

4 More than two Treatments

Introduction

In the last chapter we saw how to compare the means of two groups. Here we extend the same ideas to the case where the results from more than two levels of an experimental factor are to be compared. We will build on both the design and analysis tools of the previous chapters. The design tools we have already met, **replication**, **control** and **randomization** will continue to play a fundamental role in our experimental designs. The fourth tool, **blocking**, will be the subject of Chapter 5.

For the analysis, we will extend the t -test to what is known as the Analysis of Variance (ANOVA), a tool that we will continue to use for the rest of the book. Because the ANOVA is so important, we will take several takes on it, explaining it from different points of view.

The central idea is to build on the structure of the t -test by again constructing a signal to noise ratio that compares the variation in the means across several treatment levels to the error variation. What's different about 3 or more groups? What do you replace the numerator, $\bar{y}_1 - \bar{y}_2$ with for 3 groups?

In 1925, when Sir Ronald Fisher discovered the answer, it was certainly not obvious. It was his insight that led to the statistical test that we will describe in this chapter.

Specifically, in this chapter, we will:

- test if differences in means across several treatment levels are distinguishable from noise.
- quantify how big the differences are.
- make decisions as to which treatment level is best. This decision depends not only on the statistical evidence: the observed differences in the means and its associated p -value, but on real world considerations like the *cost* of the treatment levels and many other factors not directly measured in the experiment (reliability of vendors, stability of treatment etc.).

4 Comparing more than two treatments

Let's start with an example. A market researcher wants to test four different prepared scripts for use in a telephone solicitation at an 800 number call center. Callers use the 800 number to ask questions about their accounts, but the call center uses the opportunity to inform the customers of current offers they might be interested in. This is a polite way of saying that they use the call to initiate a sales pitch. To be successful, not only should a script result in a reasonable number of sales, but it must also be efficient. That is, it shouldn't take too long to get through, or the customer will hang up. So, for this preliminary investigation, we'll look at the amount of time it takes a solicitor to read one of several prepared scripts. The researcher has 24 telephone solicitors available, and so, decides to assign 6 of them at random to each script by drawing names out of a hat. After a few days of training, she starts the test. To minimize the disruption to the call center, she conducts the test during the normal working day of the solicitors. Each solicitor is told to start reading the script to the incoming caller as soon as the caller's initial question has been answered. As usual, phone numbers are randomly given to the solicitors as they come in. In order to reduce the possibility of a time of day bias in the experiment, each solicitor starts the test at the same time.

The results, the amount of time it takes to read the script in the 24 calls are shown in table 4.1, arranged by the script used.

Script	A	B	C	D
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
	63	65	68	63
	59	66	68	64
Averages	61	66	68	61

Table 4.1 Times in seconds to read script

The first question is: do the different scripts have different mean reading times? The *obvious* answer is: of course they do. Just look at the data. Scripts B and C took longer on the average than A and D. But, the statistical question asks: how will I be sure that this is likely to happen if I run the experiment again? Will scripts B and C continue to take more time? Were the observed differences really due to the different scripts? Couldn't the observed differences be attributed to less systematic factors like the different phone numbers each solicitor happen to get by chance, or some characteristics, like age or experience, of the different solicitors themselves? If I make a recommendation to the marketing department as to which script to use, based on this experiment, will my results hold true when given to the entire group of future solicitors?

Don't forget:

OBSERVATIONS VARY!!!

What appears to be a difference due to a systematic change that we introduce (the factor level) may actually be the result of chance variation due to changes that we did not control. The statistical analysis attempts to determine whether this is the case or not.

Graphics first

Our first attack on answering the question of whether the treatments are different is, of course, to invoke DTDP. Looking at figure 4.1, it *appears* that scripts B and C take longer on average than scripts A and D. Could this be due simply to chance variation? Moreover, can we conclude that C is worse than B?

4 Comparing more than two treatments

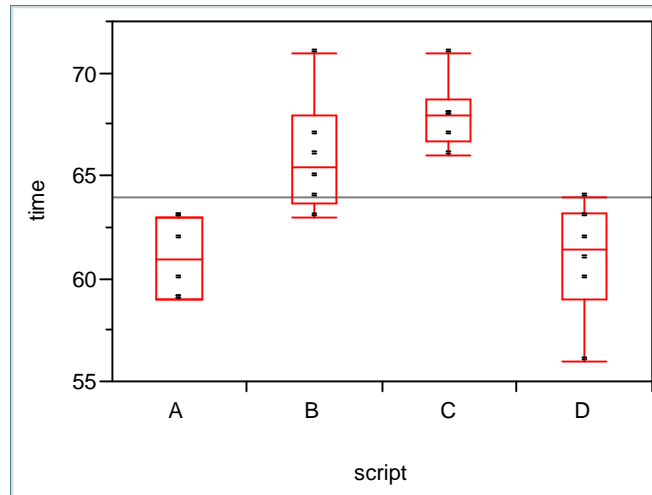


Figure 4.1 Box plot of Times in seconds for 4 Scripts

From figure 4.1, is it *obvious* and *beyond doubt* that the differences between the scripts did not occur by chance? Would your answer change if the picture looked like figure 4.2 instead?

