9	North	48	older	27900	25501 - 33300
10	10	North	48	older	24000	<= 25500

2.3 *Measures of Central Tendency and Dispersion*

Measures of central tendency and dispersion are the most common measures used to summarize the distribution of variables. We give a brief description of each of these measures below.

Measures of Central Tendency

Statistical measures of central tendency give that one number that is often used to summarize the distribution of a variable. They may be referred to generically as the "average." There are three main

central tendency measures: mode, median, and mean. In addition, Tukey devised the 5% trimmed mean.

- **Mode:** The mode for any variable is merely the group or class that contains the most cases. If two or more groups contain the same highest number of cases, the distribution is said to be "multimodal." This measure is more typically used on nominal or ordinal data and can easily be determined by examining a frequency table.
- **Median** - If all the cases for a variable are arranged in order according to their value, the median is that value that splits the cases into two equally sized groups. The median is the same as the 50th percentile. Medians are resistant to extreme scores, and so are considered robust measures of central tendency.
- **Mean**: - The mean is the simple arithmetic average of all the values in the distribution (i.e., the sum of the values of all cases divided by the total number of cases). It is the most commonly reported measure of central tendency. The mean along with the associated measures of dispersion are the basis for many statistical techniques.
- **5% trimmed mean** - The 5% trimmed mean is the mean calculated after the extreme upper 5% and the extreme lower 5% of the data values are dropped. Such a measure is resistant to extreme values.

The specific measure that you choose will depend on a number of factors, most importantly the level of measurement of the variable. The mean is considered the most "powerful" measure of the three classic measures of central tendency. However, it is good practice to compare the median, mean, and 5% trimmed mean to get a more complete understanding of a distribution.

Measures of Dispersion

Measures of dispersion or variability describe the degree of spread, dispersion, or variability around the central tendency measure. You might think of this as a measure of the extent to which observations cluster within the distribution. There are a number of measures of dispersion, including simple measures such as maximum, minimum, and range, common statistical measures, such as standard deviation and variance, as well as the interquartile range (IQR).

- **Maximum:** Simply the highest value observed for a particular variable. By itself, it can tell us nothing about the shape of the distribution, merely how high the top value is.
- **Minimum:** The lowest value in the distribution and, like the maximum, is only useful when reported in conjunction with other statistics.
- **Range:** The difference between the maximum and minimum values gives a general impression of how broad the distribution is. It says nothing about the shape of a distribution and can give a distorted impression of the data if just one case has an extreme value.
- **Variance:** Both the variance and standard deviation provide information about the amount of spread around the mean value. They are overall measures of how clustered around the mean the data values are. The variance is calculated by summing the square of the difference between the value and the mean for each case and dividing this quantity by the number of cases minus 1. If all cases had the same value, the variance (and standard deviation) would be zero. The variance measure is expressed in the units of the variable squared. This can cause difficulty in interpretation, so more often the standard deviation is used. In general terms, the larger the variance, the more spread there is in the data, the smaller the variance, the more the data values are clustered around the mean.
- **Standard Deviation:** The standard deviation is the square root of the variance which restores the value of variability to the units of measurement of the original variable. It is therefore easier to interpret. Either the variance or standard deviation is often used in conjunction with the mean as a basis for a wide variety of statistical techniques.
- **Interquartile Range (IQR)** - This measure of variation is the range of values between the 25th and 75th percentile values. Thus, the IQR represents the range of the middle 50 percent of the sample and is more resistant to extreme values than the standard deviation.

Like the measures of central tendency, these measures differ in their usefulness with variables of different measurement levels. The variability measures, variance and standard deviation, are used in conjunction with the mean for statistical evaluation of the distribution of a scale variable. The other measures of dispersion, although less useful statistically, can provide useful descriptive information about a variable.

Apply Your Knowledge

1. True or false? The mode is that value that splits the cases into two equally sized groups.

2. True or false? Consider the table depicted below. The salaries of men are clustered tighter around their mean than the salaries of women around their mean?

Report

Current Salary

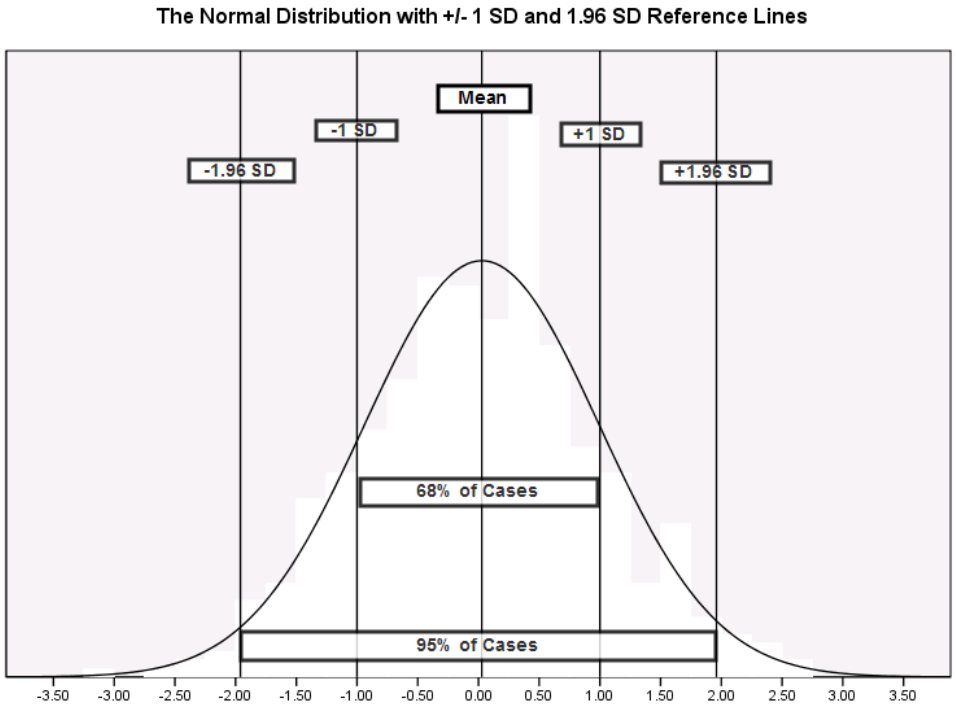
Gender	Mean	Std. Deviation
Female	26693.9	12436.9
Male	41557.0	19612.4

2.4 *Normal Distributions*

An important statistical concept is that of the normal distribution. This is a frequency (or probability) distribution which is symmetrical and is often referred to as the normal bell-shaped curve. The histogram below illustrates a normal distribution. The mean, median and mode exactly coincide in a perfectly normal distribution. And the proportion of cases contained within any portion of the normal curve can be exactly calculated mathematically.

Its symmetry means that 50% of cases lie to either side of the central point as defined by the mean. Two of the other most frequently-used representations are the portions lying between plus and minus one standard deviation of the mean (containing approximately 68% of cases) and that between plus and minus 1.96 standard deviations (containing approximately 95% of cases, sometimes rounded up to 2.00 for convenience). Thus, if a variable is normally distributed, we expect 95% of the cases to be within roughly 2 standard deviations from the mean.

Figure 2.1 The Normal Distribution



Many naturally occurring phenomena, such as height, weight and blood pressure, are distributed normally. Random errors also tend to conform to this type of distribution. It is important to understand the properties of normal distributions and how to assess the normality of particular distributions because of their theoretical importance in many inferential statistical procedures.

2.5 *Standardized (Z-) Scores*

The properties of the normal distribution allow us to calculate a *standardized score*, often referred to as a *z-score*, which indicates the number of standard deviations above or below the sample mean for each value. Standardized scores can be used to calculate the relative position of each value in the distribution. Z-scores are most often used in statistics to standardize variables of unequal scale units for statistical comparisons or for use in multivariate procedures.

For example, if you obtain a score of 68 out of 100 on a test of verbal ability, this information alone is not enough to tell how well you did in relation to others taking the test. However, if you know the mean score is 52.32, the standard deviation 8.00 and the scores are normally distributed, you can calculate the proportion of people who achieved a score at least as high as your own.

The standardized score is calculated by subtracting the mean from the value of the observation in question (68-52.32 = 15.68) and dividing by the standard deviation for the sample (15.68/8 = 1.96).

$$\text{Standardized Score} = \frac{\text{Case Score - Sample Mean}}{\text{Standard Deviation}}$$

Therefore, the mean of a standardized distribution is 0 and the standard deviation is 1.

In this case, your score of 68 is 1.96 standard deviations above the mean.

The histogram of the normal distribution above displays the distribution as a Z-score so the values on the x-axis are standard deviation units. From this figure, we can see only 2.5% of the cases are likely to have a score above 68 on the verbal ability test (1.96 standardized score). The normal distribution table (see below), found in an appendix of most statistics books, shows proportions for z-score values.

Table 2.5 Normal Distribution Table

Z-Score:	*Probability:*			*Z-Score:*	*Probability:*	
	One-tailed	*Two-tailed*			*One-tailed*	*Two-tailed*
0.0	.50000	1.00000		2.5	.00621	.01242
0.1	.46017	.92034		2.6	.00466	.00932
0.2	.42074	.84148		2.7	.00347	.00693
0.3	.38209	.76418		2.8	.00256	.00511
0.4	.34458	.68916		2.9	.00187	.00373
0.5	.30854	.61708		3.0	.00135	.00270
0.6	.27425	.54851		3.1	.00097	.00194
0.7	.24196	.48393		3.2	.00069	.00137
0.8	.21186	.42371		3.3	.00048	.00097
0.9	.18406	.36812		3.4	.00034	.00067
1.0	.15866	.31731		3.5	.00023	.00047
1.1	.13567	.27133		3.6	.00016	.00032
1.2	.11507	.23014		3.7	.00011	.00022
1.3	.09680	.19360		3.8	.00007	.00014
1.4	.08076	.16151		3.9	.00005	.00010
1.5	.06681	.13361		4.0	.00003	.00006
1.6	.05480	.10960		4.1	.00002	.00004
1.7	.04457	.08913		4.2	.00001	.00003
1.8	.03593	.07186		4.3	.00001	.00002
1.9	.02872	.05743		4.4	.00001	.00001
1.96	.02500	.05000		4.5	.00000	.00001
2.0	.02275	.04550		4.6	.00000	.00000
2.1	.01786	.03573		4.7	.00000	.00000
2.2	.01390	.02781		4.8	.00000	.00000
2.3	.01072	.02145		4.9	.00000	.00000
2.4	.00820	.01640		5.0	.00000	.00000

A score of 1.96, for example, corresponds to a value of .025 in the “one-tailed” column and .050 in the “two-tailed” column. The former means that the probability of obtaining a z-score at least as large as +1.96 is .025 (or 2.5%), the latter that the probability of obtaining a z-score of more than +1.96 or less than -1.96 is .05 (or 5%) or 2.5% at each end of the distribution. You can see these cutoffs in the histogram above.

As we mentioned, another advantage of standardized scores is that they allow for comparisons on variables measured in different units. For example, in addition to the verbal test score, you might have a mathematics test score of 150 out of 200 (or 75%). Although it appears that you did better on the mathematics test from the percentages alone, you would need to calculate the z-score for the mathematics test and compare the z-scores in order to answer the question.
You might want to compute z-scores for a series of variables and determine whether certain subgroups of your sample are, on average, above or below the mean on these variables by requesting descriptive statistics or using the Case Summaries procedure. For example, you might want to compare a customer’s yearly revenue using z-scores.

2.6 *Requesting Standardized (Z-) Scores*

The **Descriptives** procedure has an option to calculate standardized score variables. A new variable containing the standardized values is calculated for the specified variables. Creating standardized scores is accomplished by following these steps:

1) Choose variables to transform into standardized-scores.
2) Review the new variables that were created.

2.7 *Standardized (Z-) Scores Output*

The **Descriptives** procedure provides descriptive statistics of the original variables. The standardized variables will appear in the Data Editor.

Figure 2.2 Example of Descriptives Output

Descriptive Statistics

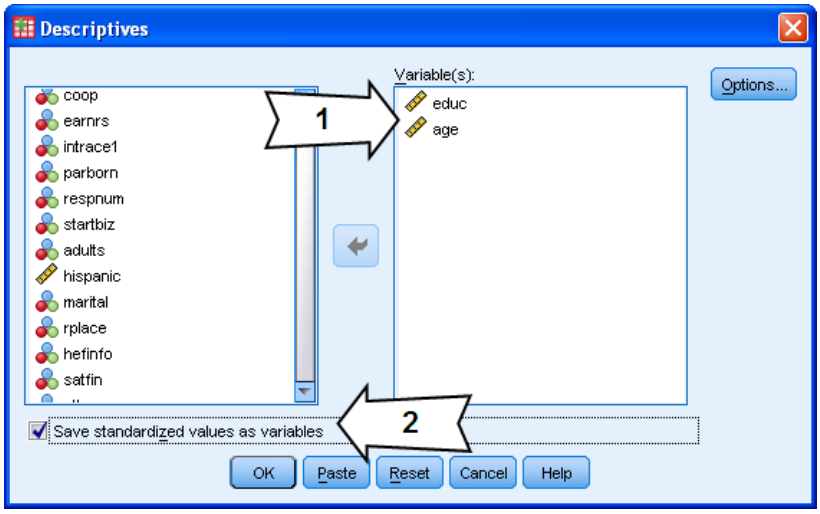
	N	Minimum	Maximum	Mean	Std. Deviation
HIGHEST YEAR OF SCHOOL COMPLETED	2018	0	20	13.43	3.079
AGE OF RESPONDENT	2013	18	89	47.71	17.351
Valid N (listwise)	2008				

2.8 *Procedure: Descriptives for Standardized (Z-) Scores*

The **Descriptives** procedure is accessed from the **Analyze…Descriptive Statistics…Descriptives** menu choice. With the **Descriptives** dialog box open:

1) Place one or more scale variables in the Variable(s) box.
2) Select the Save standardized values as variables box.

Figure 2.3 Descriptives Dialog Box to Create Z-Scores



2.9 *Demonstration: Descriptives for Z-Scores*

We will work with the *Census.sav* data file in this example. We create standardized scores for number of years of education (*educ*) and age of respondent (*age*). We would like to determine where respondents fall on the distribution of these variables.

Detailed Steps for Z-Scores

1) Place the variable ***educ*** and ***age*** in the Variable(s) box.
2) Select the **Save standardized values as variables** box.

Results from Z-Scores

By default, the new variable name is the old variable name prefixed with the letter "Z". Two new variables, *zeduc* and *zage,* containing the z-scores of the two variables, are created at the end of the data file. These variables can be saved in your file and used in any statistical procedure.

We observe that:

- The first person (row) in the data file is below the average on education but above the average on age.

Figure 2.4 Two Z-score Variables in the Data Editor

*Census.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

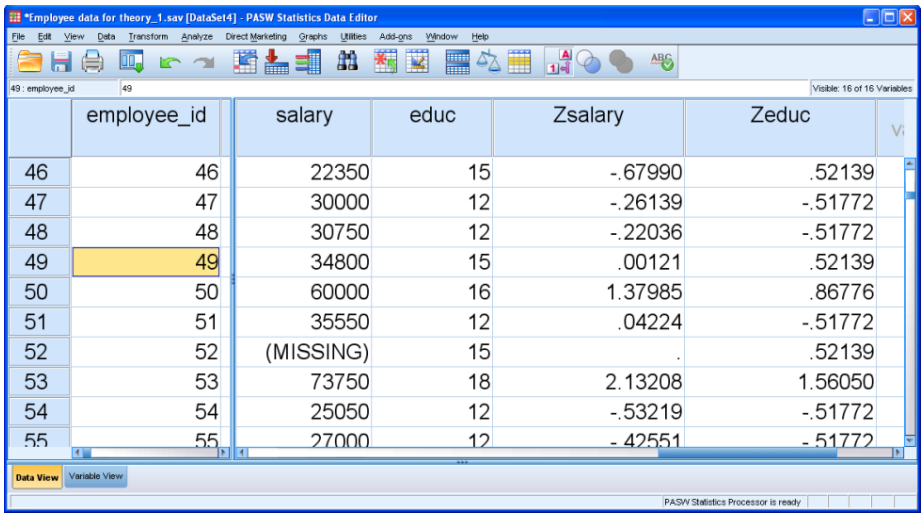
1 : id | 1

	INCOME_ACTUAL	speduc	Zeduc	Zage
1	.	97	-.46513	.07444
2	.	19	2.13315	.01681
3	999999.00	14	-.14034	-.04083
4	44039.89	97	-1.11470	-.90534
5	18890.12	16	.83401	1.40002
6	.	16	1.15880	.82368
7	70966.91	97	1.80836	-.44427

Apply Your Knowledge

1. True or false? Only for variables of measurement level scale in PASW Statistics is it meaningful to calculate standardized scores?

2. Consider the data below, where we computed standardized values for the variables *educ* (highest year of education) and *salary* (salary in dollars). Which of the following statements are correct?
    a. The observation with employee_id=49 has a salary very close to the mean salary.
    b. The observation with employee_id=50 has a salary that is more than one standard deviation above the mean.
    c. The observation with employee_id=46 is more extreme in her education than in salary.



Additional Resources

For additional information on Level of Measurement and Statistical Tests, see:

Andrews, Frank M, Klem, L., Davidson, T.N., O'Malley, P.M. and Rodgers, W.L. 1981. *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Further Info

Velleman, Paul F. and Wilkinson, L. 1993. "Nominal, Ordinal and Ratio Typologies are Misleading for Classifying Statistical Methodology," *The American Statistician*, vol. 47, pp. 65-72.

2.10 *Lesson Summary*

We explored the concept of the level of measurement and the appropriate summary statistics given level of measurement. We also discussed the normal distribution and z-scores.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Determine the level of measurement of variables and obtain appropriate summary statistics based on the level of measurement

To support the achievement of the primary objective, students should now also be able to:

- Describe the levels of measurement used in PASW Statistics
- Use measures of central tendency and dispersion
- Use normal distributions and z-scores

2.11 *Learning Activity*

The overall goal of this learning activity is to create standardized (Z-) scores for several variables. In this set of learning activities you will use the *Drinks.sav* data file.

Supporting Materials

The file *Drinks.sav*, a PASW Statistics data file that contains hypothetical data on 35 beverages. Included is information on their characteristics (e.g., % alcohol), price, origin, and a rating of quality.

1. Create standardized scores for all scale variables (*price* through *alcohol*). Which beverages have positive standardized scores on every variable? What does this mean?

2. What is the most extreme z-score on each variable? What is the most extreme z-score across all variables?

3. What beverage is most typical of all beverages, that is, has z-score values closest to 0 for these variables?

4. If the variable is normally distributed, what percentage of cases should be above 1 standard deviation from the mean or below 1 standard deviation from the mean? Calculate this percentage for a couple of the variables. Is the percentage of beverages with an absolute z-score above 1 close to the theoretical value?

Lesson 3: Data Distributions for Categorical Variables

3.1 *Objectives*

After completing this lesson students will be able to:

- Run the **Frequencies** procedure to obtain appropriate summary statistics for categorical variables

To support the achievement of this primary objective, students will also be able to:

- Use the options in the **Frequencies** procedure
- Interpret the results of the **Frequencies** procedure

3.2 *Introduction*

As a first step in analyzing data, one must gain knowledge of the overall distribution of the individual variables and check for any unusual or unexpected values. You often want to examine the values that occur in a variable and the number of cases in each. For some variables, you want to summarize the distribution of the variable by examining simple summary measures including the mode, median, and minimum and maximum values. In this chapter, we will review tables and graphs appropriate for describing categorical (nominal and ordinal) variables.

Business Context

Summaries of individual variables provide the basis for more complex analyses. There are a number of reasons for performing single variable analyses. One would be to establish base rates for the population sampled. These rates may be of immediate interest: What percentage of our customers is satisfied with services this year? In addition, studying a frequency table containing many categories might suggest ways of collapsing groups for a more succinct and statistically appropriate table. When studying relationships between variables, the base rates of the separate variables indicate whether there is a sufficient sample size in each group to proceed with the analysis. A second use of such summaries would be as a data-checking device—unusual values would be apparent in a frequency table.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

3.3 *Using Frequencies to Summarize Nominal and Ordinal Variables*

The most common technique for describing categorical data is a frequency analysis which provides a summary table indicating the number and percentage of cases falling into each category of a variable, as well as the number of valid and missing cases. We can also use the mode, which indicates the category with the highest frequency, and, if there is a large number of categories, the median (for ordinal variables), which is the value above and below which half the cases fall.

Figure 3.1 Typical Frequencies Table

RESPONDENTS SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	929	45.9	45.9	45.9
	FEMALE	1094	54.1	54.1	100.0
	Total	2023	100.0	100.0	

To represent the frequencies graphically we use bar or pie charts.

- A pie chart displays the contribution of parts to a whole. Each slice of a pie chart corresponds to a group that is defined by a single grouping variable.
- A bar chart displays the count for each distinct value or category as a separate bar, allowing you to compare categories vertically.

Figure 3.2 Pie Chart illustrated

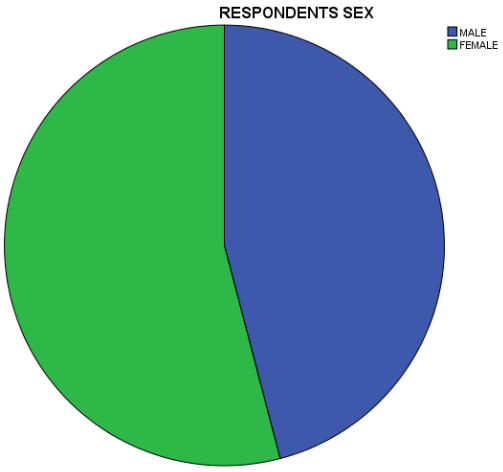
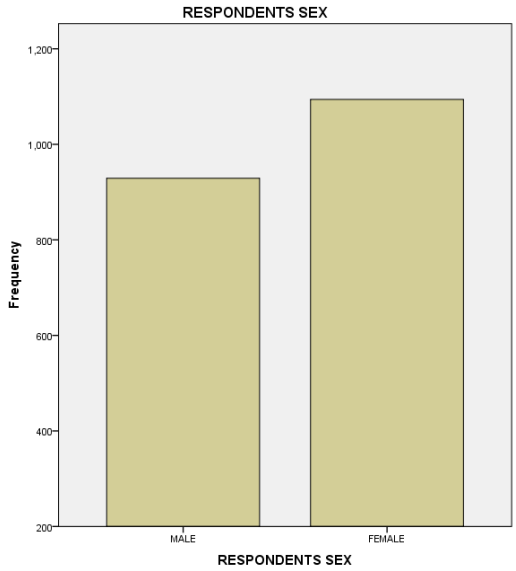


Figure 3.3 Bar Chart Illustrated



3.4 *Requesting Frequencies*

Requesting **Frequencies** is accomplished by following these steps:

1) Choose variables for the **Frequencies** procedure.
2) Request additional summary statistics and graphs.
3) Review the procedure output to investigate the distribution of the variables including:
    a. Frequency Tables
    b. Graphs

3.5 *Frequencies Output*

The information in the frequency table is comprised of counts and percentages:

- The Frequency column contains counts, i.e., the number of occurrences of each data value.
- The Percent column shows the percentage of cases in each category relative to the number of cases in the entire data set, including those with missing values.
- The Valid Percent column contains the percentage of cases in each category relative to the number of valid (non-missing) cases.
- The Cumulative percentage column contains the percentage of cases whose values are less than or equal to the indicated value. Cumulative percent is only useful for variables that are ordinal.

Figure 3.4 Example of Frequency Output

GENERAL HAPPINESS

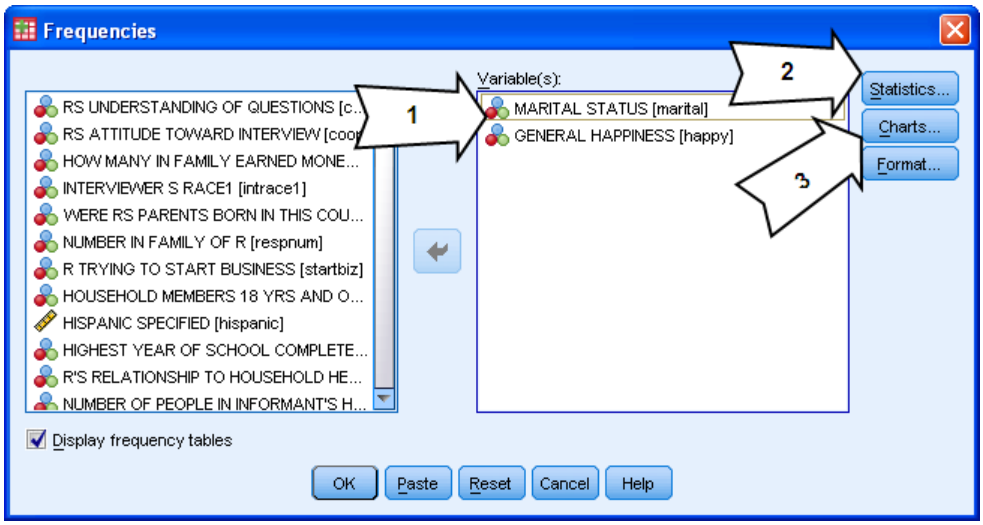
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	599	29.6	29.7	29.7
	PRETTY HAPPY	1100	54.4	54.6	84.3
	NOT TOO HAPPY	316	15.6	15.7	100.0
	Total	2015	99.6	100.0	
Missing	DK	2	.1		
	NA	6	.3		
	Total	8	.4		
Total		2023	100.0		

3.6 *Procedure: Frequencies*

The **Frequencies** procedure is accessed from the **Analyze…Descriptive Statistics…Frequencies** menu choice. With the **Frequencies** dialog box open:

1) Place one or more variables in the Variable(s) box.
2) Open the Statistics dialog to request summary statistics.
3) Open the Charts dialog to request graphs.

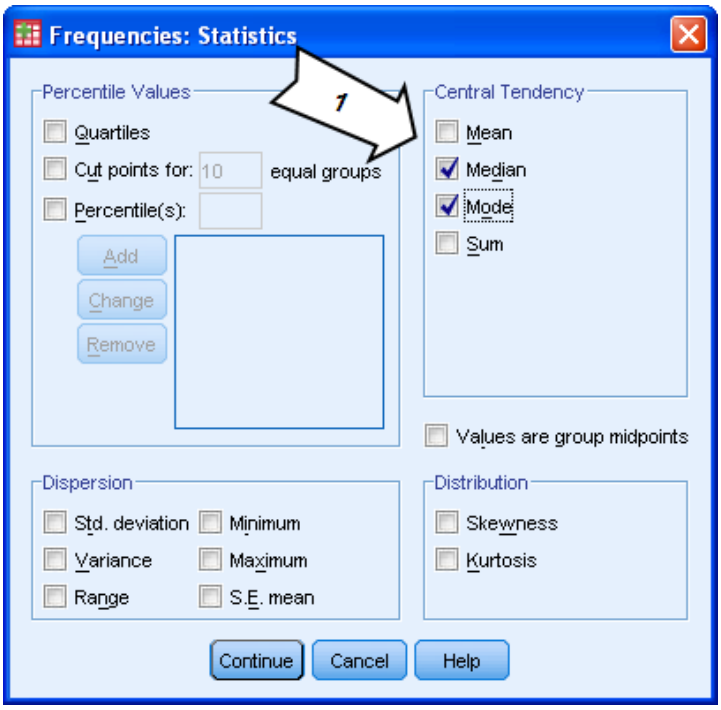
Figure 3.5 Frequencies Dialog Box



In the Statistics dialog box:

1) Ask for the appropriate measures of central tendency and dispersion.

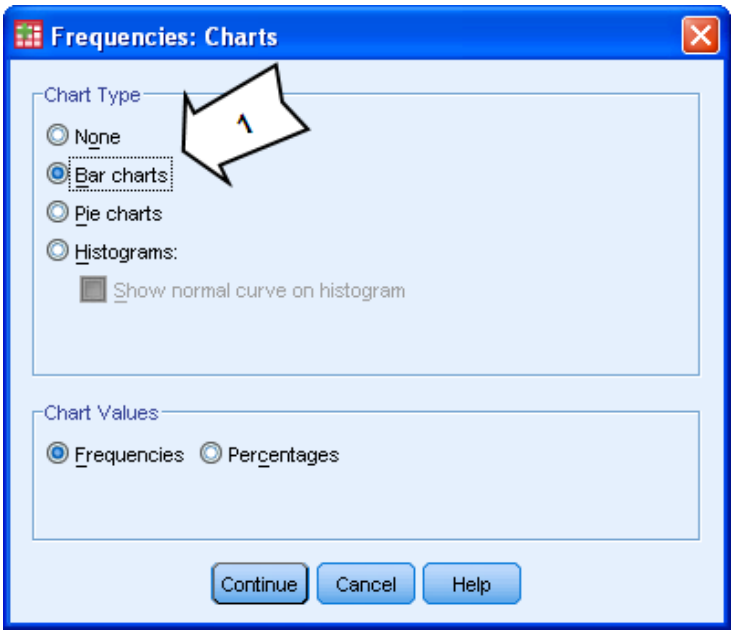
Figure 3.6 Frequencies: Statistics Dialog Box



In the Charts dialog:

1) Ask for the appropriate chart based on the scale of measurement of the variable.

Figure 3.7 Frequencies: Charts Dialog Box



3.7 *Demonstration: Frequencies*

We will work with the *Census.sav* data file in this lesson. In this example we examine the distribution of the variables *marital* and *happy*. These variables are either nominal or ordinal in scale of measurement.

Detailed Steps for Frequencies

1) Place the variables ***marital*** and ***happy*** in the Variable(s) box
2) In the Statistics dialog, select **Mode** and **Median** in the Statistics dialog
3) In the Charts dialog, select **Bar Chart** in the Chart Types area and **Percentages** in the Chart Value area

Results from Frequencies

The first table produced is the table labeled Statistics.

Figure 3.8 Statistics for Marital Status and General Happiness

Statistics

		MARITAL STATUS	GENERAL HAPPINESS
N	Valid	2018	2015
	Missing	5	8
Median		2.00	2.00
Mode		1	2

This table shows the number of cases having a valid value on *Marital Status* (2018) and *General Happiness* (2015), the number of cases having a (user- or system-) missing value (5 and 8, respectively) and the Mode and Median. The mode, the category that has the highest frequency, is a value of 1 and 2 respectively, and represents the category of “Married” for *marital* and the “Pretty Happy” group for *happy*. The median, the middle point of the distribution (50th percentile), is a value of 2 for both variables.

The second table shows the frequencies and percentages for each variable. This table confirms that almost half of the respondents are married. Since there is almost no missing data for marital status, the percentages in the Percent column and in the Valid Percent column are almost identical.

Figure 3.9 Frequency Table of Marital Status

MARITAL STATUS

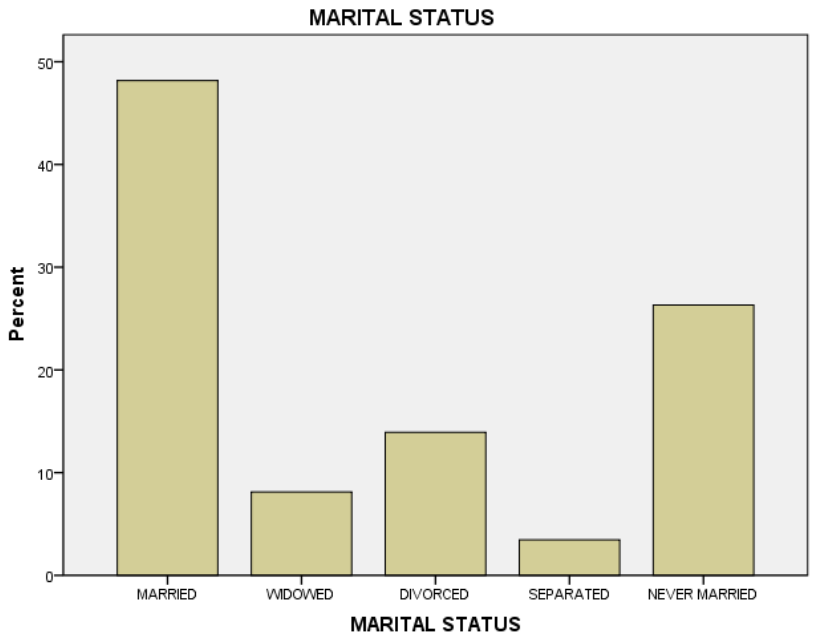
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MARRIED	972	48.0	48.2	48.2
	WIDOWED	164	8.1	8.1	56.3
	DIVORCED	281	13.9	13.9	70.2
	SEPARATED	70	3.5	3.5	73.7
	NEVER MARRIED	531	26.2	26.3	100.0
	Total	2018	99.8	100.0	
Missing	NA	5	.2		
Total		2023	100.0		

Examine the table. Note the disparate category sizes. About half of the sample is married, and there is one category that has less than 5% of the cases. Before using this variable in a crosstabulation

analysis, should you consider combining some of the categories with fewer cases? Decisions about collapsing categories usually have to do with which groups need to be kept distinct in order to answer the research question asked, and the sample sizes for the groups. For example, could we create a "was previously married" group?

The bar chart summarizes the distribution that we observed in the frequency table and allows us to "see" the distribution.

Figure 3.10 Bar Chart of Marital Status



Tip

For a nominal variable (where the order of the categories is arbitrary) sorting the table and graph descending on counts gives better insight in what the main categories are (use the Format subdialog box to sort descending on counts).

For the variable *happy*, over half of the people fall into one category, pretty happy. Might it be interesting to look at the relationship between this variable and marital status: to what extent is general happiness related to marital status?

Figure 3.11 Frequency Table of General Happiness

GENERAL HAPPINESS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	599	29.6	29.7	29.7
	PRETTY HAPPY	1100	54.4	54.6	84.3
	NOT TOO HAPPY	316	15.6	15.7	100.0
	Total	2015	99.6	100.0	
Missing	DK	2	.1		
	NA	6	.3		
	Total	8	.4		
Total		2023	100.0		

Next we view a bar chart based on the general happiness variable. Does the picture make it easier to understand the distribution?

Figure 3.12 Bar Chart of General Happiness

Note

For an ordinal variable, sorting the categories on descending/ascending counts (which was useful for nominal variables) will disturb the natural order of categories and so is not as useful for an ordinal variable.

Apply Your Knowledge

1. See the output below. Which statements are correct?
    a. The median is an appropriate statistic to report for the variable *region*.
    b. The region that has the highest frequency is the North.
    c. The cumulative percent is meaningful for *region*.
    d. The columns Percent and Valid Percent are identical because there are no missing values on *region*

Statistics

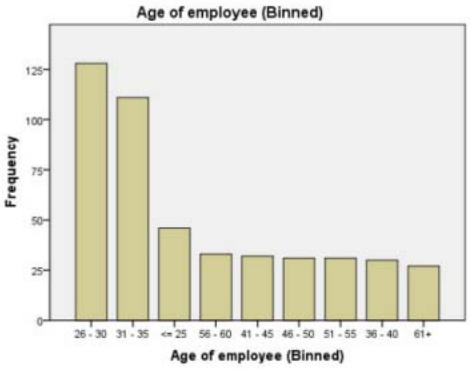
region

N	Valid	474
	Missing	0
Median		2.00
Mode		1

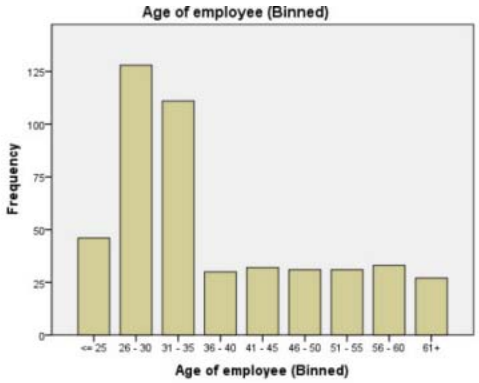
region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	North	133	28.1	28.1	28.1
	East	127	26.8	26.8	54.9
	South	114	24.1	24.1	78.9
	West	100	21.1	21.1	100.0
	Total	474	100.0	100.0	

2. See the output below. Which bar chart is best to present the distribution of the ordinal variable *age of employees*, in categories? Bars sorted descending on count (A) or sorted on ascending value (B)?



A



B

3. See the table below (a frequency table of *HAPPINESS OF MARRIAGE*, with those not married defined as missing). Which statements are correct?
    a. 29.5% of those married are very happy in their marriage.
    b. The mode is the category pretty happy.
    c. 96.9% of those married are pretty happy or very happy

HAPPINESS OF MARRIAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	596	29.5	61.5	61.5
	PRETTY HAPPY	343	17.0	35.4	96.9
	NOT TOO HAPPY	30	1.5	3.1	100.0
	Total	969	47.9	100.0	
Missing	NOT MARRIED	1054	52.1		
Total		2023	100.0		

Additional Resources

Further Info

For additional information on how to present data in tables and graphs, see:

Few, Stephen. 2004. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press

3.8 *Lesson Summary*

In this lesson we used the **Frequencies** procedure to explore the distribution of categorical variables, via both tables and graphs.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Run the **Frequencies** procedure to obtain appropriate summary statistics for categorical variables

To support the achievement of the primary objective, students should now also be able to:

- Use the options in the **Frequencies** procedure
- Interpret the results of the **Frequencies** procedure

3.9 Learning Activity

The overall goal of this learning activity is to run **Frequencies** to explore the distributions of several variables. In the exercises you will use the data file *Census.sav*.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. Run the **Frequencies** procedure on the following variables: *sex*, *wrkstat* (Labor Force Status), *paeduc* (Father's highest degree), and *satjob* (Job or Housework). What is the scale of measurement for each? Request appropriate summary statistics and charts.

2. For which of these variables is it appropriate to use the median? What conclusions can you draw about the distributions of these variables?

3. What percent of respondents have a bachelor's degree, or higher? What percent of respondents are working?
4. How might you combine some of the categories of *wrkstat* to insure that there are a sufficient number of respondents in each category?

Lesson 4: Data Distributions for Scale Variables

4.1 *Objectives*

After completing this lesson students will be able to:

- Request and interpret appropriate summary statistics for scale variables

To support the achievement of this primary objective, students will also be able to:

- Use the options in the **Frequencies**, **Descriptives**, and **Explore** procedures
- Interpret the results of the **Frequencies**, **Descriptives**, and **Explore** procedures

4.2 *Introduction*

As a first step in analyzing your data, you must first gain knowledge of the overall distribution of the individual variables and check for any unusual or unexpected values. You often want to examine the values that occur in a variable and the number of cases in each. For some variables, you want to summarize the distribution of the variable by examining simple summary measures including minimum and maximum values for the range. Frequently used summary measures describe the central tendency of the distribution, such as the arithmetic mean, and dispersion, the spread around the central point. In this lesson, we will review tables and graphs appropriate for describing scale (interval and ratio) variables.

Business Context

Summaries of individual variables provide the basis for more complex analyses. There are a number of reasons for performing single variable analyses. One would be to establish base rates for the population sampled. These rates may be of immediate interest: What is the average customer satisfaction? In addition, studying distributions might suggest ways of collapsing information for a more succinct and statistically appropriate table. When studying relationships between variables, the base rates of the separate variables indicate whether there is a sufficient sample size in each group to proceed with the analysis. A second use of such summaries would be as a data-checking device, as unusual values would be apparent in tables.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

4.3 *Summarizing Scale Variables Using Frequencies*

When working with categorical variables, frequency tables containing counts and percentages are appropriate summaries. For a scale variable, counts and percentages may still be of interest, especially when the variables can take only a limited number of distinct values. For example, when working with a one to five point rating scale we might be very interested in knowing the percentage of respondents who reply “Strongly Agree.” However, as the number of possible response values

increases, frequency tables based on interval scale variables become less useful. Suppose we asked respondents for their family income to the nearest dollar? It is likely that each response would have a different value and so a frequency table would be quite lengthy and not particularly helpful as a summary of the variable. In data cleaning, you might find a frequency table useful for examining possible clustering of cases on specific values or looking at cumulative percentages. But, beware of using frequency tables for scale variables with many values as they can be very long.

If the variables of interest are scale we can expand the summaries to include means, standard deviations and other statistical measures. You will want to spend some time looking over the summary statistics you requested. Do they make sense, or is something unusual?

For a categorical variable, we request a pie chart or a bar chart to graphically display the distribution of the variable. For a scale variable, a histogram is used to display the distribution.

Tip

A normal curve can be superimposed on the histogram and helps you to judge whether the variable is normally distributed.

4.4 *Requesting Frequencies*

Requesting statistics and a graphical display is accomplished by following these steps:

1) Select variables in the **Frequencies** procedure.
2) Request additional summary statistics and graphs.
3) Review the procedure output to investigate the distribution of the variables including:
    a. Frequency tables (if requested)
    b. Statistics tables
    c. Graphs

4.5 *Frequencies Output*

Statistics for the variable are presented in a separate table.

Figure 4.1 Example of Summary Statistics for Frequencies Output

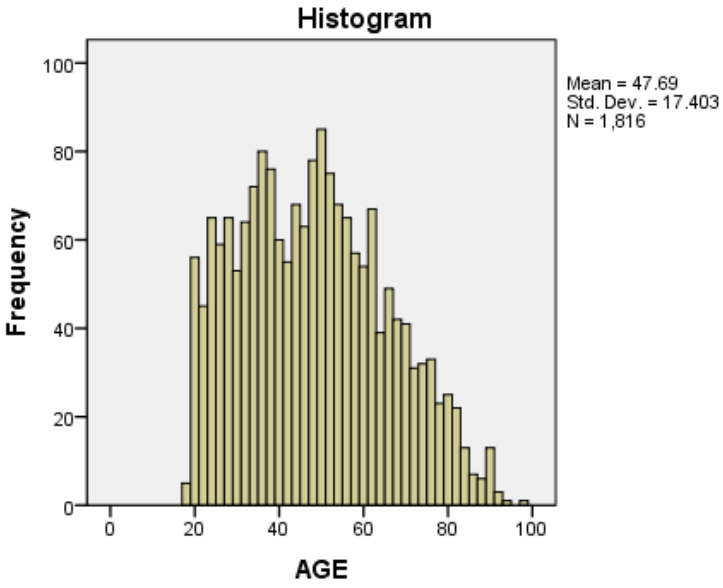
Statistics

NUMBER OF BROTHERS AND SISTERS

N	Valid	2021
	Missing	2
Mean		3.66
Median		3.00
Mode		2
Std. Deviation		3.188
Variance		10.164
Range		55
Minimum		0
Maximum		55

A histogram shows the distribution graphically. A histogram has bars, but, unlike the bar chart, they are plotted along an equal interval scale. The height of each bar is the count of values of a quantitative variable falling within the interval. A histogram shows the shape, center, and spread of the distribution.

Figure 4.2 Example of Histogram for Frequencies Output

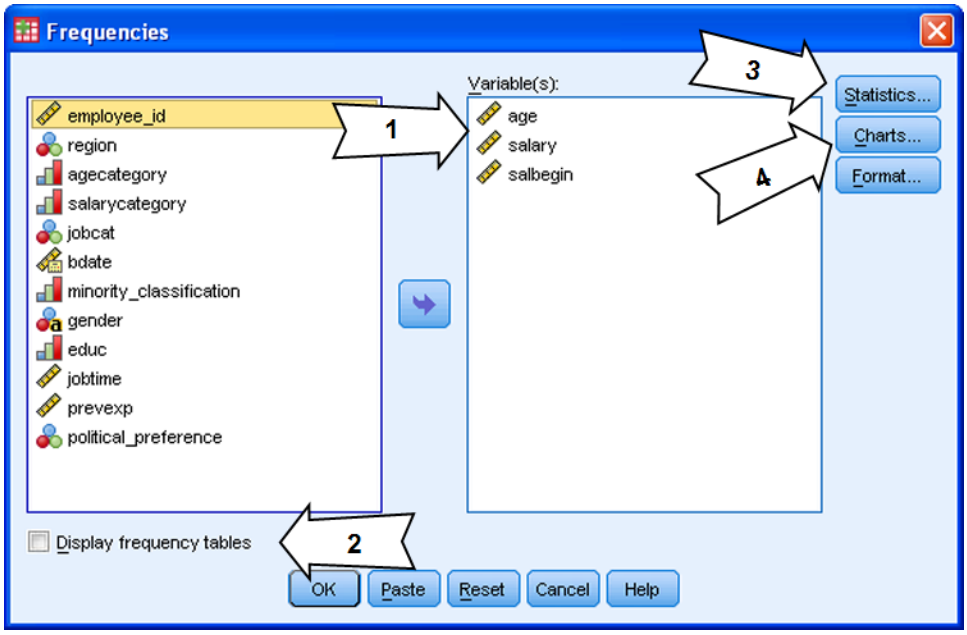


4.6 *Procedure: Frequencies*

The **Frequencies** procedure is accessed from the **Analyze…Descriptive Statistics…Frequencies** menu choice. With the **Frequencies** dialog box open:

1) Place one or more variables in the Variable(s) box.
2) Deselect the Display frequency tables check box for variables with many values.
3) Open the Statistics dialog to request summary statistics.
4) Open the Charts dialog to request graphs.

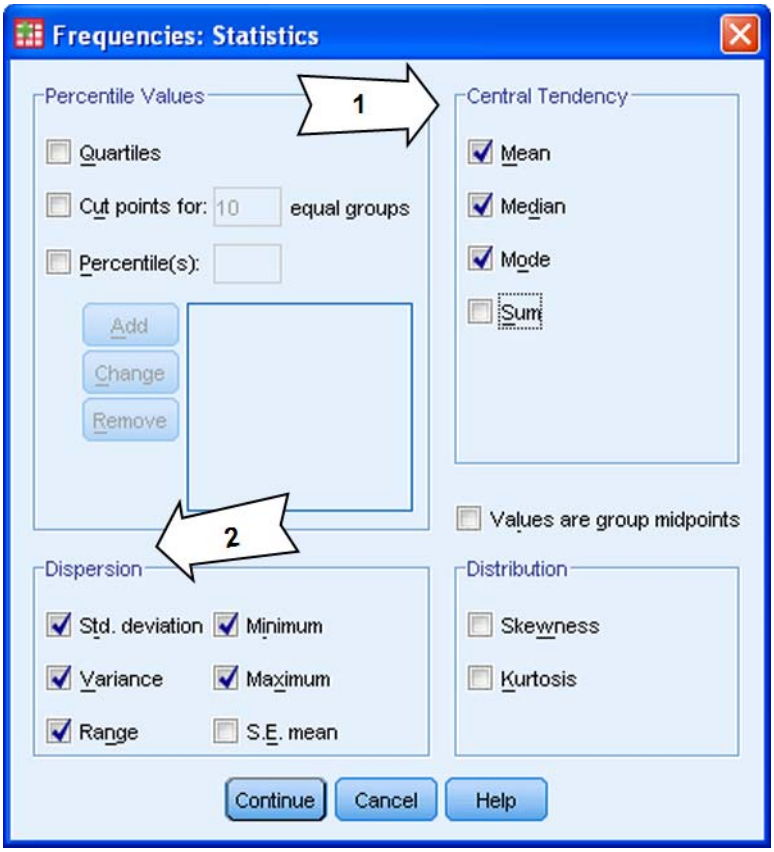
Figure 4.3 Frequencies Dialog Box



In the Statistics dialog:

1) Select appropriate measures of central tendency
2) Select appropriate measures of dispersion

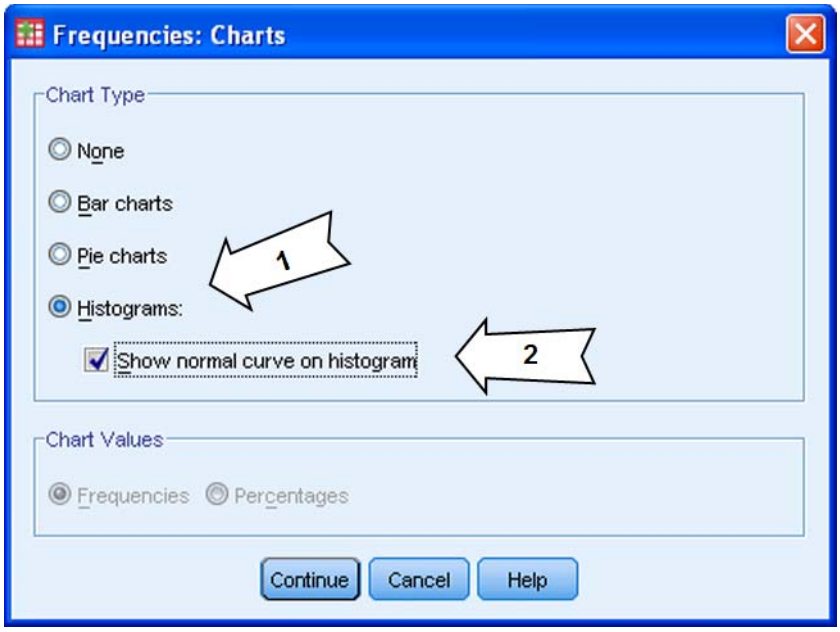
Figure 4.4 Frequencies: Statistics Dialog Box



In the Charts dialog:

1) Ask for a histogram for scale variables.
2) Optionally, superimpose a normal curve on the histogram.

