

3 Comparing Two Treatments

Where are we going?

In this chapter we really get started with the basics of experimental design. Specifically we'll learn how to test if the observed *difference* between the means of two groups can be distinguished from chance variation, and how to develop a confidence interval for that difference. In the last chapter we met the first weapon of experimental design, **replication**. In this chapter we'll meet the other three: **randomization**, **control** and **blocking**.

In the last chapter, we made statements about a single batch or sample of numbers. Now, we turn our attention to questions about comparisons between two groups. Perhaps the most fundamental question in any scientific investigation can be phrased as some variant on the following:

Is my experimental treatment better than another?

A **treatment** is an experimental condition that can be assigned to the individuals (humans, rats, plots of land, *etc.* in the experiment

What's the problem?

Let's go back to our RUX-7000 experiment. When we last left our heroes they were trying to decide if the new formulation for RUX-7000 performed better (on average) than the standard 20 mpg. But, how do we know that the RUX-7000 test was fair? We didn't give the standard formula a chance to go *head to head* against RUX-7000. Instead, we assumed its performance from past information, information that may have been based on very different conditions. How do we know that the engines we tested wouldn't have given the standard formula better mileage than 20 mpg as well, under similar conditions?

In order to ensure that our new additive is better than the standard, we should test both under *similar* conditions. If not, maybe our conditions were too favorable; maybe our experiment was biased toward the new additive. In spite of our best efforts to be fair, it is hard not to introduce bias toward the results that we want. Even the most well-intentioned researchers consciously or unconsciously find optimal conditions under which to test and develop their new products. (Just walk into any research laboratory. How many of them approximate the same

Why isn't it a waste of money to run a known standard at the same time as our experimental treatment?

3 Comparing Two Treatments

The subjects in a **control group** are given either no treatment, or a standard treatment so that the treatment of interest can be compared to the standard under the same conditions.

conditions found on the shop floor? The clean room may be fine for developing a new product, when we don't want the noise (literal and figurative) and dirt to interfere with our research. But when it comes time to *test* our new product, we might want to level the playing field again. We need to test it against the competitor(s) under the *same conditions*.)

We need to establish some sort of **control** against which we will compare the results of our new product. The idea of a control is one of the most fundamental ideas in experimental design and has helped scientific experimentation immeasurably since its introduction in the late 19th century. The experimental units that get the control is called the **control group**.

Ways to avoid bias

A **placebo** is an treatment designed to have no effect, but with human subjects, designed to look like the experimental treatment so that the subjects won't know which treatment they're receiving. The **placebo effect** is the name for the tendency of subjects to respond in the way they think they should to an experimental treatment.

Bias can creep into your experiment from a variety of sources in spite of your best intentions, especially when your measurement is subjective and/or your subjects are humans. Doctors have long been aware that people may begin to feel better just because they think they are getting a treatment. A study on productivity at a major corporation allegedly found that increasing the wattage of all the lights in their facilities increased production by 10%. Later it was discovered that *lowering* the wattage by the same amount also increased production by 10%. It turned out that nearly *any* small change in lighting, accompanied by an announcement that management was looking for an effect on productivity, produced a positive effect *in the short run*. However, as time went on, and the psychological benefit of management's scrutiny wore off, the productivity returned to previous levels. Such effects are known as **placebo** effects.

When the possibility of placebo effects is present, it is important to measure the effect of any treatment against a **control group**, in which no real treatment was given, but in which the subjects have no knowledge of whether they received a treatment or not. This is called giving the subjects a **placebo**.

When subjects are unaware of what treatment (if any) they are getting, they are referred as *blinding* the subjects. In addition, it is often beneficial to have the person measuring the effect of the treatment to be unaware of what treatment the subject received in order to eliminate any possible bias in the measurement. In this case, the evaluators are said to be blinded as well and the experiment is referred to as a **double-blind** experiment.

Even when your subjects are not human, it is important to use a control group to evaluate a potential treatment in order to eliminate biases.

