DOUGLAS BORS

DATA ANALYSIS FOR THE SOCIAL SCIENCES

Integrating Theory and Practice

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Jai Seaman
Assistant editor: Alysha Owen
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Proofreader:
Indexer:
Marketing manager:
Cover design:
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

First published 2018

Library of Congress Control Number: 00000000

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-4462-9847-3
ISBN 978-1-4462-9848-0 (pbk)

2

Chapter contents

DESCRIPTIVE STATISTICS

KEY CONCEPTS: scales of measurement, measurement and frequency data, frequency histograms, measures of central tendency, measures of spread, outliers.

2.1 PURPOSE

Once researchers have collected their data, the first step in the analysis process is to summarize the observations, both pictorially and numerically. The purpose of this chapter is to examine the most commonly employed graphs and statistics for summarizing different types of data. The focus is on graphs and statistics used to describe individual variables: *univariate* graphs and statistics. We begin by examining the types of numbers researchers use to record their data. As will be seen, a 2 is not necessarily 1 plus 1. Then we introduce the notation system used throughout this book. Following that is a presentation of the most common pictorial forms researchers use to summarize their data, focusing on the histogram. Next is an examination of the numerical descriptions used for summarizing different types of data: mean, median, mode, variance, and index of qualitative variation. Throughout the chapter the instructions for the relevant SPSS procedures are illustrated.

2.2 INTRODUCTION

Below are the hypothetical quiz scores from two sections of a statistics course. Are you able to scan them and summarize the performances of the two sections?

> Section 1: 7, 8, 5, 5, 9, 10, 2, 7, 6, 4, 8, 7, 8, 7, 3, 9, 6, 8, 6, 9, 8, 9, 5, 7, 6, 7, 6, 7, 9, 5, 5, 9, 7, 11, 7, 6, 4, 7, 8, 10
>
> Section 2: 6, 3, 6, 8, 5, 7, 6, 7, 8, 7, 8, 5, 5, 8, 7, 2, 7, 6, 4, 7, 7, 8, 5, 5, 8, 10, 1, 7, 6, 4, 8, 7, 2, 7, 3, 4, 6, 8, 10, 8

If someone began reading aloud the litany of individual scores, would you have a sense of how well the two sections performed or how they performed comparatively? Even with this relatively small data set it is virtually impossible. Therefore, the first step in any analysis is to describe and summarize the data. A description needs to be more than a listing of all of the observations, however. A listing of the quiz scores tells us how each student performed on the quiz, but it tells us little of a section's overall performance or of how the two sections compare. Researchers need ways of summarizing the individual observations without distorting the data's overall structure, such as average, and without losing too much information, such as the important details. Some descriptive statistics form the basis for testing important assumptions necessary for answering *inferential* questions, such as whether there are any reliable differences between the two sections. *Reliable* in this case means that any observed difference between the two sections is likely *not* due to chance alone.

Any summary comes at a cost. Think of the last time you tried to summarize a novel or a film for a friend. While you were attempting to capture the plot you needed to decide which details to ignore and which to include. A statistical summary faces the same challenge. With the appropriate use of graphs and descriptive statistics the overall quantitative and qualitative character of the observations can be captured, with minimal distortion and loss of vital information.

2.3 NUMERICAL SCALES

Statistics is all about numbers: the manipulation of numbers and their comparison. A single number, such as an interest rate, by itself means very little. Furthermore, not all numerical calculations are appropriate in all circumstances. Appropriateness depends upon the nature of the numbers used to record the observations. The differences in the nature of numbers are often related to four different scales of measurement. These scales were first articulated by Stevens (1946) when he defined measurement as the assignment of numbers to objects according to rules. Today we describe these scales in terms of four properties.

1. Numbers on a *nominal scale* merely name categories of observations and have no intrinsic value.
2. In addition to providing a name, numbers on an *ordinal scale* have the property of ordering categories in terms of 'more or less'.
3. In addition to the property of ordering, numbers on an *interval scale* have the property of equal size intervals between the adjacent categories or numbers.
4. Numbers on a *ratio scale* have all three of the previous properties – naming, ordering, and equal intervals – plus the additional property of a true zero.

Nominal scales

The word *nominal*, which comes from the Latin word *nominalis*, refers to the property of naming. Numbers on a nominal scale name categories of observations that are *mutually exclusive*. For categories to be mutually exclusive no single observation can be a member of more than one. The simplest example of mutually exclusive categories is that of heads versus tails. A coin toss will either come up heads or tails. The principle is simple. If it is heads, it cannot be tails. Mutual exclusiveness means that if the observation is a member of one category in a set of categories, it necessarily *excludes* that observation from being a member of any other category in that set. Mutual exclusiveness does not recognize dual citizenship or fusion cuisine. We will discuss the notion of mutually exclusive categories in more detail in Chapter 3.

All four scales contain the property of naming. The key is that all the observations will fall into one and only one of a scale's categories. Numbers on a nominal scale are only names and are completely arbitrary. If we were researching music preferences, we might assign a '1' to those who prefer jazz, a '2' to those who prefer rock, and a '3' to those who prefer classical music. Or we could capriciously change our mind and assign a '1' to those who prefer classical music, a '2' to those who prefer jazz, and a '3' to those who prefer rock. The 1, 2, and 3 do not refer to amounts or rankings. They are only names. Thus, for most purposes, the actual numbers are irrelevant. A '–0.0000017' could be assigned to those who prefer jazz, a '1.000001' to those who prefer rock, and a '9' to those who prefer classical music. In fact, numbers (names) can be randomly assigned to each category.

The number used for your car's licence plate and the number on a baseball jersey are everyday examples of numbers used only as names. We do not produce player '13' by adding players '8' and '5'.

Because the numbers on a nominal scale are arbitrary, when analysing such data we are not interested in the values assigned to the categories, rather, we are interested in the *number of instances* or *relative frequencies* we observe in each category. It would make no sense to add, subtract, multiply, or divide such arbitrarily assigned names any more than it would be to subtract Douglas from Alexander. Some exceptions to this limitation can be made, as we will see later in the book.

Ordinal scales

The word *ordinal*, which comes from the Latin word *ordinalis*, indicates that in addition to the number providing a name, there is an underlying *order* to the numbers. The property of order relates to the comparative relationship of 'more or less'. 'More' and 'less' are only qualitative comparisons, the quantitative extent of the comparison is unknown and likely changeable. One example of an ordinal scale with which many students are familiar is the assigning of numerical grades: 1, 2, 3, 4, and 5. These numerical grades often indicate failure, barely passing, average, above average, and excellent, respectively. The numbers are names, but they also indicate a relation of 'more or less' in terms of a student's performance: the larger the number the better the performance. We know that a 5 is greater than a 4 and a 4 is greater than a 3, but we do not know if the difference between a 5 and 4 is the same as the difference between a 4 and 3. Nominally (in name) – pun intended – both differences are 1, but are the two 1s quantitatively equal? The values assigned to each category are in part arbitrary. We could have just as easily used –2, –1, 0, 1 and 2 for our grading scale, as long as the transformation maintains the order of performance, so that the –2 represents failure, the –1 represents barely passing, and so on.

Furthermore, because there is no assurance that the difference between failure and barely passing is equal to the difference between barely passing and average performance, the values chosen do not need to be numerically adjacent, as long as they reflect an underlying rank ordering. For example, the numbers could be –1, 2, 10, 45, and 100, as long as that order indicates failure, barely passing, average, above average, and excellent performance, respectively.

Ordinal number scales allow comparisons of 'more and less' to be made. Because the *apparent equal differences* in the numbers are likely *quantitatively unequal*, however, the numbers on an ordinal scale cannot be added, subtracted, multiplied, or divided. We have only a qualitative grasp of the relations between the named categories. Similarly to when analysing nominal data, when analysing ordinal data the researcher is not particularly interested in the values assigned to the categories, but rather in the number of instances observed in each category.

Rankings, such as league tables or standings, are examples of ordinal scales. Is the difference between the top team and the second placed team the same as the difference between the second placed team and the third placed team? Perhaps it is, but not necessarily. How could we know? If there were 20 teams in the league, would the team at the top be 5% better than the second placed team, or be 95% better than the team at the bottom? We may manipulate the numbers, but any such conclusions would have little validity. What about the rankings of hockey, tennis, or soccer players? Are the differences between adjacently ranked players all equal?

Interval scales

The word *interval* comes from the Latin word *intervallum*. During the Middle Ages this referred to the spaces between the ramparts on castle walls (Figure 2.1), which tended to be of equal width.

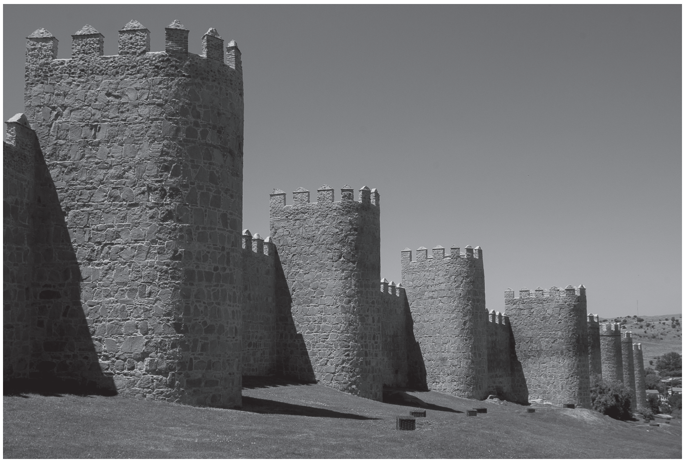


Figure 2.1

In terms of measurement scales, the term 'interval' refers to the distance between numbers. Like the nominal and ordinal scales, interval scales name mutually exclusive categories. And like ordinal scales, numbers on interval scales indicate comparative relations of 'more or less' – the larger the number, the more of something. The additional property associated with interval scales is that of *equal intervals* between numbers such that one difference of 3 (6 – 3) is the same as any other difference of 3 (99 – 96). In this case, a rose is always a rose.

The property of equal intervals allows for addition and subtraction as well as relevant comparisons, but it allows for only some forms of multiplication. A common example of an interval scale with which we are all familiar is temperature. The difference between –30° and 0° Celsius is equal to the difference between 0° and +30° Celsius. We need to be careful, however. If the temperature yesterday was 10°C and today it is 20°C, it does not mean that today it is twice as warm as it was yesterday: 20/10 = 2. To illustrate this, convert the temperatures from Celsius to Fahrenheit: F = C(9/5) + 32. The two temperatures become 50°F yesterday and 68°F today. It no longer appears to be twice as warm today as it was yesterday: 68/50 = 1.36.

PERSONALITY AND IQ

Some say that true interval scales in the behavioural and biological sciences are rare. Consequently, care must be taken. Researchers typically treat scores on personality and IQ tests as if they were interval in nature. Such scores are not the number of items answered correctly or answered in a particular manner. Performance on such tests is reported as *standardized scores*,

(Continued)

and it is unlikely that the intervals are equal. For example, the difference between an IQ of 95 and an IQ of 100 is unlikely to represent the same difference in intellectual ability as the difference between an IQ of 120 and 125. Many would argue that scores on personality and IQ tests are best considered ordinal in nature. Treating ordinal scores as interval in nature will result in distortions when inappropriate manipulations are performed. How much distortion will result? It depends, if the intervals between consecutive scores vary greatly, there will be considerable distortion. If the intervals vary only slightly, then the distortion will be minimal. It is common, particularly in educational research, to convert percentile score, which are ordinal in nature, into something called *normal curve equivalent* scores, which will be presented in Chapter 3.

Ratio scales

Although intervals may be equal within a scale, they may not be equal across scales. Also notice that the zero on the Celsius scales is not a zero on the Fahrenheit scale.

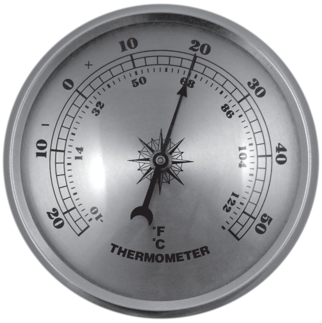


Figure 2.2

The word *ratio* comes from the Latin verb *reri*, refers to think, reckon, or calculate. A ratio scale has all three of the previously mentioned properties – naming, ordering, and equal intervals – plus the additional property of a true zero. It is the presence of a true zero that allows meaningful ratios to be *calculated*. Unlike the case with interval scales, when there is a true zero, the units of measurement can be converted, but any computed ratios will be equivalent. Time is a good example. The ratio of 2 minutes to 1 minute is 2/1 = 2. If we convert the minutes into seconds the ratio is unchanged: 120/60 = 2. In experimental areas of psychology, because they are ratio in nature, *reaction time* (RT) or latency and *accuracy* have been the two most common *dependent variables*. RT is the time it takes someone to complete a task, be it simple or complex. Accuracy is the number of correct responses observed over a period of time. Both have true zeros.

Some researchers have treated personality, IQ, and social surveys as if they represented ratio scales. Treating these measures as such will result in distortions and possible erroneous conclusions when some statistical analyses are performed. How much distortion? Again, it depends. If the data have little resemblance to a ratio scale, there will be considerable distortion in any calculation. The closer the data approximate a ratio scale, the less the distortion.

The time it takes runners to complete the 100 metre dash is one obvious example of a ratio scale. Scores (times) can be converted into ratios. Someone can be twice as fast as someone else.

REVIEW QUESTION

With which of the four scales of measurement do you associate the following measures? Using the four properties, explain why.

1 Your year of birth.
2 Your marks in secondary school.
3 Your shoe size and width.
4 The number of the row in which you sat during the last lecture (from front to back).
5 Your income last month.

Web Link 2.1 for a discussion of possible responses.

As stated at the outset, this chapter surveys the graphs and statistics that researchers use to summarize and describe individual variables: *univariate* statistics and graphs. The term 'data' usually refers to the set of individual observations the researcher collects with respect to the *dependent* or *criterion* variable. [Insert Page cross reference for a review of the distinction between dependent and criterion variables in the previous chapter].

REVIEW QUESTION

What is the difference between an experiment and a quasi-experiment?

Web Link 2.2 for an answer to the review question.

Other ways of categorizing data

As mentioned in the previous chapter, although there are four scales of measurement, researchers discuss data as being one of two types: measurement or categorical. Measurement data are sometimes called *quantitative* data, and categorical data are sometimes called *qualitative* or frequency data. With respect to measurement data, each of a researcher's observations can result in a unique value. Measurement data are associated with interval and ratio scales and they can be either *continuous* or *discrete*. Reaction time (RT) is a continuous variable. It is continuous in the sense that there are an infinite number of values between any two RTs. There are an infinite number of units of time between 1 and 2 seconds just as there are an infinite number of units between 1 and 2 milliseconds. The number of items correctly answered on a multiple-choice examination is an example of a discrete variable. It is discrete because there are adjacent values which have no values between them. On a true-or-false test, a student answers either 21 or 22 items correctly. There are no scores of 21.5.

Continuous variables, like sliding boards, are smooth from top to bottom. Discrete variables have a fixed number of steps, like staircases.

Categorical data, usually associated with nominal and ordinal scales, by their nature are discrete. When examining frequency data the focus is not on the particular values (names) assigned to the categories of the observations, but rather on the number of observations (frequencies) observed in each category. As we will see, there are different descriptive statistics associated with these different types of data.

2.3 NOTATION

At this point it is necessary to begin introducing the notation system used throughout this book. When computing statistics we often carry out simple computations iteratively. That is, we repeat a procedural step or computation. Often a computation is performed on all observations. Rather than write out each instance, we have a system for notating the process. The Greek symbol $\sum$ (Sigma) is used for this purpose. Assume we have a data set (y) of five observations: 2, 1, 4, 5, and 3. If we wish to obtain the total of the y values, it could be written out 'add up all of the observations': $2 + 1 + 4 + 5 + 3 = 15$. Or we can simply express this as $\Sigma y = 15$. Latin letters, usually y, are used to indicate a variable. Σy indicates the sum of the y values. $\Sigma(y - C)$ instructs us to subtract a constant from each of the y values and sum the differences: $(2 - C) + (1 - C) + (4 - C) + (5 - C) + (3 - C)$. These forms of summation notation, as well as elaborations on them, will be used throughout this book. It should be noted that

$$\sum_{i=1}^{n} y_i$$

is the more formal expression of Σy, where i indicates where to begin the summing. Unless stated otherwise, begin with the first observation. n is the total number of observations and, thus, the last observation in the set. With each iteration, i is incremented by one. Because summations almost always involve all observations in a data set or in a subset, the notation is simplified by dropping the i and the n. Notational variations will be introduced as necessary.

Web Link 2.3 for a more comprehensive review of summation notation and related rules.

In addition to summation notation, there is a need to clarify how letters will be used to symbolize variables and statistics. Figure 2.3 summarizes how some of the Latin and Greek letters are used in the text.

For example, Latin letters are used to indicate variables. The letter y usually is used for dependent or criterion variables and the letter x for independent or predictor variables. Most of the letters in Figure 2.3 are introduced in this chapter, the others appear in subsequent chapters. Statistics are said to be *descriptive* when they are used to merely describe a set of observations. Statistics are said to be *inferential* when they are used to make inferences about populations and their parameters or to test hypotheses. The reporting of all research results begins with the presentation of descriptive statistics.

Type of letter		Function	Data type
Roman			
	y	Dependent variable or Criterion	Measurement
	x	Independent variable or Predictor[1]	Measurement
	$\bar{Y}$, S, S^2, r	Sample descriptive statistics	Measurement
	VR, IQV	Sample descriptive statistics	Categorical
	t, F	Inferential test statistic	Measurement
Greek			
	μ, σ, σ^2, ρ	Population parameters	Measurement
	ϕ	Sample descriptive statistic	Categorical
	χ^2, λ, τ	Inferential test statistics	Categorical

Figure 2.3 Some letters used as symbols

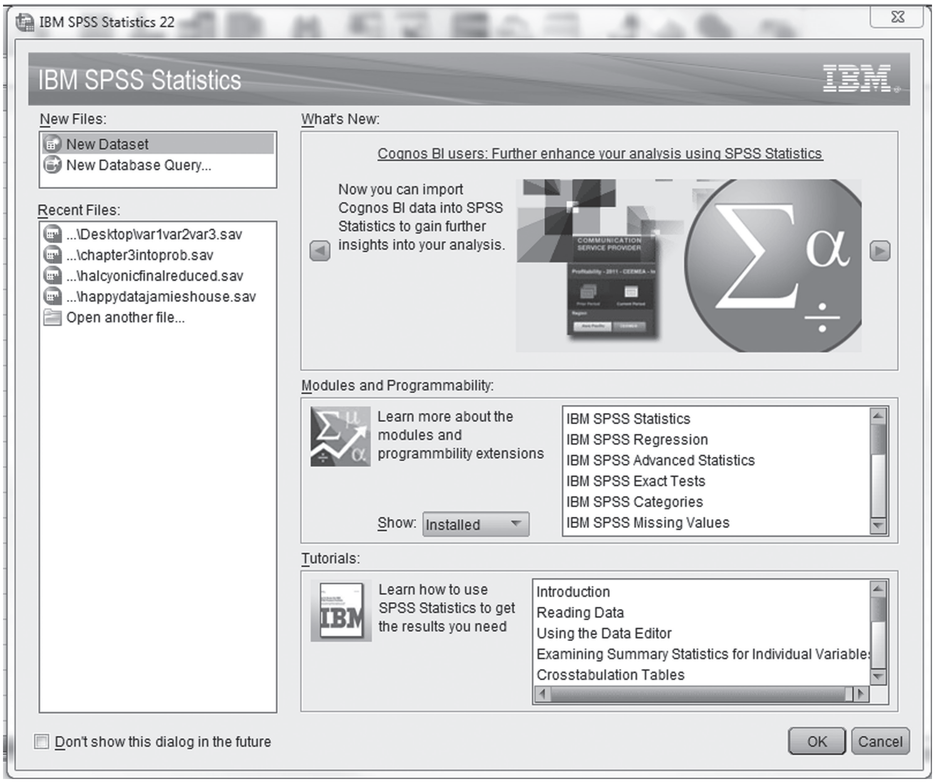
2.4 HISTOGRAMS

Entering data

Although there are many ways to pictorially summarize data, the histogram is the most common univariate graph. One reason for its wide use is that it provides a visual summary of a variable with a minimal loss of information. Furthermore, the histogram allows for easy identification of any anomalies that may need to be addressed. As we will see, anomalies include such things as a problematic shape of the distribution of the observations or an observation that is considerably greater or smaller than all others. Problematic shapes and extreme scores will restrict the type of statistical analysis that may appropriately be applied. Some of the ways to correct these anomalies are discussed at points throughout the book.