Figure 4.5 Frequencies: Charts Dialog Box



4.7 *Demonstration: Frequencies*

We will work with the *Census.sav* data file in this lesson.

In this demonstration we examine the distribution of number of brothers and sisters (*sibs*) and respondent's age. We would like to see the distribution of these variables.

Detailed Steps for Frequencies

1) Place the variables ***sibs*** and ***age*** in the Variable(s) box
2) Deselect the **Display frequency tables** check box for variables with many values
3) Select **Mode**, **Median**, **Mean**, **Minimum**, **Maximum** and **Standard Deviation** in the Statistics dialog
4) Select **Histograms** and **Show normal curve on histogram** in the Charts dialog

Note

If you request histograms and summary statistics for scale variables with many categories, you might want to uncheck (turn off) *Display frequency tables* in the **Frequencies** dialog box, as there may be almost as many distinct values as there are cases in the data file.

Results from Frequencies

The table labeled Statistics shows the requested statistics.

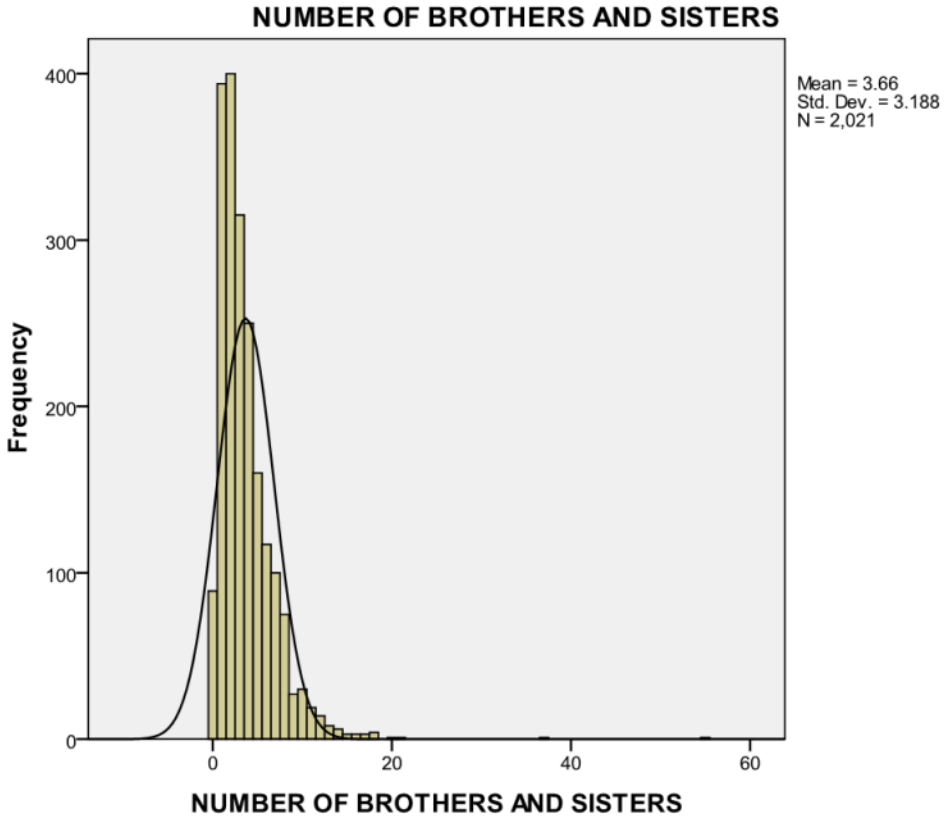
Figure 4.6 Summary Statistics for Number of Brothers and Sisters and Age of Respondent

Statistics

		NUMBER OF BROTHERS AND SISTERS	AGE OF RESPONDENT
N	Valid	2021	2013
	Missing	2	10
Mean		3.66	47.71
Median		3.00	47.00
Mode		2	50
Std. Deviation		3.188	17.351
Minimum		0	18
Maximum		55	89

This table shows the number of cases having a valid value on *sibs* (2021) and *age* (2013), the number of cases having a (user- or system-) missing value (2 and 10, respectively) and measures of central tendency and dispersion. The minimum value is 0 and the maximum value is 55 (seems unusual) for number of siblings. For *age*, the minimum value is 18 and the maximum value 89. Note, that the means and medians within each variable are similar, indicating that the variables are roughly normally distributed within the defined range. We can visually check the distribution of these variables with a histogram.

Figure 4.7 Histogram of Number of Brothers and Sisters



We can see that the lower range of values is truncated at 0 and the number of people is greatest between 0 to 6 siblings, although we do have some extreme values. The distribution is not normal.

Figure 4.8 Histogram of Age of Respondent

We can see that the lower range of values is truncated at 18 and the number of people is highest in the middle age values (the "baby boomers") with the number of cases tapering off at the higher ages as we would expect. Thus, the *age* variable for respondents of this sample of adults is roughly normally distributed.

Apply Your Knowledge

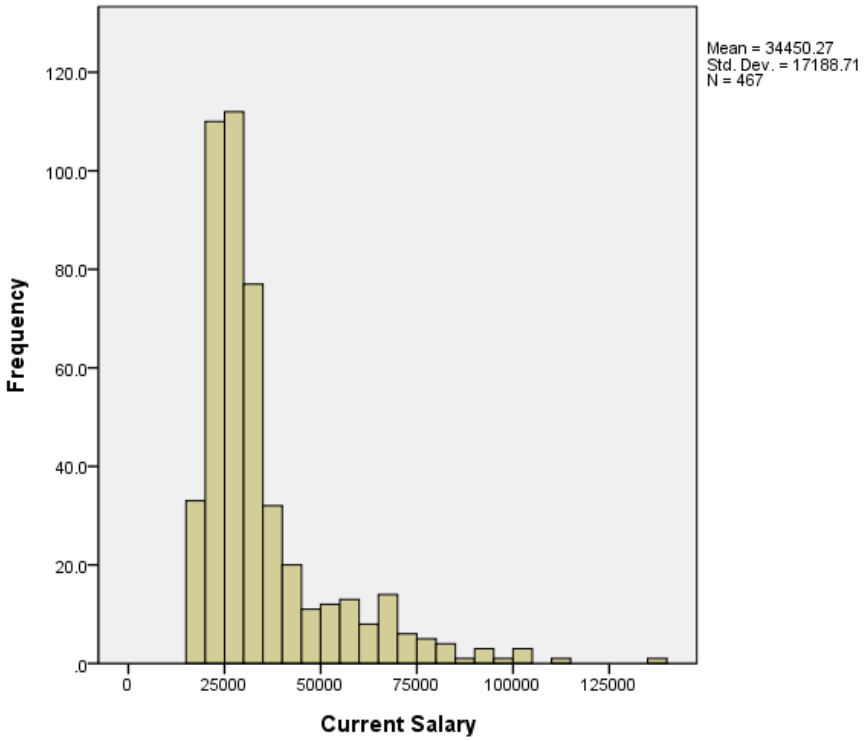
1. Suppose we have a variable *region* (with the categories north/east/south/west). Which of these statements is true?
    a. The mean is a meaningful statistic for *region*
    b. The standard deviation is a meaningful statistic for *region*
    c. The median is a meaningful statistic for *region*

2. See output below, with statistics for two variables: Current Salary and Beginning Salary (data collected on employees). Which statements are correct?
    a. There are 474 cases in the dataset
    b. Both variables are skewed to the right (meaning: there are employees with some large salaries compared to the average
    c. Half of the employees have a current salary below 30,750.
    d. The highest current salary is 135,000.

Statistics

		Current Salary	Beginning Salary
N	Valid	467	471
	Missing	7	3
Mean		34450.27	16997.72
Median		28650.00	15000.00
Mode		30750	15000
Std. Deviation		17188.710	7854.872
Variance		295451752.986	61699011.003
Range		119250	70980
Minimum		15750	9000
Maximum		135000	79980

3. See the histogram below for *Current Salary* (data collected on employees). Which of these statements is correct?
    a. The variable seems normally distributed
    b. The variable is skewed to the right
    c. The standard deviation would be smaller, if the case with salary of 135,000 would be removed from the histogram.



4.8 *Summarizing Scale Variables using Descriptives*

The **Descriptive** procedure is a good alternative to **Frequencies** when the objective is to summarize scale variables. **Descriptives** is usually used to provide a table of statistical summaries (means, standard deviations, variance, minimum, maximum, etc.) for several scale variables. The **Descriptives** procedure also provides a succinct summary of the number of cases with valid values for each variable included in the table as well as the number of cases with valid values for all variables included in the table. These summaries are quite useful in evaluating the extent of missing values in your data and in identifying variables with missing values for a large proportion of the data.

4.9 *Requesting Descriptives*

Running **Descriptives** is accomplished by following these steps:

1) Select variables for the **Descriptives** procedure.
2) Review the procedure output to investigate the distribution of the variables.

4.10 *Descriptives Output*

The figure below shows the **Descriptives** output table for a few variables.

Figure 4.9 Example Descriptives Output

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RESPONDENTS SEX	2023	1	2	1.54	.498
AGE	1816	18	97	47.69	17.403
R'S AGE WHEN 1ST CHILD BORN	1489	13	61	24.28	5.033
Valid N (listwise)	1333				

The minimum and maximum provide an efficient way to check for values outside the expected range. In general, this is a useful check for categorical variables as well. Thus, although mean and standard deviation are not relevant for respondent's sex, minimum and maximum for this variable show that there are no values outside the expected range.

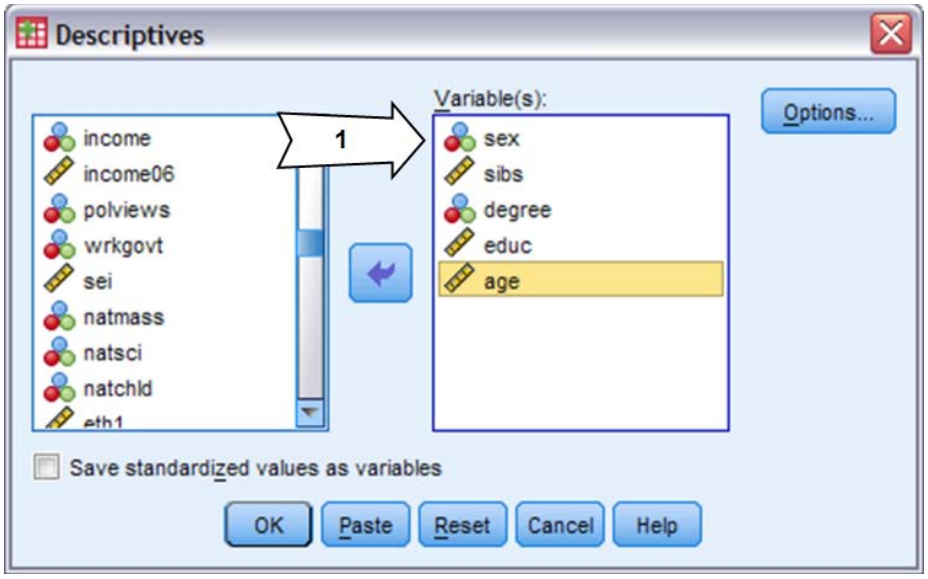
The last row in the table labeled Valid N (listwise) gives the number of cases that have a valid value on all of variables appearing in the table. In this example, 1333 cases have valid values for all three variables listed. Although this number is not particularly useful for this set of variables, it would be useful for a set of variables that you intended to use for a specific multivariate analysis. As you proceed with your analysis plans, it is helpful to know how many cases have complete information and which variables are likely to be the main sources of potential problems.

4.11 *Procedure: Descriptives*

The **Descriptives** procedure is accessed from the **Analyze…Descriptive Statistics…Descriptives** menu choice. With the **Descriptives** dialog box open:

1) Place one or more variables in the Variable(s) box.

Figure 4.10 Descriptives Dialog Box



Only numeric variables appear in the **Descriptives** dialog box. The Save standardized values as variables feature creates new variables that are standardized forms of the original variables. These new variables, referred to as z-scores, have values standardized to a mean of 0 and standard deviation of 1.

The Options button allows you to select additional summary statistics to display. You can also select the display order of the variables in the table (for example, by ascending or descending mean value).

4.12 *Demonstration: Descriptives*

We will work with the *Census.sav* data file in this lesson.

In this example we examine the summary statistics of number of siblings, respondent's age, education and respondent´s gender. We would like to see the summary statistics for these variables, as well as how much missing data there is, and if there are unusual cases.

Detailed Steps for Descriptives

1) Place the variables ***sex***, ***sibs***, ***educ***, and ***age*** in the Variable(s) box.

Results from Descriptives

The table labeled Descriptive Statistics contains the statistics.

Figure 4.11 Descriptives Output

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RESPONDENTS SEX	2023	1	2	1.54	.498
NUMBER OF BROTHERS AND SISTERS	2021	0	55	3.66	3.188
HIGHEST YEAR OF SCHOOL COMPLETED	2018	0	20	13.43	3.079
AGE OF RESPONDENT	2013	18	89	47.71	17.351
Valid N (listwise)	2006				

The column labeled N shows the number of valid observations for each variable in the table. We see there is little variation in the number of valid observations.

The number of valid cases can be a useful check on the data and help us determine which variables might be appropriate for specific analyses. Here 2006 cases have valid values for the entire set of questions.

The minimum and maximum provide an efficient way to check for values outside the expected range. Here the maximum for the variable *sibs* seems high and deserves further investigation.

4.13 *Summarizing Scale Variables using the Explore Procedure*

Exploratory data analysis (EDA) was primarily developed by John Tukey. He devised several statistical measures and plots designed to reveal data features that might not be readily apparent from standard statistical summaries. Exploratory data analysis can be viewed either as an analysis in its own right, or as a set of data checks that investigators perform before applying inferential testing procedures.

These methods are best applied to variables with at least ordinal (more commonly interval) or scale properties and which can take many different values. The plots and summaries would be less helpful for a variable that takes on only a few values (for example, a five-point scale).

4.14 *Requesting Explore*

Running **Explore** is accomplished with these steps:

1) Select variables on which to report statistics in the Dependent List box
2) Select grouping variables in the Factor box
3) Request additional summary statistics and graphs.
4) Review the procedure output to investigate the summary statistics and distribution of the variables including tables and graphs

Explore Output

The Descriptives table displays a series of descriptive statistics for *age*. From the previous table (not shown), we know that these statistics are based on 1763 respondents.

Figure 4.12 Summaries for Age of Respondent

Descriptives

			Statistic	Std. Error
AGE OF RESPONDENT	Mean		47.71	.387
	95% Confidence Interval for Mean	Lower Bound	46.95	
		Upper Bound	48.47	
	5% Trimmed Mean		47.25	
	Median		47.00	
	Variance		301.052	
	Std. Deviation		17.351	
	Minimum		18	
	Maximum		89	
	Range		71	
	Interquartile Range		26	
	Skewness		.308	.055
	Kurtosis		-.713	.109

First, several measures of central tendency appear: the *Mean, 5% Trimmed Mean*, and *Median*. These statistics attempt to describe with a single number where data values are typically found, or the center of the distribution. Useful information about the distribution can be gained by comparing these values to each other. If the mean were considerably above or below the median and trimmed mean, it would suggest a skewed or asymmetric distribution.

The measures of central tendency are followed in the table by several measures of dispersion or variability. These indicate to what degree observations tend to cluster or be widely separated. Both the standard deviation (*Std.Deviation*) and *Variance* (standard deviation squared) appear. The standard error (*Std.Error*) is an estimate of the standard deviation of the mean if repeated samples of the same size (here 1763) were taken. It is used in calculating the *95% confidence interval* for the sample mean. Technically speaking, if we would draw 100 samples of this size (1763) and construct a 95% confidence for the mean for each of the 100 samples, then the expectation is that 95 out of these 100 intervals will contain the (unknown) population mean. Also appearing is the *Interquartile Range (*often abbreviated to IQR*)* which is essentially the range between the 25th and the 75th percentile values. It is a variability measure more resistant to extreme scores than the standard deviation. We also see the *Minimum, Maximum* and *Range*.

The final two statistical measures, *Skewness* and *Kurtosis*, provide numeric summaries about the shape of the distribution of the data.

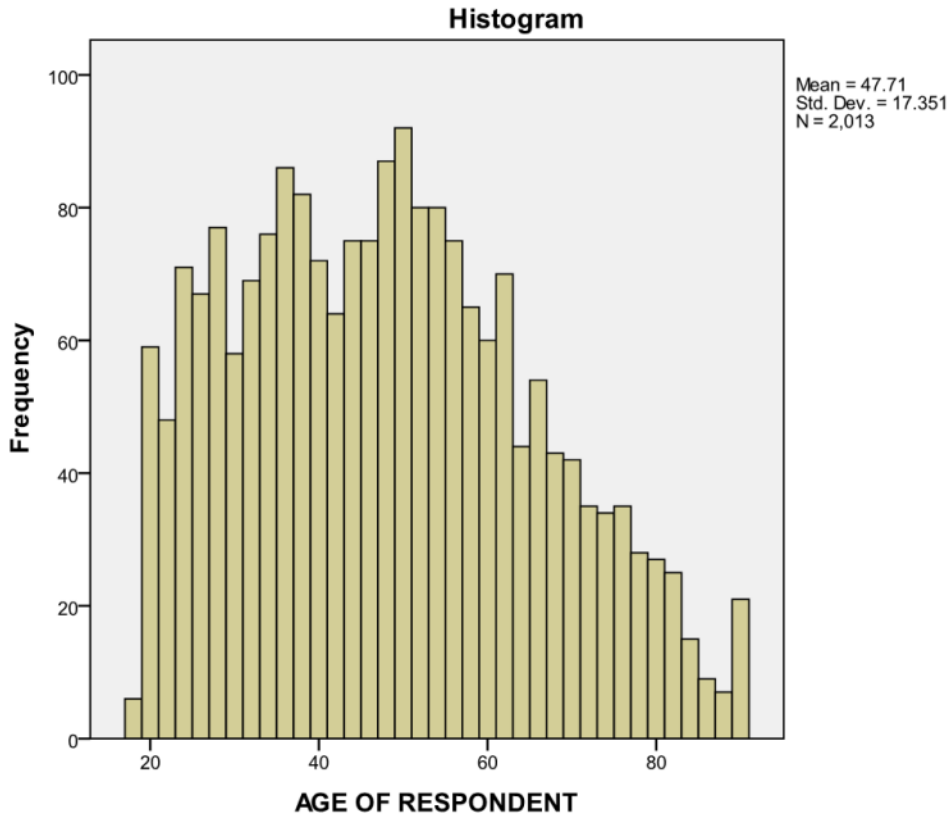
Skewness is a measure of the symmetry of a distribution. It measures the degree to which cases are clustered towards one end of the distribution. It is normed so that a symmetric distribution has zero skewness. A positive skewness value indicates bunching on the left and a longer tail on the right (for example, income distribution); negative skewness follows the reverse pattern. The standard error of skewness also appears in the Descriptives table and we can use it to determine if the data are significantly skewed. One method is to use the standard errors to calculate the 95% confidence interval around the skewness. If zero is not in this range, we could conclude that the distribution was skewed. A second method is to compare the skewness value to 1.96*(Standard error of skewness) from zero.

Kurtosis also has to do with the shape of a distribution and is a measure of how much of the data is concentrated near the center, as opposed to the tails, of the distribution. It is normed to the normal curve (for which kurtosis is zero). As an example, a distribution with longer tails and more peaked in the middle than a normal is referred to as a *leptokurtic* distribution and would have a positive kurtosis

measure. On the other hand, a *platykurtic* distribution is a flattened distribution and has negative kurtosis values. A standard error for kurtosis also appears. The same methods used for evaluating skewness can be used to evaluate the kurtosis values.

Since most analysts are content to view histograms in order to make judgments regarding the distribution of a variable, skewness and kurtosis are infrequently used.

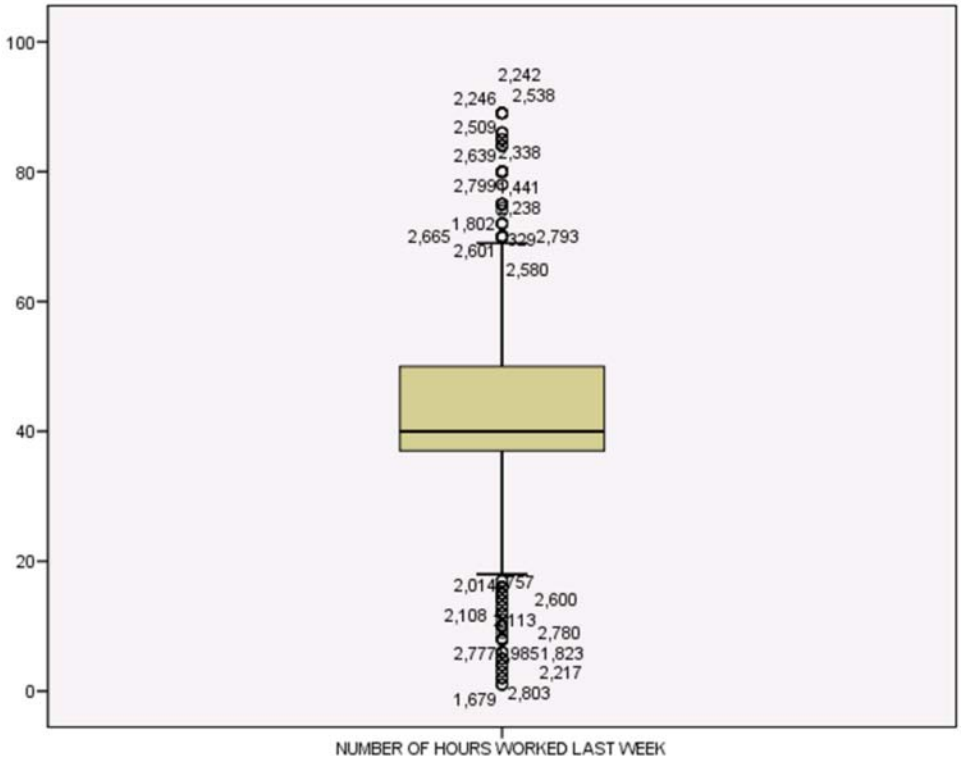
Figure 4.13 Histogram of Age of Respondent



The histogram shows the shape of the distribution. For this sample of adults, *age* is roughly normally distributed, except that the distribution is truncated below 18 years.

Boxplots, also referred to as box & whisker plots, are a more easily interpreted plot to convey the same information about the distribution of a variable. In addition, the boxplot graphically identifies outliers. Below we see the boxplot for hours worked.

Figure 4.14 Boxplot of Hours Worked



The vertical axis represents the scale for the number of hours worked. The solid line inside the box represents the median or 50th percentile. The top and bottom borders (referred to as “hinges”) of the box correspond to the 75th and 25th percentile values of hours worked and thus define the interquartile range (IQR). In other words, the middle 50% of data values fall within the box. The “whiskers” (vertical lines extending from the top and bottom of the box) are the last data values that lie within 1.5 box lengths (or IQRs) of the respective hinges (borders of box). Tukey considers data points more than 1.5 box lengths from a hinge to be "outliers." These points are marked with a circle. Points more than 3 box lengths (IQR) from a hinge are considered by Tukey to be “far out” points and are marked with an asterisk symbol (there are none here). This plot has many outliers. If a single outlier exists at a data value, the case sequence number appears beside it (an ID variable can be substituted), which aids data checking.

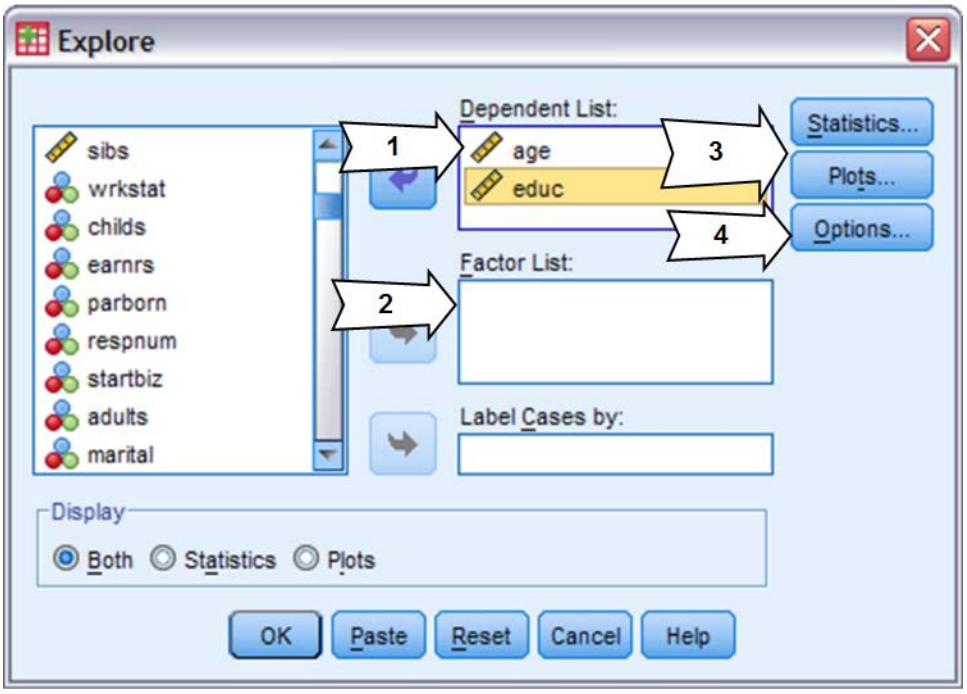
If the distribution were symmetric, the median would be centered within the box. In the plot above, the median is toward the bottom of the box, indicating a positively skewed distribution.

4.15 *Procedure: Explore*

The **Explore** procedure is accessed from the **Analyze…Descriptive Statistics…Explore** menu choice. With the **Explore** dialog box open:

1) Place one or more scale variables to be summarized in the Dependent list box.
2) The Factor list box can contain one or more categorical variables, and if used would cause the procedure to present summaries for each category of the factor variable(s).
3) We can request specific statistical summaries and plots using the Statistics and Plots buttons.
4) The Options button specifies how missing data will be handled.

Figure 4.15 Explore Dialog Box

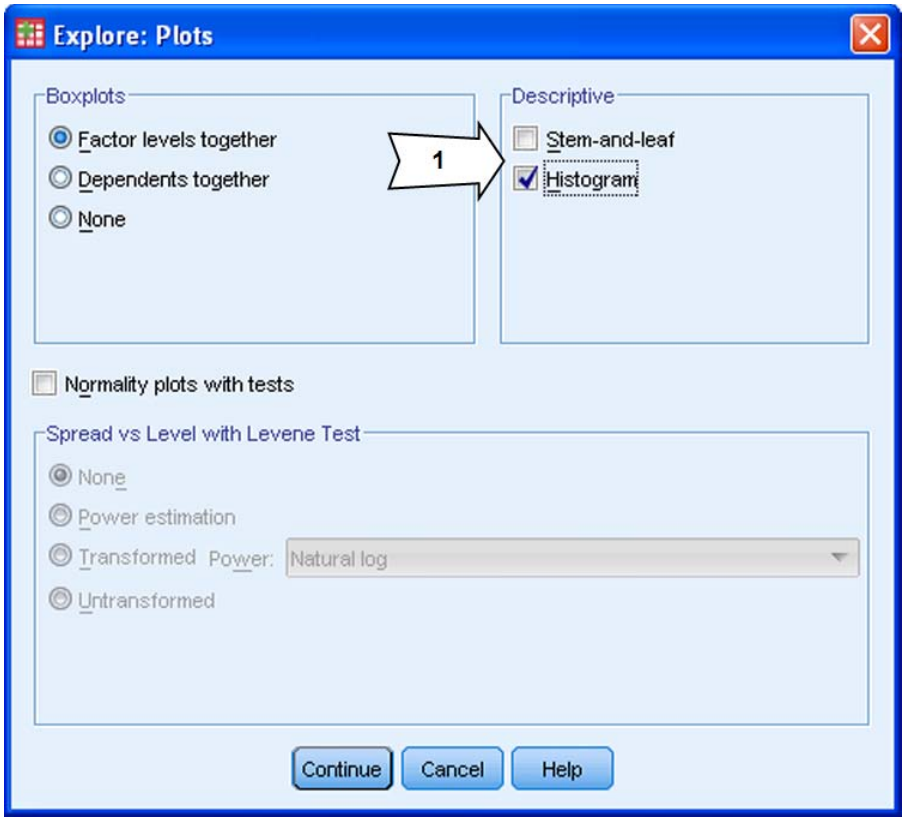


In the Plots dialog:

1) Request a histogram rather than a stem and leaf plot.

The stem & leaf plot (devised by Tukey) is modeled after the histogram, but contains more information. For most purposes, the histogram is easier to interpret and more useful. By default, a boxplot will be displayed for each scale variable.

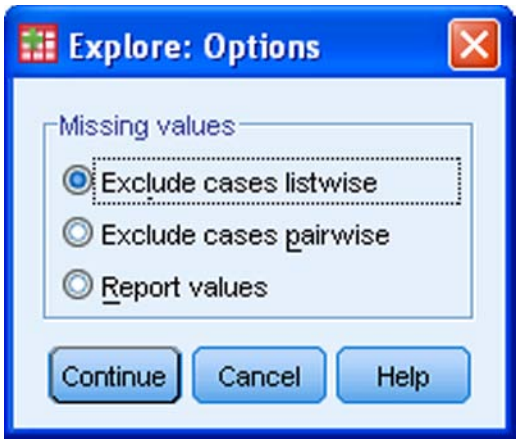
Figure 4.16 Explore: Plots Dialog Box



In the Options dialog the user specifies how to deal with missing values:

1) Request pairwise rather than listwise deletion.

Figure 4.17 Explore: Options Dialog Box



When several variables are used you have a choice as to whether the analysis should be based on only those observations with valid values for all variables in the analysis (called *listwise* deletion), or whether missing values should be excluded separately for each variable (called *pairwise* deletion). When only a single variable is considered both methods yield the same result, but they will not give identical answers when multiple variables are analyzed in the presence of missing values.

Rarely used, the *Report values* option includes cases with user-defined missing values in frequency analyses, but excludes them from summary statistics and charts.

4.16 *Demonstration: Explore*

We will work with the *Census.sav* data file in this lesson.

In this example we examine the summary statistics of *age*, and *educ*. We would like to see summary statistics for the mentioned variables, as well as how much missing data we have, and if there are unusual cases.

Detailed Steps for Explore

1) Place the variables ***age,*** and ***educ*** in the Dependent list box
2) Deselect the **Stem and leaf** and select **Histogram** in the Plots dialog
3) Request **Exclude cases pairwise** method in the Options dialog

Results from Explore

The **Explore** procedure produces two tables followed by the requested charts for each variable. The first table, Case Processing Summary, displays the number of valid and missing cases for each variable. Each variable has little missing data. For example, 2013 cases (respondents) had valid values for *age*, while 0.5% were missing.

Figure 4.18 Explore Case Processing Table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
AGE OF RESPONDENT	2013	99.5%	10	.5%	2023	100.0%
HIGHEST YEAR OF SCHOOL COMPLETED	2018	99.8%	5	.2%	2023	100.0%

The Descriptives table displays a series of descriptive statistics for *age* and *educ*. From the previous table, we know that these statistics are based on 2013 and 2018 respondents, respectively.

Comparing measures of central tendency can provide useful information about the distribution. Here the mean, median and 5% trimmed mean are very close within each variable and this suggests either that there are not many extreme scores, or that the number of high and low scores is balanced. If the mean were considerably above or below the median and trimmed mean, it would suggest a skewed or asymmetric distribution. A perfectly symmetric distribution, the normal distribution, would produce identical means, medians and trimmed means.

Figure 4.19 Summaries for Age of Respondent and Highest Year of School Completed

Descriptives

			Statistic	Std. Error
AGE OF RESPONDENT	Mean		47.71	.387
	95% Confidence Interval for Mean	Lower Bound	46.95	
		Upper Bound	48.47	
	5% Trimmed Mean		47.25	
	Median		47.00	
	Variance		301.052	
	Std. Deviation		17.351	
	Minimum		18	
	Maximum		89	
	Range		71	
	Interquartile Range		26	
	Skewness		.308	.055
	Kurtosis		-.713	.109
HIGHEST YEAR OF SCHOOL COMPLETED	Mean		13.43	.069
	95% Confidence Interval for Mean	Lower Bound	13.30	
		Upper Bound	13.57	
	5% Trimmed Mean		13.52	
	Median		13.00	
	Variance		9.480	
	Std. Deviation		3.079	
	Minimum		0	
	Maximum		20	
	Range		20	
	Interquartile Range		4	
	Skewness		-.425	.054
	Kurtosis		1.320	.109

The standard deviation of age, 17.351, indicates a variation around the mean of plus or minus 17 years. The standard error is used in calculating the 95% confidence interval for the sample mean. The interquartile range of 26 indicates that the middle 50% of the sample lies within a range of 26 years.

The shape of the distribution can be of interest in its own right. Also, assumptions are made about the shape of the data distribution within each group when performing significance tests on mean differences between groups. The shape of the distributions of both these variables does not appear problematic.

Figure 4.20 Histogram of Age of Respondent

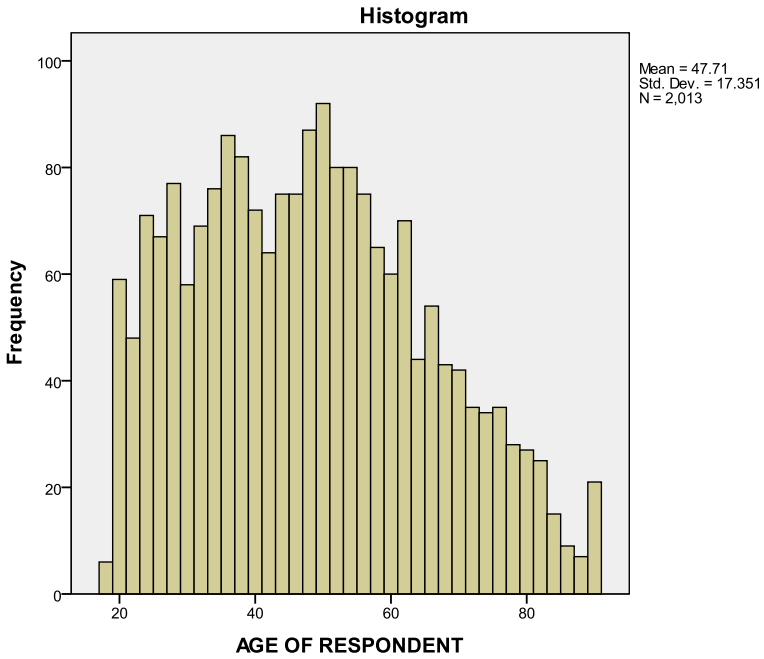
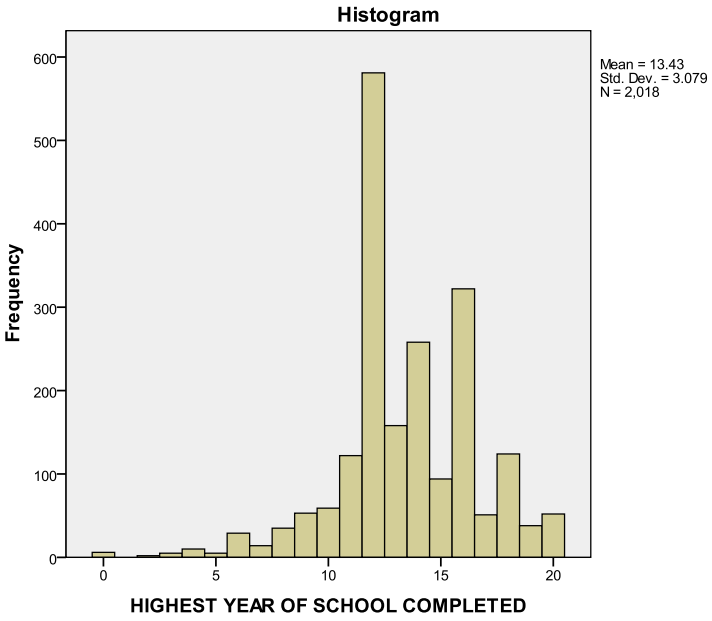
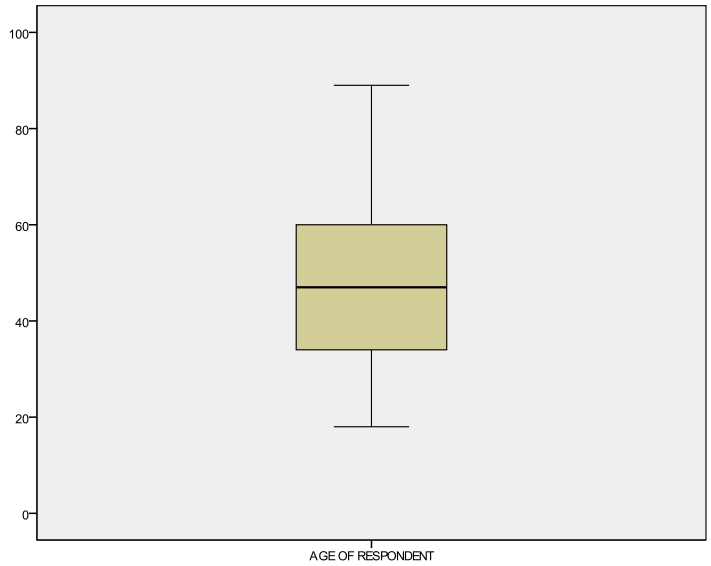


Figure 4.21 Histogram of Highest Year of School Completed



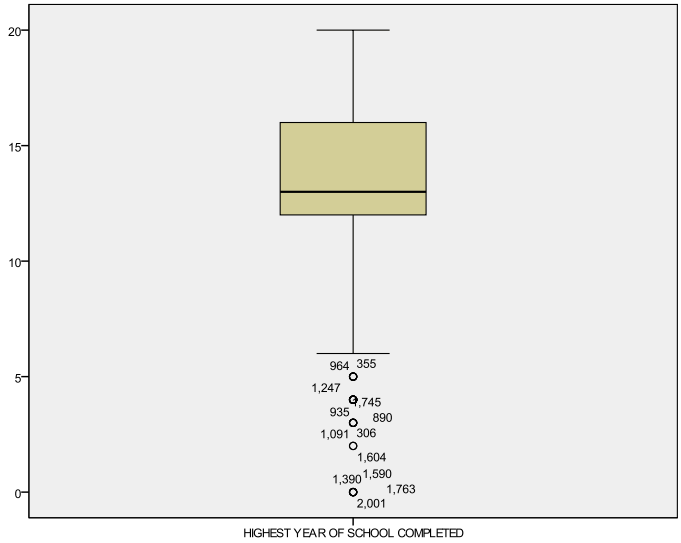
Below we see the boxplots for *age* and *educ*. Boxplots are particularly useful for obtaining an overall "feel" for a distribution. The median tells us the location or central tendency of the data. The length of the box indicates the amount of spread within the data, and the position of the median in relation to the box tells us something of the nature of the distribution. Boxplots are also useful when comparing several groups.

Figure 4.22 Boxplot of Age of Respondent



The plot for *age* has no outliers while the plot for *educ* has a few outliers on the lower end of the distribution.

Figure 4.23 Boxplot of Highest Year of School Completed



If suspicious outliers appear in your data you should check whether they are data errors. If not, you need to consider whether you wish them included in your analysis. This is especially problematic when dealing with a small sample (not the case here), since an outlier can substantially influence the analysis.

We would not argue that something of interest always appears through use of the methods of exploratory data analysis. However, you can quickly glance over these results, and if anything strikes

your attention, pursue it in more detail. The possibility of detecting something unusual encourages the use of these techniques.

Apply Your Knowledge

1. See the output below, with statistics for *Current Salary* (data collected on employees). Which statements are correct?
    a. If repeated samples of the same sample size (here 467) were taken, and we record the sample mean for each sample, then we expect a standard deviation of 795.399 for these sample means.
    b. The median is lower than the mean, indicating that the distribution is skewed to the left.
    c. 50% of the salaries are in a range of salary of 13,500.

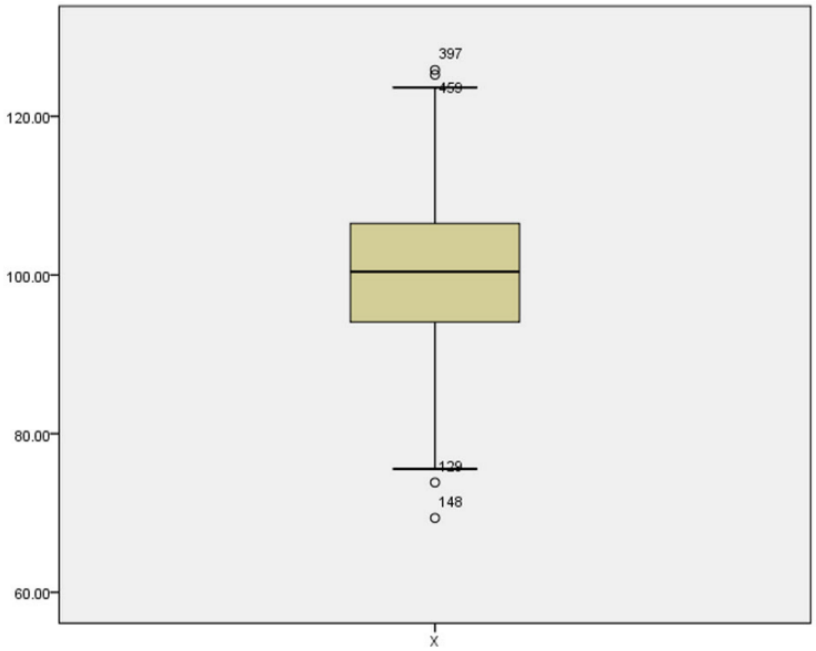
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Current Salary	467	98.5%	7	1.5%	474	100.0%

Descriptives

			Statistic	Std. Error
Current Salary	Mean		34450.27	795.399
	95% Confidence Interval for Mean	Lower Bound	32887.26	
		Upper Bound	36013.28	
	5% Trimmed Mean		32477.41	
	Median		28650.00	
	Variance		295451752.986	
	Std. Deviation		17188.710	
	Minimum		15750	
	Maximum		135000	
	Range		119250	
	Interquartile Range		13500	
	Skewness		2.109	.113
	Kurtosis		5.268	.225

2. True or false? See the boxplot for variable *X*. The boxplot of this variable indicates a normal distribution?



Additional Resources

Further Info

For additional information on Exploratory data analysis, see:

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

4.17 Lesson Summary

In this lesson we explored how to obtain summary statistics for scale variables. We reviewed three procedures—**Frequencies**, **Descriptives** and **Explore**—that can help us obtain that information.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Request and interpret appropriate summary statistics for scale variables

To support the achievement of the primary objective, students should now also be able to:

- Use the options in the **Frequencies**, **Descriptives**, and **Explore** procedures
- Interpret the results of the **Frequencies**, **Descriptives**, and **Explore** procedures

4.18 Learning Activity

In this set of learning activities you will use the *Drinks.sav* data file. The overall goal of this learning activity is to obtain summary statistics for scale variables.

Supporting Materials

The file *Drinks.sav*, a PASW Statistics data file that contains data on 35 beverages. Included is information on their characteristics (e.g., % alcohol), price, origin, and a rating of quality.

1. Run **Frequencies** on the variable *alcohol*, requesting the summary statistics median and mean, plus a histogram with a superimposed normal curve. Suppress the display of the frequency table.

2. What is the value of value of *alcohol* that splits the distribution in half? Is the median the same as the mean? Which value is lower? What does that tell you about the shape of the distribution of *alcohol*?

3. Does the histogram verify your description of the distribution of *alcohol*? How does it differ from a normal distribution?

4. Run **Descriptives** to obtain default statistics for *price* and *calories*. On which variable is there more dispersion? Is it even realistic to compare these two variables since they are on different scales?

5. Continuing your analysis of *price* and *calories*, run the **Explore** procedure for these two variables. Request a histogram in addition to the defaults.

6. Does the standard error of each variable help you better determine which variable has more dispersion?

7. Review the boxplots and histogram for each variable. Which one has more outliers? What are the outliers on each? Which variable now appears to have more dispersion, based on these graphs? Does that match what you expected based on the statistics?

Lesson 5: Making Inferences about Populations from Samples

5.1 *Objectives*

After completing this lesson students will be able to:

- Explain how to make inferences about populations from samples

To support the achievement of the primary objective, students will also be able to:

- Explain the influence of sample size
- Explain the nature of probability
- Explain hypothesis testing
- Explain different types of statistical errors and power
- Explain differences between statistical and practical importance

5.2 *Introduction*

Ideally, for any analysis we would have data about everyone we wished to study (i.e., the whole population). In practice, we rarely have information about all members of our population and instead collect information from a representative sample of the population. However, our goal is to make generalizations about various characteristics of the population based on the known facts about the sample. In this lesson we discuss the requirements of making inferences to populations from samples.

Business Context

Understanding how to make inferences from a sample to a population is the basis of inferential statistics. This allows us to reach conclusions about the population without the need to study every single individual. Hypothesis testing allows researchers to develop hypotheses which are then assessed to determine the probability or likelihood of the findings.

Supporting Materials
None

5.3 *Basics of Making Inferences about Populations from Samples*

We choose a sample with the intention of using the data from that sample to make inferences about the “true” values in the population. These population measures are referred to as **parameters** while the equivalent measures from samples are known as **statistics**. It is unlikely that we will know the population parameters; therefore we use the sample statistics to **infer** what these population values will be.

An important distinction between parameters and statistics is that parameters are *fixed* (although often not known) while statistics *vary* from one sample to another. Due to the effects of random variability, it is unlikely that any two samples drawn from the same population will produce the same statistics. By plotting the values of a particular statistic (e.g., the mean) from a large number of samples, it is possible to obtain a **sampling distribution** of the statistic. For small numbers of samples, the mean of the sampling distribution may not closely resemble that of the population. However, as the number of samples taken increases, the closer the mean of the sampling distribution

(the mean of all the means, if you like) gets to the population mean. For an infinitely large number of samples, the mean will be exactly the same as the population mean. Additionally, as sample size increases, the amount of variability in the distribution of sample means decreases. If you think of variability in terms of the error made in estimating the mean, it should be clear that the more evidence you have (i.e., the more cases in your sample), the smaller will be the error in estimating the mean.

If repeated random samples of size N are drawn from any population, then as N becomes large, the sampling distribution of sample means approaches normality—a phenomenon known as the **Central Limit Theorem**. This is an extremely useful statistical concept as it does not require that the original population distribution is normal. In the next section, we'll take a closer look at the influence of sample size on the precision of the statistics.

5.4 *Influence of Sample Size*

In statistical analysis sample size plays an important role, but one that can easily be overlooked since a minimum sample size is not required for the most commonly used statistical tests. Here we will demonstrate the effect of sample size in two common data analysis situations: crosstabulation tables and mean summaries.

Precision of Percentages

Precision is strongly influenced by the sample size. In the figures below we present a series of crosstabulation tables containing identical percentages, but with varying sample sizes. We will observe how the test statistics change with sample size and relate this result to the precision of the measurement.