Let's use a control and redo the experiment. First question -- how do we do it? How do we make sure that the comparison is fair? We will spend a lot of time talking about various strategies in this book. To start, we'll use a very simple strategy of just adding 10 more engines to the experiment, using 20 engines in total, 10 for the control and 10 for our new formula. This is certainly not the cleverest or most efficient design for this experiment. There are many other alternatives that we will discuss later in the book, but let's keep things simple for now.

To illustrate some of the concepts we'll need, let's imagine that instead of a bank of identical test engines, we only have a parking lot full of cars to test the gasolines. Okay, then, but given 20 cars, which ones get the new additive and which get the control?

Randomization ensures that on the average the conditions of the experiment are similar by using a chance device to assign individuals to the different treatment groups.

Scientists at the beginning of the century argued, that in the best interests of science, they should be the ones to decide which subjects should get which treatment since, after all, they knew the most about the treatment and the subjects' potential responses to it. This was known as a **judgement sample**. We now know that the best mechanism to eliminate such potential biases *on the average* is to **randomize** the experimental units to the two groups. I emphasize here that this is only the best strategy only *on the average*. In small samples, it may *not* be best to trust randomization to make things fair. You may be asking yourself, what could be fairer than deciding who gets what treatment on the basis of flipping a fair coin? Well, suppose I was faced with making up two soccer teams of size 5 from a group of 10 boys. In this group of 10, suppose that there are eight 7 year olds and two 12 year olds.

Randomization would imply that, on the average, I would do best by picking all 10 names out of a hat to get the fairest teams. I doubt very many people would be happy with the results. The chances of both 12 year olds winding up on the same team is just too great (in fact it's just a little less than $1/2$). A much more reasonable idea would be to place one 12 year old on each team, drawing the rest of the names out of the hat (unless even more information were available on the skill levels of the remaining 8). This idea of *purposefully* rather than *randomly* making the groups fair lies at the heart of the concept of **blocking**. Blocking is the central idea of chapter 5 and is an important alternative to simple randomization. We'll come back to blocking in great detail then. For now, though, we'll assign the treatments *at random* to the 20 engines.

Blocks are groups of similar units. When using blocking, conditions are assigned to the units at random within each block.

Now that the cars have been randomly assigned to the two groups, what order should they be tested? Again, we will do it at random. Why

3 Comparing Two Treatments

might it not be a good idea to test all the cars with standard gas before we test all the cars with our new additive? What kinds of *biases* might creep into the experiment? Randomizing the order of the experimental runs attempts to alleviate the effects of these biases by making them come out evenly *on the average*. Other ways to ensure fairness for the two methods include enforcing uniformity of conditions *controlling* the conditions explicitly, or purposefully changing the conditions, but making sure they are balanced (blocking).

The four main weapons of experimental design are **replication**, **randomization**, **control** and **blocking**.

Different aspects of the experiment require different strategies. For example, we might control the driver by keeping the driver the same throughout the experiment. We might “control” the weather by running the experiment indoors. Or, we might use randomization as a strategy for the weather by randomizing the order and hoping that the weather effects will come out roughly evenly on the average. Alternatively, we might use blocking to ameliorate the weather influence, performing the same number of runs of both groups in a number of different weather conditions.

In our experiment, we've decided to use 20 cars on which we'll test the gasolines. We've assigned 10 of them at random to the control group (standard gasoline) and 10 to the RUX-7000 group. We have decided to run the cars in random order. (To ensure this, we might label the cars 1 through 20 and then pick numbers 1 through 20 out of a hat.) After finishing the experiment, the data might look like the data in the table below:

	Regular	RUX-7000
	14.2	36.2
	21.9	11.5
	44.8	5.2
	9.5	10.1
	16.2	40.2
	38.7	17.1
	9.2	10.6
	5	45.5
	11.1	21.7
	35.4	16.1
Average	20.6	21.42
Std. Dev	14.067	14.161

Table 3.1 Twenty runs of an experiment testing regular gas against RUX-7000.