Histograms can be appropriately applied to most forms of data, including measurement, categorical, discrete, and continuous data. Histograms are often employed to summarize the dependent variable in experimental studies and to depict both criterion and predictor variables in observational research. Let us examine the data given in Section 2.2 above.

Begin by entering the data into the SPSS *Data Editor*. Go to this *SPSS* link at https://study.sagepub.com/bors **. Once SPSS opens you will see a window labelled *IBM SPSS STATISTICS* (Figure 2.4). In the small white box below the label you will see an option called *New Dataset*. Select that option and then click on the *OK* button at the bottom right of the window.**



Now you can import Cognos BI data into SPSS Statistics to gain further insights into your analysis.

Learn more about the modules and programmbility extensions

Learn how to use SPSS Statistics to get the results you need

Figure 2.4

A new window labelled *IBM SPSS Statistics Data Editor* appears with a row of menu tabs and a matrix of empty boxes (Figure 2.5). The columns are labelled *var* for 'variable'. The rows are labelled 1–39 to begin with. The rows usually represent subjects and the numbers in the first column often serve as subject numbers.

At the bottom left of the screen there are two tabs: *Data View* and *Variable View*. We are currently in the *Data View* window. By clicking the *Variable View* tab we can switch to the *Variable View* window and name our variable and define our *variable type*. When the window opens, move the cursor to the *Name* cell in the first row and type our variable's name. Let us use *section1* as the name (Figure 2.6). Notice that when you hit return other cells in the row

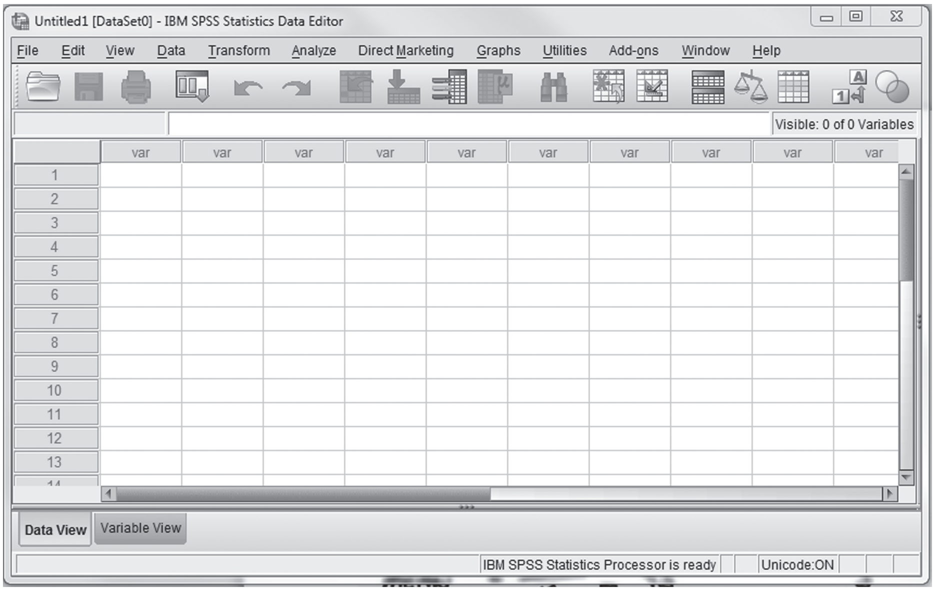
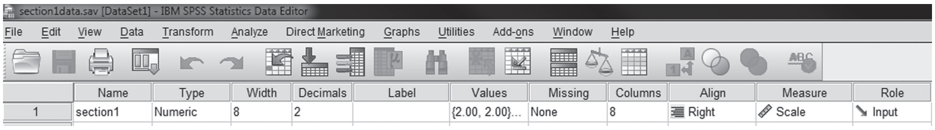

Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var var var var var var var var var var
1 2 3 4 5 6 7 8 9 10 11 12 13
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Figure 2.5


section1data.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

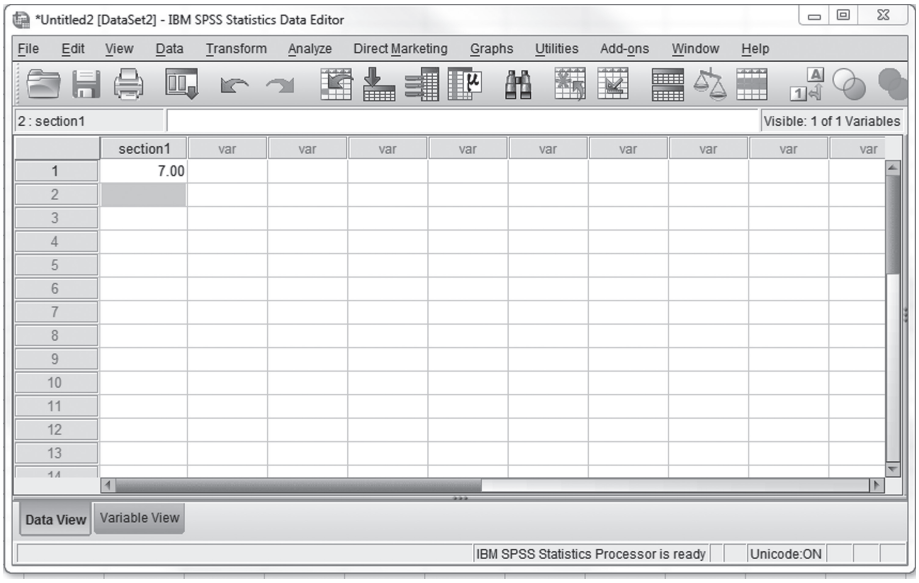
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	section1	Numeric	8	2		{2.00, 2.00}...	None	8	Right	Scale	Input

Figure 2.6

are automatically filled in. Ensure that the variable is listed as *Numeric* under the *Type* heading and as *Scale* under *Measure* heading. Although the number of items correct on the quiz is an example of a ratio scale, SPSS does not differentiate ratio and interval scales. The term *Scale* under the heading *Type* is used for either ratio or interval data. For now we will use the other default settings.

Because we are only exploring one variable, click the *Data View* tab and return to the *Data Editor* window (Figure 2.7). Move the cursor to the cell immediately under the variable labelled *section1* and enter the first score. After entering a 7 in the first row either arrow down or hit the enter key to move to the second row. Continue doing so until all 40 observations for *section1* are entered in the first column of the matrix.

Once the data are entered, we can construct a *frequency table*. From the menu row in the *Data Editor* click *Analyze*, scroll down to the *Descriptive Statistics* option and then over to *Frequencies*. When the *Frequencies* window appears, highlight and move (using the arrow between the boxes) *section1* over to the *Variable(s)* box (Figure 2.8). Ensure the *Display frequency table* box in the bottom left corner is checked. Then click *OK*. The *Statistics Viewer* or output window will appear (Figure 2.9).



2 : section1

Visible: 1 of 1 Variables

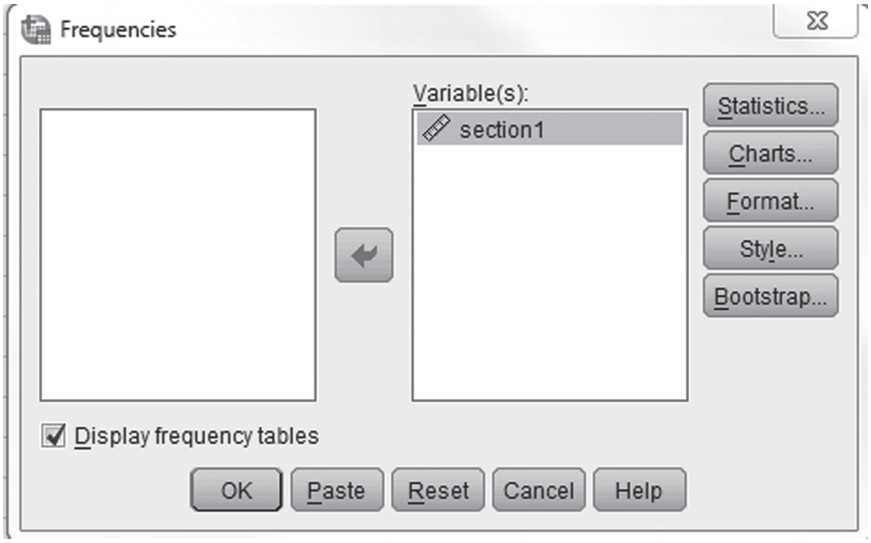
	section1	var	var	var	var	var	var	var	var	var
1	7.00									
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										

IBM SPSS Statistics Processor is ready

Unicode:ON

Figure 2.7

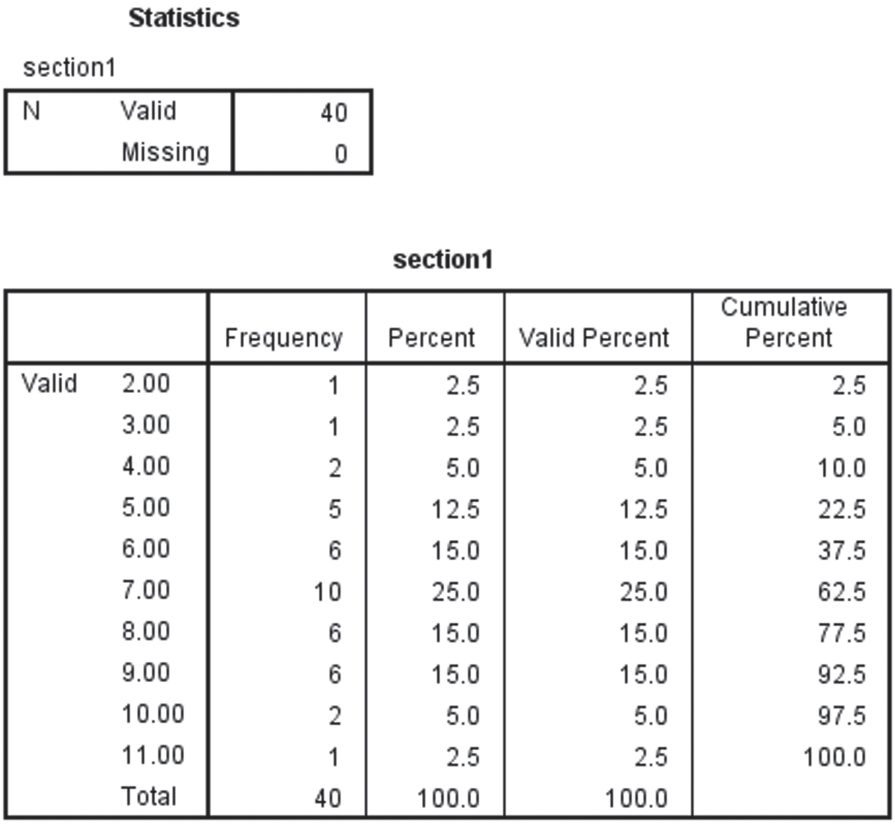
The small box at the top of the window reports the number of observations analysed (*N*) and the number of subjects with missing data. We have 40 observations (subjects) and no missing data. Below that in the output window is a table with five columns. The first column lists the categories of the observations; we had scores ranging from 2 to 11. The second column indicates



Variable(s):
section1

☑ Display frequency tables

Figure 2.8



Statistics

section1

N	Valid	40
	Missing	0

section1

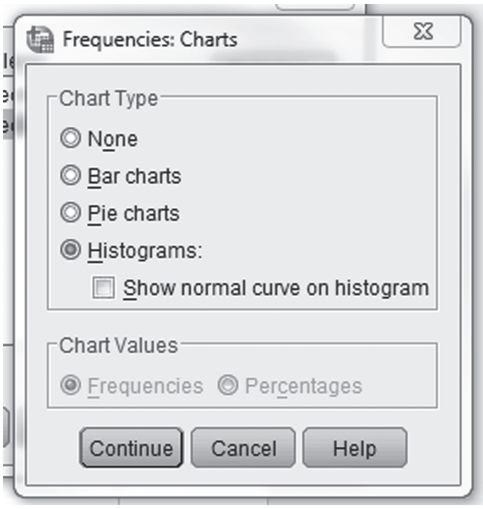
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	2.5	2.5	2.5
	3.00	1	2.5	2.5	5.0
	4.00	2	5.0	5.0	10.0
	5.00	5	12.5	12.5	22.5
	6.00	6	15.0	15.0	37.5
	7.00	10	25.0	25.0	62.5
	8.00	6	15.0	15.0	77.5
	9.00	6	15.0	15.0	92.5
	10.00	2	5.0	5.0	97.5
	11.00	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Figure 2.9

the frequency or the number of observations found in each category. The third column reports the percentage of the observations associated with each score or category. We will ignore the fourth column for now. The fifth column provides the percentage as it accumulates from the lowest score to the highest score.

Although a frequency table provides important summary information, such as the lowest and highest scores as well as the frequencies for each of the scores, researchers in many areas often prefer to view frequencies in pictorial form: a histogram.

To produce a *frequency histogram*, return to the *Frequencies* window and select the *Charts* tab. Click the *Histogram* button (Figure 2.10)



Frequencies: Charts

Chart Type

○ None
○ Bar charts
○ Pie charts
● Histograms:
□ Show normal curve on histogram

Chart Values

● Frequencies ○ Percentages

Figure 2.10

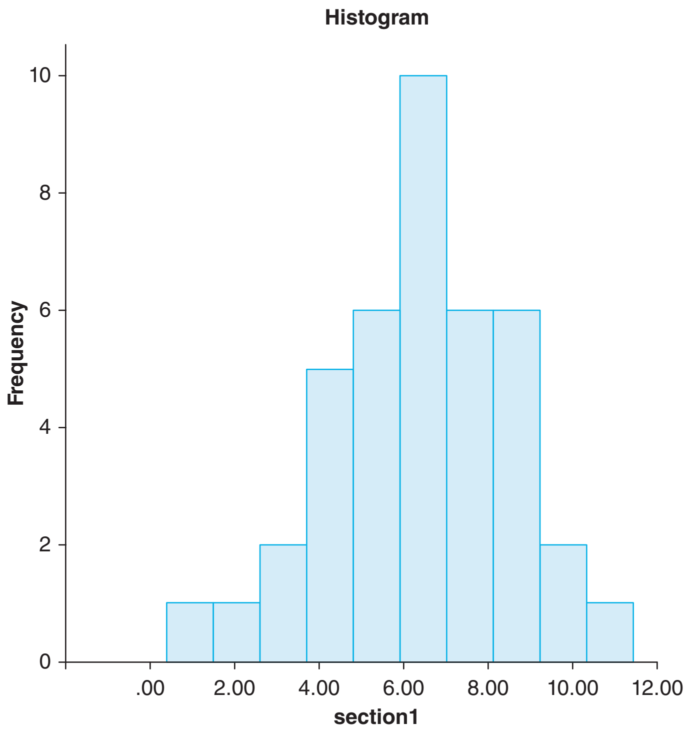


Figure 2.11

and then *Continue*. When you return to the *Frequencies* window, click *OK*.

In addition to the previous output, you now will find a frequency histogram (Figure 2.11). On the horizontal axis, or *x*-axis, the observed scores or categories are listed in ascending order. The vertical or *y*-axis represents the frequency or number of observations. For example, we can see from the histogram that there were two scores of 4 and five scores of 5. Summing the frequency of all of the categories gives us the total number of observations. Most of the information from the frequency table is contained in the frequency histogram. The percentages and cumulative percentages would need to be calculated, however. SPSS does provide histograms with that information. As an example, we will construct a *relative frequency histogram*. For that we need to switch to a different option in the *Data Editor* window.

In the *Data Editor* window we move over from *Analyze* to the *Graphs* option and scroll down and select *Chart Builder*. For our purposes, when the *Chart Builder* window appears (Figure 2.12) click *OK*.

Two windows will simultaneously appear: a second *Chart Builder* window and an *Element Properties* window (Figure 2.13). From the bottom left of the *Chart Builder* window select the *Histogram* option and then drag the leftmost of the four examples (*Simple Histogram*) into the large empty rectangular box. Next select and drag the variable name (*section1*) from the *Variables* box into the *X-Axis* field in the box to the right. Next go back into the *Element Properties* window in the *Chart Builder* and click the little down arrow that appears across from *Histogram* under *Statistic*. Choose the *Histogram Percent* option and at the bottom of the window click *Apply*. Move over to the *Chart Builder* window and click *OK*.