The Chi-square test of independence will be presented for each table as part of the presentation of the effect of changing sample size. (See the lesson “Combining Categorical Variables, Crosstabulations and the Chi-Square Statistic for a detailed discussion of the chi-square test.)

Sample Size of 100

The table below displays responses of men and women to a question asking for which candidate they would vote.

Figure 5.1 Crosstab Table with Sample of 100

Which Candidate ? by Gender N=100

			Gender		
			Male	Female	Total
Vote For?	Candidate A	Count	23	27	50
		Column %	46.0%	54.0%	50.0%
	Candidate B	Count	27	23	50
		Column %	54.0%	46.0%	50.0%
Total		Count	50	50	100
		Column %	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	.640[b]	1	.424
N of Valid Cases	100		

b.

- 46 % of the men and 54 % of the women choose candidate A—an 8% difference
- The Chi-square test assesses whether men differ from women in the population
- Significance value of .424 indicates probability that men and women share the same view (do not differ significantly) concerning candidate choice
- Conclude there is no gender difference in a sample of 100 people

Sample Size of 400

Now we view a table with percentages identical to the previous one, but based on a sample of 400 people, four times as large as before.

Figure 5.2 Crosstabulation Table with Sample of 400

Which Candidate? by Gender N=400

			Gender		
			Male	Female	Total
Vote For?	Candidate A	Count	92	108	200
		Column %	46.0%	54.0%	50.0%
	Candidate B	Count	108	92	200
		Column %	54.0%	46.0%	50.0%
Total		Count	200	200	400
		Column %	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	2.560[b]	1	.110
N of Valid Cases	400		

b.

- 46 % of the men and 54 % of the women choose candidate A—an 8% difference
- Significance value of .110 for Chi-Square test indicates probability that men and women share the same view (do not differ significantly) concerning candidate choice
- Conclude there is no gender difference in a sample of 400 people

Sample Size of 1,600

Finally we present the same table of percentages, but increase the sample size to 1,600; the increase is once again by a factor of four.

Figure 5.3 Crosstabulation Table with Sample of 1,600

Which Candidate ? by Gender N=1600

			Gender		
			Male	Female	Total
Vote For?	Candidate A	Count	368	432	800
		Column %	46.0%	54.0%	50.0%
	Candidate B	Count	432	368	800
		Column %	54.0%	46.0%	50.0%
Total		Count	800	800	1600
		Column %	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	10.240[b]	1	.001
N of Valid Cases	1600		

b.

- 46 % of the men and 54 % of the women choose candidate A-- an 8% difference
- Significance value of .001 for Chi-Square test indicates probability that men and women share the same view (do differ significantly) concerning candidate choice
- Conclude there is a gender difference in a sample of 1600 people
- We have more precise estimates as our sample size increases and the 8% sample difference can more accurately be estimated

Sample Size and Precision

In the series of crosstabulation tables above we saw that as the sample size increased we were more likely to conclude there was a statistically significant difference between two groups when the magnitude of the sample difference was constant (8%). This is because the precision with which we estimate the population percentage increases with increasing sample size. This relation can be approximated (see note for the exact relationship) by a simple equation: the precision of a sample proportion is approximately equal to one divided by the square root of the sample size. The table below displays the precision for the sample sizes used in our examples.

Table 5.1 Sample Size and Precision for Different Sample Sizes

Sample Size	Precision	
100	1/sqrt(100) = 1/10	.10 or 10%
400	1/sqrt (400) = 1/20	.05 or 5%
1600	1/sqrt(1600) = 1/40	.025 or 2.5%

And to obtain a precision of 1%, we would need a sample of 10,000 (1/sqrt(10,000) = 1/100).

Since precision increases as the square root of the sample size, in order to double the precision we must increase the sample size by a factor of four. This is an unfortunate and expensive fact of research. In practice, samples between 500 and 1,500 are often selected for national studies.

Further Info

Formally, for a binomial or multinomial distribution (a variable measured on a nominal or ordinal scale), the standard error of the sample proportion (P) is equal to

$$StdErr(P) = \sqrt{P*(1-P)/N}$$

Thus the standard error is a maximum when P = .5 and reaches a minimum of 0 when P = 0 or 1. A 95% confidence band is usually determined by taking the sample estimate plus or minus twice the standard error. Precision (*pre*) here is simply two times the standard error. Thus precision (*pre*) is

$$pre(P) = 2*\sqrt{P*(1-P)/N}\ .$$

If we substitute for P the value .5 which maximizes the expression (and is therefore conservative) we have

$$\begin{aligned} pre(0.5) &= 2*\sqrt{0.5*(1-0.5)/N} \\ &= 2*\sqrt{(0.5)*(0.5)}/\sqrt{N} \\ &= 2*(0.5)/\sqrt{N} \\ &= 1/\sqrt{N} \end{aligned}$$

This validates the rule of thumb used in the chapter. Since the rule of thumb employs the value of P=.5, which maximizes the standard deviation and thus the standard error, in practice, greater precision would be obtained when P departs from .5.

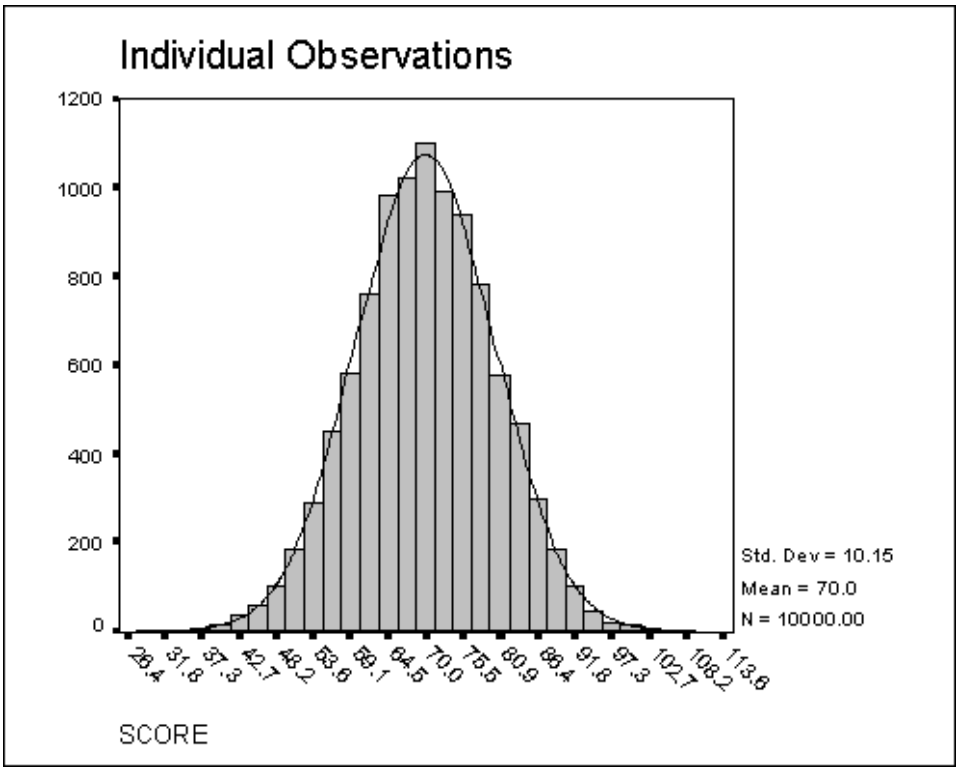
Precision of Means

The same basic relation—that precision increases with the square root of the sample size—applies to sample means as well. To illustrate this we display histograms based on different samples from a normally distributed population with mean 70 and standard deviation 10.

A Large Sample of Individuals

Below is a histogram of 10,000 observations drawn from a normal distribution of mean 70 and standard deviation 10. This is for one sample.

Figure 5.4 Histogram of 10,000 Observations

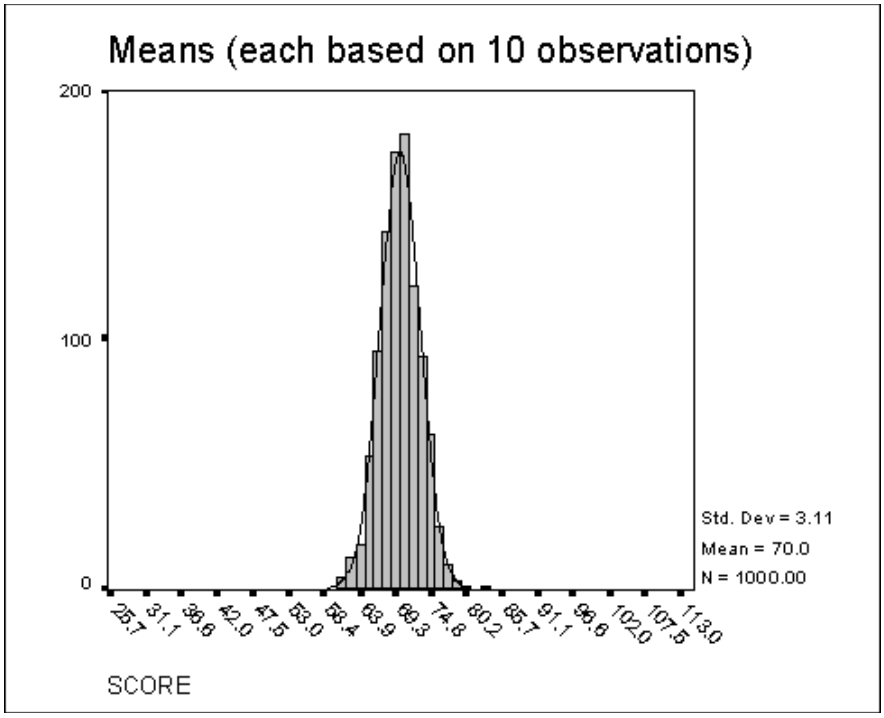


- Sample of this size closely matches its population
- Sample mean is very close to 70
- Sample standard deviation is near 10
- Shape of the distribution is normal

Means Based on Samples of 10

The second histogram displays 1,000 sample means drawn from the same population (mean 70, standard deviation 10). Here each observation is a mean based on only 10 data points. In other words we select 1,000 samples of ten observations each and plot their means in the histogram.

Figure 5.5 Histogram of Means Based on Samples of 10

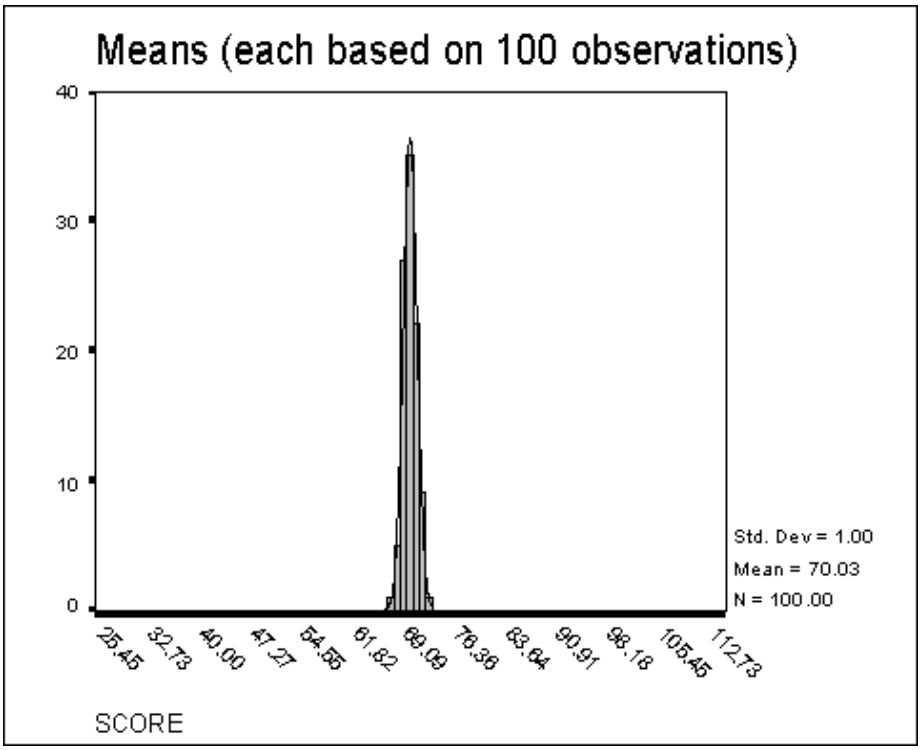


- Sample mean is very close to 70
- Sample standard deviation is reduced to 3.11
- Less variation among means based on groups of observations then among the observations themselves
- Shape of the distribution is normal

Means Based on Samples of 100

The next histogram is based on a sample of 100 means where each mean represents 100 observations. Here each observation is a mean based on 100 data points. In other words we select 100 samples of 100 observations each and plot their means in the histogram.

Figure 5-6 Histogram of Means Based on Samples of 100

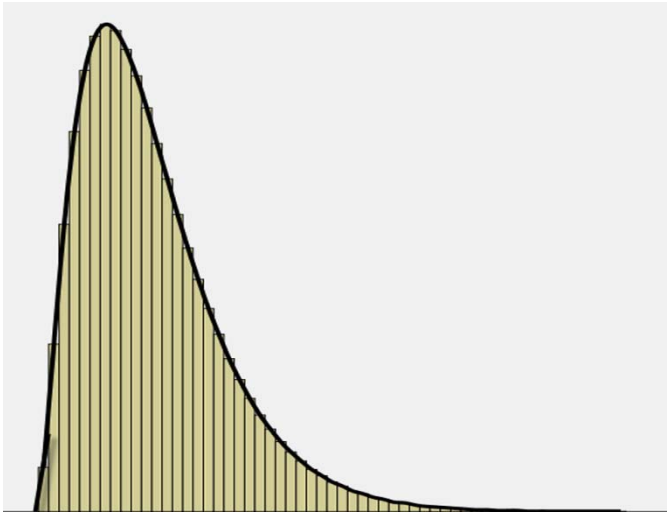


- Sample mean is very close to 70
- Sample standard deviation is reduced to 1.00
- Less variation among means based on groups of observations then among the observations themselves
- Shape of the distribution is normal

Thus with means as well as percents, precision increases with the square root of the sample size.

Apply Your Knowledge

1. Consider the population distribution of the variable *income*, depicted below. Note, this distribution is highly skewed. If we are sampling repeatedly from this distribution, each time with sample size = 1000, then what is the distribution of the sample means?
    a. The sample means show the same distribution as *income*
    b. The sample means show a normal distribution
    c. The distribution of the sample means is unknown



2. Which of these statements is correct?
    a. The bigger the sample, the higher the precision with which we estimate a population percentage
    b. The bigger the sample, the lower the precision with which we estimate a population mean
    c. A sample size of 2000 will give a precision for a population percentage that is twice as precise as the precision for this percentage with sample size 1000
    d. If a population percentage is 0 or if it is 100, the standard error for the sample percentage will be 0

3. True or false? A population parameter varies from sample to sample?

5.5 *Hypothesis Testing*

Whenever we wish to make an inference about a population from our sample, we must specify a hypothesis to test. It is common practice to state two hypotheses: the **null hypothesis** (also known as H0) and the **alternative hypothesis** (H1). The null hypothesis being tested is conventionally the one in which no effect is present. For example, we might be looking for differences in mean income between males and females, but the (null) hypothesis we are testing is that there is no difference between the groups. If the evidence is such that this null hypothesis is ***unlikely*** to be true, the alternative hypothesis should be accepted. Another way of thinking about the problem is to make a comparison with the criminal justice system. Here, a defendant is treated as innocent (i.e.. the null hypothesis is accepted) until there is enough evidence to suggest that they perpetrated the crime beyond any reasonable doubt (i.e., the null hypothesis is rejected).

The alternative hypothesis is generally (although not exclusively) the one we are really interested in and can take any form. In the above example, we might hypothesis that males will have a higher mean income than females. When the alternative hypothesis has a “direction” (we expect a specific result), the test is referred to as **one-tailed**. Often, you do not know in which direction to expect a difference and may simply wish to leave the alternative hypothesis open-ended. This is a **two-tailed** test and the alternative hypothesis would simply be that the mean incomes of males and females are different.

Whichever option you choose will have implications when interpreting the probability levels. In general, the probability of the occurrence of a particular statistic for a one-tailed test will be half that of a two-tailed test as only one extreme of the distribution is being considered in the former type of test.

5.6 *The Nature of Probability*

Descriptive statistics describe the data in our sample through the use of a number of summary procedures and statistics. Inferential statistics allow us ***to infer the results from the sample on which we have data to the population which the sample represents.*** To do this, we use procedures that involve the calculation of ***probabilities***. The fundamental issue with inferential statistical tests concerns whether any "effects" (relationships or differences between groups) we have found are genuine or are a result of sampling variability (in other words, mere chance). So we have two hypotheses and we want to know which hypothesis is true. The way hypotheses are assessed is by calculating the probability or the likelihood of finding our result. A probability value, which can range from 0 to 1 (corresponding to 0% to 100% in terms of percentages), can be defined as *"**the mathematical likelihood of a given event occurring***," and as such we can use such values to assess whether the likelihood that any differences we have found are the result of random chance.

Now in statistics, we want to be sure of our conclusions, so having formally stated your hypotheses, you must then select a criterion for acceptance or rejection of the null hypothesis. With probability tests such as the chi-square test or the t-test, you are testing the likelihood that a statistic of the magnitude obtained (or greater) would have occurred by chance assuming that the null hypothesis is true. You always assess the null hypothesis, which is the hypothesis that states there is no difference or relationship. In other words, we only wish to reject the null hypothesis when we can say that the result would have been extremely unlikely under the conditions set by the null hypothesis. In this case, the alternative hypothesis should be accepted. *It is worth noting that this does not "prove" the alternative hypothesis beyond doubt, it merely tells us that the null hypothesis is unlikely to be true.*

But what criterion (or **alpha level**, as it is often known) should we use? Unfortunately, there is no easy answer! Traditionally, a 5% level is chosen, indicating that a statistic of the size obtained would only be likely to occur on 5% of occasions (or once-in-twenty) should the null hypothesis be true. This also means that, by choosing a 5% criterion, you are accepting that you will make a mistake in rejecting the null hypothesis 5% of the time.

5.7 *Types of Statistical Errors*

Recall that when performing statistical tests we are generally attempting to draw conclusions about the larger population based on information collected in the sample. There are two major types of errors in this process. False positives, or **Type I** errors, occur when no difference (or relation) exists in the population, but the sample tests indicate there are significant differences (or relations). Thus the researcher falsely concludes a positive result. This type of error is explicitly taken into account when performing statistical tests. When testing for statistical significance using a .05 criterion (alpha level), we acknowledge that if there is no effect in the population then the sample statistic will exceed the criterion on average 5 times in 100 (.05).

Type II errors, or false negatives, are mistakes in which there is a true effect in the population (difference or relation) but the sample test statistic is not significant, leading to a false conclusion of no effect. To put it briefly, a true effect remains undiscovered. The probability of making this type of error is often referred to as the **beta level**. Whereas you can select your own alpha levels, beta levels are dependent upon things such as the alpha level and the size of the sample. It is helpful to note that statistical power, the probability of detecting a true effect, equals 1 minus the Type II error and the higher the power the better.

Table 5.2 Types of Statistical Errors in Hypothesis Testing

		Statistical Test Outcome	
		Not Significant	**Significant**
Population	**No Difference** (Ho is True)	Correct	*Type I error* (α) False positive
	True Difference (H1 is True)	*Type II error* (β) False negative	Correct (Power)

Statistical Power Analysis

With increasing precision we are better able to detect small differences that exist between groups and small relationships between variables. Power analysis was developed to aid researchers in determining the minimum sample size required in order to have a specified chance of detecting a true difference or relationship of a given size. To put it more simply, power is used to quantify your ability to reject the null hypothesis when it is false. For example, suppose a researcher hopes to find a mean difference of .8 standard deviation units between two populations. A power calculation can determine the sample size necessary to have a 90% chance that a significant difference will be found between the sample means when performing a statistical test at a specified significance level. Thus a researcher can evaluate whether the sample is large enough for the purpose of the study. Books by Cohen (1988) and Kraemer & Thiemann (1987) discuss power analysis and present tables used to perform the calculation for common statistical tests. In addition specialty software is available for such analyses, such as SamplePower®. Power analysis can be very useful when planning a study, but does require such information as the magnitude of the hypothesized effect and an estimate of the variance.

5.8 *Statistical Significance and Practical Importance*

A related issue involves drawing a distinction between statistical significance and practical importance. When an effect is found to be statistically significant we conclude that the population effect (difference or relation) is not zero. However, this allows for a statistically significant effect that is not quite zero, yet so small as to be insignificant from a practical or policy perspective. This notion of practical or real world importance is also called ecological significance. Recalling our discussion of precision and sample size, very large samples yield increased precision, and in such samples very small effects may be found to be statistically significant. In such situations, the question arises as to whether the effects make any practical difference. For example, suppose a company is interested in customer ratings of one of its products and obtains rating scores from several thousand customers. Furthermore, suppose mean ratings on a 1 to 5 satisfaction scale are 3.25 for male and 3.15 for female customers, and this difference is found to be significant. Would such a small difference be of any practical interest or use?

When sample sizes are small (say under 30), precision tends to be poor and so only relatively large (and ecologically significant) effects are found to be statistically significant. With moderate samples (say 50 to one or two hundred), small effects tend to show modest significance while large effects are highly significant. For very large samples, several hundreds or thousands, small effects can be highly significant; thus an important aspect of the analysis is to examine the effect size and determine if it is important from a practical, policy or ecological perspective.

Apply Your Knowledge

1. Which of these statements is correct?

   a. The nature of the null hypothesis is: *there is no effect/difference*
   b. The nature of the null hypothesis is: *there is an effect/difference*
   c. The nature of the alternative hypothesis is: *there is no effect/difference*

2. Which of these statements is correct?
   a. If we reject the null hypothesis while the null hypothesis is true in reality, then we make a Type II error.
   b. If we reject the null hypothesis while the null hypothesis is true in reality, then we make a Type I error.
   c. If we reject the null hypothesis, while the alternative hypothesis is true in reality, then we make a Type I error.
   d. If we reject the null hypothesis, while the alternative hypothesis is true in reality, then we make a Type II error.

3. Which of these statements is false?
   a. The probability of detecting a true effect is known as “power.”
   b. Probability of making a Type II error + the probability of detecting a true effect=1
   c. With an alpha level of .01, the probability of a Type II error will always be lower than that.
   d. With Alpha=.000001 the probability to commit a Type I error is lower than with Alpha=0.1.

5.9 *Lesson Summary*

In this lesson, we explored how to make inferences from a sample to a population. This allows us to reach conclusions about the population without the need to study every single individual. Hypothesis testing allows researchers to develop hypotheses which are then assessed to determine the probability or likelihood of the findings.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Explain how to make inferences about populations from samples

To support the achievement of the primary objective, students should now also be able to:

- Explain the influence of sample size
- Explain the nature of probability
- Explain hypothesis testing
- Explain different types of statistical errors and power
- Explain differences between statistical and practical importance

5.10 *Learning Activity*

In this set of learning activities you won’t need any supporting material.

State the null and alternative hypotheses when assessing each of the following scenarios. Would you do a one- or two-tailed test for each?

1. The relationship between gender and belief in the afterlife.
2. The relationship between number of years of education and income.
3. The difference in mean income between men and women.
4. The difference between happiness in marriage between married persons in America, Europe, Asia, Australia and Africa.

Lesson 6: Relationships Between Categorical Variables

6.1 *Objectives*

After completing this lesson students will be able to:

- Perform crosstab analysis on categorical variables
- Perform a statistical test to determine whether there is a statistically significant relationship between categorical variables

To support the achievement of this primary objective, students will also be able to:

- Use the options in the **Crosstabs** procedure
- Request appropriate statistics for a crosstabulation
- Interpret cell counts and percents in a crosstabulation
- Use the Chi-Square test, interpret its results, and check its assumptions
- Use the **Chart Builder** to visualize a crosstabulation

6.2 *Introduction*

Many data analysts consider crosstabulations the core of data analysis. Crosstabulations display the joint distribution of two or more categorical variables. In this lesson we will provide examples and advice on how best to construct and interpret crosstabulations. With PASW Statistics, statistical tests are used to determine whether a relationship between two, or more, variables is statistically significant in a crosstabulation. Another way to state this is that a statistical test is done to determine whether a relationship observed in the sample is caused by chance sampling variation or instead is likely to exist in the population of interest. To support the analysis, we also show examples of graphical displays of crosstabulations.

Business Context

When analyzing data, the focus is on both descriptive and causal relationships. When we examine a table with categorical variables, we would like to know whether a relationship we observe is likely to exist in our target population or instead is caused by random sampling variation. We might want to know whether:

- Satisfaction with the instructor in a training workshop was related to satisfaction with the course material.
- Eating more often at fast-food restaurants was related to more frequent shopping at convenience stores.
- Certain types of people are more likely to buy laptop versus desktop computers.

Statistical testing tells us whether two categorical variables are related. Without that, we might make decisions based on observed category percentage differences that are not likely to exist in a population of customers.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

6.3 *Crosstabs*

Crosstabulations are commonly used to explore how demographic characteristics are related to attitudes and behaviors, but they are also used to see how one attitude is related to another. The key point is that crosstabulations are used to study the relationships between two, or more, **categorical** variables.

Crosstabs Illustrated

To provide context, consider the table depicted below. This table shows counts and percentages in the cells. For example, there are 206 female clerical employees. The percentages are calculated within *Gender* (the column variable). Thus, 60.9% (157/258) of the men are clerical.

To assess whether the two variables are related, there is a standard procedure to follow, depending on which percentages we are using.

1) If using percentages based on the row variable, compare percentages within each column. Look to see if the percentages are the same or different across rows within each column. If percentages are the same, there is no relationship between the variables.
2) If using percentages based on the column variable, compare percentages within each row. Look to see if the percentages are the same or different across columns within each row. Again, identical percentages indicate no relationship between the variables.

Here, percentages are based on *Gender*, so to assess whether *Employment Category* and *Gender* are related, we should compare the percentages 95.4% with 60.9% (95.4% of women are clerical versus 60.9% of men), 0% with 10.5% and 4.6% with 28.7%. Here, percentages differ substantially.

Figure 6.1 Crosstabs Illustrated

Employment Category * Gender Crosstabulation

			Gender		
			Female	Male	Total
Employment Category	Clerical	Count	206	157	363
		% within Gender	95.4%	60.9%	76.6%
	Custodial	Count	0	27	27
		% within Gender	.0%	10.5%	5.7%
	Manager	Count	10	74	84
		% within Gender	4.6%	28.7%	17.7%
Total		Count	216	258	474
		% within Gender	100.0%	100.0%	100.0%

6.4 *Crosstabs Assumptions*

To use **Crosstabs**, one condition has to be met:

1) Variables used in **Crosstabs** must be categorical (nominal, ordinal).

6.5 *Requesting Crosstabs*

Requesting **Crosstabs** is accomplished by following these steps:

1) Select variables for the **Crosstabs** procedure, at least one for the row and one for the column dimension; more than one variable can be used in a dimension of the table
2) Select percentage options.
3) Review the procedure output to investigate the relationship between the variables including:
   a. Cell counts
   b. Cell percentages.

6.6 *Crosstabs Output*

The **Crosstabs** table has at least two rows and two columns. The information contained in the cells of this table will depend on what you requested:

- By default, the (cell) Count is the only statistic displayed
- Normally, you will ask for either a row or column percent, or both
- These percentages are based on the variable categories in the row or column, respectively.
- The last row and column of the table display marginal totals for each row and column
- The lower right hand cell contains table total statistics, including the number of valid cases for the table.
- The percentages are used to determine whether one variable is statistically associated with another

Figure 6.2 Example of Crosstabs Output

OPINION OF HOW PEOPLE GET AHEAD * RESPONDENTS SEX Crosstabulation

			RESPONDENTS SEX		
			MALE	FEMALE	Total
OPINION OF HOW PEOPLE GET AHEAD	HARD WORK	Count	375	511	886
		% within OPINION OF HOW PEOPLE GET AHEAD	42.3%	57.7%	100.0%
		% within RESPONDENTS SEX	61.9%	69.4%	66.0%
	BOTH EQUALLY	Count	142	152	294
		% within OPINION OF HOW PEOPLE GET AHEAD	48.3%	51.7%	100.0%
		% within RESPONDENTS SEX	23.4%	20.7%	21.9%
	LUCK OR HELP	Count	89	73	162
		% within OPINION OF HOW PEOPLE GET AHEAD	54.9%	45.1%	100.0%
		% within RESPONDENTS SEX	14.7%	9.9%	12.1%
Total		Count	606	736	1342
		% within OPINION OF HOW PEOPLE GET AHEAD	45.2%	54.8%	100.0%
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

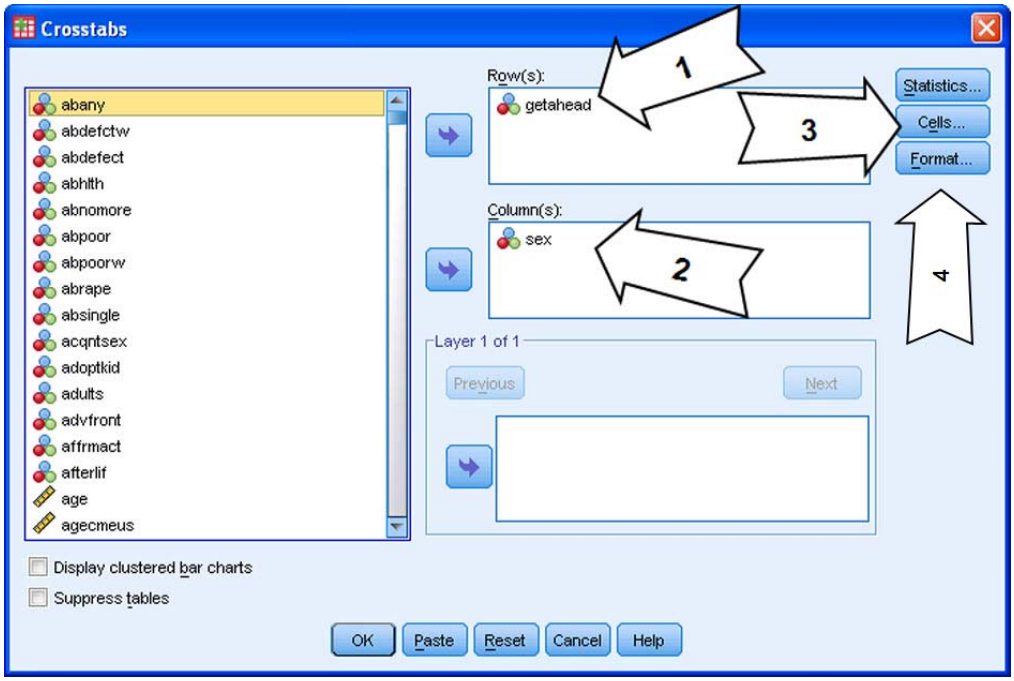
6.7 *Procedure: Crosstabs*

The **Crosstabs** procedure is accessed from the **Analyze…Descriptives…Crosstabs** menu choice. With the **Crosstabs** dialog box open:

1) Place one or more variables in the Row(s) box.
2) Place one or more variables in the Column(s) box.
3) Open the Cells dialog to specify percents and other cell statistics .
4) The order of categories can be changed in the Format dialog.

A separate table will be created for all combinations of variables in the Rows and Columns boxes.

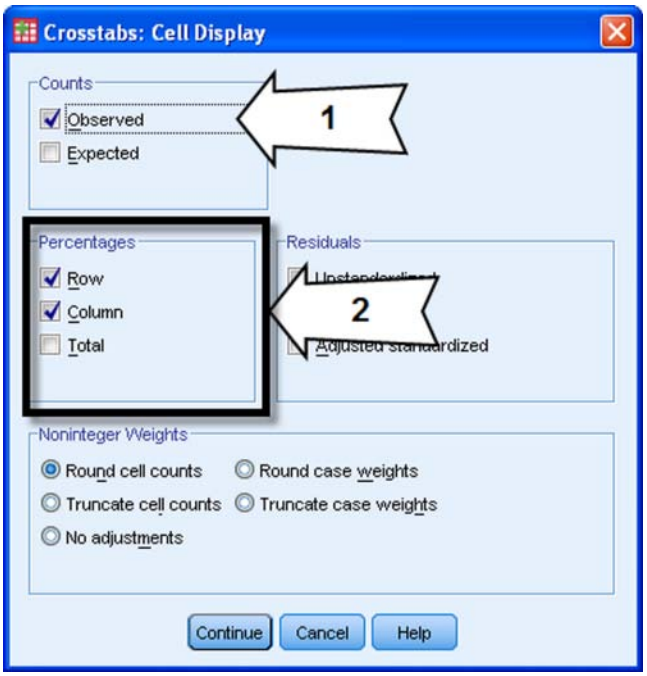
Figure 6.3 Crosstabs Dialog



In the Cells Display dialog:

1) By default the cell count is displayed (the *Observed* check box in the Counts area).
2) Typically one or both of row and column percents are selected with the appropriate check boxes in the Percentages area.
3) Other more non-standard statistics available include residuals which help to understand where there are deviations from the expected counts if there is no relationship.

Figure 6.4 Crosstabs Cell Display Dialog



6.8 *Example: Crosstabs*

We will work with the *Census.sav* data file in this lesson.

In this example we want to study how overall happiness (*happy*) is related to marital status (*marital*). We want to know what the percentage of overall happiness for each of the marital status groups, so we will percentage based on *marital*. The variable *marital* is nominal while *happy* is ordinal, so both variables are categorical and appropriate for Crosstabs.

Detailed Steps for Crosstabs

1) Place ***happy*** in the Row(s) box.
2) Place the variable ***marital*** in the Column(s) box.
3) Select **Column** in the Cells Display dialog.

The percentages are based on the categories of *marital*. This indicates that we expect marital status to affect general happiness.

Results from Crosstabs

As with all analyses, you should first look to see how many cases are missing from an analysis. The first table in the Viewer window provides this information (not shown). There is actually very little missing data here.

To examine a crosstabulation table:

- First focus on the marginal totals, which are equivalent to frequency distributions for each variable. We see that most people are either very happy (597) or pretty happy (1099) with their life overall, and most respondents are either married (969) or never married (528).

- Next turn to the counts in the cells of the table. Focus on the upper left-hand cell where 398 respondents indicated they were very happy with their life and married. Conversely, from the bottom right-hand cell, we see that 123 people are not too happy with their life and single.
- In the upper left-hand cell the first percentage is 41.1% (calculated within marital status). This tells us that, of the people who are married (398), 41.1% (398/969) are very happy in their life. The other percentages are calculated in a similar manner.

We observe that:

- 41.1% of those who are married say they are very happy. This is much larger than any other category. The next largest is divorced, at 20.0%.
- Married people also have the lowest percentage of people who say they are not too happy at 8.4%. Those who are separated have the highest percentage (25.7%).

These differences would certainly lead us to the conclusion that marital status is related to general happiness, and that married people are happier than others.

Figure 6.5 Crosstabulation of *marital* and *happy*

GENERAL HAPPINESS * MARITAL STATUS Crosstabulation

			MARITAL STATUS					
			MARRIED	WIDOWED	DIVORCED	SEPARATED	NEVER MARRIED	Total
GENERAL HAPPINESS	VERY HAPPY	Count	398	31	56	11	101	597
		% within MARITAL STATUS	41.1%	19.0%	20.0%	15.7%	19.1%	29.7%
	PRETTY HAPPY	Count	490	95	169	41	304	1099
		% within MARITAL STATUS	50.6%	58.3%	60.4%	58.6%	57.6%	54.7%
	NOT TOO HAPPY	Count	81	37	55	18	123	314
		% within MARITAL STATUS	8.4%	22.7%	19.6%	25.7%	23.3%	15.6%
Total		Count	969	163	280	70	528	2010
		% within MARITAL STATUS	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Apply Your Knowledge

1. What level of measurement should variables have to be used in **Crosstabs**?
    a. Any level of measurement
    b. Ordinal and Scale
    c. Nominal and Ordinal

2. Consider a **Crosstabs** table with the variable *sex* on the rows and variable *happiness* on the columns. If we wanted to see how *sex* might affect *happiness*, how should the table be percentaged?
    a. By *happiness*
    b. By *sex*
    c. By both

3. Consider the output depicted below. Which statements are correct?
    a. 11.9% of all women are managers
    b. 4.6% of all women are managers
    c. 60.9% of all men are clericals
    d. 2.1% of all cases are female managers

Employment Category * Gender Crosstabulation

			Gender		
			Female	Male	Total
Employment Category	Clerical	Count	206	157	363
		% within Employment Category	56.7%	43.3%	100.0%
		% within Gender	95.4%	60.9%	76.6%
		% of Total	43.5%	33.1%	76.6%
	Custodial	Count	0	27	27
		% within Employment Category	.0%	100.0%	100.0%
		% within Gender	.0%	10.5%	5.7%
		% of Total	.0%	5.7%	5.7%
	Manager	Count	10	74	84
		% within Employment Category	11.9%	88.1%	100.0%
		% within Gender	4.6%	28.7%	17.7%
		% of Total	2.1%	15.6%	17.7%
Total		Count	216	258	474
		% within Employment Category	45.6%	54.4%	100.0%
		% within Gender	100.0%	100.0%	100.0%
		% of Total	45.6%	54.4%	100.0%

6.9 *Chi-Square Test*

Comparing percentages is only part one of the story. What we don't know is whether differences in percentages are due to sampling variation, or instead are likely to be real and exist in the population. For this we turn to the Chi-Square test.

Every statistical test has a *null hypothesis*. In most cases, the null hypothesis is that there is no relationship between two variables. This is also true for the null hypothesis for a crosstabulation, so we use the Chi-Square test to determine whether there is a relationship between marital status and general happiness. If the significance is small enough, the null hypothesis is rejected. In the context of a crosstabulation, the null hypothesis would be that percentages across categories of an independent variable are statistically equivalent.

As with other statistical tests, sample size has an effect on the calculated significance. Larger sample sizes, everything else being equal, will reduce the significance value of a test statistic and thus make it easier to reject the null hypothesis. Chi-Square is particularly sensitive to the effect of increased sample size. In large samples, say 1000 to 1500 cases and above, it is fairly easy to show that two variables are related using the .05 significance level. So in large samples it is probably best to require a lower significance level before rejecting the null hypothesis (this is appropriate for the data file *Census.sav*, which contains over 2000 cases).

Note

Substantive Versus Statistical Significance
A critical distinction to make is whether a relationship is statistically significant versus being substantively, or practically, significant. In very large samples, small differences will be statistically significant. You should not let statistical significance be the overriding determinant in deciding whether a relationship or pattern you have discovered is interesting or important.

Chi-Square Test Assumptions

To correctly use a Chi-Square statistical test in **Crosstabs**, one condition has to be met:

1) At most, 20% of the expected values in the cells of the table should be below 5.

Chi-Square is calculated based on the expected values in each cell. If too many expected values are below 5, the reported significance can be incorrect. Because of this, PASW Statistics adds a footnote to the Chi-Square table noting the number of cells with expected counts less than 5. If more than 20 percent of the cells are in this condition, you should consider grouping categories of one or both variables.

Note

Another option to deal with sparse cells is to use exact statistical tests (available in the PASW Statistics Exact Tests option).

6.10 *Requesting the Chi-Square Test*

The procedure to request a Chi-Square test is to:

1) Select one or more row variables.
2) Select one or more column variables.
3) Request appropriate percentages.
4) Request the Chi-Square statistic.
5) Inspect the Chi-Square test output, if the test results in a significant result, conclude the null hypothesis of independence is rejected and report the differences in percentages. If the Chi-Square is not significant, then conclude there are sample differences, but equality of percentages in the population cannot be rejected.

6.11 *Chi-Square Output*

The table titled Chi-Square Tests shows the Chi-Square test.

There are three Chi-Square values listed, the first two of which are used to test for a relationship. We concentrate on the Pearson Chi-Square statistic, which is adequate for almost all purposes. The actual value (here 153.16), is not important, nor is the number of degrees of freedom (df), which is related to the number of cells in the table. These values are used to calculate the significance for the Chi-Square statistic, labeled “Asymp. Sig. (2-sided).” We interpret the significance value as the chance, *if the null hypothesis is true*, of finding a Chi-Square value at least this large. It is this value that we use to test the null hypothesis. If the significance value is smaller than the preset alpha, then we reject the null hypothesis of independence.

Figure 6.6. Chi-Square Tests Output

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	153.155[a]	8	.000
Likelihood Ratio	157.888	8	.000
Linear-by-Linear Association	117.739	1	.000
N of Valid Cases	2010		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.94.

Best Practice

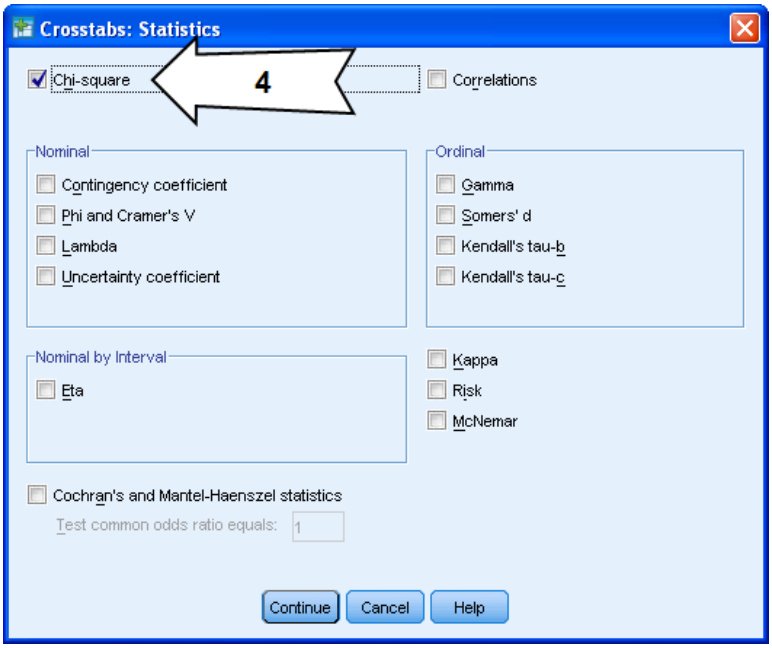
In terms of formal testing, we often use significance values of .05 or .01 to determine whether we should reject the null hypothesis.

6.12 *Procedure: Chi-Square Test*

To include the Chi-Square test in a Crosstabs, with the **Crosstabs** dialog open:

1) Place one or more variables in the Row(s) box.
2) Place one or more variables in the Column(s) box.
3) Optionally, include appropriate percentages in the Cells dialog.
4) Select Chi-square in the Statistics dialog.

Figure 6.7 Statistics Dialog



6.13 *Example: Chi-Square Test*

We will work with the *Census.sav* data file in this lesson.

In this example we continue to examine the relationship between marital status (*marital*) and happiness with one's life overall (*happy*). We would like to determine whether the percentage differences we observed above are likely to be observed in the total population.

Detailed Steps for Crosstabs with Chi-Square Test

1) Place the variable ***happy*** in the Row(s) box.
2) Place the variable ***marital*** in the Column(s) box.
3) Select **Column** in the Cells Display dialog.
4) Select **Chi-square** in the Statistics dialog.

Results from Crosstabs with Chi-Square Test

The **Crosstabs** table percentages are based on the columns, or on *marital*. We observe that the marital groups differ substantially in their happiness (e.g. the percentage not too happy is 8.4% for those married versus 25.7% for those married).

Figure 6.8 Crosstabulation of Marital Status and General Happiness

GENERAL HAPPINESS * MARITAL STATUS Crosstabulation

			MARITAL STATUS					
			MARRIED	WIDOWED	DIVORCED	SEPARATED	NEVER MARRIED	Total
GENERAL HAPPINESS	VERY HAPPY	Count	398	31	56	11	101	597
		% within MARITAL STATUS	41.1%	19.0%	20.0%	15.7%	19.1%	29.7%
	PRETTY HAPPY	Count	490	95	169	41	304	1099
		% within MARITAL STATUS	50.6%	58.3%	60.4%	58.6%	57.6%	54.7%
	NOT TOO HAPPY	Count	81	37	55	18	123	314
		% within MARITAL STATUS	8.4%	22.7%	19.6%	25.7%	23.3%	15.6%
Total		Count	969	163	280	70	528	2010
		% within MARITAL STATUS	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

The table titled Chi-Square Tests shows that the significance value is quite small and is rounded off to .000, which means that the actual value is less than .0005. With the significance being so low, it is rather unlikely that the null hypothesis is true. Therefore, it is reasonable to conclude that there is a relationship between marital status and general happiness. The exact form of the relationship is as described above, i.e., as depicted in the table.

Figure 6.9 Chi-Square Tests

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	153.155[a]	8	.000
Likelihood Ratio	157.033	8	.000
Linear-by-Linear Association	117.739	1	.000
N of Valid Cases	2010		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.94.

Apply Your Knowledge

1. A chi-square value from a test of two variables is 53.4. What should we conclude about the relationship of these two variables?
    a. There is no relationship
    b. There is a relationship
    c. We need to know the significance of this value to reach a conclusion
    d. Both A and C

2. True or false? A Chi-Square test is done differently for nominal compared to ordinal variables?

3. See the output below. True or false? There is a statistically significant relationship between Minority Classification and Employment Category (alpha=0.05).

Employment Category * Minority Classification Crosstabulation

			Minority Classification		
			No	Yes	Total
Employment Category	Clerical	Count	274	87	361
		% within Minority Classification	74.7%	83.7%	76.6%
	Custodial	Count	14	13	27
		% within Minority Classification	3.8%	12.5%	5.7%
	Manager	Count	79	4	83
		% within Minority Classification	21.5%	3.8%	17.6%
Total		Count	367	104	471
		% within Minority Classification	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	25.893[a]	2	.000

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.96.

4. See the table below. True or false? It is incorrect to do the Chi-Square test, because more than 3 out of the 6 cells have an observed count less than 5.

Y * X Crosstabulation

			X			
			A	B	C	Total
Y	A	Count	100	0	100	200
		% within X	100.0%	.0%	100.0%	50.0%
	B	Count	0	200	0	200
		% within X	.0%	100.0%	.0%	50.0%
Total		Count	100	200	100	400
		% within X	100.0%	100.0%	100.0%	100.0%

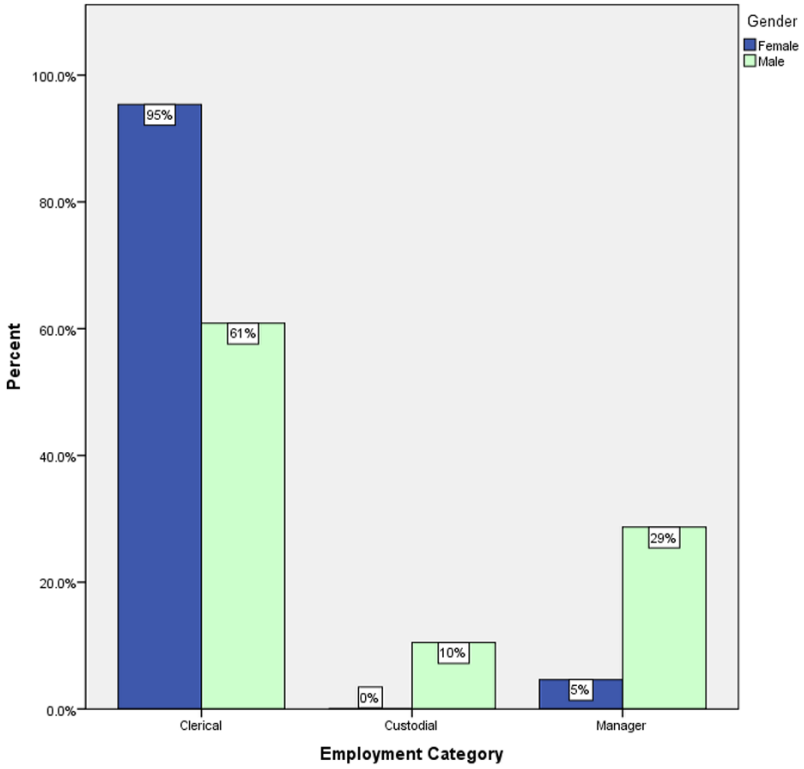
6.14 *Clustered Bar Chart*

For presentations it is often useful to show a graph of the relationship between two categorical variables. Clustered bar charts are the most effective method of doing this for **Crosstabs**.