Now it's time to analyze the data from the experiment. Invoking DTDP, let's first look at a boxplot of the two formulations

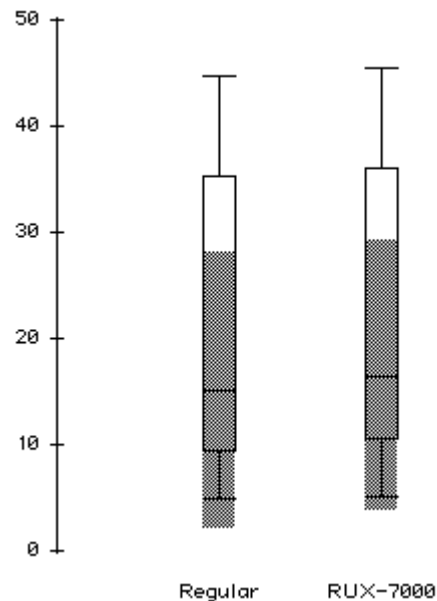


Figure 3.1 A boxplot of the mileage from the two groups. Is RUX-7000 clearly better than regular? (The bar in the middle of the box is the median of each group.)

Hmm.... Do the two gasoline's performance *look* different? Not really. We would be hard pressed to convince anyone that RUX-7000 delivers consistently more mileage on the average on the basis of this plot. Here's an experiment you could do yourself: go down to your local Walmart and collect a random group of people from aisle 11. Show them the boxplots and ask how much more they would be willing to pay for RUX-7000! What we will see below is that the statistical test associated with this picture simply confirms this impression. (This is the way statistical tests *should* work. Generally they will enhance your intuition -- not work against it. If they do, then either something is very interesting, or just plain wrong).

Later, when we perform the statistical test, we'll see that this difference of 0.82 mpg will turn out to be statistically indistinguishable from 0. Or, as it is usually described: "the difference is not *statistically significant*". However, we should note that for this sample RUX-7000 did deliver 0.82 more mpg on the average, or about a 4% increase in the performance. We've decided that this difference of 0.82 is not "statistically significant".

3 Comparing Two Treatments

But suppose our marketing department tells us that each percent improvement translates into about \$10,000,000/year in increased profit, so that 0.82 translates to just over \$8,000,000. What does this say about “statistical” versus “practical” significance?

The question about whether 0.82 mpg is significant is really two questions. The first is a statistical one. Having observed a difference of 0.82 mpg in the averages of two batches, can we conclude that this difference is real, or is it just due to the random variation of the engines, the measurement process, the phases of the moon etc.? That is, given that observations vary, is the difference of 0.82 mpg that we see really distinguishable from 0?

The second question is not a statistical one at all: is a difference of 0.82 mpg *important*, from a scientific, economic or any other viewpoint? The relevance of this question does not depend on the outcome of the statistical test. It’s an equally important question, but it lies outside our abilities as statisticians to answer.

Let's try to answer the statistical question. In looking at the data in Figure 3.1 and Table 3.1, why are we reluctant to put money on the statement that RUX-7000 really increases mileage? Because, the cars performances vary so much it’s hard to see if the difference is real. In the presence of so much noise, we don't believe that RUX-7000 will deliver more mileage *consistently*. Yes, it did it for these 20 engines, but what about the next batch of 20? We get the feeling by looking at the box plot that it is difficult to say what it might do next time. It might deliver 1.2 mpg *less*, next time, or 2.3 mpg more, or 0.54 mpg less, or in fact, almost anything.

What our intuition is telling us is that there's not much evidence here to say anything about how the treatments differ. We can't conclude that RUX-7000 outperformed Regular gas. (Nor can we really conclude that it didn't!) Before doing a formal hypothesis test on these data, let's develop our intuition further by looking at what might have happened if we had *controlled* the type of car used and run the experiment on 20 identical test engines. This experiment might produce data like:

Regular	RUX-7000
20.83	21.39
20.61	21.38

	20.6	21.51
	20.68	21.63
	20.49	21.33
	20.38	21.38
	20.72	21.19
	20.6	21.45
	20.45	21.37
	20.64	21.57
Average	20.6	21.42
Std. Dev.	0.132665	0.126139

Table 3.2. Another experiment from two groups. This time the engine type has been controlled. Note the change in standard deviations as compared with Table 3.1.

