Chapter 3
Analysis of Quantitative Data

Contents

3.1 Introduction 59
3.2 What Is Data Analysis? 60
3.3 Data Analysis Tools 61
3.4 Data Analysis for Two Data Sets 64
3.4.1 Time Series Data: Visual Analysis 66
3.4.2 Cross-Sectional Data: Visual Analysis 68
3.4.3 Analysis of Time Series Data: Descriptive Statistics 71
3.4.4 Analysis of Cross-Sectional Data: Descriptive Statistics 72
3.5 Analysis of Time Series Data: Forecasting/Data Relationship Tools 75
3.5.1 Graphical Analysis 76
3.5.2 Linear Regression 80
3.5.3 Covariance and Correlation 86
3.5.4 Other Forecasting Models 87
3.5.5 Findings 88
3.6 Analysis of Cross-Sectional Data: Forecasting/Data Relationship Tools 89
3.6.1 Findings 96
3.7 Summary 97
Key Terms 98
Problems and Exercises 99

3.1 Introduction

In Chap. 2, we explored types and uses of data, and we also performed data analysis on quantitative data with graphical techniques. We continue our study of data analysis, particularly, the analysis of quantitative data through the use of statistical methods. This chapter will delve more deeply into the various tools for quantitative data analysis contained in Excel, providing us with a strong foundation and a preliminary understanding of the results of a data collection effort. Not all Excel statistical tools will be introduced, but more powerful tools will follow in later chapters.

H. Guerrero, *Excel Data Analysis*, https://doi.org/10.1007/978-3-030-01279-3_3

3.2 What Is Data Analysis?

If you perform an internet web search on the term "Data Analysis," it will take years for you to visit every site that is returned, not to mention encountering a myriad of different types of sites, each claiming the title of "data analysis." Data analysis means many things to many people, but its goal is universal: to answer one very important question—what does the data reveal about the underlying system or process from which the data is collected? For example, suppose you gather data on customers that shop in your retail operation—data that consists of detailed records of purchases and demographic information on each customer transaction. As a data analyst, you may be interested in investigating the buying behavior of different age groups. The data might reveal that the dollar value of purchases by young men is significantly smaller than those of young women. You might also find that one product is often purchased in tandem with another. These findings can lead to important decisions on how to advertise or promote products. When we consider the findings above, we may devise sales promotions targeted at young men to increase the value of their purchases, or we may consider the co-location of products on shelves that makes tandem purchases more convenient. In each case, the decision maker is examining the data for clues of the underlying behavior of the consumer.

Although Excel provides you with numerous internal tools designed explicitly for data analysis, some of which we have seen already, the user is also capable of employing his own ingenuity to perform many types of analytical procedures by using Excel's basic mathematical functions. Thus, if you are able to understand the basic mathematical principles associated with an analytical technique, there are few limits on the type of techniques that you can apply. This is often how an **add-in** is born: an individual creates a clever analytical application and makes it available to others.

An add-in is a program designed to work within the framework of Excel. They use the basic capabilities of Excel (for example, either Visual Basic for Applications (VBA) or Visual Basic (VB) programming languages) to perform internal Excel tasks. These programming tools are used to automate and expand Excel's reach into areas that are not readily available. In fact, there are many free and commercially available statistical, business, and engineering add-ins that provide capability in user-friendly formats.

Now, let us consider what we have ahead of us in Chap. 3. We are going to focus on the built-in data analytical functionality of Excel and how to apply it to quantitative data. Also, we will carefully demonstrate how to apply these internal tools to a variety of data, but throughout our discussions, it will be assumed that the reader has a rudimentary understanding of statistics. Furthermore, recall that the purpose of this chapter (and book, for that matter) is not to make you into a statistician, but rather, to give you some powerful tools for gaining insight about the behavior of data. I urge you to experiment with your own data, even if you just make it up, to practice the techniques we will study.

3.3 Data Analysis Tools

There are a number of approaches to perform data analysis on a data set stored in an Excel workbook. In the course of data analysis, it is likely that all approaches will be useful, although some are more accessible than others. Let us take a look at the three principle approaches available:

1. Excel provides resident add-in utilities that are extremely useful in basic statistical analysis. The *Data* ribbon contains an *Analysis* group with almost 20 statistical *Data Analysis* tools. Figure 3.1 shows the location of the *Data Analysis* add-in tool, and Fig. 3.2 shows some of the contents of the *Data Analysis* menu. These tools allow the user to perform relatively sophisticated analyses without having to create the mathematical procedures from basic cell functions; thus, they usually require interaction through a dialogue box as shown in Fig. 3.2. Dialogue boxes are the means by which the user makes choices and provides instructions, such as entering parameter values and specifying ranges containing the data of interest. In Fig. 3.2 we see a fraction of the analysis tools available, including *Descriptive Statistics*, *Correlation*, etc. You simply select a tool and click the OK button.

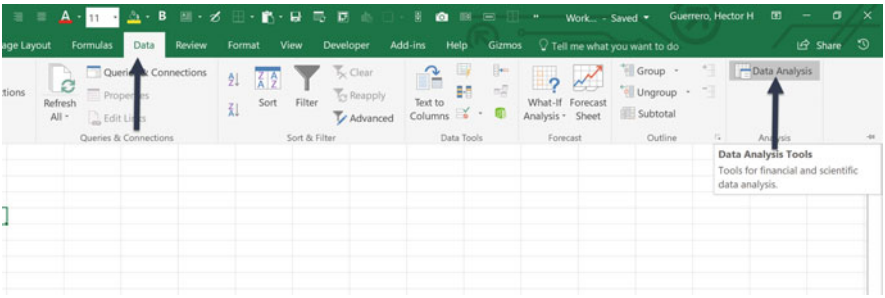


Fig. 3.1 Data analysis add-in tool

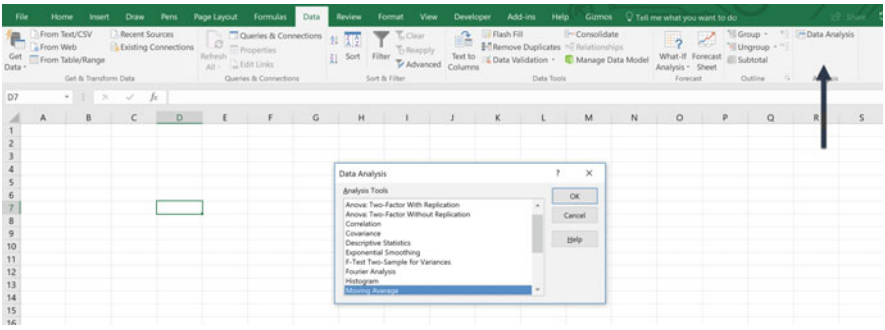


Fig. 3.2 Data analysis dialogue box

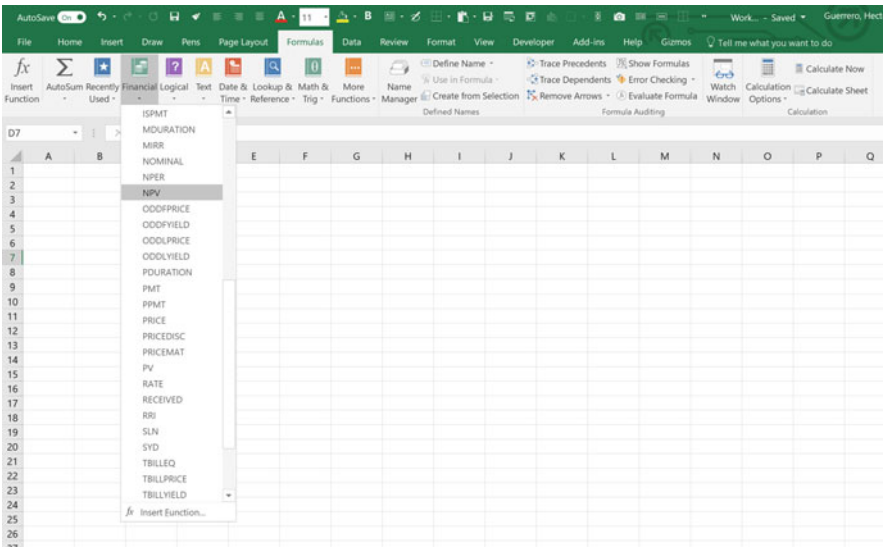


Fig. 3.3 The insert function

There will be more on this process in the next section—*Data Analysis for Two Data Sets*.

2. In a more direct approach to analysis, Excel provides dozens of statistical functions through the function utility (***fx*** Insert Function) in the *Formulas* ribbon. Simply choose the *Statistical* category of functions in the *Function Library* group, select the function you desire, and insert the function in a cell. The statistical category contains almost 100 functions that relate to important theoretical data distributions and statistical analysis tools. In Fig. 3.3, you can see that the Financial function category has been selected, NPV (net present value) in particular. Once the function is selected, Excel takes you to a dialogue box for insertion of the NPV data, as shown Fig. 3.4. The dialogue box requests two types of inputs—discount rate (Rate) and values (Value1, Value2, etc.) to be discounted to the present. The types of functions that can be inserted vary from Math & Trig, Date and Time, Statistical, Logical, to even Engineering, just to name a few. By selecting the *fx* Insert Function at the far left of the Function Library group, you can also select specific, or often used, functions. Figure 3.5 shows the dialogue box where these choices are made from the *Or select a category*: pull-down menu. As you become familiar with a function, you need only begin the process of keying in the function into a cell, preceded by an equal sign; thus, the process of selection is simplified. You can see from Fig. 3.6 that by placing the cursor in cell C3 and typing = *NPV(*, a small box opens that guides you through the data entry required by the function. The process also provides **error checking** to ensure that your data entry is correct.
3. Finally, there are numerous commercially available add-ins—functional programs that can be loaded into Excel that permit many forms of sophisticated analysis. For example, Solver is an add-in that is used in constrained optimization.

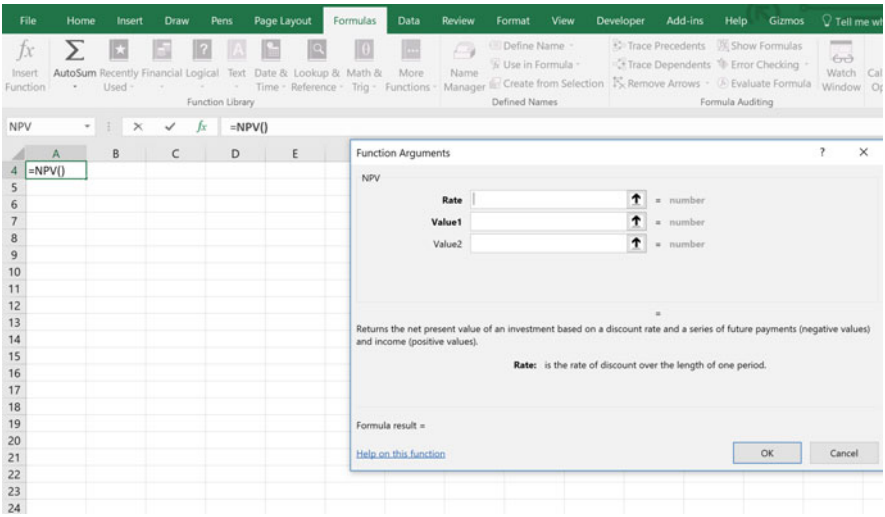


Fig. 3.4 Example of a financial NPV function

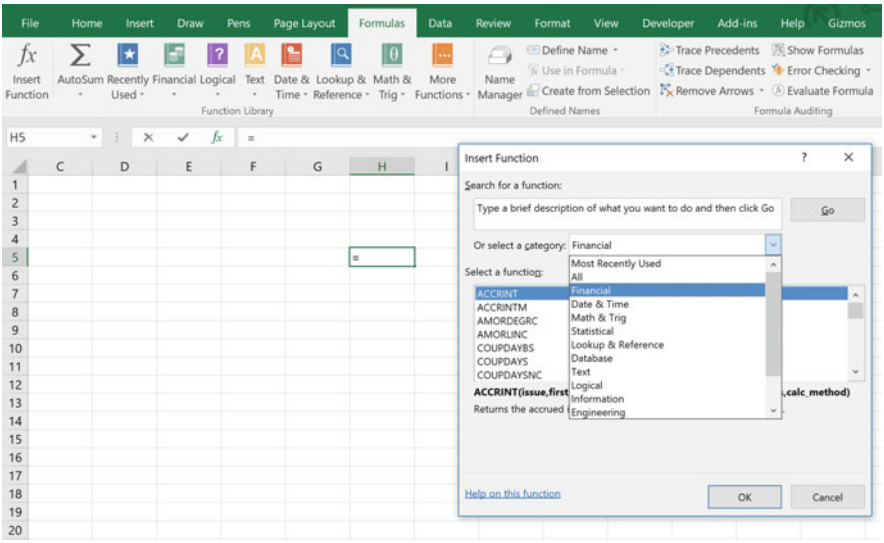


Fig. 3.5 Function categories

Although it is impossible to cover every available aspect of data analysis that is contained in Excel in this chapter, we will focus on techniques that are useful to the average entry-level user, particularly those discussed in (1) above. Once you have mastered these techniques, you will find yourself quite capable of exploring many

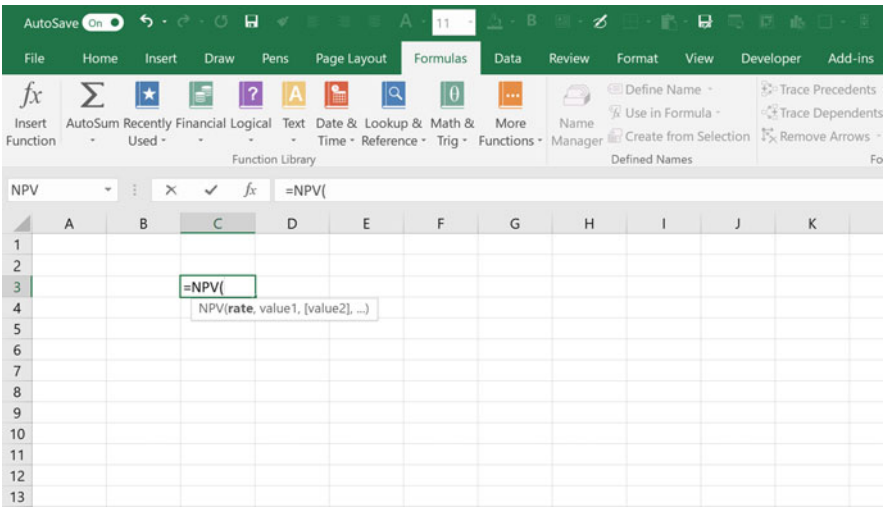


Fig. 3.6 Typing-in the NPV function in a cell

other more advanced techniques on you own. The advanced techniques will require that you have access to a good advanced statistics and/or data analytics reference.

3.4 Data Analysis for Two Data Sets

Let us begin by examining the *Data Analysis* tool in the *Analysis* group. These tools (regression, correlation, descriptive statistics, etc.) are statistical procedures that answer questions about the relationship between multiple data *series*, or provide techniques for summarizing characteristics of a single data set.

