Second Edition

Measuring the User Experience

Collecting, Analyzing, and Presenting Usability Metrics

MK
MORGAN KAUFMANN

TOM TULLIS • BILL ALBERT

CHAPTER 2

Background

CONTENTS

This chapter covers background information about data, statistics, and graphs that apply to just about any user experience metrics. Specifically, we address the following:

- The basic *types of variables and data* in any user experience study, including independent and dependent variables, and nominal, ordinal, interval, and ratio data.

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00002-9

- Basic *descriptive statistics* such as the mean and median, standard deviation, and the concept of *confidence intervals*, which reflect how accurate your estimates of measures such as task times, task success rates, and subjective ratings actually are.
- Simple *statistical tests* for comparing means and analyzing relationships between variables.
- Tips for *presenting your data visually* in the most effective way.

We use Microsoft Excel 2010 for all of the examples in this chapter (and really in most of this book) because it is so popular and widely available. Most of the analyses can also be done with other readily available spreadsheet tools such as Google Docs or OpenOffice.org.

2.1 INDEPENDENT AND DEPENDENT VARIABLES

At the broadest level, there are two types of variables in any usability study: independent and dependent. Independent variables are the things you manipulate or control for, such as designs you're testing or the ages of your participants. Dependent variables are the things you measure, such as success rates, number of errors, user satisfaction, completion times, and many more. Most of the metrics discussed in this book are dependent variables.

When designing a user experience study, you should have a clear idea of what you plan to manipulate (independent variables) and what you plan to measure (dependent variables). The most interesting outcomes of a study are at the intersection of the independent and dependent variables, such as whether one design resulted in a higher task success rate than the other.

2.2 TYPES OF DATA

Both independent and dependent variables can be measured using one of four general types of data: nominal, ordinal, interval, and ratio. Each type of data has its own unique characteristics and, most importantly, supports specific types of analyses and statistics. When collecting and analyzing user experience 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 (also called categorical) 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 user experience, nominal data might be characteristics of different types of users, such as Windows versus Mac users, users in different geographic locations, or males vs females. These are typically independent variables that allow you to segment data by these different groups. Nominal data also include some

commonly used dependent variables, such as task success, the number of users who clicked on link A instead of link B, or users 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% of the users are female, there are 200 users with blue eyes, or 95% were successful on a particular task.

CODING NOMINAL DATA

One important thing to consider when working with nominal data is how to code it. In analyzing nominal data, it's not uncommon to represent the membership in each group using numbers. For example, you might code males as group "1" and females as group "2." But 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 you're using for your analysis 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 a failure as "0," the average will represent the proportion of users who were successful.

2.2.2 Ordinal Data

Ordinal data are ordered groups or categories. As the name implies, data are organized in a certain way. However, the intervals between measurements are not meaningful. Some people think of ordinal data as ranked data. For example, the list of the top 100 movies, as rated by the American Film Institute (AFI), shows that their 10th best movie of all time, *Singing in the Rain*, is better than their 20th 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 user experience, the most common examples of ordinal data come from self-reported data. For example, a user 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. Or if you were to ask the participants in a usability study to rank order four different designs for a web page according to which they prefer, that would also be ordinal data. There's no reason to assume that the distance between the page ranked first by a participant and the page ranked second is the same as the distance between the

page ranked second and the one ranked third. It could be that the participant really loved one page and hated all three of the others.

The most common way to analyze ordinal data is by looking at frequencies. For example, you might report that 40% of the users rated the site as excellent, 30% as good, 20% percent as fair, and 10% 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 differences between the values are meaningful, but there is no natural zero point. An example of interval data familiar to most of us is temperature. Defining 0° Celsius or 32° Fahrenheit based on when water freezes is completely arbitrary. The freezing point of water does not mean the absence of heat; it only identifies a meaningful point on the scale of temperatures. But the differences between the values are meaningful: the distance from 10° to 20° is the same as the distance from 20° to 30° (using either scale). Dates are another common example of interval data.

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 and standard deviation). 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 if you can treat it as being interval. Consider these two rating scales:

o Poor o Fair o Good o Excellent

Poor o o o o Excellent

At first glance, you might say those two scales are the same, but the difference in presentation makes them different. Putting explicit labels on 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," which is why 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 o o o o o o o o o Excellent

Presenting it that way, with 9 points along the scale, makes it even more obvious that the data can be treated as if it 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 this 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 but with the addition of an absolute zero. This means that the zero value is not arbitrary, as with interval data, but has some inherent meaning. With ratio data, 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 user experience, the most obvious example of ratio data is time. Zero seconds left to complete a task 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 user 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. Aside from that calculation, there really aren't many differences between interval and ratio data in terms of the available statistics.

2.3 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.

The most common types of descriptive statistics are measures of central tendency (such as the mean), measures of variability (such as the standard deviation), and confidence intervals, which pull the other two together. The following sections use the sample data shown in Table 2.1 to illustrate these statistics. These data represent the time, in seconds, that it took each of 12 participants in a usability study to complete the same task.

2.3.1 Measures of Central Tendency

Measures of central tendency are simply a way of choosing a single number that is in some way representative of a set of numbers. The three most common measures of central tendency are the mean, median, and mode.

The mean is what most people think of as the average: the sum of all values divided by how many values there are. The mean of most user experience metrics is extremely useful and is probably the most common statistic cited in a usability report. For the data in Table 2.1, the mean is 35.1 seconds.

Participant	Task Time (seconds)
P1	34
P2	33
P3	28
P4	44
P5	46
P6	21
P7	22
P8	53
P9	22
P10	29
P11	39
P12	50

Table 2.1 Time to complete a task, in seconds, for each of 12 participants in a usability study.

EXCEL TIP

The mean of any set of numbers in Excel can be calculated using the "=AVERAGE" function. The median can be calculated using the "=MEDIAN" function, and the mode can be calculated using the "=MODE" function. If the mode can't be calculated (which happens when each value occurs an equal number of times), Excel returns "#N/A".

The median is the middle number if you put them in order from smallest to largest: half the values are below the median and half are above the median. If there is no middle number, the median is halfway between the two values on either side of the middle. For the data in Table 2.1, the median is equal to 33.5 seconds (halfway between the middle two numbers, 33 and 34). Half of the users were faster than 33.5 seconds and half were slower. In some cases, the median can be more revealing than the mean. For example, let's assume the task time for P12 had been 150 seconds rather than 50.That would change the mean to 43.4 seconds, but the median would be unchanged at 33.5 seconds. It's up to you to decide which is a more representative number, but this illustrates the reason that the median is sometimes used, especially when larger values (or so-called "outliers") may skew the distribution.