The box plots now look like this:

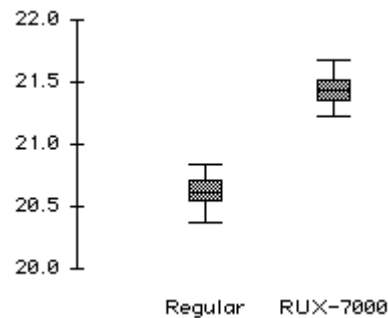


Figure 3.2. How does RUX-7000 look now?

Go back to aisle 11 and ask the same group to look at figure 3.2. Most people would be convinced that the difference in the means of these two groups is now real. Why does the difference look more “statistically significant” now? Simply because the difference between the groups is large compared to the variation. The “signal” (the difference between the group averages) is loud compared to the “noise” (the variation around the averages).

This is not only the basis of most people's intuition in looking at plots like this, but it is exactly the idea behind the statistical test. Notice the changes in the standard deviations as well as in the box plots. In table

3 Comparing Two Treatments

3.1, the standard deviations are about 200 whereas in table 3.2 they're around 0.13 . This is why the box plots in figure 3.2 are so separated as compared to those in figure 3.1. This example is an exaggeration, but it shows the effect that controlled variation can have.

Recall from the last chapter that in testing the hypothesis, we calculated a test statistic (the sample average) and saw where it lay on an appropriate reference distribution. This gave us the *p-value* of the statistic which we interpreted as the degree of evidence against the null hypothesis. The hypothesis test here does the same thing. Now, the null hypothesis is that the treatment groups (treatment vs. control) have the *same mean*. We again compute a test statistic, the *difference between the sample averages*, and see where it lies on an appropriate reference distribution. If it looks unusual enough (*i.e.* it has a small *p-value*), this is evidence for rejecting the null hypothesis. If the *p-value* is large, then it's a typical value given the null hypothesis and provides no evidence against it. In this case we would retain the null hypothesis.

Of course, all of this still says nothing about whether the observed difference is of any practical importance. It still takes \$\$, not statistics, to tell us that. For the two groups here, notice that the difference, as "obvious" as it is, is the same as it was in the first example,-- 0.82 mpg. It's important to keep the two questions about significance (practical and statistical) separate.

Exercises

1. A study, by a leading consumer organization, is considering the design for an experiment on a new spot removing detergent for clothes. They'd like to see whether it's really better at removing stains than a leading detergent as claimed by the manufacturer. To run the test, a measured amount of stain is to be put on 16 white t-shirts, which after washing, are measured for whiteness with an optical instrument. You are a researcher for the consumer group
 - (a) How would you run the experiment? Keep in mind the issues of a control group, blinding, and randomization.

After your report is read, several of your colleagues have ideas for improving the experiment that you have designed. Comment on each of the following ideas:

- (b) All 16 t-shirts should be used on the spot removing detergent and the data compared to previous data gathered on the standard detergent.
- (c) All 16 runs should be done in very hot water with a 30 second washing time to speed up the experiment.
- (d) A previous study demonstrated that soft water is better for washing clothes. Therefore, use only soft water for this experiment.
- (e) If we have only one washing machine, run all the standard detergent runs first to keep things simple.

The t -Test: The practical application

The statistical test used to formally test the hypothesis that two samples have the same mean is a numerical version of what our eye does in looking at two side by side box plots. Let's think for a minute why the means of the two groups in Figure 3.2 are so obviously different while it's almost impossible to say anything about the means of the two groups in Figure 3.1. Our intuition says that the key factor is the *ratio* of how big the difference between the two means is compared to how much variation there is within the groups. This ratio is what statisticians call the *signal to noise ratio*. Here, the signal is the difference of the two group averages: $\bar{y}_1 - \bar{y}_2$. The noise has to do with how much that difference varies (its standard error) which, in turn, is related to the standard deviation of the observations within each group. (The hard part of the formal test is computing that standard error). So, for the same difference in the averages, the more the observations in each group vary, the less confident we feel that the two means are really different. That's the difference in the two figures.