A **series**, as the name implies, is a series of data points that are collected and ordered in a specific manner. The ordering can be chronological or according to some other **treatment**: a characteristic under which the data is collected. For example, a treatment could represent customer exposure to varying levels of media advertising. These tools are useful in prediction or the description of data. To access Data Analysis, you must first enable the Analysis ToolPak box by opening the Excel Options found in the File group. Figure 3.7 shows the location of (1) the File group, and (2) Options menu. In Fig. 3.8, the arrows indicate where the menus for selecting the *Analysis ToolPak* can be found. Once enabled, a user has access to the *Analysis ToolPak.*

We will apply these tools on two types of data: **time series** and **cross-sectional**. The first data set, time series, is data that was introduced in Chap. 2, although the data set has now been expanded to provide a more complex example. Table 3.1 presents sales data for five products, A–E, over 24 quarters (6 years) in thousands of

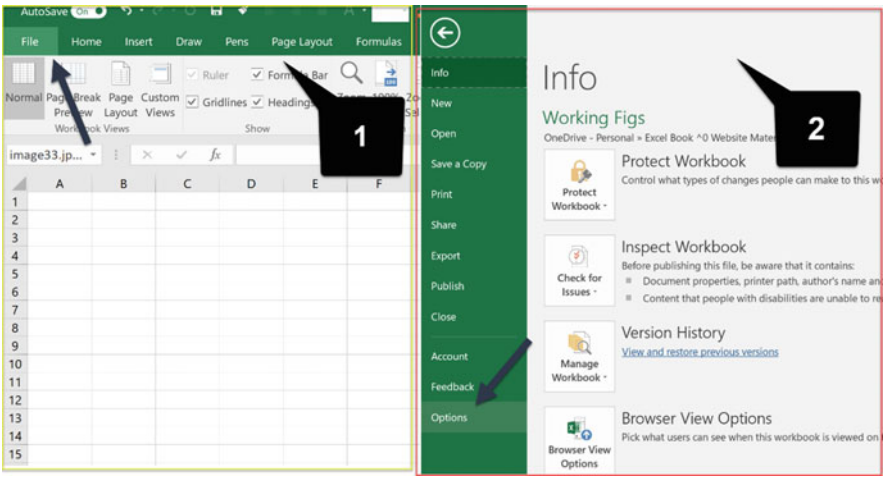


Fig. 3.7 Excel options in the file ribbon

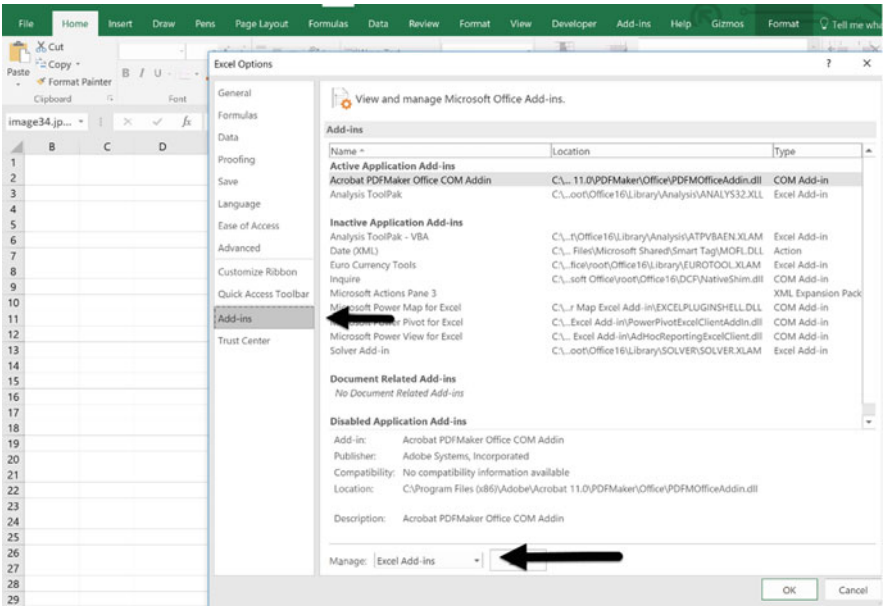


Fig. 3.8 Enabling the analysis ToolPak add-in

dollars. In Fig. 3.9, we use some of the graphing skills we learned in Chap. 2 to display the data graphically. Of course, this type of visual analysis is a preliminary step that can guide our efforts for understanding the behavior of the data and

Table 3.1 Sales[a] data for products A–E

Quarter	A	B	C	D	E
1	98	45	64	21	23
2	58	21	45	23	14
3	23	36	21	31	56
4	43	21	14	30	78
1	89	49	27	35	27
2	52	20	40	40	20
3	24	43	58	37	67
4	34	21	76	40	89
1	81	53	81	42	34
2	49	27	93	39	30
3	16	49	84	42	73
4	29	30	70	46	83
1	74	60	57	42	43
2	36	28	45	34	32
3	17	52	43	45	85
4	26	34	34	54	98
1	67	68	29	53	50
2	34	34	36	37	36
3	18	64	51	49	101
4	25	41	65	60	123
1	68	73	72	67	63
2	29	42	81	40	46
3	20	73	93	57	125
4	24	53	98	74	146

[a]Thousands of dollars

suggesting further analysis. A trained analyst can find many interesting leads to the data's behavior by creating a graph of the data; thus, it is always a good idea to begin the data analysis process by graphing the data.

3.4.1 Time Series Data: Visual Analysis

Time series data is data that is chronologically ordered and one of the most frequently encountered types of data in business. Cross-sectional data is data that is taken at a single point in time or under circumstances where time, as a dimension, is irrelevant. Given the fundamental differences in these two types of data, our approach for analyzing each will be different. Now, let us consider a preliminary approach for time series data analysis.

With time series data, we are particularly interested in how the data varies over time and in identifying patterns that occur systematically over time. A graph of the data, as in Fig. 3.9, is our first step in the analysis. As the British Anthropologist,

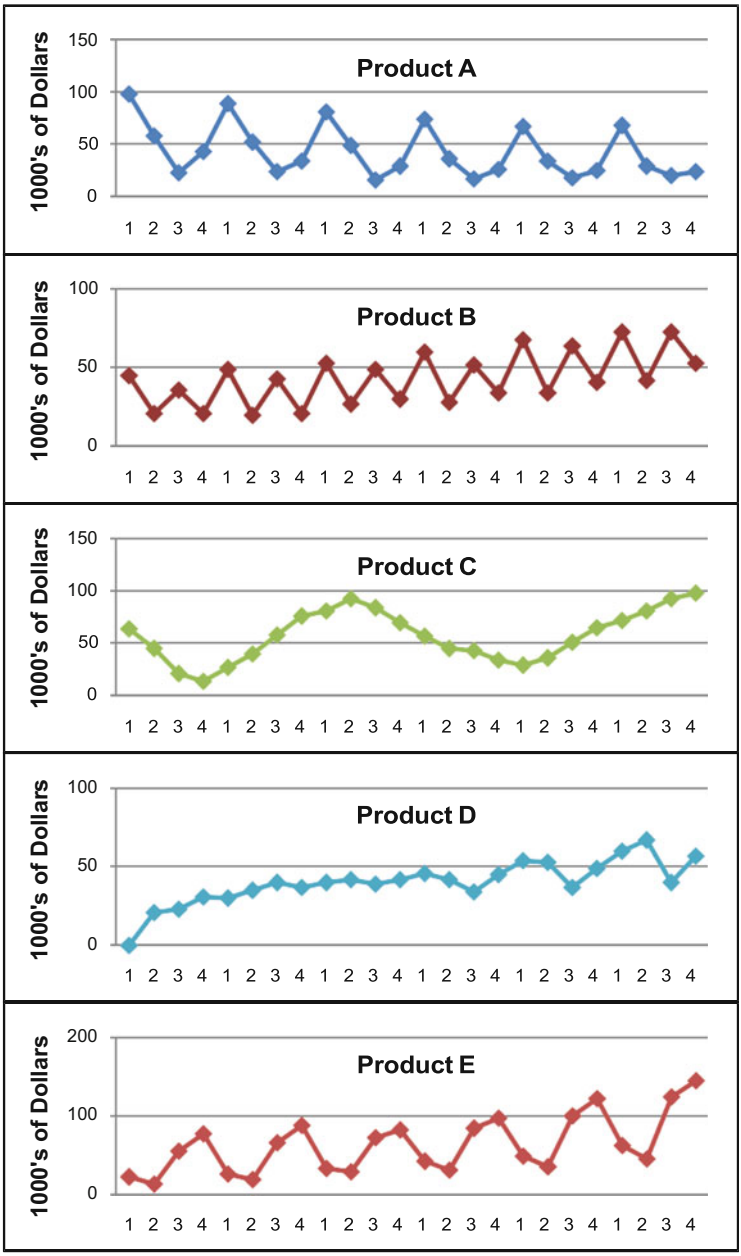


Fig. 3.9 Graph of sales data for products A–E

John Lubbock, wrote: "*What we see depends mainly on what we look for,"* and herein we see the power of Excel's charting capabilities. We can carefully scrutinize—*look for*—patterns of behavior before we commit to more technical analysis. Behavior like seasonality, co-relationship of one series to another, or one series displaying leading or lagging time behavior with respect to another are relatively easy to observe.

Now, let us investigate the graphical representation of data in Fig. 3.9. Note that if many series are displayed simultaneously, the resulting graph can be very confusing; as a result, we display each series separately. The following are some of the interesting findings for our sales data:

1. It appears that all of the product sales have some **cyclicality** except for D; that is, the data tends to repeat patterns of behavior over some relatively fixed time length (a cycle). Product D may have a very slight cyclical behavior, but is it not evident by graphical observation.
2. It appears that A and E behave relatively similarly for the first 3 years, although their cyclicality is out of phase by a single quarter. Cyclicality that is based on a yearly time frame is referred to as **seasonality**, due to the data's variation with the seasons of the year.
3. The one quarter difference between A and E (phase difference) can be explained as E **leading** A by a period. For example, E peaks in quarter 4 of the first year and A peaks in quarter 1 of the second year, thus the peak in E leads A by one quarter. The quarterly lead appears to be exactly one period for the entire 6-year horizon.
4. Product E seems to behave differently in the last 3 years of the series by displaying a general tendency to increase. We call this pattern **trend**, and in this case, a *positive* trend over time. We will, for simplicity's sake, assume that this is **linear trend**; that is, it increases or decreases at a constant rate. For example, a linear trend might increase at a rate of 4000 dollars per quarter.
5. There are numerous other features of the data that can and will be identified later.

We must be careful not to extend the findings of our visual analysis too far. Presuming we know all there is to know about the underlying behavior reflected in the data without a more formal analysis can lead to serious problems. That is precisely why we will apply more sophisticated data analysis once we have visually inspected the data.

3.4.2 Cross-Sectional Data: Visual Analysis

Now, let us consider another set of data that is collected by a web-based **e-tailer** (retailers that market products via the internet) that specializes in marketing to teenagers. The e-tailer is concerned that their website is not generating the number of **page-views** (website pages viewed per customer visit) that they desire. They suspect that the website is just not attractive to teens. To remedy the situation, they hire a web designer to redesign the site with teen's preferences and interests in mind.

An experiment is devised that randomly selects 100 teens that have not previously visited the site and exposes them to the old *and* new website designs. They are told to interact with the site until they lose interest. Data is collected on the number of web pages each teen views on the old site and on the new site.

In Table 3.2, we organize page-views for individual teens in columns. We can see that teen number 1 (top of the first column) viewed 5 pages on the old website and 14 on the new website. Teen number 15 (the bottom of the third column) viewed 10 pages on the old website and 20 on the new website. The old website and the new website represent *treatments* in the context of statistical analysis.

Our first attempt at analysis of this data is a simple visual display—a graph. In Fig. 3.10, we see a frequency distribution for our pages viewed by 100 teens, before (old) and after (new) the website update. A **frequency distribution** is simply a count of the number of occurrences of a particular quantity. For example, if in Table 3.2 we count the occurrence of two page views on the old website, we find that there are three occurrences—teen 11, 34, and 76. Thus, the frequency of two-page views is 3 and can be seen as a bar 3 units high in Fig. 3.10. Note that Fig. 3.10 counts all possible values of page views for old and new websites to develop the distribution. The range (low to high) of values for old is 1–15. It is also possible to create categories of values for the old, for example 1–5, 6–10 and 11–15 page views. This distribution would have all observations in only three possible outcomes, and appear quite different from Fig. 3.10.

We can see from Fig. 3.10 that the old website is generally located to the left (lower values of page views) of the new website. Both distributions appear to have a **central tendency**; that is, there is a central area that has more frequent values of page views than the extreme values, either lower or higher. Without precise calculation, it is likely that the average of the pages viewed will be near the center of the distributions. It is also obvious that the average, or mean, pages viewed for the old web site will be less than the average pages viewed for the new web site. Additionally, the **variation**, or spread, of the distribution for the new website is slightly larger than that of the old website: the range of the new values extends from 5 to 21, whereas the range of the old values is 1–15.

In preparation for our next form of analysis, descriptive statistics, we need to define a number terms:

1. The average, or **mean**, of a set of data is the sum of the observations divided by the number of observations.
2. A frequency distribution organizes data observations into particular categories based on the number of observations in a particular category.
3. A frequency distribution with a central tendency is characterized by the grouping of observations near or about the center of a distribution.
4. A **standard deviation** is the statistical measure of the degree of variation of observations relative to the mean of all the observations. The calculation of the standard deviation is the square root of the sum of the squared deviations for each value in the data from the mean of the distribution, which is then divided by the number of observations. If we consider the observations collected to be a

Table 3.2 Old and new website pages visited

Old website																			
5	6	2	4	11	4	8	12	10	4	6	15	8	7	5	2	3	9	5	6
4	7	11	7	6	5	9	10	6	8	10	6	4	11	8	8	15	8	4	11
7	6	12	8	1	5	6	10	14	11	4	6	11	6	8	6	11	8	6	6
5	12	5	5	7	7	2	5	10	6	7	5	12	8	9	7	5	8	6	6
7	7	10	10	6	10	6	10	8	9	14	6	13	11	12	9	7	4	11	5
New website																			
14	5	18	19	10	11	11	12	15	10	9	9	11	9	10	11	8	5	21	8
10	10	16	10	14	15	9	12	16	14	20	5	10	12	21	12	16	14	17	15
12	12	17	7	9	8	11	12	12	12	8	12	11	14	10	16	8	5	6	10
5	16	9	9	14	9	12	11	13	6	15	11	14	14	16	9	7	17	10	15
9	13	20	12	11	10	18	9	13	12	19	6	9	11	14	10	18	9	11	11

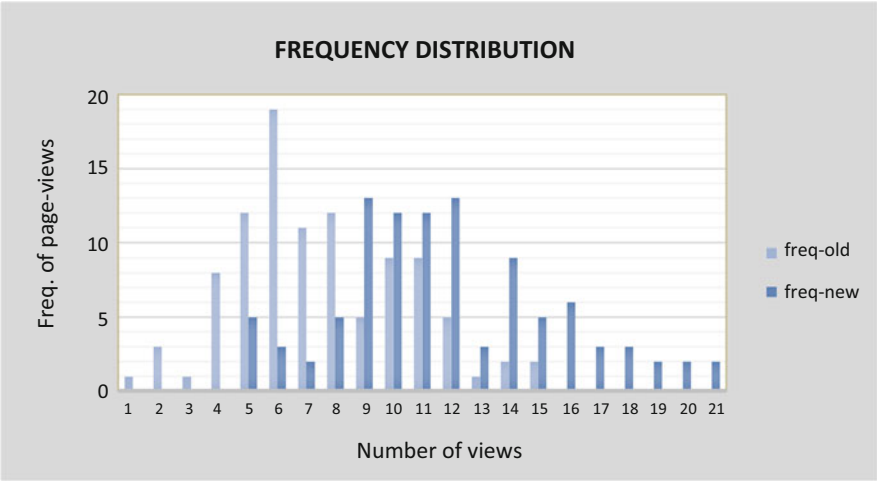


Fig. 3.10 Frequency distribution of page-views

sample, then the division is by the number of observations minus 1. The standard deviation formula in Excel for observations that are assumed to be a sample is *STDEV.S(number1, number2…)*. In the case where we assume our observations represent a **population** (all possible observations), the formula is *STDEV.P(number1, number2,…)*.