An alternative version of the frequency histogram appears in the output window: a relative frequency histogram. These two histograms look identical. There is only one difference. Notice that the label of the *y*-axis has been changed from *Frequency* to *Frequency Percent*. The current histogram provides the percentage of the total number of observations represented by each category on the *x*-axis. With a relative frequency histogram, an important piece of information

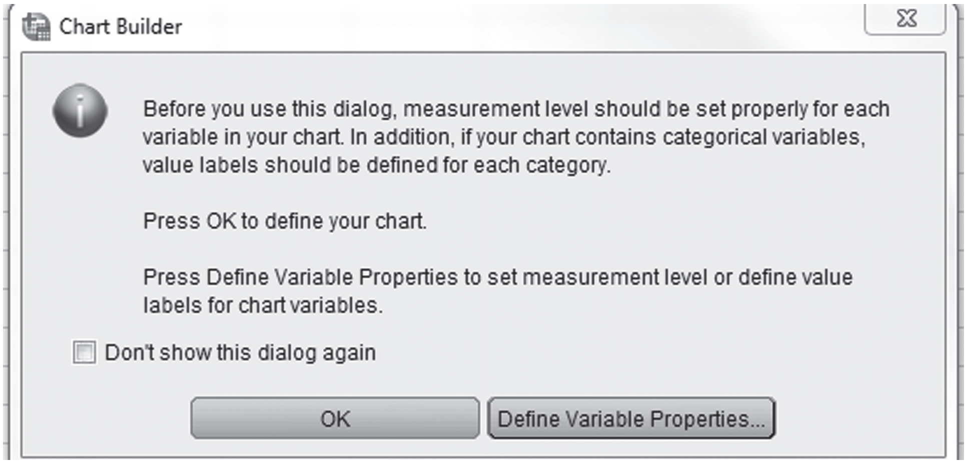


Chart Builder

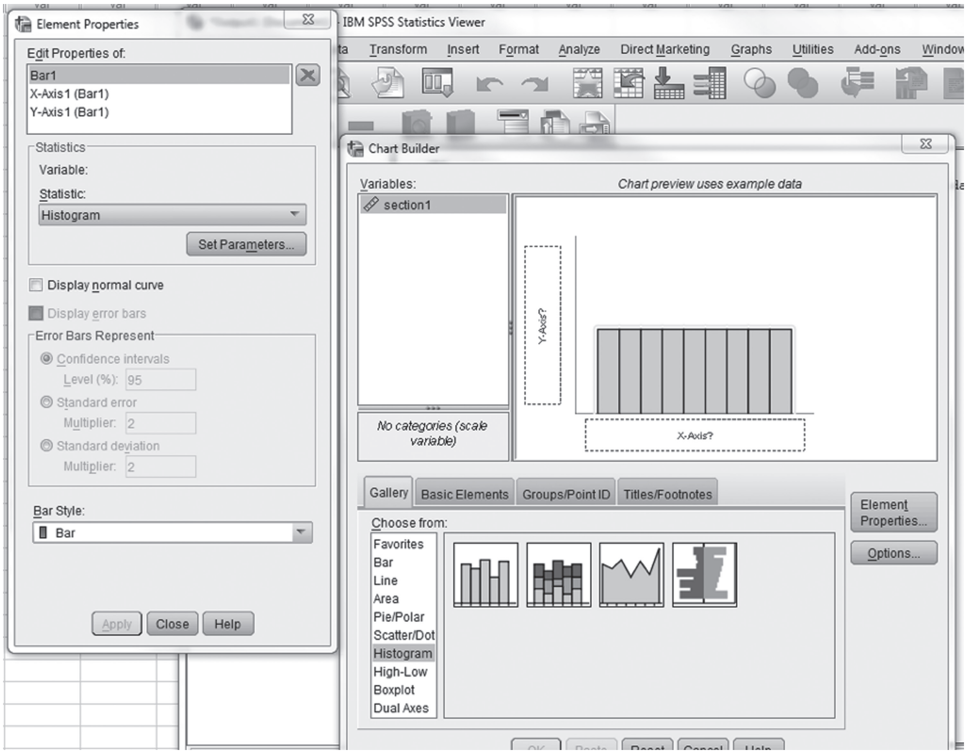
Before you use this dialog, measurement level should be set properly for each variable in your chart. In addition, if your chart contains categorical variables, value labels should be defined for each category.

Press OK to define your chart.

Press Define Variable Properties to set measurement level or define value labels for chart variables.

□ Don't show this dialog again

Figure 2.12



Element Properties

Edit Properties of:
Bar1
X-Axis1 (Bar1)
Y-Axis1 (Bar1)

Statistics
Variable:
Statistic: Histogram

□ Display normal curve
■ Display error bars

Error Bars Represent
● Confidence intervals
Level (%): 95
○ Standard error
Multiplier: 2
○ Standard deviation
Multiplier: 2

Bar Style: Bar

Chart Builder

Variables: section1

Chart preview uses example data

No categories (scale variable)

Choose from: Favorites, Bar, Line, Area, Pie/Polar, Scatter/Dot, Histogram, High-Low, Boxplot, Dual Axes

Figure 2.13

is lost. If I tell you that 20% of the people surveyed preferred coffee over tea, you have no way of knowing, unless I tell you, how many people were surveyed. I may have surveyed only 10 people or I may have surveyed 10,000.

Remember that an important factor when we evaluate the reliability of numerical information is *sample size*, or the number of observations upon which a summary is based. My data may indicate that 50% of Americans surveyed earned over $5 million a year. My summary, however, may have been based on only two observations: a professional football player and myself. Thus, when evaluating percentages, be they in a relative frequency histogram or merely reported in a news article, it is important to know the total number of observations upon which the percentages were calculated. Of course factors other than sample size are important. For example, an income survey may be conducted in the New York Yankees locker room or it may be conducted by randomly selecting income tax statements.

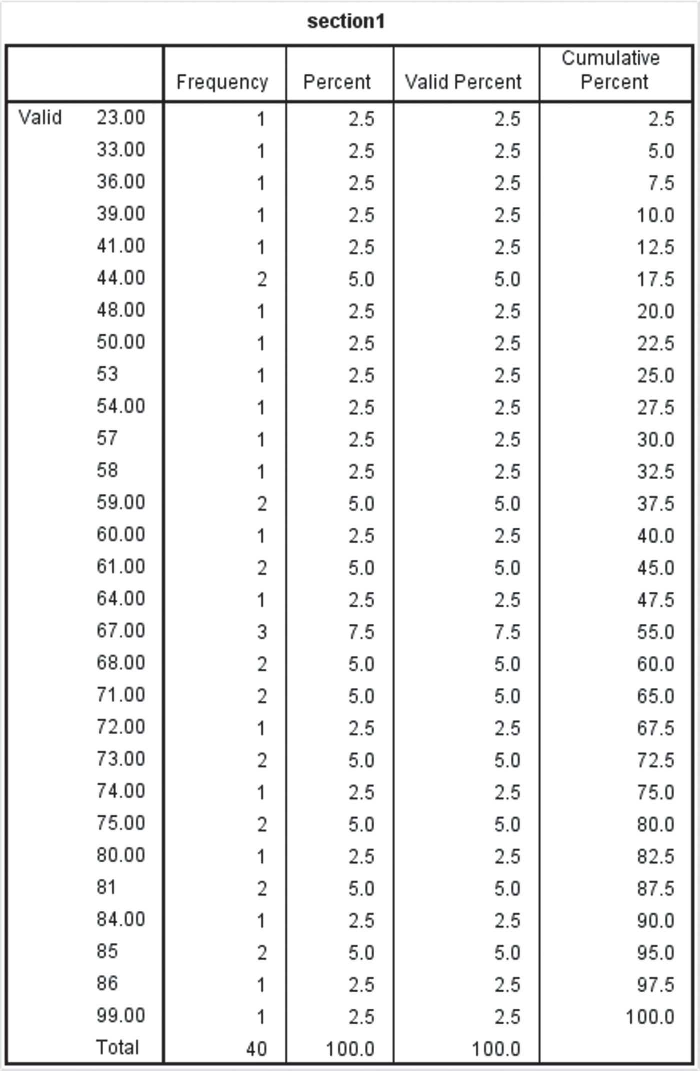
The above procedures for constructing a histogram work well when there are a limited number of observed values or categories on the *x*-axis. In the current set of discrete data there are only ten different observed values: 2 to 11. Thus there are ten categories on the *x*-axis with a number of observations in several of the categories. But what if there were considerably more categories (e.g., exam scores ranging from 23 to 99), with rarely more than one observation in any category? A glance at the hypothetical data in the frequency table in Figure 2.14 illustrates that a histogram with all observed values represented on the *x*-axis is only marginally more informative than merely listing the individual scores. Moreover, when the dependent variable is continuous in nature, such a limitation is always the case. This limitation is resolved by grouping the data into categories or *bins*.

When the data in the frequency table are categorized into ten bins and the bins are used to construct the histogram in Figure 2.15, the data are found to be more than a litany of random individual values. A pattern emerges that was not originally apparent. Like the earlier histogram when the values ranged from 2 to 11, the bulk of the observations cluster in the middle of the range, with fewer and fewer observations as we move towards the highest and the lowest scores.

Rules for constructing bins

It is clear that binning data can be quite helpful when pictorially summarizing data where there are either a large number of discrete scores or the scores are continuous in nature. Binning must follow certain rules or the resulting histogram can be misleading, however.

1 The bins should be of equal width. In the above example there are ten bins each with a width of eight. If the bins were of unequal width, the relative frequency of certain categories on the *x*-axis may be exaggerated and misinterpreted. There are justifiable exceptions. There may be instances where the bins are unequal because of the nature of the scale of interest. For example, marks in a course may be based on a final examination with scores ranging from 23 to 99. For the purpose of examining the students' performance, bins should be kept equal. If we are assigning grades 1, 2, 3, 4, and 5 to represent failure, barely passing, average performance, above average, and excellent, with scores 0–49, 50–59, 60–69, 70–79 and 80–99, respectively, then the bins for depicting the data will appear unequal: the first and last category are wider than the others. In this second case, for analytical purposes, we are changing the measurement scale from ratio to ordinal.



section1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	23.00	1	2.5	2.5	2.5
	33.00	1	2.5	2.5	5.0
	36.00	1	2.5	2.5	7.5
	39.00	1	2.5	2.5	10.0
	41.00	1	2.5	2.5	12.5
	44.00	2	5.0	5.0	17.5
	48.00	1	2.5	2.5	20.0
	50.00	1	2.5	2.5	22.5
	53	1	2.5	2.5	25.0
	54.00	1	2.5	2.5	27.5
	57	1	2.5	2.5	30.0
	58	1	2.5	2.5	32.5
	59.00	2	5.0	5.0	37.5
	60.00	1	2.5	2.5	40.0
	61.00	2	5.0	5.0	45.0
	64.00	1	2.5	2.5	47.5
	67.00	3	7.5	7.5	55.0
	68.00	2	5.0	5.0	60.0
	71.00	2	5.0	5.0	65.0
	72.00	1	2.5	2.5	67.5
	73.00	2	5.0	5.0	72.5
	74.00	1	2.5	2.5	75.0
	75.00	2	5.0	5.0	80.0
	80.00	1	2.5	2.5	82.5
	81	2	5.0	5.0	87.5
	84.00	1	2.5	2.5	90.0
	85	2	5.0	5.0	95.0
	86	1	2.5	2.5	97.5
	99.00	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Figure 2.14

2 The number of bins a researcher chooses can make a big difference to how the data appear. When the bulk of the observations are clustered in the middle of the observed values, with fewer and fewer observations further towards the highest and lowest scores, the number of bins makes little difference to the nature of the histogram. When, however, the distribution of scores diverges from that pattern, as will be seen, a change in the number of bins can make a great difference to the shape of the histogram.

3 Additionally, the appropriate number of bins is also related to sample size. The more observations upon which the histogram is based, the more categories a researcher is free to employ. If you have relatively too few observations for the number of bins, the observed shape of the histogram will be less reliable. If there are 20 observations and ten bins are used to visually summarize the data, a change in one or two

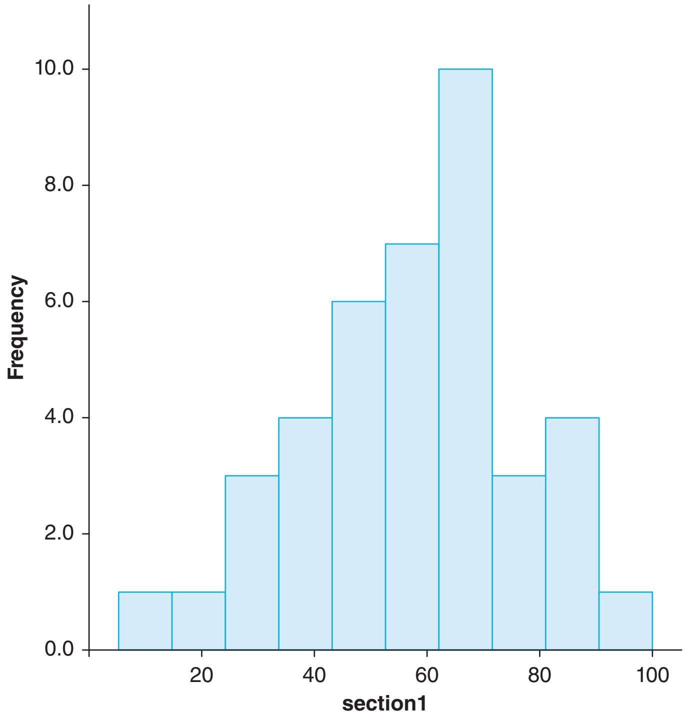


Figure 2.15

of the observations or a change in the number of bins may drastically change the shape of the histogram.

Web Link 2.4 for SPSS instruction and practice with binning.

Other forms of graphing and pictorially displaying data are described later on in this book, as needed.

Descriptive statistics

Descriptive statistics are another way to summarize the researcher's observations or data. Where graphs pictorially display the data, descriptive statistics provide numerical summaries. These numerical summaries are often utilized in further analyses of the data. We first examine the descriptive statistics used with measurement data; these are summarized in Figure 2.16.

Descriptive statistics	Measurement	Data
Measures of central tendency		
	Mean	The arithmetic average of the observations
	Median	The value that divides the rank-ordered observations in half
	Mode	The most common value of the observations among the data
Measures of spread		
	Range	The difference between the largest and the smallest value in the data
	IQR	The range of the middle 50% of the rank-ordered observations
	Variance	The average squared difference between each observation and the mean of the observations
	Standard deviation	The square root of the variance

Figure 2.16

2.5 MEASURES OF CENTRAL TENDENCY: MEASUREMENT DATA

Mean

The most basic statistics are those of central tendency or the values that represent the centre of the distribution of the observed scores. These statistics are commonly referred to as averages. These statistics can represent a variable's *expected values*. The most common measure of central tendency is the *mean*. The mean is what most people think of when they hear the word *average*. The mean is calculated by summing all of the observations and dividing this sum by the number of observations. We can write this as

$$\bar{y} = \frac{\Sigma y}{n},$$

where $\bar{y}$ is the mean of y, Σy is the sum of the observations, and n is the number of observations. For example, assume the following 19 scores are from a ten-item multiple-choice quiz:

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

To obtain the mean the quiz scores are summed and divided by the number of scores:

$$\bar{y} = \frac{\Sigma y}{n} = \frac{95}{19} = 5 .$$

Thus, the mean of the 19 scores is 5. The mean cuts the *total value* in half. Half of the *value* will be below the mean score of 5 and half will be above it. This does not necessarily mean that half of the observations will be below the mean and half will be above it. That is the function of our next measure of central tendency.

Return to the *Data Editor* window with the variable *section1* and select *Analyze* from the menu bar. From the drop-down menu, select *Descriptive Statistics*, move to the next drop-down menu and click *Frequencies*. Highlight the variable *section1* and click the arrow moving *section1* to the *Variable(s)* box (Figure 2.17). Select the button on the right labelled *Statistics*. In the menu that pops up (Figure 2.18) click on the *Mean*, *Median*, and *Mode* boxes and then on the *Continue* button at the bottom. When you return to the previous menu, click *OK*.

In the output window labelled *Statistics Viewer* the small upper box (Figue 2.19) lists several things: *N Valid* (40, which is the number of observations), *N Missing* (0, how many subjects are missing a score), mean, median, and mode. We see that the mean ($\bar{y}$) is 6.9250, the median is 7.0000, and the mode is 7.00.

Means are often referred to as descriptive statistics. Once the mean is used for any purpose other than describing the particular observations on which it is based, it becomes an *inferential statistic*. That is, the mean is being used to infer something beyond the sample, even if it is simply considered an estimate of the population mean from which the sample was drawn. Although first and foremost descriptive, means and other 'descriptive' statistics are also used *inferentially* for comparative or predictive purposes.

The related population parameter, the mean of the population, is expressed as

$$\mu = \frac{\sum Y}{N},$$

where μ (pronounced 'myu') is the true mean of the population from which samples may be drawn. $\sum Y$ is the sum of all the observations in the population, and N is the total number of observations in the population. μ indicates the true mean of the population, regardless of whether the population is finite or infinite.

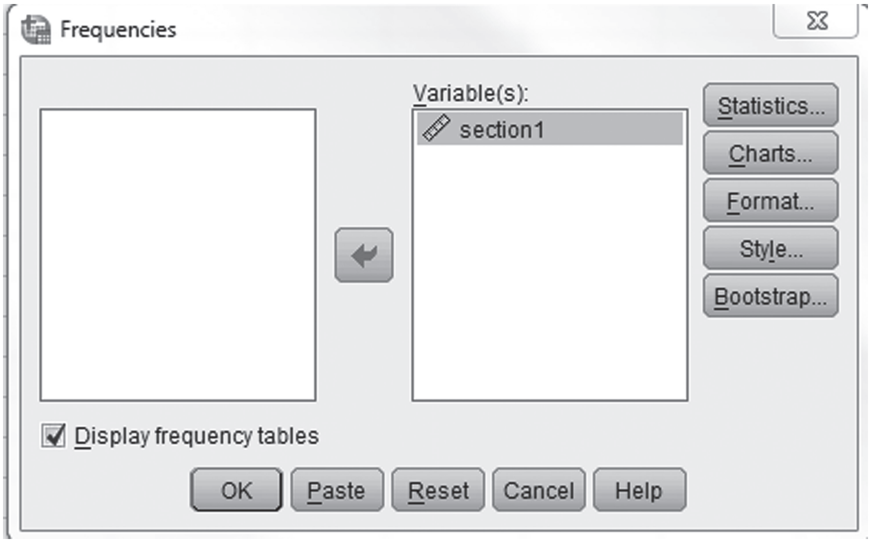


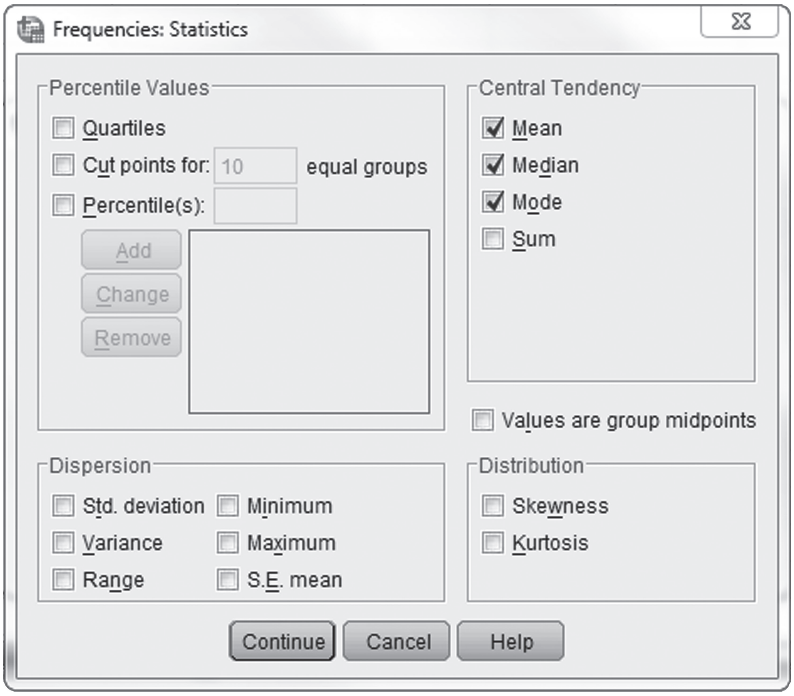
Figure 2.17

Median

The median is the value which divides the observations in half after they have been arranged in order from the smallest to the largest. Fifty per cent of the scores – not 50% of the value – will be below the median and 50% of the scores will be above the median. To find the median of the scores from the ten-item multiple-choice quiz we looked at earlier, let us begin by arranging the scores in ascending order:

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

Which value divides these scores in half? Suppose there are n scores. The median score is the one in the $\frac{1}{2}(n+1)$ th position.