The mode is the most commonly occurring value in the set of numbers. For the data in Table 2.1, the mode is 22 seconds, because two participants completed the task in 22 seconds. It's not common to report the mode in usability test results. When data are continuous over a broad range, such as the task times shown in Table 2.1, 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.

NUMBER OF DECIMAL PLACES TO USE WHEN REPORTING DATA

One of the most common mistakes many people make is reporting data from a usability test (mean times, task completion rates, etc.) with more precision than it really deserves. For example, the mean of the times in Table 2.1 is technically 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 accuracy of the original data, its magnitude, and its variability. The original data in Table 2.1 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.

2.3.2 Measures of Variability

Measures of variability reflect how much the data are spread or dispersed across the range of values. For example, these measures help answer the question, "Do most users have similar task completion times or is there a wide range of times?" In most usability studies, variability is caused by individual differences among your participants. There are three common measures of variability: range, variance, and standard deviation.

The range is the distance between the minimum and maximum values. For the data in Table 2.1, 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 useful because it will help 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.

EXCEL TIP

The minimum of any set of numbers in Excel can be determined using the "=MIN" function and the maximum using the "=MAX" function. The range can then be determined by MAX-MIN. The variance can be calculated using the "=VAR" function and the standard deviation using the "=STDEV" function.

Variance 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. For the data in Table 2.1, the variance is 126.4.

Once you know the variance, you can calculate the standard deviation easily, which is the most commonly used measure of variability. The standard deviation is simply the square root of the variance. The standard deviation of the data shown in Table 2.1 is 11.2 seconds. Interpreting the standard deviation is a little easier than interpreting the variance, as the unit of the standard deviation is the same as the original data (seconds, in this example).

EXCEL TIP: DESCRIPTIVE STATISTICS TOOL

An experienced Excel user might be wondering why we didn't just suggest using the "Descriptive Statistics" tool in the Excel Data Analysis ToolPak. (You can add the Data Analysis ToolPak using "Excel Options">"Add-Ins".) This tool will calculate the mean, median, range, standard deviation, variance, and other statistics for any set of data you specify. It's a very handy tool. However, it has what we consider a significant limitation: the values it calculates are static. If you go back and update the original data, the statistics don't update. We like to set up our spreadsheets for analyzing the data from a usability study before we actually collect the data. Then we update the spreadsheet as we're collecting the data. This means we need to use formulas that update automatically, such as MEAN, MEDIAN, and STDEV, instead of the "Descriptive Statistics" tool. But it can be a useful tool for calculating a whole batch of these statistics at once. Just be aware that it won't update if you change the data.

2.3.3 Confidence Intervals

A confidence interval is an estimate of a range of values that includes the true population value for a statistic, such as a mean. For example, assume that you need to estimate the true population mean for a task time whose sample times are shown in Table 2.1. You could construct a confidence interval around that mean to show the range of values that you are reasonably certain will include the true population mean. The phrase "reasonably certain" indicates that you will need to choose how certain you want to be or, put another way, how willing you are to be wrong in your assessment. This is what's called the confidence level that you choose or, conversely, the alpha level for the error that you're willing to accept. For example, a confidence level of 95%, or an alpha level of 5%, means that you want to be 95% certain, or that you're willing to be wrong 5% of the time.

There are three variables that determine the confidence interval for a mean:

- The sample size, or the number of values in the sample. For the data in Table 2.1, the sample size is 12, as we have data from 12 participants.
- The standard deviation of the sample data. For our example, that is 11.2 seconds.

- The alpha level we want to adopt. The most common alpha levels (primarily by convention) are 5 and 10%. Let's choose an alpha of 5% for this example, which is a 95% confidence interval.

The 95% confidence interval is then calculated using the following formula:

$$\text{Mean} \pm 1.96 * [\text{standard deviation}/\text{sqrt}(\text{sample size})]$$

The value "1.96" is a factor that reflects the 95% confidence level. Other confidence levels have other factors. This formula shows that the confidence interval will get smaller as the standard deviation (the variability of data) decreases or as the sample size (number of participants) increases.

EXCEL TIP

You can calculate the confidence interval quickly for any set of data using the CONFIDENCE function in Excel. The formula is easy to construct:

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

Alpha is your significance level, which is typically 5% (0.05) or 10% (0.10). The standard deviation can be calculated using the STDEV function. The sample size is simply the number of cases or data points you are examining, which can be calculated using the COUNT function. Figure 2.1 shows an example. For the data in Table 2.1, the result of this calculation is 6.4 seconds. Since the mean is 35.1 seconds, the 95% confidence interval for that mean is 35.1±6.4, or 28.7 to 41.5 seconds. So you can be 95% certain that the true population mean for this task time is between 28.7 and 41.5 seconds.

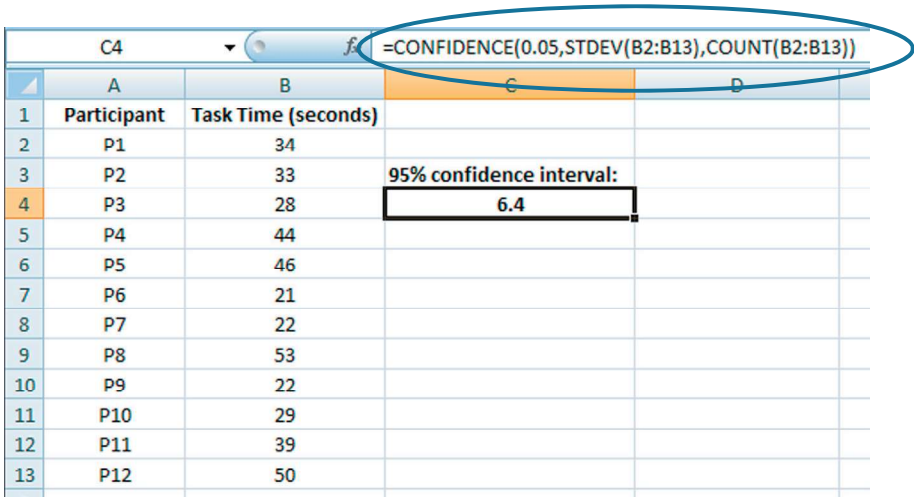


Figure 2.1 Example of how to calculate a 95% confidence interval using the "confidence" function in Excel.