5. A **range** is a simple, but useful, measure of variation, which is calculated as the high observation value minus the low value.
6. A population is the set of all possible observations of interest.
7. The **median** is the data point in the middle of the distribution of all data points. There are as many values below as above the median.
8. The **mode** is the most often occurring value in the data observations.
9. The **standard error** is the sample standard deviation divided by the square root of the number of data observations.

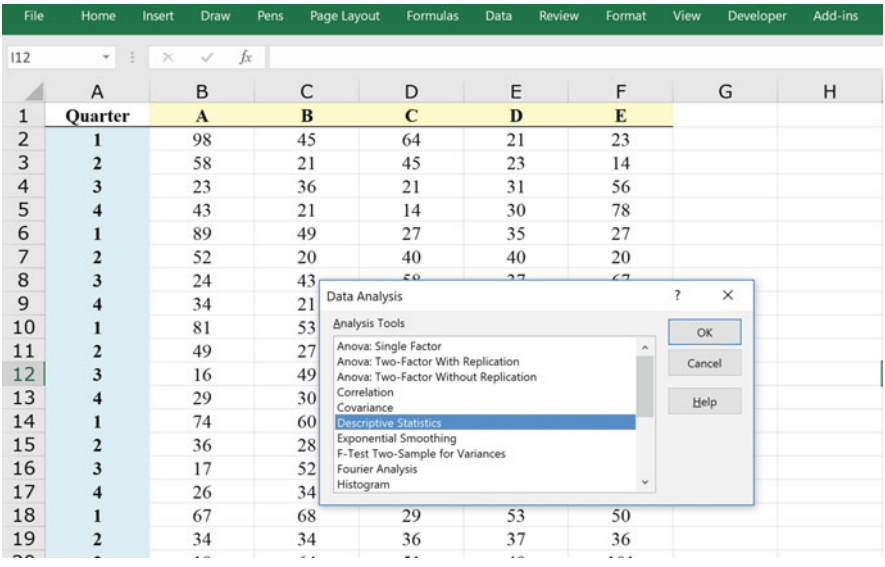
10. **Sample variance** is the square of the sample standard deviation of the data observations.
11. **Kurtosis** ("peaked-ness") and **skewness** (asymmetry) are measures related to the shape of a data set organized into a frequency distribution.

In most cases, it is likely we are *not* interested in viewing our time series data as a distribution of points, since frequency distributions generally ignore the time element of a data point. We might expect variation and be interested in examining it, but usually with a specific association to time. A frequency distribution does not provide this time association for data observations.

Let us examine the data sets by employing *descriptive statistics* for each type of data: time series and cross-sectional. We will see in the next section that some of Excel's descriptive statistics are more appropriate for some types of data than for others.

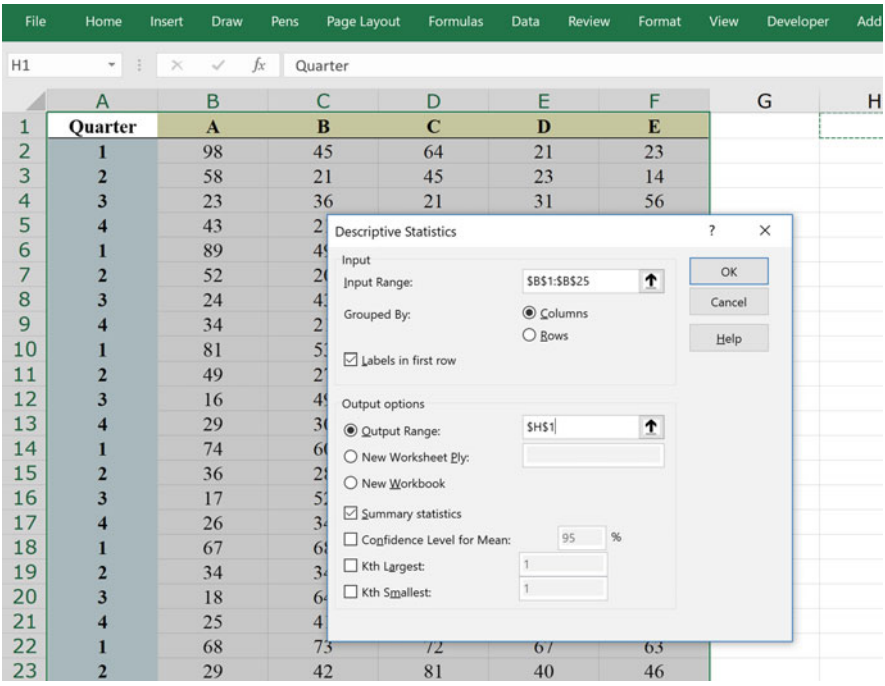
3.4.3 Analysis of Time Series Data: Descriptive Statistics

Consider the time series data for our Sales example. We will perform a very simple type of analysis that generally describes the sales data for each product—Descriptive Statistics. First, we locate our data in columnar form on a worksheet. To perform the analysis, we select the Data Analysis tool from the Analysis group in the Data ribbon. Next, we select the Descriptive Statistics tool as shown in Fig. 3.11. A dialogue box will appear that asks you to identify the input range containing the data. You must also provide some choices regarding the output location of the



	A	B	C	D	E	F	G	H
1	Quarter	A	B	C	D	E		
2	1	98	45	64	21	23		
3	2	58	21	45	23	14		
4	3	23	36	21	31	56		
5	4	43	21	14	30	78		
6	1	89	49	27	35	27		
7	2	52	20	40	40	20		
8	3	24	43					
9	4	34	21					
10	1	81	53					
11	2	49	27					
12	3	16	49					
13	4	29	30					
14	1	74	60					
15	2	36	28					
16	3	17	52					
17	4	26	34					
18	1	67	68	29	53	50		
19	2	34	34	36	37	36		

Fig. 3.11 Descriptive statistics in data analysis



	A	B	C	D	E	F
1	Quarter	A	B	C	D	E
2	1	98	45	64	21	23
3	2	58	21	45	23	14
4	3	23	36	21	31	56
5	4	43	2			
6	1	89	4			
7	2	52	2			
8	3	24	4			
9	4	34	2			
10	1	81	5			
11	2	49	2			
12	3	16	4			
13	4	29	3			
14	1	74	6			
15	2	36	2			
16	3	17	5			
17	4	26	3			
18	1	67	6			
19	2	34	3			
20	3	18	6			
21	4	25	4			
22	1	68	73	72	67	63
23	2	29	42	81	40	46

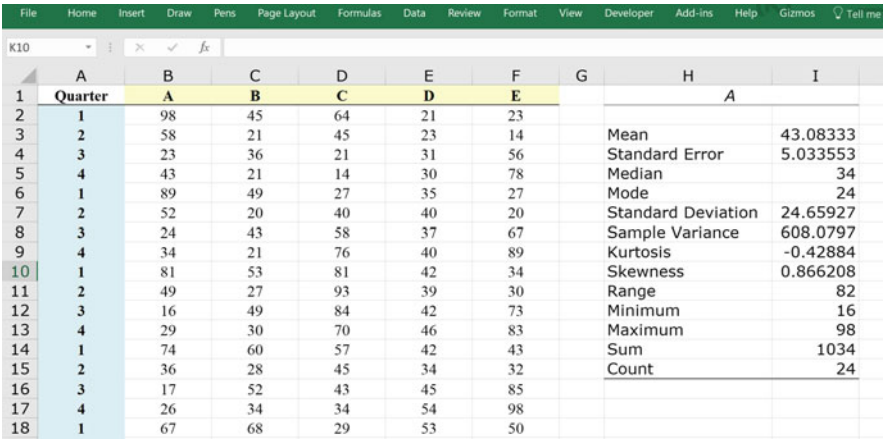
Fig. 3.12 Dialogue box for descriptive statistics

analysis and the types of output you desire (check the summary statistics box). In our example, we select data for product A (see Fig. 3.12.). We can also select all of the products (A–E) and perform the same analysis. Excel will automatically assume that each column represents data for a different product. The output of the analysis for product A is shown in Fig. 3.13.

Note that the mean of sales for product A is approximately 43 (thousand). As suggested earlier, this value, although of moderate interest, does not provide much useful information. It is the 6-year average. Of greater interest might be a comparison of each year's average. This would be useful if we were attempting to identify a yearly trend, either up (positive) or down (negative). Progressively higher means would suggest a positive trend; conversely, progressively lower means would suggest a negative trend. There will be more to come on the summary statistics for product A.

3.4.4 Analysis of Cross-Sectional Data: Descriptive Statistics

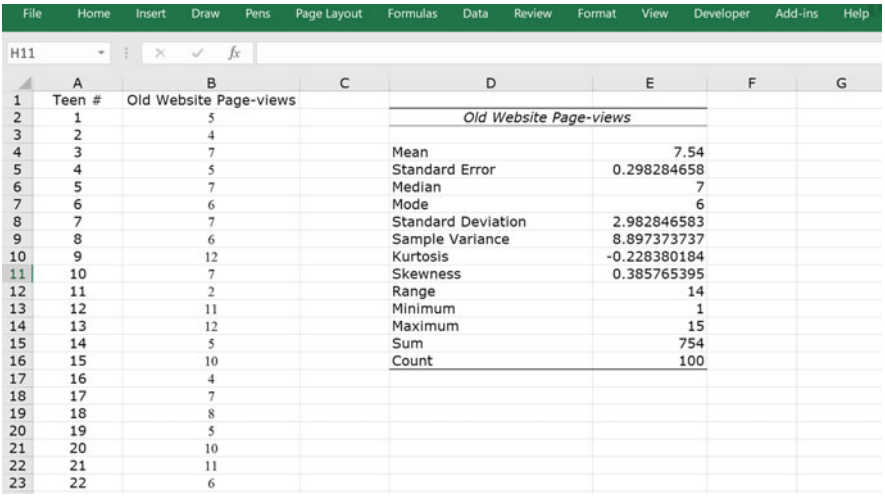
Our website data is cross-sectional; thus, the time context is not an important dimension of the data. The descriptive statistics for the old website are shown in



	A	B	C	D	E	F
1	Quarter	A	B	C	D	E
2	1	98	45	64	21	23
3	2	58	21	45	23	14
4	3	23	36	21	31	56
5	4	43	21	14	30	78
6	1	89	49	27	35	27
7	2	52	20	40	40	20
8	3	24	43	58	37	67
9	4	34	21	76	40	89
10	1	81	53	81	42	34
11	2	49	27	93	39	30
12	3	16	49	84	42	73
13	4	29	30	70	46	83
14	1	74	60	57	42	43
15	2	36	28	45	34	32
16	3	17	52	43	45	85
17	4	26	34	34	54	98
18	1	67	68	29	53	50

	H	I
	A	
	Mean	43.08333
	Standard Error	5.033553
	Median	34
	Mode	24
	Standard Deviation	24.65927
	Sample Variance	608.0797
	Kurtosis	-0.42884
	Skewness	0.866208
	Range	82
	Minimum	16
	Maximum	98
	Sum	1034
	Count	24

Fig. 3.13 Product A descriptive statistics



	A	B	C	D	E	F	G
1	Teen #	Old Website Page-views					
2	1	5			*Old Website Page-views*		
3	2	4					
4	3	7		Mean	7.54		
5	4	5		Standard Error	0.298284658		
6	5	7		Median	7		
7	6	6		Mode	6		
8	7	7		Standard Deviation	2.982846583		
9	8	6		Sample Variance	8.897373737		
10	9	12		Kurtosis	-0.228380184		
11	10	7		Skewness	0.385765395		
12	11	2		Range	14		
13	12	11		Minimum	1		
14	13	12		Maximum	15		
15	14	5		Sum	754		
16	15	10		Count	100		
17	16	4					
18	17	7					
19	18	8					
20	19	5					
21	20	10					
22	21	11					
23	22	6					

Fig. 3.14 Descriptive statistics of old website data

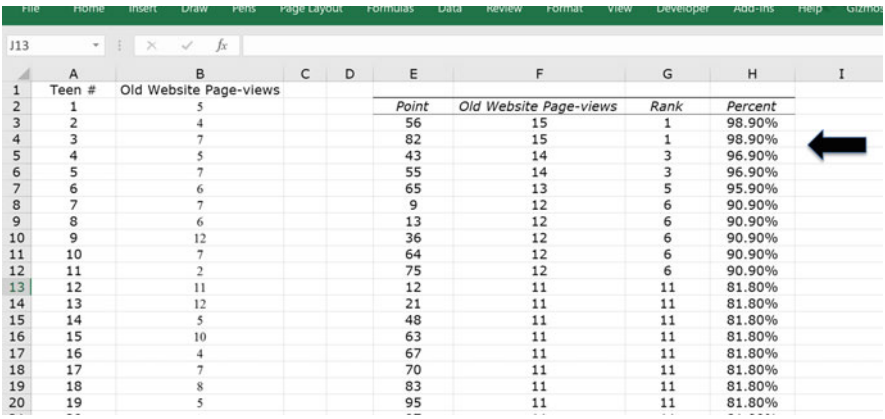
Fig. 3.14. It is a quantitative summary of the old website data graphed in Fig. 3.10. To perform the analysis, it is necessary to rearrange the data shown in Table 3.2 into a single column of 100 data points, since the *Data Analysis* tool assumes that data is organized in either rows *or* columns. Table 3.2 contains data in rows *and* columns; thus, we need to *stretch-out* the data into either a row *or* a column. This could be a tedious task if we are rearranging a large quantity of data points, but the *Cut* and *Paste* tools in the *Home* ribbon and *Clipboard* group will make quick work of the changes. It is important to keep track of the 100 teens as we rearrange the data, since

the old website will be compared to the new, and tracking the change in specific teens will be important. Thus, whatever cutting and pasting is done for the new data must be done similarly for the old data. This will ensure that comparison between old and new are for the same teen.

Now let us consider the measures shown in the *Descriptive Statistics*. As the graph in Fig. 3.10 suggested, the mean or average for the old website appears to be between 6 and 8, probably on the higher end given the positive skew of the graph—the frequency distribution tails off in the direction of higher or *positive* values. In fact, the mean is 7.54. The skewness is positive, 0.385765, indicating the right tail of the distribution is longer than the left, as we can see from Fig. 3.10. The measure of kurtosis (peaked or flatness of the distribution relative to the normal distribution), −0.22838, is slightly negative, indicating mild relative flatness. The other measures are self-explanatory, including the measures related to samples: standard error and sample variance. We can see that these measures are more relevant to cross-sectional data than to our time series data since the 100 teens are a randomly selected *sample* of the entire population of visitors to the old website for a particular period of time.

There are several other tools related to descriptive statistics—Rank, Percentile, and Histogram—that can be very useful. Rank and Percentile generates a table that contains an ordinal and percentage rank of each data point in a data set (see Fig. 3.15). Thus, one can conveniently state that of the 100 viewers of the old website, individuals number 56 and 82 rank highest (number 1 in the table shown in Fig. 3.15). Also, they hold the percentile position 98.9%, which is the percent of teens that are at or below their level of views (15). Percentiles are often used to create thresholds; for example, a score on an exam below the 60th percentile is a failing grade.