Frequencies: Statistics

Percentile Values
- □ Quartiles
- □ Cut points for: 10 equal groups
- □ Percentile(s):

Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- □ Sum

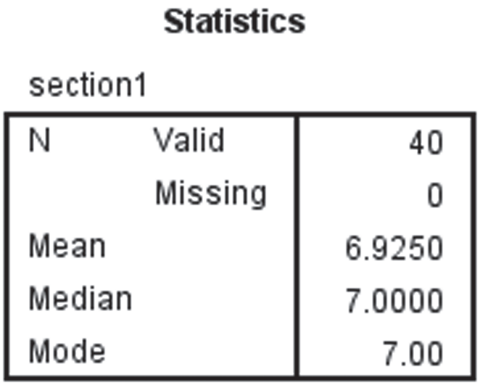
□ Values are group midpoints

Dispersion
- □ Std. deviation
- □ Variance
- □ Range
- □ Minimum
- □ Maximum
- □ S.E. mean

Distribution
- □ Skewness
- □ Kurtosis

Figure 2.18

For a data set with an odd number of observations, like the one above where n = 19, the position of the median is $\frac{1}{2}(n+1)=\frac{1}{2}(19+1)=10$. The answer, 10, represents the 10th position, *not* the value of the median. As can be seen in the above rank ordering, the value associated with the 10th position is 5. Thus, the median is 5.


Statistics

section1

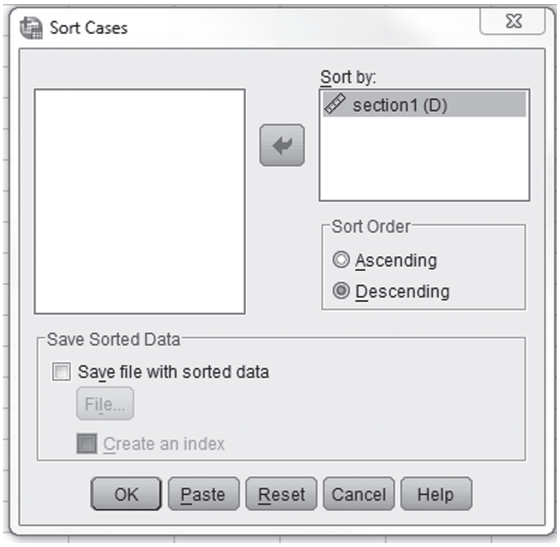
N	Valid	40
	Missing	0
Mean		6.9250
Median		7.0000
Mode		7.00

Figure 2.19

For a data set with an even number of observations, such as

1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7,
8, 8, 9, 10,

there is one additional step. We now have n = 20, so $\frac{1}{2}(n+1)=10.5$. Again, the answer, 10.5, represents the 10.5th position, *not* the value of the median. But the 10.5th position falls between



Sort by: section1 (D)

Sort Order
○ Ascending
● Descending

Save Sorted Data
□ Save file with sorted data
File...
□ Create an index

Figure 2.20

the 10th and the 11th scores, between a 5 and a 6. In such cases the average of the two values is taken: (5 + 6)/2 = 5.5. Thus, in this case, the median is 5.5.

The median is the point at which half of the observations will fall below and half will fall above. This is a different sense of the term 'average' from that of the mean. Where the mean is based on the *total value* of the observations, the median is based on the *total number* of observations. For the calculation of the median, only the one value or the two values in the middle are important. For the mean, all values play a role. In technical terms a statistic that uses all of the sample's observations is called *sufficient* (Howell, 2002).

Looking at the previous SPSS output window (Figure 2.19), we find that the median score associated with the variable *section1* is 7.0. In this data set the median is very close to the mean (6.9250). **You can gain some more SPSS experience and check the accuracy of this median by returning to the *Data Editor* window, selecting the *Data* option from the menu row, and then *Sort Cases*. When the *Sort Cases* window appears, select and move the variable *section1* from the left-hand box over to the *Sort by* box (Figure 2.20). Select the descending *Sort Order* option and then click *OK*.**

Notice that the data in the *Data Editor* have been quickly sorted, beginning with 11 and ending with 2. A question remains, however. Which of the 7s in the data set is the median?

Remember, the median of a data set will be in position. $\frac{1}{2}(n+1)$ In the case of the variable *section1* this will be position. $\frac{41}{2} = 20.5$ The 20.5th position, however, falls between the 20th and the 21st scores, between a 7 and a 7. When we take the average of the two values, we get (7 + 7)/2 = 7. In effect, the median value of 7 is not one of the actually observed 7s found in the variable *section1*, but an invisible value found between two of the observed 7s. When we count the number of observations above and below that invisible 7, we find 20 observations below it and 20 above it.

Mode

The mode provides yet another sense of the term 'average'. When the numbers are discrete, the simplest definition of the mode is that it is the most frequent value in the data set. Let us again using the above rank-ordered data from the hypothetical quiz ($n = 19$):

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

We find one 1, one 2, two 3s, three 4s, five 5s, three 6s, two 7s, one 8, and one 9. The most frequent value in the data is 5. Thus, the mode is 5.

Section 1: 7, 8, 5, 5, 9, 10, 2, 7, 6, 4, 8, 7, 8, 7, 3, 9, 6, 8, 6, 9, 8, 9, 5, 7, 6, 7, 6, 7, 9, 5, 5, 9, 7, 11, 7, 6, 4, 7, 8, 10.

Returning to the SPSS output window for variable *section1*, the *Statistics* box indicates that the mode of those 40 observations is 7.0. This is identical to the median and very close to the mean for this data set. To the chagrin of researchers, this is not always the case.

The computation of the mode is not always straightforward. When there are a great many different values and perhaps only one or two observations with the same value it makes little sense to speak of the mode. An obvious example of this could be the marks on the final examination in a biochemistry course (let us give statistics a rest). There may be 50 students in the class and no two with the same mark on the examination. In such cases, as when constructing a frequency histogram, it is common to cluster or *bin* the marks before calculating a mode. We may wish to use bins with a width of 20, for example, where we look at the number of students who scored between 81 and 100, between 61 and 80, between 41 and 60, between 21 and 40, and between 0 and 20. If we found 7 students scoring between 81 and 100, 17 students between 61 and 80, 14 students between 41 and 60, 10 students between 21 and 40, and 2 students scoring between 0 and 20, then 61–80 could be identified as the *modal category*. Notice the mode is a bin and not an individual score. The rules and limitations regarding binning presented earlier also apply here.

Whereas a sample can have only one mean and one median, a sample may have two or more modes. Not all variables are unimodal. Look at the following two samples of data that have been rank-ordered to make examination easier.

Sample I: 1, 2, 2, 3, 3, 3, 3, 4, 5, 6, 7, 7, 8, 8, 8, 8, 9, 9, 10.

Sample II: 1, 2, 2, 3, 3, 3, 3, 4, 5, 6, 7, 7, 8, 8, 8, 8, 8, 9, 10.

If you calculate the mean for each of the two samples, you find them to be 5.58 and 5.53, respectively. The two corresponding medians are 6.00 and 6.00. When Sample I is examined, two values or scores that are equally the most common are found: 3 and 8. Such a distribution of observations is deemed *bimodal*. When Sample II is inspected, 8 is found to be the most common value, but just by one observation. For most purposes, even though 8 is the most common score in Sample II, this distribution of observations also is deemed bimodal. The reason for the designation is that the frequencies of the two scores (3 and 8) stand out from the others. Of course there can be more than two modes. For most purposes researchers are concerned with knowing that the distribution of their data is unimodal.

After examining the two sets of observations again you may begin to understand why researchers are often hoping that the data are unimodal. Are the means of these two samples (5.58 and 5.53) as meaningful (pun intended) as the mean of the *section1* data (7), which was unimodal? In the case of bimodal data, do means and medians reflect *average* or *expected* scores? Clearly they do not. Should we randomly select an observation from Sample I or II, we *expect* it to be either above or below the centre. Excuse this coarse but instructive example: after calculating the mean of a large sample of young adults we find that the average human being has one testicle.

The mean and the median are the measures of central tendency most often employed to describe the *average* or the *centre* of a set of observations. As seen above, all three (mean, median, and mode) can be the same value. This occurs only in one particular circumstance: when the observations are *unimodal* and they are distributed symmetrically around the mode. In terms of measurement data, the mode is the least applied measure of central tendency. The mode is most useful when denoting the most frequent observation among a small set of options, particularly when summarizing nominal data. For example, when describing voter preferences, political scientists may wish to identify the most *popular* party. The *expected* choice of political party is not some sort of *middle* choice. Remember, numbers that are assigned to the political parties are nominal in nature and thus not amenable to calculating a mean or median.

Mean versus median

The mean and the median both describe data and form the basis for further inferential analyses. As will be seen later, the mean is typically employed when researchers analyse their data with *parametric* statistics, whereas the median is often employed when *nonparametric* statistics are used. At this point let us simply say that nonparametric statistical procedures do not use sample statistics, such as the mean ($\bar{y}$), to estimate corresponding population parameters (μ). This is an important distinction to which we will return in Parts II and III of the book. Here the groundwork is laid, however, through an exploration of how the mean and the median are differentially affected by small modifications to a data set.

To examine how the mean and the median react differently to small modifications in a data set, open a new SPSS *Data Editor* window and enter the set of 19 scores from the hypothetical ten-item multiple-choice quiz which we first encountered earlier:

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

Run the *Descriptive Statistics* (mean, median, and mode) on these 19 observations four times. Begin by running the program with the above data as given, then three more times: first, after changing the highest score, the 9, to a 10; next, after changing the 9 to a 20; and finally, after changing the 9 to 100. Notice that across the four analyses the median is constant. The value in the middle, 5, remains the value in the middle, regardless of the increases to the highest value. Rounding to the nearest two decimal places, however, the mean increases from 5.00 to 5.05, then to 5.58, and finally to 9.80. In the last case, the mean of the 19 observations is now greater than the second highest value in the data set. Note that there are also no changes in the mode. A statistic insensitive to outliers is said to be *resistant*. While the median and mode are resistant, the mean is not.

CHALLENGE QUESTION

As you have seen, the mean of a data set may change simply by changing a single value in the set. Here are several challenge questions.

1. What is the minimum number of values you would need to change the median?
2. How many values can you change without changing the median?
3. How many values can you change without changing the mean?
4. Can you change the value of only one score without changing the mean?

You might wish to return to the SPSS data set and play around with changing some of the values.

Web Link 2.5 for an answer to the challenge questions.

There are other differences between the mean and the median. An important one mentioned here involves minimizing the *total distance* the individual observations are from a single value (the absolute deviations) versus minimizing the *total squared distance* the individual observations are from a single value (the squared deviations). The value that minimizes the *absolute value* of the deviations is the median. For example, if a data set is comprised of 1, 2, and 9, the median would be 2. If we subtract 2 from our three scores – which we will call *deviations* or *errors* – we get the absolute values of, $|1|$, $|0|$ and. $|7|$ We have $\Sigma(|y - \text{median}|) = 8$. If any value other than the median is used, the *total absolute distance* will be greater than 8. This includes using the mean of our three scores:

$$\bar{y} = \frac{\Sigma y}{n} = \frac{12}{3} = 4\ .$$

If the mean of 4 is subtracted from the three scores the absolute distances are, $|3|$, $|2|$ and $|5|$, and. $\Sigma(y - \bar{y}) = 10$ If you were allowed to use only one value to guess all of the values hidden in a hat, and you wished to minimize the total absolute value of your error (total absolute distance), then you should choose the median of the values. If, however, you wish to minimize the *squared deviations* or squared total errors using a single value, then you will want to know the mean of the values in the hat. Analysing the three scores of 1, 2, and 9, we find the following. Recall the mean is 4. Then

$$\Sigma(y - \bar{y})^2 = (1-4)^2 + (2-4)^2 + (9-4)^2 = 38\ .$$

When any value other than their mean is used the total of the squared errors will be greater than 38. This includes using the median of the three scores.

$$\Sigma(y - \text{median})^2 = (1-2)^2 + (2-2)^2 + (9-2)^2 = 50\ .$$

These differences between the mean and median are important to note. They will become important for tests of statistical significance and the problem of assumptions.

REVIEW QUESTION

When will it make no difference whether we use the mean or the median for calculating the total absolute difference and the total squared difference?

Web Link 2.6 for an answer to the review question.

• • • • •

Note that the mean and the median may be defined as being the *expected values*, and the three scores (1, 2, and 9) described as the *observed values*. This is the simplest use of these terms (expected and observed) introduced in Chapter 1.

Composite mean: mean of means

To extend the application of the notation and symbol systems found throughout this book, the concept of *composite mean* is explored. A composite mean is defined as the overall mean or grand mean of all observations across several groups or samples. For example, there may be three sections of students in your statistics course. Imagine you are told each section's mean score on a quiz and you wish to know the overall or grand mean (*GM*) of all students across all three sections. There are two circumstances in which the solution is simple. First, if we have the marks from all of the individual students, we can simply sum them – ignoring sections – and divide by the total number of students. This can be written as

$$\text{GM} = \frac{\Sigma y_{ij}}{N},$$

where *GM* represents the grand mean across all sections or samples, and Y_{ij} represents the *i*th subject in the *j*th section or sample. In simple terms, Σy_{ij} is the sum of the marks of all students across all sections of the course. *N* represents the total number of students across all sections, and lowercase n_j represents the number of students in a given section. Thus, $N = \Sigma n_{\text{j}}$. The notation denotes summing all n_j scores across all sections or subsamples.

Second, if there is an equal number of students in each section of the course, we can sum the section means and divide them by the number of sections. This is the simple mean of the means, written

$$\text{GM} = \frac{\Sigma \bar{y}_j}{K},$$

where K is the number of sections, and $\bar{y}_j$ represents the mean of the jth section. We can write $\Sigma \bar{y}_j$ more formally as $\Sigma_{j=1}^{k} \bar{y}_j$.

A problem arises when the individual observations are not available and the section sizes are unequal. When the n are unequal, it is highly unlikely that the GM is a simple mean of the means. For example, sample sizes in Figure 2.21 vary greatly.

The first three columns are the scores in three samples. The fourth column is all scores from the three samples: $\sum y_{ij}$. The means for the three samples are 2, 4, and 3, respectively. The simple mean of the means is $GM = \left(\sum \bar{y}_j\right)/k = 3$. But notice that the simple mean of the means is not the true mean of all 13 scores. The mean of all scores across the three samples is actually 3.3077. The problem is not insurmountable, however. Remember what a mean is: the total value divided by the number of observations.

With the sample means and the number of observations for each sample, it is possible to reconstruct $GM = \left(\sum y_{ij}\right)/N$. The numerator, Σy_{ik} $(\Sigma y_{i1} + \Sigma y_{i2} + \ldots + \Sigma y_{ij} + \ldots + \Sigma y_{ik})$, represents the total value of the observations in the last groups . Because K is the total number of groups, it is also the number associated with the last group. Next, recall that $\bar{y} = \left(\sum y\right)/n$. Thus, with a little algebra, $\sum y = n\bar{y}$. That is, the total of the marks of the students in any section can be recovered by multiplying the number of students in the section by the mean of the section. The result of completing this for all sections and summing the products is the total of all of the individual marks across all sections: $\Sigma n_j \bar{y}_j = \Sigma y_{ij}$. The individual marks are not recoverable, but the total is. Summing the number of observations across all sections provides the total number of observations: $N = \Sigma n_j$. Thus,

$$GM = \frac{\sum n_j \bar{y}_j}{\sum n_j} = \frac{\sum y_{ij}}{N} .$$

For our example,

$$\frac{(2\times 3)+(4\times 7)+(3\times 3)}{3+7+3} = \frac{43}{13} = 3.3077 .$$

Enter the data from Figure 2.21 into SPSS using the four columns as variables and obtain the means for the four variables for practice and for comparison with the subsequent calculations in the text. Note: there may be small differences due to rounding.

Sample 1	Sample 2	Sample 3	All scores
1	1	2	1
2	2	3	2
3	3	4	3
	4		1
	5		2
	6		3
	7		4
			5
			6

Figure 2.21

CHALLENGE QUESTION

When the sample *n* are unequal, is it ever possible to calculate the grand mean by summing the means of the subgroups and dividing by the number of groups? If it is never possible for this to happen, then why not? If it is possible, when is it?

Web Link 2.7 for an answer to the challenge question.

2.6 MEASURES OF SPREAD: MEASUREMENT DATA

This section covers another crucial dimension of descriptive statistics: measures of spread. Locating the centre of a set of observations is important when summarizing data, but the centre alone is insufficient. How much of a description can the mean alone provide? For example, if an instructor informs her class that the mean score on an examination was 75, this provides some information about how well the class performed. But knowing the centre says nothing of the spread of the scores. If the instructor informs the class that the marks range from 70 to 80, this gives a very different impression than if the instructor says the marks range from 30 to 90. In this section various measures of spread (range, inter-quartile range, variance, and standard deviation) along with their uses and limitations are surveyed.

Either calculate or enter into a new SPSS *Data Editor* window the following two variables.

Var1: 0, 2, 4, 4, 5, 5, 5, 6, 6, 8, 10.

Var2: 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6.

They can be labelled *var1* and *var2*.

As earlier, select *Analyze* from the *Data Editor* menu in SPSS. Then select *Descriptive Statistics* and then *Frequencies*. From the *Statistics* menu calculate the mean, median, and mode of both *var1* and *var2*.

Notice that both groups have a mean of 5.0, a median of 5.0, and a mode of 5.0. A visual examination of *Var1* and *Var2* reveals a clear difference in the two sets of observations. The two groups of scores are considerably different in terms of their *spread*.

Range and inter-quartile range

In *Var1* the observations range from 0 to 10. In *Var2* the observations range from 4 to 6. The *range* is defined as the difference between the highest and the lowest values in a set of observations. For *Var1*, range = 10 – 0 = 10. For *Var2*, range = 6 – 4 = 2. Thus, although the two groups of observations have the same means, medians, and modes, they differ considerably in their range. Combining the mean and the range offers a more informative summary description of the two variables than either alone does. The measure of range has its utility, but it also has limitations. Consider the following three groups of data:

var1: 2, 4, 5, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 9, 10, 12.

var2: 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 10, 10, 11, 11, 12, 12, 12, 12, 12, 12.

var3: 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 16.

The observations are rank-ordered for ease of visual examination. If you enter the 20 observations of these three variables into SPSS and examine the descriptive statistics, you will discover that all three variables have a mean and median of 7.0 and a range of 10. But the shapes of the distributions of the observations about their means and across their range differ greatly. This is one reason why pictorial summaries, particularly histograms, of data are always valuable. Look at the difference in these three variables, all with the same mean, median, and range.