Confidence intervals are incredibly useful. We think you should calculate and display them routinely for just about any means that you report from a usability study. When displayed as error bars on a graph of means, they make it visually obvious how accurate the measures actually are.

WHAT CONFIDENCE LEVEL SHOULD YOU USE?

How should you decide what confidence level to use? Traditionally, the three commonly used confidence levels are 99, 95, and 90% (or their corresponding alpha levels of 1, 5, and 10%). The history behind the use of these three levels goes back to the days before computers and calculators, when you had to look up values for confidence levels in printed tables. The people printing these tables didn't want to print a bunch of different versions, so they made just these three. Today, of course, all of these calculations are done for us so we could choose any confidence level we want. But because of their longstanding use, most people choose one of these three.

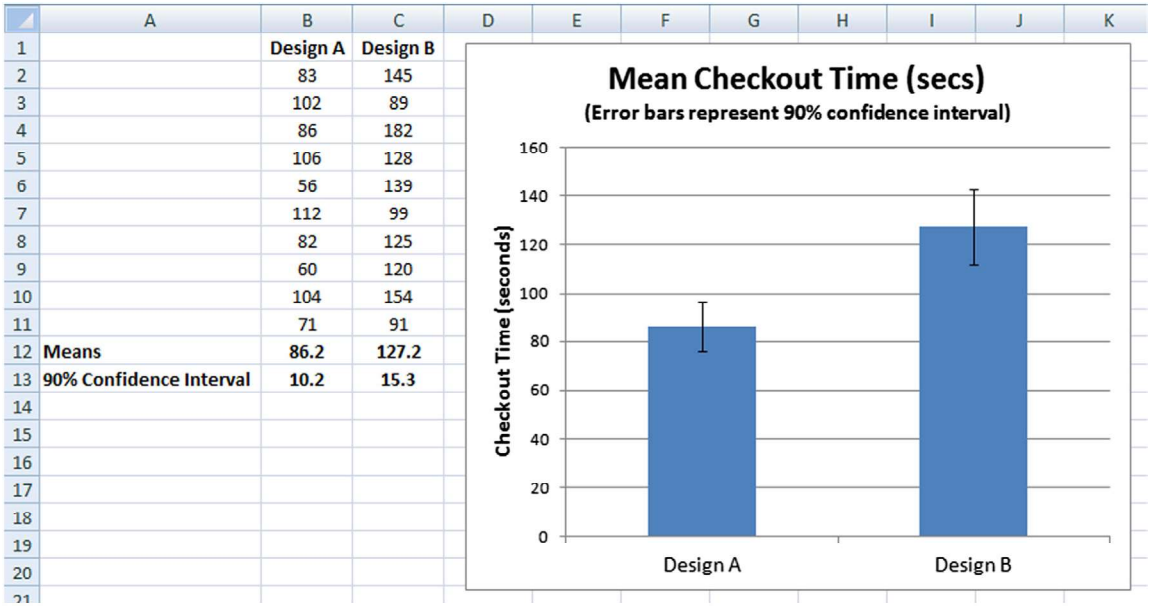
The level you choose really does depend on how certain you need to be that the confidence interval contains the true mean. If you're trying to estimate how long it will take someone to administer a life-saving shock using an automated external defibrillator, you probably want to be very certain of your answer, and would likely choose at least 99%. But if you're simply estimating how long it will take someone to upload a new photo to their Facebook page, you probably would be satisfied with a 90% confidence level. In our day-to-day use of confidence intervals, we find that we use a 90% confidence level most of the time, sometimes a 95% level, and rarely a 99% level.

2.3.4 Displaying Confidence Intervals as Error Bars

Let's now consider the data in Figure 2.2, which shows the checkout times for two different designs of a prototype website. In this study, 10 participants performed the checkout task using Design A and another 10 participants performed the checkout task using Design B. Participants were assigned randomly to one group or the other. The means and 90% confidence interval for both groups have been calculated using the AVERAGE and CONFIDENCE functions. The means have been plotted as a bar graph, and the confidence intervals are shown as error bars on the graph. Even just a quick glance at this bar graph shows that the error bars for these two means don't overlap with each other. When that is the case, you can safely assume that the two means are significantly different from each other.

EXCEL TIP

Once you've created a bar graph showing the means, such as Figure 2.2, you then want to add error bars to represent the confidence intervals. First, click on the chart to select it. Then, in the Excel button bar, choose the "Layout" tab under "Chart Tools." On the "Layout" tab, choose "Error Bars>More Error Bars Options." In the resulting dialog box, select the "Custom" option near the bottom of the dialog box. Then click on the



	A	B	C
1		Design A	Design B
2		83	145
3		102	89
4		86	182
5		106	128
6		56	139
7		112	99
8		82	125
9		60	120
10		104	154
11		71	91
12	Means	86.2	127.2
13	90% Confidence Interval	10.2	15.3

Figure 2.2 Illustration of displaying confidence intervals as error bars on bar graph.

"Specify Value" button. The resulting small window allows you to specify the values for the positive and negative portions of the error bars, which will both be the same. Click on the button to specify the Positive Error Value and then select **both** of the values for the 90% confidence interval on the spreadsheet (cells B13 and C13 in Figure 2.2). Then click on the button for the Negative Error Value and select the exact same cells again. Close both windows and your error bars should be on the graph.

2.4 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 ratings than another or if the number of errors is higher for one group of users compared to another, your best approach is through statistics.

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

1. Is the comparison *within* the same set of users or *across* different users? For example, if you are comparing some data for men vs women, it is highly likely that these are different users. When comparing different samples like this, it's called independent samples. But if you're comparing the *same* group of users on different products or designs, you will use something called paired samples.

2. 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).

2.4.1 Independent Samples

Perhaps the simplest way to compare means from independent samples is using confidence intervals, as shown in the previous section. In comparing the confidence intervals for two means, you can draw the following conclusions:

- If the confidence intervals *don't overlap*, you can safely assume the two means are significantly different from each other (at the confidence level you chose).
- If the confidence intervals *overlap slightly*, the two means might still be significantly different. Run a *t* test to determine if they are different.
- If the confidence intervals *overlap widely*, the two means are not significantly different.

