Background

2

In this chapter we review some basic statistical concepts, walk you through a few common statistical procedures, and discuss how to design a well-thought-out usability study. We won't burden you with too many formulas or complicated statistics. Instead, we focus on the practical side to give you a step-by-step guide to analyzing your usability data in the simplest possible way. We want to get you comfortable with these methods as easily and quickly as possible before you continue with the material in the rest of the book.

The first part of this chapter covers basic information about understanding data and designing a usability study. The second part includes a step-by-step guide to analyzing different types of usability data. We show you how to calculate and interpret descriptive statistics, how to compare means, how to examine relationships between variables, and how to use some nonparametric statistics. We address the techniques you will use most often as part of everyday usability testing. In our examples, we use Microsoft Excel for our calculations because it is so widely available. Many similar statistical software packages provide the same calculations with only minor variations in presentation and use.

2.1 DESIGNING A USABILITY STUDY

Many factors must be considered when you design a usability study. A well-thought-out study can save you time and effort and answer your research questions clearly. A poorly designed study can be just the opposite: a waste of time, money, and effort, without giving you the answers you need. To design a well-thought-out usability study, you need to answer these questions:

- What type of participants do I need?
- How many participants do I need?
- Am I going to compare the data from a single group of participants or from several different groups?
- Do I need to counterbalance (adjust for) the order of tasks?

We examine each of these questions and try to help you find the answers.

2.1.1 **Selecting Participants**

Selecting participants for a usability study should be a set of deliberate decisions based on factors such as cost, availability, appropriateness, and study goals. One of the most common criticisms of some usability studies is that the participants are not representative of the larger population or don't match the target audience. You should consider carefully how you select participants for your study and try to get all the interested parties to agree on the sampling strategy *before* you begin.

The first question you must answer is how well your participants should reflect your target audience. We recommend you try to recruit truly representative participants whenever you can. For example, if you're designing a new medical application for doctors to use in their practices, try to get practicing physicians as your participants. In some situations, however, you may have to settle for participants who are only close approximations of the target users. In those cases, be aware of the limitations of the data you collect. Many of the statistics you will calculate assume that the sample data reflects the larger population.

The second important question when selecting participants is whether you're going to divide your data by different types of participants. If you plan to separate your participants into distinct groups, think about what those groups are and about how many participants you want in each group. In usability, there are a few common types of groups or segments:

- Self-reported expertise in some domain (novice, intermediate, expert)
- Frequency of use (e.g., number of web visits or interactions per month)
- Amount of experience with something relevant (days, months, years)
- Demographics (gender, age, location, etc.)
- Activities (use of particular functionality or features)

The third question involves sampling strategy. The goal of larger (or quantitative) usability studies is to be able to generalize the findings to the larger population. To attain that goal, you need to develop a sampling strategy that allows you to say something about the overall user population. Here are some different sampling techniques:

Random sampling: Everyone in the population has a roughly equal probability of being selected to participate in the study. Random sampling involves numbering everyone on a list of all potential recruits and using a random number generator to select participants according to the desired sample size.

Systematic sampling: You select each participant based on predefined criteria. For example, you might select every tenth person from a list of users or every one hundredth person who passes through a turnstile at a sporting event.

Stratified sampling: You create subsamples of the entire population and ensure that certain sample sizes are achieved for each subgroup. The goal is to ensure that the sample reflects the larger population. For example, you might recruit 50 percent males and 50 percent females, or recruit a group of participants in which 20 percent are older than 65.

Samples of convenience: This approach, which is very common is usability studies, includes anyone willing to participate in a study. Locating participants for these samples might be done through an advertisement or by using a list of people who came to the lab for past testing. It's important to know how well a sample of convenience reflects the general population and to be aware of any special biases that may be reflected in their feedback or data.

2.1.2 Sample Size

One of the most commonly asked questions in the usability field concerns the sample size for a usability study. Everyone involved—usability professionals, project managers, market researchers, developers, designers—wants to know how many participants are enough in a usability study. There's no rule that says if you don't have at least x number of participants in a study, the data won't be valid. The sample size you choose should be based on two factors: the goals of your study and your tolerance for a margin of error.

If you are interested only in identifying major usability issues as part of an iterative design process, you can get useful feedback from three or four representative participants. This small sample means that you will not identify all or even most of the usability issues, but you can identify some of the more significant ones. If you have lots of tasks or several different parts of the product to evaluate, then you will certainly need more than four participants. As a general rule of thumb, during the early stages of design, you need fewer participants to identify the major usability issues. As the design gets closer to completion, you need more participants to identify the remaining issues. (Chapter 5 provides more discussion of this issue, as well as some simple statistics for determining sample sizes.)

The other issue to consider is how much error you're willing to accept. Table 2.1 shows how the degree of confidence (or confidence interval) changes based on an average 80 percent success rate for different sample sizes. We'll talk more about confidence intervals later, but the basic idea is that they show what projections you can make about the true value of a statistic for the whole population based on what you observed in your sample. For example, if eight out of ten participants in a usability test successfully completed a task, can you say that 80 percent of users in the larger population will be able to complete it? No, you can't.

As Table 2.1 shows, you can say (with 95 percent confidence) that somewhere between 48 and 95 percent of the people in the larger population will be able to successfully complete the task. But you can see that as the sample size increases, the lower and upper bounds of the 95 percent confidence interval move closer together. So if you ran 100 participants, and 80 of them completed the task successfully, you can say that 71 to 86 percent of the larger population will be able to complete the task (with 95 percent confidence). We'll talk more later about how you calculate confidence intervals.

Table 2.1 Example of How Confidence Intervals Change as a Function of Sample Size

Number Successful	Number of Participants	Lower 95% Confidence	Upper 95% Confidence
4	5	36%	98%
8	10	48%	95%
16	20	58%	95%
24	30	62%	91%
40	50	67%	89%
80	100	71%	86%

Note: These numbers indicate how many participants in a usability test successfully completed a given task and the confidence interval for that mean completion rate in the larger population.

2.1.3 Within-Subjects or Between-Subjects Study

Another important decision to make is whether you are going to be comparing different data for each participant (such as success rates for different designs of the product) or data from each participant to the other participants (such as success rates for different age groups). The first approach is commonly referred to as a within-subjects design, and the second is known as a between-subjects design. Both approaches have their strengths and weaknesses.

A within-subjects, or repeated-measures, design is most commonly used in studies when you want to evaluate how easily a participant can learn to use a particular product. By comparing metrics such as task completion times or errors across several trials with the same set of participants, you can determine how quickly and easily the participant becomes familiar with the product. A within-subjects study does not require as large a sample size, and you don't have to worry about differences across groups. Because each participant is being compared to himself, the differences you observe in the data cannot be attributed to differences between participants. The downside of a within-subjects design is that you may need to worry about "carryover effects," where performance in one condition impacts performance in another condition. A carryover effect might be the result of practice (improving performance) or fatigue (decreasing performance). If there is a possible carryover effect, you can counterbalance for this effect as you design the study and analyze the data.

A between-subjects study is used to compare results for different participants, such as differences in satisfaction between novices and experts or in task completion times for younger versus older participants. In another type of between-subjects design, participants are randomly assigned to groups and then receive

different treatments, such as different prototype designs for the same product. Because there is generally more variance across a group of participants than within a single participant, a between-subjects design requires a larger sample size. One advantage of a between-subjects design is the elimination of carryover effects, because any potential carryover effects would impact both groups equally.

If neither a between-subjects nor a within-subjects design meets your needs, consider a mixed design. A mixed design contains a between-subjects factor, such as gender, *and* a within-subjects factor, such as three trials distributed over time. For example, you could use a mixed-design study to find out if there is a difference in the way men and women perform some task across several trials. Mixed designs can be a very powerful technique in a usability study, and because they may eliminate the need for separate studies for each question that arises, they can also save time and money.

2.1.4 Counterbalancing

Sometimes the order in which participants perform their tasks has a significant impact on the results. Participants usually learn the product as their experience with it grows. As a result, you must consider the order in which the data are collected, which is usually the order of tasks. It's possible that you may see improvement in performance or satisfaction as the usability session continues. Can you determine if the improvement occurs because the fifth task is easier than the first task or if some learning takes place between the first and fifth tasks that makes the fifth task easier to perform? The only way to address this question is to control for order effects through a technique called counterbalancing.