The observations in *var1* (Figure 2.22) are clustered near the mean, median, and mode, and are symmetrically spread out around the central value. There are no observations near the mean or median in *var2* (Figure 2.23), nor is there a clear single mode, although 12 is slightly more frequent than 2. Where the observations in *var1* are clustered near the centre, the observations in *var2* are clustered at the extremes. When a variable's scores have two or more clusters as we see in *var2*, the mean and the median have little value as measures of central tendency. The *expected value* is certainly not the centre. In effect, there are two expected values, one at either extreme. The observations in *var3* (Figure 2.24), like those in *var1*, have a mean, median, and mode all of 7.0. And like those observations in *var1*, those of *var3* also have a range of 10. The observations

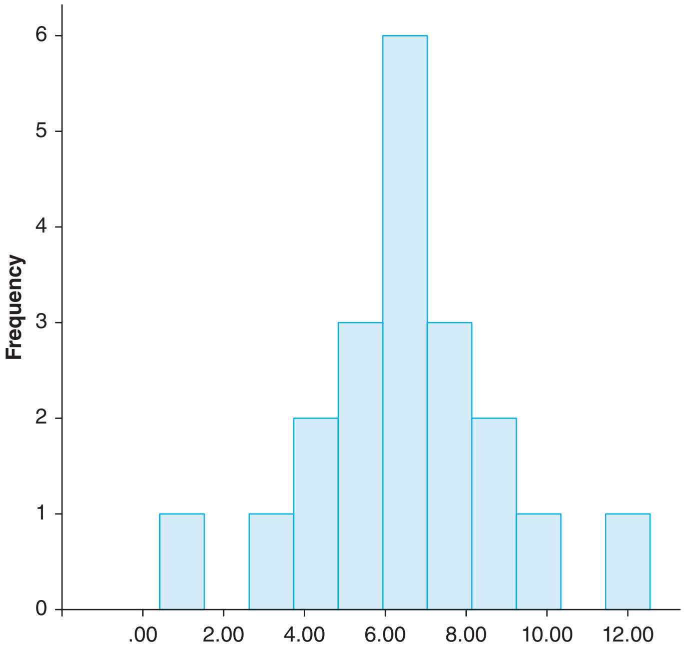


Figure 2.22 *var1* histogram

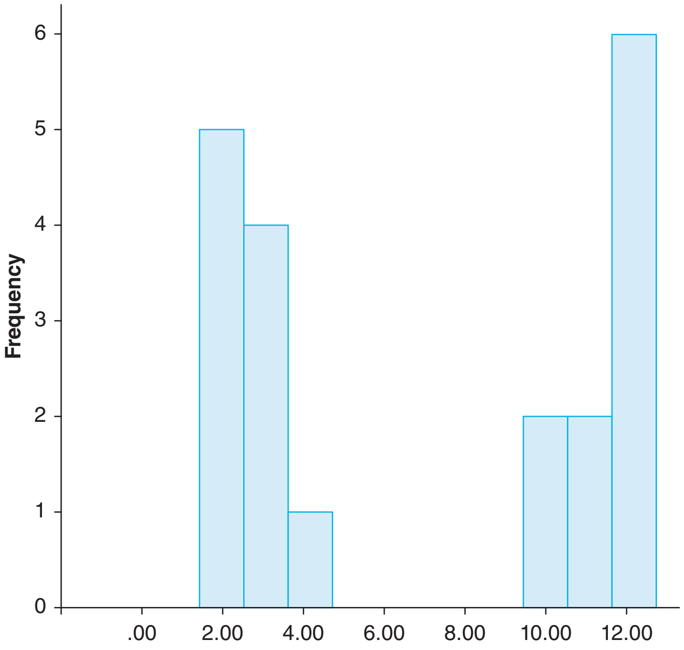


Figure 2.23 *var2* histogram

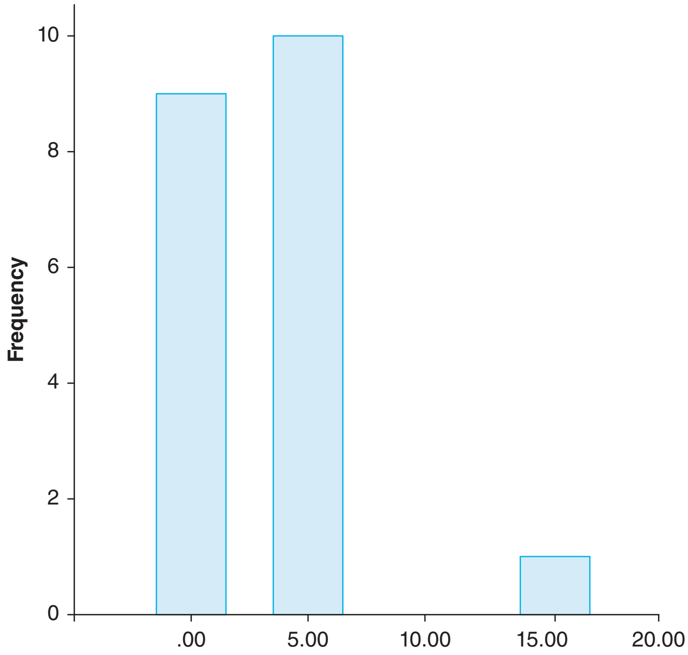


Figure 2.24 *var3* histogram

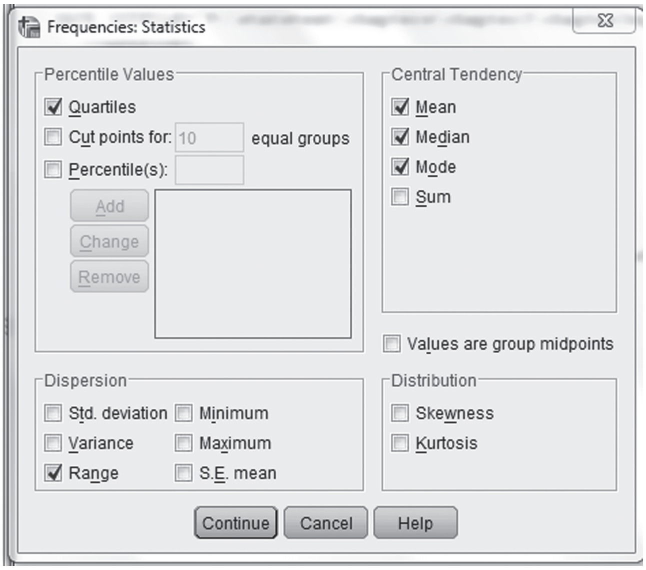
in *var3*, however, are clustered at the low end of the range, with the single exception at the highest value.

Although the centre and spread of *var1* is what we typically assume to be the case, *var2* and *var3* illustrate that it is important to examine how the observations are clustered in order to avoid misinterpretation. *var3* illustrates the effect of a single extreme value. Remember, all three variables have a mean of 7.0 and a range of 10. Despite this equivalence in these measures of central tendency, the distribution of the observed values in these three variables differs greatly. In this case, using only the mean and range to describe the data fails to adequately summarize the three variables and leads to erroneous comparative conclusions. For example, if these were quiz marks from three sections of a class, using only the mean and the range to describe the performance of the three sections leads to the conclusion that the three sections performed similarly. Clearly, this is not the case. Is it this sort of dilemma which drove Mark Twain and Benjamin Disraeli to say that there were 'lies, damned lies, and statistics?'

To avoid some of the problems associated with clustering differences and extreme scores, some researchers report the *inter-quartile range* (IQR) when describing their data. The IQR range can be computed by segmenting the rank-ordered values into quartiles. This is done by determining the median of the scores and then determining a median for the upper and lower halves of the rank-ordered scores. This divides the scores into four equal segments or *quartiles*. (*Note*: SPSS uses a different procedure for determining the quartile cut-offs. This can result in slightly different values at times.) Once this is done, the median of the lower half is subtracted from the median of the upper half to produce the inter-quartile range: IQR = $Q_3 - Q_1$. Half of the observations will fall between the median of the lower half and the median of the upper half. Returning to our three variables used for examining the overall range, it is clear how the IQR helps to differentiate the three sets of observations.

If you entered the data into SPSS, you can return to the *Data Editor* window, select *Analyze*, then *Descriptive Statistics*, and finally *Frequencies*. From the *Statistics* menu click the box in the upper left-hand corner next to *Quartiles* (Figure 2.25). This will add the values associated with the 25th, 50th, and 75th percentiles for each variable to the output (Figure 2.26). The value associated with the 50th percentile is the median; the value associated with the 25th percentile (Q_1) is the median of the lower half; and, the value associated with the 75th percentile (Q_3) is the median of the upper half.

For *var1*, the values associated with the 75th and 25th percentiles are 8.0 and 6.0, respectively, and the IQR is 8.0 – 6.0 = 2.0. For *var2*, the values associated with the 75th and 25th percentiles are 12.0 and 2.25, respectively, and the IQR is 9.75. For *var3*, the values associated with the 75th and the 25th percentiles are 7.0 and 6.0, respectively, and the IQR is 1.0. Remember, 50% of the scores will always fall within the IQR. With respect to the three variables, the IQRs indicate that the observations in *var2* have the greatest spread of the three and those of *var3* are spread out the least. Looking at the histograms again, the IQRs correspond better with the data than do the three overall ranges. Of course, the IQR does not indicate the spread of the observations beyond the middle two quartiles. Nor does it specify how the observations are spread out within the middle two quartiles.



Frequencies: Statistics

Percentile Values
- ☑ Quartiles
- ☐ Cut points for: 10 equal groups
- ☐ Percentile(s):

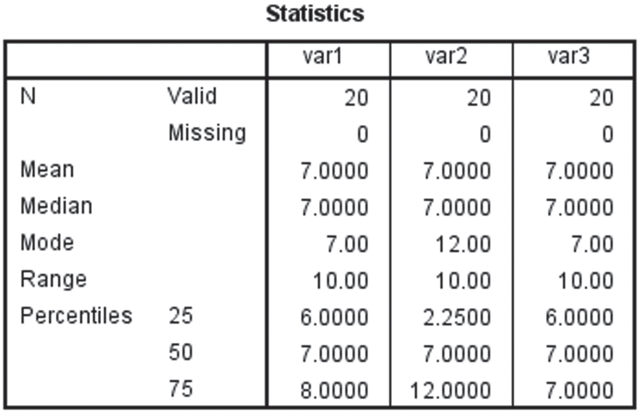
Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- ☐ Sum

☐ Values are group midpoints

Dispersion
- ☐ Std. deviation
- ☐ Minimum
- ☐ Variance
- ☐ Maximum
- ☑ Range
- ☐ S.E. mean

Distribution
- ☐ Skewness
- ☐ Kurtosis

Figure 2.25



Statistics

		var1	var2	var3
N	Valid	20	20	20
	Missing	0	0	0
Mean		7.0000	7.0000	7.0000
Median		7.0000	7.0000	7.0000
Mode		7.00	12.00	7.00
Range		10.00	10.00	10.00
Percentiles	25	6.0000	2.2500	6.0000
	50	7.0000	7.0000	7.0000
	75	8.0000	12.0000	7.0000

Figure 2.26

In addition to being a descriptive statistic of spread, the IQR is used to construct another pictorial representation of data: the box-and-whisker plot (Tukey, 1977). Box-and-whisker plots are graphic representation of a variable's centre, IQR, and range. Box-and-whisker plots are often used to graphically compare the distributions of two or more variables. Figure 2.27 depicts the box-and-whisker plots for *var1*, *var2* and *var3*.

The *y*-axis is scaled in the units of measurement used to record the variables depicted in the graph. If multiple variables are depicted in a single graph, they must be recorded using the same scale of measurement (e.g., the number of correct items on a quiz). The bottom of each rectangular box represents the 25th percentile. The top of each box represents the 75th percentile. The horizontal line inside a box denotes the median. The whiskers extend as far above and as far below the box to the highest and lowest score within 1.5 IQRs from the median.

Any scores beyond the 1.5 IQRs above or below the median are represented by dots or special characters. The dot above the whisker in *var1* represents the score of 12. The '20' next to it identifies it as case (row) 20 for *var1* in the *Data Editor*. The star above the whisker in *var3* represents the score of 16. The '20' next to it again identifies it as case (row) 20 for *var3* in the

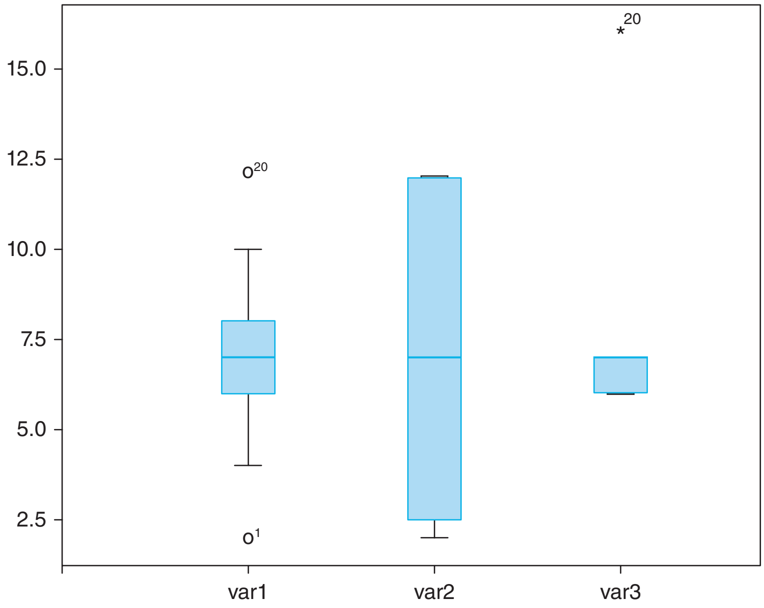


Figure 2.27 Box-and-whisker plots for *var1*, *var2* and *var3*

Data Editor. The nearly complete (virtually invisible) lack of whiskers for *var2* and *var3* reflects the constricted clustering of the observations.

To create the box-and-whisker plots for *var1*, *var2* and *var3* select the *Graphs* category in the *Data Editor*, cursor down to the *Legacy Dialogs* and then move over and down to the *Boxplot* option. Because we are summarizing more than one variable, select the *Summaries of separate variables* button when the *Boxplot* window appears (Figure 2.28) and click *Define*. In the *Define Simple Boxplot* window highlight and move the three variables over to the *Boxes Represent* area (Figure 2.29) and click *OK*. The box-and-whisker plots will appear in the output window.

REVIEW QUESTION

In var2 there are no scores above the IQR. The whisker is equal to the top of the box. Why? Why are there no whiskers on the box for var3?

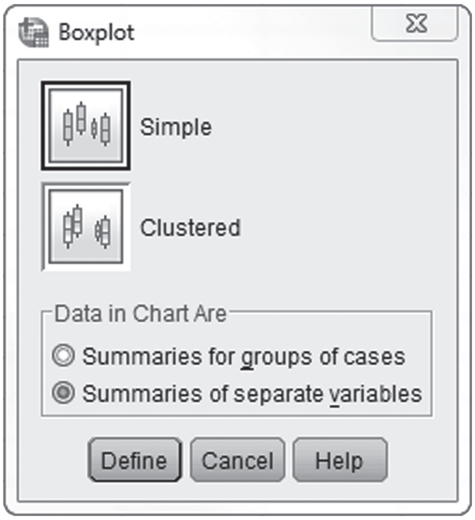
Web Link 2.8 for an answer to the review question.

Variance and standard deviation

For measurement data, the most common measure of central tendency is the mean. When the mean is used, the most common measure of spread reported by researchers is the *variance*. Like the range and IQR, the variance is a single value that indicates something about how the observations are spread out. It is not directly based on the span of the observations, which only involves two values. Rather, it is based on each observation's relation to the mean. It is the *average squared distance* or deviation of a set of scores from their mean. Remember, a mean is the value that minimizes those squared distances, so using those squared distances as a measure of spread should not be a surprise. The formula for the sample variance is

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1} = \frac{\text{sum of squared deviations}}{\text{degree of freedom}}.$$

When $n - 1$ is in the denominator, rather than n, variance is being used as an inferential statistic. That is, it is employed as an estimate of the population variance, σ^2. When variance is an inferential statistic, because $\bar{y}$ is only an estimate of μ, we need to subtract 1 from the number of observations. Degrees of freedom are presented in more detail later in the chapter.

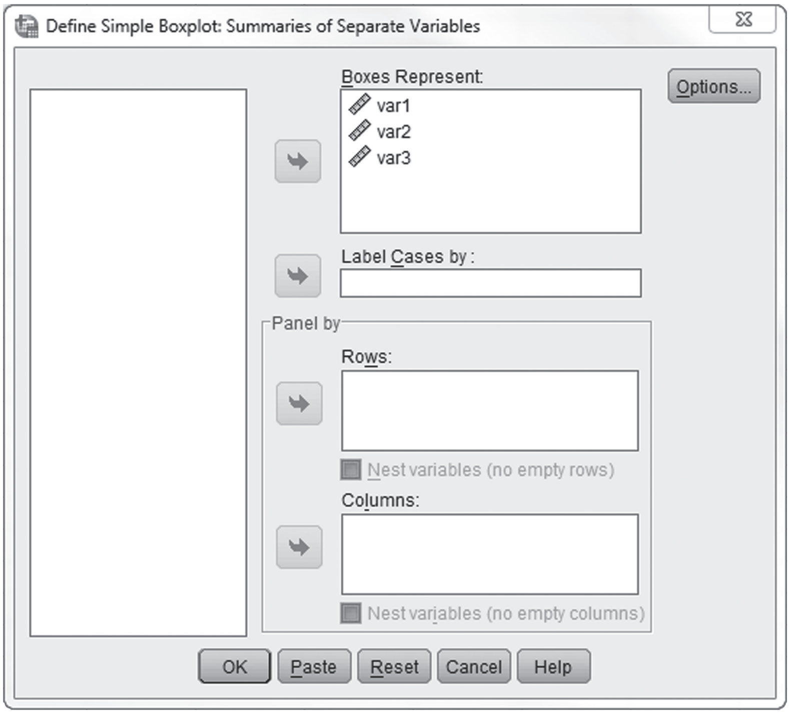


Data in Chart Are

○ Summaries for groups of cases

● Summaries of separate variables

Figure 2.28


Define Simple Boxplot: Summaries of Separate Variables
Options...
Panel by
Nest variables (no empty rows)
Nest variables (no empty columns)
OK
Paste
Reset
Cancel
Help

Boxes Represent:
var1
var2
var3

Label Cases by:

Rows:

Columns:

Figure 2.29

Using the mean and variance to summarize a set of observations results in the following understanding. If an observation (observed score) is randomly selected from a given sample, we would *expect* it to be the mean. We intuitively know, however, that the chance of randomly selecting a score that is equivalent to the mean is extremely small. In fact, a sample's mean may not correspond to any of the actual observations in the sample. How wrong will our expectation be? In other words, what is the *expected difference* between the mean and any randomly observed score? In terms of squared units, on average, we will *expect* to err by the value of the variance. The mean is the *expected score* and the variance is the *expected error*. This is most applicable for those cases where the variable has a single mode and the observations are roughly symmetrically distributed about the mean.

REVIEW QUESTIONS

1 Why are the variances of the following two sets of scores the same?

Set A: 1, 2, 3.
Set B: 101,102,103.

2 Can you create a data set of five observations that has a mean of 3 and a variance of 4?

Web Link 2.9 for the answers to the two review questions.

The greater the variance, the greater the spread in the scores about their mean. It is this last element – *about their mean* – that makes the variance different from the range and IQR. Variance creates a link between the measure of central tendency and spread. If you entered *var1, var2* and *var3* into SPSS, return to the *Data Editor* window.

Select *Analyze*, then *Descriptive Statistics*, and then *Frequencies*. From the *Statistics* menu select *Std. deviation* and *Variance* in the *Dispersion* box.

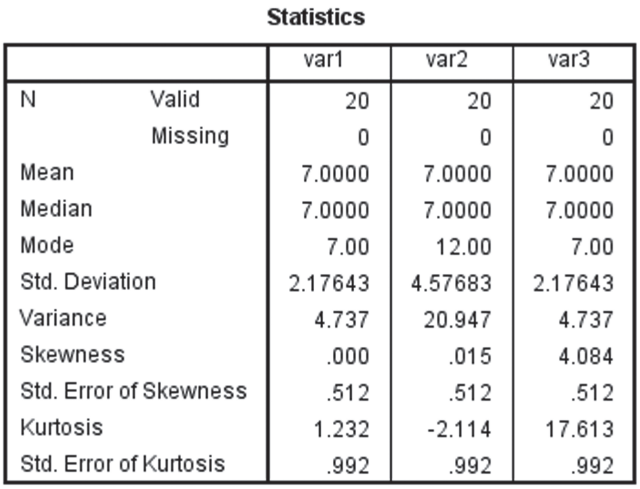
In the output for *var1*, *var2*, and *var3*, we find that the variances are 4.737, 20.947, and 4.737, respectively (Figure 2.30). Why is the variance of *var2* so much greater than the variance of the other two variables? Look back at the histograms. It is because most of the scores in *var2* are clustered further from their mean than are the scores in the other two variables.