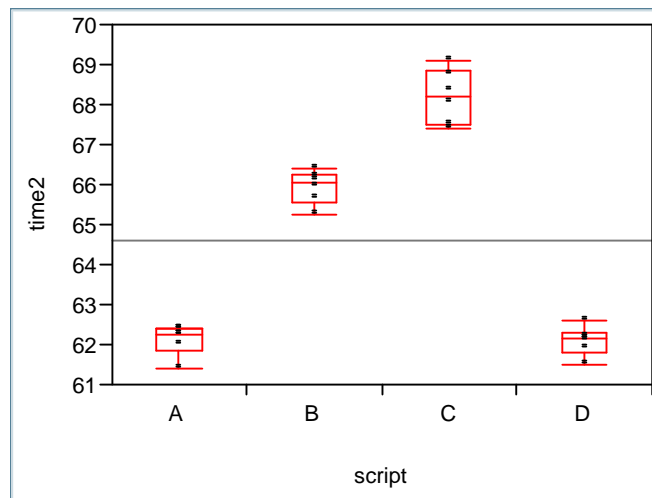


Figure 4.2 Box plot of Times by Scripts. The differences among the scripts are the same as in figure 4.1, but the variation within each script is much less.

Isn't it more obvious now that the differences are real and repeatable and *not* due to chance? What if the picture had looked like figure 4.3 though?

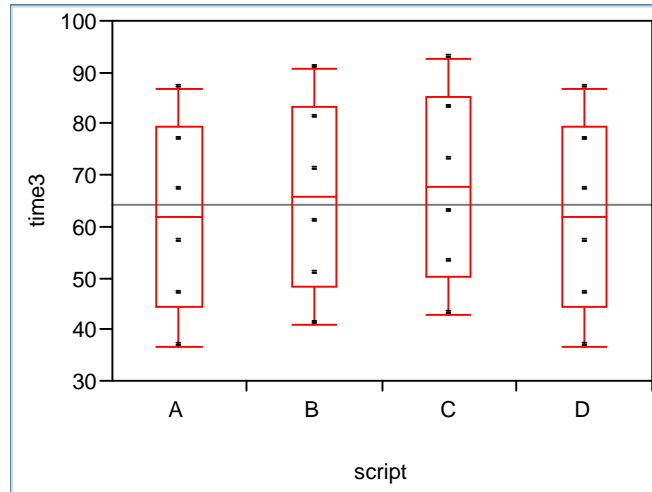


Figure 4.3 Box plot of Times by Scripts. Here the differences among them are still the same, but the variation has been increased.

Notice that the absolute *size* of the differences in all three pictures remains the same. However, the noise (the part not due to the different scripts) changes dramatically. Just as when we compared two groups, we still compare the size of the difference between the treatment levels (the signal) to the spread of the data within each treatment (the noise).

If the difference between treatment levels (scripts in this case) is large enough, you should be convinced that the difference is probably due to the treatments (as in Figure 4.2). Otherwise, you may decide that there's not enough evidence to convince you that the treatment means really are not all the same. In this case you either accept the null hypothesis of no differences between treatment means, or decide to collect more data. Remember, since we start by *assuming* that there are no differences between treatment means, we can never actually *prove* it. All we can do is fail to reject it. But, we can perform a power calculation. This will show us what the probability of rejecting the null hypothesis is for any hypothetical real difference, or *effect size*. So, if we fail to reject the null hypothesis, we can't say the means are really equal, but we can say that had an average difference of say 2 seconds across the four treatments really existed, we would have detected it with a certain probability, like

4 Comparing more than two treatments

90%. We'll come back to the details of how to do such a calculation in a later section.

Putting numbers on it

Before we plow into the formulas for the analysis of variance, let's first look at the output and talk about the concepts behind the test. We'll look at the algebra and details of the test later in the chapter. To get started, some summary statistics from our solicitation experiment are shown in table 4.2:

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
A	6	61.0000	1.89737	0.7746	59.384	62.616
B	6	66.0000	2.82843	1.1547	63.591	68.409
C	6	68.0000	1.67332	0.6831	66.575	69.425
D	6	61.0000	2.82843	1.1547	58.591	63.409

Table 4.2 Summary Statistics from the four scripts

Our null hypothesis is that the means of all the treatment levels are equal. Symbolically, we write:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

The alternative (H_a) is that *any* of the treatment means differ. Looking back at figures 4.1,2 and 3, how do we decide whether we think the means are different? We judge whether the differences *between* the group means is large compared with the variation *within* each treatment. The statistical procedure is based on exactly the same idea. We break up the variation in the experiment into two parts: the part due to differences in the treatment averages (the signal) and the part due to experimental error (the noise). Then, we create a ratio out of these two

quantities. The idea is that, if the treatment differences are large, compared to the experimental error, the ratio will be large enough to reject the null hypothesis.