Counterbalancing involves simply changing the order in which different tasks are performed. There are a few ways to do this. You can randomize the order of tasks by "shuffling" the task order prior to each participant, or you can create various orders ahead of time so that each participant performs each task in a different order. Table 2.2 shows one example of how to counterbalance task order. Notice how each task appears in each position only once. Task 2 (T2) is performed as the second task only once by participant 1. Participant 2 performs task 2 as the last task.

Table 2.2 Example of How to Counterbalance Task Order for Four Participants and Four Tasks

Participant	First Task	Second Task	Third Task	Fourth Task
P1	T1	T2	T3	T4
P2	T3	T1	T4	T2
P3	T2	T4	T1	T3
P4	T4	T3	T2	T1

If you suspect that there might be an order effect, we recommend you counter-balance task order. However, there are a few situations when it is not necessary or may even be detrimental to your findings. First, if the tasks are totally unrelated to each other, learning between tasks is unlikely. Performing one task successfully will not help in any other task. Second, counterbalancing is not appropriate when a natural order of tasks is present. Sometimes the order cannot be juggled because the test session would not make sense. In this situation, acknowledge order effects as part of the general learning process rather than as a symptom of a poorly designed usability study.

2.1.5 Independent and Dependent Variables

In any usability study you must identify both independent and dependent variables. An independent variable of a study is an aspect that you manipulate. Choose the independent variables based on your research questions. For example, you may be concerned with differences in performance between males and females, or between novices and experts, or between two different designs. All of these are independent variables that can be manipulated to answer specific research questions.

Dependent variables (also called outcome or response variables) describe what happened as the result of the study. A dependent variable is something you measure as the result of, or as dependent on, how you manipulate the independent variables. Dependent variables include metrics or measurements such as success rates, number of errors, user satisfaction, completion times, and many more. Most of the metrics we discuss in this book are dependent variables.

When you design a usability study, you must have a clear idea of what you plan to manipulate (independent variables) and what you plan to measure (dependent variables). If you don't have a clear idea of those, go back to the research goals. If you can draw a logical connection between the research goals and your independent and dependent variables, you and your study should be successful.

2.2 TYPES OF DATA

Data, the cornerstone of usability metrics, exist in many forms. In the world of usability, types of data include task completion rates, web traffic, responses to a satisfaction survey, or the number of problems a participant encounters in a lab test. To analyze usability data, you need to understand the four general types of data: nominal, ordinal, interval, and ratio. Each type of data has its own strengths and limitations. When collecting and analyzing usability data, you should know what type of data you're dealing with and what you can and can't do with each type.

2.2.1 Nominal Data

Nominal data are simply unordered groups or categories. Without order between the categories, you can say only that they are different, not that one is any better

than the other. For example, consider apples, oranges, and bananas. They are just different; no one fruit is inherently better than any other.

In usability, nominal data might be characteristics of different types of users, such as Windows versus Mac users, users in different geographic locations, or males as opposed to females. These are typically independent variables that allow you to segment the data by these different groups. Nominal data also include some commonly used dependent variables, such as task success. Nominal data could also be the number of participants who clicked on link A instead of link B, or participants who chose to use a remote control instead of the controls on a DVD player itself.

Among the statistics you can use with nominal data are simple descriptive statistics such as counts and frequencies. For example, you could say that 45 percent of the participants are female, or 200 participants have blue eyes, or 95 percent were successful on a particular task.

One important thing to remember when you work with nominal data is how you code the data. In statistical analysis programs such as Excel, it's common to represent the membership in each group using numbers. For example, you might code males as group "1" and females as group "2." Remember that those figures are not data to be analyzed as numbers: An average of these values would be meaningless. (You could just as easily code them as "F" and "M.") The software can't distinguish between numbers used strictly for coding purposes, like these, and numbers whose values have true meaning. One useful exception to this is task success. If you code task success as a "1" and task failure as "0," the average will be the same as the proportion of users who were successful.

2.2.2 Ordinal Data

Ordinal data are ordered groups or categories. As the name implies, the data are organized in a certain way. However, the intervals between the measurements are not meaningful. Some people think of ordinal data as ranked data. The list of the top 100 movies, as rated by the American Film Institute (AFI), shows that the tenth best movie of all time, *Singing in the Rain,* is better than the twentieth best movie of all time, *One Flew Over the Cuckoo's Nest.* But these ratings don't say that *Singing in the Rain* is *twice* as good as *One Flew Over the Cuckoo's Nest.* One film is just *better* than the other, at least according to the AFI. Because the distance between the ranks is not meaningful, you cannot say one is twice as good as the other. Ordinal data might be ordered as better or worse, more satisfied or less satisfied, or more severe or less severe. The relative ranking (the order of the rankings) is the only thing that matters.

In usability, the most common occurrence of ordinal data comes from self-reported data on questionnaires. For example, a participant might rate a website as excellent, good, fair, or poor. These are relative rankings: The distance between excellent and good is not necessarily the same distance between good and fair. Severity ratings are another example of ordinal data. A usability specialist might assign a severity rating of high, medium, or low for each problem that the

participant encountered. The distance between high and medium is not necessarily the same as that between medium and low.

The most common way to analyze ordinal data is by looking at frequencies. For example, you might report that 40 percent of the participants rated the site as excellent, 30 percent as good, 20 percent as fair, and 10 percent as poor. Calculating an average ranking may be tempting, but it's statistically meaningless.

2.2.3 Interval Data

Interval data are continuous data where the differences between the measurements are meaningful but there is no natural zero point. An example of interval data familiar to most of us is temperature, either Celsius or Fahrenheit. Defining zero as the freezing point (in the case of Celsius) or as the point at which pipes begin to burst (in the case of Fahrenheit) is completely arbitrary. Zero degrees does not mean the absence of heat; it only identifies a meaningful point on the scale of temperatures.

In usability, the System Usability Scale (SUS) is one example of interval data. SUS (described in detail in Chapter 6) is based on self-reported data from a series of questions about the overall usability of any system. Scores range from 0 to 100, with a higher SUS score indicating better usability. The distance between each point along the scale is meaningful in the sense that it represents an incremental increase or decrease in perceived usability.

Interval data allow you to calculate a wide range of descriptive statistics (including averages, standard deviation, etc.). There are also many inferential statistics that can be used to generalize about a larger population. Interval data provide many more possibilities for analysis than either nominal or ordinal data. Much of this chapter will review statistics that can be used with interval data.

One of the debates you can get into with people who collect and analyze subjective ratings is whether you must treat the data as purely ordinal or you can treat it as being interval. Consider these two rating scales:

○ Poor ○ Fair ○ Good ○ Excellent

Poor ○ ○ ○ ○ ○ Excellent

At first glance, you might say the two scales are the same, but the difference in presentation makes them different. Putting explicit labels on the items in the first scale makes the data ordinal. Leaving the intervening labels off in the second scale and only labeling the end points make the data more "interval-like." That's the reason that most subjective rating scales only label the ends, or "anchors," and not every data point.

Consider a slightly different version of the second scale:

Poor ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Excellent

Presenting it this way, with 10 points along the scale, makes it even more obvious that the data can be treated as if they were interval data. The reasonable interpretation of this scale by a user is that the distances between all the data points along the scale are equal. A question to ask yourself when deciding whether you

can treat some data like these as interval or not is whether a point halfway between any two of the defined data points makes sense. If it does, then it makes sense to analyze the data as interval data.

2.2.4 Ratio Data

Ratio data are the same as interval data, with the addition of an absolute zero. These data mean that the zero value is not arbitrary, as with interval data, but has some inherent meaning. With ratio data, the differences between the measurements are interpreted as a ratio. Examples of ratio data are age, height, and weight. In each example, zero indicates the absence of age, height, or weight.

In usability, the most obvious example of ratio data is time to completion. Zero seconds left would mean no time or duration remaining. Ratio data let you say something is twice as fast or half as slow as something else. For example, you could say that one participant is twice as fast as another user in completing a task.

There aren't many additional analyses you can do with ratio data compared to interval data in usability. One exception is calculating a geometric mean, which might be useful in measuring differences in time (Nielsen, 2001a). Aside from that calculation, there really aren't many differences between interval and ratio data in terms of the available statistics.