Note that the variances in *var1* and *var3* are identical, despite the fact that the spread of the scores is very different. This points to the limitations of variance. That is, the variances are most safely comparable when the variables have a single mode and when the observations are roughly symmetrically distributed about the means. Despite such limitations, variance is a crucial element in most forms of statistical analysis.

REVIEW QUESTION

How is it possible that *var1* and *var3* can look so different in terms of their spread and have the same variance?

Web Link 2.10 for an answer to the review question.



Statistics

		var1	var2	var3
N	Valid	20	20	20
	Missing	0	0	0
Mean		7.0000	7.0000	7.0000
Median		7.0000	7.0000	7.0000
Mode		7.00	12.00	7.00
Std. Deviation		2.17643	4.57683	2.17643
Variance		4.737	20.947	4.737
Skewness		.000	.015	4.084
Std. Error of Skewness		.512	.512	.512
Kurtosis		1.232	-2.114	17.613
Std. Error of Kurtosis		.992	.992	.992

Figure 2.30

The formula for the true population variance is

$$\sigma^2 = \frac{\Sigma(y-\mu)^2}{N},$$

where N is the number of observations that comprise the population of interest. Because variance is computed in squared units, it is difficult to picture it as a measure of spread. Variance does have important advantages, but transparency is not one of them. For this reason, particularly for descriptive purposes, variance is often converted into *standard deviation*.

By taking the square root of variance we return to the initial units of measurement. Thus the sample standard deviation is given by

$$\sqrt{s^2} = s = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}.$$

The corresponding formula for true population standard deviation is

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

where N is the number of observations that comprise the population of interest.

The standard deviations of *var1*, *var2*, and *var3* are 2.18, 4.58, and 2.18, respectively. These values inform us that if we randomly select observations from *var1* and use their mean (7.0) to blindly guess the observations' values, on average we would be wrong by approximately 2.18 quiz items. The same is true for *var3*. With respect to *var2*, we would be wrong on average by 4.58 items.

It is not easy to understand what a variance of 143.36 means with respect to the height of the above nine characters shown in Figure 2.31 (1 inch equals 2.54 cm). Variance is in squared units of measurement: squared seconds, squared dollars, the squared number of correct answers on an exam, etc.

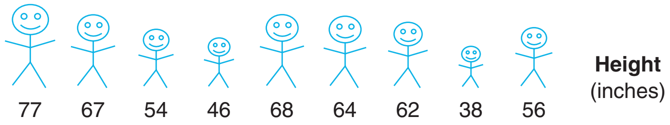


Figure 2.31

It may be asked why we square the difference between the mean and the scores in the first place, if subsequently the results will be unsquared. First, because the mean cuts the total value of the sample in half, half of the total value will be positive and half will be negative. Thus, the total difference between the mean and the individual observations will always be zero. This can be rectified, of course, by calculating the absolute value of the differences. Will this result in the same value as first squaring the deviations and then taking the square root? No, it will almost always be different. Furthermore, as we will see, by first squaring the deviations, our measure of spread gains an important quality.

Web Link 2.11 for a discussion regarding the difference between the squared and the absolute deviations.

The effect of a single score on variance

Like the mean, variance is not a *resistant* statistic and is sensitive to extreme scores. To examine this sensitivity we will explore the set of 19 scores from the hypothetical ten-item multiple-choice quiz which we first encountered in Section 2.5:

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

If you wish, open the SPSS *Data Editor* window and enter the scores. First calculate the mean, median, standard deviation, and variance with the data as given. Then calculate those statistics three more times: after changing the 9 first to a 10; after changing it to a 20; and finally, after changing it to 100. Notice that there are no changes in the median. Note also the dramatic increase in variance (from 4.00, to 4.50, to 15.26, to 480.29). Because the mean and the variance are sensitive to extreme scores, vigilance is required to ensure that such scores do not unduly influence our analyses and conclusions. Regarding the fourth data set, do you think that stating that it has a mean of 9.79 and a variance of 480.29 is a fair description? Are the scores in the fourth version of the data set really that different from those in the first? Your answer in both cases should be *no*.

Besides being extreme, researchers worry that such scores may be erroneous. Perhaps there was a mistake recording the score or some special circumstance existed that produced the score, a special circumstance with which the researcher is unconcerned. Stated differently, perhaps the observation does not arise from the same population as do the other scores. Researchers certainly do not want recording errors or special circumstances to unduly influence the outcome of their research. When scores are found to be too extreme it is standard practice to either eliminate them or to transform them.

Imagine you are recording the speed with which a sample of ten university students can name primary colours that flash onto a computer screen. We would normally expect that students will be able to do so in less than half a second. Now imagine that unbeknown to you one of the students is not a native English speaker. Having to name the colours in English rather than in his or her mother tongue may slow that person down substantially. As a consequence, the mean and variance of your sample of 10 students will be greater than it would be without the non-native English speaker. This is a simple example where one observation originates from a different population and reflects a circumstance that we were not interested in exploring (e.g., differences in mother tongue). The term 'different population' does not refer to a geographical location. A *different population* refers to observations that have a *different* μ from the population from which you believe you are sampling. In this case, you assume that you are sampling native English speakers.

The first question that needs answering is, when is a score too extreme? This question needs to be answered before data are collected. It is *not* appropriate to establish the criterion after viewing the data. Typically, researchers use a distance of either 3 or 4 standard deviations from the mean as the criterion. Unless stated otherwise, throughout this book we use 4 standard deviations as the cut-off. That is, if an observation is more than 4 standard deviations above or below its mean, it will be considered too extreme, an *outlier*. This can be tested by subtracting the mean from any questionable score and then dividing the difference by the sample's standard deviation. If we examine the four versions of the above data set,

1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9,(10, 20, 100),

we find that only in the final version, where the highest score is 100, do we find an outlier:

$$\frac{100-9.79}{21.92}=4.23\ .$$

Because 4.23 is greater than 4.0, we conclude that a score of 100 is an outlier in this sample and it must be either deleted or transformed. If we test the largest value (20) in the third version of the data set, we find that the 20 is not an outlier,

$$\frac{20-5.58}{3.88}=3.70\ .$$

Remember, scores in themselves are not outliers. They are only outliers – too large or too small – for a given sample with a particular mean and standard deviation.

The second question which needs to be answered is, what should be done with an outlier? The first, and perhaps the most common, strategy is to simply delete the outlier from the sample. After doing so, the descriptive statistics must be recalculated. In the example above, once we have removed the 100 from the fourth version of the sample, the new mean, variance, and standard deviation are 4.78, 3.24, and 1.80, respectively. This is a great change from the mean and standard deviation of 9.79 and 480.29 when the 100 was included in the sample. Note that the median and mode continue to remain unchanged: 5.00 and 5.00. The outlier's effect is on the mean and variance. This strategy of removing an observation is not always the most desirable, especially with small sample sizes.

Another common strategy is to winsorize the observation. Winsorizing (a process named after the biostatistician C. P. Winsor) recodes the outlier to the nearest acceptable higher or lower value

(Hastings et al., 1947). In our current working example, we calculate the winsorized score using the original mean and standard deviation. The highest acceptable score is 4 standard deviations (21.92 × 4 = 87.68) above the mean (9.79). The new winsorized score (9.79 + 87.68) is 97.47 (rounded to 97), which is not that different from the original score of 100. Again a new mean, variance, and standard deviation must be calculated. In this case they become 9.63, 450.69, and 21.23, respectively.

REVIEW QUESTION

What would be the new mean and variance had we winsorized the score of 100 to three standard deviations above the mean rather than to four?

Web Link 2.12 for the answer to the review question.

The choice of strategy for addressing outliers can make a big difference to the findings. In our example, when we chose the strategy of deleting the outlier, the descriptive statistics were much closer to the original statistics when the highest score was 9 than they were when the highest score was 100. On the other hand, after winsorizing, the statistics were not greatly different from those produced when the outlier was included. The choice of criterion for identifying outliers also makes a difference in the results of analyses discussed in Parts II and III of this book.

Sampling distribution of the mean

Not only is there variance in observations or scores, there is also variance in statistics that summarize observations. Recall that statistics, such as the mean, are used as estimates of population parameters. As a consequence, there will be sample-to-sample differences in those estimates. Imagine randomly sampling the accumulated debt of 100 students graduating from your university this year. Now imagine taking another random sample of 100 students, again from this year. It is extremely unlikely that the means of the two samples will be the same. Therefore a mean, variance, and standard deviation of the means can be computed. Conceptually, with a very large number of sample means all drawn from the same population, the formulae for the mean of the means, the variance of the means, and the standard deviation of the means are straightforward. These formulae assume that sample size is held constant. The mean of the means is

$$u_{\bar{y}} = \frac{\Sigma \bar{y}}{k},$$

where k is the number of samples, which theoretically could be infinite . The variance of the means is

$$\sigma_{y}^{2} = \frac{\Sigma(\bar{y} - \mu)^{2}}{k}$$

and the standard deviation of the means is

$$\sigma_{\bar{y}} = \frac{\Sigma(\bar{y} - \mu)^2}{k} .$$

It can be demonstrated that with only one sample mean the variance of the mean, referred to as the sampling distribution of the mean, and the standard deviation of the mean, referred to as the standard error of the mean, can easily be estimated. The sampling distribution of the mean is

$$\sigma^2_{\bar{y}} = \frac{s^2}{n} ,$$

where s^2 is the variance of the sample's observations and n is the sample size; the standard error of the mean is

$$\sigma_{\bar{y}} = \frac{s}{\sqrt{n}} ,$$

where s is the standard deviation of the sample's observations.

These 'statistics of statistics' are central to almost all inferential tests which will be presented in Parts II and III of this book. The reason for the change in terminology is context. While *variance* and *standard deviation* indicate that the statistics refer to observations, *sampling distribution* and *standard error* indicate that the statistics refer to statistics.

2.7 WHAT CREATES VARIANCE?

The answer to this question is one key to understanding research design and statistical testing. Why is there variance? There is variance because people, animals, plants – all members of any class of things – differ. No two people are exactly the same. But that simply begs the question, why do they differ? From the perspective of research and statistical analysis the answer requires a bit of a digression.

Anything you wish to study is a variable, either a dependent or a criterion variable. Memory is an example. If you test the memories of a group of people, you will find that they vary. There will be a mean and a variance in the memory scores. Like any activity, memory performance will be influenced by an unknown number of other variables. Some of these will have a positive effect on a person's memory and some will have a negative effect. Those with more positive effects than negative effects will be above the average performance and those with more negative effects will be below the average performance. The greater the preponderance of positive effects, the further above average the subject's memory performance. The greater the preponderance of negative effects, the further below average the subject's memory performance. Memory for a list of words may be influenced by such factors as the person's familiarity with the words, the time spent studying the words, as well as a host of other factors such as how much sleep he or she had the night before, their memorization strategy, and how much coffee he or she has consumed. There is a host of influences of which you are unaware.

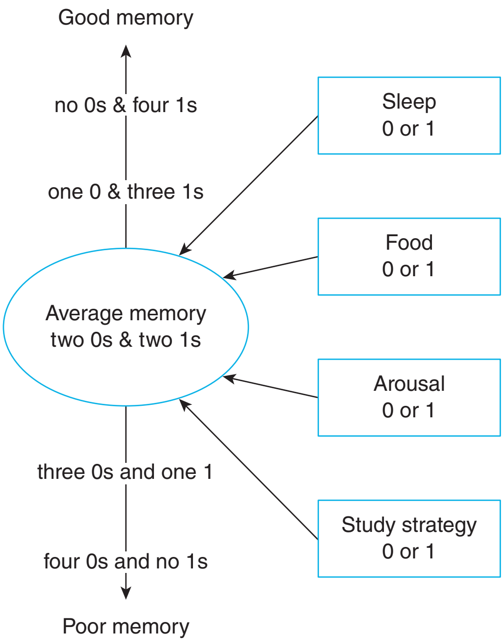


Figure 2.32

Let us assume that variable y (memory test score) is influenced by four factors: x_1 (sleep), x_2 (food), x_3 (arousal), and x_4 (study strategy). Insufficient sleep and nutrition negatively affect memory performance. Being unmotivated or being overly anxious (arousal level) also negatively affect memory performance. Finally, having a strategy other than simply using rote memory to study improves memory performance. Assume that each of the four factors can only take the value of either 0 or 1. The 0 represents the negative effect of the factor and the 1 the positive effect. This hypothetical model of memory performance is depicted in Figure 2.32 .

Someone may have the following pattern of 0s and 1s: $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, and $x_4 = 1$. The total of the person's 1s and 0s is 3. Another person may have two 1s and two 0s and thus a total of 2. Yet another person, albeit someone very unlucky, may have no 1s and four 0s (assume that it is the 1 that improves memory). This person would have a total of 0. The individual totals are the memory scores that we observe. Our group's observed scores will range from 0 to 4. Of course there are many more factors than four which affect performance on a memory test, and not all are either a 0 or a 1. The more factors that are affecting their memories, the greater the range of scores and the greater the variance. If in the current four-factor model you hold one of those four factors constant at 0, however, then the scores will only range from 0 to 3, and the resulting variance will be reduced.

One important goal of experimental control mentioned in the previous chapter is to hold constant as many of the sources of variance as possible, and to reduce the variance in the observed scores. Reducing the variance is important when researchers are evaluating the reliability of the observed differences or the associations. In this regard, variance is a key factor for answering a key question posed in Chapter 1: what is expected due to chance alone?

Web Link 2.13 for an interactive demonstration concerning the source of variance based on the logic of Figure 2.32.

Skewness and kurtosis

Two other descriptive statistics associated with spread are *skewness* and *kurtosis*.

Skewness indicates how symmetrical the observations are about their mean. A value of 0 indicates that the spread of the observations is perfectly symmetrical about their mean. Negative values indicate that the observations below the mean stretch out further than do the observations above the mean. Positive skewness values indicate the opposite. One easy estimate of skewness allows for some conceptual understanding of the measure:

$$\text{Skewness} = 3\,\frac{\text{sample mean} - \text{sample median}}{\text{sample standard deviation}}.$$

Remember that when the distribution is symmetrical, the mean and median will be the same value. Looking at the above formula, it can be seen that if the sample standard deviation is held constant, the greater the difference between the sample mean and the sample median, the greater the skewness. It can also be seen that if the difference between the sample mean and the sample median is held constant, the smaller the sample standard deviation, the greater the skewness.

It is misleading to speak of the 'average' or 'per capita' income or wealth in most societies today. As this fictional distribution in Figure 2.33 illustrates, income is markedly skewed. A few people have much of the wealth and many people have little of the wealth. Such skewness is disclosed by many economic, social, political, and cultural indices. When describing such indices, it is important to be specific about which 'average' is being reported. When data are substantially skewed, it is always best to report all three measures of central tendency.

Kurtosis tells us how clustered or how flat is a set of scores. A kurtosis value of 0 indicates that the observations are mesokurtic, neither overly clustered nor overly flat. Negative values indicate that the spread of the observations is relatively flat. Positive kurtosis values mean that the spread of the observations is clustered and very pointy. The larger the absolute values of skewness and kurtosis, the greater the degree of skewness and kurtosis.

Skewness and kurtosis are related to the famous *normal distribution* (skewness = 0 and kurtosis = 0), a topic dealt with in more detail in the next chapter. At this point, let us simply say that a *normal distribution* is one that is unimodal (the mean, median, and mode, are all the same value), it is symmetrical, and it is neither overly flat (platykurtic) nor overly clustered (leptokurtic). It is the normal distribution that is often assumed for using parametric statistics. Nonparametric statistics make no such assumptions about the distribution of the scores.

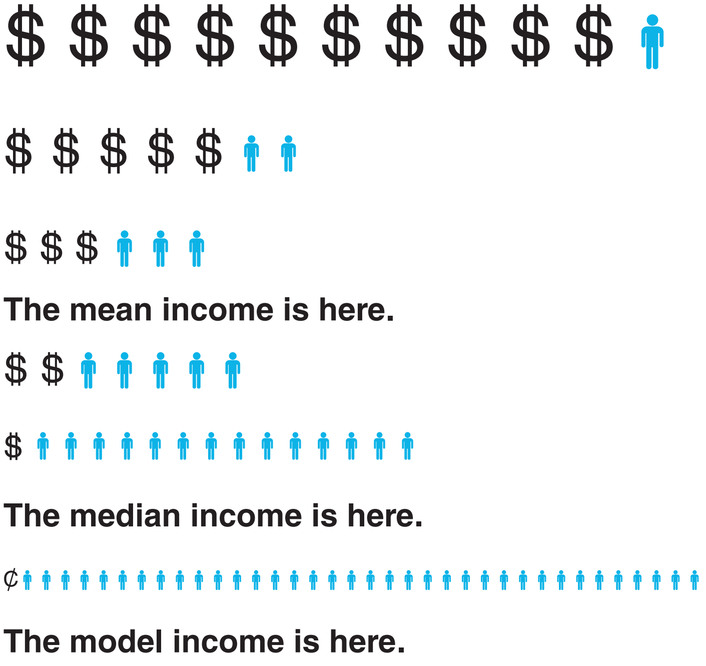
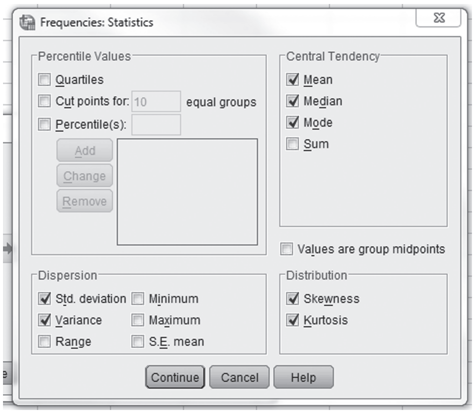


Figure 2.33

Kurtosis as defined originally by Kark Pearson is a measure of the frequency of extreme deviations from the mean and the standard normal distribution would have a Kurtosis of 3. It is common practice now to subtract three from Pearson's measure, giving the standard normal distribution a Kurtosis of zero.