These quantities are usually displayed in a table, called the ANOVA table, as shown in table 4.3. The variation due to the treatment differences is called the *treatment mean square* and is denoted MS_T , while the variation due to error is called the *error mean square* or *residual mean square*, and is denoted MS_R . The new signal to noise ratio is constructed by dividing the treatment mean square by the error mean square. This produces a test statistic similar to the t statistic of Chapter 3, although now, the reference distribution for the test statistic, when the null hypothesis is true, is no longer the t -distribution, but a distribution called the **F** distribution.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Treatment (Scripts)	3	228.00	MS_T 76.00	F = 13.5714	<.0001
Error	20	112.00	MS_R 5.60		
C. Total	23	340.00			

Table 4.3. The Analysis of Variance Table

In table 4.4, the $MS_T = 76$, while the $MS_R = 5.6$, and so the **F** statistic is 13.57. This **F** statistic, like the t -statistic, becomes larger as the differences between the treatment means increase, since the differences between treatment means are in MS_T the numerator.

The p -value is *interpreted* exactly the same way as before. Given that the null hypothesis is true, the p -value gives the likelihood of getting a test statistic as large (or larger) as the one observed from our experiment. Here, the p -value gives the probability of getting treatment averages at

4 Comparing more than two treatments

least as different from each other as the ones we've observed, given that all the treatment means are really the same. A small p-value means that we've managed to produce some very unlikely data if the treatment means are the same, and so provides evidence *against* the hypothesis of equal means. By contrast, a *large* p-value indicates that the observed differences are *consistent* with the null hypothesis, meaning that if the groups did have the same mean, we would be quite likely to see differences this big just because of chance variation. In other words, with a large p-value, we have *no* evidence against the null hypothesis of equal means.

What does the ANOVA have to say about the data shown in Figure 4.1? Although, we'll examine, in detail, what all the numbers mean in the next section, and the *assumptions* under which the F test is valid, for now, let's just look at the p-value. The statistical question and judgement is whether the differences in sample averages provides enough evidence to reject the hypothesis that the means are equal.

The p-value in table 4.3 is $< .0001$. What this says is that if the mean time of the four different scripts are really equal, the chance of getting differences as large as those seen in figure 4.1 from a sample of 24 is less than 1 out of 10,000.

What does that say about the null hypothesis? It says that **if** the scripts do not affect the mean time, then we got an experimental result that is very unlikely to happen. There are only two possibilities:

- The null hypothesis is true, and the observed differences are just due to chance. The probability of getting data this extreme in this case is the p-value. OR,
- The null hypothesis is wrong.

At this point you have to make a decision as to which statement seems more plausible.

You are the jury and the closing arguments have all been made -- no more evidence. You've got to make the decision. Which of the two possibilities above seems more plausible? The evidence says: ``If the null hypothesis is true, observed differences as large as those obtained

from your experiment have less than a 1 out of 10,000 chance of occurring". Is this *ironclad* proof that the null hypothesis is wrong? Of course not. Something with a 1 out 10,000 chance will occur about 1 out of 10,000 times! But, if we are never willing to let go of the null hypothesis, if there is no p-value small enough to convince us that we should drop the null hypothesis, then we shouldn't bother running the experiment! The strength of the evidence, the p-value at which **you** are willing to be convinced that the null hypothesis is wrong, is a personal decision, and is highly context dependent. At the very least, there should be some consideration of the cost of making the wrong decision (in both directions).

It's also important to recall at this point, that, as tempting as it is to say, the p-value above **does not** say that the probability that the means are equal is less than 1 out of 10,000. What it does say is that if the treatment means are all equal, the probability of seeing differences as large as we got is less than 1 out of 10,000. This sounds very similar, but we just can't switch the probability statement around. The p-value is $P(\text{data as extreme as ours} \mid \text{given that the null hypothesis true})$, whereas the other statement would be $P(\text{The null hypothesis true} \mid \text{given data as extreme as ours})$.

These are *not* the same statement. The probability that the sun comes up tomorrow given that you win the lottery today (a very likely probability), is certainly *not* the same as the probability that you win the lottery today given that the sun comes up tomorrow. (If it were, I would stop writing this book right now and head down to the nearest 7-11).

Of course, we don't look just at the p-value in order to make a statistical judgement. There are assumptions upon which the F-test in the Analysis of Variance is based. If they're not satisfied, at least approximately, then all our conclusions are suspect and possibly very misleading. Let's make this explicit.

WARNING: The F-test in the Analysis of Variance depends on some assumptions about the data that we will discuss in the section called *Assumptions about the Analysis of Variance*.