Let's consider the data in Figure 2.3 to illustrate running a *t* test for independent samples. This shows the ratings of ease of use on a 1 to 5 scale for two different designs as rated by two different groups of participants (who were assigned randomly to one group or the other). We've calculated the means and confidence intervals and graphed those. But note that the two confidence intervals overlap slightly: Design 1's interval goes up to 3.8, whereas Design 2's goes down to 3.5. This is a case where you should run a *t* test to determine if the two means are significantly different.

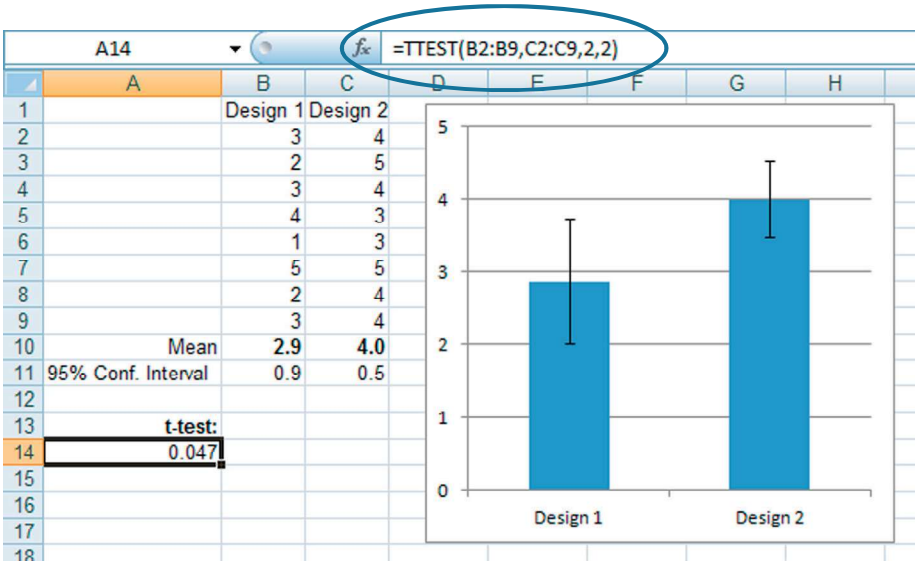


Figure 2.3 Example of a *t* test on independent samples.

EXCEL TIP

As illustrated in Figure 2.3, you can use the TTEST function in Excel to run a *t* test:

=TTEST(Array 1, Array 2, Tails, Type)

Array 1 and Array 2 refer to the sets of values that you want to compare. In Figure 2.3, Array 1 is the set of ratings for Design 1 and Array 2 is the set of ratings for Design 2. Tails refer to whether your test is one-tailed or two-tailed. This relates to the tails (extremes) of the normal distribution and whether you're considering one end or both ends. From a practical standpoint, this is asking whether it is theoretically possible for the difference between these two means to be in either direction (i.e., Design 1 *either* higher or lower than Design 2). In almost all cases that we deal with, the difference could be in either direction, so the correct choice is "2" for two tailed. Finally, Type indicates the type of *t* test. For these independent samples (not paired), the Type is 2.

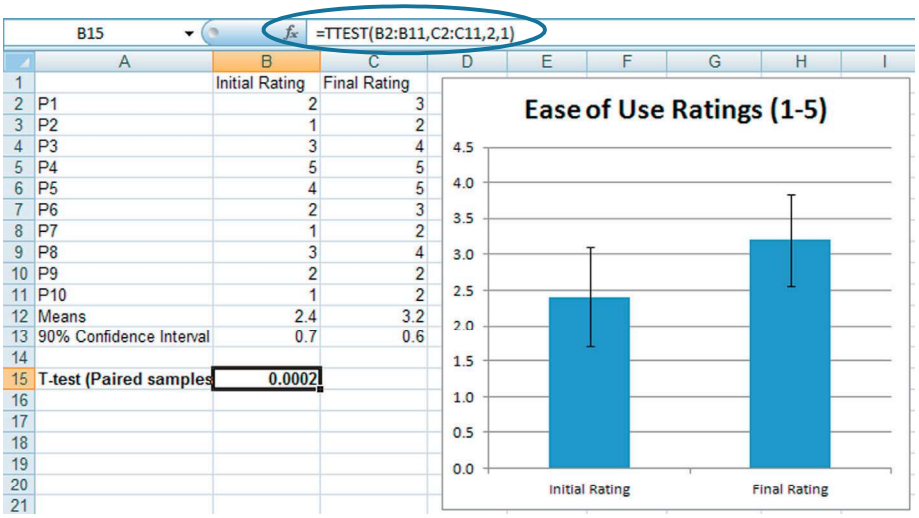
This *t* test returns a value of 0.047. So how do you interpret that? It's telling you the probability that the difference between these two means is simply due to chance. So there's a 4.7% chance that this difference is *not* significant. Since we were dealing with a 95% confidence interval, or a 5% alpha level, and this result is less than 5%, we can say that the difference is statistically significant at that level.

2.4.2 Paired Samples

A paired samples *t* test is used when comparing means within the same set of users. For example, you may be interested in knowing whether there is a difference between two prototype designs. If you have the same set of users 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.

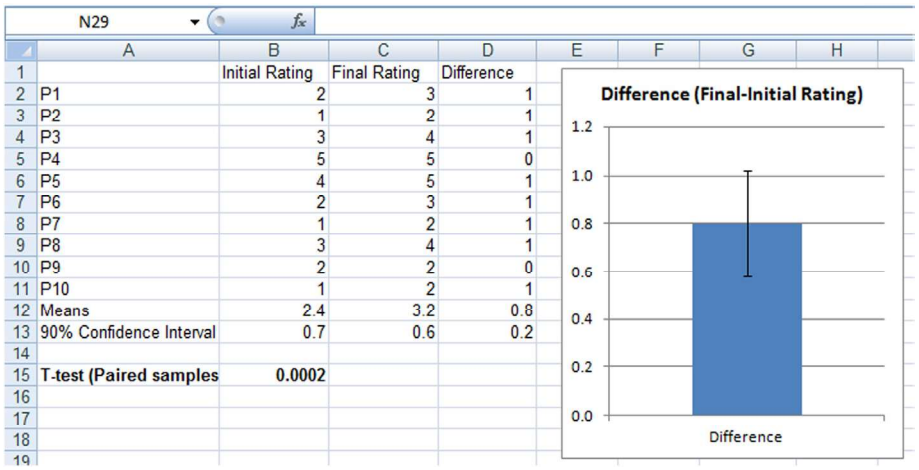
With paired samples like these, the key is that you're comparing each person to themselves. Technically, you're looking at the difference in each person's data for the two conditions you're comparing. Let's consider the data shown in Figure 2.4, which shows "Ease of Use" ratings for an application after their initial use and then again at the end of the session. So there were 10 participants who gave two ratings each. The means and 90% confidence intervals are shown and have been graphed. Note that the confidence intervals overlap pretty widely. If these were independent samples, you could conclude that the ratings are not significantly different from each other. However, because these are paired samples, we've done a *t* test on paired samples (with the "Type" as "1"). That result, 0.0002, shows that the difference is highly significant.