The *Histogram* tool in the *Data Analysis* group creates a table of the frequency of the values relative to your selection of *bin* values. The results could be used to create the graphs in Fig. 3.10. Figure 3.16 shows the dialogue box entries necessary to



	A	B	C	D	E	F	G	H	I
1	Teen #	Old Website Page-views							
2	1	5			*Point*	*Old Website Page-views*	*Rank*	*Percent*	
3	2	4			56	15	1	98.90%	
4	3	7			82	15	1	98.90%	
5	4	5			43	14	3	96.90%	
6	5	7			55	14	3	96.90%	
7	6	6			65	13	5	95.90%	
8	7	7			9	12	6	90.90%	
9	8	6			13	12	6	90.90%	
10	9	12			36	12	6	90.90%	
11	10	7			64	12	6	90.90%	
12	11	2			75	12	6	90.90%	
13	12	11			12	11	11	81.80%	
14	13	12			21	11	11	81.80%	
15	14	5			48	11	11	81.80%	
16	15	10			63	11	11	81.80%	
17	16	4			67	11	11	81.80%	
18	17	7			70	11	11	81.80%	
19	18	8			83	11	11	81.80%	
20	19	5			95	11	11	81.80%	

Fig. 3.15 Rank and percentile of old website data

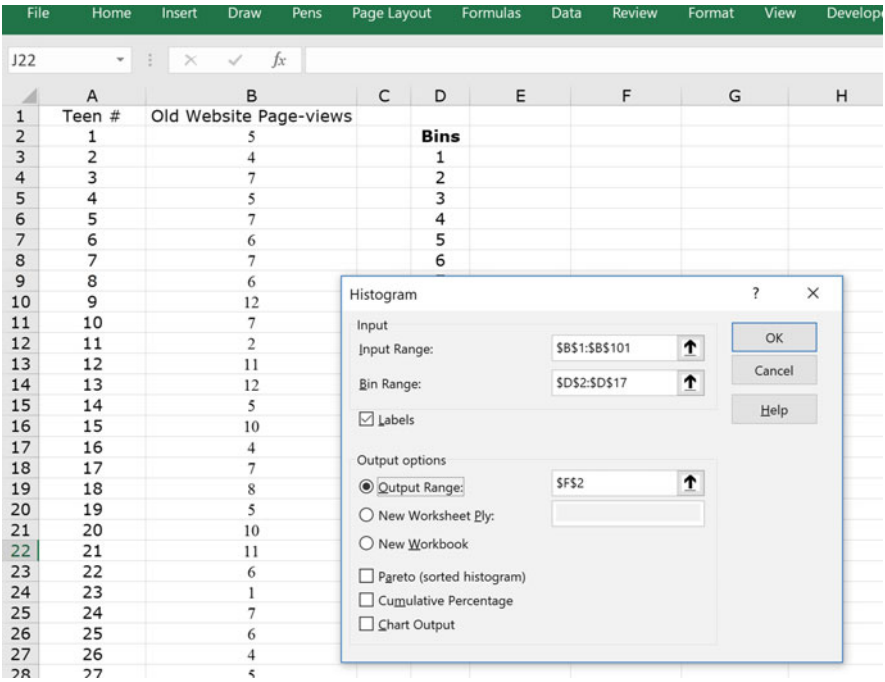


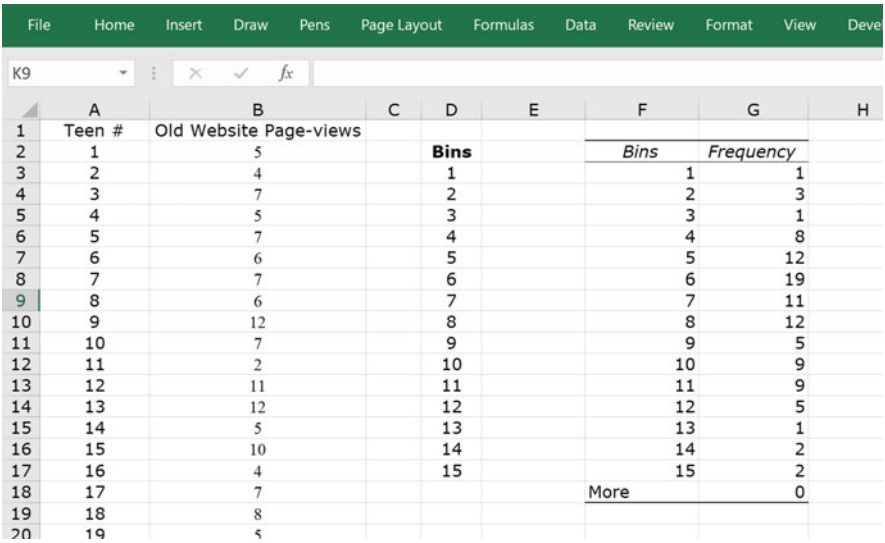
Fig. 3.16 Dialogue box for histogram analysis

create the histogram. Just as the bin values used to generate Fig. 3.10 are values from the lowest observed value to the largest in increments of one, these are the entry values in the dialogue box in Fig. 3.16—D2:D17. (Note the *Labels* box is checked to include the title *Bins*— in cell D2). The results of the analysis are shown in Fig. 3.17. It is now convenient to graph the histogram by selecting the *Insert* ribbon and the *Charts* group. This is equivalent to the previously discussed frequency distribution in Fig. 3.10.

3.5 Analysis of Time Series Data: Forecasting/Data Relationship Tools

We perform data analysis to answer questions and gain insight. So, what are the central questions we would like to ask about our time series data? Put yourself in the position of a data analyst, and consider what might be important to you. Here is a list of possible questions you may want to answer:

1. Do the data for a particular series display a repeating and systematic pattern over time?
2. Does one series move with another in a predictable fashion?



	A	B	C	D	E	F	G	H
1	Teen #	Old Website Page-views						
2	1	5		**Bins**		*Bins*	*Frequency*	
3	2	4		1		1	1	
4	3	7		2		2	3	
5	4	5		3		3	1	
6	5	7		4		4	8	
7	6	6		5		5	12	
8	7	7		6		6	19	
9	8	6		7		7	11	
10	9	12		8		8	12	
11	10	7		9		9	5	
12	11	2		10		10	9	
13	12	11		11		11	9	
14	13	12		12		12	5	
15	14	5		13		13	1	
16	15	10		14		14	2	
17	16	4		15		15	2	
18	17	7				More	0	
19	18	8						
20	19	5						

Fig. 3.17 Results of histogram analysis for old website views

3. Can we identify behavior in a series that can predict systematic behavior in another series?
4. Can the behavior of one series be incorporated into a forecasting model that will permit accurate prediction of the future behavior of another series?

Although there are many questions that can be asked, these four are important and will allow us to investigate numerous analytical tools in Data Analysis. As a note of caution, let us keep in mind that this example is based on a very small amount of data; thus, we must be careful to not overextend our perceived insight. The greater the amount of data, the more secure one can be in his or her observations. Let us begin by addressing the first question.

3.5.1 Graphical Analysis

Our graphical analysis of the sales data has already revealed the possibility of **systematic behavior** in the series; that is, there is an underlying system that influences the behavior of the data. As we noted earlier, all of the series, except for product D, display some form of cyclical behavior. How might we determine if systematic behavior exists? Let us select product E for further analysis, although we could have chosen any of the products.

In Fig. 3.18, we see that the product time series does in fact display repetitive behavior; in fact, it is *quite* evident. Since we are interested in the behavior of both the yearly demand and quarterly demand, we need to rearrange our time series data

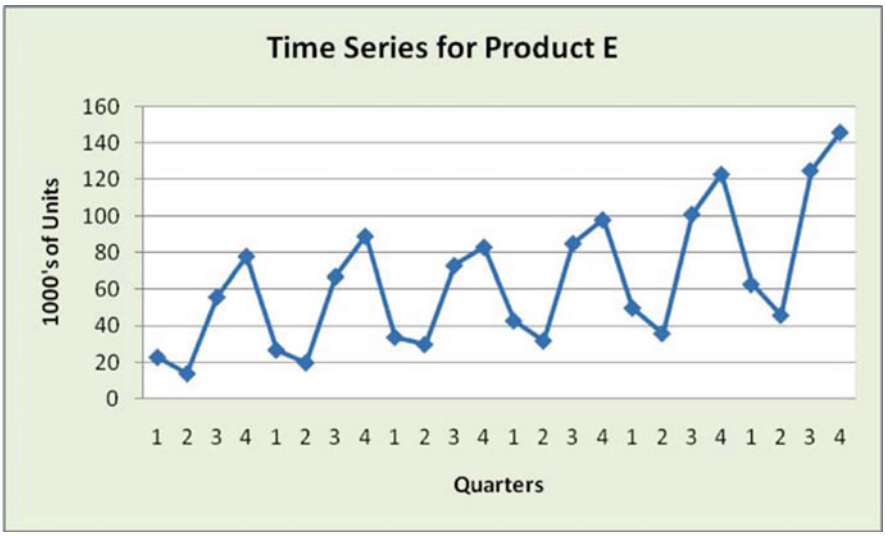


Fig. 3.18 Product E time series data

Table 3.3 Modified quarterly data[a] for product E

	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Yearly total
Yr1	23	14	56	78	171
Yr2	27	20	67	89	203
Yr3	34	30	73	83	220
Yr4	43	32	85	98	258
Yr5	50	36	101	123	310
Yr6	63	46	125	146	380

[a]Sales in thousands of dollars

to permit a different type of graphical analysis. Table 3.3 shows the data from Table 3.1 in a modified format: each row represents a year (1–6) and each column a quarter (1–4); thus, the value 101 represents quarter 3 in year 5. Additionally, the rightmost vertical column of the table represents yearly totals. This new data configuration will allow us to perform some interesting graphical analysis.

Now, let us proceed with the analysis. First, we will apply the *Histogram* tool to explore the quarterly data behavior in greater depth. There is no guarantee that the tool will provide insight that is useful, but that's the challenge of data analysis—it can be as much an art as a science. In fact, we will find the *Histogram* tool will be of little use. Why? It is because the tool does not distinguish between the various quarters. As far as the *Histogram* tool is concerned, a data point is a data point, without regard to its related quarter; thus, we see the importance of the *context* of data points. Had the data points for each quarter been clustered in distinct value groups (e.g. all quarter 3 values clustered together), the tool would have been much more useful. See Fig. 3.19 for the results of the histogram with bin values in

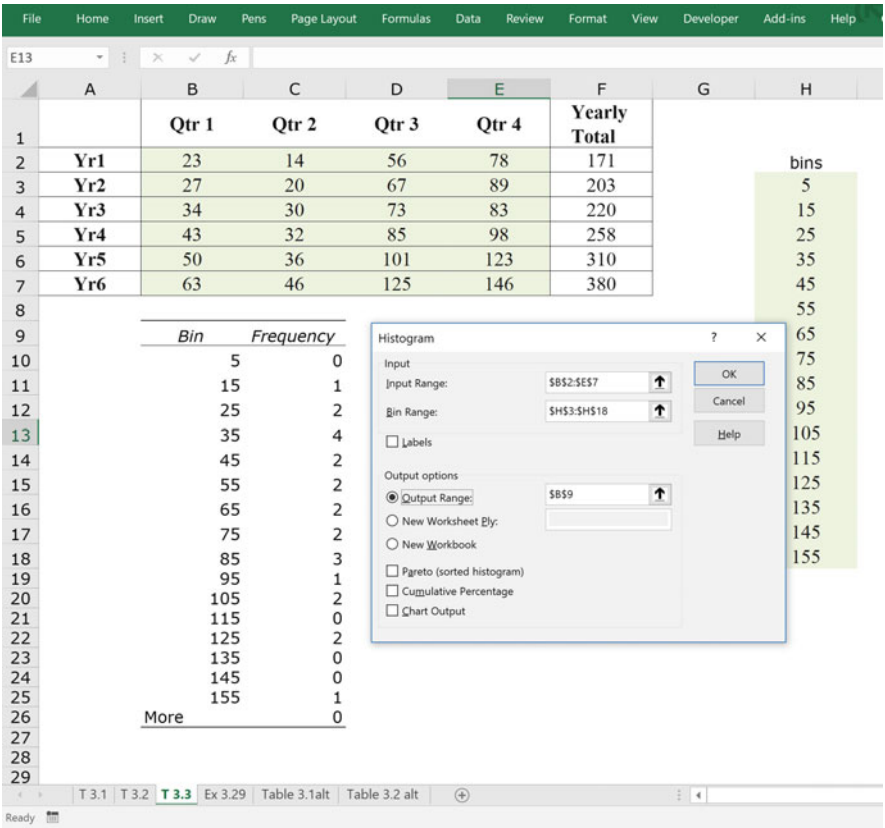


Fig. 3.19 Histogram results for all product E adjusted data

increments of 10 units, starting at a low value of 5 and a high of 155. There are clearly no clusters of data representing distinct quarters that are easily identifiable. For example, there is only 1 value (14) that falls into the 15 bin (values $>$ 5 and $<=$ 15). That value is the second quarter of year 1. Similarly, there are 3 data values that fall into the 85 bin (values $>$75 to $<=$ 85): quarters 4 of year 1, quarter 4 of year 3, and quarter 3 of year 4. It may be possible to adjust the bins to capture clusters of values more effectively, but that is not the case for these data. But don't despair, we still have other graphical tools that will prove useful.

Figure 3.20 is a graph that explicitly considers the quarterly position of data by dividing the time series into four quarterly sub-series for product E. See Fig. 3.21 for the data selected to create the graph. It is the same as Table 3.3. From Fig. 3.20, it is evident that all the product E time series over 6 years display important data behavior: the fourth quarter in all years is the largest sales value, followed by quarters 3, 1, and 2, respectively. Note that the Yearly Total is increasing consistently over time (measured on the right vertical scale-Yrly Total), as are all other series, except for quarter 4, which has a minor reduction in year 3. This suggests that

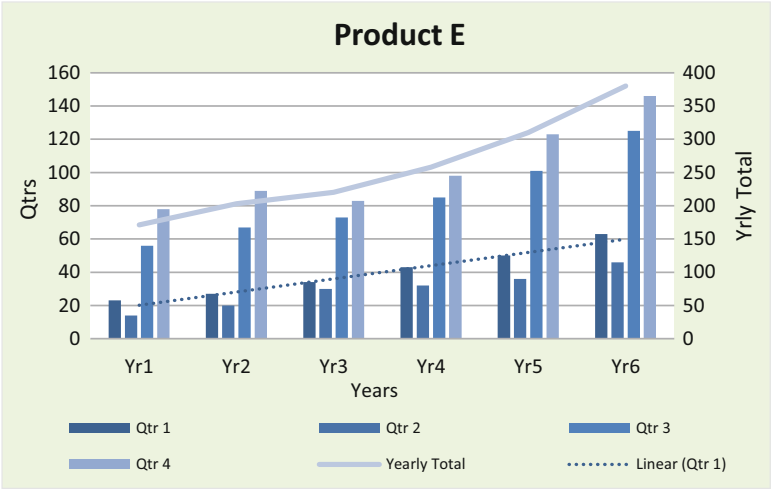


Fig. 3.20 Product E quarterly and yearly total data

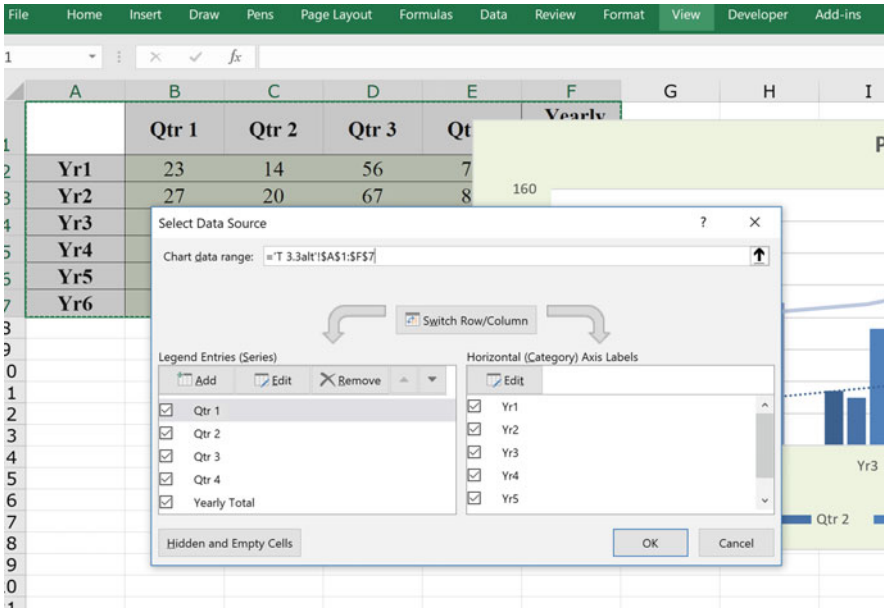


Fig. 3.21 Selected data for quarters and yearly total

there is a seasonal effect related to our data, as well as an almost consistent trend for all series. Yet, it may be wise to reserve judgment on quarterly sales behavior into the future since our data set represents a relatively small amount of data.

