Chapter 6

The t-test and Basic Inference Principles

The t-test is used as an example of the basic principles of statistical inference.

One of the simplest situations for which we might design an experiment is the case of a nominal two-level explanatory variable and a quantitative outcome variable. Table 6.1 shows several examples. For all of these experiments, the treatments have two levels, and the treatment variable is nominal. Note in the table the various experimental units to which the two levels of treatment are being applied for these examples.. If we randomly assign the treatments to these units this will be a randomized experiment rather than an observational study, so we will be able to apply the word “causes” rather than just “is associated with” to any statistically significant result. This chapter only discusses so-called “between subjects” explanatory variables, which means that we are assuming that each experimental unit is exposed to only one of the two levels of treatment (even though that is not necessarily the most obvious way to run the fMRI experiment).

This chapter shows one way to perform statistical inference for the two-group, quantitative outcome experiment, namely the independent samples t-test. More importantly, the t-test is used as an example for demonstrating the basic principles of statistical inference that will be used throughout the book. The understanding of these principles, along with some degree of theoretical underpinning, is key to using statistical results intelligently. Among other things, you need to really understand what a p-value and a confidence interval tell us, and when they can
An alternative inferential procedure is one-way ANOVA, which always gives the same results as the t-test, and is the topic of the next chapter.

As mentioned in the preface, it is hard to find a linear path for learning experimental design and analysis because so many of the important concepts are interdependent. For this chapter we will assume that the subjects chosen to participate in the experiment are representative, and that each subject is randomly assigned to exactly one treatment. The reasons we should do these things and the consequences of not doing them are postponed until the Threats chapter. For now we will focus on the EDA and confirmatory analyses for a two-group between-subjects experiment with a quantitative outcome. This will give you a general picture of statistical analysis of an experiment and a good foundation in the underlying theory. As usual, more advanced material, which will enhance your understanding but is not required for a fairly good understanding of the concepts, is shaded in gray.
6.1. Case study from the field of Human-Computer Interaction (HCI)

This (fake) experiment is designed to determine which of two background colors for computer text is easier to read, as determined by the speed with which a task described by the text is performed. The study randomly assigns 35 university students to one of two versions of a computer program that presents text describing which of several icons the user should click on. The program measures how long it takes until the correct icon is clicked. This measurement is called “reaction time” and is measured in milliseconds (ms). The program reports the average time for 20 trials per subject. The two versions of the program differ in the background color for the text (yellow or cyan).

The data can be found in the file background.sav on this book’s web data site. It is tab delimited with no header line and with columns for subject identification, background color, and response time in milliseconds. The coding for the color column is 0=yellow, 1=cyan. The data look like this:

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Color</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYP</td>
<td>0</td>
<td>859</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTS</td>
<td>1</td>
<td>1005</td>
</tr>
</tbody>
</table>

Note that in SPSS if you enter the “Values” for the two colors and turn on “Value labels”, then the color words rather than the numbers will be seen in the second column. Because this data set is not too large, it is possible to examine it to see that 0 and 1 are the only two values for Color and that the time ranges from 291 to 1005 milliseconds (or 0.291 to 1.005 seconds). Even for a dataset this small, it is hard to get a good idea of the differences in response time across the two colors just by looking at the numbers.

Here are some basic univariate exploratory data analyses. There is no point in doing EDA for the subject IDs. For the categorical variable Color, the only useful non-graphical EDA is a tabulation of the two values.
### Frequencies

<table>
<thead>
<tr>
<th>Background Color</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid yellow</td>
<td>17</td>
<td>48.6</td>
<td>48.6</td>
</tr>
<tr>
<td>Valid cyan</td>
<td>18</td>
<td>51.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The “Frequency” column gives the basic tabulation of the variable’s values. Seventeen subjects were shown a yellow background, and 18 were shown cyan for a total of 35 subjects. The “Percent Valid” vs. “Percent” columns in SPSS differ only if there are missing values. The Percent Valid column always adds to 100% across the categories given, while the Percent column will include a “Missing” category if there are missing data. The Cumulative Percent column accounts for each category plus all categories on prior lines of the table; this is not very useful for nominal data.

This is non-graphical EDA. Other non-graphical exploratory analyses of Color, such as calculation of mean, variance, etc. don’t make much sense because Color is a categorical variable. (It is possible to interpret the mean in this case because yellow is coded as 0 and cyan is coded as 1. The mean, 0.514, represents the fraction of cyan backgrounds.) For graphical EDA of the color variable you could make a pie or bar chart, but this really adds nothing to the simple 48.6 vs 51.4 percent numbers.

For the quantitative variable Reaction Time, the non-graphical EDA would include statistics like these:

<table>
<thead>
<tr>
<th>Reaction Time (ms)</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35</td>
<td>291</td>
<td>1005</td>
<td>670.03</td>
<td>180.152</td>
</tr>
</tbody>
</table>

Here we can see that there are 35 reactions times that range from 291 to 1005 milliseconds, with a mean of 670.03 and a standard deviation of 180.152. We can calculate that the variance is $180.152^2 = 32454$, but we need to look further at the data to calculate the median or IQR. If we were to assume that the data follow a Normal distribution, then we could conclude that about 95% of the data fall within mean plus or minus 2 sd, which is about 310 to 1030. But such an assumption is most likely incorrect, because if there is a difference in reaction times between the two colors, we would expect that the distribution of reaction times ignoring color would be some bimodal distribution that is a mixture of the two individual
6.1. CASE STUDY FROM THE FIELD OF HUMAN-COMPUTER INTERACTION (HCI)

reaction time distributions for the two colors.