Let's look at the data in Figure 2.4 in a slightly different way, as shown in Figure 2.5. This time we've simply added a third column to the data in which the initial rating was subtracted from the final rating for each participant. Note that for 8 of the 10 participants, the rating increased by one point, whereas for 2 participants it stayed the same. The bar graph shows the mean of those differences (0.8) as well as the confidence interval for that mean difference. In a



	A	B	C
1		Initial Rating	Final Rating
2	P1	2	3
3	P2	1	2
4	P3	3	4
5	P4	5	5
6	P5	4	5
7	P6	2	3
8	P7	1	2
9	P8	3	4
10	P9	2	2
11	P10	1	2
12	Means	2.4	3.2
13	90% Confidence Interval	0.7	0.6
14			
15	T-test (Paired samples	0.0002	

Figure 2.4 Data showing paired samples in which each of 10 participants gave an ease of use rating (on a 1–5 scale) to an application after an initial task and at the end of the study.



	A	B	C	D
1		Initial Rating	Final Rating	Difference
2	P1	2	3	1
3	P2	1	2	1
4	P3	3	4	1
5	P4	5	5	0
6	P5	4	5	1
7	P6	2	3	1
8	P7	1	2	1
9	P8	3	4	1
10	P9	2	2	0
11	P10	1	2	1
12	Means	2.4	3.2	0.8
13	90% Confidence Interval	0.7	0.6	0.2
14				
15	T-test (Paired samples	0.0002		

Figure 2.5 Same data as in Figure 2.4, but also showing the difference between initial and final ratings, the mean of those differences, and the 90% confidence interval.

paired-samples test like this, you're basically testing to see if the confidence interval for the mean difference includes 0 or not. If not, the difference is significant.

Note that in a paired samples test, you should have an equal number of values in each of the two sets of numbers that 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.

2.4.3 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 ANOVA lets you determine whether there is a significant difference across more than two groups.

Excel lets you perform three types of ANOVAs. We will give an example for just one type of ANOVA, called a single-factor ANOVA. A single-factor ANOVA is used when you just have one variable you want to examine. For example, you might be interested in comparing task completion times across three different prototypes.

Let's consider the data shown in Figure 2.6, which shows task completion times for three different designs. There were a total of 30 participants in this study, with each using only one of the three designs.

EXCEL TIP

To run an ANOVA in Excel requires the Analysis ToolPak. From the "Data" tab, choose the "Data Analysis" button, which is probably on the far right of the button bar. Then choose "ANOVA: Single Factor." 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 B, C, and D. We have set an alpha level to 0.05 and have included our labels in the first row.

Results are shown in two parts (the right-hand portion of Figure 2.6). The top part is a summary of the data. As you can see, the average time for Design 2 is quite a bit slower, and Designs 1 and 3 completion times are faster. Also, the variance is greater for Design 2 and less for Designs 1 and 3. The second part of the output lets us know whether this difference is significant. The *p* value of 0.000003 reflects the statistical significance of this result. Understanding exactly what this means is important: It means that there is a significant effect of the "designs" variable.

	A	B	C	D	E	F	G	H	I	J	K	L
1		Design 1	Design 2	Design 3		Anova: Single Factor						
2		34	49	22								
3		33	54	28		SUMMARY						
4		28	52	21		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5		44	39	30		Design 1	10	335	33.5	43.16667		
6		21	60	32		Design 2	10	490	49	63.33333		
7		40	58	36		Design 3	10	302	30.2	38.62222		
8		36	49	27								
9		29	34	40								
10		32	46	37		ANOVA						
11		38	49	29		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Means	33.5	49.0	30.2		Between Groups	2015.267	2	1007.633	20.83003	0.000003	3.354131
13	90% Conf Interval	3.4	4.1	3.2		Within Groups	1306.1	27	48.37407			
14												
15						Total	3321.367	29				

Figure 2.6 Task completion times for three different designs (used by different participants) and results of a single-factor ANOVA.

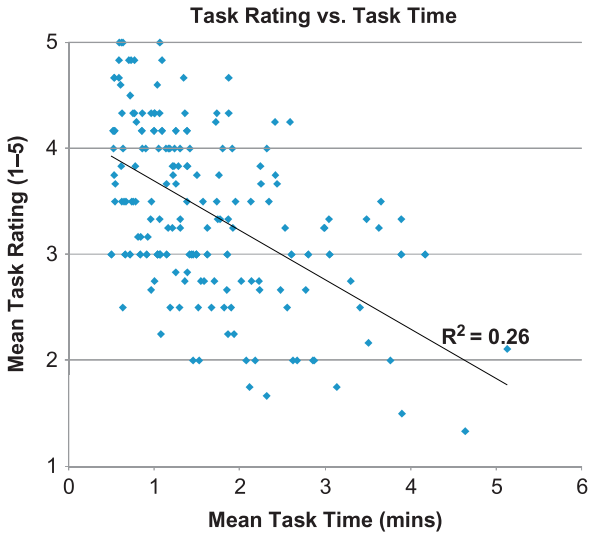


Figure 2.7 An example of a scatterplot (with trend line) in Excel.

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 sample *t* test on just those two sets of values.

2.5 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 users say and what they do don't always correspond with each other. Many users 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. This section provides examples of how to perform analyses that investigate these kinds of relationships (or lack thereof).