2.3 METRICS AND DATA

Choosing the right statistics is critical. Choosing the wrong statistical test and ending up with an incorrect conclusion could invalidate your results and negate your entire usability test. Table 2.3 shows some of the common statistical tests

Table 2.3 Choosing the Right Statistics for Different Data Types and Usability Metrics

Data Type	Common Metrics	Statistical Procedures
Nominal (categories)	Task success (binary), errors (binary), top-2-box scores	Frequencies, crosstabs, Chi-square
Ordinal (ranks)	Severity ratings, rankings (designs)	Frequencies, crosstabs, chi-square, Wilcoxon rank sum tests, Spearman rank correlation
Interval	Likert scale data, SUS scores	All descriptive statistics, t-tests, ANOVAs, correlation, regression analysis
Ratio	Completion time, time (visual attention), average task success (aggregated)	All descriptive statistics (including geometric means), t-tests, ANOVAs, correlation, regression analysis

described in the rest of this chapter. The type of data you are examining will dictate different tests. In Chapter 3 we discuss the reasons you should choose one metric over another.

2.4 DESCRIPTIVE STATISTICS

Descriptive statistics are essential for any interval or ratio-level data. Descriptive statistics, as the name implies, describe the data without saying anything about the larger population. Inferential statistics let you draw some conclusions or infer something about a larger population above and beyond your sample. Descriptive statistics are very easy to calculate using most statistical software packages. Let's assume we have the time data shown in Figure 2.1.

For our examples, we use Excel to analyze the data. First click on ''Tools,'' then on ''Data Analysis.'' (*Note:* If you do not see a ''Data Analysis'' option on the ''Tools'' menu, you must first install it by choosing ''Tools'' > ''Add-ins'' and then checking ''Analysis Toolpak.'') In the ''Data Analysis'' window, choose ''Descriptive Statistics'' from the list of analysis options. Next, define the range of data on which you want to run the descriptive statistics. In our example, we identify as the input range of data column B from row 1 through row 13. (Notice that the first row in the spreadsheet data is the label, so we check the box indicating ''labels in first row.'' Including the label is optional, but it helps you keep your data more organized.) Next, select the output range (the top, left corner of the area where you want the results to be placed). Finally, we indicate that we want to see the summary statistics and the 95 percent confidence interval. (The 95 percent confidence interval is the default selection in Excel.)

The results of the descriptive statistics are shown in Figure 2.2. Our hypothetical raw task time data for 12 participants are shown on the left side (columns A and B). The descriptive statistics are shown on the right side (columns D and E). We will review this output in the next several sections.

	A	B
1	Participant	Task Time
2	P1	34
3	P2	33
4	P3	28
5	P4	44
6	P5	46
7	P6	21
8	P7	22
9	P8	53
10	P9	22
11	P10	29
12	P11	39
13	P12	50
14		

FIGURE 2.1

Time to complete a task, in seconds, for 12 participants in a usability study.

	A	B	C	D	E
1	Participant	Task Time		*Task Time*	
2	P1	34			
3	P2	33		Mean	35.08333333
4	P3	28		Standard Error	3.246112671
5	P4	44		Median	33.5
6	P5	46		Mode	22
7	P6	21		Standard Deviation	11.24486415
8	P7	22		Sample Variance	126.4469697
9	P8	53		Kurtosis	−1.321525965
10	P9	22		Skewness	0.251441718
11	P10	29		Range	32
12	P11	39		Minimum	21
13	P12	50		Maximum	53
14				Sum	421
15				Count	12
16				Confidence Level(95.0%)	7.144645813
17					

FIGURE 2.2

Example data showing the output of the descriptive statistics option in Excel.

2.4.1 Measures of Central Tendency

Measures of central tendency are the first thing you should look at when you run descriptive statistics. Central tendency is simply the middle, or central, part of any distribution. The three most common measures of central tendency are mean, median, and mode. In Figure 2.2, the mean, or average, is 35.1, so the average time to complete this task was just over 35 seconds. The mean of most usability metrics is extremely useful and is probably the most common statistic cited in a usability report.

The median is the midway point in the distribution: Half the participants are below the median and half are above the median. In Figure 2.2, the median is equal to 33.5 seconds: Half of the participants were faster than 33.5 seconds, and half of the participants were slower than 33.5 seconds. In some cases, the median can be more revealing than the mean. In an example of salaries, median salaries for a company are more commonly reported because the higher executive salaries will skew the mean value so much that the average salary appears much higher than the majority really are. In such cases involving possible extreme values (as is sometimes the case with time data), consider using the median.

The mode is the most commonly occurring value. In Figure 2.2, the mode is 22 seconds: Two participants completed the task in 22 seconds. It's not common to report the mode in usability test results, but it may be useful to know it. When the data are continuous over a broad range, such as the completion times shown in Figure 2.2, the mode is generally less useful. When data have a more limited set of values (such as subjective rating scales), the mode is more useful.

HOW MANY DECIMAL PLACES TO USE WHEN REPORTING DATA

One of the most common mistakes usability specialists make is reporting the data from a usability test (mean times, task completion rates, etc.) with more precision than they really deserve. For example, Figure 2.2 shows that the mean time was 35.08333333 seconds. Is that the way you should report the mean? Of course not. That many decimal places may be mathematically correct, but it's ridiculous from a practical standpoint. Who cares whether the mean was 35.083 or 35.085 seconds? When you're dealing with tasks that took *about* 35 seconds to complete, a few milliseconds or a few hundredths of a second make no difference whatsoever. So how many decimal places should you use?

There's no universal answer, but some of the factors to consider are the accuracy of the original data, their magnitude, and its variability. The original data in Figure 2.2 appear to be accurate to the nearest second. One rule of thumb is that the number of significant digits you should use when reporting a statistic, such as the mean, is no more than one additional significant digit in comparison to the original data. So in this example, you could report that the mean was 35.1 seconds.

2.4.2 Measures of Variability

Measures of variability show how the data are spread or dispersed across the range of all the data. These measures help answer the question "Do most users have similar completion times, or is there a wide range of times?" Determining the variability is critical if you want to know how confident you can be of the data. The greater the variability or spread in the data, the less dependable the data are relative to understanding the general population. The less the variability or spread, the more confidence you can have in relating the findings to a larger population. There are three common measures of variability: the range, the variance, and the standard deviation.

The range is the distance between the minimum and maximum data points. In Figure 2.2, the range is 32, with a minimum time of 21 seconds and a maximum time of 53 seconds. The range can vary wildly depending on the metric. For example, in many kinds of rating scales, the range is usually limited to five or seven, depending on the number of values used in the scales. When you study completion times, the range is very important because it will identify "outliers" (data points that are at the extreme top and bottom of the range). Looking at the range is also a good check to make sure that the data are coded properly. If the range is supposed to be from one to five, and the data include a seven, you know there is a problem.

Variance, another common and important measure of variability, tells you how spread out the data are relative to the average or mean. The formula for calculating variance measures the difference between each individual data point and the mean, squares that value, sums all of those squares, and then divides the result by the sample size minus 1. In Figure 2.2, the variance is 126.4.

Once you know the variance, you can easily calculate the standard deviation, the most commonly used measure of variability. The standard deviation is simply

the square root of the variance. The standard deviation in the example shown in Figure 2.2 is 11 seconds. Interpreting this measure of variability is a little easier than interpreting the variance, since the unit of the standard deviation is the same as the original data (seconds in this example).

2.4.3 Confidence Intervals

Confidence intervals are extremely valuable for any usability professional. A confidence interval is a range that estimates the true population value for a statistic. For example, assume that you need to estimate the mean for the entire population and you want to be 95 percent certain about what that mean is. In Figure 2.2 the 95 percent confidence interval is just over 7 seconds. This means that you can have a 95 percent confidence that the population mean is 35 seconds plus or minus 7 seconds, or between 28 seconds and 42 seconds.

Alternatively, you can quickly calculate the confidence interval using the CONFIDENCE function in Excel. The formula is very easy to construct:

$$= \text{confidence (alpha, standard deviation, sample size)}$$

The alpha is your significance level, which is typically 5 percent, or 0.05. The standard deviation is easily calculated using the Excel "stdev" function. The sample size is simply the number of cases or data points you are examining, which is easily calculated using the "Count" function. Figure 2.3 shows an example.

The CONFIDENCE function in Excel works a little differently from the confidence level calculated as part of the descriptive statistics function shown in Figure 2.2. Remember, however, that this function is based on the standard deviation of the population, which in many cases is unknown. As a result, the

FIGURE 2.3

Example of how to calculate a 95 percent confidence interval using the CONFIDENCE function in Excel.