A histogram and/or boxplot of reaction time will further help you get a feel for the data and possibly find errors.

For bivariate EDA, we want graphs and descriptive statistics for the quantitative outcome (dependent) variable Reaction Time broken down by the levels of the categorical explanatory variable (factor) Background Color. A convenient way to do this in SPSS is with the “Explore” menu option. Abbreviated results are shown in this table and the graphical EDA (side-by-side boxplots) is shown in figure 6.1.

<table>
<thead>
<tr>
<th>Background Color</th>
<th>Statistics</th>
<th>Std.Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Std.Error</td>
</tr>
<tr>
<td>Reaction Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>Mean</td>
<td>679.65</td>
</tr>
<tr>
<td></td>
<td>95% Confidence Lower Bound</td>
<td>587.7</td>
</tr>
<tr>
<td></td>
<td>Interval for Mean Upper Bound</td>
<td>761.60</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>683.05</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>159.387</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>906</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-0.411</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.550</td>
</tr>
<tr>
<td>Cyan</td>
<td>Mean</td>
<td>660.94</td>
</tr>
<tr>
<td></td>
<td>95% Confidence Lower Bound</td>
<td>560.47</td>
</tr>
<tr>
<td></td>
<td>Interval for Mean Upper Bound</td>
<td>761.42</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>662.38</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>202.039</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>1.036</td>
</tr>
</tbody>
</table>

Very briefly, the mean reaction times for the subjects shown cyan backgrounds is about 19 ms shorter than the mean for those shown yellow backgrounds. The standard deviation of the reaction times is somewhat larger for the cyan group than it is for the yellow group.
Figure 6.1: Boxplots of reaction time by color.
EDA for the two-group quantitative outcome experiment should include examination of sample statistics for mean, standard deviation, skewness, and kurtosis separately for each group, as well as boxplots and histograms.

We should follow up on this EDA with formal statistical testing. But first we need to explore some important concepts underlying such analyses.

6.2 How classical statistical inference works

In this section you will see ways to think about the state of the real world at a level appropriate for scientific study, see how that plays out in experimentation, and learn how we match up the real world to the theoretical constructs of probability and statistics. In the next section you will see the details of how formal inference is carried out and interpreted.

How should we think about the real world with respect to a simple two group experiment with a continuous outcome? Obviously, if we were to repeat the entire experiment on a new set of subjects, we would (almost surely) get different results. The reasons that we would get different results are many, but they can be broken down into several main groups (see section 8.5) such as measurement variability, environmental variability, treatment application variability, and subject-to-subject variability. The understanding of the concept that our experimental results are just one (random) set out of many possible sets of results is the foundation of statistical inference.

The key to standard (classical) statistical analysis is to consider what types of results we would get if specific conditions are met and if we were to repeat an experiment many times, and then to compare the observed result to these hypothetical results and characterize how “typical” the observed result is.
6.2.1 The steps of statistical analysis

Most formal statistical analyses work like this:

1. Use your judgement to choose a model (mean and error components) that is a reasonable match for the data from the experiment. The model is expressed in terms of the population from which the subjects (and outcome variable) were drawn. Also, define parameters of interest.

2. Using the parameters, define a (point) null hypothesis and a (usually complex) alternative hypothesis which correspond to the scientific question of interest.

3. Choose (or invent) a statistic which has different distributions under the null and alternative hypotheses.

4. Calculate the null sampling distribution of the statistic.

5. Compare the observed (experimental) statistic to the null sampling distribution of that statistic to calculate a p-value for a specific null hypothesis (and/or use similar techniques to compute a confidence interval for a quantity of interest).

6. Perform some kind of assumption checks to validate the degree of appropriateness of the model assumptions.

7. Use your judgement to interpret the statistical inference in terms of the underlying science.

Ideally there is one more step, which is the power calculation. This involves calculating the distribution of the statistic under one or more specific (point) alternative hypotheses before conducting the experiment so that we can assess the likelihood of getting a “statistically significant” result for various “scientifically significant” alternative hypotheses.

All of these points will now be discussed in more detail, both theoretically and using the HCI example. Focus is on the two group, quantitative outcome case, but the general principles apply to many other situations.
Classical statistical inference involves multiple steps including definition of a model, definition of statistical hypotheses, selection of a statistic, computation of the sampling distribution of that statistic, computation of a p-value and/or confidence intervals, and interpretation.

6.2.2 Model and parameter definition

We start with definition of a model and parameters. We will assume that the subjects are representative of some population of interest. In our two-treatment-group example, we most commonly consider the parameters of interest to be the population means of the outcome variable (true value without measurement error) for the two treatments, usually designated with the Greek letter mu (µ) and two subscripts. For now let’s use µ₁ and µ₂, where in the HCI example µ₁ is the population mean of reaction time for subjects shown the yellow background and µ₂ is the population mean for those shown the cyan background. (A good alternative is to use µ_Y and µ_C, which are better mnemonically.)