Frequencies: Statistics

Percentile Values
- □ Quartiles
- □ Cut points for: 10 equal groups
- □ Percentile(s):

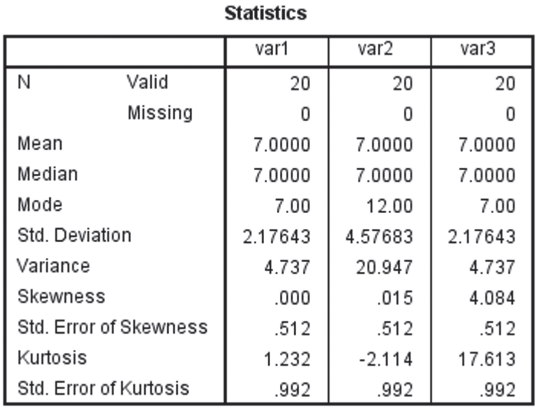
Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- □ Sum

□ Values are group midpoints

Dispersion
- ☑ Std. deviation
- □ Minimum
- ☑ Variance
- □ Maximum
- □ Range
- □ S.E. mean

Distribution
- ☑ Skewness
- ☑ Kurtosis

Figure 2.34



Statistics

		var1	var2	var3
N	Valid	20	20	20
	Missing	0	0	0
Mean		7.0000	7.0000	7.0000
Median		7.0000	7.0000	7.0000
Mode		7.00	12.00	7.00
Std. Deviation		2.17643	4.57683	2.17643
Variance		4.737	20.947	4.737
Skewness		.000	.015	4.084
Std. Error of Skewness		.512	.512	.512
Kurtosis		1.232	-2.114	17.613
Std. Error of Kurtosis		.992	.992	.992

Figure 2.35

Return to the *Data Editor* window where we have entered the data for *var1*, *var2*, and *var3*. Again select *Analyze*, then *Descriptive Statistics*, and then *Frequencies*. In addition to the previous choices, from the *Statistics* menu, select *Skewness* and *Kurtosis* (Figure 2.34).

The output window (Figure 2.35) reveals that *var1* has a skewness of 0.0. Inspecting the observations in *var1*, we see that they are symmetrical about the mean. *var2* has a slight positive skewness: 0.015. Finally, *var3* is considerably skewed: 4.084. The positive kurtosis for *var1* (1.232) indicates that the observations in *var1* are somewhat flat. *var2* (–2.114) is somewhat clustered. Finally, *var3* is substantially clustered (17.613). How important are the deviations from 0 and when are they important? We will address these questions in a later chapter. There are rules of thumb for the severity of skewness and kurtosis. Notice in the SPSS output that along with *Skewness* and *Kurtosis* there are their *standard errors*: *Std. Error of Skewness* and *Kurtosis*. These standard errors are a type of variance, the *expected variances* in the statistics themselves. More pertinent is the fact that these standard errors represent the variability in these statistics that is *expected due to chance alone*. According to one rule of thumb, if you divide the statistic (skewness or kurtosis) by its standard error you can evaluate the severity. One rule of thumb is that if the quotient is greater than ±1.96, then the key assumption for *parametric* tests is called into question. Figure 2.36 presents the results of the evaluations of skewness and kurtosis for the three variables. The skewness in *var3* and the kurtosis in *var2* and *var3* are problematic.

As will be seen in the last few sections of Chapter 3, 1.96 is an important and almost magical number. Many other numbers used for evaluating statistics are derived from it.

	Variable	Statistic	Std. Error	Quotient
Skewness				
	Var1	0.000	0.512	0.00
	Var2	0.015	0.512	0.03
	Var3	4.084	0.512	7.98
Kurtosis				
	Var1	1.232	0.992	1.24
	Var2	–2.114	0.992	–2.13
	Var3	17.613	0.992	17.76

Figure 2.36

An outlier can be, at least in part, responsible for a variable's skewness by creating a tail at one end of the distribution. Addressing the outlier often reduces any problem regarding skewness. Earlier in the chapter we saw how a single score can greatly influence a sample's mean and variance. Returning to the data depicted in the histogram (Figure 2.37) and calculating the mean, variance, skewness, and kurtosis with and without the highest observation (16) makes clear the potential impact of a single observation on skewness and kurtosis.

Simply removing the single observation changes the mean from 7.00 to 6.53; it changes the variance from 4.74 to 0.26; it changes the skewness from 4.08 to –0.12; and it changes the kurtosis from 17.61 to –2.24. Not only are there quantitative changes in the skewness and the kurtosis, but in some cases changes in kind. The skewness went from being positive to negative and the kurtosis went from being clustered to being nearly flat.

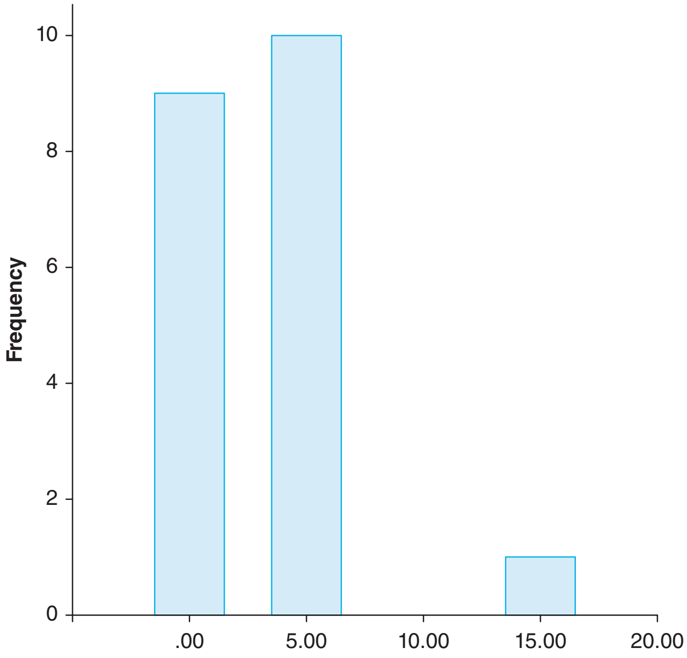


Figure 2.37

2.8 MEASURES OF CENTRAL TENDENCY: CATEGORICAL DATA

Nominal data

For nominal data, where numbers are merely names, the calculation of means and medians makes little sense. If we can change all of the 3s to 1s, change the 1s to 2s, and change the 2s to 3s, what sense does it make to sum all of the 1s, 2s, and 3s and calculate a mean or median? For example, imagine that you are examining the students' choices of beverage in your class. You assign all of those students choosing coffee a score of 1; you assign all of those choosing tea a 2; and you assign all of those choosing juice a 3. Then you find that 10 students chose coffee, 15 chose tea, and 5 chose juice. You might wish to enter these 30 observations into the SPSS *Data Editor* window and run the descriptive statistics of mean and median. If we calculate the mean and median, we find that they are 1.83 and 2.00, respectively. If, however, you recoded the 5 juices as 1, the 10 coffees as 2, and the 15 teas as 3, you find a mean of 2.33 and a median of 2.5. The median goes from being in the tea range to being between tea and coffee.

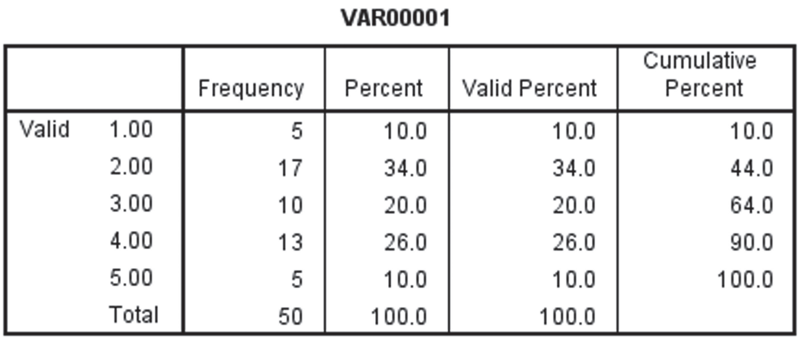
Furthermore, because both the numbers (names) and their rank ordering are arbitrary, it makes no sense to calculate the usual measures of spread such as range and variance. Imagine what would happen not only to the mean and median but also to the range and variance if we decided to name the choice of juice '100'. Would the spread in the choice of beverage really change?

The mode, as an indicator of the most common category of response, does have a role to play in summarizing nominal observations. In this context, however, the mode is not a measure of central tendency; it is simply an indicator of the most common observation. In the beverage example, tea was the most frequently chosen beverage.

Will the real mean and median choice of beverage please stand up? Do we really wish to say that the mean choice of beverage is either 1.83 or 2.33 and that the median choice is either 2.00 or 2.50?

Ordinal data

For ordinal data, where numbers represent a rank ordering of categories, both the mode and the median are useful measures. The mode again reveals the most common category. The numbers



VAR00001

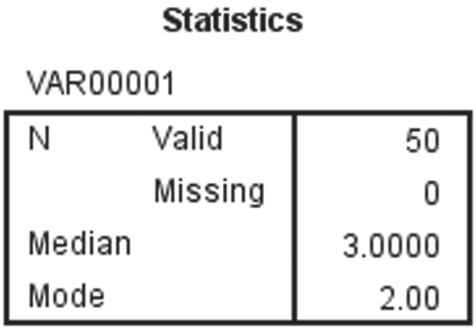
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	5	10.0	10.0	10.0
	2.00	17	34.0	34.0	44.0
	3.00	10	20.0	20.0	64.0
	4.00	13	26.0	26.0	90.0
	5.00	5	10.0	10.0	100.0
	Total	50	100.0	100.0	

Figure 2.38

as names of categories are still arbitrary, but because in an ordinal scale their ordering from smallest to largest is not arbitrary, the median can be useful. The median indicates where in the rank ordering of the observations the middle is to be found. For example, an employer records the level of education of her employees and finds that 10% have not completed high school, 34% have completed high school, 20% spent some time at university, 26% have completed university, and 10% have completed at least one postgraduate programme. The most common category of level of education, or the mode, of her employees is 'completed high school'. Because the median divides the observations in half (the 50th percentile), the median would be the category 'spent some time at university'.

Let us enter the following 50 observations into the SPSS *Data Editor* reflecting the above information, where the coding for not completing high school, completing high school, some university, completed university, and completed postgraduate program is 1, 2, 3, 4, and 5, respectively (the data are rank-ordered for ease of entry and analysis):

> 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5.



Statistics

VAR00001

N	Valid	50
	Missing	0
Median		3.0000
Mode		2.00

Figure 2.39

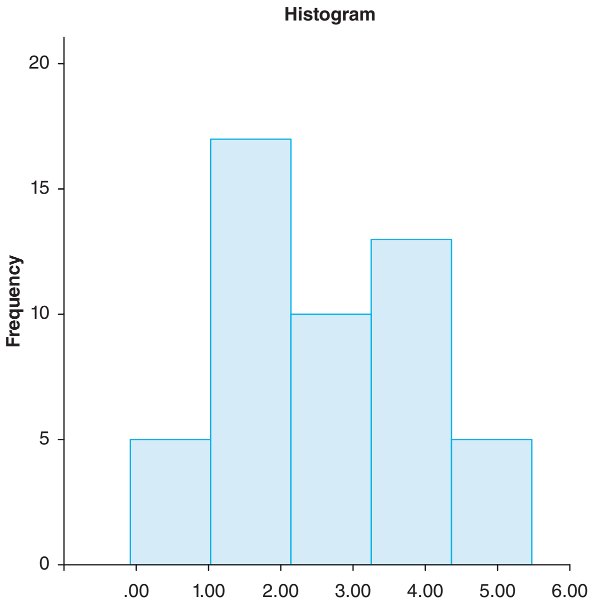


Figure 2.40

Be sure that the *Display Frequency Tables* option is selected along with the statistical options of median and mode. The results in Figure 2.38–2.40 confirm that the mode is category 2 (completed high school) and that the median is category 3 (some university). That is, although the most common level of education is *completed high school*, at least half of her employees had at least some university education.

2.9 MEASURES OF SPREAD: CATEGORICAL DATA

Because both nominal and categorical data have neither equal intervals between categories nor a true zero, the usual measures of spread have little relevance. There have been, however, many attempts to develop single-number indexes of *spread* for categorical data. Many of these indexes involve the relative predominance of the mode. Also, rather than being indices of distance or squared distance from a measure of central tendency, they are indices of the proportion of observations that differ from the mode, relative to the number of categories observed.

Perhaps the simplest measure of spread for nominal data is Freeman's (1965) *variation ratio* (VR). It is most frequently used when there are only two nominal categories, such as men and women. The variation ratio is

$$\mathrm{VR} = 1 - \frac{fm}{N},$$

where *fm* is the frequency of the modal category and *N* is the total number of observations. Thus, if the modal category is 60 out of a 100 observations, then $fm/N = 60/100 = 0.6$. The resulting VR = 1 – 0.6 = 0.4. The greater the VR, the more equally distributed are the observations. If the modal category in the previous example was 75 out of 100, then $\mathrm{VR} = 1 - \frac{75}{100} = 1 - 0.75 = 0.25$, less equally distributed than the previous example.

Another useful index of spread for categorical data that is applicable for both nominal and ordinal scales is the *index of qualitative variation* (IQV):

$$\mathrm{IQV} = K\,(100^2 - \sum pct^2)\,/\,\{100^2(K-1)\},$$

where *K* is number of categories and Σpct^2 is the sum of the squared frequencies of all of the categories. The IQV provides a standardized value between 0 and 1: a 0 indicates a complete lack of diversity where one category totally dominates; a 1 indicates the maximum amount of diversity where all categories have the same frequency. When the IQV is multiplied by 100, the result can be interpreted as the percentage of the maximum possible diversity present in the observations.

In the above examples of employees' education level the researcher finds the following IQV:

$$\mathrm{IQV} = 5\big((10{,}000) - \sum(10^2 + 34^2 + 20^2 + 26^2 + 10^2)\big)\,/\,(10{,}000 \times 4) = 0.946 = 94.6\%$$

This means that 94.6% of the diversity possible in the observations is present. An examination of the data's histogram illustrates that although there is a clear mode, it certainly does not dominate the distribution of the observations.

CHALLENGE QUESTION

Would changing the numbers that label the categories in the above education level example to 1, 3, 5, 7, and 9 change the median and IQV? If you think the change will alter the median and IQV, why and how? If you think the change will make no difference, why not?

Web Link 2.14 for an answer to the challenge question.

2.10 UNBIASED ESTIMATORS

μ is the value to *expect* (best guess) when an observation is randomly selected from a population, and $\bar{y}$ is an estimation of μ. s^2, often called *error*, describes the *observed* unpredictability of the observations which are, at least initially, *due to chance alone*. The word 'initially' in the previous sentence captures the starting point for all forms of data analysis. Although it is unexplained, what is *due to chance alone* is not *unexplainable*. One way to understand empirical research is to frame it as an effort to explain some of the as yet unexplained: an attempt to reduce the *error* in our predictions. The reason for s^2 is not supernatural. s^2 has its sources, and the goal of research is to identify those sources.

In the above section on variance and standard deviation, $n - 1$ rather than n was used in the denominator to calculate the average sum of squared deviations or variance. Also, the mean and the standard deviation were both found to be sensitive to extreme scores or outliers. Although the mean and the variance are not *resistant* statistics, they are important when conducting a parametric analysis. The reason for their importance is that, unlike other measures of central tendency and spread, the sample mean ($\bar{y}$) and variance (s^2) – with $n - 1$ in the denominator – are unbiased estimators of their corresponding population parameters, μ and σ^2.

An unbiased estimator is a statistic whose expected value (E) is the true population parameter. An E is a type of mean. It is a mean of a statistic rather than a mean of individual observations. Furthermore, it is the mean of an infinite number of instances of a statistic or of all possible instances of a statistic. The sample size for these instances must be held constant.

Without going through the algebraic proof here, it may be worthwhile to illustrate in other ways the unbiased nature of the sample mean and variance. First, let us examine a very small population of three observations: 101, 102, and 103. Other authors, such as Howell (2002), have found this type of empirical example to be effective for illustrating the point. The μ of the population is (101 +102 +103)/3 = 102. The σ^2 of the population is $[(101 - 102)^2 + (102 - 102)^2 + (103 - 102)^2]/ 3 = (1 + 0 + 1)/3 = 0.67$. These are the true μ and true σ^2 of the small population. The first column of Figure 2.41 shows all possible randomly selected samples of two observations from the population in question. The second column reports the mean of each of those samples. The third column reports the variance as calculated with n in the denominator. The fourth column reports the sample variance with $n - 1$

Sample	$\bar{y}$	$s^2(n)$	$s^2(n–1)$
101,101	101.00	0.00	0.00
101,102	101.50	0.25	0.50
101,103	102.00	1.00	2.00
102,101	101.50	0.25	0.50
102,102	102.00	0.00	0.00
102,103	102.50	0.25	0.50
103,101	102.00	1.00	2.00
103,102	102.50	0.25	0.50
103,103	103.0	0.00	0.0
E =	102.0	0.33	0.67

Figure 2.41

in the denominator. The bottom row of the table reports the expected values (E) for the sample mean and variance from both forms of calculation. In this case the expected value is the mean of all possible two-observation samples as reported in the table.

Note that the E($\bar{y}$) = μ: 102.0. Also note that the average s^2 with $n - 1$ in the denominator equals σ^2 (0.67), whereas the average s^2 with n in the denominator underestimates σ^2 (0.33). This does not mean that a sample variance with $n - 1$ in the denominator will always be closer to the population's true value. This systematic underestimate when n is used in the denominator is related to the fact that each sample mean minimizes the sum of squared deviations for each individual sample.

Any value (including μ) other than $\bar{y}$ results in a greater total of the squared deviations. If μ is known, why use a different estimate for each sample? In fact, we should not. Look at Figure 2.42. The first column of Figure 2.42 again enumerates all possible randomly selected two-observation samples from the population in question. The second column reports μ, which here is used to calculate all of the sample variances. The third column reports the variance for each sample as calculated with n in the denominator. The fourth column reports the sample variance with $n - 1$ in the denominator. Again, the bottom row of the table reports the expected values (E) for the sample mean and variance from both forms of calculation. Not surprisingly, again E($\bar{y}$) = μ: 102.0. Now, however, the s^2 version with $n - 1$ in the denominator on average overestimates $\sigma^2(1.33)$, whereas the s^2 version with n in the denominator now equals $\sigma^2(0.67)$. These two tables illustrate that $n - 1$ in the denominator of sample variance is there to correct for the fact that we are estimating μ with $\bar{y}$. Should μ be known, there would be no need to make the correction.

Sample	μ	$s^2(n)$	$s^2(n-1)$
101,101	102.00	1.00	2.00
101,102	102.00	0.50	1.00
101,103	102.00	1.00	2.00
102,101	102.00	0.50	1.00
102,102	102.00	0.00	0.00
102,103	102.00	0.50	1.00
103,101	102.00	1.00	2.00
103,102	102.00	0.50	1.00
103,103	102.00	1.00	2.00
E =	102.00	0.67	1.33

Figure 2.42

CHALLENGE QUESTION

How does sample size affect the extent of bias when n rather than $n - 1$ is used in the denominator of the variance formula?

Web link 2.15 for the answer to and discussion of the challenge question.

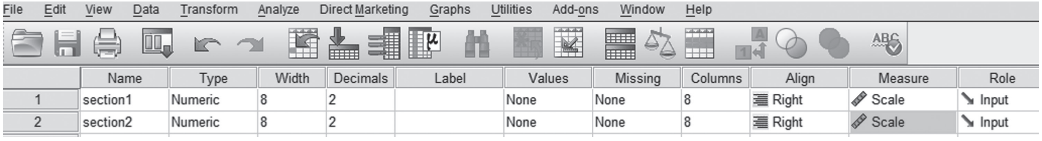
2.11 PRACTICAL SPSS SUMMARY

We now return to the problems that were posed at the outset of this chapter. You were given hypothetical quiz scores from two sections of a statistics course. You were asked, if the scores were read to you, would you have a sense of how each section performed and how their performances compared?

> Section 1: 7, 8, 5, 5, 9, 10, 2, 7, 6, 4, 8, 7, 8, 7, 3, 9, 6, 8, 6, 9, 8, 9, 5, 7, 6, 7, 6, 7, 9, 5, 5, 9, 7, 11, 7, 6, 4, 7, 8, 10.
>
> Section 2: 6, 3, 6, 8, 5, 7, 6, 7, 8, 7, 8, 5, 5, 8, 7, 2, 7, 6, 4, 7, 7, 8, 5, 5, 8, 10, 1, 7, 6, 4, 8, 7, 2, 7, 3, 4, 6, 8, 10, 8.