	A	B	C	D
	B17 ▼	f_x =CONFIDENCE(0.05,B16,B15)		
1	Participant	Task Time		
2	P1	34		
3	P2	33		
4	P3	28		
5	P4	44		
6	P5	46		
7	P6	21		
8	P7	22		
9	P8	53		
10	P9	22		
11	P10	29		
12	P11	39		
13	P12	50		
14	Mean	35.08333		
15	Count	12		
16	Standard Deviation	11.24486		
17	95% Confidence Level	6.362264		
18				

confidence intervals don't quite match up between Figures 2.2 and 2.3. We recommend that you take a more conservative perspective and use the confidence level as part of the descriptive statistics (Figure 2.2), since this does not make any assumptions about the standard deviation of the population.

Ultimately, which calculation method you choose does not matter a great deal. As your sample size increases, the difference between the two calculations gets smaller. What really matters is that you use one of them. We cannot emphasize enough the importance of calculating and presenting confidence levels with your metrics.

2.5 COMPARING MEANS

One of the most useful things you can do with interval or ratio data is to compare different means. If you want to know whether one design has higher satisfaction rates than another, or if error rates are higher for one group of participants compared to another, your best approach is through statistics. This is quite easy using Excel or many other analysis packages. Therefore, instead of providing all the formulas for the various ways of comparing means, we will explain when to use each, how to use each, and how to interpret the results.

There are several ways to compare means, but before you jump into the statistics, you should know the answers to a few questions:

1. Is the comparison *within* the same set of participants or *across* different participants? For example, if you are comparing some data for men and women, it is highly likely that these are different participants. Comparing different samples like this is called independent samples. But if you're comparing the same group of participants on two different products or designs (a within-subjects design), you will use something called repeated measures analysis or paired samples.

2. What is the sample size? If the sample size is less than 30, use a *t*-test. If the sample is 30 or more, use a *z*-test.

3. How many samples are you comparing? If you are comparing two samples, use a *t*-test. If you are comparing three or more samples, use an analysis of variance (also called ANOVA).

2.5.1 Independent Samples

Frequently in usability studies you're comparing means based on independent samples. This only means that the groups are different. For example, you might be interested in comparing satisfaction rates between expert and novice participants. The most common question is whether the two groups are different. Doing this is easy in Excel. First, go to "Data" > "Tools" and choose the option of "*t*-test: Two Samples

	A	B	C	D	E	F
1	Expert_time	Novice_time		t-Test: Two Samples Assuming Equal Variance		
2	34	45				
3	33	48			Expert_time	Novice_time
4	28	53		Mean	35.08333333	49.33333333
5	44	66		Variance	126.4469697	229.6969697
6	46	67		Observations	12	12
7	21	35		Pooled Variance	178.0719697	
8	22	39		Hypothesized Mean Difference	0	
9	53	21		df	22	
10	22	34		t Stat	-2.61572876	
11	29	55		P(T<=t) one-tail	0.007892632	
12	39	59		t Critical one-tail	1.717144335	
13	50	70		P(T<=t) two-tail	0.015785265	
14				t Critical two-tail	2.073873058	

FIGURE 2.4

Output from an independent samples t-test in Excel.

Assuming Equal Variances" (of course, assuming that the two variances are roughly equal). Next, input the data for both variables. In this example, the data are coming from columns A and B in our Excel spreadsheet (Figure 2.4). We decided to include the labels in each row, so we check the label option. Because we hypothesize that there is no difference in means, we enter 0. (You might hear this called the "null hypothesis.") This means we are testing whether there is a difference between the two variables. We select an alpha level equal to 0.05. This means we want to be 95 percent confident of our results. Another way to think of this is that we are willing to be wrong 5 percent of the time by concluding there is a difference when there really is not one.

The output from this analysis is shown in Figure 2.4. The first thing you might notice is the difference in means between the novices and the experts. Experts are faster (35 seconds) compared to novices (49 seconds). The other very important piece of data here is the p-value. Because we aren't making any assumptions ahead of time about who might be faster (experts or novices), we look at the p-value for the two-tailed distribution. The p-value is about 0.016, which is well below our 0.05 threshold. Therefore, we can say there is a statistically significant difference in completion times between novices and experts, at the 0.05 level. It's important to remember to state the alpha level because it shows how much error you are willing to accept.

2.5.2 Paired Samples

A paired samples t-test is used when you're comparing means within the same set of participants. For example, you may be interested in knowing whether there is a difference between two prototype designs. If you have the same set of participants perform tasks using prototype A and then prototype B, and you are measuring variables such as self-reported ease of use and time, you will use a paired samples t-test.

Running this test is simple in Excel. On the main menu, go to "Tools" > "Data Analysis." Choose the option "t-Test: Paired Two Samples for Means." Next, select

	A	B	C	D	E	F	G
1	Participant	Design A_SUS	Design B_SUS		t-Test: Two Samples for Means		
2	P1	80	48				
3	P2	88	55			*Design A_SUS*	*Design B_SUS*
4	P3	76	53		Mean	77.7500000	57.0833333
5	P4	90	80		Variance	125.4772727	153.7196970
6	P5	93	81		Observations	12.0000000	12.0000000
7	P6	67	51		Pearson Correlation	0.6521213	
8	P7	68	61		Hypothesized Mean Difference	0.0000000	
9	P8	55	41		df	11.0000000	
10	P9	77	55		t Stat	7.2295917	
11	P10	71	57		P(T<=t) one-tail	0.0000084	
12	P11	88	59		t Critical one-tail	1.7958848	
13	P12	80	44		P(T<=t) two-tail	0.0000169	
14					t Critical two-tail	2.2009852	

FIGURE 2.5

Output from a paired samples *t*-test in Excel.

the two columns that will be compared. In this example, columns B and C are being compared, from row 2 through row 13 (Figure 2.5). Next, indicate the "Hypothesized Mean Difference." In this example, we chose 0 because we are hypothesizing that there is no difference between the means of columns B and C. Next, we set an alpha value to 0.05, saying that we want to be 95 percent confident in our findings. The output options are where you want to see the results. This dialog box is set up exactly like the independent samples *t*-test. The main difference, of course, is that the comparisons are within the same participant as opposed to across different participants.

The outcome is shown in Figure 2.5. The raw data are on the left side, and the results are on the right side. As with the independent samples output, it's important to look at the mean and the variance. The *p*-value of 0.0000169 indicates that there is a significant difference between the two designs, since this number is considerably smaller than 0.05.

Notice that in a paired samples test, you should have an equal number of values in each of the two distributions you're comparing (although it is possible to have missing data). In the case of independent samples, the number of values does not need to be equal. You might happen to have more participants in one group than the other.

2.5.3 Comparing More Than Two Samples

We don't always compare only two samples. Sometimes we want to compare three, four, or even six different samples. Fortunately, there is a way to do this without a lot of pain. An analysis of variance (commonly referred to as an ANOVA) lets you determine whether there is a significant difference across more than two groups.

Excel lets you perform three types of ANOVA. We will give an example for just one type of ANOVA, called a single-factor ANOVA. A single-factor ANOVA is used

	A	B	C	D	E	F	G	H	I	J	K
1	Design 1_time	Design 2_time	Design 3_time		Anova: Single Factor						
2	34	45	66								
3	33	48	45		SUMMARY						
4	28	53	89		Groups	Count	Sum	Average	Variance		
5	44	66	49		Design 1_time	12	421	35.08333	126.447		
6	46	67	55		Design 2_time	12	592	49.33333	229.697		
7	21	35	77		Design 3_time	12	802	66.83333	333.4242		
8	22	39	90								
9	53	21	43								
10	22	34	56		ANOVA						
11	29	55	66		Source of Variation	SS	df	MS	F	P-value	F crit
12	39	59	69		Between Groups	6069.5	2	3034.75	13.20283	0.00006	3.284918
13	50	70	97		Within Groups	7585.25	33	229.8561			
14											
15					Total	13654.75	35				

FIGURE 2.6

Output from a single-factor ANOVA in Excel.

when you have just one variable you want to examine. For example, you might be interested in comparing task completion times across three different prototypes.

To run an ANOVA in Excel, first select "ANOVA: Single Factor" from "Tools" > "Data Analysis." This just means that you are looking at one variable (factor). Next, define the range of data. In our example (Figure 2.6), the data are in columns A, B, and C. We have set an alpha level to 0.05 and have included our labels in the first row.