Before we proceed, let us take stock of what the graphical data analysis has revealed about product E:

1. We have assumed that it is convenient to think in terms of these data having three components—a base level, seasonality effects, and a linear trend.
2. The base relates to the value of a specific quarter, and when combined with a quarterly trend for the series, results in a new base in the following year. Trends for the various quarters may be different, but all our series (possible exception of quarter 4) have a clear positive linear trend, including the total.
3. We have dealt with seasonality by focusing on specific quarters in the *yearly cycle* of sales. By noting that there is a consistent pattern or relationship within a yearly cycle (quarter 4 is always the highest value), we observe seasonal behavior.
4. If sales follow the past behavior, visual analysis suggests that we can build a model that might provide future estimates of quarterly and yearly total values. We can do this because we understand the three elements that make up the behavior of each quarterly series and yearly total—base, trend, and seasonality.

One last comment on the graph in Fig. 3.20 is necessary. Note that the graph has two vertical axis scales. This is necessary due to the large difference in the magnitude of values for the individual quarters and the Yrly Totals series. To use a single vertical axis would make viewing the movement of the series, relative to each other, difficult. By selecting any data observation associated with the Yrly Total on the graph with a right-click, a menu appears that permits you to format the data series. One of the options available is to plot the series on a secondary axis. This feature can be quite useful when viewing data series that vary in magnitude.

3.5.2 Linear Regression

Now, let me introduce a tool that is useful in the prediction of future values of a series. The tool is the forecasting technique linear regression, and although it is not appropriate for all forecasting situations, it is very commonly used. There are many other forecasting techniques that can be used to model business and economic data, but I introduce linear regression because of its very common usage and its instructive nature—understanding the concept of a linear model is quite useful in its application to other types of modeling. Just as in our graphical analysis, the choice of a model should be a methodical process, and attention to the appropriateness of the application must be considered. Use the right tool for the right technique.

Linear Regression builds a model that predicts the future behavior of a *dependent* variable based on the assumed *linear* influence of one or more independent variables. For example, say your daily mood (DM) depends on two variables—temperature (T) in degrees centigrade and amount of (S) sunshine in lumens. Then the dependent variable is DM, and the independent variables are T and S. The dependent variable is what we attempt to predict or forecast. In the case of sales value for quarters, the independent variable, number of years into the future, is what

we base our forecast on. The dependent variable is the corresponding future sales in dollars for the product.

Thus, the concept of a regression formula is relatively simple: for particular values of an independent variable, we can construct a linear relationship that permits the prediction of a dependent variable. For our product E sales data, we will create a regression model for each quarter. So, we will construct four regressions—quarter 1, quarter 2, etc. It is possible to construct a single regression for prediction of all quarters, but that will require additional independent variables to deal with seasonality. More on this topic later in the chapter.

Simple linear regression, which is the approach we will use, can be visualized on an X–Y coordinate system—a single X represents the independent variable and Y the dependent variable. **Multiple linear regression** uses more than one X to predict Y. Simple regression finds the linear, algebraic relationship that best fits the data by choosing a slope of the regression line, known as the **beta** (β), and a Y intercept (where the line crosses the Y axis), known as the **alpha** (α). If we examine the series in Fig. 3.20, it appears that all quarters, except maybe quarter 4, are a good linear fit with years as the independent variable. To more closely understand the issue of a linear fit, I have added a linear trendline for the quarter 1 series—marked Linear (Qtr1) in the legend. Notice, the line tracks the changes in the quarter 1 series nicely. The regression line is added by selecting the series and right-clicking—an option to *Add Trendline* appears and we select linear.

Before we move on with the analysis, let me caution that creating a regression model from only six data points is quite dangerous. Yet, data limitations are often a fact of life and must be dealt with, even if it means basing predictions on very little data, and assuredly, six data points are an extremely small number of data observations. In this case, it is also a matter of using what I would refer to as a *baby problem* to demonstrate the concept. So, how do we perform the regression analysis?

As with the other tools in Data Analysis, a dialogue box, shown in Fig. 3.22, will appear and query you as to the data ranges that you wish to use for the analysis: the dependent variable will be the *Input Y Range* and the independent variable will be the *Input X Range*. The data range for Y is the set of six values (C2:C8, including label) of observed quarterly sales data. The X values are the numbers 1–6 (B2:B8) representing the years for the quarterly data. Thus, regression will determine an alpha and beta that, when incorporated into a predictive formula ($Y = \beta X + \alpha$), will result in the best model available for some criteria. Frequently, the criteria that is used by regression to select alpha and beta is a method called Ordinary Least Squares (OLS). There is no guarantee that a regression will be a good fit—it could be good, bad, or anything in between. Once alpha and beta have been determined, they can then be used to create a predictive model. The resulting regression statistics and regression details are shown in Fig. 3.23.

The R-square (coefficient of determination) shown in the *Regression Statistics* of Fig. 3.23 is a measure of how well the estimated values of the regression correspond to the actual quarterly sales data; it is a guide to the *goodness of fit* of the regression model. R-square values can vary from 0 to 1, with 1 indicating perfect correspondence between the estimated value and the data, and with 0 indicating no systematic

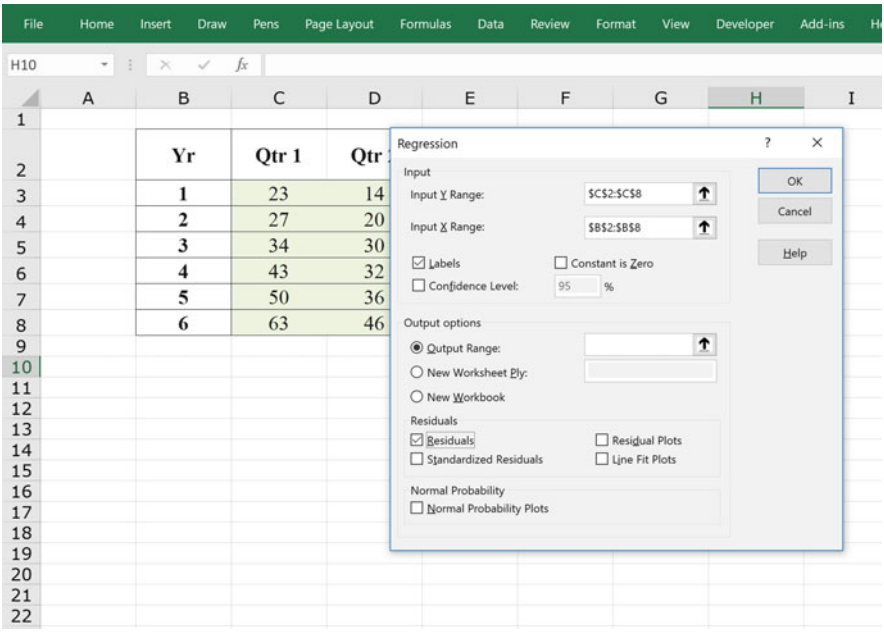
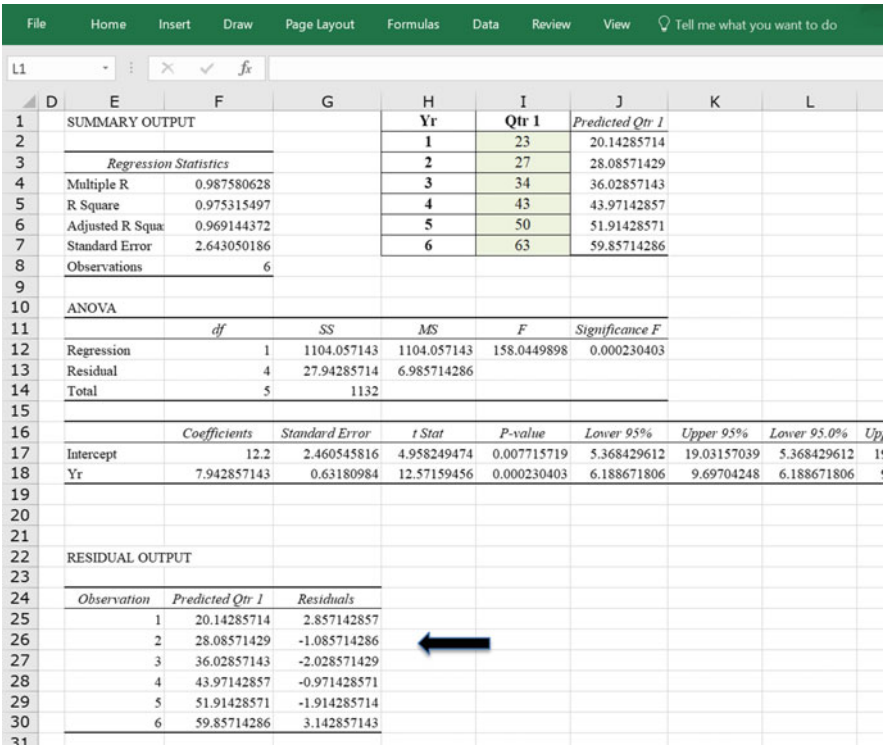


Fig. 3.22 Dialogue box for regression analysis of product E, quarter 1

correspondence whatsoever. Alternatively, we can say that R-square is the percent of variation in our dependent variable that is due to our independent variable, as provided by our regression equation. In this model, the R Square is approximately 97.53%. This is a very high R-square, implying a very good fit.

The question of what is a good R-square is a matter for professional debate. Clearly, at the extremes, we can say that 97.53% is good and 0.53% is not good; but, it is generally a relative question. Someone in marketing might be happy with an R-square of 40%, yet someone in engineering would have 80% as a threshold for acceptance. Ultimately, the answer is a matter of whether the regression is of financial benefit, or not. Finally, R-square is just one measure of fit, as we will see.

The analysis can also provide some very revealing graphs: the fit of the regression to the actual data and the residuals (the difference between the actual and the predicted values). To produce a residuals plot, check the residuals box in the dialogue box, shown in Fig. 3.22. This allows you to see the accuracy of the regression model. In Fig. 3.23, you can see the Residuals Output at the bottom of the output. The residual for the first observation (23) is 2.857… since the predicted value produced by the regression is 20.143… (23–20.143… = 2.857…). Finally, the coefficients of the regression are also specified in Fig. 3.23. The Y intercept, or $\alpha = 12.2$, is where the regression line crosses the Y-axis. The coefficient of the independent variable $\beta = 7.94\ldots$, is the slope of the linear regression for the independent variable. These coefficients *specify* the model and can be used for prediction. For example, the analyst



	E	F	G	H	I	J	K	L
1	SUMMARY OUTPUT			Yr	Qtr 1	Predicted Qtr 1		
2				1	23	20.14285714		
3	Regression Statistics			2	27	28.08571429		
4	Multiple R	0.987580628		3	34	36.02857143		
5	R Square	0.975315497		4	43	43.97142857		
6	Adjusted R Squa	0.969144372		5	50	51.91428571		
7	Standard Error	2.643050186		6	63	59.85714286		
8	Observations	6						
9								
10	ANOVA							
11		df	SS	MS	F	Significance F		
12	Regression	1	1104.057143	1104.057143	158.0449898	0.000230403		
13	Residual	4	27.94285714	6.985714286				
14	Total	5	1132					
15								
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Up
17	Intercept	12.2	2.460545816	4.958249474	0.007715719	5.368429612	19.03157039	5.368429612	1
18	Yr	7.942857143	0.63180984	12.57159456	0.000230403	6.188671806	9.69704248	6.188671806	
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	Observation	Predicted Qtr 1	Residuals						
25	1	20.14285714	2.857142857						
26	2	28.08571429	-1.085714286						
27	3	36.02857143	-2.028571429						
28	4	43.97142857	-0.971428571						
29	5	51.91428571	-1.914285714						
30	6	59.85714286	3.142857143						

Fig. 3.23 Summary output for product E quarter 1

may want to predict an estimate of the first quarterly value for the seventh year. Thus, the prediction calculation results in the following:

$$\text{Estimated Y for Year } 7 = \alpha + \beta\,(\text{Year}) = 12.1 + 7.94(7) = 67.8$$

Figure 3.24 shows the relationship resulting from the regression: a graph of the actual and predicted values for quarter 1. The fit is almost perfect, hence the R-squared of 97.53%. Note that regression can be applied to any data set, but it is only when we examine the results that we can determine if regression is a good predictive tool. When the R-square is low, and residuals are not a good fit, it is time to look elsewhere for a predictive model. Also, remember that regression, as well as other forecasting techniques, focuses on past behavior to predict the future. If a systematic change occurs in the future that makes the past behavior irrelevant, then the regression equation is no longer of value. For example, a regression predicting some economic variable might be useful until a change occurs in the overall economy, like a recession. Our OLS approach also weighs the value of observations equally, so old data is as important as new. There may be good reason to consider new data more heavily than old. There are techniques in regression to consider this situation. These are advanced topics that are not discussed here, but you can easily find information regarding these techniques in a text dedicated to regression.

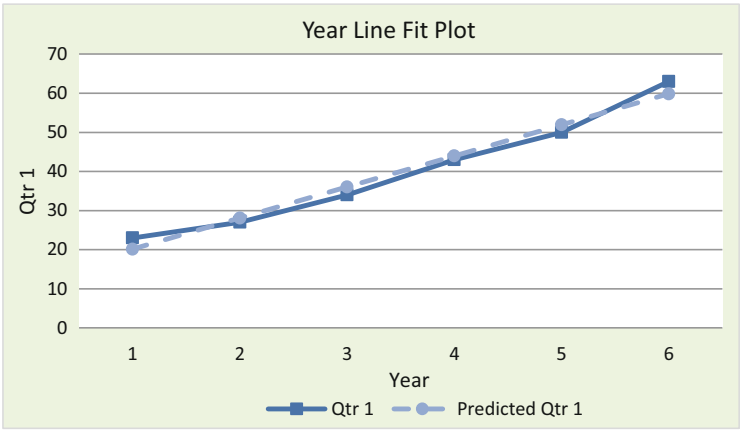


Fig. 3.24 Plot of fit for product E quarter 1

Now, let us determine the fit of a regression line for quarter 4. As mentioned earlier, a visual observation of Fig. 3.20 indicates that quarter 4 appears to be the least suitable among quarters for a linear regression model, and Fig. 3.25 indicates a less impressive R-square of approximately 85.37%. Yet, this is still a relatively high value—approximately 85% of the variation in the dependent variable is explained by the independent variables. Figure 3.26 shows the predicted and actual plot for quarter 4.

There are other important measures of fit that should be considered for regression. Although we have not discussed this measure yet, the Significance F for quarter 1 regression is quite small (0.0002304), indicating that we should conclude that there is significant association between the independent and dependent variables. (The term significant has very special meaning in the realm of statistics.) For the quarter 4 regression model in Fig. 3.25, the value is larger (0.00845293), yet there is likely to be a significant association between X and Y.