Clustered Bar Chart Illustrated

As an illustration of a clustered bar chart, see the graph below. The bars show the percentage in each of the job categories by gender. For example, 95% of the women are clerical versus 61% of the men.

Figure 6.10 Clustered Bar Chart Illustrated



6.15 *Requesting a Clustered Bar Chart with Chart Builder*

The Chart Builder procedure allows for the creation of a variety of charts, including clustered bar charts. It provides for great flexibility in creating charts, including formatting options.

To create a clustered bar chart:

1) In the Chart Builder, select the clustered bar chart type of graph.
2) Specify the appropriate variables.
3) Specify the appropriate percentage statistic.
4) Inspect the graph in the output and compare the percentages.

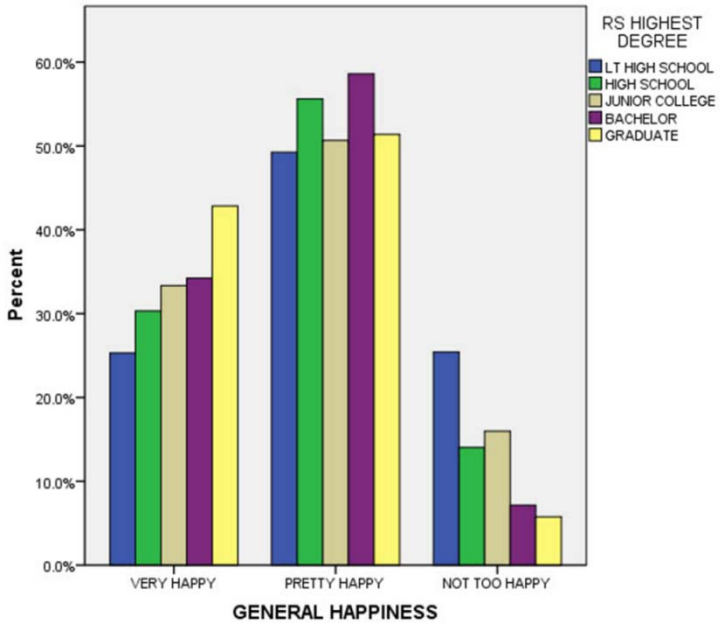
6.16 *Clustered Bar Chart from Chart Builder Output*

The clustered bar chart from Chart Builder gives the user control of:

- Which variable is used for clustering, and which variable defines the X-axis
- What variable is used for percentages
- The chart also correctly lists that the percentage is displayed, and it has labels and a legend that is placed outside the chart area.

In the graph depicted below, *RS HIGHEST DEGREE [degree]* is used for clustering (in the legend), *GENERAL HAPPINESS [happy]* defines the X-axis. Percentages are based on *degree*, so that percentages for each degree sum to 100%.

Figure 6.11 Example of a Clustered Bar Chart from Chart Builder

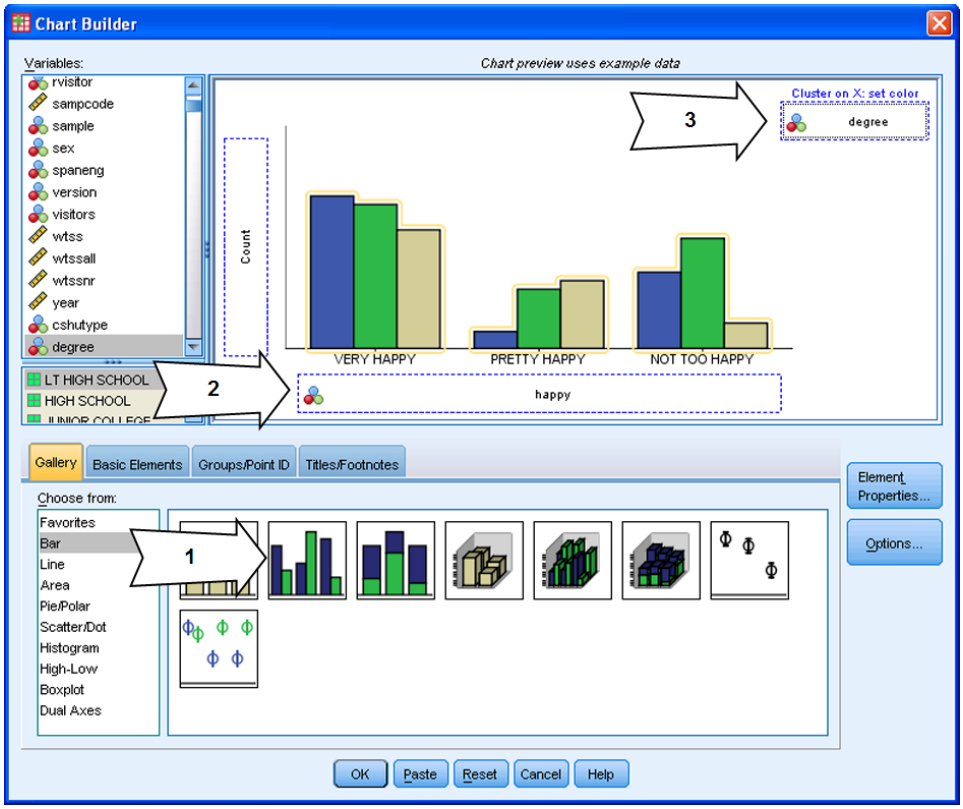


6.17 *Procedure: Clustered Bar Chart with Chart Builder*

The Chart Builder procedure is accessed from the **Graphs…Chart Builder** menu. With the Chart Builder dialog open:

1) Select a clustered bar chart in the Chart Builder and drag it into the Chart preview area
2) Select a variable for the X-axis.
3) Select a variable for the Cluster on X: set color box.

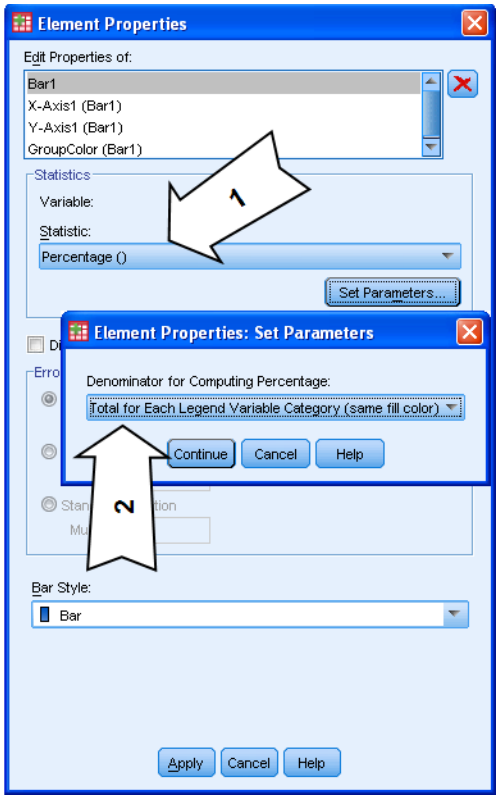
Figure 6.12 Chart Builder Dialog to Create Clustered Bar Chart



To specify the statistic:

1) In the Element Properties dialog, specify the Percentage() statistic.
2) Select the appropriate base for percentages in the Set Parameters dialog.

Figure 6.13 Setting Percentage in Element Properties Dialog



6.18 *Example: Clustered Bar Chart with Chart Builder*

We will create a clustered bar chart of the crosstab table of marital status and general happiness. We want to see how general happiness varies across categories of marital status, so we use marital status as the clustering variable. This is equivalent to how we percentage a crosstab by marital status to study this relationship.

Detailed Steps for Clustered Bar Chart

1) Select the **clustered bar chart icon** and put it in the Chart Preview pane.
2) Place ***happy*** in the X-axis box.
3) Place ***marital*** in the Set Color box.
4) Select the **Percentage()** statistic in the Element Properties Statistics drop down
5) Select **Total for Each Legend Variable Category** in the Set Parameters dialog.

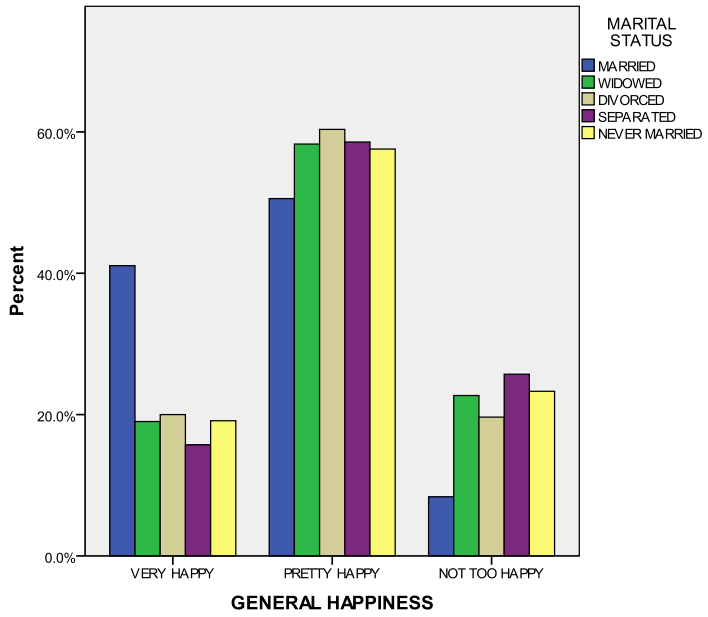
Results from the Clustered Bar Chart Created with Chart Builder

A crosstab table with percentages based from marital status lets us compare percentages of that variable within categories of general happiness, and this is mirrored in the clustered bar chart.

- There are 3 values on the X-axis for each separate value of general happiness
- At each value, the five categories of marital status are displayed with the bar representing percentages
- We can readily compare the categories of marital status in this arrangement, equivalent to comparing across rows in the crosstabulation

Married respondents are by far more likely to say they are very happy, and they are less likely to say they are not too happy. There isn't much difference between the other categories.

Figure 6.14 Clustered Bar Chart of Marital Status and General Happiness



6.19 *Adding a Control Variable*

Tables can be made more complex by adding variables to the layer dimension.

- Adding one or more layer variables to a **Crosstabs** table is known as adding *control variables*
- Using a control variable means that a subtable will be formed for every value of the control variable; thus, the relationship between the original variables will be *controlled* by examining it within values of the third variable
- You can add more than one control variable to a table by nesting variables in the layer (use the *Next* button), but unless you have a large sample size, or the variables have only a few categories, you may quickly create tables with only a few cases, or even no cases, in several cells

Control Variable Crosstabs Illustrated

The table depicted below includes *sex* as the control variable. Percentages are calculated per subtable.

Figure 6.15 A Control Variable Table

GENERAL HAPPINESS * MARITAL STATUS * RESPONDENTS SEX Crosstabulation

RESPONDENTS SEX				MARITAL STATUS					
				MARRIED	WIDOWED	DIVORCED	SEPARATED	NEVER MARRIED	Total
MALE	GENERAL HAPPINESS	VERY HAPPY	Count	183	4	26	6	52	271
			% within MARITAL STATUS	39.0%	12.9%	20.8%	22.2%	19.0%	29.3%
		PRETTY HAPPY	Count	243	17	75	15	163	513
			% within MARITAL STATUS	51.8%	54.8%	60.0%	55.6%	59.7%	55.5%
		NOT TOO HAPPY	Count	43	10	24	6	58	141
			% within MARITAL STATUS	9.2%	32.3%	19.2%	22.2%	21.2%	15.2%
	Total		Count	469	31	125	27	273	925
			% within MARITAL STATUS	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FEMALE	GENERAL HAPPINESS	VERY HAPPY	Count	215	27	30	5	49	326
			% within MARITAL STATUS	43.0%	20.5%	19.4%	11.6%	19.2%	30.0%
		PRETTY HAPPY	Count	247	78	94	26	141	586
			% within MARITAL STATUS	49.4%	59.1%	60.6%	60.5%	55.3%	54.0%
		NOT TOO HAPPY	Count	38	27	31	12	65	173
			% within MARITAL STATUS	7.6%	20.5%	20.0%	27.9%	25.5%	15.9%
	Total		Count	500	132	155	43	255	1085
			% within MARITAL STATUS	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Best Practice

Choosing the proper control variables is somewhat of an art, since there is no reason (nor is there time) to try every possible other variable as a control. Select control variables based on the goals of the study, any available theory, and also based on which variables are possibly related to one or both variables in the table. Demographic variables are often used as controls.

6.20 *Requesting a Control Variable*

Including a control variable in **Crosstabs** is accomplished with these steps:

1) Select row and column variables.
2) Add a control variable.
3) Request appropriate percentages.
4) Optionally, request the Chi-Square statistic.
5) In the output, compare the percentages within each category of the control variable to see if there is a relationship between the row variable and column variable, for that particular category.
6) In the output, inspect the Chi-Square test output for each of the subtables and compare these results.

6.21 *Control Variable Output*

The layer variable will be included as outer-left variable in the crosstabulation. Here the control variable is gender.

Figure 6.16 Crosstabs with a Control variable

R OWNS BUSINESS * RACE OF RESPONDENT * RESPONDENTS SEX Crosstabulation

RESPONDENTS SEX				RACE OF RESPONDENT			
				WHITE	BLACK	OTHER	Total
MALE	R OWNS BUSINESS	Yes	Count	126	9	12	147
			% within RACE OF RESPONDENT	17.6%	7.8%	12.6%	15.8%
		No	Count	591	107	83	781
			% within RACE OF RESPONDENT	82.4%	92.2%	87.4%	84.2%
	Total		Count	717	116	95	928
			% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%
FEMALE	R OWNS BUSINESS	Yes	Count	85	10	8	103
			% within RACE OF RESPONDENT	10.1%	6.1%	9.1%	9.4%
		No	Count	755	155	80	990
			% within RACE OF RESPONDENT	89.9%	93.9%	90.9%	90.6%
	Total		Count	840	165	88	1093
			% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%

If a Chi-Square test is requested, a Chi-Square test will be performed for each of the values of the control variable. The interpretation of the test is the same as in any table, but it applies only to each subtable (here, for the male and female subtables separately).

Figure 6.17 Chi-Square Test with a Control Variable

Chi-Square Tests

RESPONDENTS SEX		Value	df	Asymp. Sig. (2-sided)
MALE	Pearson Chi-Square	8.032[a]	2	.018
	Likelihood Ratio	9.123	2	.010
	Linear-by-Linear Association	4.540	1	.033
	N of Valid Cases	928		
FEMALE	Pearson Chi-Square	2.674[b]	2	.263
	Likelihood Ratio	2.951	2	.229
	Linear-by-Linear Association	1.073	1	.300
	N of Valid Cases	1093		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.05.

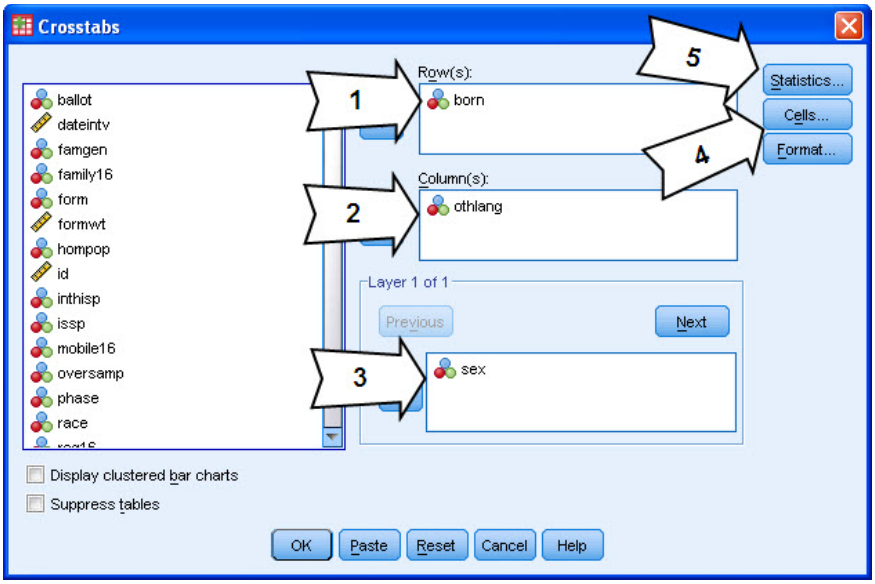
b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.29.

6.22 *Procedure: Adding a Control Variable*

With the **Crosstabs** dialog open, the procedure to include a control variable is:

1) Place one or more variables in the Rows box.
2) Place one or more variables in the Columns box.
3) Place the control variable in the Layer(s) box.
4) In the Cells dialog, select appropriate percentages.
5) Optionally, select Chi-Square in the Statistics dialog.

Figure 6.18 Crosstabs Dialog with Control Variable



6.23 *Example: Adding a Control Variable*

We'll create an entirely new table with three new variables to investigate the relationship between *sex*, *race*, and do you own a business (*ownbiz*). We'll use *sex* as the control variable to further illuminate the relationship between *race* and *ownbiz*.

To illustrate the benefit of adding a control variable, first we create the table of *race* and *ownbiz*.

Detailed Steps for the Two-Way Crosstabs

1) Select **Reset** button
2) Place ***[race*** in the Column(s) box
3) Place ***ownbiz*** in the Rows box
4) In the Cells Display dialog, select **Column** percents
5) Select **Chi-Square** in the Statistics dialog

Results for the Two-Way Crosstabs

We see that the percentages in the table suggests that whites (13.6%) are most likely to own a business, followed by those of other races (10.9%) and then blacks (6.8%).

Figure 6.19 Crosstab of Race and Owning a Business

R OWNS BUSINESS * RACE OF RESPONDENT Crosstabulation

			RACE OF RESPONDENT			
			WHITE	BLACK	OTHER	Total
R OWNS BUSINESS	Yes	Count	211	19	20	250
		% within RACE OF RESPONDENT	13.6%	6.8%	10.9%	12.4%
	No	Count	1346	262	163	1771
		% within RACE OF RESPONDENT	86.4%	93.2%	89.1%	87.6%
Total		Count	1557	281	183	2021
		% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%

Figure 6.20 Chi-Square Tests

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.510[a]	2	.005
Likelihood Ratio	11.871	2	.003
Linear-by-Linear Association	5.062	1	.024
N of Valid Cases	2021		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.64.

Detailed Steps for Control Variable Crosstabs

Now we'll see what happens when we add the variable *sex* to the layer.

1) Place ***sex*** in the Layer(s) box.

Results from Adding a Control Variable

The resulting table is much larger. Actually, there are two subtables, one for males and one for females.

Does the relationship vary by gender? To a certain extent it does.

- White males are much more likely than black males to own a business (almost a 10 percentage point difference)
- Black females are closer to white females in the percentage owning a business (a 4 percentage point difference)

Figure 6.21 Crosstab of Race, Sex, and Owning a Business

R OWNS BUSINESS * RACE OF RESPONDENT * RESPONDENTS SEX Crosstabulation

RESPONDENTS SEX				RACE OF RESPONDENT			
				WHITE	BLACK	OTHER	Total
MALE	R OWNS BUSINESS	Yes	Count	126	9	12	147
			% within RACE OF RESPONDENT	17.6%	7.8%	12.6%	15.8%
		No	Count	591	107	83	781
			% within RACE OF RESPONDENT	82.4%	92.2%	87.4%	84.2%
	Total		Count	717	116	95	928
			% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%
FEMALE	R OWNS BUSINESS	Yes	Count	85	10	8	103
			% within RACE OF RESPONDENT	10.1%	6.1%	9.1%	9.4%
		No	Count	755	155	80	990
			% within RACE OF RESPONDENT	89.9%	93.9%	90.9%	90.6%
	Total		Count	840	165	88	1093
			% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests for Control Variable

Are these differences we observe statistically significant? We can answer this, as before, with the chi-square test.

Males. The chi-square test is significant (.018). This means that, for males, race is related to owning a business.

Females. The chi-square test is not significant (.263). This means that, for females, there is no relationship between race and owning a business.

Figure 6.22 Chi-Square Tests with a Control Variable

Chi-Square Tests

RESPONDENTS SEX		Value	df	Asymp. Sig. (2-sided)
MALE	Pearson Chi-Square	8.032[a]	2	.018
	Likelihood Ratio	9.123	2	.010
	Linear-by-Linear Association	4.540	1	.033
	N of Valid Cases	928		
FEMALE	Pearson Chi-Square	2.674[b]	2	.263
	Likelihood Ratio	2.951	2	.229
	Linear-by-Linear Association	1.073	1	.300
	N of Valid Cases	1093		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.05.

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.29.

We added a control variable and discovered that the initial relationship depends somewhat on the respondent's gender. What does this mean about our analysis of the bivariate table? Should the results in this table be entirely discarded in favor of the three-way table?

The first point to make in answering these questions is that this example confirms the importance of multivariate analysis. That said, we can add these comments:

1) The bivariate table is still "correct," although that word is better replaced by "valuable." The bivariate table *is* our estimate of the relationship between owing a business and race.
2) However, if we generalize this relationship and believe it is the same for males and females, we would be incorrect.
3) Adding the control variable has given us additional insight into this relationship. It may help us understand more about the two-way relationship.

Apply Your Knowledge

1. See the output below. Which statements are correct?
    a. The table shows that there is a statistically significant relationship between gender and minority classification (alpha=0.05)
    b. The relationship between current salary (binned) and minority classification is not significant for women (alpha=0.05)
    c. The relationship between current salary (binned) and minority classification is not significant for men (alpha=0.05)

d. The relationship between current salary (binned) and minority classification is not the same for men and women

Current Salary (Binned) * Minority Classification * Gender Crosstabulation

Gender				Minority Classification		
				No	Yes	Total
Female	Current Salary (Binned)	<= 25500	Count	100	30	130
			% within Minority Classification	58.5%	75.0%	61.6%
		25501 - 33300	Count	44	9	53
			% within Minority Classification	25.7%	22.5%	25.1%
		33301+	Count	27	1	28
			% within Minority Classification	15.8%	2.5%	13.3%
	Total		Count	171	40	211
			% within Minority Classification	100.0%	100.0%	100.0%
Male	Current Salary (Binned)	<= 25500	Count	14	12	26
			% within Minority Classification	7.4%	18.8%	10.3%
		25501 - 33300	Count	64	39	103
			% within Minority Classification	33.9%	60.9%	40.7%
		33301+	Count	111	13	124
			% within Minority Classification	58.7%	20.3%	49.0%
	Total		Count	189	64	253
			% within Minority Classification	100.0%	100.0%	100.0%

Chi-Square Tests

Gender		Value	df	Asymp. Sig. (2-sided)
Female	Pearson Chi-Square	5.885[a]	2	.053
	N of Valid Cases	211		
Male	Pearson Chi-Square	28.992[b]	2	.000
	N of Valid Cases	253		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.31.

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.58.

Brainstorming Exercise

In regard to your own data, what are some potentially important control variables for your analyses?

6.24 *Extensions: Beyond Crosstabs*

Decision Tree analysis is often used by data analysts who need to predict to which group an individual can be classified, based on potentially many nominal or ordinal background variables. For example, an insurance company is interested in the combination of demographics that best predict whether a client is likely to make a claim. Or a direct mail analyst is interested in the combinations of background characteristics that yield the highest return rates. Here the emphasis is less on testing a hypothesis and more on a heuristic method of finding the optimal set of characteristics for prediction purposes. CHAID (Chi-Square automatic interaction detection), a commonly used type of decision-tree technique, along with other decision-tree methods are available in the PASW Decision Trees add-on module.

A technique called **loglinear modeling** can also be used to analyze multi-way tables. This method requires statistical sophistication and is well beyond the domain of this course. PASW Statistics has several procedures (Genlog, Loglinear and Hiloglinear) to perform such analyses. They provide a way of determining which variables relate to which others in the context of a multi-way crosstab (also called contingency) table. These procedures could be used to explicitly test for the three-way interaction suggested above. For an introduction to this methodology see Fienberg (1977). Academic researchers often use such models to test hypotheses based on many types of data.

6.25 *Association Measures*

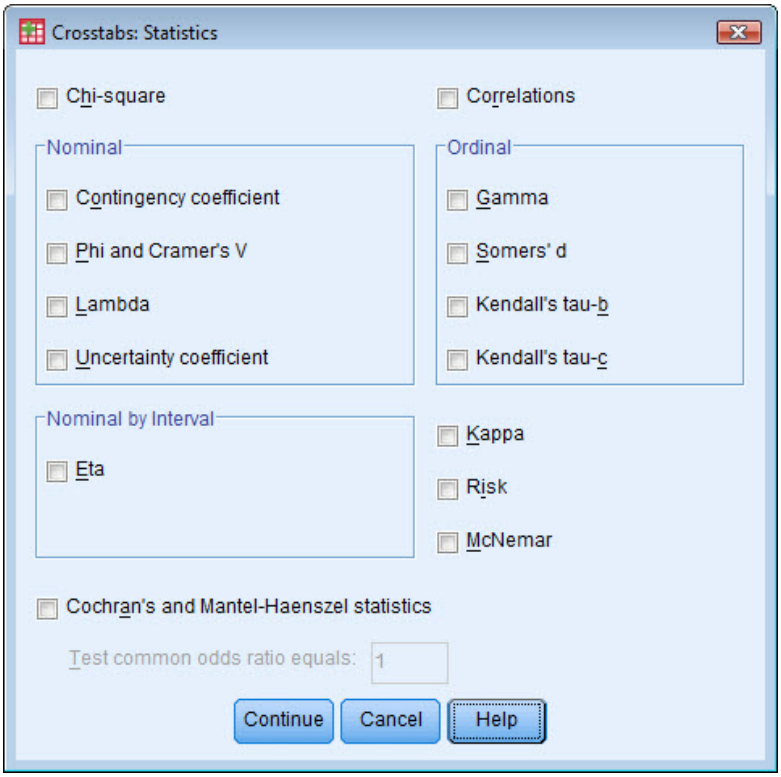
Measures of association for categorical variables have been developed to summarize the strength of a relationship in a single statistic. This allows you to compare different tables (groups) concisely.

- They are typically normed to range between 0 and 1 for variables on a nominal scale, or –1 and 1 for variables on an ordinal scale.
- The specific measure used thus depends on the level of measurement of the variables, among other things
- In general, measures of association are bivariate rather than designed for multi-way tables

The **Crosstabs** procedure provides many measures of association. Several are available because different aspects of the association are emphasized by particular measures.

There are four tests each in the Nominal and Ordinal areas. The variable at the lowest level of measurement determines which test can be used, e.g., in a table with a nominal and ordinal variable, you must use one of the nominal tests.
For a discussion of these tests, see Gibbons (2005).

Figure 6.23 Crosstabs Statistics Dialog with Measures of Association



6.26 *Lesson Summary*

We explored the use of the **Crosstabs** procedure to analyze relationships between categorical variables.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Perform crosstab analysis on categorical variables

To support the achievement of the primary objective, students should now also be able to:

- Use the options in the **Crosstabs** procedure
- Request appropriate statistics for a crosstabulation
- Interpret cell counts and percents in a crosstabulation
- Use the Chi-Square test, interpret its results, and check its assumptions
- Use the **Chart Builder** to visualize a crosstabulation

6.27 *Learning Activity*

The overall goal of this learning activity is to create two and three-way crosstabulations to explore the relationship between several variables and to use the **Chart Builder** to visualize the relationship. You'll use the file *Census.sav*.

Supporting Materials

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. Investigate the relationship between the variables *race* and self-rated health (*health*), can people be trusted (*cantrust*), and support for spending on scientific research (*natsci*). Request appropriate percentages and a Chi-Square test.

2. What is the significance of the Chi-Square test? Is there a statistically significant relationship between race and these variables? Describe the relationships that you observe in the tables.

3. Create a clustered bar chart to display the relationship of one of the tables with a significant relationship.

4. Now add *sex* as a control variable. Are there differences in the relationship by gender? Are the relationships in the subtables significant or not? Are they different or not?

5. Now remove the control variable *sex* and substitute the variable *born* (was the respondent born in the U.S. or not). Does the relationship between race and self-rated health vary depending on place of birth?

6. *For those with extra time:* Select some variables that you find interesting and explore their relationship with **Crosstabs**. Begin with two-way tables before investigating more complex relationships.

Lesson 7: The Independent- Samples T Test

7.1 *Objectives*

After completing this lesson students will be able to:

- Perform a statistical test to determine whether there is a statistically significant difference between two groups on a scale variable

To support the achievement of this primary objective, students will also be able to:

- Check the assumptions of the **Independent-Samples T Test**
- Use the **Independent-Samples T Test** to test the difference in means
- Know how to interpret the results of a **Independent-Samples T Test**
- Use the **Chart Builder** to create an error bar graph to display mean differences

7.2 *Introduction*

When our purpose is to examine group differences on scale variables, we turn to the mean as the summary statistic since it provides a single measure of central tendency. In this lesson we outline the logic involved when testing for mean differences between groups, state the assumptions, and then perform an analysis comparing two groups.

Business Context

When analyzing data, we are concerned with whether groups differ from each other. Without statistical testing, we might make decisions based on perceptions that are not likely to exist in a population of customers. An **Independent-Samples T Test** allows us to determine if two groups differ significantly on a scale variable. For example, we might want to know whether:

- One customer group purchases more items, on average, than a second
- Drug A reduces depression levels better than drug B
- Student test scores in one class are higher than in a second class

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

7.3 *The Independent-Samples T Test*

For historical reasons related to the development of statistics and lack of computing power, separate methods were developed to test for mean differences when the predictor or grouping variable had two categories versus three or more. For dichotomous predictors we use the Independent-Samples T Test; for predictors with more categories we use analysis of variance (ANOVA). So, the Independent-

Samples T Test applies when there are two separate populations to compare on a scale dependent variable (for example, males and females on salary).
As with any statistical test, the goal is to draw conclusions about the population, based on sample data. For the Independent-Samples T Test the null hypothesis states that population means are the same, and we use sample data to evaluate this hypothesis.

So, Ho (the null hypothesis) assumes that the population means are identical. We then determine if the differences in sample means are consistent with this assumption. If the probability of obtaining sample means as far (or further) apart as we find in our sample is very small (less than 5 chances in 100 or .05), assuming no population differences, we reject our null hypothesis and conclude the populations are different.

7.4 *Independent-Samples T Test Assumptions*

To correctly use the Independent-Samples T Test requires a number of assumptions to be made.

1) Only two population subgroups are compared.
2) The dependent variable has a scale measurement level.
3) The distribution of the dependent variable within each population subgroup follows the normal distribution (normality).

Note

The Independent-Samples T Test is robust to moderate violations of the normality assumption when sample sizes are moderate to large (over 50 cases per group) and the dependent measure has the same distribution (for example, skewed to the right) within each group

4) The variation is the same within each population subgroup. This is the so-called "homogeneity of variance" assumption. Violation of this assumption is more critical than violation of the normality assumption. When this assumption is violated, the significance or probability value reported by PASW Statistics is incorrect and the test statistics must be adjusted.

Note

The **Independent-Samples T Test** is robust to moderate violations of the homogeneity of variance assumption. If the ratio *(greatest variance / smallest variance)* is less than 2, with similar sample sizes of the groups, then violation is not serious.
The **Independent-Samples T Test** provides a statistical test for the assumption of homogeneity of variance and an alternative test for testing equality of means, taking into account unequal variances.

7.5 *Requesting the Independent-Samples T Test*

Requesting an **Independent-Samples T Test** is accomplished with these steps:

1) Select one or more variables to be tested that are scale in measurement level.
2) Select a grouping or test variable and define the two groups.
3) Optionally, set the confidence interval if you prefer something other than 95%.
4) Review the procedure output to investigate the relationship between the variables including:
    a. Group Statistics Table

b. Check the assumptions of the **Independent-Samples T Test**.

5) Examine the t test statistics to determine whether there is a significant difference in the means.

7.6 *Independent-Samples T Test Output*

The Group Statistics table provides sample sizes, means, standard deviations, and standard errors for the two groups. In the table below, there is a 1.4 year difference in group means on highest year of school completed. The sample standard deviations (and so the variances) are quite different, indicating potentially unequal population variances.

Figure 7.1 Example of Group Statistics

Group Statistics

	WAS R BORN IN THIS COUNTRY	N	Mean	Std. Deviation	Std. Error Mean
HIGHEST YEAR OF SCHOOL COMPLETED	YES	1756	13.61	2.792	.067
	NO	262	12.24	4.383	.271

To understand how to work with the second table, the Independent Samples Test, we need to review the assumptions for conducting an **Independent-Samples T Test**. The most serious violation is that of the assumption of equal variances, so this assumption must be tested. Levene´s Test for Equality of Variances does exactly this. Levene´s homogeneity of variances test evaluates the null hypothesis that the dependent variable's variance is the same in the two populations. Since homogeneity of variance is assumed when performing the Independent-Samples T Test, the analyst hopes to find this test to be *nonsignificant*.

In the first, left half, section of the Independent Samples Test table, Levene´s test for equality of variances is displayed.

- The null hypothesis of Levene´s test is that the variances are equal
- The F statistic is a technical detail to calculate the significance *(Sig.)*
- The significance is the likelihood that the variances have the observed difference, or a greater difference, in the target population
- We use a standard criterion value of .05 or .01 to reject the null hypothesis

Best Practice

As sample size increases, it is better to use the .01 level for the Levene's test.

Figure 7.2. Example of Levene´s Test for Equality of Variances Output

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
HIGHEST YEAR OF SCHOOL COMPLETED	Equal variances assumed	96.438	.000
	Equal variances not assumed		

Here, in the example output, we observe that the hypothesis of equal variances must be rejected because the significance value is low.

The second, right half, section of the Independent Samples Test table provides the test statistics for the null hypothesis of equal group means. The row labeled “Equal variances assumed” contains results of the standard t test. The second row labeled “Equal variances not assumed” contains an adjusted t test that corrects for lack of homogeneity of variances in the data. You would choose one or the other based on your evaluation of the homogeneity of variance question. Summarizing: the result of the Levene test tells us which of the two rows of t test statistics to use:

- If the null hypothesis of equal variances is not rejected, then use the row labeled “Equal variances assumed”
- If the null hypothesis of equal variances is rejected, then use the row “Equal Variances not assumed”

To test equality of means:

- The null hypothesis of the test is that the means are equal
- The t and df statistics are technical details to calculate the significance (*Sig. 2-tailed)*
- The significance is the likelihood that the means have the observed difference, or a greater difference, in the target population
- We use a standard criterion value of .05 or .01 to reject the null hypothesis

Figure 7.3 Example of T Test for Equality of Means Output

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HIGHEST YEAR OF SCHOOL COMPLETED	Equal variances assumed	6.768	2016	.000	1.365	.202	.970	1.761
	Equal variances not assumed	4.895	293.40	.000	1.365	.279	.816	1.914

Here, we observe that the hypothesis of equal population means must be rejected (reading the row Equal variances not assumed, since the null hypothesis of equal variances was rejected).

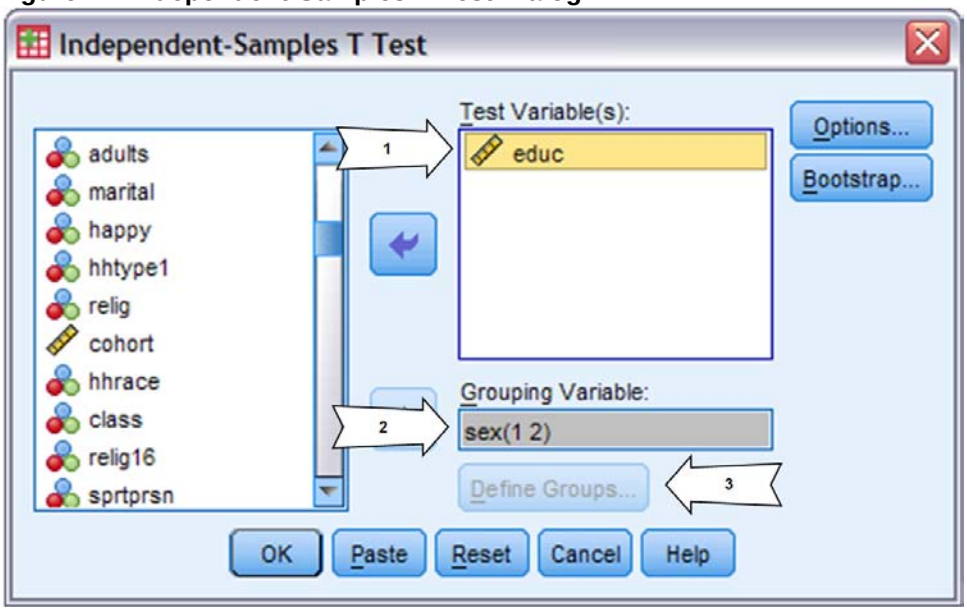
The Independent Samples Test table provides an additional bit of useful information: the 95% confidence band for the population mean difference. The 95% confidence band for the difference provides a measure of the precision with which we have estimated the true population difference. In the output shown below, the 95% confidence band for the mean difference between groups is .816 years to 1.914 years (again using the Equal variances not assumed row). Note that the difference values does not include zero, because there is a difference between groups. So, the 95% confidence band indicates the likely range within which we expect the population mean difference to fall. Speaking in a technically correct fashion, if we were to continually repeat this study, we would expect the true population difference to fall within the confidence bands 95% of the time. While the technical definition is not illuminating, the 95% confidence band provides a useful precision indicator of our estimate of the group difference.

7.7 Procedure: Independent-Samples T Test

The **Independent Samples T Test** procedure is accessed from the **Analyze…Compare Means…Independent-Samples T Test** menu choice. With the **Independent-Samples T Test** dialog box open:

1) Place one or more scale dependent variables in the Test Variable box.
2) Place one categorical independent variable in the Grouping Variable box.
3) Click the Define Groups button to specify which two groups are being compared.

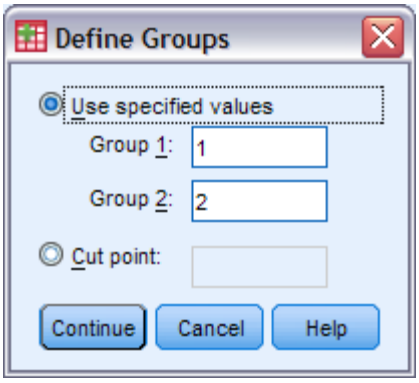
Figure 7.4 Independent-Samples T Test Dialog



In the Define Groups dialog:

4) Specify the two group values

Figure 7.5 Independent-Samples T-Test --- Define Groups Dialog

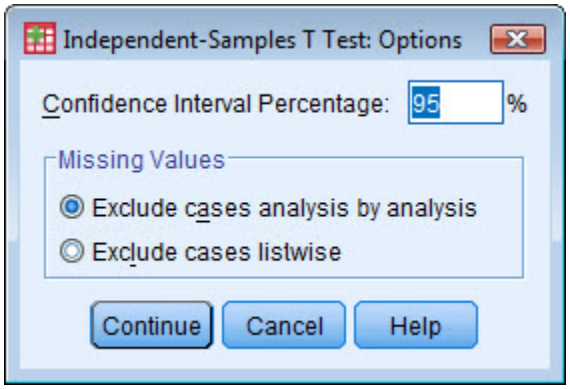


Tip

If the independent variable is a numeric variable, specify a single cut point value to define the two groups. Those cases less than or equal to the cut point go into the first group, and those greater than the cut point fall into the second group.
If the independent variable is categorical but has more than two categories, it can be still used by using only two categories in an analysis.

The Options dialog can be used to change the confidence interval percentage and change the handling of missing values to listwise.

Figure 7.6 T Test Options Dialog



7.8 *Demonstration: Independent-Samples T Test*

We will work with the *Census.sav* data file in this lesson.

In this example we examine the relationship between the respondent's gender (*sex*) and number of children (*childs*). You would expect that the means would be equal—it normally takes two to procreate—but let's investigate the question.

Before doing the actual test, we should explore the data to compare the distributions of number of children by gender. We'll use the **Explore** procedure to do so.

Note

We will not repeat details here on the Explore dialog box or the options available with the procedure. If you need a review, see the Lesson on *Understanding Data Distributions for Scale Variables*.

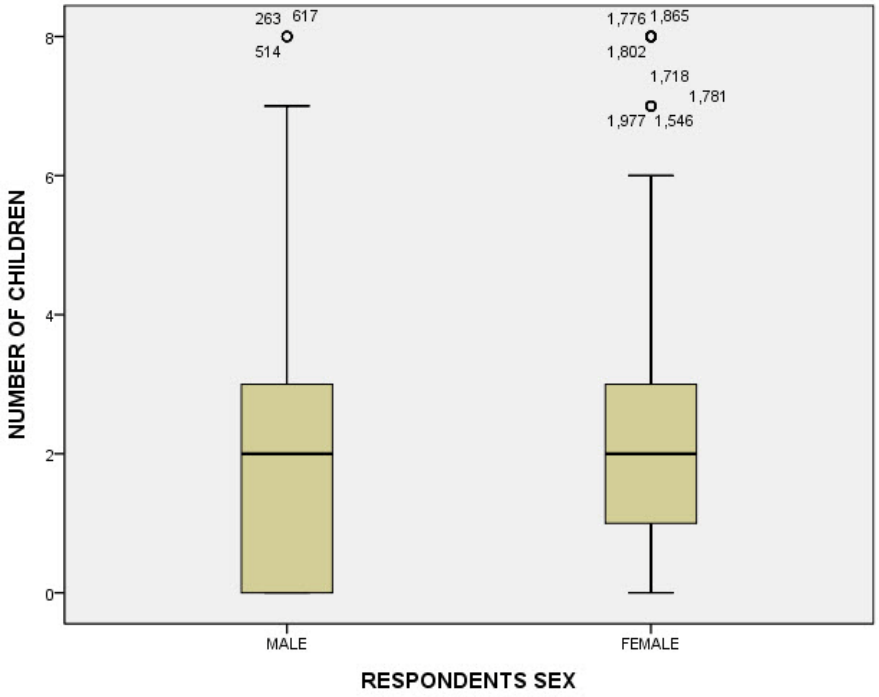
Detailed Steps for Explore Procedure for Number of Children by Gender

The **Explore** dialog box is accessed from the **Analyze…Descriptive Statistics…Explore** menu. With the dialog box open:

1) Place ***childs*** in the Dependent List: box
2) Place ***sex*** in the Factor List: box
3) In the **Plots** dialog, select **Histogram** check box

We'll concentrate first on the boxplot. The highest value of *childs* is 8 (values any higher are coded to 8). Still, there is enough variation in the data to see that the male and female distributions are not identical. The inter-quartile range for males is wider than for females, and it extends all the way to 0. The median for both genders is the same (2), and there are fewer outliers for males as a consequence of the wider IQR.

Figure 7.7 Boxplot of Number of Children by Gender



Looking next at the histograms, we observe that the distributions are not normal. However, given the sample size, this shouldn't be a deterrent to doing a t test. We might be more concerned as to whether the mean is an suitable measure of central tendency of these distributions, especially for males. But given the narrow range of *childs*, from 0 to 8, it seems reasonable to proceed.

Figure 7.8 Histogram of Number of Children for Males

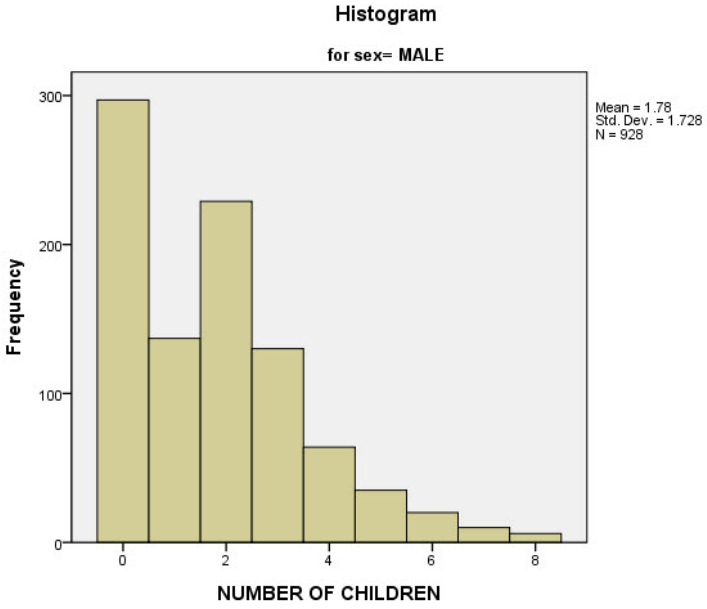
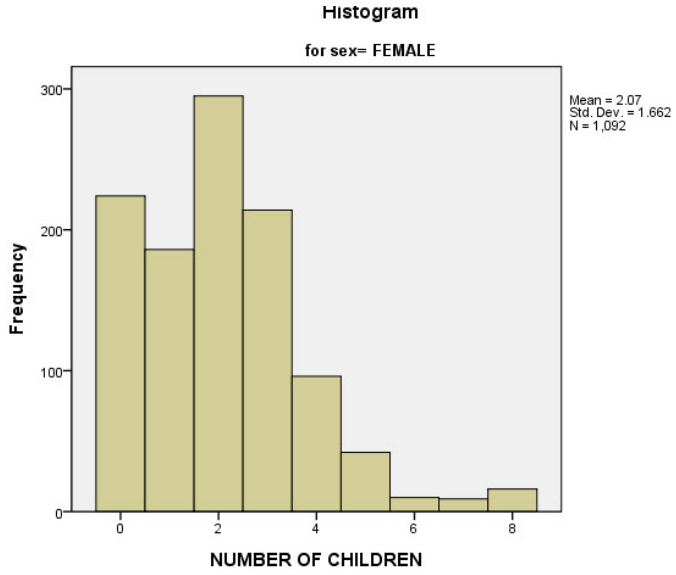


Figure 7.9 Histogram of Number of Children for Females



Detailed Steps for Independent Samples T-Test

1) Place the variable ***childs*** in the Test Variable(s) box.
2) Place the variable ***sex*** in the Grouping Variable box.
3) Select the **Define Groups** button.

Notice the question marks following *sex*. The **Independent-Samples T Test** dialog requires that you indicate which groups are to be compared, which is usually done by providing the data values for the two groups.

4) Specify that groups **1** and **2** are being compared

Results from Independent Samples T-Test

As with all analyses, you should first look to see how many cases are in each group, along with the means and standard deviations. The Group Statistics table provides this information. We have fairly large samples for each group. Intriguingly, the means are not very close for males and females. The mean number of children for females is about .30 higher than for males (The actual sample mean difference is displayed in the Independent Samples Test table). The standard deviations are similar for each group, which is a bit unexpected since the boxplots looked somewhat different.

Figure 7.10 Group Statistics Table

Group Statistics

	RESPONDENTS SEX	N	Mean	Std. Deviation	Std. Error Mean
NUMBER OF CHILDREN	MALE	928	1.78	1.728	.057
	FEMALE	1092	2.07	1.662	.050

Reading an Independent Samples Test Table

Looking at Levene´s test, the null hypothesis assuming homogeneity is not rejected at the .01 level. Given the large sample size, it makes more sense to use a more stringent alpha level. So, we may assume homogeneity of variances, and we take the result of the t test from the Equal variances assumed row (actually, the two rows gives very similar results in this example).

Figure 7.11 Independent Samples Levene's Test

Independe

		Levene's Test for Equality of Variances		
		F	Sig.	
NUMBER OF CHILDREN	Equal variances assumed	6.009	.014	-3
	Equal variances not assumed			-3

To test equality of means, move to the column labeled "Sig. (2-tailed)." This is the probability of our obtaining sample means as far or further apart, by chance alone, if the two populations (males and females) actually have the same number of children. Thus the probability of obtaining such a large difference by chance alone is quite small (.000), so we conclude there is a significant difference in mean number of children between men and women. Can you suggest how that could be true?

The 95% confidence band for the mean difference between groups is from –.438 to –.142 years.