With two samples, the usual null hypothesis will be that the difference in means is 0. The test statistic is the difference in the two averages divided by its standard errors. This ratio then tells us how many standard errors away from 0 the difference in the two sample means are. The reference distribution for this ratio is again the t distribution, as in the one sample test, although with a different number of degrees of freedom. So, if the ratio is larger than about 2, this is still a rough guide of where to begin to cast doubt on the validity of the null hypothesis. Of course, the exact *p-value* will depend on the number of degrees of freedom. And what you decide about the null hypothesis depends on your α level (and your confidence in the data, your belief in the null hypothesis and any number of unquantifiable factors). For now, let's not worry about *how* the standard error of the difference is calculated. Instead, let's build up some expertise in interpreting the output of the test, and when to use it.

In both data sets (figures 3.1 and 3.2) the signal, the difference between the two averages is the same in absolute size: 0.82 mpg. If the null hypothesis of equal means is true, then the test statistic, the difference in

the averages of the two groups, should be 0. Of course, observations vary, so it will never be exactly 0 even if the two groups really do have the same mean. But the farther from 0 the difference in the two sample means gets, the more doubt it casts on the null hypothesis. What the standard error (the denominator) and the reference distribution does is to put all possible differences between two group averages on the same map. By looking at the ratio of the difference to its standard error, we translate the observed difference to standard error units. While it's impossible to say whether an observed difference of 6.5 microns is big, knowing that it's 10 standard errors away from 0 tells us a lot. Whether it turns out to be scientifically significant or not, it tells us that the observed difference didn't occur by chance and so gives us quite a bit of evidence that the true difference is not 0.

Here's the output of the t-test for the experiment on the 20 different cars:

T-test of hypothesis that means are equal

	Difference	t-statistic	Degrees of freedom	Prob > t
Estimate	0.82	0.130	18	0.898
Std. error	6.312			
Lower limit of 95% conf int	-12.441			
Upper limit of 95% conf int	14.081			

Figure 3.3 A hypothesis test that the two groups have the same mean. Notice howwide the confidence interval is and that it covers zero.

For this experiment, the difference between the two group averages is 0.82, with a standard error of 6.312. Translating to standard error units gives $0.82/6.312$ or 0.13 standard errors. In other words, RUX-7000 performed 0.13 standard errors better than Regular. How does this strike you? Is this an impressive increase?

Let's formally *test* the null hypothesis. The p-value is 0.898. What this means is that if the null hypothesis is true, and the difference in means of the two groups is really 0, then we would expect a difference at least as large as 0.82 by chance about 89.8% of the time. In other words, even if the null hypothesis is true, these are very likely and reasonable data. So, there is essentially *no* evidence here to suggest that the null hypothesis is false.

3 Comparing Two Treatments

Does this *confirm* the null hypothesis? Does this experiment *prove* that the two formulations give the same mean performance? No!! It just says that we have no evidence against it. (Does failing to prove someone is guilty in court prove their innocence?). Let's see what additional information we can get from the confidence interval.

Remember, our best guess of the difference between the group means is 0.82 with a standard error of 6.312. So, using the rule of thumb factor of 2 to make a very rough 95% confidence interval, we can say that the true difference between the means of the two groups lies in the interval: $0.82 \pm 2 * (6.312) = (-11.804, 13.444)$ with roughly 95% confidence.

Now, because our sample size is small -- 10 observations in each group -- the actual confidence interval is even a little wider than this back of the envelope calculation. We really should use the $t_{.025}$ value with 18 degrees of freedom, which is 2.101. So, the more precise 95% confidence interval found by the software is: (-12.44, 14.08), the one shown in figure 3.3

What does this interval mean? What kind of a statement can we now make about the performance of RUX-7000? All we can say is that with 95% confidence, the performance of RUX-7000 is somewhere between 14.08 mpg better and 12.44 mpg worse than Regular!

This is almost worse than useless. I probably could have guessed that almost any formulation would deliver somewhere in that range. All we can really conclude is that our experiment didn't contain much information! Why was that? It was because our experimental design was so poor. It used 20 different cars whose mileage varied far too much from each other to see anything. In chapter 5 we'll investigate more suitable designs for this problem.