The results are shown in two parts (see Figure 2.6). The top part is a summary of the data. As you can see, the average time for Design 3 is quite a bit slower, and Design 1 completion times are faster. Also, the variance is greater for Design 3 and less for Design 1. The second part of the output lets us know whether this difference is significant. The F-value is equal to 13.20. The critical value that we need to achieve significance is 3.28. The p-value of 0.00006 reflects the statistical significance of this result. Understanding exactly what this means is important: It means that there is a significant effect due to the "designs" variable. It does not necessarily mean that each of the design means is significantly different from each of the others—only that there *is* an effect overall. To see if any two means are significantly different from each other, you could do a two samples t-test on just those two sets of values.

2.6 RELATIONSHIPS BETWEEN VARIABLES

Sometimes it's important to know about the relationship between different variables. We've seen many cases where someone observing a usability test for the first time remarks that what participants say and what they do don't always correspond with each other. Many participants will struggle to complete just a few tasks with a prototype, but when asked to rate how easy or difficult it was, they often give it good ratings. In this section we provide examples of how to perform analyses that investigate these kinds of relationships (or the lack thereof).

2.6.1 Correlations

When you first begin examining the relationship between two variables, it's important to visualize what the data look like. In Excel, it's easy to create a scatterplot of the two variables. Figure 2.7 is an example of a scatterplot that shows the relationship between months of experience (x-axis) and average errors per day (y-axis). Notice that as the months of experience increase, the average number of errors drops. This is called a negative relationship because as one variable increases (months of experience), the other variable decreases (errors per day). The line that runs through the data is called a trend line and is easily added to the chart in Excel by right-clicking on any one of the data points and selecting "Add Trend Line." The trend line helps you to better visualize the relationship between the two variables. You can also have Excel display the R^2 value (a measure of the strength of the relationship) by clicking on the "Options" tab and checking the box next to "Display R-squared value on chart."

Plotting your data in a scatterplot is just the first step. You also want to understand the degree of association between the two variables. This can be done by choosing the CORRELATION function in the "Data Analysis" tool.

Next, specify the range of data. In Figure 2.8, the data are in columns B and C. We have also included the first row, which is the label for the row. The only decision here is to choose the output range (where you see the results). The output from the correlation is something called a correlation coefficient, or r value. A correlation coefficient is a measure of the strength of the relationship between the two variables and has a range from -1 to $+1$. The stronger the relationship, the closer the value is to -1 or $+1$, and the weaker the relationship, the closer the correlation coefficient is to 0.

Figure 2.8 shows a correlation coefficient of -0.76. The negative correlation coefficient means that it is a negative relationship (as experience increases, errors decrease). This correlation also tells us that there is a pretty strong relationship
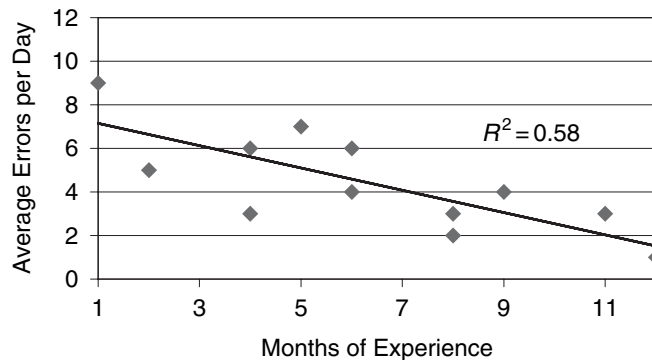

FIGURE 2.7

An example of a scatterplot (with trend line) in Excel.

	A	B	C	D	E	F	G
1	Participant	Months of Experience	Errors			*Months of Experience*	*Errors*
2	P1	6	4		Months of Exp	1	
3	P2	6	6		Errors	–0.76160463	1
4	P3	8	3				
5	P4	5	7				
6	P5	4	3				
7	P6	12	1				
8	P7	11	3				
9	P8	1	9				
10	P9	9	4				
11	P10	8	2				
12	P11	4	6				
13	P12	2	5				

FIGURE 2.8

Output data from a correlation in Excel.

between experience and errors. Another way of thinking about it is that if you knew one of the two values, you could predict the other one reasonably well.

2.7 NONPARAMETRIC TESTS

Nonparametric tests are used for analyzing nominal and ordinal data. For example, you might want to know if there is a significant difference between men and women for success and failure on a particular task. Or perhaps you're interested in determining whether there is a difference between experts, intermediates, and novices on how they ranked different websites. To answer questions that involve nominal and ordinal data, you will need to use some type of nonparametric test.

Nonparametric statistics make different assumptions about the data than the statistics we've reviewed for comparing means and describing relationships between variables. For instance, when we run t-tests and correlation analysis, we assume that data are normally distributed and variances are approximately equal. The distribution is not normal for nominal or ordinal data. Therefore, we don't make the same assumptions about the data in nonparametric tests. For example, in the case of (binary) success, when there are only two possibilities, the data are based on the binomial distribution. Some people like to refer to nonparametric tests as "distribution-free" tests. There are a few different types of nonparametric tests, but we will just cover the chi-square test because it is probably the most commonly used.

2.7.1 The Chi-Square Test

The chi-square test is used when you want to compare categorical (or nominal) data. Let's consider an example. Assume you're interested in knowing whether there is a significant difference in task success between three different groups: novices,

intermediates, and experts. You run a total of 60 participants in your study, 20 in each group. You measure task success or failure on a single task. You count the number of participants who were successful in each group. For the novices, only 6 out of 20 were successful, 12 out of 20 intermediates were successful, and 18 out of 20 experts were successful. You want to know if there is a statistically significant difference between the groups.

To perform a chi-square test in Excel, you use the CHITEST function. This function calculates whether the differences between the observed and expected values are simply due to chance. The function is relatively easy to use: =CHITEST (actual_range, expected_range). The actual range is the number of people who were successful on the task for each group. The expected range is the total number of participants successful (33) divided by the number of groups (3), or equivalent to 11 in this example. The expected value is what you would expect if there were no differences between any of the three groups.

Figure 2.9 shows what the data look like and the output from the CHITEST function. In this example, the likelihood that this distribution is due to chance is about 2.9 percent (0.028856). Because this number is less than 0.05 (95 percent confidence), we can reasonably say that there is a difference in success rates between the three groups.

In this example we were just examining the distribution of success rates across a single variable (experience group). There are some situations in which you might want to examine more than one variable, such as experience group and design prototype. Performing this type of evaluation works the same way. Figure 2.10

FIGURE 2.9

Output from a chi-square test in Excel.

FIGURE 2.10

Output from a chi-square test with two variables.

shows data based on two different variables: group and design. For a more detailed example of using the chi-square to test for differences in live website data for two alternative pages (so-called A/B tests), see section 9.1.4.

2.8 PRESENTING YOUR DATA GRAPHICALLY

You might have collected and analyzed the best set of usability data ever, but it's of little value if you can't communicate it effectively to others. Data tables are certainly useful in some situations, but in most cases you'll want to present your data graphically. A number of excellent books on the design of effective data

---

GENERAL TIPS FOR DATA GRAPHS

Label the axes and units. It might be obvious to you that a scale of 0 to 100 percent represents the task completion rate, but it may not be obvious to your audience. Or you might know that the times being plotted on a graph are minutes, but your audience may be left pondering whether they could be seconds or even hours. Sometimes the labels on an axis make it clear what the scale is (e.g., "Task 1," "Task 2," etc.), in which case adding a label for the axis itself (e.g., "Tasks") would be redundant.

Don't imply more precision in your data than they deserve. Labeling your time data with "0.00" seconds to "30.00" seconds is almost never appropriate, nor is labeling your task completion data with "0.0%" to "100.0%." Whole numbers work best in most cases. Exceptions include some metrics with a very limited range and some statistics that are almost always fractional (e.g., correlation coefficients).

Don't use color alone to convey information. Of course, this is a good general principle for the design of any information display, but it's worth repeating. Color is commonly used in data graphs, but make sure it's supplemented by positional information, labels, or other cues that help someone who can't clearly distinguish colors to interpret the graph.

Display labels horizontally whenever possible. When you try to squeeze too many items onto the horizontal axis, you may be tempted to display the labels vertically. But you don't want to give your audience a sore neck from constantly tilting their heads to read the axis! An exception is that the main title for the vertical axis is normally shown vertically.

Show confidence intervals whenever possible. This mainly applies to bar graphs and line graphs that are presenting means of individual participant data (times, ratings, etc.). Showing the 95 or 90 percent confidence intervals for the means via error bars is a good way to visually represent the variability in the data.

Don't overload your graphs. Just because you *can* create a single graph that shows the task completion rate, error rate, task times, and subjective ratings for each of 20 tasks, broken down by novice versus experienced participants, doesn't mean you *should*.