It is helpful to think about the relationship between the treatment randomization and the population parameters in terms of counterfactuals. Although we have the measurement for each subject for the treatment (background color) to which they were assigned, there is also “against the facts” a theoretical “counterfactual” result for the treatment they did not get. A useful way to visualize this is to draw each member of the population of interest in the shape of a person. Inside this shape for each actual person (potential subject) are many numbers which are their true values for various outcomes under many different possible conditions (of treatment and environment). If we write the reaction time for a yellow background near the right ear and the reaction time for cyan near the left ear, then the parameter µ₁ is the mean of the right ear numbers over the entire population. It is this parameter, a fixed, unknown “secret of nature” that we want to learn about, not the corresponding (noisy) sample quantity for the random sample of subjects randomly assigned to see a yellow background. Put another way, in essentially every experiment that we run, the sample means of the outcomes for the treatment groups differ, even if there is really no true difference between the outcome mean parameters for the two treatments in the population, so focusing on those differences is not very meaningful.
Figure 6.2 shows a diagram demonstrating this way of thinking. The first two subjects of the population are shown along with a few of their attributes. The population mean of any attribute is a parameter that may be of interest in a particular experiment. Obviously we can define many parameters (means, variances, etc.) for many different possible attributes, both marginally and conditionally on other attributes, such as age, gender, etc. (see section 3.6).

It must be strongly emphasized that statistical inference is all about learning what we can about the (unknowable) population parameters and not about the sample statistics per se.

As mentioned in section 1.2 a statistical model has two parts, the structural model and the error model. The structural model refers to defining the pattern of means for groups of subjects defined by explanatory variables, but it does not state what values these means take. In the case of the two group experiment, simply defining the population means (without making any claims about their equality or non-equality) defines the structural model. As we progress through the course, we will have more complicated structural models.

The error model (noise model) defines the variability of subjects “in the same group” around the mean for that group. (The meaning of “in the same group” is obvious here, but is less so, e.g., in regression models.) We assume that we cannot predict the deviation of individual measurements from the group mean more exactly than saying that they randomly follow the probability distribution of the error model.

For continuous outcome variables, the most commonly used error model is that for each treatment group the distribution of outcomes in the population is normally distributed, and furthermore that the population variances of the groups are equal. In addition, we assume that each error (deviation of an individual value from the group population mean) is statistically independent of every other error. The normality assumption is often approximately correct because (as stated in the CLT) the sum of many small non-Normal random variables will be normally distributed, and most outcomes of interest can be thought of as being affected in some additive way by many individual factors. On the other hand, it is not true that all outcomes are normally distributed, so we need to check our assumptions before interpreting any formal statistical inferences (step 5). Similarly, the assumption of
Figure 6.2: A view of a population and parameters.

I=IQ, W=waist size, S=soccer kick distance
Y/C=reaction times with yellow/cyan backgrounds
equal variance is often but not always true.

The structural component of a statistical model defines the means of groups, while the error component describes the random pattern of deviation from those means.

6.2.3 Null and alternative hypotheses

The null and alternative hypotheses are statements about the population parameters that express different possible characterizations of the population which correspond to different scientific hypotheses. Almost always the null hypothesis is a so-called point hypothesis in the sense that it defines a specific case (with an equal sign), and the alternative is a complex hypothesis in that it covers many different conditions with less than (<), greater than (>), or unequal (≠) signs.

In the two-treatment-group case, the usual null hypothesis is that the two population means are equal, usually written as \( H_0 : \mu_1 = \mu_2 \), where the symbol \( H_0 \), read “H zero” or “H naught” indicates the null hypothesis. Note that the null hypothesis is usually interpretable as “nothing interesting is going on,” and that is why the term null is used.

In the two-treatment-group case, the usual alternative hypothesis is that the two population means are unequal, written as \( H_1 : \mu_1 ≠ \mu_2 \) or \( H_A : \mu_1 ≠ \mu_2 \) where \( H_1 \) or \( H_A \) are interchangeable symbols for the alternative hypothesis. (Occasionally we use an alternative hypothesis that states that one population mean is less than the other, but in my opinion such a “one-sided hypothesis” should only be used when the opposite direction is truly impossible.) Note that there are really an infinite number of specific alternative hypotheses, e.g., \( |\mu_0 - \mu_1| = 1 \), \( |\mu_0 - \mu_1| = 2 \), etc. It is in this sense that the alternative hypothesis is complex, and this is an important consideration in power analysis.

The null hypothesis specifies patterns of mean parameters corresponding to no interesting effects, while the alternative hypothesis usually covers everything else.
6.2.4 Choosing a statistic

The next step is to find (or invent) a statistic that has a different distribution for the null and alternative hypotheses and for which we can calculate the null sampling distribution (see below). It is important to realize that the sampling distribution of the chosen statistic differs for each specific alternative, that there is almost always overlap between the null and alternative distributions of the statistic, and that the overlap is large for alternatives that reflect small effects and smaller for alternatives that reflect large effects.