The smaller the Significance F, the better the fit. So, when is the Significance F value significant? This value is compared to a threshold value, often 0.05. When the Significance F is at or below this value, it is significant; when it is above this value, there is no significance. When we discuss hypothesis testing in future chapters, we will learn more about this topic. For now, we can assume that 0.05 is an appropriate measure for comparison.

There are many other important measures of regression fit that we have not discussed for time series errors or residuals—e.g. independence or serial correlation, homoscedasticity, and normality. These are equally important measures to those we have discussed, and they deserve attention in a serious regression modeling effort, but they are beyond the scope of this chapter. It is often the case that a regression will meet fit standards for some, but not all measures. Forecasting professionals will make a judgement as to overall fit, and whether it is sufficient in an imperfect world.

Thus far, we have used data analysis to explore and examine our data. Each form of analysis has contributed to our overall insight. Simply because a model, such as

File Home Insert Draw Page Layout Formulas Data Review View Tell me what you want to do

O16

	A	B	C	D	E	F	G	H	I
2									
3	*Regression Statistics*								
4	Multiple R	0.923961649							
5	R Square	0.853705129							
6	Adjusted R Square	0.817131412							
7	Standard Error	11.30570863							
8	Observations	6							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	2983.557143	2983.557143	23.3420386	0.008452926			
13	Residual	4	511.2761905	127.8190476					
14	Total	5	3494.833333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	57.13333333	10.52504194	5.428323577	0.005586078	27.91113214	86.35553452	27.91113214	86.35553452
18	Yr	13.05714286	2.702581281	4.831359912	0.008452926	5.553574289	20.56071143	5.553574289	20.56071143
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	*Observation*	*Predicted Qtr 4*	*Residuals*						
25	1	70.19047619	7.80952381						
26	2	83.24761905	5.752380952						
27	3	96.3047619	-13.3047619						
28	4	109.3619048	-11.36190476						
29	5	122.4190476	0.580952381						
30	6	135.4761905	10.52380952						
31									

Fig. 3.25 Summary output for product E quarter 4

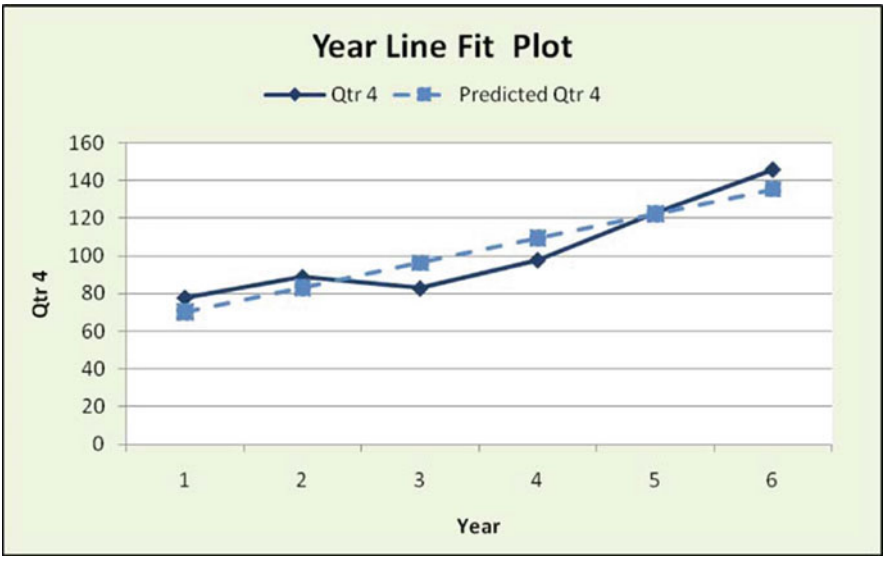


Fig. 3.26 Plot of fit for product E quarter 4

regression, does not fit our data, it does not mean that our efforts have been wasted. It is still likely that we have gained important insight, i.e., this is not an appropriate model, and there may be indicators of an alternative to explore. It may sound odd, but often we may be as well-informed by what doesn't work, as by what does.

3.5.3 *Covariance and Correlation*

Recall the original questions posed about the product sales data, and in particular, the second question: "Does one series move with another in a predictable fashion?" The Covariance tool helps answer this question by determining how the series *co-vary*. We return to the original data in Table 3.1 to determine the movement of one series with another. The Covariance tool, which is found in the Data Analysis tool, returns a matrix of values for a set of data series that you select. For the product sales data, it performs an exhaustive pairwise comparison of all six times series. As is the case with other Data Analysis tools, the dialogue box asks for the data ranges of interest, and we provide the data in Table 3.1. Each value in the matrix represents either the variance of one time series, or the covariance of one time series compared to another. For example, in Fig. 3.27 we see the covariance of product A to itself (its variance) is

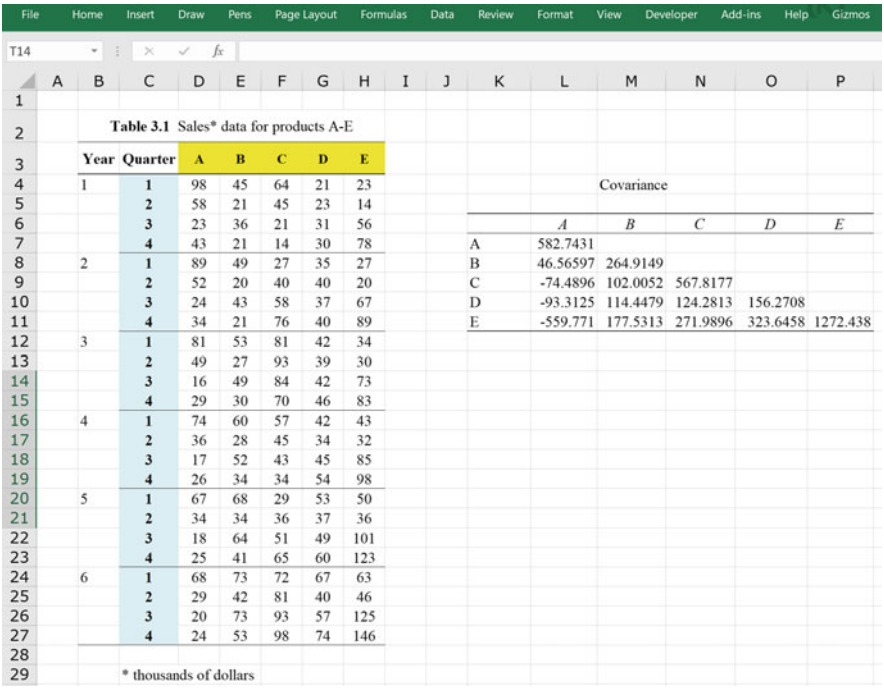


Fig. 3.27 Covariance matrix for product A–E

582,743.1 and the covariance of product A and C is −74.4896. Large positive values of covariance indicate that large values of data observations in one series correspond to large values in the other series. Large negative values indicate the inverse: small values in one series indicate large values in the other.

Figure 3.27 is relatively easy to read. The covariance of product D and E is relatively strong at 323.649, while the same is true for product A and E at −559.77. These values suggest that we can expect D and E moving together, or in the same direction, while A and E also move together, but in opposite directions due to the negative sign of the covariance. Again, we need only refer to Fig. 3.9 to see that the numerical covariance values bear out the graphical evidence. Small values of covariance, like those for product A and B (and C also), indicate little co-variation. The problem with this analysis is that it is not a simple matter to know what we mean by large or small values—large or small relative to what?

Fortunately, statisticians have a solution for this problem—Correlation analysis. Correlation analysis will make understanding the linear co-variation, or co-relation, between two variables much easier, because it is measured in values that are standardized between the range of −1 to 1, and these values are known as correlation coefficients. A correlation coefficient of 1 for two data series indicates that the two series are perfectly, positively correlated: as one variable increases so does the other. If correlation coefficient of −1 is found, then the series are perfectly, negatively correlated: as one variable increases the other decreases. Two series are said to be independent if their correlation is 0. The calculation of correlation coefficients involves the covariance of two data series, but as we mentioned, the correlation coefficient is more easily interpreted.

In Fig. 3.28 we see a correlation matrix, which is very similar to the covariance matrix. You can see that the strongest positive correlation in the matrix is between products D and E (at 0.725793), and the strongest negative correlation is between A and E, where the coefficient of correlation is −0.65006. There are also some values that indicate near linear independence (for example, products A and B with a coefficient of 0.118516). Clearly, this is a more direct method of determining the linear correlation of one data series with another than the covariance matrix.

3.5.4 Other Forecasting Models

In a more in-depth investigation of the data, we would include a search for *other* appropriate models to describe the data behavior. These models could then be used to predict future quarterly periods. Forecasting models and techniques abound and require very careful selection, but a good candidate model for this data is one that is known as **Winters' 3-factor Exponential Smoothing**, often called Winters' model. It is a member of a class of techniques known as smoothing—older data is given exponentially smaller weight in the determination of forecasts. Thus, older data becomes less important, which has serious, practical appeal. Winters' model assumes three components in the structure of a forecast model—a base (level), a

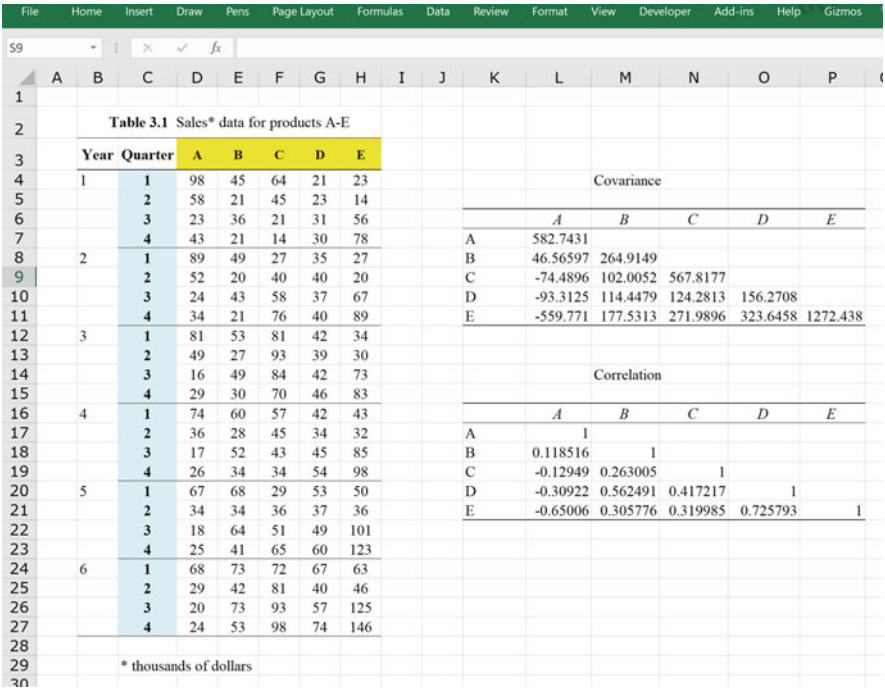


Table 3.1 Sales* data for products A-E

Year	Quarter	A	B	C	D	E
1	1	98	45	64	21	23
	2	58	21	45	23	14
	3	23	36	21	31	56
	4	43	21	14	30	78
2	1	89	49	27	35	27
	2	52	20	40	40	20
	3	24	43	58	37	67
	4	34	21	76	40	89
3	1	81	53	81	42	34
	2	49	27	93	39	30
	3	16	49	84	42	73
	4	29	30	70	46	83
4	1	74	60	57	42	43
	2	36	28	45	34	32
	3	17	52	43	45	85
	4	26	34	34	54	98
5	1	67	68	29	53	50
	2	34	34	36	37	36
	3	18	64	51	49	101
	4	25	41	65	60	123
6	1	68	73	72	67	63
	2	29	42	81	40	46
	3	20	73	93	57	125
	4	24	53	98	74	146

* thousands of dollars

Covariance

	A	B	C	D	E
A	582.7431				
B	46.56597	264.9149			
C	-74.4896	102.0052	567.8177		
D	-93.3125	114.4479	124.2813	156.2708	
E	-559.771	177.5313	271.9896	323.6458	1272.438

Correlation

	A	B	C	D	E
A	1				
B	0.118516	1			
C	-0.12949	0.263005	1		
D	-0.30922	0.562491	0.417217	1	
E	-0.65006	0.305776	0.319985	0.725793	1

Fig. 3.28 Correlation matrix for product A–E

linear trend, and some form of cyclicality. All these elements appear to be present in most of the data series for product sales, and they are also part of our previous analytical assumptions. The Winters' model also incorporates the differences between the actual and predicted values (errors) into its future calculations: so, if forecasts are too high, the model will self-correct, and the future forecasts will be lower. This self-corrective property permits the model to adjust to changes that may be occurring in underlying behavior. A much simpler version of Winters' model is found in the Data Analysis tool, **Simple Exponential Smoothing**, which only assumes a base or level component of sales.

3.5.5 Findings

So, what have we learned about our product sales data? A great deal has been revealed about the underlying behavior of the data. Some of the major findings are summarized in the list below:

1. The products display varying levels of trend, seasonality, and cyclicality. This can be seen in Fig. 3.9. Not all products were examined in depth, but the period of the cyclicality varied from seasonal for product A and E, to multi-year for product

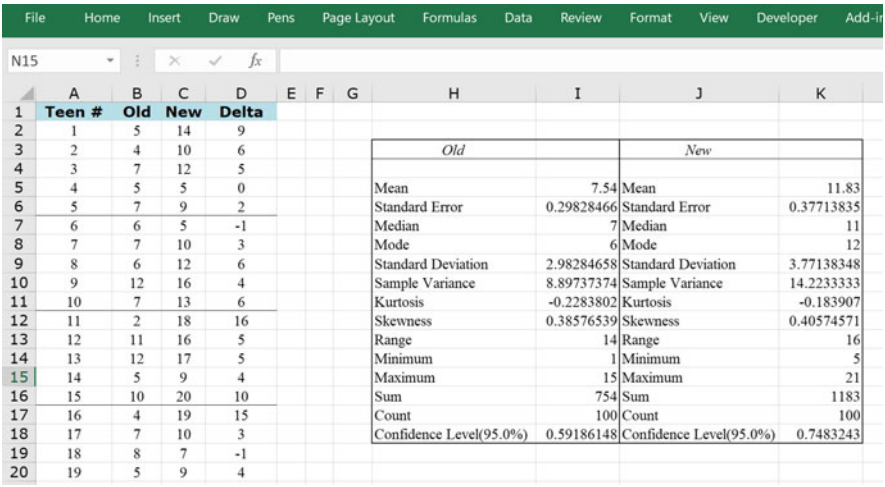
C. Product D appeared to have no cyclicality, while product B appears to have a cycle length of two quarters. These are reasonable observations, although we should be careful given the small number of data points.
2. Our descriptive statistics are not of much value for time series data, but the mean and the range could be of interest. Why? Because descriptive statistics generally ignore the time dimension of the data, and this is problematic for our time series data.
3. There are both positive (products D and E) and negative linear (products A and E) co-relations among several the time series. For some (products A and B), there is little to no linear co-relation. This variation may be valuable information for predicting behavior of one series from the behavior of another.
4. Repeating systematic behavior is evident in varying degrees in our series. For example, product D exhibits a small positive trend in early years. In later years the trend appears to increase. Products B, D, and E appear to be growing in sales. Product C might also be included, but it is not as apparent as in B, D, and E. The opposite statement can be made for product A, although its periodic lows seem to be very consistent. All these observations are derived from Fig. 3.9.
5. Finally, we examined an example of quarterly behavior for the series over 6 years, as seen in Fig. 3.20. In the case of product E, we fitted a linear regression to the quarterly data and determined a predictive model that could be used to forecast future Product E sales. The results were a relatively good model fit. Once again, this is all based on a very, very small amount of data, so we must be careful to not overextend our conclusions. In a real-world forecasting effort, we would want much more data upon which to build conclusions.