Figure 7.12 Independent Samples T Test Results

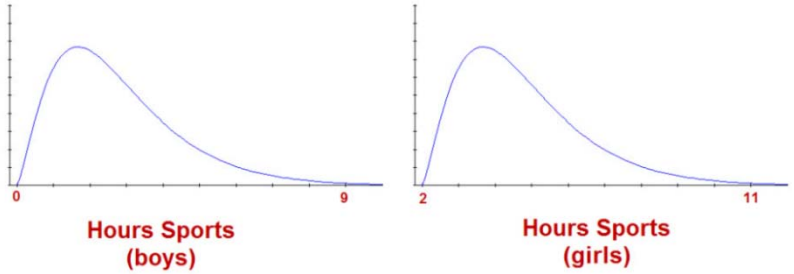
Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
NUMBER OF CHILDREN	Equal variances assumed	-3.838	2018	.000	-.290	.076	-.438	-.142
	Equal variances not assumed	-3.826	1939.092	.000	-.290	.076	-.439	-.141

Apply Your Knowledge

1. In which of the following situations can an Independent-Samples T Test be applied?
   a. Difference between men and women with respect to political preferences (liberal/ conservative/independent)
   b. Difference in mean age between liberals and conservatives
   c. Difference in mean income between three groups of political affiliation (liberals, conservatives and independents)
   d. Difference in mean expenditure between those shopping at Harrods and those not shopping at Harrods

2. True or False? Suppose we want to test whether boys and girls differ in mean hours (a week) doing sports. Below we have the distribution of this variable, for both genders. Now we draw a random sample of 50 boys and 50 girls. The Independent-Samples T Test cannot be done because the distribution of hours sport is not normal within each of the groups?



3. See the output below. Which statements are correct?
   a. To test equality of means, we use the "Equal variances assumed" row and disregard the row "Equal variances not assumed."
   b. The null hypothesis of equal group means is rejected (alpha=0.05)
   c. The 95% confidence interval for the difference contains the value 0, which indicates that the null hypothesis of equal group means cannot be rejected (alpha=0.05).

Group Statistics

	gender child	N	Mean	Std. Deviation	Std. Error Mean
Hours sports per week	boy	489	4.9937	1.01987	.04612
	girl	511	4.9690	.98973	.04378

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hours sports per week	Equal variances assumed	.489	.484	.390	998	.697	.02476	.06355	-.0999	.14947
	Equal variances not assumed			.389	992.561	.697	.02476	.06359	-.1000	.14956

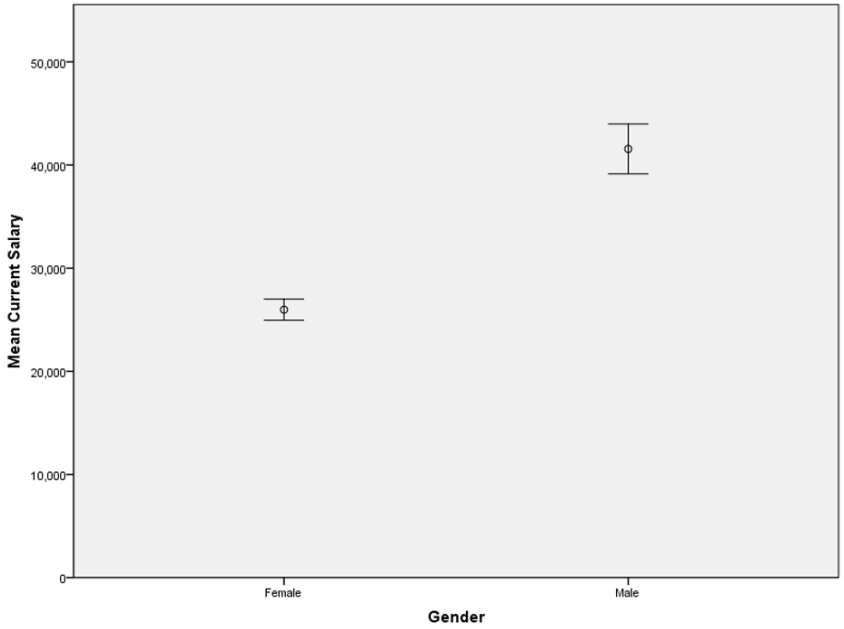
7.9 *Error Bar Chart*

Although the **Independent-Samples T Test** procedure displays the appropriate statistical test information, a summary chart is often preferred as a way to present significant results. Bar charts displaying the group sample means can be produced using the **Chart Builder** procedure. However, many people prefer an error bar chart instead. It is a chart that focuses more on the precision of the estimated mean for each group than the mean itself.

Error Bar Chart Illustrated

The figure below provides an example of an the error bar chart. The graph shows mean *salary* and its associated 95% confidence interval, for each of two groups. Mean salary for men is higher than for women and the intervals do not overlap, indicating differences between men and women in the population in mean salary.

Figure 7.13 Error Bar Chart Illustrated



7.10 *Requesting an Error Bar Chart with Chart Builder*

Requesting an Error bar chart is accomplished with these steps, using the **Chart Builder** procedure.

1) Place an Error Bar chart icon in the Chart Preview area
2) Select the scale variable for which you want a mean and confidence intervals.
3) Select the categorical variable defining the groups.
4) In the resulting graph, check if confidence intervals overlap.

Note

This method of comparing confidence intervals is not as precise as statistical testing. For example, an error bar chart does not take into account whether the homogeneity of variance assumption is met. Still, used carefully, they can be very useful.

7.11 *Error Bar Chart Output*

The error bar chart will generate a graph depicting the relationship between a scale and categorical variable. It provides a visual sense of how far the groups are separated.

- Note the means of each group
- Note if the 95% confidence intervals of the groups overlap

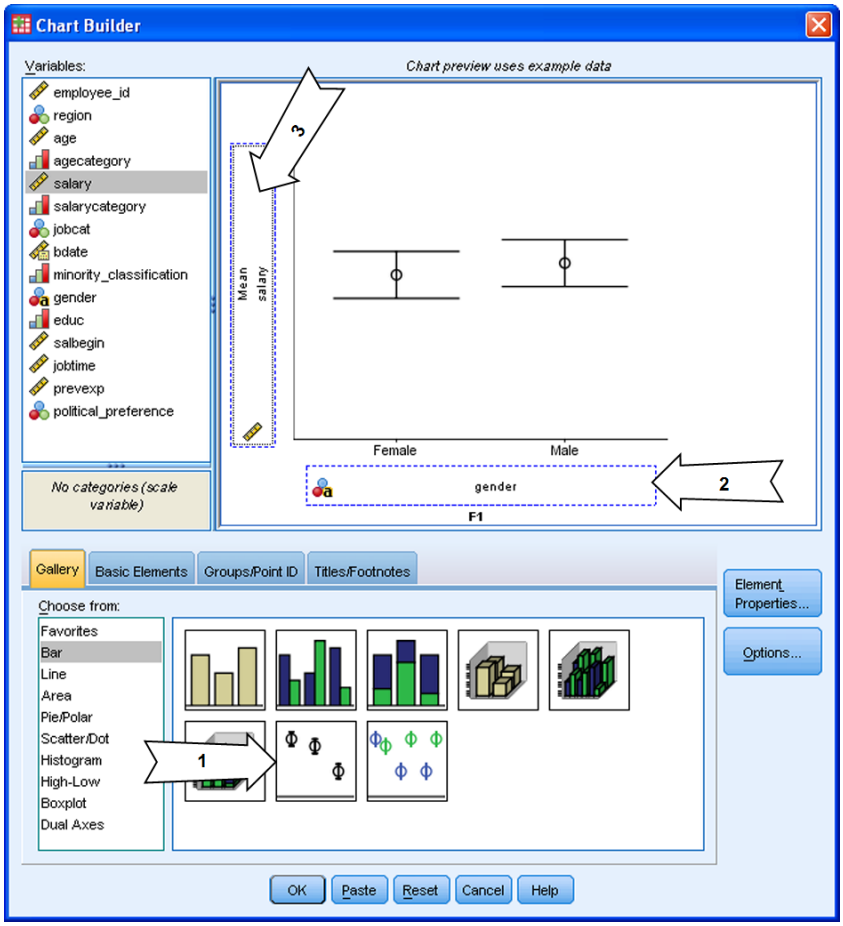
- If the error bar and statistical test lead to the same conclusion, support the statistical test with the error bar chart

Procedure: Error Bar Chart with Chart Builder

The **Chart Builder** procedure is accessed from the **Graphs…Chart Builder** menu. With the **Chart Builder** dialog open:

1) Select the simple error bar chart icon  in the Bar Choose from: group and place it in the Chart preview area
2) Specify a variable for the X-axis.
3) Specify a variable for the Y-axis. This variable should be scale in measurement.

Figure 7.14 Chart Builder Dialog to Create Error Bar Chart



7.12 *Demonstration: Error Bar Chart with Chart Builder*

We will create an error bar chart corresponding to the **Independent-Samples T-Test** of gender and number of children. We want to see how number of children varies across categories of gender, so we use *childs* as the Y-axis variable. This is equivalent to how we used **the Independent-Samples T Test** procedure to study this relationship.

Detailed Steps for Error Bar Chart

Before beginning this example, to insure that *childs* is displayed properly in the error bar chart, do the following:

1) In the Data Editor, change the Measure level for ***childs*** to **Scale**
2) Change the value of Width to **3**; change the number of Decimals to **1**

Then, with the Chart Builder dialog open:

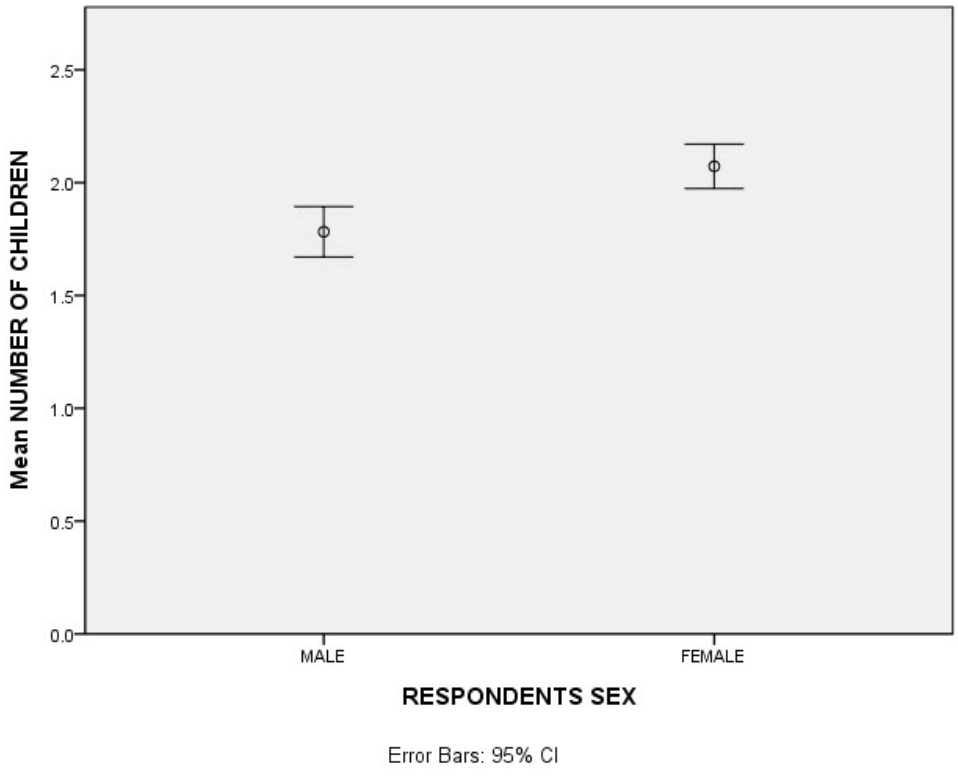
1) Select a simple error bar chart icon and put it in the Chart Preview pane.
2) Place ***childs*** in the Y-axis box.
3) Place the variable ***sex*** in the X-axis box.

Results from the Error Bar Chart

We can observe the following details from the graph.

- The mean number of children for each gender along with 95% confidence intervals is represented in this chart
- The confidence intervals for the two genders don't quite overlap, which is consistent with the result from the T Test
- The error bars have a small range compared to the range of *childs*, which indicates we are fairly precisely measuring number of children (because of large sample sizes)

Figure 7.15 Error Bar Chart of Education and Gender



7.13 *Lesson Summary*

We explored the use of the **Independent-Samples T-Test procedure** and error bar charts to test whether there are mean differences in a scale variable between two groups.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Perform a statistical test to determine whether there is a statistically significant difference between two groups on a scale dependent variable

To support the achievement of the primary objective, students should now also be able to:

- Check the assumptions of the **Independent-Samples T Test**
- Use the **Independent-Samples T Test** to test the difference in means
- Know how to interpret the results of a **Independent-Samples T Test**
- Use the **Chart Builder** to create an error bar graph to display mean differences

7.14 *Learning Activity*

In these activities you will use the file *Census.sav*. The overall goal is to run the **Independent-Samples T Test**, to interpret the output and visualize the results with an error bar chart.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. We want to see whether men and women differ in their mean socioeconomic index (*sei*) and their age when their first child was born (*agekdbrn*). First, use the Explore procedure to view the distributions of these two variables by gender. Are they similar or different? Do you see any problems with doing a t test?

2. Now do a t test for each variable, by gender. Is the homogeneity of variance assumption met, or not? What do you conclude about mean differences by gender?

3. Create an error bar chart for each variable by gender. Is the graph consistent with the result from the t test?

4. Now do the same analysis with the variable *race*, testing whether there are differences in *sei* and *agekdbrn* comparing whites to blacks. Although *race* has three categories, you can use only two categories in the t test. As before, first use the Explore procedure to view the distributions of these two variables by race? Are they similar or different? Do you see any problems with doing a t test?
5. Now do a t test for each variable, by white versus black. Is the homogeneity of variance assumption met, or not? What do you conclude about mean differences between whites and blacks?

6. Create an error bar chart for each variable by race. Is the graph consistent with the result from the t test?

7. *For those with more timed:* How could you display an error bar chart with only the categories of white and black, not other? There are at least two methods.

Lesson 8: The Paired-Samples T Test

8.1 *Objectives*

After completing this lesson students will be able to:

- Perform a statistical test to determine whether there is a statistically significant difference between the means of two scale variables

To support the achievement of this primary objective, students will also be able to:

- Use the **Paired-Samples T Test** procedure
- Interpret the results of a **Paired-Samples T Test**

8.2 *Introduction*

In this lesson we outline the logic involved when testing for a difference between two scale variables, and then perform an analysis comparing two variables. As with the Independent Samples T Test, we use the mean to compare the two variables, as the mean is an excellent measure of central tendency. As an example, we might want to compare a student´s test score before and after participating in a particular program and see if the program had an effect by calculating the mean difference between the two test scores.

Business Context

When analyzing data, we are concerned with whether there is a difference between two scale variables. Without statistical testing, we might make decisions based on perceptions that are not likely to exist in a population of customers. The **Paired-Samples T Test** allows us to determine if the difference between two variables in the sample reflects a difference in the population. For example:

- In medical research a **Paired-Samples T Test** would be used to compare means on a measure administered both before and after some type of treatment.
- In market research, if a subject were to rate the product they usually purchase and a competing product on some attribute, a **Paired-Samples T Test** would be needed to compare the mean ratings.
- In customer satisfaction studies, if a special customer care program is implemented, we can test whether satisfaction beforehand is lower, or higher, than satisfaction after the program is in place.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

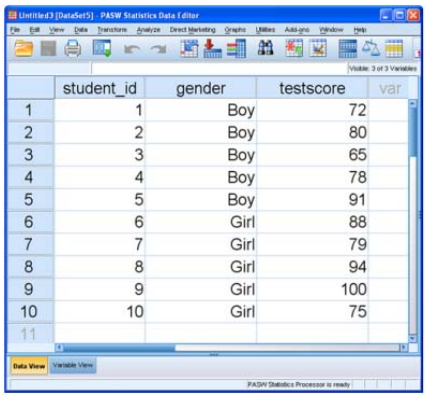
8.3 *The Paired-Samples T Test*

The **Paired-Samples T Test** is used to test for statistical significance between two population means when each observation (respondent) contributes to both means. With a **Paired-Samples T Test** each person serves as his own control. To the extent that an individual's outcomes across the two

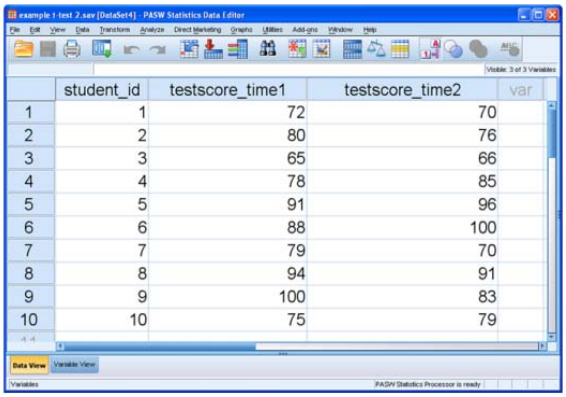
conditions are related, the **Paired-Samples T Test** provides a more powerful statistical analysis (greater probability of finding true effects) than the **Independent-Samples T Test**. Moreover, due to the assumption of independent for the independent-samples test, a paired-samples test must be used when the same subject/respondent provides both scores.

To clearly state the difference between the **Independent-Samples T Test** and the **Paired-Samples T Test**, it is instructive to compare the data structure needed for each of the two tests.

Figure 8.1 Data structure for Independent-Samples T Test (left) and Paired-Samples T Test (right)



	student_id	gender	testscore	var
1	1	Boy	72	
2	2	Boy	80	
3	3	Boy	65	
4	4	Boy	78	
5	5	Boy	91	
6	6	Girl	88	
7	7	Girl	79	
8	8	Girl	94	
9	9	Girl	100	
10	10	Girl	75	
11				



	student_id	testscore_time1	testscore_time2	var
1	1	72	70	
2	2	80	76	
3	3	65	66	
4	4	78	85	
5	5	91	96	
6	6	88	100	
7	7	79	70	
8	8	94	91	
9	9	100	83	
10	10	75	79	

8.4 *Assumptions for the Paired-Samples T Test*

There are three assumptions to apply the **Paired-Samples T Test**:

1) Variables have a scale measurement level.
2) Variables should be in the same unit of measurement.
3) The difference between the two variables is normally distributed (because the mean difference is what is being tested)

Note

The homogeneity of variance assumption that holds for the **Independent-Samples T Test** does not apply to the **Paired-Samples T Test** since we are dealing with only one group.

Note

The **Paired-Samples T Test** is robust to violations of the normality assumption when sample sizes are moderate to large (over 50 cases per group).

8.5 *Requesting a Paired-Samples T Test*

Requesting a **Paired-Samples T Test** is accomplished by following these steps:

1) Choose pairs of variables for the **Paired-Samples T Test**.
2) Review the procedure output to investigate the relationship between the variables including:
    a. Paired Samples Statistics Table
    b. Paired Samples Test Table.

8.6 *Paired-Samples T Test Output*

The Paired Samples Statistics table provides summary information including sample sizes, means, standard deviations, and standard errors for the pair of variables. The Paired Samples Correlations table displays the correlation between the pair of variables. The higher the correlations, the more statistical power there is to detect a mean difference.

Figure 8.2 Example of Paired Samples Statistics Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	RESPONDENT SOCIOECONOMIC INDEX	49.519	1491	19.5778	.5070
	R'S FATHER'S SOCIOECONOMIC INDEX	47.094	1491	18.8498	.4882

Figure 8.3 Example of Paired Samples Correlations Output

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	RESPONDENT SOCIOECONOMIC INDEX & R'S FATHER'S SOCIOECONOMIC INDEX	1491	.286	.000

The null hypothesis is that the two means are equal. The mean difference in socioeconomic index is 2.42, reported along with the sample standard deviation and standard error in the Paired Samples Test table. Although this is not a large difference, the significance value (.000) indicates that the two means are significantly different (at the .01 level). A 95% confidence interval for the mean difference is also reported, which can be quite useful.

Figure 8-4 Paired Samples Test Table

Paired Samples Test

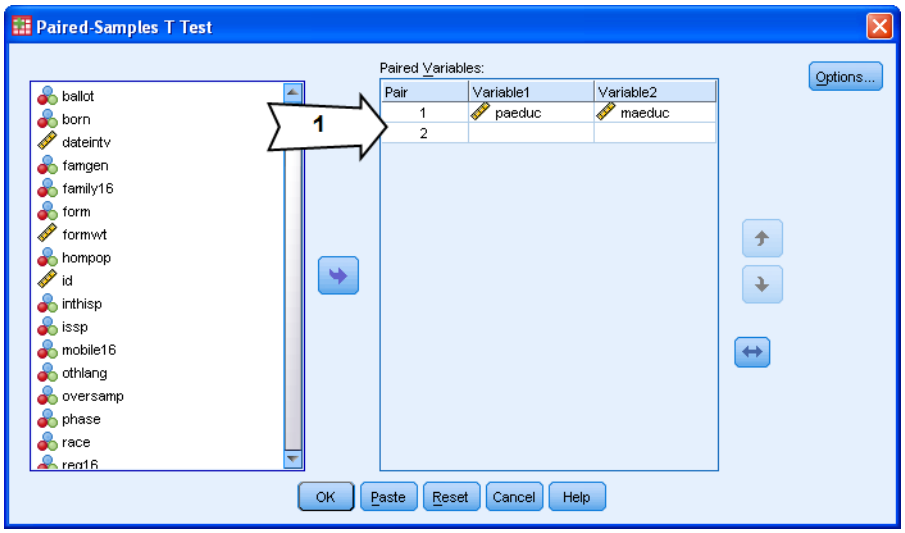
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	RESPONDENT SOCIOECONOMIC INDEX - R'S FATHER'S SOCIOECONOMIC INDEX	2.4243	22.9729	.5949	1.2573	3.5914	4.075	1490	.000

8.7 *Procedure: Paired-Samples T Test*

The **Paired-Samples T Test** procedure is accessed from the **Analyze…Compare Means…Paired-Samples T Test** menu choice. With the **Paired-Samples T Test** dialog box open:

1) Select the pair of variables to compare and place them in the Paired Variables box (more than one pair of variables can be in the Paired Variable(s) box). Hint: Use Ctrl+Click to select the pair.

Figure 8.5 Paired-Samples T Test Dialog



8.8 *Demonstration: Paired-Samples T Test*

We will work with the *Census.sav* data file in this lesson.

To demonstrate a **Paired-Samples T Test**, we will compare mean education levels of the respondent (*educ*) and his or her spouse (*speduc*). The **Paired-Samples T Test** is appropriate because we will obtain data from a single respondent regarding his/her education and that of the spouse's education. We are interested in testing whether there is a significant difference in education between spouses in the population. Although two people are involved—husband and wife—information is being obtained from only one person, so the paired-sample test is appropriate.

Detailed Steps for Paired-Samples T Test

1) Select the variables ***educ*** and ***speduc*** and place them in the Paired Variables box

Results from Paired-Samples T Test

The first table displays the mean, standard deviation and standard error for each of the variables. We see that the means for respondent and spouse's education are close, but the education for the spouse is a bit lower. This might indicate very close educational matching of people who marry.

Figure 8.6 Paired Samples Statistics Table

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	HIGHEST YEAR OF SCHOOL COMPLETED	13.76	963	3.022	.097
	HIGHEST YEAR SCHOOL COMPLETED, SPOUSE	13.57	963	3.019	.097

In the next table, the sample size (number of pairs) appears along with the correlation between the two variables. The correlation (.594) is positive, high, and statistically significant (differs from zero in the population). This suggests that the power to detect a difference between the two means is substantial.

Figure 8-7 Paired Samples Correlations Table

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	HIGHEST YEAR OF SCHOOL COMPLETED & HIGHEST YEAR SCHOOL COMPLETED, SPOUSE	963	.594	.000

The mean education difference, about .20 years, is reported along with the sample standard deviation of the difference. The significance of the difference is .024, so if we are using a criterion of .05, we would reject the null hypothesis and conclude the means are different (but we would reach a different conclusion if using an alpha value of .01).

Whether or not the means are substantively different is a separate question.

Figure 8.8 Paired Samples Test Table

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	HIGHEST YEAR OF SCHOOL COMPLETED - HIGHEST YEAR SCHOOL COMPLETED, SPOUSE	.198	2.722	.088	.026	.370	2.262	962	.024

Apply Your Knowledge

1. True or false? Before doing the Paired-Samples T Test, a test of equality of variances should be done?

2. See the output below. Which statements are correct?
   a. The sample mean difference is .6.
   b. The null hypothesis that the mean difference is 0 must be rejected (alpha=0.05)
   c. Normality of the distribution of mean differences is not a concern here.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	testscore_time1	82.20	10	10.809	3.418
	testscore_time2	81.60	10	11.539	3.649

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	testscore_time1 - testscore_time2	.600	8.369	2.647	-5.387	6.587	.227	9	.826

8.9 *Lesson Summary*

We explored the use of the **Paired-Samples T Test** to test the mean difference between two variables.

Lesson Objectives Review

After completing this lesson students will be able to:

- Perform a statistical test to determine whether there is a statistically significant difference between the means of two scale variables

To support the achievement of this primary objective, students will also be able to:

- Use the **Paired-Samples T Test** procedure
- Interpret the results of a **Paired-Samples T Test**

8.10 *Learning Activity*

The overall goal of this learning activity is to use the **Paired-Samples T Test**.

Supporting Materials

The SPSS customer satisfaction data file *SPSS_CUST.SAV*. This data file was collected from a random sample of SPSS customers asking about their satisfaction with the software, service, and other features, and some background information on the customer and their company.

1. One variable in the customer survey asked about agreement that SPSS products are a good value (*gdvalue*). A second question asked about agreement that SPSS offers high quality products (*hiqualty*). Use a paired-samples t test to see whether the means of these two questions differ (they are measured on a five-point scale). What do you conclude?

2. Then test whether there is a mean difference between agreement that SPSS products are easy to learn (*easylrn*) and SPSS products are easy to use (*easyuse*). What do you conclude?

3. Could we use a paired-sample t test to compare how long a customer has used SPSS products (*usespss*) and how frequently they use SPSS (*freqspss*)? Why or why not?

Lesson 9: One-Way ANOVA

9.1 *Objectives*

After completing this lesson students will be able to:

- Perform a statistical test to determine whether there is a statistically significant difference among three or more groups on a scale dependent variable

To support the achievement of this primary objective, students will also be able to:

- Use the options in the **One-Way ANOVA** procedure
- Check the assumptions for **One-Way ANOVA**
- Interpret the results of a **One-Way ANOVA** analysis
- Use the **Chart Builder** to create an error bar to graph mean differences

9.2 *Introduction*

Analysis of variance (ANOVA) is a general method of drawing conclusions regarding differences in population means when three or more comparison groups are involved. The **Independent-Samples T Test** applies only to the simplest instance (two groups), while the **One-Way ANOVA** procedure can accommodate more complex situations (three or more groups). In this lesson we will provide information on the assumptions of using the **One-Way ANOVA** procedure and then provide examples of its use.

Business Context

When analyzing data, we are often concerned with whether groups differ from each other. Without statistical testing, we might make decisions based on perceptions that are not likely to exist in a population of customers. **One-Way ANOVA** allows us to determine if three or more groups significantly differ on scale variables; thus we can determine which groups score higher or lower than the others. For example, we might want to know whether:

- Customer groups differ on attitude toward a product or service.
- Different drugs better reduce depression levels.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

9.3 *One-Way Anova*

The basic logic of significance testing for comparing group means on more than two groups is the same as that for comparing two group means (i.e., the **Independent-Samples T Test**). To summarize:

- The null hypothesis is that the population groups have the same means.

- We determine the probability of obtaining a sample with group mean differences as large (or larger) as what we find in our data. To make this assessment the amount of variation among group means (between-group variation) is compared to the amount of variation among observations within each group (within-group variation). Assuming in the population that the group means are identical (null hypothesis), the only source of variation among sample means would be the fact that the groups are composed of different individual observations.

Thus a ratio of the two sources of variation (between group/within group) should be about 1 when there are no population differences. When the distribution of individual observations within each group follows the normal curve, the statistical distribution of this ratio is known (F distribution) and we can make a probability statement about the consistency of our data with the null hypothesis. The final result is the probability of obtaining sample differences as large (or larger) as what we found if there were no population differences. If this probability is sufficiently small (usually less than 5 chances in 100, or .05) we conclude the population groups differ.

Once we find a difference, we have to determine which groups differ from each other. (When the null hypothesis is rejected, it does not follow that all group means differ significantly; the only thing that can be said is that not all group means are the same.)

9.4 *Assumptions of One-Way ANOVA*

To correctly use the **One-Way ANOVA** procedure requires an understanding of additional issues.

1) The dependent variable must have a scale measurement level.
2) The independent variable (named a “factor” in Anova analyses) must have a categorical measurement level.
3) The distribution of the dependent variable within each population subgroup follows the normal distribution (normality). **One-Way ANOVA** is robust to moderate violations when sample sizes are moderate to large (over 25 cases) and the dependent measure has the same distribution (for example, skewed to the right) within each comparison group.
4) The variation is the same within each population subgroup (homogeneity of variance). **One-Way ANOVA** is robust to moderate violations when sample sizes of the groups are similar.

Similar to the **Independent-Samples T Test**, violation of the assumption of homogeneity of variances is more serious than violation of the assumption of normality. And like the **Independent-Samples T Test**, **One-Way ANOVA** applies a two-step strategy for testing:

1) Test the homogeneity of variance assumption.
2) If the assumption holds, proceed with the standard test (the ANOVA F Test) to test equality of means; if the null hypothesis of equal variances is rejected, use an adjusted F test to test equality of means.

9.5 *Requesting One-Way ANOVA*

Running a **One-Way ANOVA** is accomplished by following these steps:

1) Select the scale variable on which to test equality of group means
2) Select a factor variable.
3) Request the Levene homogeneity of variance test
4) Review the procedure output to:
    a. Check on the test for homogeneity of variances
    b. Review the test on equality of means, either using the standard ANOVA F test or an adjusted F test taking into account unequal variances
5) If the null hypothesis of equal population means is rejected, extend the analysis by adding a post hoc test within the **One-Way ANOVA** procedure.

9.6 *One-Way ANOVA Output*

The first table of output is the test for homogeneity of variance. The null hypothesis is that the variances are equal, so if the significance level is low enough (as it is in the table below), we reject the null hypothesis and conclude the variances are not equal.

As with the independent-samples t test, this isn't a problem, but it does mean that we should use the tests that adjust for unequal variance.

Figure 9.1 Levene Test of Homgeneity of Variances

Test of Homogeneity of Variances

HOURS PER DAY WATCHING TV

Levene Statistic	df1	df2	Sig.
9.155	4	900	.000

Tip

You should also examine the actual standard deviations or variances for each group. In large samples, it is relatively easy to reject the null hypothesis of equal variances even when the variances are within a factor of 2 of each other.

Most of the information in the ANOVA table is technical in nature and is not directly interpreted. Rather the summaries are used to obtain the F statistic and, more importantly, the probability value we use in evaluating the population differences. The standard ANOVA table will provide the following information:

- The first column has a row for the between-groups and a row for within-groups variation.
- S*ums of squares* are intermediate summary numbers used in calculating the between- (deviations of individual group means around the total sample mean) and within- (deviations of individual observations around their respective sample group mean) group variances.
- The "*df" column* contains information about *degrees of freedom*, related to the number of groups and the number of individual observations within each group.
- *Mean Squares* are measures of the between-group and within-group variation (Sum of Squares divided by their respective degrees of freedom).
- *The F statistic* is the ratio of between to within group variation and will be about 1 if the null hypothesis is true.
- The column labeled *"Sig."* provides the *probability of obtaining the sample F ratio* (taking into account the number of groups and sample size), if the null hypothesis is true.

In practice, most researchers move directly to the significance value since the columns containing the sums of squares, degrees of freedom, mean squares and *F* statistic are all necessary for the probability calculation but are rarely interpreted in their own right. In the table below, the low significance value leads us to reject the null hypothesis of equal means.

Figure 9.2 ANOVA Table Output

ANOVA

HOURS PER DAY WATCHING TV

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	339.006	4	84.752	16.223	.000
Within Groups	4701.745	900	5.224		
Total	5040.751	904			

When the condition of equal variances is not met, an adjusted F test has to be used. PASW Statistics provides two such tests, Welch and Brown-Forsythe. The table Robust Tests of Equality of Means provides the details. Again, the columns containing test statistic and degrees of freedom are technical details to compute the significance.

Figure 9.3 Robust Tests of Equality of Means Output

Robust Tests of Equality of Means

HOURS PER DAY WATCHING TV

	Statistic[a]	df1	df2	Sig.
Welch	20.614	4	210.120	.000
Brown-Forsythe	17.440	4	350.170	.000

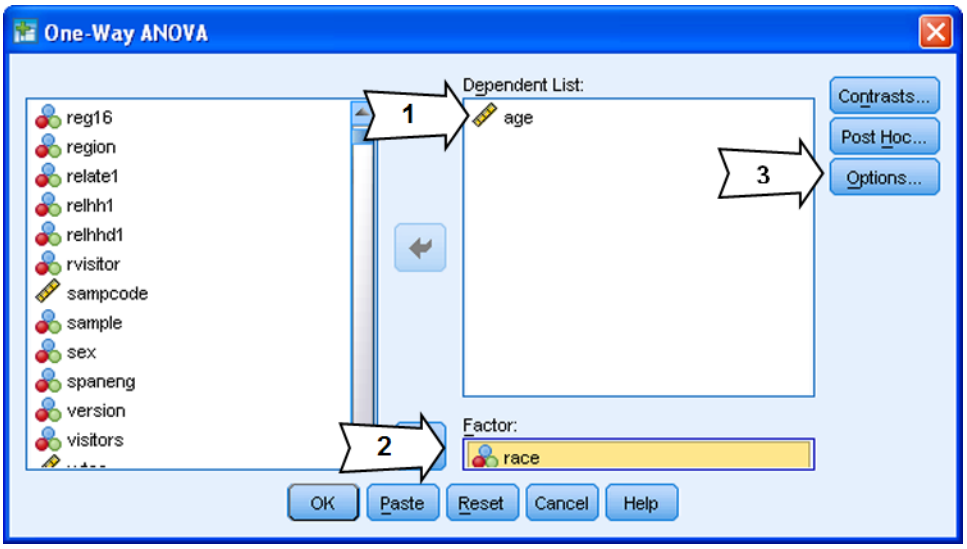
a. Asymptotically F distributed.

9.7 *Procedure: One-Way ANOVA*

The **One-Way ANOVA** procedure is accessed from the **Analyze…Compare Means…One-Way ANOVA** menu choice. With the **One-Way ANOVA** dialog box open:

1) Place one or more scale variables in the Dependent List box.
2) Place one categorical variable in the Factor box.
3) Open the Options dialog to request descriptive information and the test for homogeneity of variance.

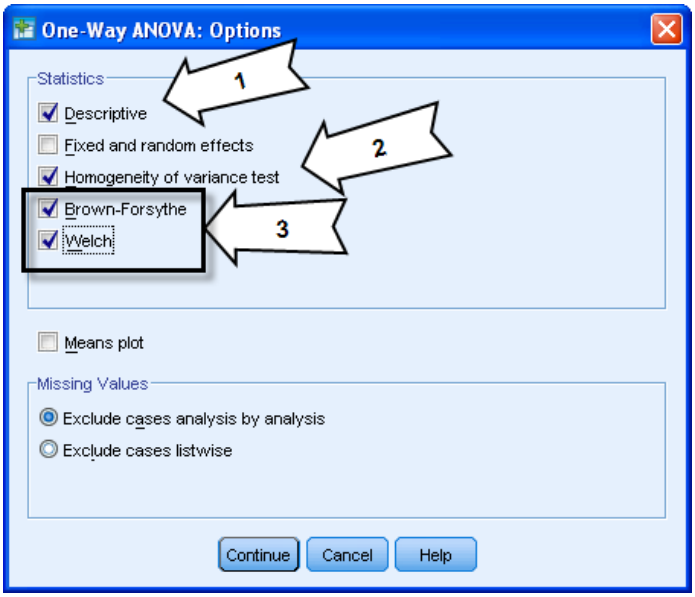
Figure 9.4 One-Way ANOVA Dialog



In the Options dialog:

1) Ask for Descriptive statistics so that group means and standard deviations are displayed.
2) The homogeneity of variance test allows one to assess the assumption of homogeneity of variance.
3) Brown-Forsythe and Welch are robust tests that do not assume homogeneity of variance and thus can be used when this assumption is not met.

Figure 9.5 One-Way ANOVA Options Dialog



9.8 *Demonstration: One-Way ANOVA*

We will work with the *Census.sav* data file in this lesson.

In this example we investigate the relationship between marital status (*marital*) and education in years (*educ*). We would like to determine whether there are educational differences among marital status groups.

Detailed Steps for One-Way ANOVA

1) Place the variable ***educ*** in the Dependent List box.
2) Place the variable ***marital*** in the Factor box.
3) In the Options dialog, select **Descriptive**, **Homogeneity of variance test**, **Brown-Forsythe**, and **Welch** check boxes.

Results from One-Way ANOVA

As with all analyses, you should first look to see how many cases are in each group, along with the means and standard deviations. The first table in the Viewer window provides this information. The size of the groups ranges from 70 to 971 people. The means vary from 11.76 to 13.73 years (the **One-Way ANOVA** procedure will assess if these means differ), while the standard deviations vary from 2.89 to 3.45 (the test of homogeneity of variance will assess if these standard deviations differ). We observe that:

- The married group has the most education (13.73) while those separated have the least education (11.76).

Figure 9.6 Table of Descriptive Statistics

Descriptives

HIGHEST YEAR OF SCHOOL COMPLETED

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MARRIED	971	13.73	3.055	.098	13.53	13.92	0	20
WIDOWED	162	12.46	3.446	.271	11.93	13.00	0	20
DIVORCED	281	13.37	2.887	.172	13.03	13.71	0	20
SEPARATED	70	11.76	3.178	.380	11.00	12.51	4	20
NEVER MARRIED	529	13.42	2.968	.129	13.17	13.68	0	20
Total	2013	13.43	3.079	.069	13.29	13.56	0	20

What we don't know is whether these differences are due to sampling variation, or instead are likely to be real and exist in the population. For this we turn to the ANOVA.

Levene Test of Homogeneity of Variance

First, we must review Levene's test of homogeneity of variance. The null hypothesis of homogeneity of within-group variance is not rejected (significance .694). This means we can use the standard ANOVA table.

Figure 9.7 Test of Homogeneity of Variances Table

Test of Homogeneity of Variances

HIGHEST YEAR OF SCHOOL COMPLETED

Levene Statistic	df1	df2	Sig.
.558	4	2008	.694

ANOVA Table

Every statistical test has a *null hypothesis*. In most cases, the null hypothesis is that there is no difference between groups. This is also true for the null hypothesis for the **One-Way ANOVA** procedure, so we test with **One-Way ANOVA** whether there is no difference in mean education among marital status groups. If the significance is small enough, we reject the null hypothesis and conclude that there are differences.

Figure 9.8 ANOVA Table for Education by Marital Status

ANOVA

HIGHEST YEAR OF SCHOOL COMPLETED

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	434.220	4	108.555	11.691	.000
Within Groups	18644.222	2008	9.285		
Total	19078.442	2012			

We see that the probability of the null hypothesis being correct is extremely small, less than .05, therefore we reject the null hypothesis and conclude that there are differences in education among these groups.

If we had not met the homogeneity of variances assumption, and given our disparate sample sizes, we would have to turn to the Brown-Forsythe and Welch tests, which test for equality of group means without assuming homogeneity of variance. These tests are shown below, although we would not report them in this situation (where equal variances may be assumed).

Figure 9.9 Robust Tests of Mean Differences

Robust Tests of Equality of Means

HIGHEST YEAR OF SCHOOL COMPLETED

	Statistic[a]	df1	df2	Sig.
Welch	10.295	4	349.664	.000
Brown-Forsythe	11.087	4	637.579	.000

a. Asymptotically F distributed.

Both of these measures mathematically attempt to adjust for the lack of homogeneity of variance.

- When calculating the between-group to within-group variance ratio, the *Brown-Forsythe test* explicitly adjusts for heterogeneity of variance by adjusting each group's contribution to the between-group variation by a weight related to its within-group variation.
- The *Welch test* adjusts the denominator of the *F* ratio so it has the same expectation as the numerator, when the null hypothesis is true, despite the heterogeneity of within-group variance.

Both tests indicate there are highly significant differences in average highest year of school completed between the marital status groups, which are consistent with the conclusions we drew from the standard ANOVA.

Having concluded that there are differences in amount of education among different marital status groups, we probe to find specifically which groups differ from which others.

Apply Your Knowledge

1. True or false? Suppose we have collected data for four groups of respondents on region (north/east/south/west) and their income categories (very low/low/moderate/high/very high). Is One-Way ANOVA the correct procedure to test whether there are differences between the regions with respect to income categories?

2. True or false? When the F-test is not significant, then we will not follow up this analysis with an analysis on which group means differ?

3. In a dataset about employees and their salaries we tested whether there are differences in mean salary according to the job category of the employee. The output is depicted below. Which statements are correct?
    a. Although the sample standard deviations are different, the null hypothesis of equal population variances cannot be rejected (alpha=0.05)
    b. The table titled ANOVA must be discarded, as the null hypothesis of equal population variances is rejected (alpha=0.05)
    c. The null hypothesis of equal population group means is rejected by both Welch and Brown-Forsythe tests (alpha=0.05)

Descriptives

Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Clerical	357	27755.71	7567.228	400.500	26968.07	28543.36	15750	80000
Custodial	26	30975.00	2147.987	421.255	30107.41	31842.59	24300	35250
Manager	84	63977.80	18244.776	1990.668	60018.44	67937.16	34410	135000
Total	467	34450.27	17188.710	795.399	32887.26	36013.28	15750	135000

Test of Homogeneity of Variances

Current Salary

Levene Statistic	df1	df2	Sig.
58.799	2	464	.000

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	89551201781.125	2	44775600890.563	431.668	.000
Within Groups	48129315110.417	464	103726972.221		
Total	137680516891.542	466			

Robust Tests of Equality of Means

Current Salary

	Statistic[a]	df1	df2	Sig.
Welch	163.082	2	112.666	.000
Brown-Forsythe	307.903	2	94.072	.000

a. Asymptotically F distributed.

9.9 *Post Hoc Tests with a One-Way ANOVA*

Post hoc tests are typically performed only after the overall *F* test indicates that population differences exist, although for a broader view see Milliken and Johnson (2004). At this point there is usually interest in discovering just which group means differ from which others. In one aspect, the procedure is quite straightforward: every possible pair of group means is tested for population differences and a summary table produced. However, a problem exists in that as more tests are performed, the probability of obtaining at least one false-positive result increases. As an extreme example, if there are ten groups, then 45 pairwise group comparisons (n*(n-1)/2) can be made. If we are testing at the .05 level, we would expect to obtain on average about 2 (.05 * 45) false-positive tests. In an attempt to reduce the false-positive rate when multiple tests of this type are done, statisticians have developed a number of methods.

Often, more than one post hoc test is used and the results are compared to provide for more evidence about potential mean differences.

9.10 *Requesting Post Hoc Tests with a One-Way ANOVA*

If the null hypothesis of equal population group means is rejected, a post hoc analysis is required, following these steps:

1) Ask for one or more appropriate post hoc tests in the Post Hoc dialog.
2) Inspect the output and report which groups differ significantly in their population mean.

9.11 *Post Hoc Tests Output*

The table labeled Multiple Comparisons provides all pairwise comparisons.

The rows are formed by every possible combination of groups. The column labeled *"Mean Difference (I-J)"* contains the sample mean difference between each pairing of groups. If this difference is statistically significant at the specified level after applying the post hoc adjustments, then an asterisk (*) appears beside the mean difference. Notice the actual significance value for the test appears in the column labeled *"Sig."*. In addition, the standard errors and 95% confidence intervals for each mean difference appear. These provide information on the precision with which we have estimated the mean differences. Note that, as you would expect, if a mean difference is not significant, the confidence interval includes 0.

Also notice that each pairwise comparison appears twice. For each such duplicate pair the significance value is the same, but the signs are reversed for the mean difference and confidence interval values.

Figure 9.10 Multiple Comparisons Output

Multiple Comparisons

HOURS PER DAY WATCHING TV
Bonferroni

(I) MARITAL STATUS	(J) MARITAL STATUS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MARRIED	WIDOWED	-1.879*	.292	.000	-2.70	-1.06
	DIVORCED	-.639*	.213	.027	-1.24	-.04
	SEPARATED	-1.221*	.417	.034	-2.39	-.05
	NEVER MARRIED	-.497*	.175	.046	-.99	.00
WIDOWED	MARRIED	1.879*	.292	.000	1.06	2.70
	DIVORCED	1.240*	.330	.002	.31	2.17
	SEPARATED	.658	.487	1.000	-.71	2.03
	NEVER MARRIED	1.383*	.307	.000	.52	2.25
DIVORCED	MARRIED	.639*	.213	.027	.04	1.24
	WIDOWED	-1.240*	.330	.002	-2.17	-.31
	SEPARATED	-.582	.444	1.000	-1.83	.67
	NEVER MARRIED	.142	.234	1.000	-.51	.80
SEPARATED	MARRIED	1.221*	.417	.034	.05	2.39
	WIDOWED	-.658	.487	1.000	-2.03	.71
	DIVORCED	.582	.444	1.000	-.67	1.83
	NEVER MARRIED	.725	.428	.905	-.48	1.93
NEVER MARRIED	MARRIED	.497*	.175	.046	.00	.99
	WIDOWED	-1.383*	.307	.000	-2.25	-.52
	DIVORCED	-.142	.234	1.000	-.80	.51
	SEPARATED	-.725	.428	.905	-1.93	.48

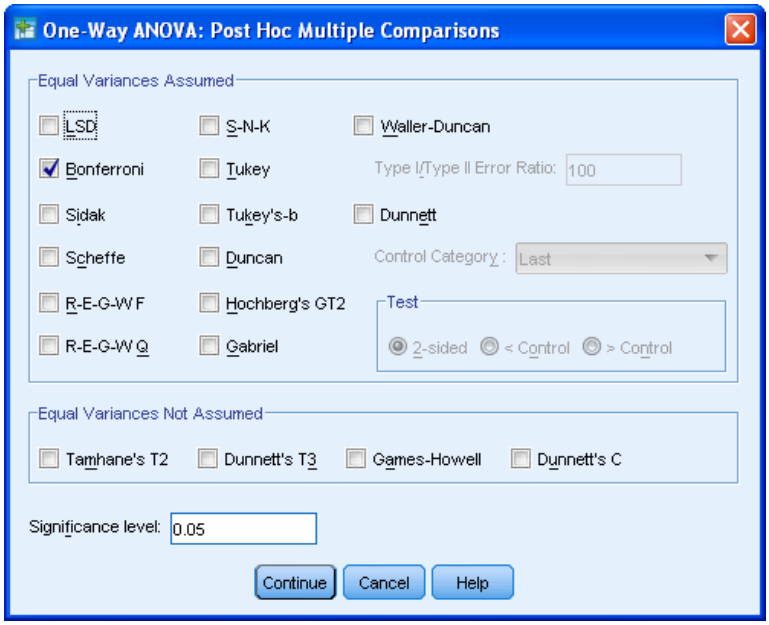
*. The mean difference is significant at the 0.05 level.

9.12 *Procedure: Post Hoc Tests with a One-Way ANOVA*

Post hoc analyses are accessed from the **One-Way ANOVA** dialog box. With the One-Way ANOVA dialog box open:

1) Open the Post Hoc Multiple Comparisons dialog
2) Select the appropriate method of multiple comparisons, which will depend on whether the assumption of homogeneity of variance has been met.

Figure 9.11 Post Hoc Testing Dialog



Why So Many Tests?

The Post Hoc dialog lists over a dozen tests just in the Equal Variances Assumed area. We need to review why so many tests are available.

The ideal post hoc test would demonstrate tight control of Type I (false-positive) error, have good statistical power (probability of detecting true population differences), and be robust over assumption violations (failure of homogeneity of variance, non-normal error distributions). Unfortunately, there are implicit tradeoffs involving some of these desired features (Type I error and power) and no current post hoc procedure is best in all these areas. Add to this the fact that pairwise tests can be based on different statistical distributions (t, F, studentized range, and others) and that Type I error can be controlled at different levels (per individual test, per family of tests, variations in between), and you have a large collection of post hoc tests.