For the two-treatment-group experiment with a quantitative outcome a commonly used statistic is the so-called “t” statistic which is the difference between the sample means (in either direction) divided by the (estimated) standard error (see below) of that difference. Under certain assumptions it can be shown that this statistic is “optimal” (in terms of power), but a valid test does not require optimality and other statistics are possible. In fact we will encounter situations where no one statistic is optimal, and different researchers might choose different statistics for their formal statistical analyses.

Inference is usually based on a single statistic whose choice may or may not be obvious or unique.

The standard error of the difference between two sample means is the standard deviation of the sampling distribution of the difference between the sample means. Statistical theory shows that under the assumptions of the t-test, the standard error of the difference is

\[
\text{SE}(\text{diff}) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

where \( n_1 \) and \( n_2 \) are the group sample sizes. Note that this simplifies to \( \sqrt{2\sigma^2/\sqrt{n}} \) when the sample sizes are equal.

In practice the estimate of the SE that uses an appropriate averaging
of the observed sample variances is used.

\[
\text{estimated SE}(\text{diff}) = \sqrt{\frac{s_1^2(df_1) + s_2^2(df_2)}{df_1 + df_2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

where \( df_1 = n_1 - 1 \) and \( df_2 = n_2 - 1 \). This estimated standard error has \( n_1 + n_2 - 2 = df_1 + df_2 \) degrees of freedom.

### 6.2.5 Computing the null sampling distribution

The next step in the general scheme of formal (classical) statistical analysis is to compute the **null sampling distribution** of the chosen statistic. The null sampling distribution of a statistic is the probability distribution of the statistic calculated over repeated experiments under the conditions defined by the model assumptions and the null hypothesis. For our HCI example, we consider what would happen if the truth is that there is no difference in reaction times between the two background colors, and we repeatedly sample 35 subjects and randomly assign yellow to 17 of them and cyan to 18 of them, and then calculate the t-statistic each time. The distribution of the t-statistics under these conditions is the null sampling distribution of the t-statistic appropriate for this experiment.

For the HCI example, the null sampling distribution of the t-statistic can be shown to match a well known, named continuous probability distribution called the “t-distribution” (see section 3.9). Actually there are an infinite number of t-distributions (a family of distributions) and these are named (indexed) by their “degrees of freedom” (df). For the two-group quantitative outcome experiment, the df of the t-statistic and its corresponding null sampling distribution is \( (n_1 - 1) + (n_2 - 1) \), so we will use the t-distribution with \( n_1 + n_2 - 2 \) df to make our inferences. For the HCI experiment, this is 17+18-2=33 df.

The calculation of the mathematical form (pdf) of the null sampling distribution of any chosen statistic using the assumptions of a given model is beyond the scope of this book, but the general idea can be seen in section 3.7.
Probability theory (beyond the scope of this book) comes into play in computing the null sampling distribution of the chosen statistic based on the model assumptions.

You may notice that the null hypothesis of equal population means is in some sense “complex” rather than “point” because the two means could be both equal to 600, 601, etc. It turns out that the t-statistic has the same null sampling distribution regardless of the exact value of the population mean (and of the population variance), although it does depend on the sample sizes, $n_1$ and $n_2$.

### 6.2.6 Finding the p-value

Once we have the null sampling distribution of a statistic, we can see whether or not the observed statistic is “typical” of the kinds of values that we would expect to see when the null hypothesis is true (which is the basic interpretation of the null sampling distribution of the statistic). If we find that the observed (experimental) statistic is typical, then we conclude that our experiment has not provided evidence against the null hypothesis, and if we find it to be atypical, we conclude that we do have evidence against the null hypothesis.

The formal language we use is to either “reject” the null hypothesis (in favor of the alternative) or to “retain” the null hypothesis. The word “accept” is not a good substitute for retain (see below). The inferential conclusion to “reject” or “retain” the null hypothesis is simply a conjecture based on the evidence. But whichever inference we make, there is an underlying truth (null or alternative) that we can never know for sure, and there is always a chance that we will be wrong in our conclusion even if we use all of our statistical tools correctly.

Classical statistical inference focuses on controlling the chance that we reject the null hypothesis incorrectly when the underlying truth is that the null hypothesis is correct. We call the erroneous conclusion that the null hypothesis is incorrect when it is actually correct a **Type 1 error**. (But because the true state of the null hypothesis is unknowable, we never can be sure whether or not we have made
A Type 1 error in any specific actual situation.) A synonym for Type 1 error is “false rejection” of the null hypothesis.

The usual way that we make a formal, objective reject vs. retain decision is to calculate a p-value. Formally, a p-value is the probability that any given experiment will produce a value of the chosen statistic equal to the observed value in our actual experiment or something more extreme (in the sense of less compatible with the null hypotheses), when the null hypothesis is true and the model assumptions are correct. Be careful: the latter half of this definition is as important as the first half.