What, if anything, then, can we conclude from this experiment? Obviously, the confidence interval is much too large to be of much use in any practical decision making. So, how could we narrow the confidence interval? The brute force method would be to increase the sample size. Although the standard error of the difference is more complicated than the standard error of one average (see the next section), it is still proportional to $1/\sqrt{n}$, the square root of the sample size. So, increasing the sample size will lower the standard error by this amount and thus proportionally decreasing the size of the confidence interval. As an example, an increase of 4 times as many observations will cut the length of the confidence interval in half. Our current confidence interval is plus or minus about 13 mpg. To get it to one tenth this big (plus or minus about 1.3 mpg) would take 100 times as many

observations (2000 cars!). But, observations are usually expensive, so, this is not usually the best strategy. Another way to improve the efficiency of the experiment is to *control* some of the other factors that are making the observations vary. This will reduce the standard deviation and, in turn, the standard error of the difference. The main problem with this experiment was that the 20 cars performed so differently. That's why we also discussed using very similar test engines. These are not only very similar to each other, but by using test engines, we avoid all the real life problems of drivers, weather, tire pressure and so on. The drawback of this strategy is twofold. First, we may not be able to generalize from these engines to all types of cars, and secondly, controlling many factors may not give a *realistic* idea of how the additives will perform in real life. But, for *comparing* two formulations, it does reduce the noise and may answer the question of how much better one is than the other.

If we use these test engines, we might get data similar to those shown in table 3.2. Now, just from looking at the side by side boxplots of these data (figure 3.2), it's pretty obvious that the means are different. You probably don't need a Ph.D. in Statistics to tell that these two groups have different means.¹ But, just to make sure, we'll perform the t -test:

T-test of hypothesis that means are equal

	Difference	t-statistic	Degrees of freedom	Prob > t
Estimate	0.82	14.165	18	3.347 e-11
Std. error	0.058			
Lower limit of 95% conf int	0.698			
Upper limit of 95% conf int	0.942			

Figure 3.4. T-test of the hypothesis that the means are zero from the controlled experiment on test engines. The difference is pretty obviously significant. Notice how small the standard error is and the range of the confidence interval.

¹ When the results are this obvious, we sometimes say that we reject the null hypothesis using the "ocular trauma test of significance" since the difference hits you right between the eyes.

3 Comparing Two Treatments

Compare the p-value and the width of the confidence interval with those in figure 3.3. Notice that the difference between the group averages is still 0.82mpg. But this time, its standard error is about 100 times smaller. Why? Because we controlled everything under the sun, and most importantly, the engine type. This, in turn, lowered the standard deviation in each group from about 14 mpg to about 0.13mpg, a factor of about 100.

By virtue of eliminating many sources of noise, this experiment has given us a much more precise estimate of the difference between the means. The confidence interval reflects this by being much, much narrower: (0.698, 0.942) mpg. In terms of its standard error, the difference, 0.82 mpg, is now 14.17 standard errors away from 0 (0.82mpg divided by 0.058 mpg). Compare this to the previous experiment where 0.82 mpg was only 0.13 standard errors away. Exactly how unusual is 14.17 standard errors? For adult U.S. males, (mean 69 inches, standard deviation 2.5 inches, this corresponds to a man about 8' 8" tall.

How likely is a difference of 14.17 standard errors to occur by chance? Not very -- in fact, the probability is $3.347 \cdot 10^{-11}$ (or about 3 and a half

Now, let's get back to the experimental design question. We controlled everything in sight in order to reduce the noise. What did we lose? Well, we tested only one engine type, so perhaps our inference is restricted to this type of engine. Secondly, we simulated the driving. Our simulation does not reflect actual driving conditions. We may want to use blocking as a compromise between these two designs, and in Chapter 5 we'll revisit this experiment.

For the rest of this chapter we'll look at the formulas for the t-statistic and the confidence interval to get some more insight into the analysis. signal to noise ratio. In the next section, we will satisfy the curiosity of those out there who want to see all the details and derivations. (The rest can skip that section).