We will briefly compare post hoc tests from the perspective of being liberal or conservative regarding control of the false-positive rate (Type 1 error). The existence of numerous post hoc tests suggests that there is no single approach that statisticians agree will be optimal in all situations.

LSD

The LSD or least significant difference method simply applies standard *t* tests to all possible pairs of group means. No adjustment is made based on the number of tests performed. The argument is that since an overall difference in group means has already been established at the selected criterion level (say .05), no additional control is necessary. This is the most liberal of the post hoc tests.

SNK, REGWF, REGWQ & Duncan

The SNK (Student-Newman-Keuls), REGWF (Ryan-Einot-Gabriel-Welsh F), REGWQ (Ryan-Einot-Gabriel-Welsh Q, based on the studentized range statistic) and Duncan methods involve sequential testing. After ordering the group means from lowest to highest, the two most extreme means are tested for a significant difference using a critical value adjusted for the fact that these are the extremes from a larger set of means. If these means are found not to be significantly different, the testing stops; if they are different then the testing continues with the next most extreme set, and so on. All are more conservative than the LSD. REGWF and REGWQ improve on the traditionally used SNK in that they adjust for the slightly elevated false-positive rate (Type I error) that SNK has when the set of means tested is much smaller than the full set.

Bonferroni & Sidak

The Bonferroni (also called the Dunn procedure) and Sidak (also called Dunn-Sidak) perform each test at a stringent significance level to insure that the family-wise (applying to the set of tests) false-positive rate does not exceed the specified value. They are based on inequalities relating the probability of a false-positive result on each individual test to the probability of one or more false positives for a set of independent tests. For example, the Bonferroni is based on an additive inequality, so the criterion level for each pairwise test is obtained by dividing the original criterion level (say .05) by the number of pairwise comparisons made. Thus with five means, and therefore ten pairwise comparisons, each Bonferroni test will be performed at the .05/10 or .005 level.

Tukey (b)

The Tukey (b) test is a compromise test, combining the Tukey (see next test) and the SNK criterion producing a test result that falls between the two.

Tukey

Tukey's HSD (Honestly Significant Difference; also called Tukey HSD, WSD, or Tukey(a) test) controls the false-positive rate family-wise. This means if you are testing at the .05 level, that when performing all pairwise comparisons, the probability of obtaining one or more false positives is .05. It is more conservative than the Duncan and SNK. If all pairwise comparisons are of interest, which is usually the case, Tukey's test is more powerful than the Bonferroni and Sidak.

Scheffe

Scheffe's method also controls the family-wise error rate. It adjusts not only for the pairwise comparisons, but also for any possible comparison the researcher might ask. As such it is the most conservative of the available methods (false-positive rate is least), but has less statistical power.

Specialized Post Hoc Tests

Hochberg's GT2 & Gabriel: Unequal Ns

Most post hoc procedures mentioned above (excepting LSD, Bonferroni & Sidak) were derived assuming equal group sample sizes in addition to homogeneity of variance and normality of error. When the subgroup sizes are unequal, PASW Statistics substitutes a single value (the harmonic mean) for the sample size. Hochberg's GT2 and Gabriel's post hoc test explicitly allow for unequal sample sizes.

Waller-Duncan

The Waller-Duncan takes an approach (Bayesian) that adjusts the criterion value based on the size of the overall *F* statistic in order to be sensitive to the types of group differences associated with the *F* (for example, large or small). Also, you can specify the ratio of Type I (false positive) to Type II (false negative) error in the test. This feature allows for adjustments if there are differential costs to the two types of errors.

Unequal Variances and Unequal Ns

Tamhane T2, Dunnett's T3, Games-Howell, Dunnett's C

Each of these post hoc tests adjusts for unequal variances and sample sizes in the groups. Simulation studies (summarized in Toothaker, 1991) suggest that although Games-Howell can be too liberal when the group variances are equal and sample sizes are unequal, it is more powerful than the others.

The bottom line is that your choice in post hoc tests should reflect your preference for the power/false-positive tradeoff and your evaluation of how well the data meet the assumptions of the analysis, and you live with the results of that choice.

9.13 *Demonstration: Post Hoc Tests with a One-Way ANOVA*

We will work with the *Census.sav* data file in this example. Previous analysis showed that the null hypothesis of equal mean education for the marital status groups was rejected. Post hoc analysis will reveal which groups differ.

Detailed Steps for a Post Hoc Test

1) Place the variable ***educ*** in the Dependent List box.
2) Place the variable ***marital*** in the Factor box.
3) In the Post hoc dialog, select **Bonferroni**.

Results for the Post Hoc Tests

We will move directly to the post hoc test results.

Figure 9.12 Bonferroni Post Hoc Results

Multiple Comparisons

HIGHEST YEAR OF SCHOOL COMPLETED
Bonferroni

(I) MARITAL STATUS	(J) MARITAL STATUS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MARRIED	WIDOWED	1.264*	.259	.000	.54	1.99
	DIVORCED	.361	.206	.808	-.22	.94
	SEPARATED	1.970*	.377	.000	.91	3.03
	NEVER MARRIED	.304	.165	.653	-.16	.77
WIDOWED	MARRIED	-1.264*	.259	.000	-1.99	-.54
	DIVORCED	-.904*	.301	.027	-1.75	-.06
	SEPARATED	.706	.436	1.000	-.52	1.93
	NEVER MARRIED	-.960*	.274	.005	-1.73	-.19
DIVORCED	MARRIED	-.361	.206	.808	-.94	.22
	WIDOWED	.904*	.301	.027	.06	1.75
	SEPARATED	1.609*	.407	.001	.47	2.75
	NEVER MARRIED	-.057	.225	1.000	-.69	.58
SEPARATED	MARRIED	-1.970*	.377	.000	-3.03	-.91
	WIDOWED	-.706	.436	1.000	-1.93	.52
	DIVORCED	-1.609*	.407	.001	-2.75	-.47
	NEVER MARRIED	-1.666*	.388	.000	-2.76	-.58
NEVER MARRIED	MARRIED	-.304	.165	.653	-.77	.16
	WIDOWED	.960*	.274	.005	.19	1.73
	DIVORCED	.057	.225	1.000	-.58	.69
	SEPARATED	1.666*	.388	.000	.58	2.76

*. The mean difference is significant at the 0.05 level.

We see the Married and Widowed groups have a mean difference of 1.26 years of education. This difference is statistically significant at the specified level after applying the post hoc adjustments. The interval [.54, 1.99] contains the difference between these two population means with 95% confidence.

Summarizing the entire table, we would say that the Married, Divorced, and Never Married groups differs in amount of education from the Separated and Widowed groups. We could not be sure of this just from examining the means.

Apply Your Knowledge

1. In a dataset about employees and their salaries we tested whether there are differences in mean salary according to the political preference of the employee. The output is depicted below. Which statements are correct?
    1) The null hypothesis of equal variances is not rejected (alpha=0.05)
    2) The null hypothesis of equal group means is rejected (alpha=0.05)
    3) Post-hoc tests show significant differences between all three groups A, B, and C (alpha=.05).

Test of Homogeneity of Variances

Current Salary

Levene Statistic	df1	df2	Sig.
2.087	2	464	.125

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1918706817.261	2	959353408.631	3.279	.039
Within Groups	135761810074.280	464	292590107.919		
Total	137680516891.542	466			

Post Hoc Tests

Multiple Comparisons

Current Salary
Bonferroni

(I) Politica...	(J) Political...	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	B	672.022	1654.785	1.000	-3303.92	4647.96
	C	-6910.534	2969.896	.061	-14046.28	225.21
B	A	-672.022	1654.785	1.000	-4647.96	3303.92
	C	-7582.555*	2987.081	.034	-14759.59	-405.52
C	A	6910.534	2969.896	.061	-225.21	14046.28
	B	7582.555*	2987.081	.034	405.52	14759.59

*. The mean difference is significant at the 0.05 level.

9.14 *Error Bar Chart with Chart Builder*

For presentations it is often useful to show a graph of the relationship between a scale and categorical variable. Error bar charts are the most effective method of doing this for ANOVAs.

The **Chart Builder** allows for the creation of a variety of charts, including error bar charts. It provides for great flexibility in creating charts, including formatting options.

9.15 *Requesting an Error Bar Chart with Chart Builder*

Follow these steps to produce an error bar using the **Chart Builder**:

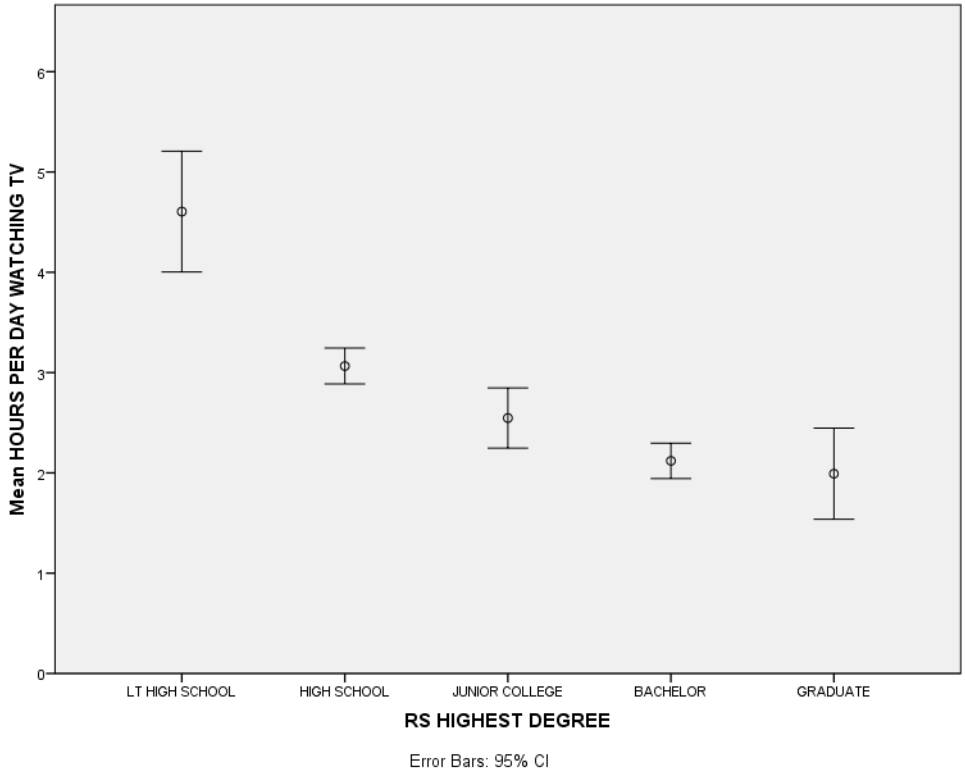
1) Select an error bar chart type of graph.
2) Specify the scale variable for which means and confidence intervals are requested.
3) Specify the categorical variable that defines the groups.
4) Inspect the error bar in the output to see which groups do (not) overlap in their confidence intervals.

9.16 *Error Bar Chart Output*

The standard error bar chart will generate a graph depicting the relationship between a scale and categorical variable. It provides a visual sense of how far the groups are separated.

- Note the means of each group (the small circle)
- Note if the 95% confidence intervals of the groups overlap

Figure 9.13 Error Bar Chart of TV Hours by Highest Degree

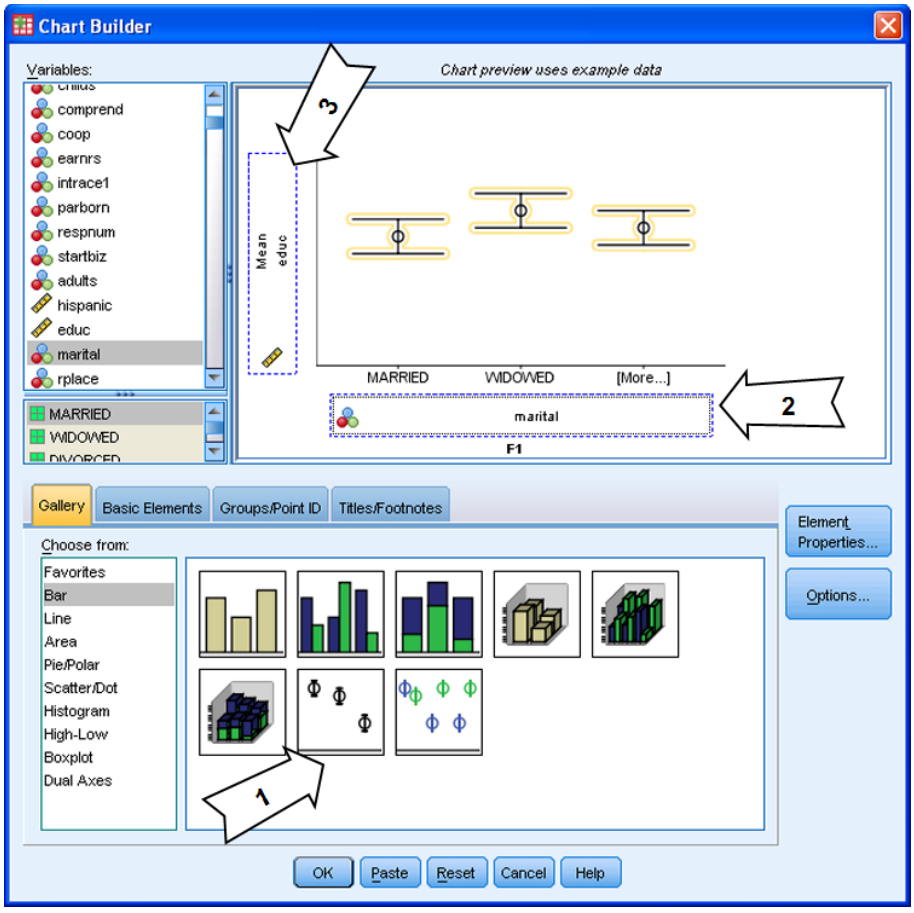


9.17 *Procedure: Error Bar Chart with Chart Builder*

The **Chart Builder** procedure is accessed from the **Graphs…Chart Builder** menu. With the **Chart Builder** dialog open:

1) Select an error bar chart icon and place it in the Chart preview area.
2) Specify a categorical variable for the X-axis.
3) Specify a scale variable for the Y-axis. The mean is calculated by default.

Figure 9.14 Chart Builder Dialog to Create Error Bar Chart



9.18 *Demonstration: Error Bar Chart with Chart Builder*

We will create an error bar chart for the ANOVA of marital status and education. We want to see visually how education varies across categories of marital status, so we use education as the Y-axis variable. This is equivalent to how we used ANOVA to study this relationship.

Detailed Steps for Error Bar Chart

1) Select an **error bar chart icon** and put it in the Chart Preview pane.
2) Place the variable ***educ*** in the Y-axis box.
3) Place the variable ***marital*** in the X-axis box.

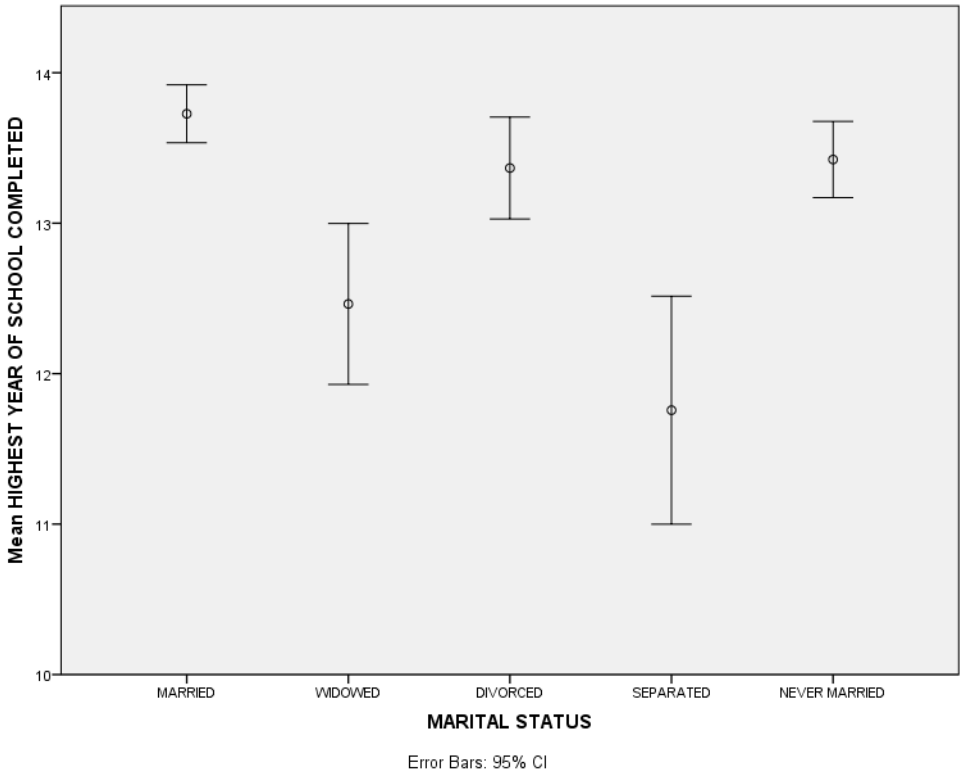
Results from the Error Bar Chart Created with Chart Builder

An ANOVA table with means based from education within categories of marital status lets us compare groups, and this is mirrored in the error bar chart.

- The mean education of each marital status group along with 95% confidence intervals is represented in this chart
- Confidence intervals that do not overlap indicate that those groups differ from each other
- Confidence intervals that do overlap indicate that those groups do not differ from each other
- We can readily compare the categories of marital status in this arrangement

Married, divorced and never married respondents have more education than the separated and widowed categories.

Figure 9.15 Error Bar Chart of Education and Marital Status



Further Info

The confidence intervals on the error bar chart are determined for each group separately and no adjustment is made based on the number of groups that are being compared, or for unequal variance. So an error bar chart should never be used without doing statistical tests.

Additional Resources

Further Info

For additional information on multiple comparison tests with ANOVA, see:

Klockars, Alan J. and Sax, G. 1986. *Multiple Comparisons*. Newbury Park, CA: Sage.

9.19 *Lesson Summary*

We explored the use of the **One-Way ANOVA** procedure to analyze relationships between a categorical and a scale variable.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Perform a statistical test to determine whether there is a statistically significant difference among three or more groups on a scale dependent variable

Lesson Objectives Review

To support the achievement of the primary objective, students should now also be able to:

- Use the options in the **One-Way ANOVA** procedure
- Check the assumptions for **One-Way ANOVA**
- Interpret the results of a **One-Way ANOVA** analysis
- Use the **Chart Builder** to create an error bar to graph mean differences

9.20 *Learning Activity*

The overall goal of this learning activity is to use One-Way ANOVA with post hoc tests to explore the relationship between several variables. You will use the PASW Statistics data file *Census.sav*.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. Investigate how the number of siblings (*sibs*) varies by highest degree (*degree*). Ask for appropriate statistics.

2. Is the assumption of homogeneity of variance met? Is the ANOVA test significant at the .01 level?

3. Do a post hoc analysis, if justified. Ask for both the Bonferroni and Scheffe tests? What do you conclude from these tests? Which education groups have different mean numbers of children? Are the Bonferroni and Scheffe tests consistent?

4. Create an error bar chart to display the mean differences for *sibs* by *degree*. Is the error bar chart a correct representation of which means are different?

5. Now do another analysis of political position (*polviews*) by *degree*. Repeat the same steps from the analysis above. Which education groups differ in their political position?

Lesson 10: Bivariate Plots and Correlations for Scale Variables

10.1 *Objectives*

After completing this lesson students will be able to:

- Perform a statistical test to determine whether two scale variables are correlated (related)

To support the achievement of the primary objective, students will also be able to:

- Visually assess the relationship between two scale variables with scatterplots, using the **Chart Builder** procedure
- Explain the options of the **Bivariate Correlations** procedure
- Explain the Pearson correlation coefficient and its assumptions
- Interpret a Pearson correlation coefficient

10.2 *Introduction*

In this lesson we examine and quantify the relationship between two scale variables. A scatterplot visually presents the relationship between two scale variables, while the Pearson correlation coefficient is used to quantify the strength and direction of the relationship between scale variables. The Pearson correlation coefficient (formally named the Pearson product-moment correlation coefficient) is a measure of the extent to which there is a linear (or straight line) relationship between two variables.

Business Context

When we examine the distributions of two scale variables, we would like to know whether a relationship we observe is likely to exist in our target population or instead is caused by random sampling variation. Statistical testing tells us whether two scale variables are related. Assessing correlation coefficients allows us to determine the direction and strength of this relationship. For example, we might want to know whether:

- Higher SAT scores are associated with higher first year college GPAs
- Eating more often at fast-food restaurants was related to more frequent shopping at convenience stores
- Lower levels of depression are associated with higher self-esteem scores

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

10.3 *Scatterplots*

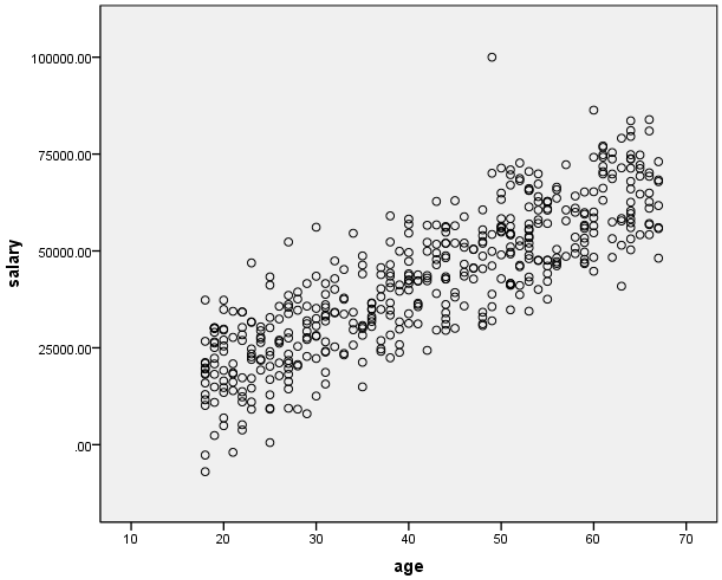
The scatterplot visually presents the relationship between two scale variables. A scatterplot displays individual observations in an area determined by a vertical and a horizontal axis, each of which

represent the variables of interest (note that the variables must be scale). In a scatterplot, look for a relationship between the two variables and note any patterns or extreme points.

Scatterplot illustrated

For an illustration, see the scatterplot below, showing the relationship between variables *age* and *salary*. The relationship is linear (a straight line describes the relationship between *age* and *salary*), and positive (if *age* increases, then so does *salary*); there is one person having a salary which is "out of line."

Figure 10.1 Scatterplot of Age and Salary



Best Practice

Place the independent variable on the x-axis and the dependent variable on the y-axis. Here, *salary* depends on *age* and not vice versa, so *salary* is the dependent variable on the y-axis, *age* the independent variable on the x-axis.
If the situation is such that no dependent or independent variable can be identified (say number of hours watching tv and number of hours of internet use), then the choice of which variable goes where is arbitrary.

10.4 *Requesting a Scatterplot*

The scatterplot is available in the **Chart Builder** procedure. The steps to create a scatterplot are:

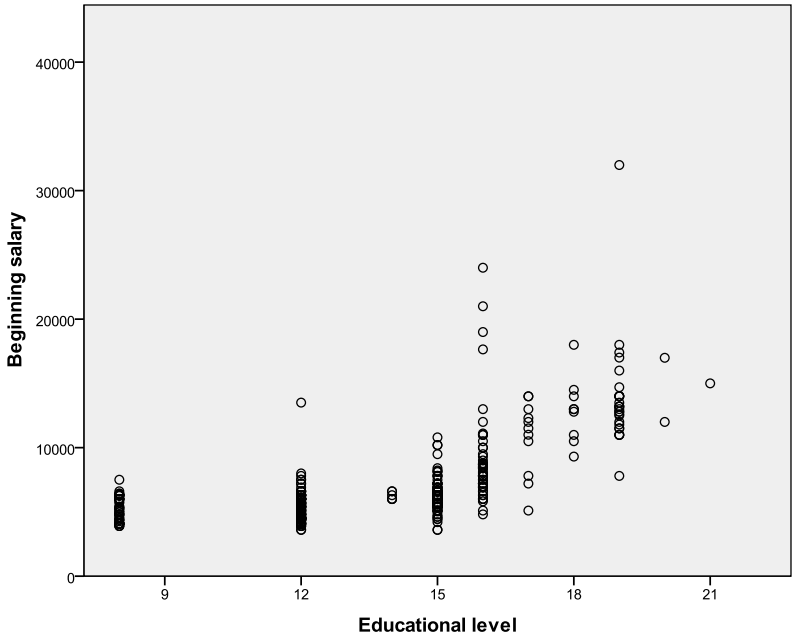
1) Select a chart (simple scatter) from the Gallery.
2) Place the dependent variable on the vertical (y) axis and the independent variable on the horizontal (x).
3) Review the procedure output to investigate the relationship between the variables including:
    a. Linearity
    b. Directionality
    c. Outliers.

10.5 *Scatterplot Output*

The standard scatterplot will generate a graph depicting the relationship between the two variables.

- Note if there is a linear relationship
- Note the direction of the relationship
- Note any outliers

Figure 10.2 Scatterplot Showing the Relationship between Beginning Salary and Education Level

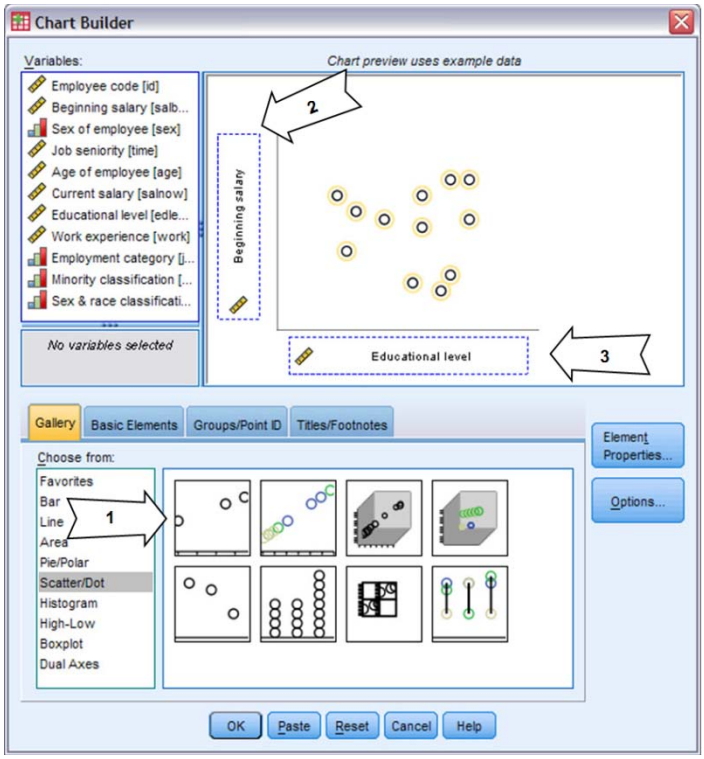


10.6 *Procedure: Scatterplot*

The scatterplot is created from the **Graphs…Chart Builder** menu choice. With the **Chart Builder** dialog box open:

1) Select a Simple Scatter graph on the canvas and drag and drop it in the Chart Preview area.
2) Place the dependent variable in the vertical axis.
3) Place the independent variable in the horizontal axis.

Figure 10.3 Chart Builder Dialog Box to Create a Scatterplot



10.7 *Demonstration: Scatterplot*

We will work with the *Census.sav* data file in this example.
In this demonstration we examine the relationship between mother's education (*maeduc*) and father's education (*paeduc*). This will allow us to investigate whether people marry those with a similar amount of education (we could also investigate this question by examining the scatterplot of the respondent' education and his/her spouse).

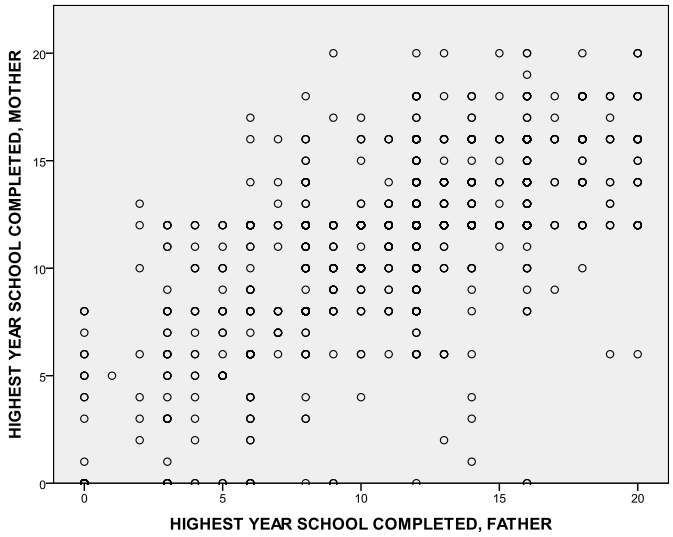
Detailed Steps for a Scatterplot

1) Place the **Simple Scatter icon** in the Chart Preview area
2) Place the variable ***maeduc*** in the Y-axis box.
3) Place the variable ***paeduc*** in the X-axis box.

Results from the Scatterplot

The scatterplot visually presents the relationship between two variables by displaying individual observations. Observe that for the most part couples with low education tend to marry each. Also, couples with high education tend to marry each other, thus there is a positive, linear relationship.

Figure 10.4 Scatterplot of Mother's and Father's Highest Year of School Completed



In preparation for our discussion of the correlation coefficient, we will edit the scatterplot to superimpose a best fit straight line to the data.

10.8 *Adding a Best Fit Straight Line to the Scatterplot*

All charts in PASW Statistics can be edited. The type of editing that can be done depends on the type of chart. For a scatterplot, one option is to add a best fit line, which can be done in one step.

Detailed Steps to Edit Scatterplot

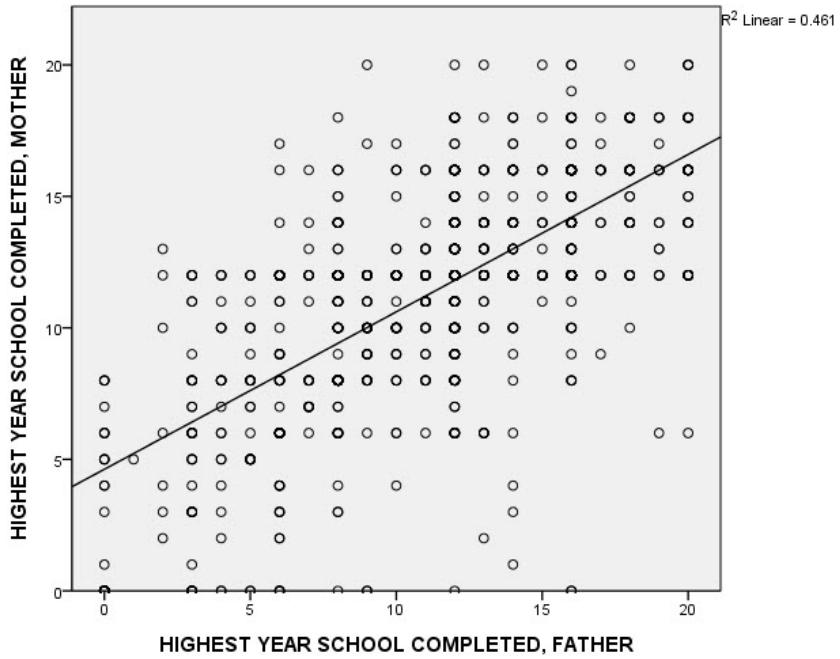
We need to open the scatterplot in the Chart Editor.

1) Double-click on the chart to open it in the Chart Editor
2) Select **Elements…Fit Line at Total**
3) Close the Chart Editor

The straight line tracks the positive relationship between father and mother's educational attainment. How well do you think it describes or models the relationship? Does it match what you would have drawn by hand?

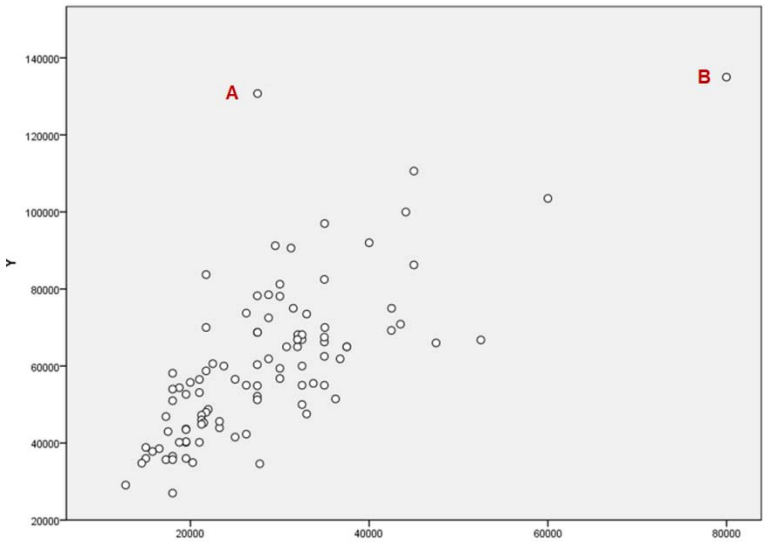
We use scatterplots to get a sense of whether or not it is appropriate to use a correlation coefficient to describe this relationship with one number. As we will learn, the correlation coefficient assumes a linear relationship.

Figure 10.5 Scatterplot with Fit Line Added



Apply Your Knowledge

1. Consider the scatterplot between *X* and *Y* depicted below. Which statements are correct?
    a. There is a linear relationship between *X* and *Y*
    b. There is a positive relationship between *X* and Y
    c. The point labeled A is far from the straight line describing the relationship between *X* and *Y*
    d. The point labeled B is far from the straight line describing the relationship between *X* and *Y*

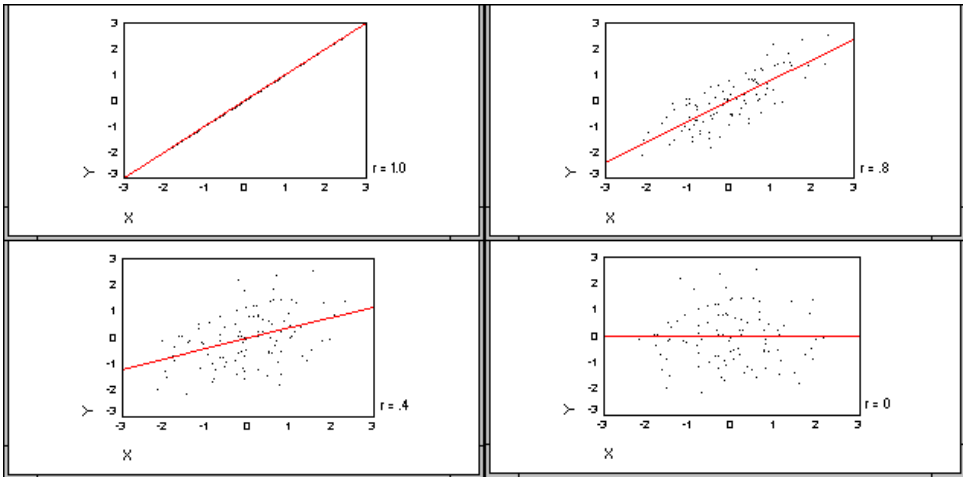


10.9 *Pearson Correlation Coefficient*

The Pearson Correlation Coefficient is a measure of the extent to which there is a linear (or straight line) relationship between two scale variables. It is normed so that a correlation of +1 indicates that the data fall on a perfect straight line sloping upwards (positive relationship), while a correlation of –1 would represent data forming a straight line sloping downwards (negative relationship). A correlation of 0 indicates there is no straight-line relationship at all.

Below are four scatterplots with their accompanying correlations, all based on simulated data following normal distributions. Four different correlations appear (here, and in general, the letter "r" is used to denote the Pearson Correlation Coefficient). All are positive, but represent the full range in strength of linear association (from 0 to 1). As an aid in interpretation, a best-fitting straight line is superimposed on each chart.

Figure 10.6 Scatterplots Based on Various Correlations



For the perfect correlation of 1.0, all points fall on the straight line trending upwards. In the scatterplot with a correlation of .8 in the upper right, the strong positive relation is apparent, but there is some variation around the line. Looking at the plot of data with correlation of .4 in the lower left, the association is clearly less pronounced than with the data correlating .8 (note greater scatter of points around the line). The final chart displays a correlation of 0: there is no linear association present and the best-fitting straight line is a horizontal line.

We use statistical tests to determine whether a relationship between two or more variables is statistically significant. That is, we want to test whether the correlation differs from zero (zero indicates no linear association) in the population, based on the sample results. In other words, the null hypothesis is the correlation coefficient is 0 in the population, and we use a statistical test to assess this hypothesis.

Pearson Correlation Coefficient Assumptions

To correctly use the Pearson correlation coefficient and apply statistical tests, three conditions must be met:

1) Variables must have a scale measurement level.
2) Variables must be linearly related.
3) Variables must be normally distributed.

10.10 *Requesting a Pearson Correlation Coefficient*

The Pearson Correlation Coefficient is available in the **Bivariate Correlations** procedure. Requesting a Pearson Correlation Coefficient is accomplished with these steps:

1) Choose variables for the correlation (We simply list the variables to be analyzed; there is no designation of dependent and independent variables. Correlations will be calculated on all pairs of variables listed.).
2) Optionally select the correlation statistic to calculate. Pearson is used for scale variables, while Spearman and Kendall's tau-b (less common) are used for non-normal data or ordinal data, as relationships are evaluated after the original data have been transformed into ranks.
3) Optionally, select a one- or two-tailed significance test to perform.
4) The Flag significant correlations check box is checked by default. When checked, asterisks appearing beside the correlations will identify significant correlations.
5) Examine the output to see which correlations are significant.

10.11 *Bivariate Correlation Output*

The standard correlation table will provide the following information:

- The Pearson Correlation, which will range from +1 to -1, the further away from 0, the stronger the relationship
- The 2-tailed significance level, the test of the null hypothesis that the correlation is 0 in the population; all correlations with a significance level less than .05 will be considered statistically significant and will have an asterisk next to the coefficient
- N, which is the sample size
- The correlations in the major diagonal will always be 1, because these are the correlations of each variable with itself
- The correlation matrix is symmetric, so that the same information is represented above and below the major diagonal

Figure 10.7 Example of Bivariate Correlations Output

Correlations

		Beginning salary	Age of employee	Educational level	Work experience
Beginning salary	Pearson Correlation	1	-.011	.633**	.045
	Sig. (2-tailed)		.811	.000	.327
	N	474	474	474	474
Age of employee	Pearson Correlation	-.011	1	-.281**	.804**
	Sig. (2-tailed)	.811		.000	.000
	N	474	474	474	474
Educational level	Pearson Correlation	.633**	-.281**	1	-.252**
	Sig. (2-tailed)	.000	.000		.000
	N	474	474	474	474
Work experience	Pearson Correlation	.045	.804**	-.252**	1
	Sig. (2-tailed)	.327	.000	.000	
	N	474	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

Further Info

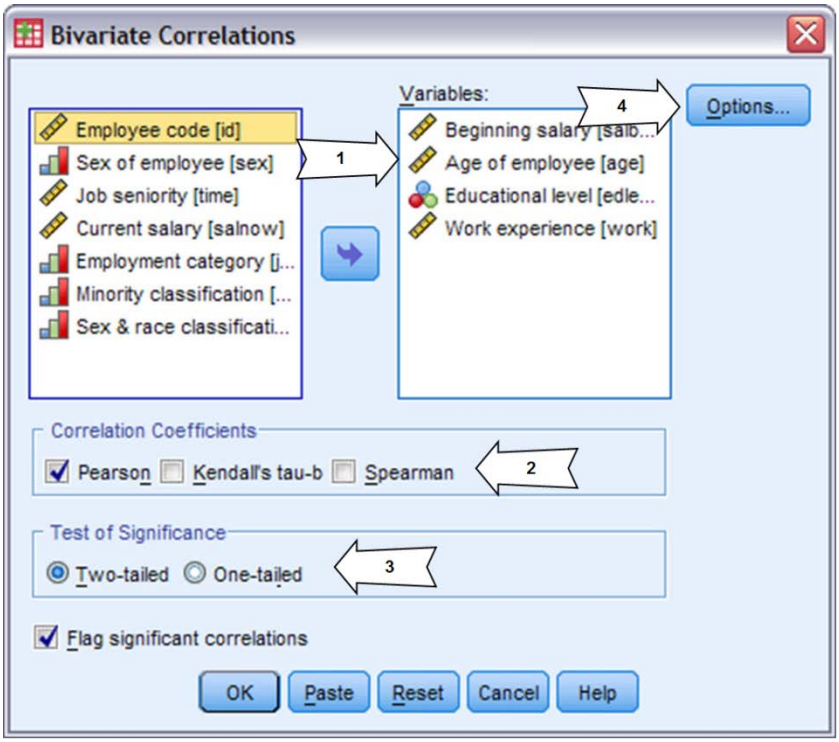
Dichotomous variables with only two categories can be used to calculate correlation coefficients. A scatterplot won't be useful to study the relationship between a dichotomous variable and a scale variable, but the correlation coefficient still provides information on the strength and direction of the relationship.

10.12 *Procedure: Pearson Correlation with Bivariate Correlations*

The **Bivariate Correlations** procedure is accessed from the **Analyze…Correlate…Bivariate** menu choice. With the **Bivariate Correlations** dialog box open:

1) Place two or more variables in the Variables box.
2) Optionally, select the correlation coefficient to calculate.
3) Optionally, select the type of significance test to perform.
4) Open the Options dialog box to display descriptive information and determine how to handle missing values.

Figure 10.8 Bivariate Correlations Dialog Box



10.13 *Demonstration: Pearson Correlation with Bivariate Correlations*

We will work with the *Census.sav* data file in this lesson.

In this demonstration we examine the relationship between mother's education (*maeduc*), father's education (*paeduc*), and the education of the respondent (*educ*). We would like to determine whether, for example, people marry people with a similar amount of education and if the children also have a similar amount of education as their parents.

Detailed Steps for Bivariate Correlations

1) Place ***maeduc***, ***paeduc*** and ***educ*** in the Variables box.

Results from Bivariate Correlations

In the Correlations table we see that the three correlations are moderate to high, and positive. Couples with high education tended to marry each other, thus we have a positive, linear relationship (r = .68). There is a similar level of association between the respondent's education and that of his father (r=.48) and that of his mother (r=.44).

Figure 10.9 Correlations Table of Education Variables

Correlations

		HIGHEST YEAR SCHOOL COMPLETED, MOTHER	HIGHEST YEAR SCHOOL COMPLETED, FATHER	HIGHEST YEAR OF SCHOOL COMPLETED
HIGHEST YEAR SCHOOL COMPLETED, MOTHER	Pearson Correlation	1	.679**	.445**
	Sig. (2-tailed)		.000	.000
	N	1780	1397	1777
HIGHEST YEAR SCHOOL COMPLETED, FATHER	Pearson Correlation	.679**	1	.481**
	Sig. (2-tailed)	.000		.000
	N	1397	1487	1485
HIGHEST YEAR OF SCHOOL COMPLETED	Pearson Correlation	.445**	.481**	1
	Sig. (2-tailed)	.000	.000	
	N	1777	1485	2018

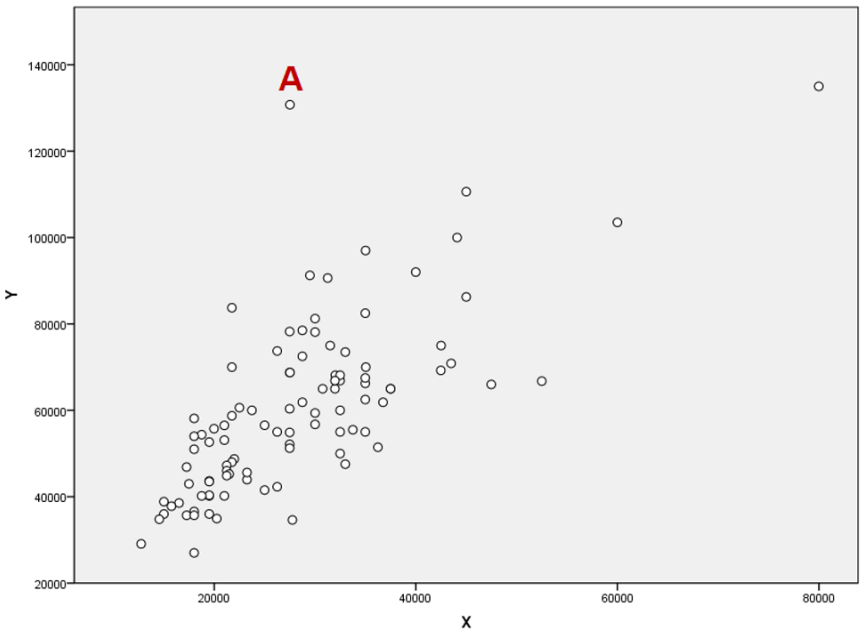
**. Correlation is significant at the 0.01 level (2-tailed).

What we don't know is whether these relationships are due to sampling variation, or instead are likely to be real and exist in the population of adults. For this we turn to the significance level of the Pearson Correlation Coefficient.

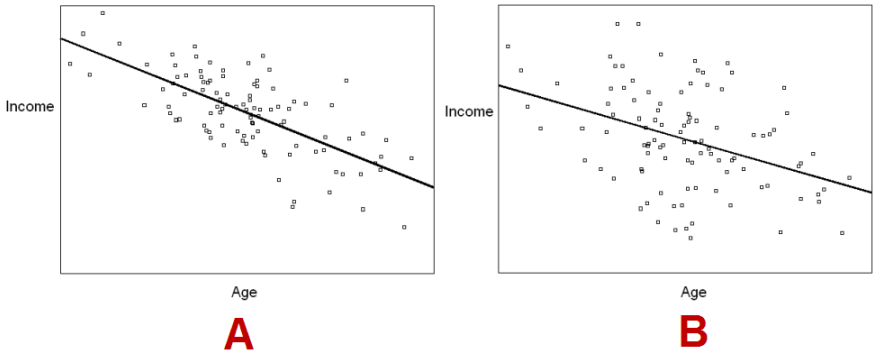
We see that the probability of the null hypothesis being true for all of these relationships is extremely small, less than .01; therefore we reject the null hypothesis and conclude that there is a positive, linear relationship between these variables.

Apply Your Knowledge

1. True or false? If we remove point A from the data, the correlation between X and Y will be lower?



2. True or false? In the two scatterplots below, the correlation between age and income is higher in A than in B?



3. What is the range of a Pearson correlation coefficient?
    a. From 0 to 1
    b. Can take on any positive or negative value
    c. From –1 to 1
    d. Depends on the standard deviation of the variables

10.14 *Lesson Summary*

We explored the use of scatterplots and correlations to examine relationships between scale variables.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Perform a statistical test to determine whether two scale variables are correlated (related)

To support the achievement of the primary objective, students should now also be able to:

- Visually assess the relationship between two scale variables with scatterplots, using the **Chart Builder** procedure
- Explain the options of the **Bivariate Correlations** procedure
- Explain the Pearson correlation coefficient and its assumptions
- Interpret a Pearson correlation coefficient

10.15 *Learning Activity*

The overall goal of this learning activity is to visualize the relationship between two scale variables creating scatterplots and to quantify this relationship with the correlation coefficient. In this set of learning activities you will use the data file *Bank.sav*.

Supporting Materials

The file *Bank.sav*, a PASW Statistics data file that contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Suppose you are interested in understanding how an employees demographic characteristics, beginning salary, and time at the bank and in the work force are related to current salary. Start by producing scatterplots of *salbeg*, *sex*, *time*, *age*, *edlevel*, and *work* with *salnow*. Add a fit line to each plot. Check on the variable labels for *time* and *work* so you understand what these variables are measuring.
2. Describe the relationships based on the scatterplots. Do they all appear to be linear? Are any relationships negative? What is the strongest relationship?