For the HCI example, the numerator of the t-statistic is the difference between the observed sample means. Therefore values near zero support the null hypothesis of equal population means, while values far from zero in either direction support the alternative hypothesis of unequal population means. In our specific experiment the t-statistic equals 0.30. A value of -0.30 would give exactly the same degree of evidence for or against the null hypothesis (and the direction of subtraction is arbitrary). Values smaller in absolute value than 0.30 are more suggestive that the underlying truth is equal population means, while larger values support the alternative hypothesis. So the p-value for this experiment is the probability of getting a t-statistic greater than 0.30 or less than -0.30 based on the null sampling distribution of the t-distribution with 33 df. As explained in chapter 3, this probability is equal to the corresponding area under the curve of the pdf of the null sampling distribution of the statistic. As shown in figure 6.3 the chance that a random t-statistic is less than -0.30 if the null hypothesis is true is 0.382, as is the chance that it is above +0.30. So the p-value equals 0.382+0.382=0.764, i.e. 76.4% of null experiments would give a t-value this large or larger (in absolute value). We conclude that the observed outcome (t=0.30) is not uncommonly far from zero when the null hypothesis is true, so we have no reason to believe that the null hypothesis is false.

The usual convention (and it is only a convention, not anything stronger) is to reject the null hypothesis if the p-value is less than or equal to 0.05 and retain
Figure 6.3: Calculation of the p-value for the HCI example
it otherwise. Under some circumstances it is more appropriate to use numbers bigger or smaller than 0.05 for this decision rule. We call the cutoff value the significance level of a test, and use the symbol alpha (\(\alpha\)), with the conventional alpha being 0.05. We use the phrase statistically significant at the 0.05 (or some other) level, when the p-value is less than or equal to 0.05 (or some other value). This indicates that if we have used a correct model, i.e., the model assumptions mirror reality and if the null hypothesis happens to be correct, then a result like ours or one even more “un-null-like” would happen at most 5% of the time. It is reasonable to say that because our result is atypical for the null hypothesis, then claiming that the alternative hypothesis is true is appropriate. But when we get a p-value of less than or equal to 0.05 and we reject the null hypothesis, it is completely incorrect to claim that there is only a 5% chance that we have made an error. For more details see chapter 12.

You should never use the word “insignificant” to indicate a large p-value. Use “not significant” or “non-significant” because “insignificant” implies no substantive significance rather than no statistical significance.

The most common decision rule is to reject the null hypothesis if the p-value is less than or equal to 0.05 and to retain it otherwise.

It is important to realize that the p-value is a random quantity. If we could repeat our experiment (with no change in the underlying state of nature), then we would get a different p-value. What does it mean for the p-value to be “correct”? For one thing it means that we have made the calculation correctly, but since the computer is doing the calculation we have no reason to doubt that. What is more important is to ask whether the p-value that we have calculated is giving us appropriate information. For one thing, when the null hypothesis is really true (which we can never know for certain) an appropriate p-value will be less than 0.05 exactly 5% of the time over repeated experiments. So if the null hypothesis is true, and if you and 99 of your friends independently conduct experiments, about five of you will get p-values less than or equal to 0.05 causing you to incorrectly reject the null hypothesis. Which five people this happens to has nothing to do with the quality of their research; it just happens because of bad luck!

And if an alternative hypothesis is true, then all we know is that the p-value will be less than or equal to 0.05 at least 5% of the time, but it might be as little
6% of the time. So a “correct” p-value does not protect you from making a lot of Type 2 errors which happen when you incorrectly retain the null hypothesis. With Type 2 errors, something interesting is going on in nature, but you miss it. See section 6.2.10 for more on this “power” problem.

We talk about an “incorrect” p-value mostly with regard to the situation where the null hypothesis is the underlying truth. It is really the behavior of the p-value over repeats of the experiment that is incorrect, and we want to identify what can cause that to happen even though we will usually see only a single p-value for an experiment. Because the p-value for an experiment is computed as an area under the pdf of the null sampling distribution of a statistic, the main reason a p-value is “incorrect” (and therefore misleading) is that we are not using the appropriate null sampling distribution. That happens when the model assumptions used in the computation of the null sampling distribution of the statistic are not close to the reality of nature. For the t-test, this can be caused by non-normality of the distributions (though this is not a problem if the sample size is large due to the CLT), unequal variance of the outcome measure for the two-treatment-groups, confounding of treatment group with important unmeasured explanatory variables, or lack of independence of the measures (for example if some subjects are accidentally measured in both groups). If any of these “assumption violations” are sufficiently large, the p-value loses its meaning, and it is no longer an interpretable quantity.

A p-value has meaning only if the correct null sampling distribution of the statistic has been used, i.e., if the assumptions of the test are (reasonably well) met. Computer programs generally give no warnings when they calculate incorrect p-values.

6.2.7 Confidence intervals

Besides p-values, another way to express what the evidence of an experiment is telling us is to compute one or more confidence intervals, often abbreviated CI. We would like to make a statement like “we are sure that the difference between $\mu_1$ and $\mu_2$ is no more than 20 ms. That is not possible! We can only make statements such as, “we are 95% confident that the difference between $\mu_1$ and $\mu_2$ is no more
than 20 ms.” The choice of the percent confidence number is arbitrary; we can choose another number like 99% or 75%, but note that when we do so, the width of the interval changes (high confidence requires wider intervals).

The actual computations are usually done by computer, but in many instances the idea of the calculation is simple.