2.5.1 Correlations

When you first begin examining the relationship between two variables, it's important to visualize what the data look like. That's easy to do in Excel using a scatterplot. Figure 2.7 is an example of a scatterplot of actual data from an online usability study. The horizontal axis shows mean task time in minutes, and the vertical axis shows mean task rating (1–5, with higher numbers being better). Note that as the mean task time increases, the average task rating drops. This is called a negative relationship because as one variable increases (task time), the other variable decreases (task rating). The line that runs through the data is called a trend line and is added easily 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 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 right-clicking on the trend line, choosing "Format Trend Line," and checking the box next to "Display R-squared value on chart."

EXCEL TIP

You can calculate the strength of the relationship between any two variables (such as task time and task rating) using the CORREL function in Excel:

= CORREL(Array 1, Array 2)

Array 1 and Array 2 are the two sets of numbers to be correlated. The result will be a correlation coefficient, or "r". For the data represented in Figure 2.7, $r = -0.53$. 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. The weaker the relationship, the closer the correlation coefficient is to 0. The negative value for "r" signifies the negative relationship between the two variables. If you square the correlation coefficient you get the same value as the R^2 value shown on the scatterplot (0.28).

2.6 NONPARAMETRIC TESTS

Nonparametric tests are used for analyzing nominal and ordinal data. For example, you might want to know if a significant difference exists between men and women for success and failure on a particular task. Or perhaps you're interested in determining whether there is a difference among 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 distributed normally and the 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 χ^2 test because it is probably the most commonly used.

2.6.1 The χ^2 Test

The χ^2 (pronounced "chi square") test is used when you want to compare nominal (or categorical) data. Let's consider an example. Assume you're interested in knowing whether there is a significant difference in task success among three different groups: novice, intermediates, and experts. You run a total of 60 people in your study, 20 in each group. You measure task success or failure on a single task. You count the number of people who were successful in each group. For 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 among the groups.

EXCEL TIP

To perform a χ^2 test in Excel, you use the "CHITEST" function. This function calculates whether differences between observed and expected values are simply due to chance. The function is relatively easy is to use:

= CHITEST(actual_range, expected_range)

The actual range is the number of people successful on the task for each group. The expected range is the total number of people successful (33) divided by the number of groups (3), or 11 in this example. The expected value is what you would expect if there were no differences among any of the three groups.

C7 =CHITEST(B2:B4,C2:C4)

	A	B	C	D
1	**Group**	**Observed**	**Expected**	
2	Novice	6	11	
3	Intermediate	9	11	
4	Experts	18	11	
5	**Total**	**33**	**33**	
6				
7		Chi Test	0.029	

Figure 2.8 Output from a χ^2 test in Excel.

C13 =CHITEST(B3:C5,B9:C11)

	A	B	C	D
1		***Observed***	***Observed***	
2	**Group**	**Design A**	**Design B**	
3	Novice	4	2	
4	Intermediate	6	3	
5	Expert	12	6	
6				
7		***Expected***	***Expected***	
8	**Group**	**Design A**	**Design B**	
9	Novice	5.5	5.5	
10	Intermediate	5.5	5.5	
11	Expert	5.5	5.5	
12				
13		Chi Test	0.003	

Figure 2.9 Output from a χ^2 test with two variables.

Figure 2.8 shows what the data look like and output from the CHITEST function. In this example, the likelihood that this distribution is due to chance is about 2.9% (0.029). Because this number is less than 0.05 (95% confidence), we can reasonably say that there is a difference in success rates among 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 2.9 shows data based on two different variables: group and design. For a more detailed example of using χ^2 to test for differences in live website data for two alternative pages (so-called A/B tests), see Chapter 9.

2.7 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 graphs are available, including those written by Edward Tufte (1990, 1997, 2001, 2006), Stephen Few (2006, 2009, 2012), and Dona Wong (2010). 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 user experience 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 or donut charts
Stacked bar or column graphs

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

GENERAL TIPS FOR DATA GRAPHS

Label the axes and units. It might be obvious to you that a scale of 0 to 100% represents the task completion rate, but it may not be obvious to your audience. 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 would be redundant.

Don't imply more precision in your data than it deserves. 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 used commonly 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.

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 95 or 90% confidence intervals for means via error bars is a good way to visually represent the variability in 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 users, 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 many cases, the use of 3D makes it harder to see the values being plotted.

2.7.1 Column or Bar Graphs

Column graphs and bar graphs (Figure 2.10) 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 used for bar graphs.

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

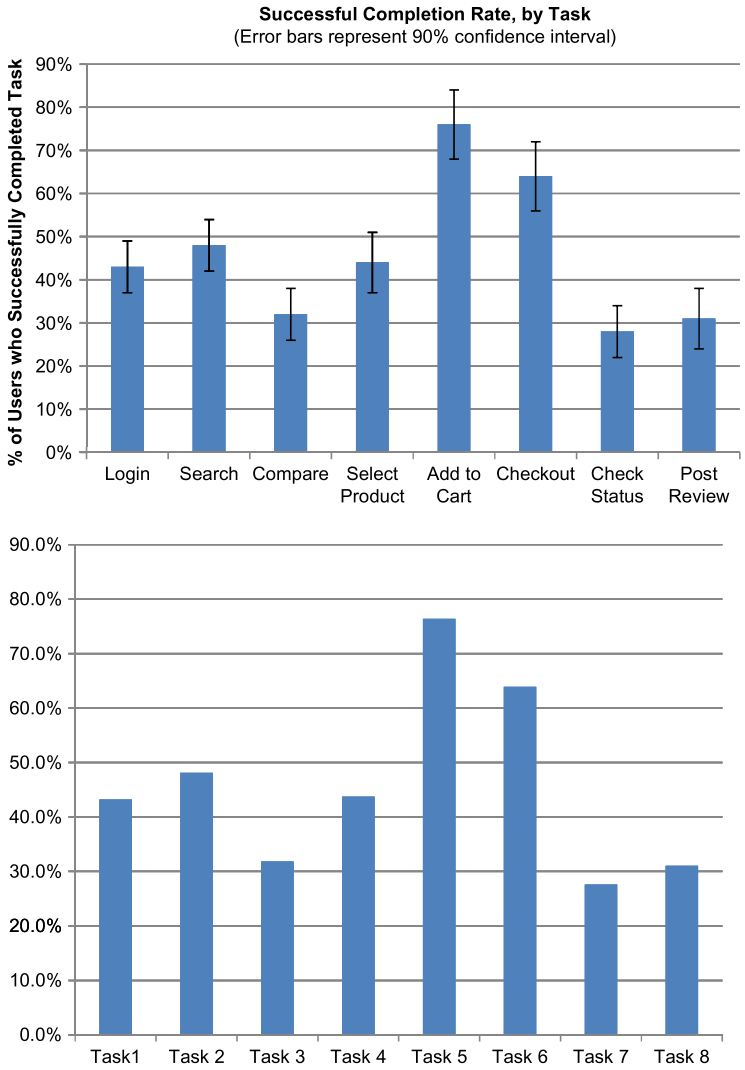


Figure 2.10 Good (top) and bad (bottom) examples of bar graphs for the same data. Mistakes in the bad version include failing to label data, not starting the vertical axis at 0, not showing confidence intervals when you can, and showing too much precision in the vertical axis labels.