3. Now produce correlations with all these variables. Which correlations with *salnow* are significant? What is the largest correlation in absolute value with *salnow*? Did this match what you thought based on the scatterplots?

4. Examine the correlations between the other variables? Which variables are most strongly related? Create scatterplots for these as well to check for linearity.

5. *For those with more time:* Go back and review the scatterplots with *salnow*. Are there any employees who are outliers—far from the fit line—in any of the scatterplots? How might they be affecting the relationship?

Lesson 11: Regression Analysis

11.1 *Objectives*

After completing this lesson students will be able to:

- Perform linear regression to determine whether one or more variables can significantly predict or explain a dependent variable

To support the achievement of the primary objective, students will also be able to:

- Explain linear regression and its assumptions
- Explain the options of the **Linear Regression** procedure
- Interpret the results of the **Linear Regression** procedure

11.2 *Introduction*

Correlations allow one to determine if two scale variables are linearly related to each other. So for example, beginning salary and education might be positively related for employees. Regression allows one to further quantify this relation by developing an equation predicting starting salary based on education. Linear regression is a statistical method used to predict a variable (a scale dependent measure) from one or more independent (scale) variables. Commonly, straight lines are used, although other forms of regression allow nonlinear functions. In this lesson we will focus on linear regression.

Business Context

When we examine the relationships between scale variables, we would like to know whether a relationship we observe is likely to exist in our target population or instead is caused by random sampling variation. Statistical testing tells us whether scale variables are related. The results of linear regression allow us to determine if one or more scale variables predict an outcome variable and the impact each independent variable has on this variable. For example, we might want to know whether:

- Higher SAT scores explain higher first year college GPAs
- Eating more often at fast-food restaurants predicts more frequent shopping at convenience stores
- Increasing income leads to more customer purchases

Supporting Materials

The file *Bank.sav*, a PASW Statistics data file that contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

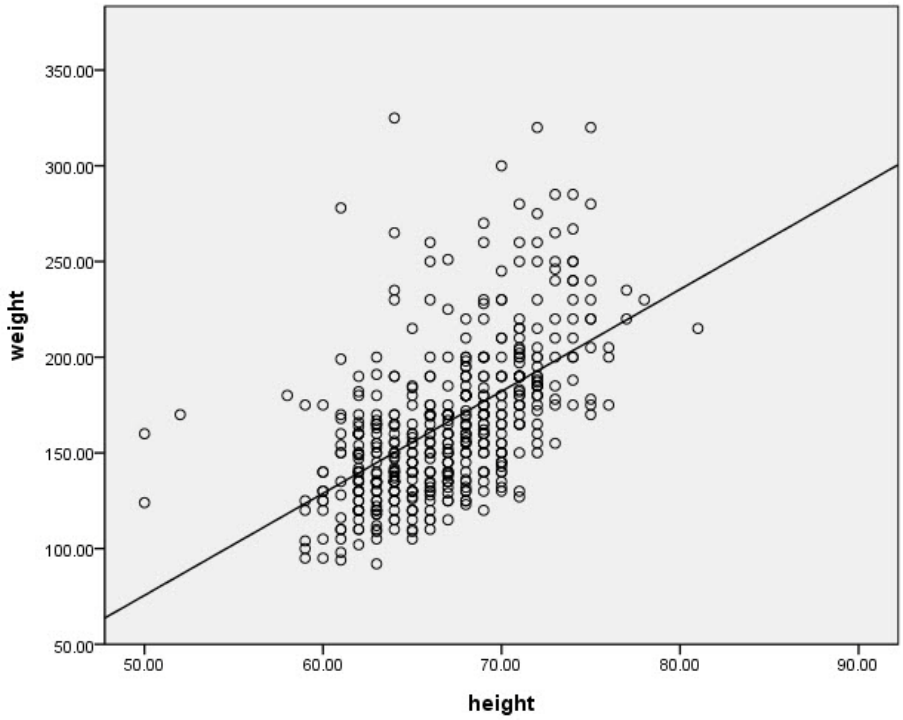
11.3 *Simple Linear Regression*

Linear regression involving a single independent (scale) variable is the simplest case and is called simple linear regression. In other words, one scale variable, say *X*, is used to predict another scale variable, say *Y*.

Simple Regression Illustrated

When there is a single independent variable, the relationship between the independent variable and dependent variable can be visualized in a scatterplot, and the concept of linear regression can be explained using the scatterplot.

Figure 11.1 Scatterplot of Height and Weight



The line superimposed on the scatterplot is the best straight line that describes the relationship. The line is represented in general form by the equation,

$$Y = a + b*X$$

where, *b* is the slope (the change in Y per unit change in X) and *a* is the intercept (the value of Y when X is zero). (Here, *Y* is *weight in pounds* and *X* is *height in inches*).

The value of the equation is linked to how well it actually describes or fits the data, and so part of the regression output includes fit measures. To quantify the extent to which the straight line fits the data, the fit measure, R Square, was developed. R^2 has the dual advantages of falling on a standardized scale and having a practical interpretation. The R Square measure (which is simply the correlation squared, or r^2, when there is a single predictor variable, and thus its name) is on a scale from 0 (no linear association) to 1 (perfect linear prediction). Also, the R Square value can be interpreted as the proportion of variation in one variable that can be predicted from the other. Thus an R Square of .50 indicates that we can account for 50% of the variance in one variable if we know values of the other. You can think of this value as a measure of the improvement in your ability to predict one variable from the other (or others if there are multiple independent variables).

We use statistical tests to determine whether a relationship between the independent variable and dependent variable is statistically significant. That is, we want to test whether the predictor can significantly explain the dependent variable.

- Linear Regression is used to determine whether the independent variable can significantly explain the dependent variable, in other words, test the null hypothesis that R Square is zero

- Once we find a relationship, we have to assess the effect of the independent variable on the dependent variable

Finally, referring to the scatterplot, we see that many points fall near the line, but some are quite a distance from it. For each point, the difference between the value of the dependent variable and the value predicted by the equation (value on the line) is called the residual (also known as the error). Points above the line have positive residuals (they were under-predicted), those below the line have negative residuals (they were over-predicted), and a point falling on the line has a residual of zero (perfect prediction). Points having relatively large residuals are of interest because they represent instances where the prediction line did poorly. Outliers, or points far from the mass of the others, are of interest in regression because they can exert considerable influence on the equation (especially if the sample size is small).

Note

While not covered in this course, PASW Statistics can provide influence statistics to aid in judging whether the equation was strongly affected by a particular observation.

11.4 *Simple Linear Regression Assumptions*

To correctly use simple linear regression and apply statistical tests, four conditions must be met:

1) Variables must have a scale measurement level.
2) Variables must be linearly related.
3) Residuals must be normally distributed.
4) Residuals are assumed to be independent of the predicted values, implying that the variation of the residuals around the line is homogeneous.

Note

A variable coded as a dichotomy (say 0 and 1) can technically be considered as a scale variable. A scale variable assumes that a one-unit change has the same meaning throughout the range of the scale. If a variable's only possible codes are 0 and 1 (or 1 and 2, etc.), then a one-unit change does mean the same change throughout the scale. Thus dichotomous variables, for example sex, can be used as predictor variables in regression. It also permits the use of nominal predictor variables if they are converted into a series of dichotomous variables; this technique is called dummy coding and is considered in most regression texts (Draper and Smith, 1998; Cohen and Cohen, 2002).

11.5 *Requesting Simple Linear Regression*

Requesting a linear regression involves these steps:

1) Select a dependent variable and an independent variable
2) Review the procedure output to investigate the relationship between the variables including:
   a. R^2
   b. Adjusted R^2.
3) Examine the regression test statistics to determine whether the observed relationship is statistically significant.
4) Determine the impact of the independent variable on the dependent variable.

11.6 *Simple Linear Regression Output*

The standard linear regression will generate three tables depicting the relationship between the two variables.

- The Model Summary table provides several measures of how well the model fits the data
- *R*, the multiple correlation coefficient, which can range from 0 to 1, is a generalization of the correlation coefficient. It is the correlation between the dependent measure and the combination of the independent variable(s), thus the closer the multiple *R* is to 1, the better the fit. If there is only one predictor, the multiple correlation coefficient is equivalent to the Pearson correlation coefficient.
- *R Square*, which can range from 0 to 1, is the correlation coefficient squared. It can be interpreted as the proportion of variance of the dependent measure that can be predicted from the independent variable(s).
- *Adjusted R Square* represents a technical improvement over R Square in that it explicitly adjusts for the number of predictor variables relative to the sample size. If *Adjusted R Square* and *R Square* differ dramatically, it is a sign that you have used too many predictor variables for your sample size.
- *Standard Error of the Estimate* is a standard deviation type summary of the dependent variable that measures the deviation of observations around the best fitting straight line. As such it provides, in the scale of the dependent variable, an estimate of how much variation remains to be accounted for after the line is fit.

Figure 11.2 Model Summary and ANOVA Tables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.563[a]	.317	.316	32.97827

a. Predictors: (Constant), height

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	247270.217	1	247270.217	227.361	.000[a]
	Residual	531819.966	489	1087.566		
	Total	779090.183	490			

a. Predictors: (Constant), height

b. Dependent Variable: weight

While the fit measures indicate how well we can expect to predict the dependent variable, they do not tell whether there is a statistically significant relationship between the dependent and independent variable(s). The analysis of variance table (**ANOVA** in the Output Viewer) presents technical summaries (sums of squares and mean square statistics) of the variation accounted for by the prediction equation. Our main interest is in determining whether there is a statistically significant (non-zero) linear relation between the dependent variable and the independent variable(s) in the population.

- The Sig. column provides the probability that the null hypothesis is true (i.e., no relationship between the independent and dependent variable).

We use the significance, as in any other hypothesis test, to determine whether or not to reject the null hypothesis. If significant results are found, we turn to the next table, **Coefficients**, to view the regression coefficients.

Figure 11.3 Regression Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-190.923	23.689		-8.059	.000
	height	5.330	.353	.563	15.078	.000

a. Dependent Variable: weight

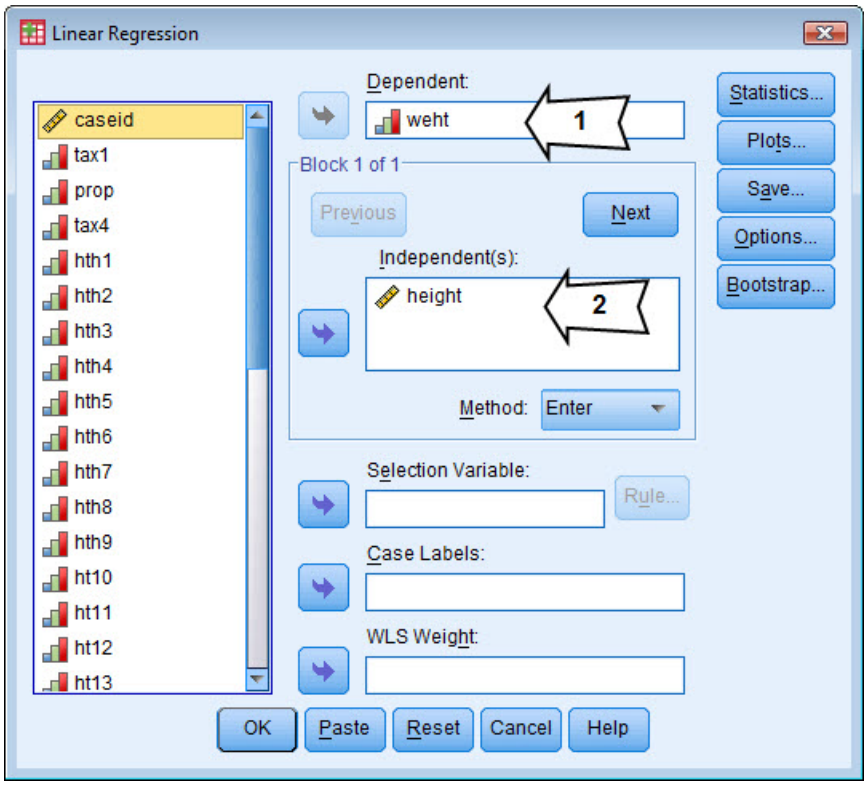
- The first column contains a list of the independent variables plus the intercept *(Constant)*. The intercept is the value of the dependent variable when the independent variable is 0.
- The column labeled *B* contains the estimated regression coefficients we would use in a prediction equation. The coefficient for *height* indicates that on average, each additional inch in height was associated with an increase in weight of 5.33 pounds.
- The *Standard Error (of B)* column contains standard errors of the regression coefficients. The standard errors can be used to create a 95% confidence interval around the *B* coefficients.
- *Betas* are standardized regression coefficients and are used to judge the relative importance of each of several independent variables.
- The *t statistics* provide a significance test for each *B* coefficient, testing whether it differs from zero in the population.

11.7 *Procedure: Simple Linear Regression*

Linear Regression is available in the **Regression…Linear** menu choice. With the **Linear Regression** dialog box open:

1) Place the dependent variable in the Dependent: box.
2) Place the independent variable in the Independent(s): box.

Figure 11.4 Linear Regression Dialog Box



In the **Linear Regression** dialog:

1) The Independent(s) list box will permit more than one independent variable, and so this dialog box can be used for both simple and multiple regression.
2) The block controls permit an analyst to build a series of regression models with the variables entered at each stage (block), as specified by the user.
3) By default, the regression Method is Enter, which means that all independent variables in the block will be entered into the regression equation simultaneously. This method is selected to run one regression based on all variables you specify. If you wish the program to select, from a larger set of independent variables, those that in some statistical sense are the best predictors, you can request the Stepwise method.
4) The Selection Variable option permits cross-validation of regression results. Only cases whose values meet the rule specified for a selection variable will be used in the regression analysis; then the resulting prediction equation will be applied to the other cases. Thus you can evaluate the regression on cases not used in the analysis, or apply the equation derived from one subgroup of your data to other groups.
5) The Statistics dialog box presents many additional (and some of them quite technical) statistics.
6) The Plots dialog box is used to generate various diagnostic plots used in regression, including residual plots.
7) The Save dialog box permits you to add new variables to the data file. These variables contain such statistics as the predicted values from the regression equation, various residuals and influence measures.
8) The Options dialog box controls the criteria when running stepwise regression and choices in handling missing data. By default, PASW Statistics excludes a case from regression if it has one or more values missing for the variables used in the analysis.

Note

The PASW Statistics Missing Values add-on module provides more sophisticated methods for handling missing values. This module includes procedures for displaying patterns of missing data and imputing (estimating) missing values using multiple variable imputation methods.

11.8 *Demonstration: Simple Linear Regression*

We will work with the *Bank.sav* data file in this example.

In this example we examine the relationship between beginning salary at the bank (*salbeg*) and highest year of school completed (*edlevel*). We would like to determine whether, for example, people with more education receive a higher initial salary, and then determine the impact of each additional year of education on salary.

Before we begin, we should view a scatterplot of these two variables.

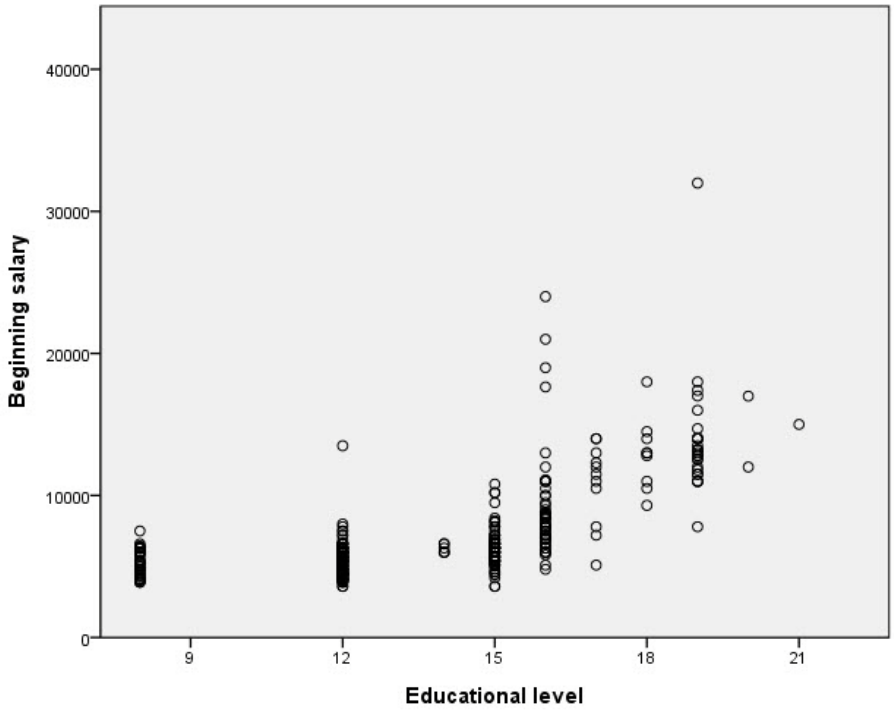
Detailed Steps for Scatterplot

In the Chart Builder dialog, accessed from **Graphs…Chart Builder**:

1) Place the **Simple Scatter icon** in the Chart Preview area
2) Place the variable ***salbeg*** in the Y-axis box.
3) Place the variable ***edlevel*** in the X-axis box.

We can observe that there is a positive relationship between education and beginning salary. This relationship, though, is not strong, as there is a fair amount of scatter in the data. The relationship does seem reasonably linear.

Figure 11.5 Scatterplot of Beginning Salary and Education Level



Note

If you wish, you can edit the chart and request a regression fit line to determine whether linear regression seems appropriate for these data.

Once we have viewed the scatterplot, we are read for the simple linear regression.

Detailed Steps for Simple Linear Regression

1) Place ***salbeg*** in the Dependent box.
2) Place ***edlevel*** in the Independent(s) box.

Results from Simple Linear Regression

Here we can observe that the multiple R or correlation coefficient between highest year of school completed and beginning salary is .633—this is the correlation between these two variables. If we square the multiple R, we get R Square, which tells us that about 40% of the variance (.401) in the dependent variable can be predicted from the independent variable. Adjusted R Square is basically the same, as there is only one independent variable in the model.

Figure 11.6 Model Summary Table for Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.633[a]	.401	.400	2439.304

a. Predictors: (Constant), Educational level

So now we know the correlation between the dependent and independent variables and the percent of variance accounted for by the model. However, we don't know whether this relationship is due to sampling variation, or instead is likely to be real and exist in the population. For this we turn to the ANOVA table.

Every statistical test has a *null hypothesis*. In most cases, the null hypothesis is that there is no relationship between two variables. This is also true for the null hypothesis for Linear Regression (the null hypothesis is that we have no relationship between the dependent and the combination of independent variable(s)). If the significance is small enough, the null hypothesis has to be rejected.

The probability of the null hypothesis being correct for this relationship is extremely small, less than .01, therefore the null hypothesis has to be rejected and the conclusion is that there is a linear relationship between these variables.

Figure 11.7 ANOVA Table

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.880E9	1	1.880E9	315.897	.000[a]
	Residual	2.808E9	472	5950202.052		
	Total	4.688E9	473			

a. Predictors: (Constant), Educational level

b. Dependent Variable: Beginning salary

Because a significant relationship has been found between our dependent and independent variable, next step is to determine what the impact on the dependent variable is.

Figure 11.8 Coefficients Table

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2516.387	536.368		-4.692	.000
	Educational level	691.011	38.879	.633	17.773	.000

a. Dependent Variable: Beginning salary

Looking at the B column, we see that for each additional year of education completed, the expected increase in beginning salary is $691.01. In fact, if we wanted to predict beginning salary based on education, we would use the following equation: *salbeg* = –2,516.387 + 691.011*(*educ*).

The column t is a technical detail and tells that the observed sample coefficient (691.011) is 17.773 times the standard error (38.879) away from the value that we expect if the null hypothesis is true, i.e., zero. Such a t value leads to a significance value of .000 (column *Sig*.) for the B coefficient.

Further Info

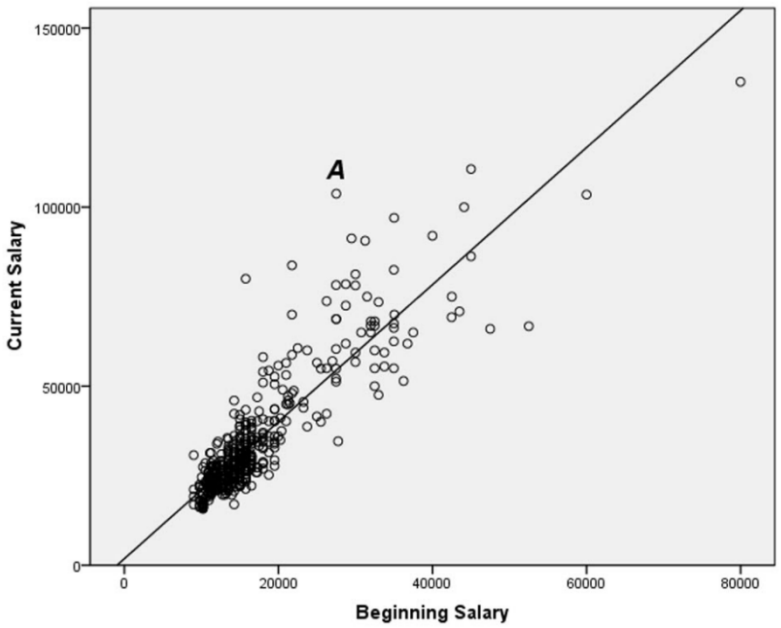
In case of simple regression the test on the null hypothesis *R Square = 0* tests the same as the test on *B = 0*. In both cases the test is whether X and Y are linearly related. This equivalence can be seen in the test statistics, as the F value in the ANOVA table is the square of the t value in the Coefficients table.

Note

The column labeled Standardized Coefficients will be discussed in the Multiple Regression section.

Apply Your Knowledge

1. True or false? In the figure depicted below the point labeled **A** has a **positive** residual?



2. True or false. Suppose we predict *Salary* with *Age* and find an R Square of .25. Then 75% of the variation in Salary cannot be accounted for by Age?

3. Which coefficient is used in the regression equation to make predictions?
   a. Beta
   b. B

11.9 *Multiple Regression*

A regression involving more than one independent variable is called multiple regression and is a direct extension of simple linear regression. When we run multiple regression we will again be concerned with how well the equation fits the data, whether a linear model is the best fit to the data, whether any of the variables are significant predictors, and estimating the coefficients for the best-fitting prediction equation. In addition, we are interested in the relative importance of the independent variables in predicting the dependent measure.

11.10 *Multiple Linear Regression Assumptions*

To correctly use multiple regression and apply statistical tests, one extra condition has to be met, in addition to the assumptions stated for simple linear regression. We restate the four assumptions for simple linear regression and add the extra assumption:

1) Variables must have a scale measurement level.
2) Variables must be linearly related.
3) Residuals must be normally distributed.
4) Residuals are assumed to be independent of the predicted values, implying that the variation of the residuals around the line is homogeneous.

5) Absence of multi-collinearity—no exact or nearly exact linear relation between the independent variables.

11.11 *Requesting Multiple Linear Regression*

Requesting multiple linear regression is accomplished with these steps:

1) Select a dependent variable and two or more independent variables
2) Request a histogram of the residuals; this allows for a check of the normality of errors
3) Request a scatterplot of the standardized residuals and standardized predicted value; this allows for a check of the homogeneity of errors
4) Review the procedure output to investigate the relationship between the variables including:
   a. R^2
   b. Adjusted R^2.
5) Examine the regression test statistics to determine whether the observed relationship is statistically significant.
6) Determine which independent variables are significantly related to the dependent variable.
7) Determine the impact of each independent variable on the dependent variable.

11.12 *Multiple Linear Regression Output*

The standard linear regression will generate the same tables depicting the relationship between the variables that were discussed earlier.

The R square and Adjusted R square are interpreted the same, except now the amount of explained variance is from a group of predictors, not just one.

Figure 11.9 Variables in the Model and Model Summary

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Education in years, Age, sex, height[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: weight

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.576[a]	.332	.326	32.13264

a. Predictors: (Constant), Education in years, Age, sex, height

The ANOVA table has a similar interpretation, except that now it tests whether *any* variable has a significant effect on the dependent variable.

Figure 11.10 ANOVA Table for Multiple Regression

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	245356.254	4	61339.063	59.408	.000[a]
	Residual	494570.496	479	1032.506		
	Total	739926.750	483			

a. Predictors: (Constant), Education in years, Age, sex, height

b. Dependent Variable: weight

In the Coefficients table, there is an additional twist to the interpretation of the B coefficient. For example, the effect here of *height* on *weight* can be stated as every additional inch of height predicts an additional 4.261 pounds of weight. However, that estimate *controls for the other variables in the model*.

Here, we must also examine the significance values for each coefficient, because a regression that is overall significant does not imply that each coefficient is statistically significant.

Figure 11.11 Multiple Regression Coefficients Table

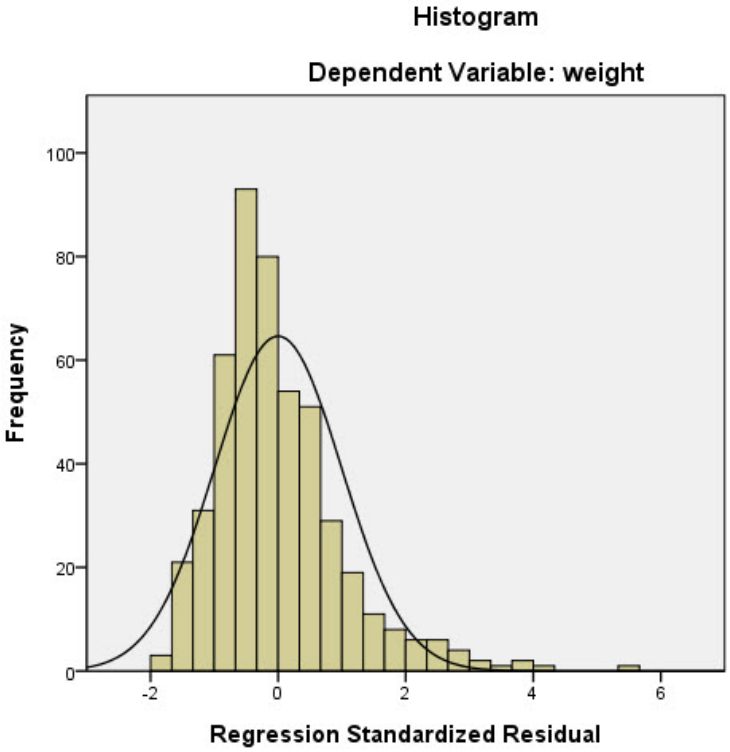
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-100.717	38.872		-2.591	.010
	height	4.261	.519	.456	8.218	.000
	Age	2.254	1.568	.056	1.438	.151
	sex	-3.311	1.082	-.168	-3.059	.002
	Education in years	-1.016	.569	-.068	-1.786	.075

a. Dependent Variable: weight

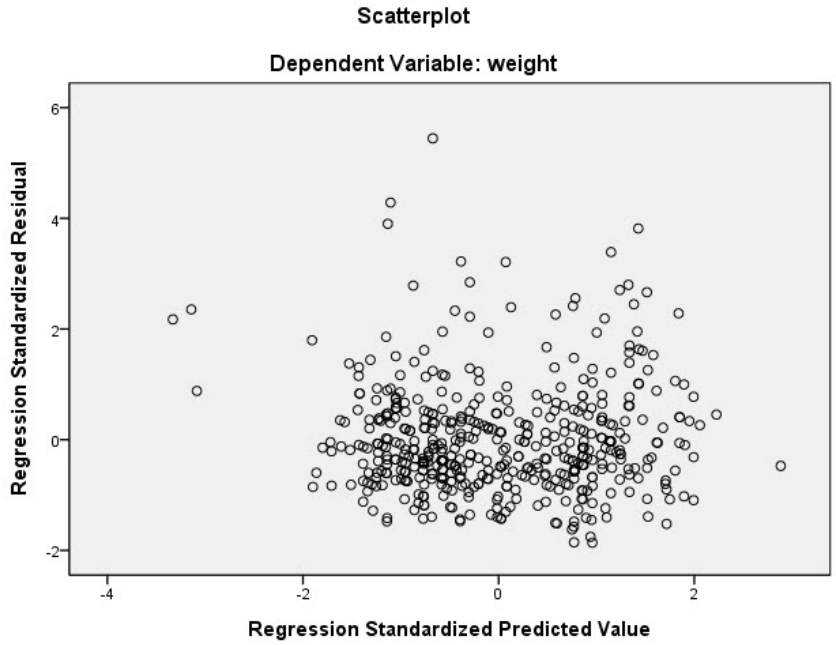
To test the assumptions of regression, we turn to the histogram of residuals. The residuals should be approximately normally distributed, which is basically true for the histogram below.

Figure 11.12 Histogram of Residuals



Additionally, the scatterplot of the standardized error (residual) and standardized predicted value (here of weight) should show no pattern if homogeneity of variance holds.

Figure 11.13 Scatterplot of Residuals and Predicted Value

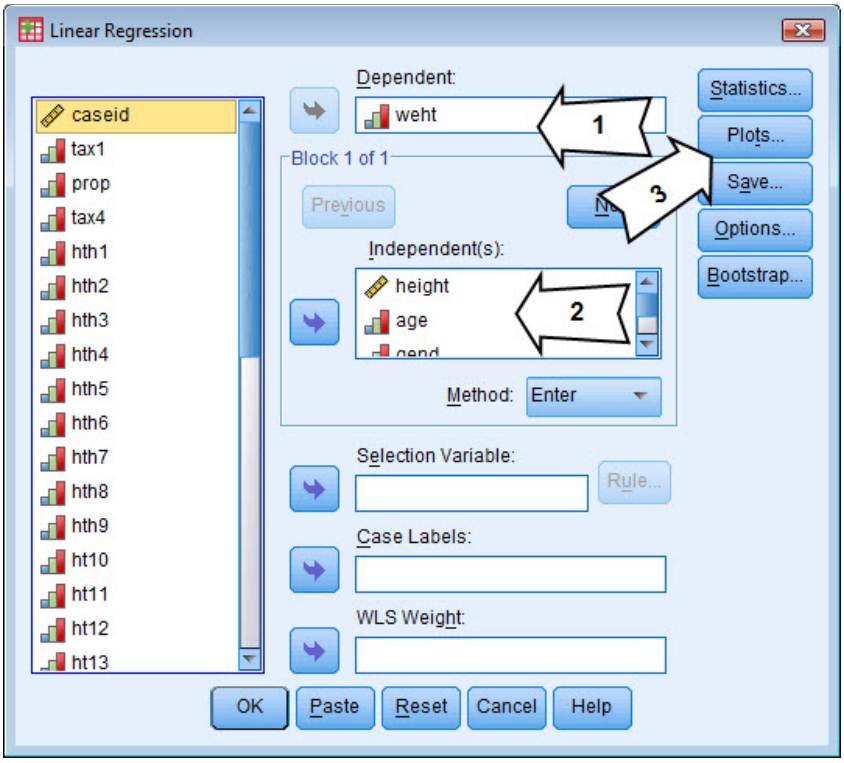


11.13 *Procedure: Multiple Linear Regression*

Multiple linear regression is available in the **Regression…Linear** menu choice. With the **Linear Regression** dialog box open:

1) Place the dependent variable in the Dependent: box.
2) Place the independent variables in the Independent(s): box.
3) Select the Plots button to open that dialog

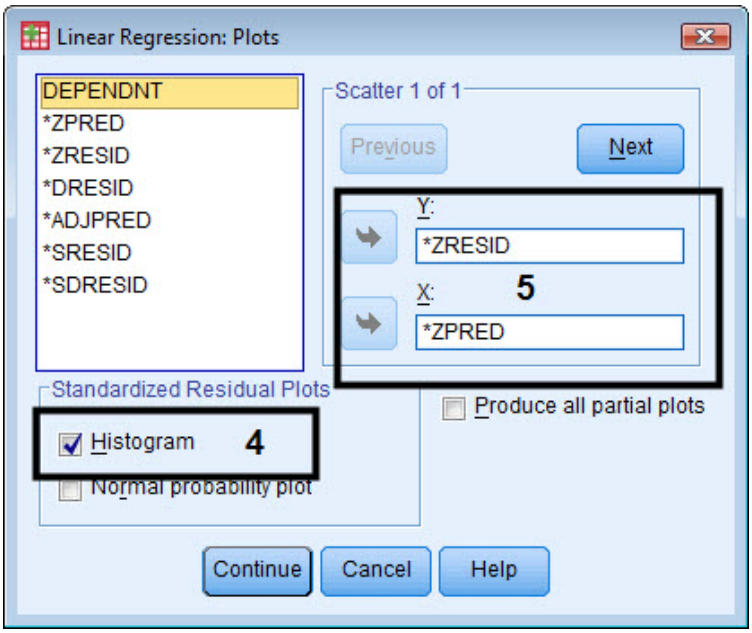
Figure 11.14 Linear Regression Dialog for Multiple Regression



In the Plots dialog:

4) Select Histogram
5) Move **ZRESID* into the Y: box in the Plots dialog and**ZPRED i*nto the X: box in the Plots dialog

Figure 11.15 Linear Regression Plots Dialog



11.14 *Demonstration: Multiple Linear Regression*

We will work with the *Bank.sav* data file in this lesson.

In this example we examine the relationship between beginning salary at the bank and several predictors, including education, years of previous work experience (*work*), *age*, and gender (*sex*). Note that gender is a dichotomous variable coded 0 for males and 1 for females, but it can be included as an independent variable (see Message Box above).

Detailed Steps for Multiple Linear Regression

1) Place ***salbeg*** in the Dependent: box.
2) Place ***edlevel***, ***sex***, ***age***, and ***work*** in the Independent(s): box.

While we can run multiple regression at this point, we will request some diagnostic plots involving residuals and information about outliers. By default no residual plots will appear.

3) Select the **Plots** button
4) Select **Histogram**.
5) Move ***ZRESID** into the Y: box in the Plots dialog.
6) Move ***ZPRED** *i*nto the X: box in the Plots dialog.

We requested a histogram of the standardized residuals because regression assumes that the residuals follow a normal distribution.

Regression can produce summaries concerning various types of residuals. We request a scatterplot of the standardized residuals (**ZRESID*) versus the standardized predicted values (**ZPRED*) because regression assumes that residuals are independent of predicted values, thus if we see any patterns (as opposed to a random blob) in this plot, it might suggest a way of adjusting and improving the analysis.

Next we request casewise diagnostics in the Statistics dialog. The Casewise Diagnostics check box requests information about all cases whose standardized residuals are more than 3 standard deviations from the fit line.

7) Select the **Statistics** button
8) Select **Casewise Diagnostics**

Results from Multiple Linear Regression

Recall that **Linear Regression** uses listwise deletion of missing data so that if a case is missing data on any of the five variables used in the regression it will be dropped from the analysis. If this results in heavy data loss, other choices for handling missing values are available in the Regression Options dialog box.

Here we can observe that the multiple R or correlation coefficient between our combination of predictors and the dependent variable is .699. If we square the multiple R, we get R square, which is .489. The Adjusted R square is .485, which is about the same.

When reporting on explained variance—R square—always report the adjusted R square.

Further Info

Therefore, about 48.5% of the variance in beginning salary can be predicted from the four independent variables.

Figure 11.16 Model Summary Table for Multiple Regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.699[a]	.489	.485	2260.153

a. Predictors: (Constant), Work experience, Sex of employee, Educational level, Age of employee

b. Dependent Variable: Beginning salary

We see that the probability of the null hypothesis being correct for this relationship is extremely small, less than .01, therefore we reject the null hypothesis and conclude that there is a linear relationship between these variables and beginning salary.

Figure 11.17 ANOVA Table for Regression of Beginning Salary

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.292E9	4	5.731E8	112.188	.000[a]
	Residual	2.396E9	469	5108291.011		
	Total	4.688E9	473			

a. Predictors: (Constant), Work experience, Sex of employee, Educational level, Age of employee

b. Dependent Variable: Beginning salary

Figure 11.18 Regression Coefficients Table to Predict Beginning Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2665.523	774.710		-3.441	.001
	Educational level	650.548	40.952	.596	15.886	.000
	Sex of employee	-1525.618	242.679	-.242	-6.287	.000
	Age of employee	33.017	15.636	.124	2.112	.035
	Work experience	20.342	21.829	.056	.932	.352

a. Dependent Variable: Beginning salary

In the Coefficients table, the independent variables appear in the order they were listed in the **Linear Regression** dialog box, not in order of importance. Although the *B* coefficients are important for prediction and interpretive purposes, analysts usually look first to the *t* test at the end of each line to determine which independent variables are significantly related to the outcome measure. Since four variables are in the equation, we are testing if there is a linear relationship between each independent

variable and the dependent variable after adjusting for the effects of the three other independent variables. Looking at the significance values we see that *edlevel*, *sex*, and *age* significant at the .05 significance level, while the work experience is not (sig = .352 after controlling for the other predictors).

The estimated regression (*B*) coefficient for *edlevel* is about $651, similar but not identical to the coefficient (691) found in the simple regression using *edlevel* alone. In the simple regression we estimated the *B* coefficient for *edlevel* ignoring any other effects, since none were included in the model. Here we evaluate the effect of *edlevel* after controlling (statistically adjusting) for *age*, *sex*, and *work*. If the independent variables are correlated, the change in *B* coefficient from simple to multiple regression can be substantial. So, after controlling (holding constant) *age*, *sex*, and *work*, one year of formal education, on average, was worth another $651 in beginning salary.

Continuing on:

- The variable *sex* has a B coefficient of about –$1,526. This means that a one-unit change in gender (which means moving from male status to female, or comparing females to males), controlling for the other variables, is associated with a drop in beginning salary of –$1,526.
- The variable *age* has a B coefficient of $33, so each additional year increases beginning salary by $33.
- Since we found work experience's coefficient to be not significantly different from zero, we treat it as 0.

If we simply look at the estimated *B* coefficients we might think that *sex* is the most important variable. However, the magnitude of the *B* coefficient is influenced by the unit of measurement (or standard deviation if you like) of the independent variable. The *Beta coefficients* explicitly adjust for such standard deviation differences in the independent variables.

- They indicate what the regression coefficients would be if all variables were standardized to have means of 0 and standard deviations of 1.
- A *Beta* coefficient thus indicates the expected change (in standard deviation units) of the dependent variable per one standard deviation unit increase in the independent variable (after adjusting for other predictors). This provides a means of assessing relative importance of the different predictor variables in multiple regression.
- The *Betas* are normed so that the maximum should be less than or equal to one in absolute value (if any *Betas* are above 1 in absolute value, it suggests a problem with the data: multi-collinearity).

Examining the *Betas*, we see that *edlevel* is the most important predictor, followed by *sex*, and then *age*. The *Beta* for *work* is near zero, as we would expect.

If we needed to predict *salbeg* from these background variables (dropping *work*) we would use the *B* coefficients. Rounding to whole numbers, we would say:

salbeg = –2,667 + 651**edlevel* – 1526**sex* + 33**age*.

Diagnostic Statistics

The request for casewise diagnostics produces two tables, the most important of which is shown below. The Casewise Diagnostics table lists those observations more than three standard deviations (in error) from the regression fit line. Assuming a normal distribution, this would happen less than 1% of the time by chance alone. In this data file that would be about 5 outliers (.01*474), so the six cases does not seem excessive. Residuals should normally be balanced between positive and negative values; when they are not, you should investigate the data further. In these data, all six residuals are positive, so this does indicate that some additional investigation is required. We could, for example see if these observations have anything in common (very high education which may be out of line

with others). Since we know their case numbers (an ID variable can be substituted), we could find them easily at them more closely.

We also don't want to discover very large prediction errors, but here residuals are very, very high, over 6 standard deviations above the fit line.

Figure 11.19 Casewise Listing of Outliers

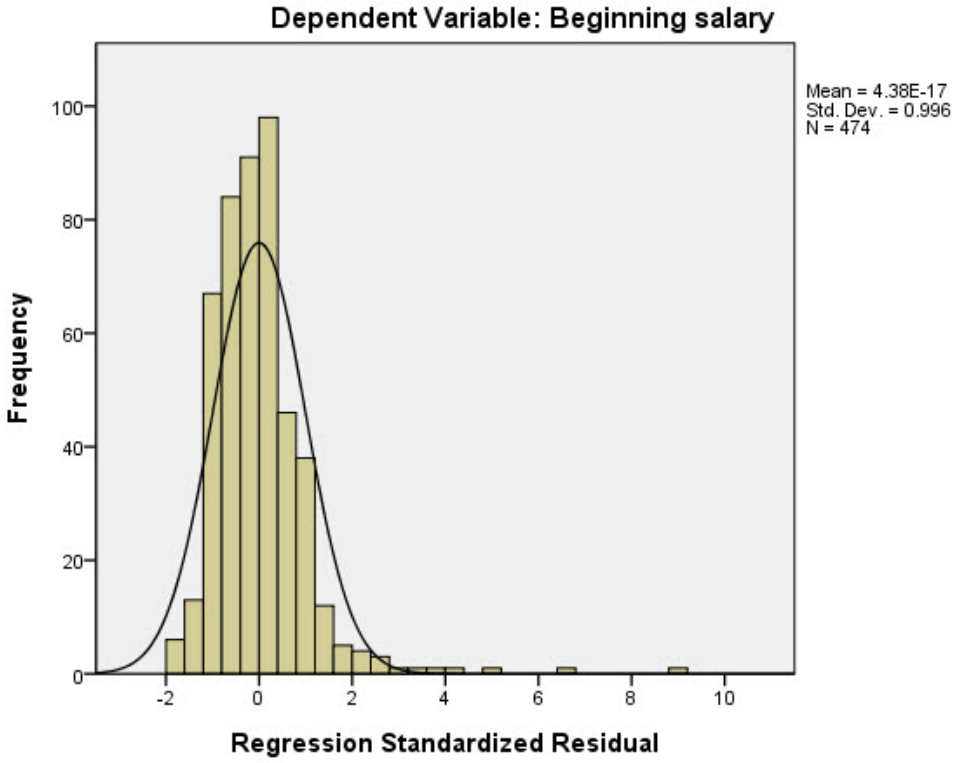
Casewise Diagnostics[a]

Case Number	Std. Residual	Beginning salary	Predicted Value	Residual
2	6.491	24000	9329.08	14670.917
55	3.425	18000	10259.99	7740.010
56	8.992	31992	11669.13	20322.873
67	4.247	18996	9398.09	9597.912
122	4.943	21000	9828.79	11171.210
132	3.005	18000	11208.08	6791.918
415	3.767	17640	9126.77	8513.226

a. Dependent Variable: Beginning salary

In the diagnostic plots involving residuals we see the distribution of the residuals with a normal bell-shaped curve superimposed, depicted in the figure below. The residuals are fairly normal, although they are a bit too concentrated in the center. They are also somewhat positive skewed. Given this pattern, a data analyst might try a data transformation on the dependent measure, which might improve the properties of the residual distribution, e.g., the log. However, just as with ANOVA, larger sample sizes protect against moderate departures from normality.

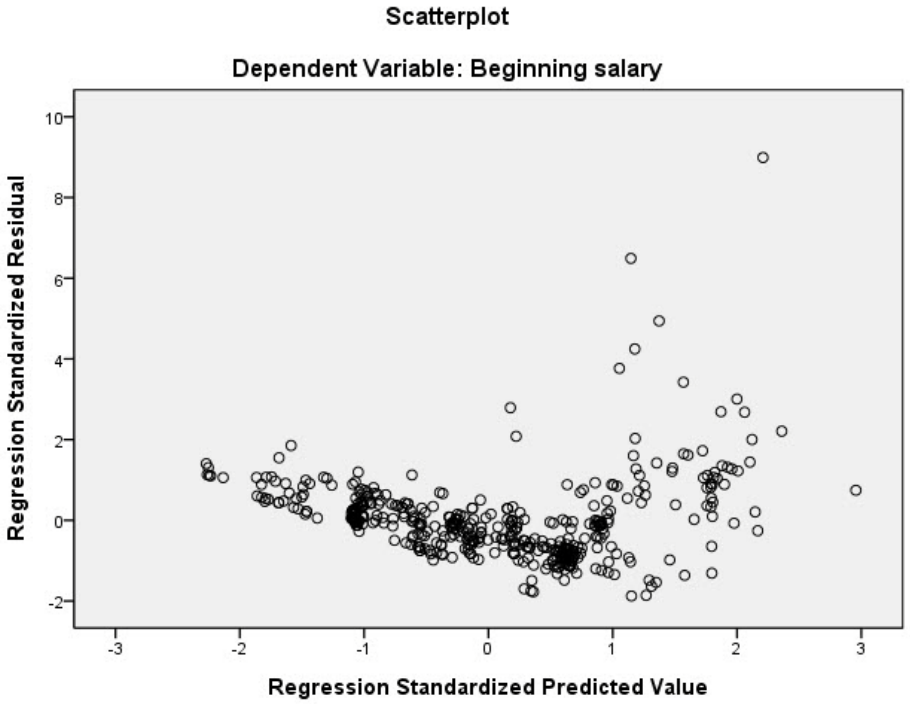
Figure 11.20 Histogram of the Residuals



In the scatterplot of residuals, we hope to see a horizontally oriented blob of points with the residuals showing the same spread across different predicted values. Unfortunately, we see a hint of a curving pattern: the residuals seem to slowly decrease, then swing up at higher salaries. This type of pattern can mean the relationship is curvilinear.

Also, the spread of the residuals is much more pronounced at higher predicted salaries. This suggests lack of homogeneity of variance.

Figure 11.21 Scatterplot of Residuals and Predicted Value for Beginning Salary



Apply Your Knowledge

1. Consider the output below, for the regression where Miles per gallon for a vehicle is predicted from Engine Size, Horsepower, Weight and American car or not (coded as 1 for American cars and 0 for cars from other countries). Which statements are correct?
    a. We predict lower mpg for an American car than for a non-American car
    a. All predictors have an effect significantly different from 0
    b. The most important predictor for Miles per Gallon is Vehicle Weight
    c. The fact that there is an Unstandardized coefficient greater than 1 (in absolute value) indicates a problem due to multi-collinearity.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45.396	1.216		37.334	.000
	Engine Displacement (cube inches)	.004	.007	.057	.579	.563
	Horsepower	-.054	.013	-.268	-4.058	.000
	Vehicle Weight (lbs.)	-.005	.001	-.579	-7.494	.000
	American car	-1.977	.608	-.123	-3.251	.001

a. Dependent Variable: Miles per Gallon

Additional Resources

For additional information on linear regression analysis, see:

Allison, Paul D. 1998. *Multiple Regression: A Primer*. Thousand Oaks, CA: Pine Forge.

Further Info

Draper, Norman and Smith, Harry. 1998. *Applied Regression Analysis.* 3rd ed. New York: Wiley.

11.15 *Lesson Summary*

We explored the use of **Linear Regression** to test relationships between scale variables and develop prediction equations to prediction the dependent variable and determine the impact of each independent variable on the dependent variable.

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Perform linear regression to determine whether one or more variables can significantly predict or explain a dependent variable

To support the achievement of the primary objective above, students should also be able to:

- Explain linear regression and its assumptions
- Explain the options of the **Linear Regression** procedure
- Interpret the results of the **Linear Regression** procedure

11.16 *Learning Activity*

The overall goal of this learning activity is to run linear regressions and to interpret the output. You will use the PASW Statistics data file *Census.sav*.

Supporting Materials

The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. Run a linear regression to predict total family income (*income06*) with highest year of education (*educ*). First, do a scatterplot of these two variables and superimpose a fit line. Does the relationship seem linear? How would you characterize the relationship?

2. Now run the linear regression. What is the Adjusted R square value? Is the regression significant? What is the B coefficient for *educ*? Interpret it.

3. Next add the variables *born* (born in the U.S. or overseas), *age*, *sex*, and number of brothers and sisters (*sibs*). Check the coding on *born* so you can interpret its coefficient. First, do a scatterplot of *age* and *sibs* with *income06*. Superimpose a fit line. Does the relationship seem linear? How would you characterize the relationship? Why not do scatterplots of *income06* with *sex* and *born*?