If the underlying data are normally distributed, or if we are looking at a sum or mean with a large sample size (and can therefore invoke the CLT), then a confidence interval for a quantity (statistic) is computed as the statistic plus or minus the appropriate “multiplier” times the estimated standard error of the quantity. The multiplier used depends on both the desired confidence level (e.g., 95% vs. 90%) and the degrees of freedom for the standard error (which may or may not have a simple formula). The multiplier is based on the t-distribution which takes into account the uncertainty in the standard deviation used to estimate the standard error. We can use a computer or table of the t-distribution to find the multiplier as the value of the t-distribution for which plus or minus that number covers the desired percentage of the t-distribution with the correct degrees of freedom. If we call the quantity 1-(confidence percentage)/100 as alpha (α), then the multiplier is the 1-α/2 quantile of the appropriate t-distribution.

For our HCI example the 95% confidence interval for the fixed, unknown, “secret-of-nature” that equals \( \mu_1 - \mu_2 \) is \([-106.9, 144.4]\). We are 95% confident that the mean reaction time is between 106.9 ms shorter and 144.4 ms longer for the yellow background compared to cyan. The meaning of this statement is that if all of the assumptions are met, and if we repeat the experiment many times, the random interval that we compute each time will contain the single, fixed, true parameter value 95% of the time. Similar to the interpretation a p-value, if 100 competent researchers independently conduct the same experiment, by bad luck about five of them will unknowingly be incorrect if they claim that the 95% confidence interval that they correctly computed actually contains the true parameter value.

Confidence intervals are in many ways more informative than p-values. Their greatest strength is that they help a researcher focus on substantive significance in addition to statistical significance. Consider a bakery that does an experiment
to see if an additional complicated step will reduce waste due to production of unsaleable, misshapen cupcakes. If the amount saved has a 95% CI of [0.1, 0.3] dozen per month with a p-value of 0.02, then even though this would be statistically significant, it would not be substantively significant.

In contrast, if we had a 95% CI of [-30, 200] dozen per month with p=0.15, then even though this not statistically significant, the inclusion of substantively important values like 175 dozen per month tells us that the experiment has not provided enough information to make a good, real world conclusion.

Finally, if we had a 95% CI of [-0.1, 0.2] dozen per month with p=0.15, we would conclude that even if a real non-zero difference exists, its magnitude is not enough to add the complex step to our cupcake making.

Confidence intervals can add a lot of important real world information to p-values and help us complement statistical significance with substantive significance.

The slight downside to CIs and substantive significance is that they are hard to interpret if you don’t know much about your subject matter. This is usually only a problem for learning exercises, not for real experiments.

6.2.8 Assumption checking

We have seen above that the p-value can be misleading or “wrong” if the model assumptions used to construct the statistic’s sampling distribution are not close enough to the reality of the situation. To protect against being mislead, we usually perform some assumption checking after conducting an analysis but before considering its conclusions.

Depending on the model, assumption checking can take several different forms. A major role is played by examining the model residuals. Remember that our standard model says that for each treatment group the best guess (the expected or predicted value) for each observation is defined by the means of the structural model. Then the observed value for each outcome observation is deviated higher or lower than the true mean. The error component of our model describes the distribution of these deviations, which are called errors. The residuals, which are
defined as observed minus expected value for each outcome measurement, are our best estimates of the unknowable, true errors for each subject. We will examine the distribution of the residuals to allow us to make a judgment about whether or not the distribution of the errors is consistent with the error model.

Assumption checking is needed to verify that the assumptions involved in the initial model construction were good enough to allow us to believe our inferences.

Defining groups among which all subjects have identical predictions may be complicated for some models, but is simple for the 2-treatment-group model. For this situation, all subjects in either one of the two treatment groups appear to be identical in the model, so they must have the same prediction based on the model. For the t-test, the observed group means are the two predicted values from which the residuals can be computed. Then we can check if the residuals for each group follow a Normal distribution with equal variances for the two groups (or more commonly, we check the equality of the variances and check the normality of the combined set of residuals).

Another important assumption is the independence of the errors. There should be nothing about the subjects that allows us to predict the sign or the magnitude of one subject’s error just by knowing the value of another specific subject’s error. As a trivial example, if we have identical twins in a study, it may well be true that their errors are not independent. This might also apply to close friends in some studies. The worst case is to apply both treatments to each subject, and then pretend that we used two independent samples of subjects. Usually there is no way to check the independence assumption from the data; we just need to think about how we conducted the experiment to consider whether the assumption might have been violated. In some cases, because the residuals can be looked upon as a substitute for the true unknown errors, certain residual analyses may shed light on the independent errors assumption.

You can be sure that the underlying reality of nature is never perfectly captured by our models. This is why statisticians say “all models are wrong, but some are useful.” It takes some experience to judge how badly the assumptions can be bent before the inferences are broken. For now, a rough statement can be made about the independent samples t-test: we need to worry about the reasonableness of the inference if the normality assumption is strongly violated, if the equal vari-
ance assumption is moderately violated, or if the independent errors assumption is mildly violated. We say that a statistical test is robust to a particular model violation if the p-value remains approximately “correct” even when the assumption is moderately or severely violated.

All models are wrong, but some are useful. It takes experience and judgement to evaluate model adequacy.

6.2.9 Subject matter conclusions

Applying subject matter knowledge to the confidence interval is one key form of relating statistical conclusions back to the subject matter of the experiment. For p-values, you do something similar with the reject/retain result of your decision rule. In either case, an analysis is incomplete if you stop at reporting the p-value and/or CI without returning to the original scientific question(s).