3.6 Analysis of Cross-Sectional Data: Forecasting/Data Relationship Tools

Now, let us return to our cross-sectional data, and let us apply some of the *Data Analysis* tools to the page-views data. Which tools shall we apply? We have learned a considerable amount about what works and why, so let us use our new-found knowledge and apply techniques that make sense.

First, recall that this is cross-sectional data; thus, the time dimension of the data is not a factor to consider in our analysis. Thus, consider the questions that we might ask about our data:

1. Is the average number of page-views higher or lower for the new website?
2. How does the frequency distribution of *new* versus *old* page-views compare?
3. Can the results for our sample of 100 teen subjects be generalized to the population of all possible teen visitors to our website?
4. How *secure* are we in our generalization of the sample results to the population of all possible teen visitors to our website?



	A	B	C	D
1	Teen #	Old	New	Delta
2	1	5	14	9
3	2	4	10	6
4	3	7	12	5
5	4	5	5	0
6	5	7	9	2
7	6	6	5	-1
8	7	7	10	3
9	8	6	12	6
10	9	12	16	4
11	10	7	13	6
12	11	2	18	16
13	12	11	16	5
14	13	12	17	5
15	14	5	9	4
16	15	10	20	10
17	16	4	19	15
18	17	7	10	3
19	18	8	7	-1
20	19	5	9	4

Old		*New*	
Mean	7.54	Mean	11.83
Standard Error	0.29828466	Standard Error	0.37713835
Median	7	Median	11
Mode	6	Mode	12
Standard Deviation	2.98284658	Standard Deviation	3.77138348
Sample Variance	8.89737374	Sample Variance	14.2233333
Kurtosis	-0.2283802	Kurtosis	-0.183907
Skewness	0.38576539	Skewness	0.40574571
Range	14	Range	16
Minimum	1	Minimum	5
Maximum	15	Maximum	21
Sum	754	Sum	1183
Count	100	Count	100
Confidence Level(95.0%)	0.59186148	Confidence Level(95.0%)	0.7483243

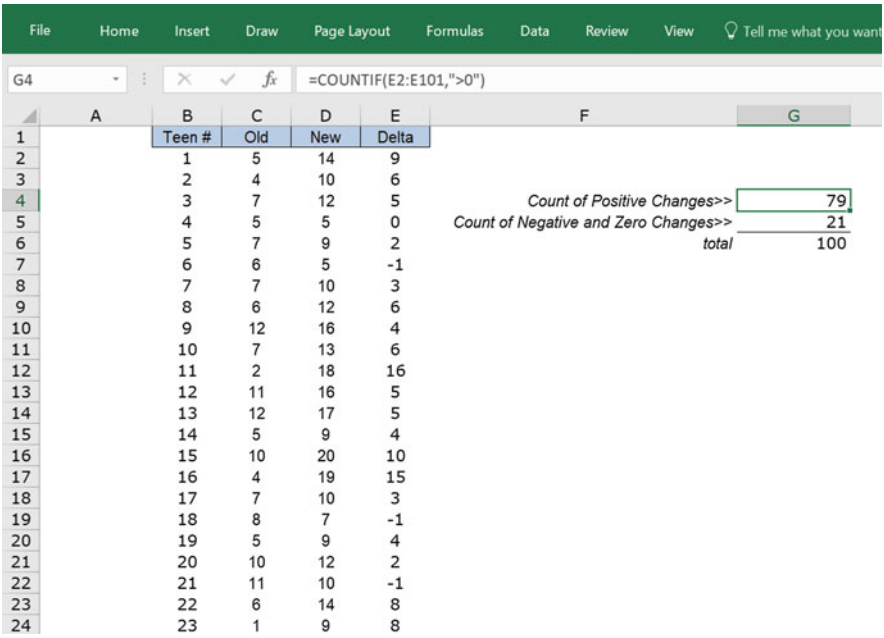
Fig. 3.29 New and old website page-views descriptive statistics

As with our time series data, there are many other questions we could ask, but these four questions are certainly important to our understanding of the effectiveness of the *new* website design. Additionally, as we engage in the analysis, other questions of interest may arise. Let us begin with a simple examination of the data. Figure 3.29 presents the descriptive statistics for the new and old website data.

Notice that the mean of the views at the old website is 7.54 and the new website mean is 11.83. This appears to be a considerable difference—an increase of 4.29 pages visited. But, the difference could simply be a matter of the sample of 100 individuals we have chosen for our experiment; that is, the 100 observations may not be representative of the universe of potential website visitors. I will say that in the world of statistics, a random sample of 100 is often a substantial number of observations; that is, it is likely to represent the major population summary statistics adequately.

The website change in page-views represents an approximately 57% increase from the old page-views. Can we be sure that a 4.29 change is indicative of what will be seen in the universe of all potential teen website visitors? Fortunately, there are statistical tools available for examining the question of our confidence in the outcome of the 100 teens experiment—hypothesis tests, confidence intervals, etc. We will return to this question momentarily, but in the interim, let us examine the changes in the data a bit more carefully.

Each of the randomly selected teens has two data points associated with the data set: old and new website page-views. We begin with a very fundamental analysis: a calculation of the difference between the old and new page-views. Specifically, we count the number of teens that increase their number of views, and conversely, the number that reduce or remain at their current number of views. Figure 3.30 provides this analysis for these two categories of results. For the 100 teens in the study, 21 had



	A	B	C	D	E	F	G
1		Teen #	Old	New	Delta		
2		1	5	14	9		
3		2	4	10	6		
4		3	7	12	5	*Count of Positive Changes>>*	79
5		4	5	5	0	*Count of Negative and Zero Changes>>*	21
6		5	7	9	2	*total*	100
7		6	6	5	-1		
8		7	7	10	3		
9		8	6	12	6		
10		9	12	16	4		
11		10	7	13	6		
12		11	2	18	16		
13		12	11	16	5		
14		13	12	17	5		
15		14	5	9	4		
16		15	10	20	10		
17		16	4	19	15		
18		17	7	10	3		
19		18	8	7	-1		
20		19	5	9	4		
21		20	10	12	2		
22		21	11	10	-1		
23		22	6	14	8		
24		23	1	9	8		

Fig. 3.30 Change in each teen's page-views

fewer or the same number of page-views for the new design, while 79 viewed more pages. The column labeled *Delta*, column E, is the difference between the new and old website page-views, and the logical criteria used to determine if a cell will be counted is >0, which must be placed in quotes. It is shown in the formula bar as the formula—*Countif (E3:E102, ">0")*.

Again, this appears to be relatively convincing evidence that the website change has had an effect, but the strength and the certainty of the effect may still be in question. This is the problem with sampling—we can never be absolutely certain that the sample is representative of the population from which it is taken.

Sampling is a fact of life, and living with its shortcomings is unavoidable. We are often forced to sample because of convenience and the cost limitations associated with performing a census, and samples can lead to unrepresentative results for our population. This is one of the reasons why the mathematical science of statistics was invented: to help us quantify our level of comfort with the results from samples.

Fortunately, we have an important tool available in our *Descriptive Statistics* that helps us with sampling results—*Confidence Level*. This is the standard tool for determining how confident we are that the sample is sampled from the assumed population. We choose a particular **level of confidence**, 95% in our case, and we create an interval about the sample mean, above and below. Assume that the span for a 95% level of confidence is 6.5. If we find a mean of 50 for a sample, the interval is a span above and below the sample mean—43.5 to 56.5. If we repeatedly sample 100 teens many times, say 1000, from our potential teen population, approximately 950 of the CI's will capture the true population, mean and approximately 50 will not.

In Fig. 3.29 we can see the **Confidence Interval** (CI) for 95% at the bottom of the descriptive statistics. Make sure to check the *Confidence Level for Mean* box in the Descriptive Statistics dialogue box to return this value. A confidence level of 95% is very common and suitable for our application. So, our 95% confidence interval for the mean of the new website is $11.83 \pm 0.74832\ldots$, or approximately the range 11.08168 to 12.57832. For the old website, the confidence interval for the mean is $7.54 \pm 0.59186\ldots$, or the range 6.94814 to 8.13186. Note that the low end of the mean for the new website page-views (11.08168) is larger than the high end of the mean for the old page-views (8.13186). When this occurs, it very strongly suggests, with statistical confidence, that there is indeed a significant difference in the page views.

Next, we can expand on the analysis by not only considering the two categories, positive and non-positive differences, but also the magnitude of the differences. This is an opportunity to use the *Histogram* tool in *Data Analysis*. We will use bins values from −6 to 16 in one-unit intervals. These are the minimum and maximum observed values, respectively. Figure 3.31 shows the graphed histogram results of the column

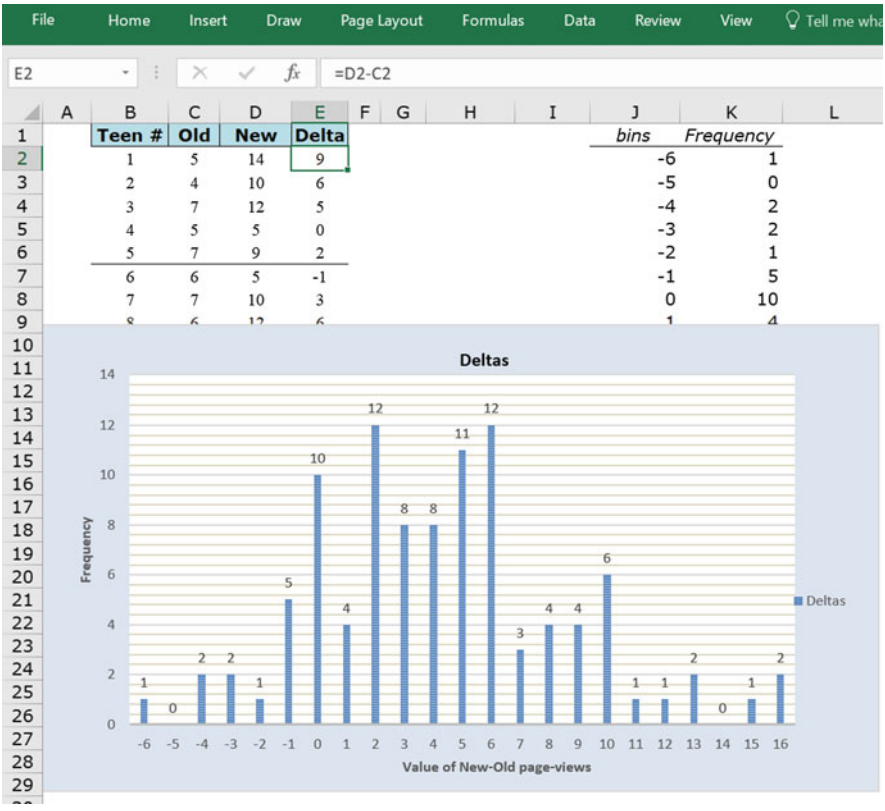


Fig. 3.31 Histogram of difference in each teen's page views

E (Delta). The histogram appears to have a central tendency around the range 2–6 page-views, and the calculated mean of page-views is 4.29. It also has a very small positive skew. For *perfectly* symmetrical distributions, the mean, the median, and the mode of the distribution are the same and the skewness measure is zero. Finally, if we are relatively confident about our sample of 100 teens being representative of all potential teens, we are ready to make a number of important statements about our data, given our current analysis:

1. If our sample of 100 teens is representative, we can expect an average improvement of about 4.29 pages after the change to the *new* web-site design.
2. There is considerable variation in the difference between new and old (Delta) evidenced by the range, −6 to 16. There is a central tendency in the graph that places many of the Delta values between 2 and 6.
3. We can also make statements such as: (1) I believe that approximately 21% of teens will respond negatively, or not at all, to the web-site changes; (2) approximately 51% of teens will increase their page views by 2–6 pages; (3) approximately 24% of teens will increase page views by 7 or more pages. These statements are based on observing the 100 teens and their distribution—simple count the teens in the range of interest. This will likely vary somewhat if another sample is taken. If these numbers are important to us, then we may want to take a much larger sample to improve of chances of stability in these percentages.
4. Our 95% CI about the new website mean is 11.83 ± 0.74832. This is a relatively tight interval. If a larger number of observations is taken in our sample, the interval will be even tighter ($< 0.74832\ldots$). The larger the sample, the smaller the interval for a given confidence interval, and vice versa.

Let us now move to a more sophisticated form of analysis, which answers questions related to our ability to generalize the sample result to the entire teen population. In the *Data Analysis* tool, there is an analysis called a **t-Test**. A t-test examines whether the means from two samples are equal or different; that is, whether they come from population distributions with the same mean, or not. Of special interest for our data is the **t-Test: Paired Two Sample for Mean**. It is used when *before* and *after* data is collected from the same sample group; that is, the same 100 teens being exposed to both the *old* web-site and the *new*.

By selecting *t-Test: Paired Two Sample for Means* from the *Data Analysis* menu, the two relevant data ranges can be selected, along with a hypothesized mean difference (0 in our case), because we will hypothesize *no* difference. Finally, an *alpha* value is requested. The value of alpha must be in the range 0–1. Alpha is the significance level related to the probability of making a **type 1 error** (rejecting a true hypothesis); the more certain you want to be about not making a type 1 error, the smaller the value of alpha that is selected. Often, an alpha of 0.05, 0.01, or 0.001 is appropriate, and we will choose 0.05. Once the data is provided in the dialogue box, a table with the resulting analysis appears. See Fig. 3.32 for the dialogue box inputs and Fig. 3.33 for the results.