Be careful with 3D graphs. If you're tempted to use a 3D graph, ask yourself whether it really helps. In some cases, the use of 3D makes it harder to see the values being plotted.

graphs are available, including those written by Edward Tufte (1990, 1997, 2001, 2006), Stephen Few (2004, 2006), and Robert Harris (1999). Our intent in this section is simply to introduce some of the most important principles in the design of data graphs, particularly as they relate to usability data.

We've organized this section around tips and techniques for five basic types of data graphs:

- Column or bar graphs
- Line graphs
- Scatterplots
- Pie charts
- Stacked bar graphs

We will begin each of the following sections with one good example and one bad example of that particular type of data graph.

2.8.1 Column or Bar Graphs

Column graphs and bar graphs (Figure 2.11) are the same thing; the only difference is their orientation. Technically, column graphs are vertical and bar graphs are horizontal. In practice, most people refer to both types simply as bar graphs, which is what we will do.

Bar graphs are probably the most common way of displaying usability data. Almost every presentation of data from a usability test that we've seen has included at least one bar graph, whether it was for task completion rates, task times, self-reported data, or something else. The following are some of the principles for using bar graphs:

- Bar graphs are appropriate when you want to present the values of continuous data (e.g., times, percentages, etc.) for discrete items or categories (e.g., tasks, participants, designs, etc.). If both variables are continuous, a line graph is appropriate.

- The axis for the continuous variable (the vertical axis in Figure 2.11) should normally start at 0. The whole idea behind bar graphs is that the lengths of the bars represent the values being plotted. By not starting the axis at 0, you're artificially manipulating their lengths. The bad example in Figure 2.11 gives the impression that there's a larger difference between the tasks than there really is. A possible exception is when you include error bars, making it clear which differences are real and which are not.

- Don't let the axis for the continuous variable go any higher than the maximum value that's theoretically possible. For example, if you're plotting percentages of users who successfully completed each task, the theoretical maximum is 100 percent. If some values are close to that maximum, Excel and other packages will tend to automatically increase the scale beyond the maximum, especially if error bars are shown.
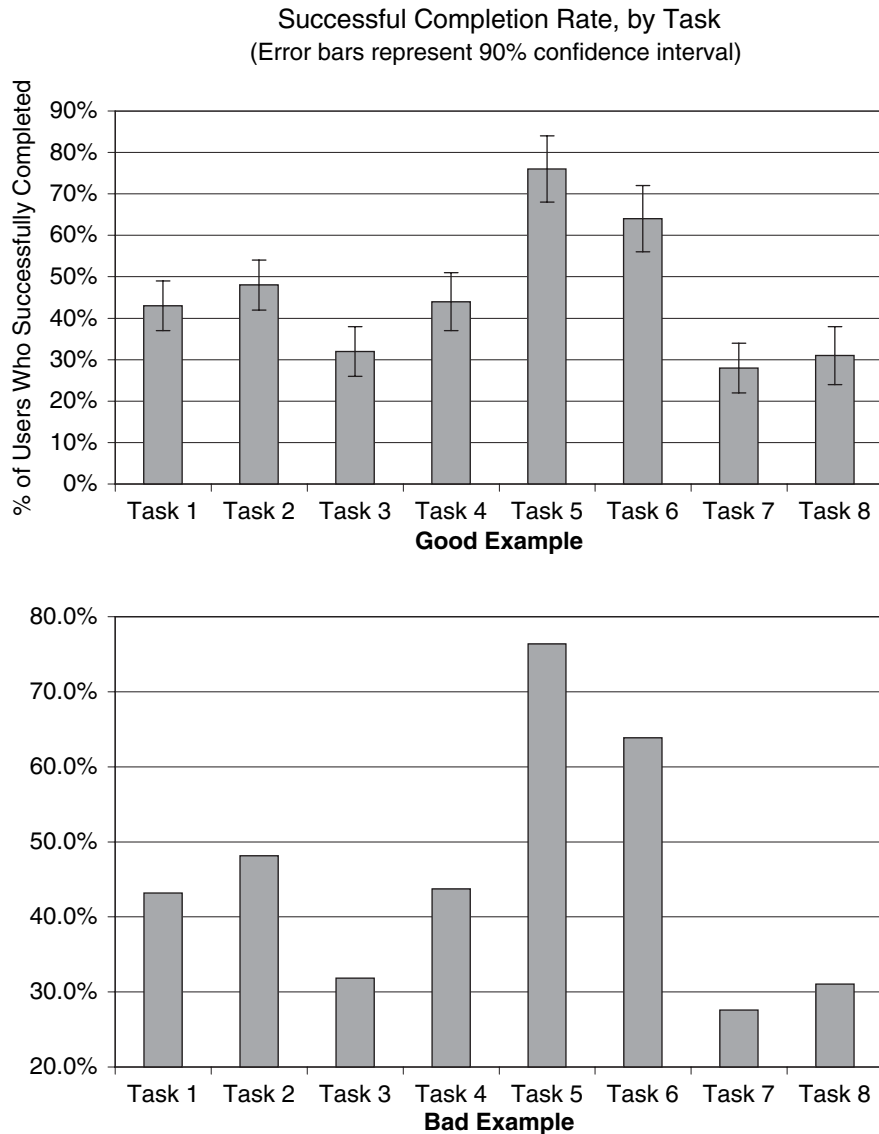
FIGURE 2.11

Good and bad examples of bar graphs for the same data. The mistakes in the bad version (*bottom*) include failing to label the data, not starting the vertical axis at 0, not showing confidence intervals when you can, and showing too much precision in the vertical axis labels.

LINE GRAPHS VERSUS BAR GRAPHS

Some people have a hard time deciding whether it's appropriate to use a line graph or a bar graph to display a set of data. Perhaps the most common data-graph mistake we see is using a line graph when a bar graph is more appropriate. If you're considering presenting some data with a line graph, ask yourself a simple question: Do the places along the line *between* the data points make sense? In other words, even though you don't have data for those locations, would they make sense if you did? If they don't make sense, a bar graph is more appropriate.

For example, it's technically possible to show the data in Figure 2.11 as a line graph, as follows.

Successful Completion Rate
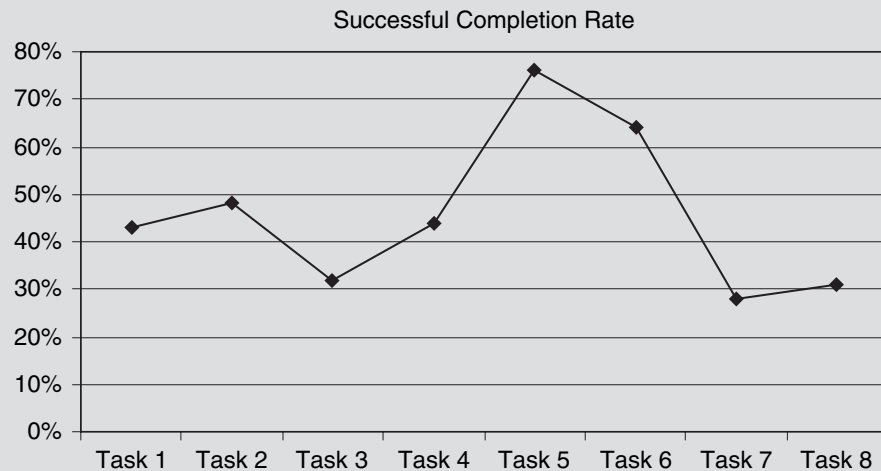
However, you should ask yourself whether things like "Task 1½" or "Task 6¾" make any sense, because the lines imply that they should. Obviously, they don't, so a bar graph is the correct representation. The line graph might make an interesting picture, but it's a misleading picture.