6.2.10 Power

The power of an experiment is defined for specific alternatives, e.g., $|\mu_1 - \mu_2| = 100$, rather than for the entire, complex alternative hypothesis. The power of an experiment for a given alternative hypothesis is the chance that we will get a statistically significant result (reject the null hypothesis) when that alternative is true for any one realization of the experiment. Power varies from $\alpha$ to 1.00 (or $100\alpha$% to 100%). The concept of power is related to Type 2 error, which is the error we make when we retain the null hypothesis when a particular alternative is true. Usually the rate of making Type 2 errors is symbolized by beta ($\beta$). Then power is $1-\beta$ or 100-100$\beta$%. Typically people agree that 80% power ($\beta$=20%) for some substantively important effect size (specific magnitude of a difference as opposed to the zero difference of the null hypothesis) is a minimal value for good power.

It should be fairly obvious that for any given experiment you have more power to detect a large effect than a small one.

You should use the methods of chapter 12 to estimate the power of any experiment before running it. This is only an estimate or educated guess because some
needed information is usually not known. Many, many experiments are performed which have insufficient power, often in the 20-30% range. This is horrible! It means that even if you are studying effective treatments, you only have a 20-30% chance of getting a statistically significant result. Combining power analysis with intelligent experimental design to alter the conduct of the experiment to maximize its power is a quality of a good scientist.

**Poor power is a common problem. It cannot be fixed by statistical analysis. It must be dealt with before running your experiment.**

For now, the importance of power is how it applies to inference. If you get a small p-value, power becomes irrelevant, and you conclude that you should reject the null hypothesis, always realizing that there is a chance that you might be making a Type 1 error. If you get a large p-value, you “retain” the null hypothesis. If the power of the experiment is small, you know that a true null hypothesis and a Type 2 error are not distinguishable. But if you have good power for some reasonably important sized effect, then a large p-value is good evidence that no important sized effect exists, although a Type 2 error is still possible.

**A non-significant p-value and a low power combine to make an experiment totally uninformative.**

**In a nutshell:** All classical statistical inference is based on the same set of steps in which a sample statistic is compared to the kinds of values we would expect it to have if nothing interesting is going on, i.e., if the null hypothesis is true.

### 6.3 Do it in SPSS

Figure 6.4 shows the Independent Samples T-test dialog box.
Before performing the t-test, check that your outcome variable has Measure “scale” and that you know the numeric codes for the two levels of your categorical (nominal) explanatory variable.

To perform an independent samples t-test in SPSS, use the menu item “Independent Samples T-Test” found under Analyze/CompareMeans. Enter the outcome (dependent) variable into the Test Variables box. Enter the categorical explanatory variable into the Grouping Variable box. Click “Define Groups” and enter the numeric codes for the two levels of the explanatory variable and click Continue. Then click OK to produce the output. (The t-statistic will be calculated in the direction that subtracts the level you enter second from the level you enter first.)

For the HCI example, put Reaction Time in the Test Variables box, and Background Color in the Grouping Variable box. For Define Groups enter the codes 0 and 1.

6.4 Return to the HCI example

The SPSS output for the independent samples (two-sample) t-test for the HCI text background color example is shown in figure 6.5.

The group statistics are very important. In addition to verifying that all of
the subjects were included in the analysis, they let us see which group did better. Reporting a statistically significant difference without knowing in which direction the effect runs is a cardinal sin in statistics! Here we see that the mean reaction time for the “yellow” group is 680 ms while the mean for the “cyan” group is 661 ms. If we find a statistically significant difference, the direction of the effect is that those tested with a cyan background performed better (faster reaction time). The sample standard deviation tells us about the variability of reaction times: if the reaction times are roughly Normal in distribution, then approximately 2/3 of the people when shown a yellow background score within 159 ms of the mean of 680 ms (i.e., between 521 and 839 ms), and approximately 95% of the people shown a yellow background score within $2\times159=318$ ms of 680 ms. Other than some uncertainty in the sample mean and standard deviation, this conclusion is unaffected by changing the size of the sample.

The means from “group statistics” show the direction of the effect and the standard deviations tell us about the inherent variability of what we are measuring.
The standard error of the mean (SEM) for a sample tells us about how well we have “pinned down” the population mean based on the inherent variability of the outcome and the sample size. It is worth knowing that the estimated SEM is equal to the standard deviation of the sample divided by the square root of the sample size. The less variable a measurement is and the bigger we make our sample, the better we can “pin down” the population mean (what we’d like to know) using the sample (what we can practically study). I am using “pin down the population mean” as a way of saying that we want to quantify in a probabilistic sense in what possible interval our evidence places the population mean and how confident we are that it really falls into that interval. In other words we want to construct confidence intervals for the group population means.

When the statistic of interest is the sample mean, as we are focusing on now, we can use the central limit theorem to justify claiming that the (sampling) distribution of the sample mean is normally distributed with standard deviation equal to \( \frac{\sigma}{\sqrt{n}} \) where \( \sigma \) is the true population standard deviation of the measurement. The standard deviation of the sampling distribution of any statistic is called its standard error. If we happen to know the value of \( \sigma \), then we are 95% confident that the interval \( \bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right) \) contains the true mean, \( \mu \). Remember that the meaning of a confidence interval is that if we could repeat the experiment with a new sample many times, and construct a confidence interval each time, they would all be different and 95% (or whatever percent we choose for constructing the interval) of those intervals will contain the single true value of \( \mu \).