4. Use all these variables to predict *income06*. Request residual statistics including the histogram of errors and the scatterplot of standardized values. Also request casewise diagnostics. What is the Adjusted R square? How much has it increased from above?

5. Which variables are significant predictors? What is the effect of each on *income06*? Which variable is the strongest predictor? The weakest?
6. Examine the casewise diagnostics. Do you see any pattern? Are there more cases with large errors than we would expect?

7. Examine the histogram and scatterplot. Are the errors normally distributed? Do you see any pattern in the scatterplot? What might that mean?

8. What is the prediction equation for *income06*?

9. *For those with more time:* Add additional variables to the regression equation for *income06*. Examples are father and mother's education, or number of children. Be careful to add variables that are at least on an interval scale of measurement. Repeat the exercise above. Are the new variables significant predictors? Does adding variables change the effects of the variables already in the model from above?

Lesson 12: Nonparametric Tests

12.1 *Objectives*

After completing this lesson students will be able to:

- Perform non-parametric tests on data that don't meet the assumptions for standard statistical tests

To support the achievement of this primary objective, students will also be able to:

- Describe when non-parametric tests should and can be used
- Describe the options in the Nonparametric procedure dialog box and tabs
- Interpret the results of several types of nonparametric tests

12.2 *Introduction*

Parametric tests, such as the t test, ANOVA, or the Pearson correlation coefficient, make several assumptions about the data. They typically assume normality of the variable(s) distribution for scale variables, and they often assume that the variance is equal within categories of a grouping or factor variable. If these assumptions are clearly violated, the results of these tests are in doubt. Fortunately there are alternatives.

There are a whole family of various tests and methods that make fewer assumptions about the data. These tests fall under the class of *nonparametric* statistics.

1) These methods generally don't assume normality or variance equality.
2) They are generally less powerful than parametric tests, which means they have a lower chance of finding true differences.
3) These tests are most useful with questions using a short response scale with 3, 4, or 5 points. These scales are not truly interval in measurement.
4) They are also useful when variables have very skewed distributions and so the normality assumption is violated.
5) These tests are also commonly used when sample size is small.

Business Context

Nonparametric tests allow us to determine if we have relationships in our data when we do not meet important distributional assumptions. They permit us to do standard analysis—looking at group differences, or assessing associations between variables—with all types of data, thus extending data analysis capabilities.

Supporting Materials

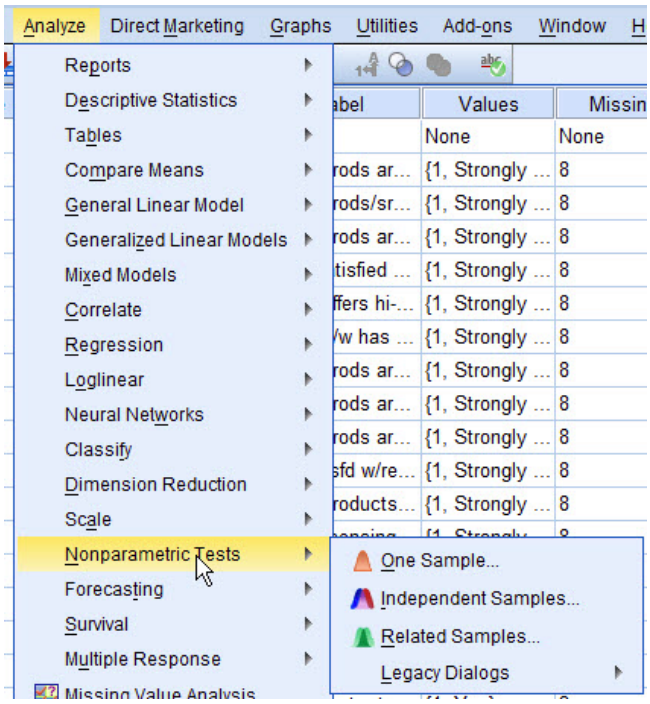
The file *Census.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

12.3 *Nonparametric Analyses*

PASW Statistics includes a wizard to guide you through selecting the appropriate nonparametric test for a particular set of variables. You need to know whether the general situation is:

- One Sample: A dependent variable with no grouping variable
- Independent Samples: A dependent variable with a grouping (factor) variable
- Related Samples: Two (or more) dependent variables whose association you wish to test (such as experiments with pre- and post-test measurements)

Figure 12.1 Nonparametric Menu Choices



Important

By default, the Nonparametric Tests procedures will use the declared scale of measurement of variables to determine how they are used. In particular, recall that nonparametric tests are an alternative to parametric methods such as **One-Way ANOVA**, and these parametric procedures assume that a dependent variable is scale. Therefore, to conduct the equivalent nonparametric test, the dependent variable must have scale level of measurement in PASW Statistics..
If you use the default settings in the Nonparametric Tests Wizard, it is critical that the level of measurement be set correctly for each variable in the analysis.

12.4 *The Independent Samples Nonparametric Analysis*

When the dependent variable is scale and we want to test equality of population means for two, or three or more groups, the procedures that we have used were the **Independent-Samples T Test** and

One-Way ANOVA, respectively. Certain conditions (normality, homogeneity of variances) had to be satisfied to use these procedures.

If the assumptions are violated, or if the variable is ordinal in nature, these tests cannot be used and an alternative is needed. Nonparametric Independent Samples tests provide this alternative. The wizard will select the appropriate test, depending on whether there are two groups or three or more groups. If there are more than two groups, a post hoc analysis can be run to determine which groups differ significantly, analogous to the post hoc pairwise comparisons in the **One-Way ANOVA** procedure.

Independent Samples Nonparametric Assumptions

The nonparametric tests for two or more independent samples only assume:

1) There is a categorical independent variable defining two or more groups
2) There is an ordinal or scale dependent variable on which group differences are tested.

12.5 *Requesting an Independent Samples Nonparametric Analysis*

Requesting a nonparametric test for two or more independent samples is accomplished with these steps:

1) Select **Independent Samples** from the Nonparametric Tests menu entry.
2) Select whether you want to compare the shape of the distributions, the medians, or customize the analysis.
3) Specify the Test field and Group variables.
4) To see the equivalent of post hoc pairwise comparisons, you can request a specific test and the appropriate comparison.
5) Review the significance test and other output in the Model Viewer.

12.6 *Independent Samples Nonparametric Tests Output*

The output from the **Independent Samples** tests, and for all procedures in the Nonparametric Tests menu entry, is produced in the Model Viewer. Initially, all that is displayed is the actual test result.

- The null hypothesis being tested is described in plain language
- The specific test used is noted, here the Kruskal-Wallis test
- The significance level is listed
- The decision about the null hypothesis is listed, using the .05 level of significance

Figure 12.2 Nonparametric Independent Samples Hypothesis Test Summary

Hypothesis Test Summary

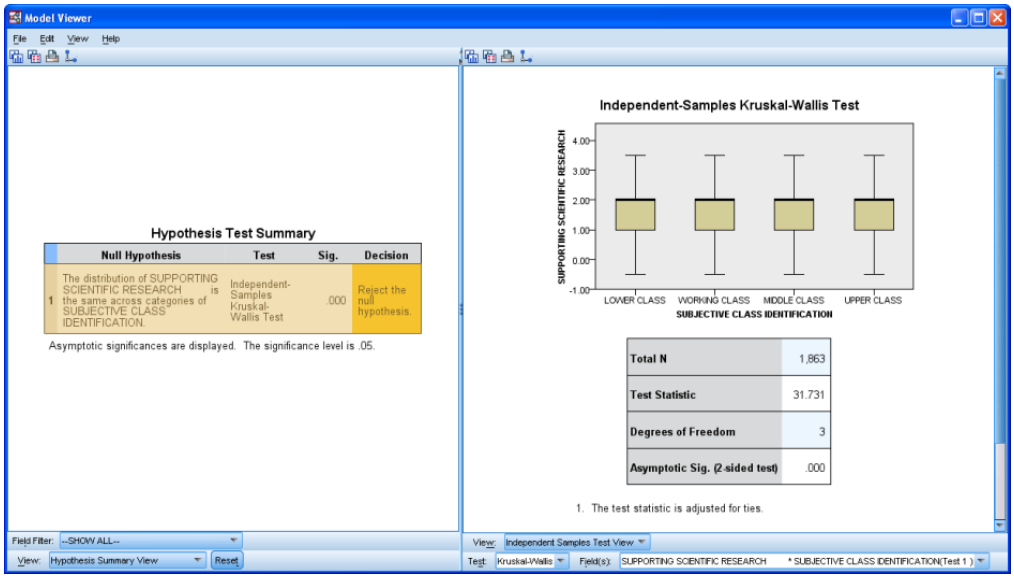
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SUPPORTING SCIENTIFIC RESEARCH is the same across categories of SUBJECTIVE CLASS IDENTIFICATION.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

By double-clicking on the Model output, the Model Viewer window is opened, with additional output. The Model Viewer contains two panes:

1) The Main View appears on the left side. It displays general information about the tests, and there is a dropdown list at the bottom of the pane that allows you to switch between views about the test/model.
2) The Auxiliary view appears on the right side. This view displays a more detailed visualization (including tables and graphs) of the model compared to the general visualization in the main view. Like the main view, the auxiliary view may have more than one element that can be displayed. Initially, in the main view, the overall test is listed, identical to what was displayed in the Output Viewer. In the Auxiliary view, the distribution of the dependent variable is displayed via a boxplot, and details about the test are listed.

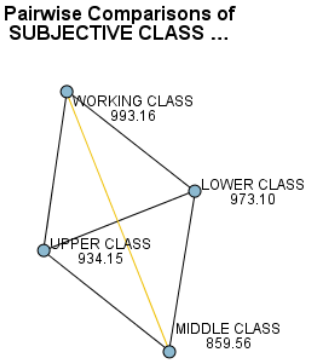
Figure 12.3 Model Viewer Panes for Nonparametric Independent Samples Test



In the Auxiliary view, we can open the Pairwise Comparisons view, as shown in the figure below, where all the possible pairwise comparisons are listed (with no redundancy). The tests are adjusted for the fact that multiple comparisons are being done (6 tests in this example). Here only one comparison is significant at the .05 level, that between the middle and working class groups. They differ in their attitude toward spending on supporting scientific research.

The distance network chart—the graph above the table—lists the average rank for each category of the grouping variable. The rank is used because this is a nonparametric test, and data can be ranked without making many assumptions about the dependent variable. It is a graphical representation of the comparisons table in which the distances between nodes in the network correspond to differences between samples. Yellow lines correspond to statistically significant differences; black lines correspond to non-significant differences. Hovering over a line in the network displays a tooltip with the adjusted significance of the difference between the nodes connected by the line.

Figure 12.4 Pairwise Comparison Tests



Each node shows the sample average rank of SUBJECTIVE CLASS IDENTIFICATION.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
MIDDLE CLASS-UPPER CLASS	-74.587	60.993	-1.223	.221	1.000
MIDDLE CLASS-LOWER CLASS	113.535	43.181	2.629	.009	.051
MIDDLE CLASS-WORKING CLASS	133.602	24.174	5.527	.000	.000
UPPER CLASS-LOWER CLASS	38.948	70.608	.552	.581	1.000
UPPER CLASS-WORKING CLASS	59.014	60.871	.970	.332	1.000
LOWER CLASS-WORKING CLASS	-20.066	43.008	-.467	.641	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

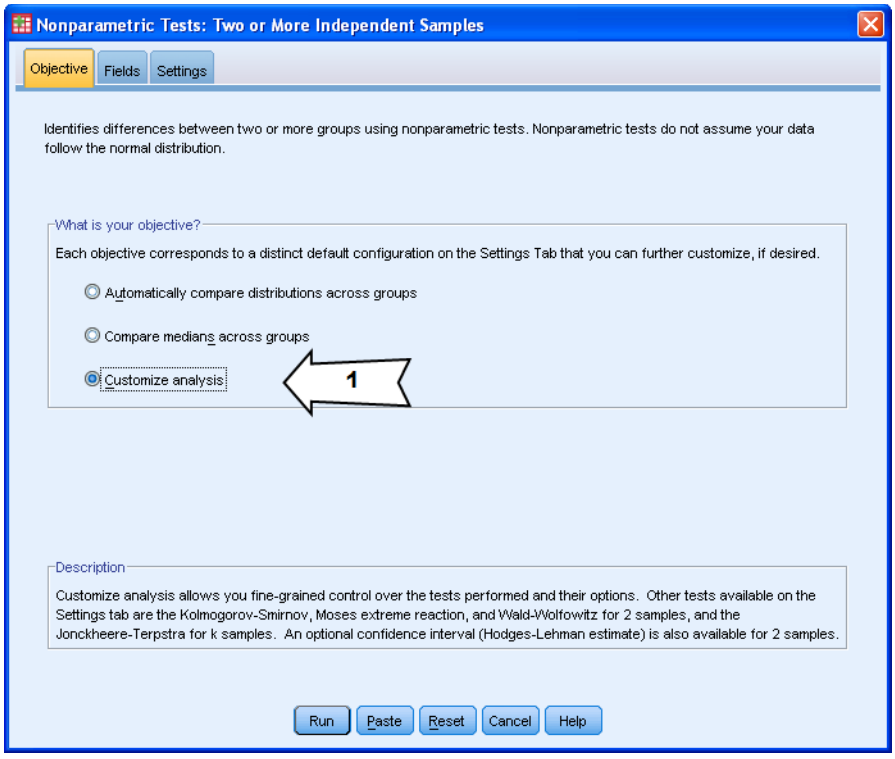
12.7 *Procedure: Independent Samples Nonparametric Tests*

The nonparametric **Independent Samples** tests are accessed from the **Analyze…Nonparametric Tests…Independent Samples** menu choice. With the dialog box open:

In the Objectives tab:

1) Select Customize analysis

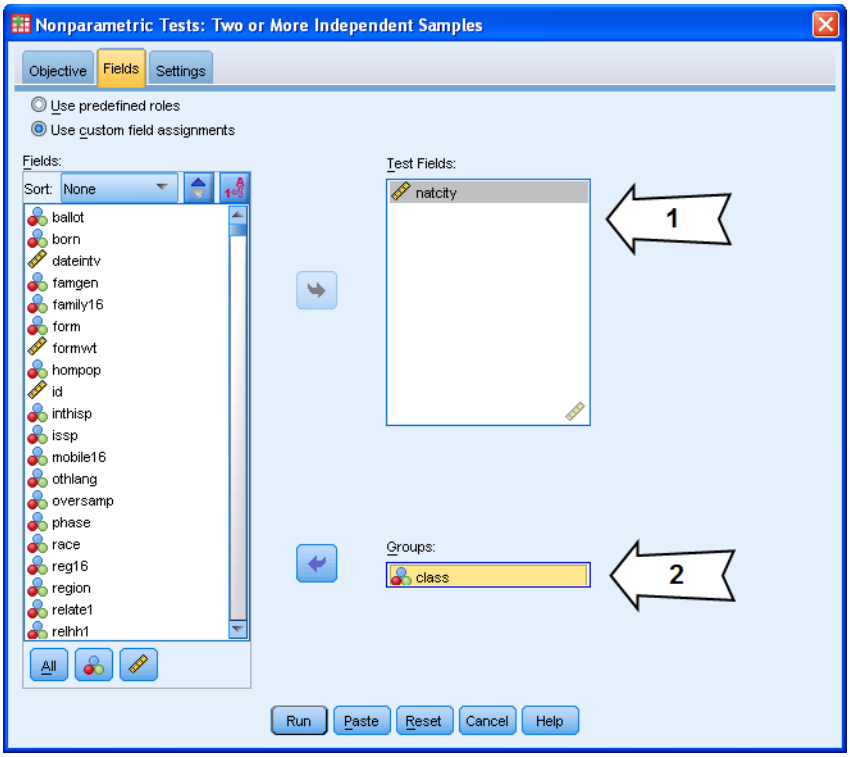
Figure 12.5 Nonparametric Tests Two or More Independent Samples Dialog



In the Fields tab:

1) Specify the test variable(s)
2) Specify the Groups variable

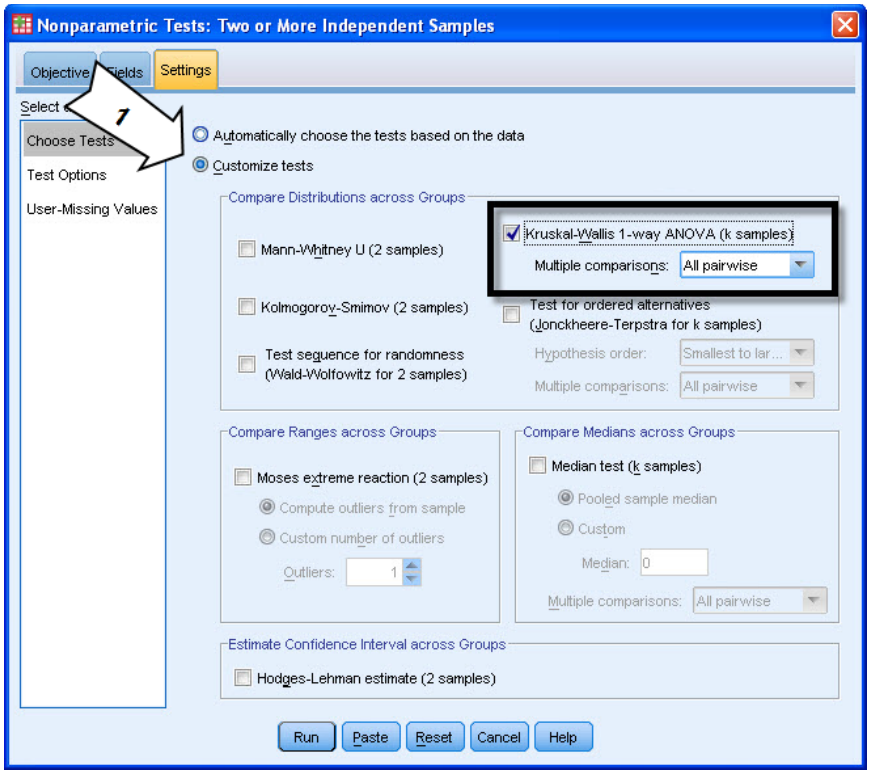
Figure 12.6 Two or More Independent Samples Fields Tab Dialog



In the Settings tab:

1) Select Customize tests and specify the desired test and, if applicable, any multiple comparisons option.

Figure 12.7 Two or More Independent Samples Settings Tab Dialog



12.8 *Demonstration: Independent Samples Nonparametric Tests*

In this example we will use the file *Census.sav*. Our objective is to see how one's political position (from liberal to conservative; the variable *polviews*) is related to marital status. For example, are married people more conservative than others?

The variable *polviews* is measured on a seven-point scale, and it is truly ordinal, not interval, in measurement scale. Therefore, testing for differences by marital status is best done with a nonparametric method.

Note

The variable *polviews* may not be Scale in measurement level in the data. If not, change its measurement level before beginning this example.

Detailed Steps for Independent Samples Nonparametric Test

1) In the Objectives tab, select **Customize analysis**.
2) In the Fields tab, specify ***polviews*** as the Test Fields variable.
3) Specify ***marital*** as the Groups variable.
4) In the Settings tab, select **Customize tests**

5) Select the **Kruskal-Wallis 1-way ANOVA** test, and also select the **All pairwise** option on the Multiple comparisons: dropdown

Results from Independent Samples Nonparametric Test

The model view output, initially condensed in the Viewer window, shows that we reject the null hypothesis, as the significance of the Kruskal-Wallis test is .000. This test uses the ranks of cases on the dependent variable to determine whether there are differences between categories, and we conclude that there are.

Figure 12.8 Nonparametric Test of Political Position by Marital Status

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of THINK OF SELF AS LIBERAL OR CONSERVATIVE is the same across categories of MARITAL STATUS.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

To decide which categories differ from others, we need to open the Model viewer and look in the Auxiliary pane at the Pairwise Comparisons view. There are many possible comparisons to make, and those significant at the .05 level are highlighted. We find that:

- Those who are never married are significantly different than those who are divorced, widowed, or married
- There are no other pairs that are significantly different (although the separated-married pair has an adjusted significance of .063).

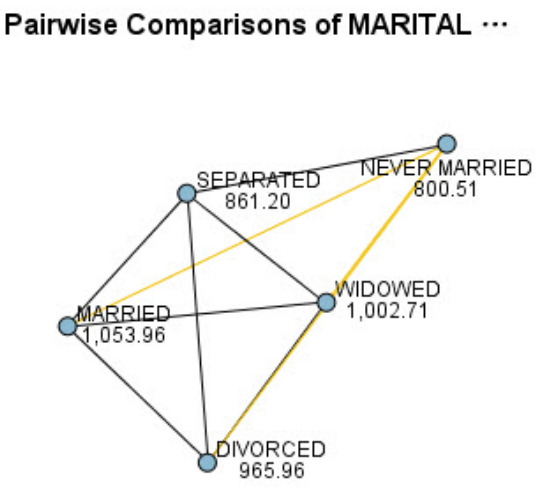
Figure 12.9 Pairwise Comparison Tests for Marital Status

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
NEVER MARRIED-SEPARATED	60.691	72.400	.838	.402	1.000
NEVER MARRIED-DIVORCED	165.448	40.601	4.075	.000	.000
NEVER MARRIED-WIDOWED	202.194	49.895	4.052	.000	.001
NEVER MARRIED-MARRIED	253.454	29.662	8.545	.000	.000
SEPARATED-DIVORCED	104.757	75.802	1.382	.167	1.000
SEPARATED-WIDOWED	141.504	81.160	1.744	.081	.812
SEPARATED-MARRIED	192.763	70.550	2.732	.006	.063
DIVORCED-WIDOWED	36.746	54.714	.672	.502	1.000
DIVORCED-MARRIED	88.005	37.201	2.366	.018	.180
WIDOWED-MARRIED	51.259	47.170	1.087	.277	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

To see the average rank on *polviews* for the pairs that differ, we can look at the distance network chart. The average rank for the never married category is 800.51, lower than any other category. And lower values on *polviews* indicate more liberal attitudes, so this implies that those who have never been married are more liberal than married, divorced, and widowed respondents.

Figure 12.10 Distance Network Chart for Marital Status



Each node shows the sample average rank of MARITAL STATUS.

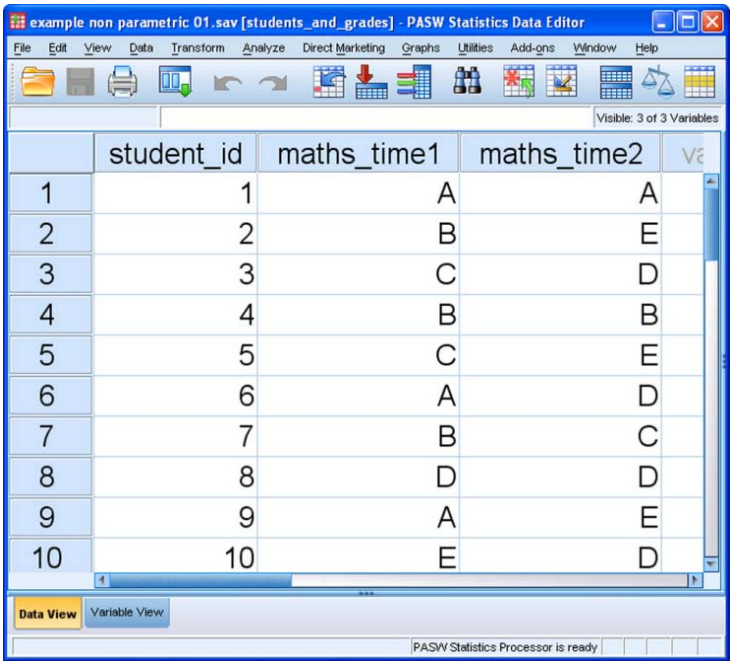
Further Info

A One-Way ANOVA would find similar results as the Kruskal-Wallis test, but the nonparametric test is more appropriate for the data. However, sometimes an analyst will perform both tests, and if the nonparametric test is consistent with the parametric test results, report only on the latter.

Apply Your Knowledge

1. True or false? In order to assess if there is a relationship between region and gender (two nominal variables), a nonparametric Independent Samples test can be run?

2. See the dataset shown below, with data collected on students, with grade for mathematics at two different points in time. What is the appropriate test to see whether grade differs at time 2 from time 1?
   a. Nonparametric tests: **One Sample**
   b. Nonparametric tests: **Independent Samples**
   c. Nonparametric tests: **Related Samples**
   d. Parametric test: **Independent Samples T-Test**



	student_id	maths_time1	maths_time2	Va
1	1	A	A	
2	2	B	E	
3	3	C	D	
4	4	B	B	
5	5	C	E	
6	6	A	D	
7	7	B	C	
8	8	D	D	
9	9	A	E	
10	10	E	D	

12.9 *The Related Samples Nonparametric Analysis*

As the parametric **Independent-Samples T Test** and **One-Way ANOVA** have their analog in the nonparametric **Independent Samples** tests, so has the parametric **Paired-Samples T Test** its equivalent in the nonparametric **Related Samples** test. (Actually the parametric **Paired-Samples T Test** is done with two paired variables, while the nonparametric **Related Samples** procedure allows for two or more paired variables.)

The nonparametric tests for two or more paired samples only assume that:

1) The measurement level of the variables is ordinal or scale (depending on the specific test chosen).

12.10 *Requesting a Related Samples Nonparametric Analysis*

Requesting a nonparametric analysis is accomplished with these steps:

1) Select the **Related Samples** menu selection.
2) Select whether you want to compare observed data to hypothesized or customize the analysis.
3) Specify the Test fields.
4) Review the significance test and other output in the Model Viewer.

12.11 *Related Samples Nonparametric Tests Output*

The output from the nonparametric **Related Samples** procedure is produced in the Model Viewer. Initially, all that is displayed is the actual test result.

- The null hypothesis being tested is described in plain language
- The specific test used is noted, here the Wilcoxon Signed Ranks test
- The significance level is listed
- The decision about the null hypothesis is listed, using the .05 level of significance

Figure 12.11 Nonparametric Related Samples Test Summary

Hypothesis Test Summary

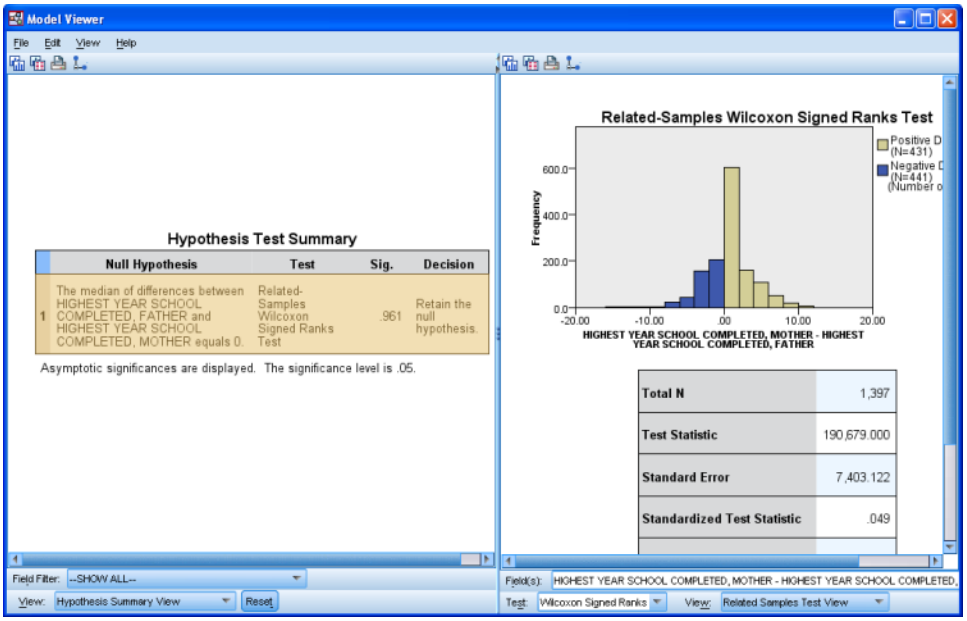
	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between HIGHEST YEAR SCHOOL COMPLETED, FATHER and HIGHEST YEAR SCHOOL COMPLETED, MOTHER equals 0.	Related-Samples Wilcoxon Signed Ranks Test	.961	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

By double-clicking on the Model output, the Model Viewer window is opened, with additional output. The Model Viewer contains two panes:

1) The Main View appears on the left side. It displays general information about the tests, and there is a dropdown list at the bottom of the pane that allows you to switch between views about the test/model.
2) The Auxiliary view appears on the right side. This view displays a more detailed visualization (including tables and graphs) of the model compared to the general visualization in the main view. Like the main view, the auxiliary view may have more than one element that can be displayed. Initially, in the main view, the overall test is listed, identical to what was displayed in the Output Viewer. In the auxiliary view, the distribution of the variables is displayed via a histogram, and details about the Wilcoxon Signed Ranks test are listed.

Figure 12.12 Model Viewer Panes Related Samples Test



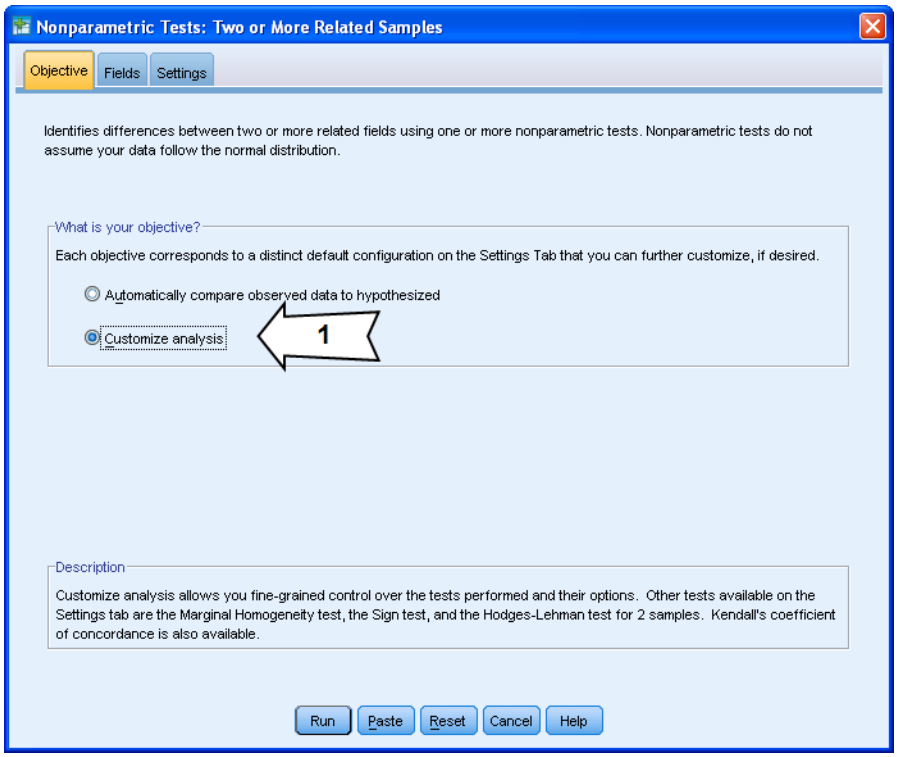
12.12 *Procedure: Related Samples Nonparametric Tests*

The nonparametric **Related Samples** procedure is accessed from the **Analyze…Nonparametric Tests…Related Samples** menu choice.

In the Objectives tab:

1) Select Customize analysis.

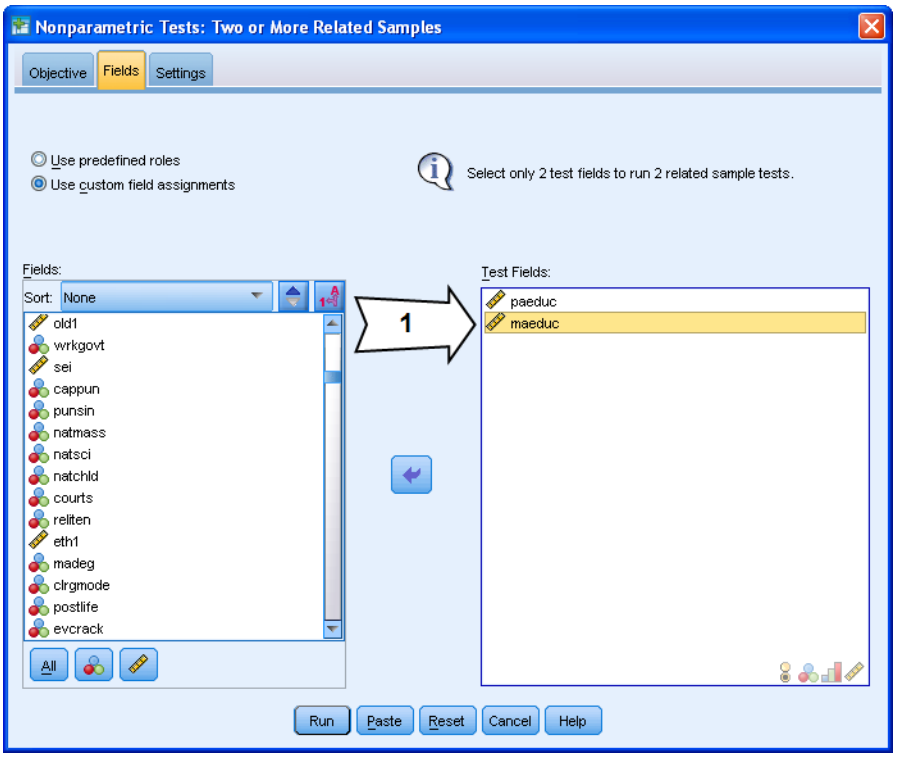
Figure 12.13 Nonparametric Tests Two or More Related Samples Dialog



In the Fields tab:

1) Specify the Test Fields variables

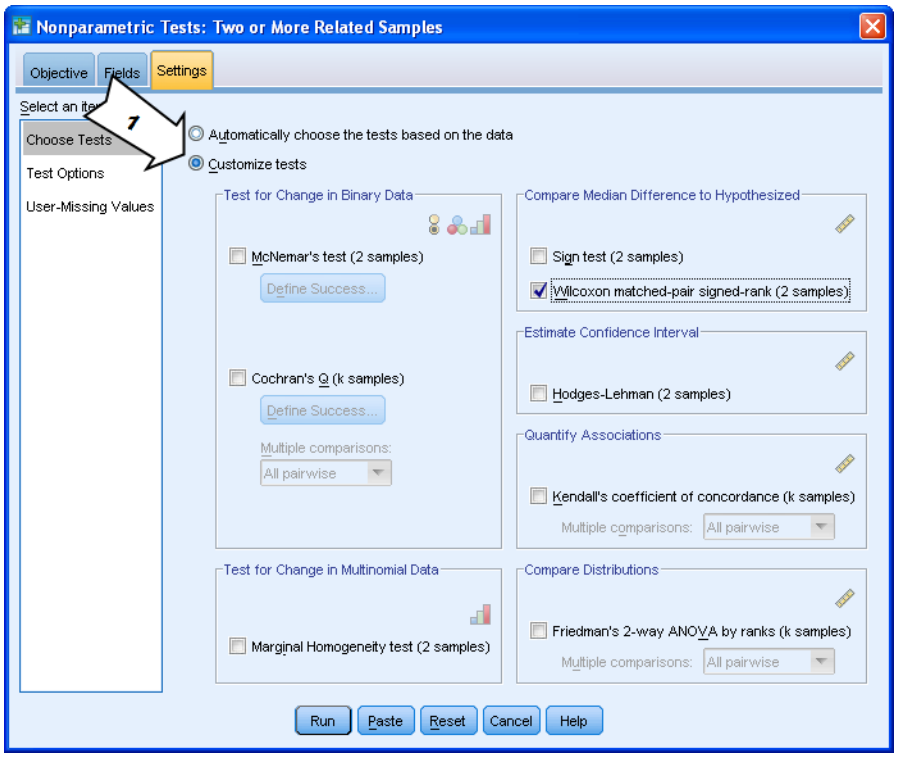
Figure 12.14 Field Tab Dialog Related Samples



In the Settings tab:

1) Select Customize tests and specify the desired test

Figure 12.15 Settings Tab Dialog Related Samples



12.13 *Demonstration: Related Samples Nonparametric Tests*

In this example we will continue to use the data file *Census.sav*. Several questions asked about the respondent's interest in various subjects, on a three-point scale from *Very Interested* to *Not at all interested*. We would like to see whether the respondents have more interest in medical discoveries (*intmed*) or in scientific discoveries (*intsci*).

To understand what we are testing, below are the two frequency tables for these two variables. These variables are measured on an ordinal scale, so a paired-sample t test not justified. We can see that the percentage of people saying they are Very Interested in medical discoveries is higher than for scientific discoveries, but we want to see if this difference is statistically significant. We will use a test based on the median of the distribution.

Figure 12.16 Frequencies for Interest in Medicine and Interest in Science

INTERESTED IN MEDICAL DISCOVERIES

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very interested	907	44.8	60.6	60.6
	Moderately interested	496	24.5	33.2	93.8
	Not at all interested	93	4.6	6.2	100.0
	Total	1496	73.9	100.0	
Missing	IAP	518	25.6		
	DONT KNOW	7	.3		
	NA	2	.1		
	Total	527	26.1		
Total		2023	100.0		

INTERESTED IN NEW SCIENTIFIC DISCOVERIES

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very interested	607	30.0	40.6	40.6
	Moderately interested	675	33.4	45.2	85.8
	Not at all interested	213	10.5	14.2	100.0
	Total	1495	73.9	100.0	
Missing	IAP	518	25.6		
	DONT KNOW	9	.4		
	NA	1	.0		
	Total	528	26.1		
Total		2023	100.0		

Detailed Steps for Related Samples Nonparametric Test

1) Change the level of measurement to **Scale** for ***intmed*** and ***intsci***
2) In the Objectives tab of the **Nonparametric tests: Two or More Related Samples** dialog, select **Customize analysis.**
3) In the Fields tab, specify ***intmed*** and ***intsci*** as the Test Field variables
4) In the Settings tab, select the **Wilcoxon matched-pair signed-rank** test.

Results from Related Samples Nonparametric Test

The model view output, initially condensed in the Viewer window, shows that we reject the null hypothesis, as the significance of the Wilcoxon Signed Ranks test is .000. This test uses the ranks of cases on the variables to determine whether there are differences between them.

Figure 12.17 Nonparametric Test of Interest in Medicine and Science

Hypothesis Test Summary

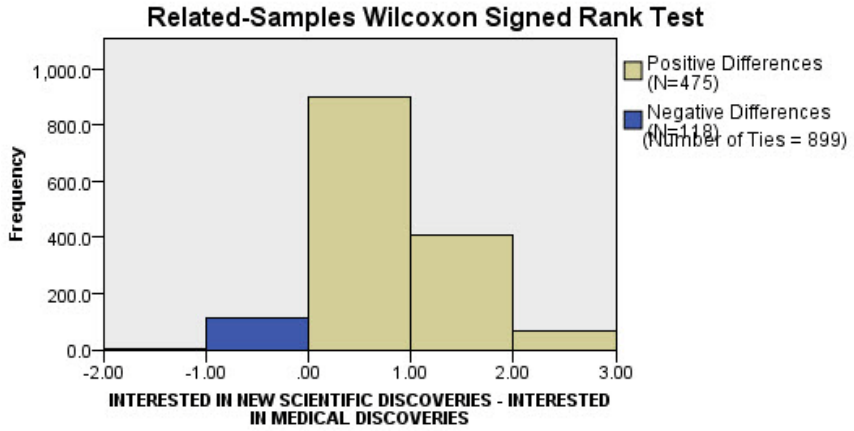
	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between INTERESTED IN MEDICAL DISCOVERIES and INTERESTED IN NEW SCIENTIFIC DISCOVERIES equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

To see the distributions of the differences between these two variables, we need to open the Model viewer and look in the Auxiliary pane.

The bar chart of differences varies from –2 to 2 because the variables are coded from 1 to 3, so the difference can vary from (3-1) to (1-3). It is clear that there are more differences in one direction than another (the actual direction depends on how the variables are coded and the order of variables in the dialog box). It is the number of positive and negative differences that is used to calculate the test statistic.

Figure 12.18 Distribution of Differences between Interest in Medicine and Science



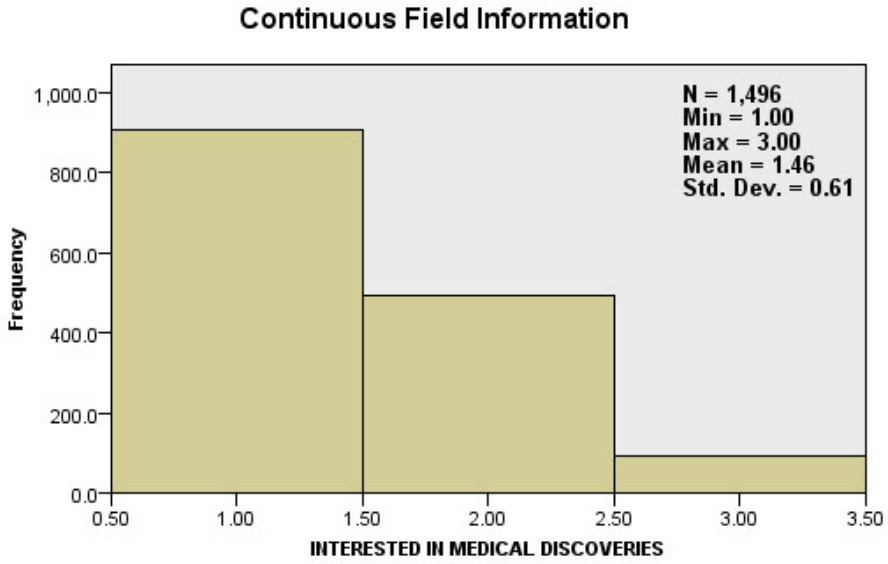
Total N	1,492
Test Statistic	143,899.500
Standard Error	3,805.769
Standardized Test Statistic	14.672
Asymptotic Sig. (2-sided test)	.000

The actual output doesn't tell us which variable has a higher level of interest, which is why we viewed the frequency tables first. We now know that there is more interest in medical discoveries than scientific discoveries.

The distribution of the each variable can be viewed within the Model Viewer, though.

1) Select **Continuous Field Information** from the View dropdown
2) Select **INTERESTED IN MEDICAL DISCOVERIES** from the Field(s): dropdown

Figure 12.19 Distribution of Interest in Medical Discoveries



Apply Your Knowledge

1. Would you use a related -samples nonparametric test in the following situations? Select all that apply.
    a. When related variables are not truly interval/scale in measurement.
    b. When we want to compare two groups of respondents
    c. When we want to compare three ordinal variables measured on the same response scale

Additional Resources

Further Info

For additional information on nonparametric tests, see:

Daniel, Wayne W. 2000. *Applied Nonparametric Statistics.* 2nd ed. Boston: Duxbury Press.

Gibbons, Jean D. 2005. *Nonparametric Measures of Association*. Newbury Park, CA: Sage.

12.14 *Lesson Summary*

We demonstrated the use of the Nonparametric Tests procedure in this lesson for data that don't meet the assumptions for parametric tests.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Perform non-parametric tests on data that don't meet the assumptions for standard statistical tests

To support the achievement of this primary objective, students should now also be able to:

- Describe when non-parametric tests should and can be used
- Describe the options in the Nonparametric procedure dialog box and tabs
- Interpret the results of several types of nonparametric tests

12.15 *Learning Activity*

The overall goal of this learning activity is to use nonparametric tests to explore the relationship between several variables, using the data file *SPSS_CUST.SAV*.

Supporting Materials

The SPSS customer satisfaction data file *SPSS_CUST.SAV*. This data file was collected from a random sample of SPSS customers asking about their satisfaction with the software, service, and other features, and some background information on the customer and their company.

1. Most of the questions asking for customer evaluation of the software and service are measured on a five-point scale from Strongly Agree to Strongly Disagree (lower values are more agreement, equivalent to higher satisfaction). Review the data file.

2. Test whether overall customer satisfaction (*satcust*) is different by highest degree earned (*degree*). Be sure to code *satcust* as Scale beforehand. Try both the Kruskal-Wallis test for independent samples and the Median test. What do you conclude? Are the test results consistent?

3. Look at the pairwise results for each test. Which degree groups are more satisfied overall? Are the pairwise results consistent? If not, how do they differ? Is there something odd about the pairwise results for the Kruskal-Wallis test?

4. To help think further about these results, use Crosstabs to request a table of *satcust* by *degree*. Request appropriate percentages and a chi-square test. Is this analysis consistent with the nonparametric analysis?

5. *For those with more time:* Temporarily remove the five respondents with only some high school education and rerun the nonparametric tests. Does this change any of the results?

6. Two questions on the survey ask whether SPSS products are easy to learn (*easylrn*) and easy to use (*easyuse*). Test whether customers think that products are easier to learn than use, or vice-versa with a related samples test. Use both the related-samples Wilcoxon signed rank test and the related-samples Friedman test. Are the results consistent for both tests? What do you conclude?

Lesson 13: Course Summary

13.1 *Course Objectives Review*

Now that you have completed the course, you should be able to:

- Perform basic statistical analysis using selected statistical techniques with PASW Statistics

And you should be able to:

- Explain the basic elements of quantitative research and issues that should be considered in data analysis
- Determine the level of measurement of variables and obtain appropriate summary statistics based on the level of measurement
- Run the Frequencies procedure to obtain appropriate summary statistics for categorical variables
- Request and interpret appropriate summary statistics for scale variables
- Explain how to make inferences about populations from samples
- Perform crosstab analysis on categorical variables
- Perform a statistical test to determine whether there is a statistically significant relationship between categorical variables
- Perform a statistical test to determine whether there is a statistically significant difference between two groups on a scale variable
- Perform a statistical test to determine whether there is a statistically significant difference between the means of two scale variables
- Perform a statistical test to determine whether there is a statistically significant difference among three or more groups on a scale dependent variable
- Perform a statistical test to determine whether two scale variables are correlated (related)
- Perform linear regression to determine whether one or more variables can significantly predict or explain a dependent variable
- Perform non-parametric tests on data that don't meet the assumptions for standard statistical tests

13.2 *Course Review: Discussion Questions*

1. Is there a "correct" order to the steps for a statistical analysis?
2. What factors would help you to decide whether to use a table or a chart to report on an analysis? Or to use both?
3. Which procedure do you prefer to analyze scale variables, **Frequencies** or **Descriptives**?
4. When should you use nonparametric methods of analysis?
5. What should you do when a relationship that is substantively significant is not statistically significant?

13.3 *Next Steps*

Thought Starters

How might you use Paired T Tests in analyzing data for your organization?
How might you use ANOVA in data analysis?

How might you use Regression in data analysis?

Next Courses

This course discussed many statistical techniques. In this section we provide direction for what courses you can attend to broaden your knowledge in specific areas.

If you want to learn more about this:	Take this course:
Regression and related methods	Advanced Techniques: Regression
ANOVA and related methods	Advanced Techniques: ANOVA
A variety of advanced statistical methods	Advanced Statistical Analysis Using PASW Statistics

Appendix A: Introduction to Statistical Analysis References

1.1 *Introduction*

This appendix lists only references that cover several of the techniques and statistical methods discussed in this course. Some of these references, such as the book by Field, also cover advanced statistics. More detailed references for specific techniques are included in some of the lessons.

1.2 *References*

INTRODUCTORY AND INTERMEDIATE STATISTICS TEXTS

Berenson, Mark L., Timothy C. Krehbiel, David M. Levine. 2005. *Basic Business Statistics: Concepts and Applications.* 10th ed. New York: Prentice Hall.

Burns, Robert P and Burns, Richard. 2008 . *Business Research Methods and Statistics Using SPSS.* London: Sage Publications Ltd.

Field, Andy. 2009. *Discovering Statistics Using SPSS*. 3rd ed. London: Sage Publications Ltd.

Knoke, David, Bohrnstedt, George W. and Mee, Alisa Potter. 2002. *Statistics for Social Data Analysis*. 4th ed. Wadsworth Publishing.

Norusis, Marija J. 2009. *SPSS 17.0 Guide to Data Analysis*. New York: Prentice-Hall.

Norusis, Marija J. 2010. *PASW Statistics 18.0 Statistical Procedures Companion*. (forthcoming) New York: Prentice-Hall.