2.8.2 Line Graphs

Line graphs (Figure 2.12) are most commonly used to show trends in continuous variables, often over time. Although not as common as bar graphs in presenting usability data, they certainly have their place. The following are some of the key principles for using line graphs:

■ Line graphs are appropriate when you want to present the values of one continuous variable (e.g., percent correct, number of errors, etc.) as a function of another
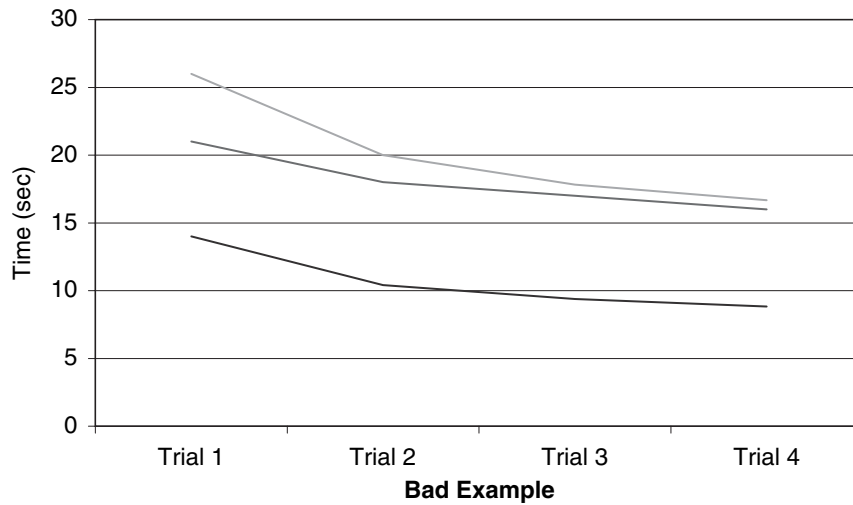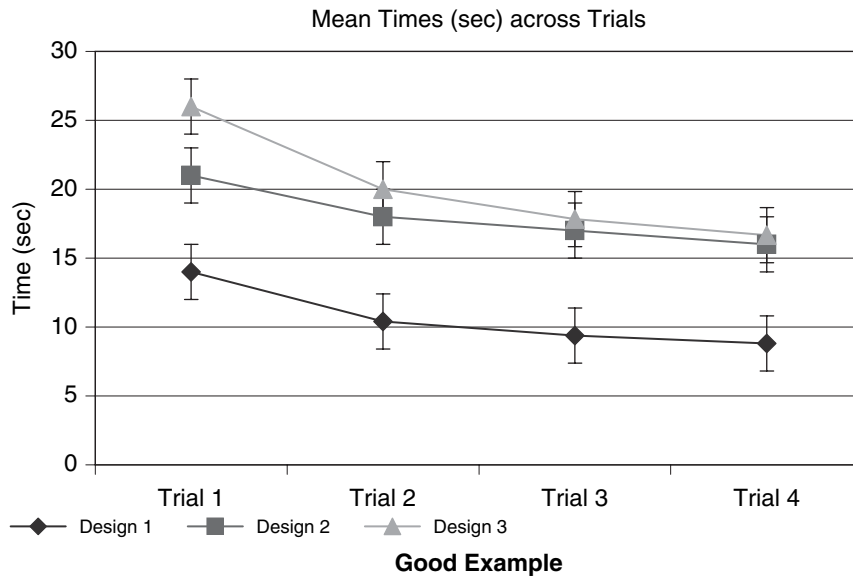
FIGURE 2.12

Good and bad examples of line graphs for the same data. The mistakes in the bad version (*bottom*) include failing to label the vertical axis, not showing the data points, not including a legend, not showing confidence intervals, and using lines that are too lightweight.

continuous variable (e.g., age, trial, etc.). If one of the variables is discrete (e.g., gender, participant, task, etc.), then a bar graph is more appropriate.

■ Show your data points. Your actual data points are the things that really matter, not the lines. The lines are just there to connect the data points and to make the trends more obvious. You may need to increase the default size of the data points in Excel.

■ Use lines that have sufficient weight to be clear. Very thin lines are not only hard to see, but it's harder to detect their color and they may imply a greater precision in the data than is appropriate. You may need to increase the default weight of lines in Excel.

■ Include a legend if you have more than one line. In some cases, it may be clearer to manually move the labels from the legend into the body of the graph and put each label beside its appropriate line. It may be necessary to do this in Power-Point or some other drawing program.

■ As with bar graphs, the vertical axis normally starts at 0, but it's not as important with a line graph to always do that. There are no bars whose length is important, so sometimes it may be appropriate to start the vertical axis at a higher value. In that case, you should mark the vertical axis appropriately. The traditional way of doing this is with a "discontinuity" marker on that axis: ⧸. Again, it may be necessary to do that in a drawing program.

2.8.3 Scatterplots

Scatterplots (Figure 2.13), or X/Y plots, show pairs of values. Although they're not very common in usability reports, they can be very useful in certain situations. Here are some of the key principles for using scatterplots:

■ You must have paired values that you want to plot. A classic example is heights and weights of a group of people. Each person would appear as a data point, and the two axes would be height and weight.

■ Normally, both of the variables would be continuous. In Figure 2.13, the vertical axis shows mean values for a visual appeal rating of 42 web pages (from Tullis & Tullis, 2007). Although that scale originally had only four values, the means come close to being continuous. The horizontal axis shows the size, in k pixels, of the largest nontext image on the page, which truly is continuous.

■ You should use appropriate scales. In Figure 2.13, the values on the vertical axis can't be any lower than 1.0, so it's appropriate to start the scale at that point rather than 0.

■ Your purpose in showing a scatterplot is usually to illustrate a relationship between the two variables. Consequently, it's usually helpful to add a trend line to the scatterplot, as in the good example in Figure 2.13. You may want to include the R^2 value to indicate the goodness of fit.
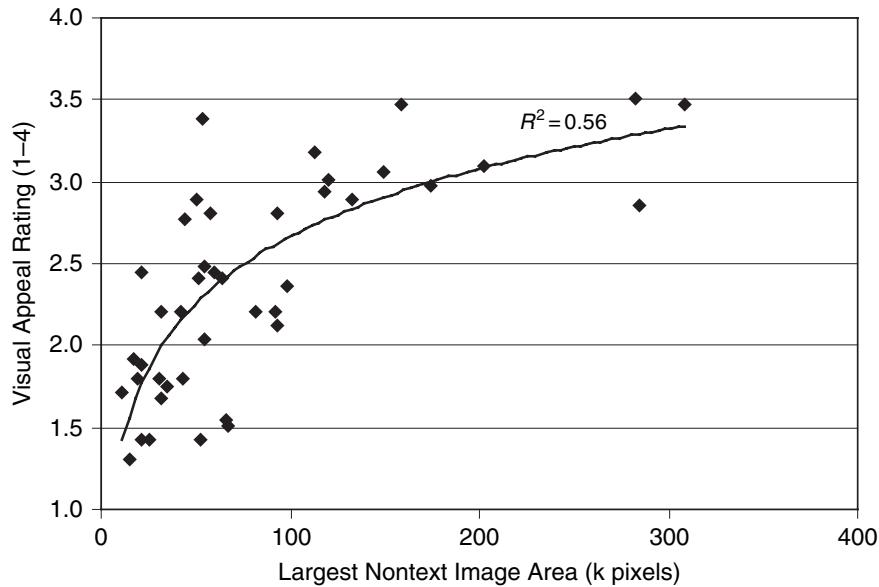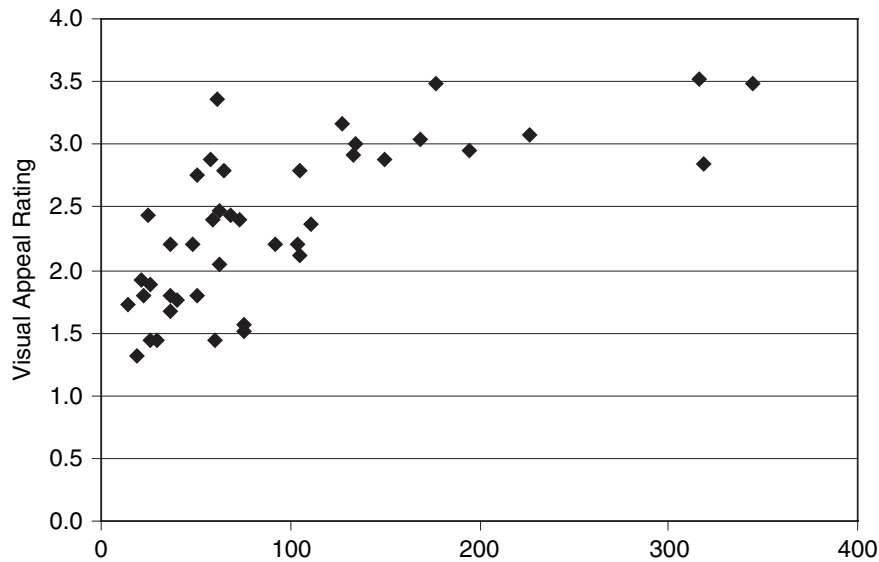
FIGURE 2.13

Good and bad examples of scatterplots for the same data. The mistakes in the bad version include an inappropriate scale for the vertical axis, not showing the scale for the visual appeal ratings (1–4), not showing a trend line, and not showing the goodness of fit (R^2).