Technically, if the original distribution of the data is normally distributed, then the sampling distribution of the mean is normally distributed regardless of the sample size (and without using the CLT). Using the CLT, if certain weak technical conditions are met, as the sample size increases, the shape of the sampling distribution of the mean approaches the Normal distribution regardless of the shape of the data distribution. Typically, if the data distribution is not too bizarre, a sample size of at least 20 is enough to cause the sampling distribution of the mean to be quite close to the Normal distribution.

Unfortunately, the value of \( \sigma \) is not usually known, and we must substitute the sample estimate, \( s \), instead of \( \sigma \) into the standard error formula, giving an
estimated standard error. Commonly the word “estimated” is dropped from the phrase “estimated standard error”, but you can tell from the context that $\sigma$ is not usually known and $s$ is taking its place. For example, the estimated standard deviation of the (sampling) distribution of the sample mean is called the standard error of the mean (usually abbreviated SEM), without explicitly using the word “estimated”.

Instead of using $1.96$ (or its rounded value, 2) times the standard deviation of the sampling distribution to calculate the “plus or minus” for a confidence interval, we must use a different multiplier when we substitute the estimated SEM for the true SEM. The multiplier we use is the value (quantile) of a t-distribution that defines a central probability of 95% (or some other value we choose). This value is calculated by the computer (or read off of a table of the t-distribution), but it does depend on the number of degrees of freedom of the standard deviation estimate, which in the simplest case is $n - 1$ where $n$ is the number of subjects in the specific experimental group of interest. When calculating 95% confidence intervals, the multiplier can be as large as 4.3 for a sample size of 3, but shrinks towards 1.96 as the sample size grows large. This makes sense: if we are more uncertain about the true value of $\sigma$, we need to make a less well defined (wider) claim about where $\mu$ is.

So practically we interpret the SEM this way: we are roughly 95% certain that the true mean ($\mu$) is within about 2 SEM of the sample mean (unless the sample size is small).

The mean and standard error of the mean from “group statistics” tell us about how well we have “pinned down” the population mean based on the inherent variability of the measure and the sample size.

The “Independent Samples Test” box shows the actual t-test results under the row labeled “Equal variances assumed”. The columns labeled “Levene’s Test for Equality of Variances” are not part of the t-test; they are part of a supplementary test of the assumption of equality of variances for the two groups. If the Levene’s Test p-value (labeled “Sig”, for “significance”, in SPSS output) is less than or equal to 0.05 then we would begin to worry that the equal variance assumption is violated, thus casting doubt on the validity of the t-test’s p-value. For our example, the Levene’s test p-value of 0.272 suggests that there is no need for worry about
that particular assumption.

The seven columns under “t-test for Equality of Means” are the actual t-test results. The t-statistic is given as 0.30. It is negative when the mean of the second group entered is larger than that of the first. The degrees of freedom are given under “df”. The p-value is given under “Sig. (2-tailed)”. The actual difference of the means is given next. The standard error of that difference is given next. Note that the t-statistic is computed from the difference of means and the SE of that difference as difference/(SE of difference). Finally a 95% confidence interval is given for the difference of means. (You can use the Options button to compute a different sized confidence interval.)

SPSS (but not many other programs) automatically gives a second line labeled “Equal variances not assumed”. This is from one of the adjusted formulas to correct for unequal group variances. The computation of a p-value in the unequal variance case is quite an unsettled and contentious problem (called the Behrens-Fisher problem) and the answer given by SPSS is reasonably good, but not generally agreed upon. So if the p-value of the Levene’s test is less than or equal to 0.05, many people would use the second line to compute an adjusted p-value (“Sig. (2-tailed)”), SEM, and CI based on a different null sampling distribution for the t-statistic in which the df are adjusted an appropriate amount downward. If there is no evidence of unequal variances, the second line is just ignored.

For model assumption checking, figure 6.6 shows separate histograms of the residuals for the two groups with overlaid Normal pdfs. With such a small sample size, we cannot expect perfectly shaped Normal distributions, even if the Normal error model is perfectly true. The histograms of the residuals in this figure look reasonably consistent with Normal distributions with fairly equal standard deviation, although normality is hard to judge with such a small sample. With the limited amount of information available, we cannot expect to make definite conclusions about the model assumptions of normality or equal variance, but we can at least say that we do not see evidence of the kind of gross violation of these assumptions that would make us conclude that the p-value is likely to be highly misleading. In more complex models, we will usually substitute a “residual vs. fit” plot and a quantile-normal plot of the residuals for these assumption checking plots.
In a nutshell: To analyze a two-group quantitative outcome experiment, first perform EDA to get a sense of the direction and size of the effect, to assess the normality and equal variance assumptions, and to look for mistakes. Then perform a t-test (or equivalently, a one-way ANOVA). If the assumption checks are OK, reject or retain the null hypothesis of equal population means based on a small or large p-value, respectively.