Exercises

2. One of the cheapest ways to improve fuel efficiency in automobiles is to keep the tire pressure from getting too low. In order to quantify this effect, Wayne Collier devised the following experiment: To compare the effect of increased tire pressure, I chose two values, one at 24 psi, the suggested value given by the auto manufacturer, and the other at 38 psi, the value that I generally use on long distance trips. In order to make the measurements of mileage more accurate and easier to obtain, I made some alterations in [sic] the normal flow of gasoline to the engine. I inserted a T-junction into the fuel line just before the fuel filter, and ran a line into the passenger compartment of my car, where it joined with a graduated 2 liter Rubbermaid[®] bottle that I mounted in a box where the passenger seat is normally fastened. Then I sealed off the fuel-return line, which under normal operation sends excess fuel from the fuel pump back to the fuel tank. Valves on the two inputs to the T-junction allowed me to drive until I was in the correct location to begin the runs. To perform the tests, I chose a section of four-lane highway about 12 miles long, between the fire districts of Mills River and Pisgah Forest, North Carolina. I decided to do the tests after dark, since the traffic on the road would be reduced to about 2 cars per mile, and the weather conditions were more likely to remain constant. After arriving at the location where I was to begin the run, I closed off the fuel line from the main tank, opened up the line to the small bottle in the passenger compartment, and let the engine run until it drained the fuel line and ran out of gas. To keep the accelerator pedal in a constant position, I attached a block of wood between it and the floorboard."

3 Comparing Two Treatments

Wayne's data appear in table 3.4 along with some summary statistics.

	Miles Travelled	Tire Pressure	Run Order	Miles Travelled	Tire Pressure	Run Order
	9.3	Low	10	10.55	High	5
	9.85	Low	2	9.8	High	14
	9.7	Low	7	10.3	High	11
	9.9	Low	13	9.6	High	12
	9.4	Low	16	9.8	High	4
	9.65	Low	9	10.2	High	6
	9.4	Low	8	9.7	High	15
	9.7	Low	1	9.9	High	3
Average	9.6125			9.98125		
Std dev	0.2216			0.3316		
Difference of aveages	0.369		Std error of difference	0.1410		

Table 3.4 Data from Wayne's tire pressure experiment.

- What factors did Wayne control? What factors did he randomize?
- What factors did Wayne not consider that might affect the fuel efficiency?
- How might changes in those factors during the experiment affect Wayne's results?
- State the null and alternative hypotheses. Is the alternative one or two sided?
- Test the null hypothesis at $\alpha=.05$.
- Construct an approximate 95% confidence interval for the difference in mean miles travelled under the two tire pressures.
- What assumptions about the data have you made in order to answer (d) and (e)? Do they appear reasonable given the data?

3. An experiment to test battery life was designed by Bob Osborne. To make the experiment simple, he used a discman© CD player with the same CD running continuously (using the repeat key). He tested six

pairs of AA alkaline batteries from two major battery manufacturers, a well-known brand name and a generic brand. For each trial he fixed the volume control to 5 and waited until no more music was heard through the headphones. (He ran an initial trial to find out approximately how long that would take, so that he didn't have to spend the first 3 hours listening to the same CD). The response measured was the time it took until no more music was heard. The data are shown in table 3.5.

	Brand Name	Generic
	169.5	220.7
	205.5	233.5
	184.0	233.5
	172.4	236.5
	194.0	252.5
	199.2	239.4
Average	187.433	236.017
Std dev	14.611	10.302
Difference	48.583	
Std error of Difference	7.299	

Table 3.5. Data from Bob Osborne's Battery experiment in minutes until batteries ran out of power on CD discman.

- State the null and alternative hypotheses. Is the alternative one or two sided. If one sided, which way?
- Bob decided to control several factors. Discuss how this might limit the inference that one could draw from the results of his experiment.
- Test the null hypothesis at $\alpha=.01$.
- Construct a rough 95% confidence interval for the true mean difference in time the batteries will last under the conditions of Bob's experiment.
- Given that the brand name batteries cost 33% more than the generic ones, does the difference found seem to be a financially significant one?