2.8.4 Pie Charts

Pie charts (Figure 2.14) illustrate the parts or percentages of a whole. They can be useful any time you want to illustrate the relative proportions of the parts of a whole to each other (e.g., how many participants in a usability test succeeded, failed, or gave up on a task). Here are some key principles for their use:

- Pie charts are appropriate only when the parts add up to 100 percent. You have to account for all cases. In some situations, this might mean creating an ''other'' category.

- Minimize the number of segments in the pie chart. Even though the bad example in Figure 2.14 is technically correct, it's almost impossible to make any sense of it because it has so many segments. Try to use no more than six segments. Logically combine segments, as in the good example, to make the results clearer.

- In almost all cases, you should include the percentage and label for each segment. Normally these should be next to each segment, connected by leader lines if necessary. Sometimes you have to manually move the labels to prevent them from overlapping.
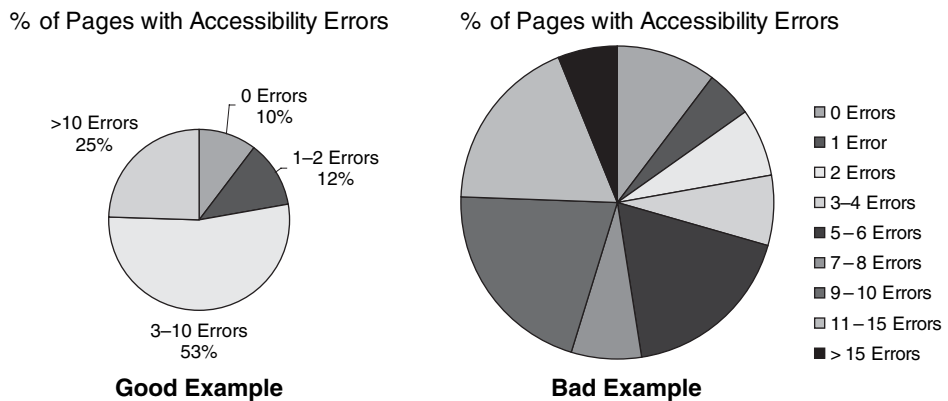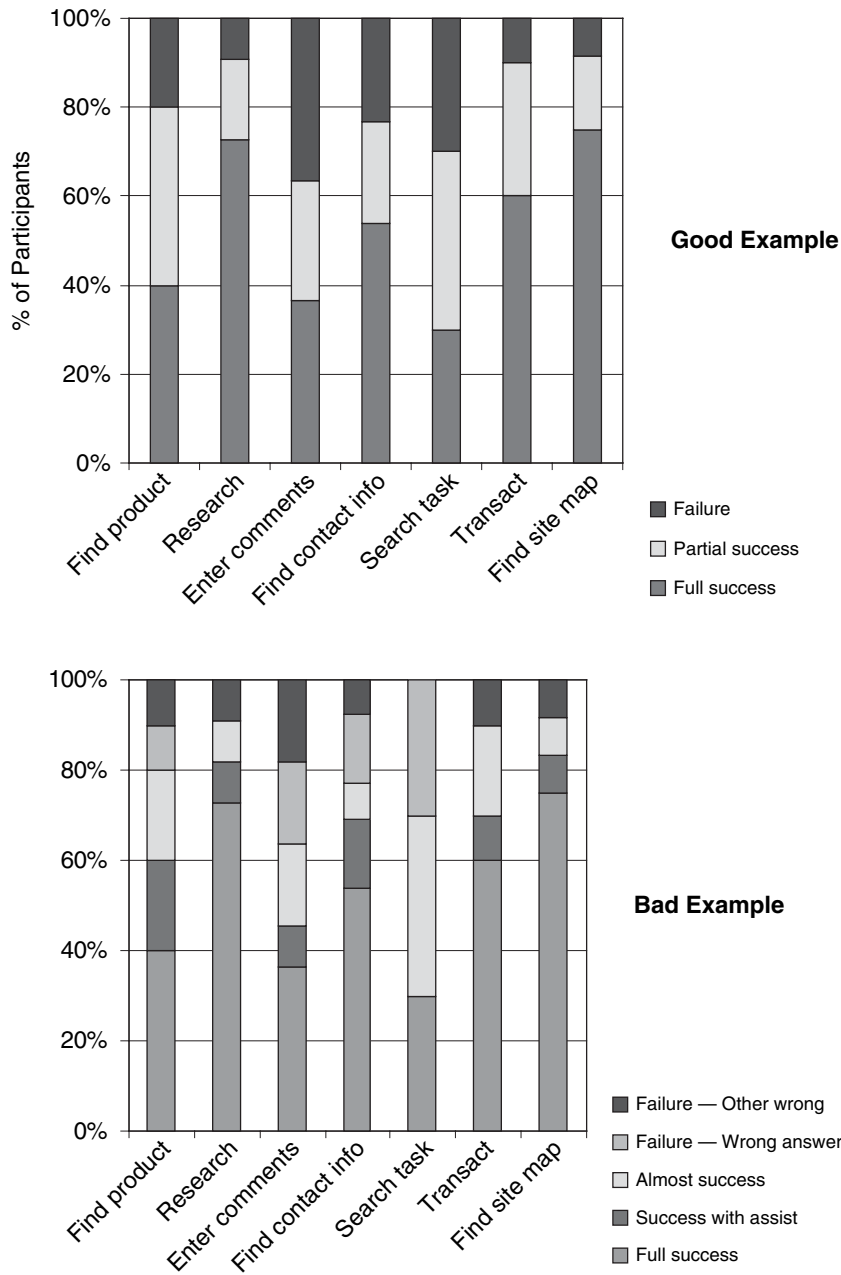

FIGURE 2.14

Good and bad examples of pie charts for the same data. The mistakes in the bad version (*right*) include too many segments, poor placement of the legend, and not showing percentages for each segment.

2.8.5 Stacked Bar Graphs

Stacked bar graphs (Figure 2.15) are basically multiple pie charts shown in bar form. They're appropriate whenever you have a series of datasets, each of which represents parts of the whole. Their most common use in usability data is to show different task completion states for each task. Here are some key principles for their use:

- Like pie charts, stacked bar graphs are only appropriate when the parts for each item in the series add up to 100 percent.

FIGURE 2.15

Good and bad examples of stacked bar graphs for the same data. The mistakes in the bad version (*bottom*) include too many segments, poor color coding, and failing to label the vertical axis.

- The items in the series are normally categorical (e.g., tasks, participants, etc.).

- Minimize the number of segments in each bar. More than three segments per bar can make it difficult to interpret. Combine segments as appropriate.

- When possible, make use of color coding conventions that your audience is likely to be familiar with. For many U.S. audiences, green is good, yellow is marginal, and red is bad. Playing off these conventions can be helpful but don't rely solely on them.

2.9 SUMMARY

In a nutshell, this chapter is about knowing your data. The better you know your data, the more likely you are to clearly answer your research questions. The following are some of the key takeaways from this chapter.

1. It's important to consider questions about how to select participants for your study, how to order tasks, what participants perform what tasks, and how many participants you need to get reasonably reliable feedback.

2. Knowing your data is critical when analyzing your results. The specific type of data you have will dictate what statistics you can (and can't) perform.

3. Nominal data are categorical, such as binary task success or males and females. Nominal data are usually expressed as frequencies or percentages. Chi-square tests can be used when you want to learn whether the frequency distribution is random or there is some underlying significance to the distribution pattern.

4. Ordinal data are rank orders, such as a severity ranking of usability issues. Ordinal data are also analyzed using frequencies, and the distribution patterns can be analyzed with a chi-square test.

5. Interval data are continuous data where the intervals between each point are meaningful but without a natural zero. The SUS score is one example. Interval data can be described by means, standard deviations, and confidence intervals. Means can be compared to each other for the same set of users (paired samples *t*-test) or across different users (independent samples *t*-test). ANOVA can be used to compare more than two sets of data. Relationships between variables can be examined through correlations.

6. Ratio data are the same as interval data but with a natural zero. One example is completion times. Essentially, the same statistics that apply to interval data also apply to ratio data.

7. When presenting your data graphically, use the appropriate types of graphs. Use bar graphs for categorical data and line graphs for continuous data. Use pie charts or stacked bar graphs when the data sum to 100 percent.