- The axis for the continuous variable (the vertical axis in Figure 2.10) 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 manipulating their lengths artificially. The bad example in Figure 2.10 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 completed each task successfully, the theoretical maximum is 100%. 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.

2.7.2 Line Graphs

Line graphs (Figure 2.11) are used most commonly 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.

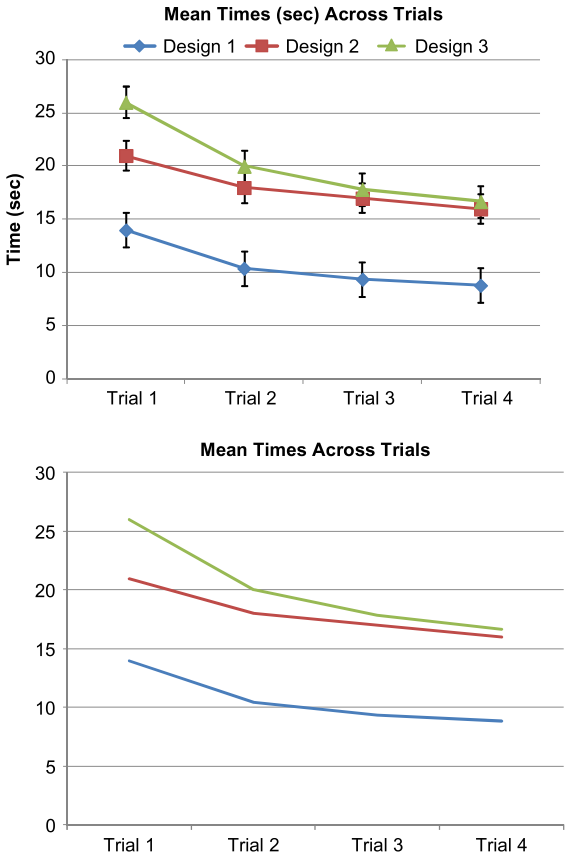


Figure 2.11 Good (top) and bad (bottom) examples of line graphs for the same data. Mistakes in the bad version include failing to label the vertical axis, not showing data points, not including a legend, and not showing confidence intervals.

- Line graphs are appropriate when you want to present the values of one continuous variable (e.g., percent correct, number of errors) as a function of another continuous variable (e.g., age, trial). If one of the variables is discrete (e.g., gender, participant, task), 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 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 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 move the labels manually 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 PowerPoint 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.

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 2.10 as a line graph, as shown in Figure 2.12.However, you should ask yourself whether things such as "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.

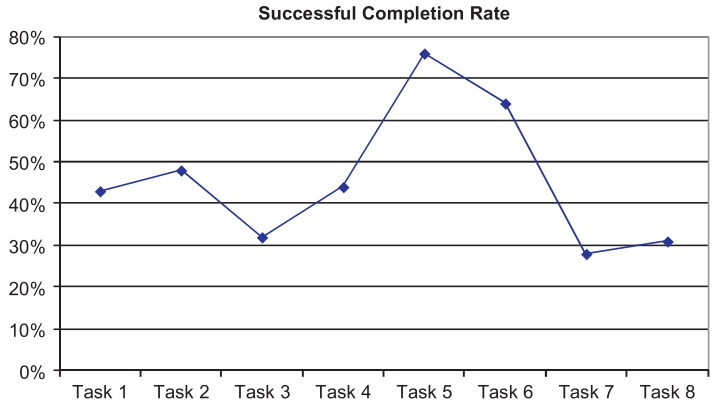


Figure 2.12 An inappropriate line graph of data shown in Figure 2.10. Lines imply that the tasks are a continuous variable, which they are not.

2.7.3 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, especially to illustrate relationships between two variables. 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

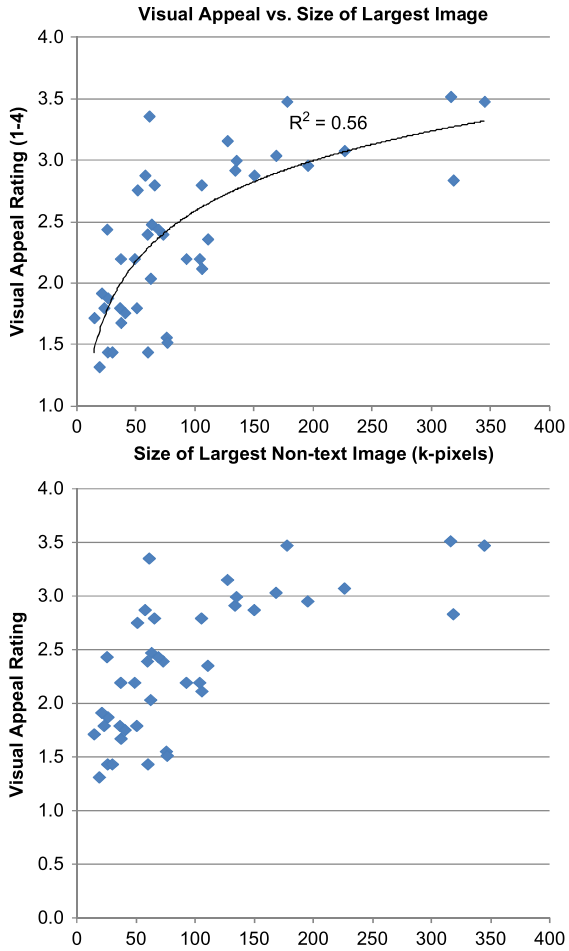


Figure 2.13 Good (top) and bad (bottom) examples of scatterplots for the same data. Mistakes in the bad version include an inappropriate scale for the vertical axis, not showing the scale for visual appeal ratings (1–4), not showing a trend line, and not showing goodness of fit (R^2).