You begin by summarizing and exploring each section's 40 scores separately. In many cases it is best to begin with frequency histograms.

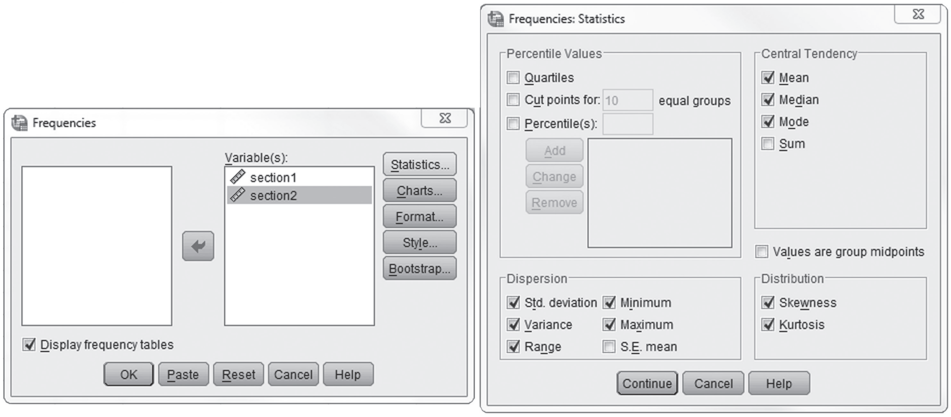
Enter the data from the two variables into the SPSS *Data Editor* window, if you have not done so already. Temporarily switch to the *Variable View* (tab at lower left of screen) and ensure that the variables are listed as *Numeric* under *Type* and as *Scale* under *Measure* (Figure 2.43). The number of items correct on the quiz is an example of a ratio scale. SPSS does not differentiate ratio and interval scales. The term *Scale* under the heading *Type* is used for either ratio or interval data.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	section1	Numeric	8	2		None	None	8	Right	Scale	Input
2	section2	Numeric	8	2		None	None	8	Right	Scale	Input

Figure 2.43

After returning to the *Data Editor* window, select the *Frequencies* option from *Descriptive Statistics* under the *Analyze* menu. Highlight and move both variables over to the *Variable(s)* field (Figure 2.44). After opening the *Statistics* window select the *Mean*, *Median*, and *Mode*



Frequencies

Variable(s):
section1
section2

☑ Display frequency tables

Frequencies: Statistics

Percentile Values
☐ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

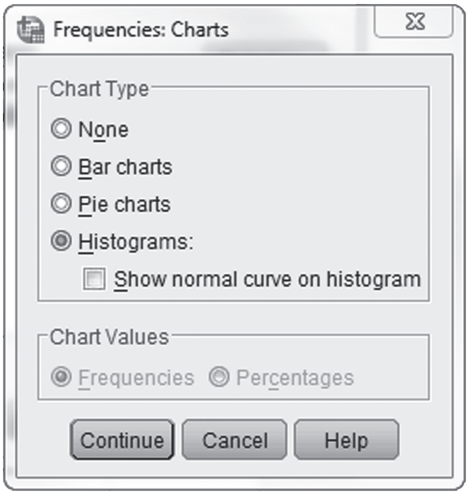
Central Tendency
☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

Dispersion
☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☑ Range ☐ S.E. mean

Distribution
☑ Skewness
☑ Kurtosis

Figure 2.44



Frequencies: Charts

Chart Type

○ None

○ Bar charts

○ Pie charts

● Histograms:

□ Show normal curve on histogram

Chart Values

● Frequencies ○ Percentages

Figure 2.45

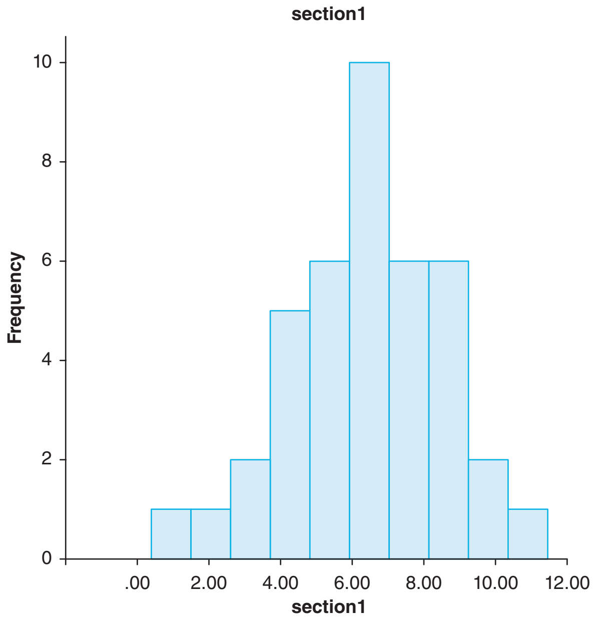


Figure 2.46

central tendency options. Also select the *Std. deviation*, *Variance*, *Range*, *Minimum*, and *Maximum* from the dispersion options. Finally, from the distribution area, select the *Skewness* and *Kurtosis* options.

Press *Continue* and return to the *Frequencies* window and open the *Charts* window. Select the *Histograms* chart type (Figure 2.45) and then *Continue*. When you return to the *Frequencies* window click *OK*.

Scroll down the output window and find the histograms. The histogram for *section1* clearly is unimodal (Figure 2.46) with most scores being clustered in the middle. It appears to be quite symmetrical. The histogram does not suggest the presence of any unusually low or high scores well beyond the others (possible outliers). The histogram for *section2* also is unimodal (Figure 2.47). Its distribution is less symmetrical than that of *section1*, however, and it appears to be negatively skewed. Like the *section1* histogram, the *section2* histogram does not suggest the presence of extreme scores or outliers.

This type of histogram is appropriate for the type of data in these two variables. When there are many possible scores (*x*-axis categories) and very few observations in any one score, binning is necessary. Frequency histograms are not very suitable when there are very few observations (e.g., less than 20). In such cases researchers rely solely on the descriptive statistics.

Next the descriptive statistics are examined (Figure 2.48). The fact that all three measures of central tendency (mean, median, and mode) are so similar indicates that *section1* is symmetrically distributed. Using 4 standard deviations as the criterion, the scores can be examined for outliers. None are found.

The skewness of *section1*'s distribution divided by the standard error is –0.303/0.374 = –0.81. The kurtosis of *section1*'s distribution divided by the standard error is 0.082/0.733 = 0.11. Because both values are less than 1.96, you may conclude that there is no problem with either of these measures.

The scores for *section2* had a mean of 6.15, a median of 7.00, and a mode of 7.00. The fact that the mean is somewhat lower that the median and the mode indicates that *section2*'s scores may be somewhat asymmetrically distributed. The skewness of *section2*'s distribution divided by the standard error is –0.605/0.374 = –1.62. The kurtosis of *section2*'s distribution divided by the standard error is 0.122/0.733 = 0.17. The skewness for *section2* is greater than the corresponding values for *section1*, as initially suggested by the histograms. Because both values again are less than 1.96, you may conclude that there is no problem with either of these measures.

At the outset of Chapter 1 it was asked if ten digits had been randomly sorted into two columns or groups. Here we are faced with a more concrete incarnation of that problem: has one group's performance been statistically different from the other? Imagine the two sections were given different textbooks from which to study. Did the difference in the textbook make a difference in performance? Some of the descriptive statistics suggest that there was no difference in performance: the medians, modes and ranges are identical, and the variances and standard deviations are quite similar. The difference in the means (6.93 versus 6.15) and the fact that the minimum and maximum scores are higher in *section1* than in *section2* suggest a difference in performance. But of course, as we know from Chapter 1, these differences could be due chance alone. The question is, how reliable are those differences? Can we conclude that textbook makes a difference?

To answer these inferential questions, we need procedures for determining the likelihood of observing such a difference (e.g., 6.93 – 6.15 = 0.78). To answer such inferential questions, it is necessary to combine descriptive statistics with procedures for calculating probabilities, which is the focus of the next chapter.

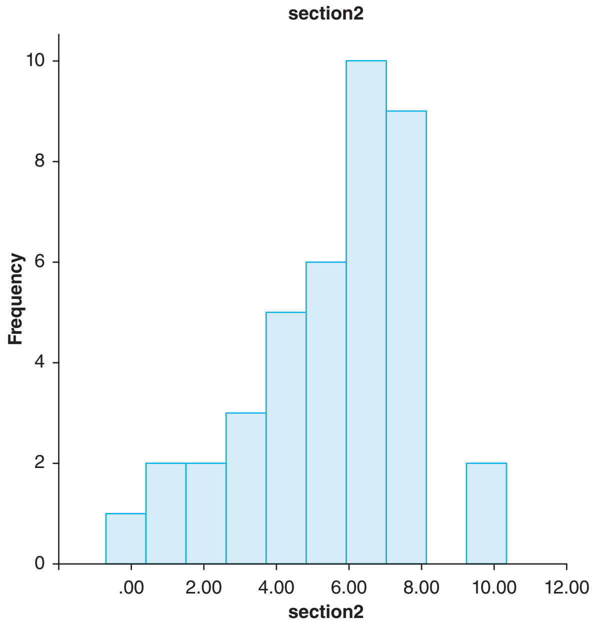
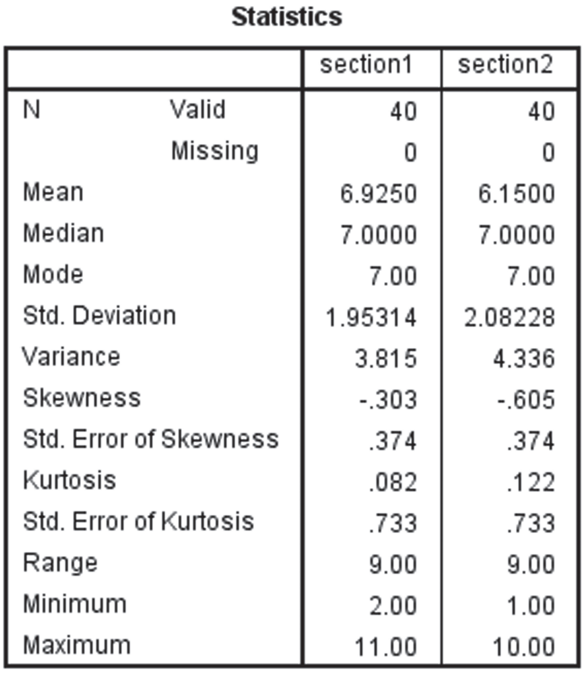


Figure 2.47



Statistics

		section1	section2
N	Valid	40	40
	Missing	0	0
Mean		6.9250	6.1500
Median		7.0000	7.0000
Mode		7.00	7.00
Std. Deviation		1.95314	2.08228
Variance		3.815	4.336
Skewness		-.303	-.605
Std. Error of Skewness		.374	.374
Kurtosis		.082	.122
Std. Error of Kurtosis		.733	.733
Range		9.00	9.00
Minimum		2.00	1.00
Maximum		11.00	10.00

Figure 2.48

2.12 CHAPTER SUMMARY

In this chapter the most common *univariate* graphs and statistics used to summarize both *measurement* and *categorical* data were examined. Measurement data include *interval* and *ratio* number scales; categorical data include *nominal* and *ordinal* scales. This chapter focused on the importance of descriptively examining data as a first step in the process of analysis. Various versions of the most important univariate graphic display, the *histogram*, were explored. The most common *descriptive statistics* were defined and compared. For measurement data, the *mean* and the *median* are two important *measures of central tendency*. For categorical data, the *mode* is the most common measure of central tendency. In terms of *measures of spread*, while *variance*, being an unbiased estimator, is central for measurement data, the *variation ratio* is the typical measure of spread for categorical data.

The importance of the shape of a variable's distribution was discussed and the need to examine the data for *outliers* was examined. The notions of *skewness* and *kurtosis*, indices of the symmetry and the flatness of a distribution respectively, were considered. Of particular note was the distorting effect of outliers and skewness on the mean and the variance. The material covered in this chapter provides half of the fundamentals for what is needed to understand the *inferential statistical* tests presented in Parts II and III of this book. The other half of the fundamentals is the topic of the next chapter: probability.

2.13 RECOMMENDED READINGS

Phillips, J. L. (1999). *How to think about statistics* (6th ed.). New York: W. H. Freeman.

Chapters 1 –4 provide a simple summary of the core material covered in this chapter.

Weisberg, H. F. (1992). *Central tendency and variability*. London: Sage.

Weisberg's book provides more detail about some of the descriptive statistics discussed in this chapter. He also covers some descriptive measures not presented in this chapter.

Wheelan, C. (2013). *Naked statistics: Stripping the dread from the data*. New York: W.W. Norton.

Wheelan provides a light-hearted approach to the material covered in this and subsequent chapters, along with examples from popular culture.

Holcomb, Z. C. (1998). *Fundamentals of descriptive statistics*. New York: Pyrczak Publishing.

This book is almost entirely dedicated to a detailed exposition of the graphs and statistics presented in this chapter. It elaborates on everything from graphs to *z*-scores, from means to outliers.

2.14 CHAPTER REVIEW QUESTIONS

Multiple-choice questions

1 A researcher is interested in examining the voting behaviour of individuals in a small town. He contacted those eligible to vote to set up interviews with them. Of the people living in the town 7000 are eligible to vote. The researcher contacted 5000 of them; 5% of those contacted agreed to an interview with the researcher. What is the population?

- **a** Everyone in the small town
- **b** The 7000 eligible voters
- **c** The 5000 individuals contacted
- **d** The 250 individuals who were interviewed
- **e** None of the above

2 A researcher is interested in examining the voting behaviour of individuals in a small town. He contacted those eligible to vote to set up interviews with them. Of the people living in the town 7000 are eligible to vote. The researcher contacted 5000 of them; 5% of those contacted agreed to an interview with the researcher. What is the sample?

- **a** Everyone in the small town
- **b** The 7000 eligible voters
- **c** The 5000 individuals contacted
- **d** The 250 individuals who were interviewed
- **e** None of the above

3 During the interviews the researcher questions the interviewees about their income and how many times they had voted previously. Income is what type of variable?

- **a** Ratio and discrete
- **b** Nominal and discrete
- **c** Nominal and continuous
- **d** Interval and continuous
- **e** Categorical and discrete

4 During the interviews the researcher questions the interviewees about their income and how many times they had voted previously. Previous voting behaviour is what type of variable?

- **a** Ratio and discrete
- **b** Nominal and discrete
- **c** Nominal and continuous
- **d** Interval and continuous
- **e** Categorical and discrete

5 Counting the number of patients who are categorized into one of several diagnostic categories for the sake of comparison is an example of ____________________.

- **a** a continuous variable
- **b** a categorical data
- **c** measurement data
- **d** an ordinal scale
- **e** a leptokurtic scale

6 If we attached numbers to the labels for the disorders used in Question 2, those numbers would be an example of _________________.

- **a** an ordinal scale
- **b** frequency data
- **c** a nominal scale
- **d** a ratio scale
- **e** a continuous variable

7 The ______________ is more sensitive to outliers than is the _____________.

- a median; mean
- b mode; median
- c mode; mean
- d a continuous variable; a discrete variable
- e standard deviation; mode

8 The most common measure of central tendency for nominal data is the _______.

- a median
- b mean
- c variance
- d variation ratio
- e mode

9 A common measure of spread for nominal and ordinal data is the _________.

- a median
- b standard error of the mean
- c variance
- d variation ratio
- e mode

10 When a distribution is overly flat it is said to be _________.

- a positively skewed
- b negatively skewed
- c leptokurtic
- d platykurtic
- e bimodal

11 When a distribution is positively skewed, the ______ will be greater than the _______.

- a mean, median
- b mode, median
- c mean, variance
- d mode, median
- e None of the above

12 The sampling distribution of the mean indicates ______.

- a how much variance there is in your data due to chance alone
- b how much variance in your observed mean is due to chance alone
- c how much variance you expect due to chance alone in means sampled from the same population
- d how much variance you expect due to chance alone in observations sampled from the same population
- e how much variance you expect due to chance alone in variances sampled from the same population

Short-answer questions

1 When could we use *n* rather than *n* – 1 in the denominator for sample variance? Why?

2 What is the difference between a frequency histogram and a relative frequency histogram?

3 When binning data is required for a histogram, what determines the number of bins?

4 What is the difference between a variance and a sampling distribution?

5 What causes variance?

6 What does it mean to say a statistic is resistant?

7 What does it mean to say a statistic is unbiased?

8 When will a single new observation added to a data set leave the mean unchanged?

9 What is the primary difference between parametric and nonparametric statistics?

10 Why is the mean not very informative when a distribution is bimodal?

Data questions

1 With the following data, construct a frequency distribution table and a frequency histogram with bin widths of 10. Observations: 44, 46, 47, 49, 63, 64, 66, 68, 72, 72, 75, 76, 81, 84, 88.

2 With the data below, create a frequency histogram with five categories (bins) on the *x*-axis. Data: 24, 21, 2, 5, 8, 11, 13, 18, 17, 21, 20, 20, 12, 12, 10, 3, 6, 15, 11, 15, 25, 11, 14, 1, 6, 3, 10, 7, 19, 17, 18, 9, 18, 12, 15.

3 What are the mean, the variance, sampling distribution of the mean, and the standard error of the mean for the data in Question 2?

4 For the data in Question 2, are the skewness and kurtosis values a concern for the researcher who is assuming a normal distribution? (You will need SPSS to answer this question.)

5 Create a population of three numbers (e.g., 10, 11, 12). Then analyse *all* possible samples of two, including samples such as 10 and 10. For all samples calculate variance using both *n* and *n* – 1. Then repeat this analysis using the population mean for each calculation, rather than the individual sample means. In the two series of analyses, which formula (using *n* or *n* – 1) produces an unbiased estimator and why?

6 Students are often asked to rate their professor, typically on a 1–5 scale, 1 being the lowest ranking and 5 being the highest. In an educational psychology class of 25 students, 3 gave their instructor a rating of 1, 4 students gave a rating of 2, 8 students gave a rating of 3, 7 students gave a rating of 4, and 3 students gave a rating of 5.

(a) What are the mean and median ratings?

(b) What are the variance and standard deviation of the ratings?

(c) What might be a problem with computing the statistics in (a) and (b)?

(d) What are alternative descriptive statistics for those in (a) and (b)?

7 A charity hired three groups of clowns (balloon-twisters, magicians, and jugglers) to perform at a fund raising event. Figure 2.49 shows the number of clowns and the average amount of donations (per clown) raised by the three groups. The jugglers raised a total of $800.

Clown type	No. of clowns	Average $/clown
Balloon-twister	20	75
Magician	20	70
Juggler		80

Figure 2.49

(a) How many clowns were there in total?

(b) What was the total amount of donations raised by the three groups?

(c) How much did the average clown raise?

8 Create two distributions with identical means, medians, and ranges. One distribution should be platykurtic and the other leptokuric.

9 There are three sections of quiz scores in your class. One has 10 students and a mean of 7. The second has 5 students and a mean of 9. The third has only 5 students and a mean of 5. What is the composite or grand mean of the 20 students?

10 If you took the three means in Question 9 (7, 9, and 5) and simply divided by 3 (the number of sections), how would that compare with the composite mean computed in Question 9. Why?