The resulting **t-Stat**, 9.843008, is compared with a **critical value** of 1.660392 and 1.984217 for the one-tail and two-tail tests, respectively. This comparison amounts

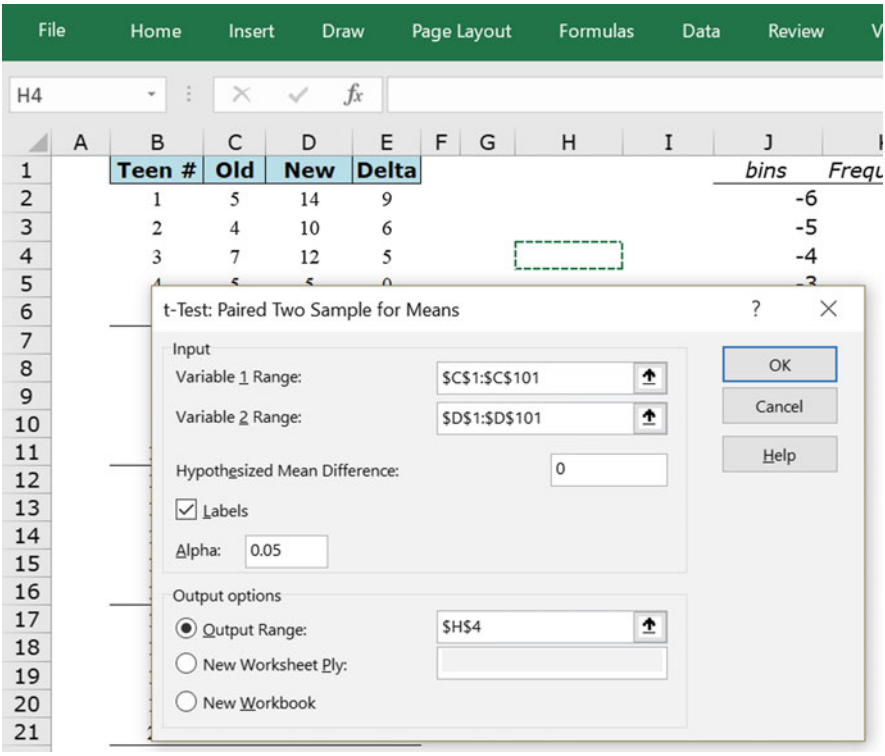
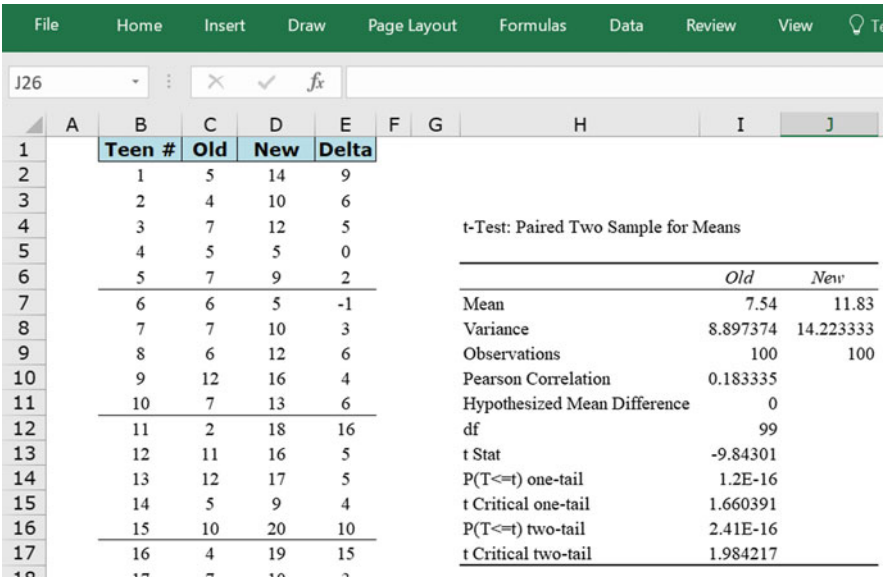


Fig. 3.32 t-test: Paired two sample for means dialogue box

to what is known as a **test of hypothesis**. In hypothesis testing, a **null hypothesis** is established: the means of the underlying populations are the *same,* and therefore their difference is equal to 0. If the calculated t-stat value is larger than the critical values, then the hypothesis that the difference in means is equal to 0 is *rejected* in favor of the alternative that the difference is *not* equal to 0. For a **one-tail test**, we assume that the result of the rejection implies an alternative in a particular direction (higher or lower). In our case, we compare the one-tail critical value (1.660392) to the resulting *t-Stat* (9.843008), where we assume that if we reject the hypothesis that the means are equal. We then favor that the *new* page-view mean is in fact *greater* than the *old*. The analysis that gave us a 4.29-page increase would strongly suggest this alternative. The one-tail test does in fact reject the null hypothesis since $9.843008 > 1.660392$. So, this implies that the difference in means is *not* zero. We will discuss details for hypothesis tests in future chapters.

If we decide not to impose a direction for the alternative hypothesis, a **two-tail test** of hypothesis is selected. We might be interested in results in both directions: a possible higher mean suggesting that the new website improves page views or a lower mean suggesting that the number of new site views is lower than before.



Teen #	Old	New	Delta
1	5	14	9
2	4	10	6
3	7	12	5
4	5	5	0
5	7	9	2
6	6	5	-1
7	7	10	3
8	6	12	6
9	12	16	4
10	7	13	6
11	2	18	16
12	11	16	5
13	12	17	5
14	5	9	4
15	10	20	10
16	4	19	15

t-Test: Paired Two Sample for Means

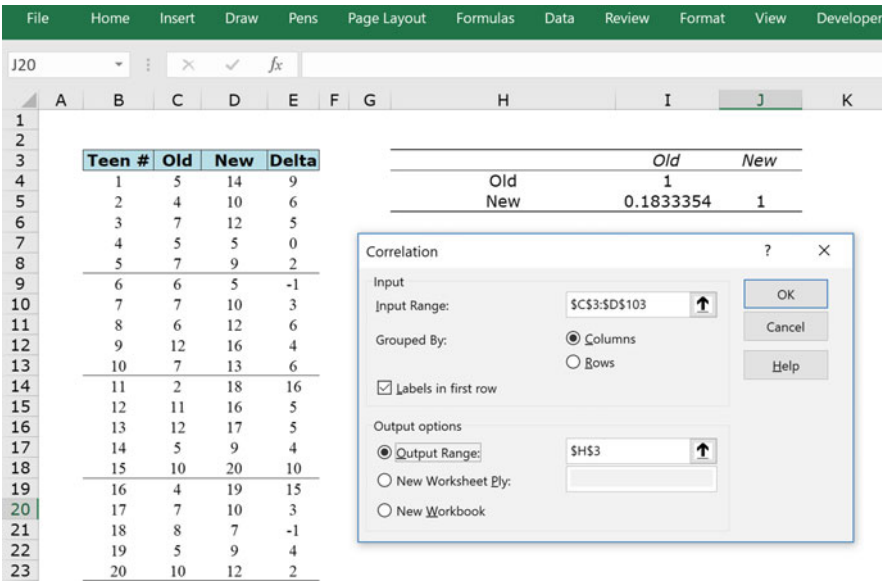
	Old	*New*
Mean	7.54	11.83
Variance	8.897374	14.223333
Observations	100	100
Pearson Correlation	0.183335	
Hypothesized Mean Difference	0	
df	99	
t Stat	-9.84301	
P(T<=t) one-tail	1.2E-16	
t Critical one-tail	1.660391	
P(T<=t) two-tail	2.41E-16	
t Critical two-tail	1.984217	

Fig. 3.33 t-test: Paired two sample for means results

The critical value (1.9842...) in this case is also much smaller than the t-Stat (9.843...). This indicates that we can *reject* the notion that the means for the new and old page views are equal. Thus, outcomes for both the one-tail and two-tail tests suggest that we should believe that the web-site has indeed improved page views.

Although this is not the case in our data, in situations where we consider more than two means and more than a single factor in the sample (currently we consider a visitor's status as a *teen* as a single factor), we can use **ANOVA** (Analysis of Variance) to do similar analysis as we did in the t-tests. For example, what if we determine that gender of the teens might be an important factor, and we have an additional third website option? In that case, there are two new alternatives. We might randomly select 100 teens each (50 men and 50 women) to view three websites—the old website, a new one from web designer X, and a new one for web designer Y. This is a very different and more complex problem than our paired t-test data analysis, and certainly more interesting. ANOVA is more sophisticated and powerful statistical test than t-tests, and they require a basic understanding of inferential statistics. We will see more of these tests in later chapters.

Finally, we might wonder if most teens are *equally* affected by the new website—Is there a *predictable* number of additional web-pages that most teens will visit while viewing the new site? Our initial guess would suggest *no* because of the wide distribution of the histogram in Fig. 3.31. If every teen had been influenced to view exactly four more web-pages after viewing the *new* site, then the



Teen #	Old	New	Delta
1	5	14	9
2	4	10	6
3	7	12	5
4	5	5	0
5	7	9	2
6	6	5	-1
7	7	10	3
8	6	12	6
9	12	16	4
10	7	13	6
11	2	18	16
12	11	16	5
13	12	17	5
14	5	9	4
15	10	20	10
16	4	19	15
17	7	10	3
18	8	7	-1
19	5	9	4
20	10	12	2

Fig. 3.34 Correlation matrix for new and old page views

histogram would indicate a single value of 4 for all 100 observations. This is certainly not the results that we see. One way to statistically test this question is to examine the correlation of the two series. Just as we did for the product sales data, we can perform this analysis on the *new* and *old* web-page views. Figure 3.34 shows a correlation matrix for the analysis. The result is a relatively low positive correlation (0.183335), indicating a very slight linear movement of the series in the same direction. So, although there is an increase in page views, the increase is quite different for different individuals—some are greatly affected by the *new* site, others are not.

3.6.1 Findings

We have completed a thorough analysis of the cross-sectional data and we have done so using the *Data Analysis* tools in the *Analysis* group. So, what have we learned? The answer is similar to the analysis of the time series data—a great deal. Some of the major findings are presented below:

1. It appears that the change in the new website has had an effect on the number of page-views for teens. The average increase for the sample is 4.29.
2. There is a broad range in the difference of data (Delta), with 51% occurring from 2 to 6 pages and only 21% of the teens not responding positively to the new

website. This is determined by counting the occurrence of observations in the histogram in Fig. 3.31.
3. The 95% confidence interval for our sample of 100 is approximately 0.75 units about (±) the sample mean of 11.83. In a sense, the interval gives us a measure of how uncertain we are about the population mean for a given sample size: larger intervals suggest greater uncertainty; smaller intervals suggest greater certainty.
4. A *t-Test: Paired Two Sample for Means* has shown that it is *highly unlikely* that the means for the *old* and *new* views are equal. This reinforces our growing evidence that the website changes have indeed made a positive difference in page views among teens.
5. To further examine the extent of the change in views for individual teens, we find that our *Correlation* tool in *Data Analysis* suggests a relatively low value of positive correlation. This suggests that although we can expect a positive change with the *new* website, the *magnitude* of change for individuals is not a predictable quantity—it is highly variable.

3.7 Summary

Data analysis can be performed at many levels of sophistication, ranging from simple graphical examination of the data to far more complex statistical methods. This chapter has introduced the process of thorough examination of data. The tools we have used are those that are often employed in an initial or preliminary examination of data. They provide an essential basis for a more critical examination, in that they guide our future analyses by suggesting new analytical paths that we may want to pursue. In some cases, the analysis preformed in this chapter may be sufficient for an understanding of the data's behavior; in other cases, the techniques introduced in this chapter are simply a beginning point for further analysis.

There are several issues that we need to keep in mind as we embark on the path to data analysis:

1. Think carefully about the type of data you are dealing with and ask critical questions to clarify where the data comes from, the conditions under which it was collected, and the measures represented by the data.
2. Keep in mind that not all data analysis techniques are appropriate for all types of data: for example, sampling data versus population data, cross-sectional versus time series, and multi-attribute data versus single-attribute data.
3. Consider the possibility of data transformation that may be useful. For example, our cross-sectional data for the new and old website was combined to create a *difference*, our Delta data set. In the case of the time series data, we can adjust data to account for outliers (data that are believed to be unrepresentative) or one-time events, like promotions.
4. Use data analysis to generate further questions of interest. In the case of the teen's data, we made no distinction between male and female teens, or the actual ages of

the teens. It is logical to believe that a 13-year-old female web visitor may behave quite differently than a 19-year-old male. This data may be available for analysis, and it may be of critical importance for understanding behavior.

Often our data is in qualitative form rather than quantitative, or is a combination of both. In the next chapter, we perform similar analyses on qualitative data. It is important to understand the value of both types of data, because they both serve our goal of gaining insight. In some cases, we will see similar techniques applied to both types of data, but in others, the techniques will be quite different. Developing good skills for both types of analyses is important for anyone performing data analysis.

Key Terms

Add-in	Dependent variable
Error checking	Independent variable
Series	Linear regression
Treatment	Simple linear regression
Time series data	Multiple linear regression
Cross-sectional data	Beta
Cyclicality	Alpha
Seasonality	R-square
Leading	Residuals
Trend	Significance F
Linear trend	Covariance
E-tailer	Correlation
Page-views	Perfectly positively correlated
Frequency distribution	Perfectly negatively correlated
Central tendency	Winters' 3-factor exponential smoothing
Variation	Simple exponential smoothing
Descriptive statistics	Level of confidence
Mean	Confidence interval
Standard deviation	t-Test
Population	t-Test: Paired two sample for mean
Range	Type 1 error
Median	t-Stat
Mode	Critical value
Standard error	Test of hypothesis
Sample variance	Null hypothesis
Kurtosis	One-tail test
Skewness	Two-tail test
Systematic behavior	ANOVA

Problems and Exercises

1. What is the difference between time series and cross-sectional data? Give examples of both?
2. What are the three principle approaches we discussed for performing data analysis in Excel?
3. What is a frequency distribution?
4. Frequency distributions are often of little use with time series data. Why?
5. What are three statistics that provide location information of a frequency distribution?
6. What are two statistics describing the dispersion or variation of frequency distributions?
7. What does a measure of positive skewness suggest about a frequency distribution?
8. If a distribution is perfectly symmetrical, what can be said about its mean, median, and mode?
9. How are histograms and frequency distributions related?
10. What is the difference between a sample and a population?
11. Why do we construct confidence intervals?
12. Are we more of less confident that a sampling process will capture the true population mean if the level confidence is 95% or 99%?
13. What happens to the overall length of a confidence interval as we are required to be more certain about capturing the true population mean?
14. What is the difference between an independent variable and dependent variable in regression analysis?
15. You read in a newspaper article that a Russian scientist has announced that he can predict the fall enrollment of students at Inner Mongolia University (IMU) by tracking last spring's wheat harvest in metric tons in Montana, USA.

    (a) What are the scientist's independent and dependent variables?
    (b) You are dean of students at IMU, so this announcement is of importance for your planning. But you are skeptical, so you call the scientist in Moscow to ask him about the accuracy of the model. What measures of fit or accuracy will you ask the scientist to provide?

16. The Russian scientist provides you with an alpha (1040) and a beta (38.8) for the regression. If the spring wheat harvest in Montana is 230 metric tons, what is your prediction for enrollment?
17. The Russian scientist claims the sum of all residuals for his model is zero and therefore it is a perfect fit. Is he right? Why or why not?
18. What *Significance F* would you rather have if you are interested in having a model with a significant association between the independent and dependent variables—0.000213 or 0.0213?

19. In the covariance matrix below, answer the following questions:

    (a) What is the variance of C?
    (b) What is the covariance of B and D?

	A	B	C	D
A	432.10			
B	−345.10	1033.1		
C	19.23	−543.1	762.4	
D	123.81	−176.4	261.3	283.0

20. What is the correlation between amount of alcohol consumption and the ability to operate an automobile safely—Negative or positive?
21. Consider the sample data in the table below.

Obs. #	Early	Late
1	3	14
2	4	10
3	7	12
4	5	7
5	7	9
6	6	9
7	7	10
8	6	12
9	2	16
10	1	13
11	2	18
12	4	16
13	3	17
14	5	9
15	2	20

    (a) Perform an analysis of the descriptive statistics for each data category (Early and Late).
    (b) Graph the two series and predict the correlation between Early and Late—positive or negative?
    (c) Find the correlation between the two series.
    (d) Create a histogram for the two series and graph the results.
    (e) Determine the 99% confidence interval for the Early data.

22. Assume the Early and Late data in problem 21 represent the number of clerical tasks correctly performed by college students, who are asked to perform the tasks Early in the morning and then Late in the morning. Thus, student 4 performs 5 clerical tasks correctly Early in the morning and 7 correctly Late in the morning.

(a) Perform a test to determine if the means of the two data categories come from population distributions with the same mean. What do you conclude about the one-tail test and the two-tail test?
(b) Create a histogram of the differences between the two series—Late minus Early. Are there any insights that are evident?

23. *Advanced Problem*—Assume the Early and Late data in problem 21 is data relating to energy drinks sold in a college coffee shop on individual days—on day 1 the Early sales of energy drinks were 3 units and Late sales were 14 units, etc. The manager of the coffee shop has just completed a course in data analytics, and believes she can put her new-found skills to work. In particular, she believes she can use one of the series to predict future demand for the other.

(a) Create a regression model that might help the manager of the coffee shop to predict the Late purchases of energy drinks. Perform the analysis and specify (the regression equation and all its coefficients) the predictive formula.
(b) Do you find anything interesting about the relationships between Early and Late?
(c) Is the model a good fit? Why?
(d) Assume you would like to use the Late of a particular day to predict the Early of the next day—on day 1 use Late to predict Early on day 2. How will the regression model change?
(e) Perform the analysis and specify the predictive formula.