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 often 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.

2.7.4 Pie or Donut Charts

Pie or donut 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 or donut charts are appropriate only when the parts add up to 100%. You have to account for all the cases. In some situations, this might mean creating an "other" category.
- Minimize the number of segments in the chart. Even though the bad example in Figure 2.14 is technically correct, it's almost impossible to make any sense out 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 move the labels manually to prevent them from overlapping.

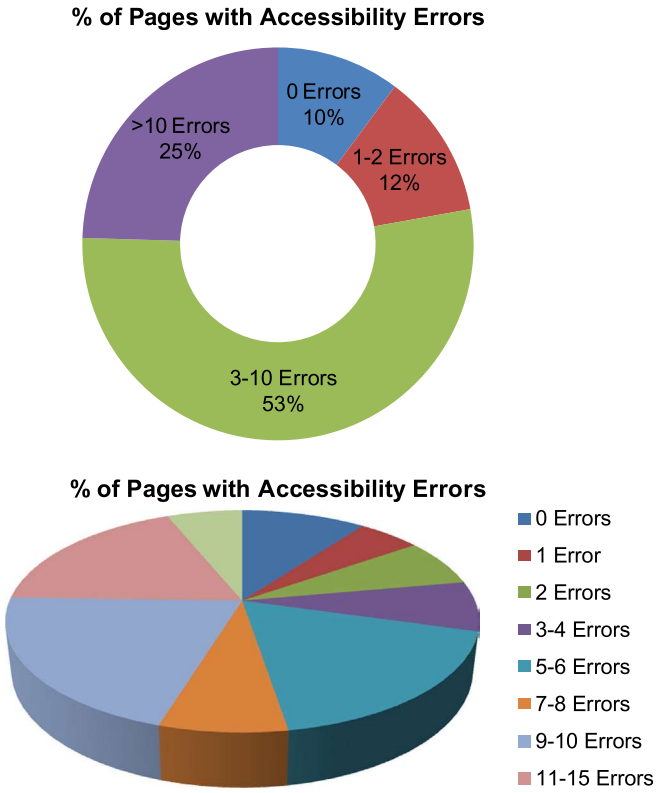


Figure 2.14 Good (top) and bad (bottom) examples of pie or donut charts for the same data. Mistakes in the bad version include too many segments, poor placement of the legend, not showing percentages for each segment, and using 3D, for which the creator of this pie chart should be pummeled with a wet noodle.

2.7.5 Stacked Bar or Column Graphs

Stacked bar graphs (Figure 2.15) are basically multiple pie charts shown in bar or column form. They're appropriate whenever you have a series of data sets, each of which represents parts of the whole. Their most common use in user experience 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%.
- The items in the series are normally categorical (e.g., tasks, participants).

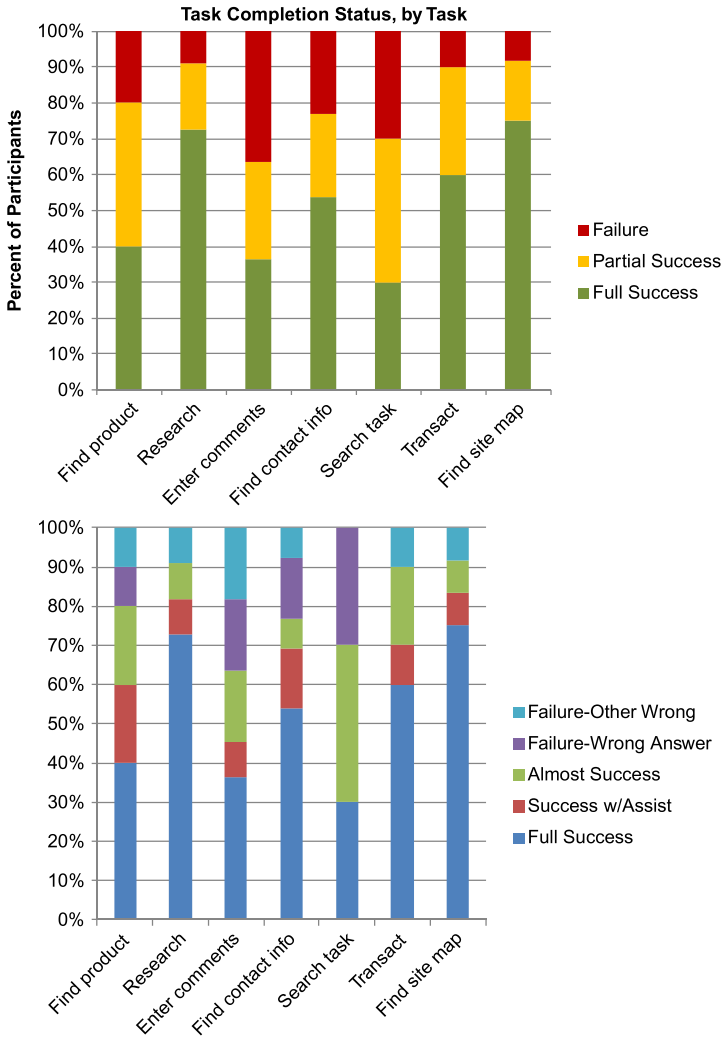


Figure 2.15 Good and bad examples of stacked bar graphs for the same data. Mistakes in the bad version include too many segments, poor color coding, and failing to label the vertical axis.

- 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 of these conventions can be helpful, as in the good example in Figure 2.15, but don't rely solely on them.

2.8 SUMMARY

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

1. When analyzing your results, it's critical to know your data. The specific type of data you have will dictate what statistics you can (and can't) perform.
2. Nominal data are categorical, such as binary task success or males and females. Nominal data are usually expressed as frequencies or percentages. χ^2 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.
3. 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 χ^2 test.
4. 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.
5. Ratio data are the same as interval but with a natural zero. One example is completion times. Essentially, the same statistics that apply to interval data also apply to ratio data.
6. Any time you can calculate a mean, you can also calculate a confidence interval for that mean. Displaying confidence intervals on graphs of means helps the viewer understand the accuracy of the data and to see quickly any differences between means.
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 data sum to 100%.