Chapter 8
Comparing more than two continuous variables: Introduction to ANOVA

James Myers
2022/4/8 DRAFT

1. Introduction

We've discussed a lot of useful statistical techniques so far, ones that you can use in real research reports, from standard deviations to *z* scores to correlations to *t* tests to chi-squared tests. But probably the most commonly used statistical test, especially for experimental data, is one that so far we've only mentioned in passing: **ANOVA**, which stands for **analysis of variance** (called 變異數分析 in Chinese Excel, but 方差分析 in Chinese Wikipedia).

Why is ANOVA so popular? There are two important reasons. The first reason is that it's like a *t* test (so it's a parametric test for continuous, normally distributed values), except that it can compare more than two means at a time. In fact, as we'll see, the *t* test is just a special case of ANOVA. The extra power of ANOVA is useful because it doesn't always make sense to divide your research question exactly in half. For example, suppose you want to compare the effect of first language on something or other, so you compare Chinese native speakers with English native speakers. But why restrict yourself just to those two languages? With an ANOVA, you can include more languages (e.g., English vs. Chinese vs. Navajo). So now we have a **factor**, language, that has multiple **levels** or **treatments** (both terms refer to the subcategories defining the categorical factor; note that "level" here does not imply any hierarchy, and "treatment" doesn't mean that your data must come from a medical experiment). This is sketched in Table 1.

Language		
Chinese	English	Navajo

Table 1. One factor that has three levels

Another reason for its popularity is that an ANOVA can also analyze the **interaction** between two or more factors (called 交互作用 in Chinese Excel, but 互動 in Chinese Wikipedia). For example, suppose you want to know not just the effect of language (Chinese vs. English), but also the effect of gender (female vs. male). Maybe Chinese-speaking men and women behave differently, while English-speaking men and women behave the same, or maybe the Chinese-English difference goes one way for men, but the opposite way for women. In other words, here we have two factors, each with two levels, but the two factors are fully **crossed**, creating four **cells** (or **conditions**).

No problem, ANOVA can analyze all that. In fact, in this case it would give you three results at the same time, testing for the two **main effects**, language and gender, plus their interaction. We've already had a hint of the importance of interactions when we looked at the two-way chi-squared test, which is designed to test the interaction between two factors, but not the main effects of each factor. That's why each combination of factor levels in ANOVA is called a **cell**, since we're actually dealing here with a sort of contingency table (though what we care about now are the cell means, not the cell frequencies). This is sketched in Table 2.

		Language	
		Chinese	English
Gender	female	Chinese female	English female
	male	Chinese male	English male

Table 2. Two crossed two-level factors

Both of these advantages of ANOVA are related, not just mathematically (the secret lies in unpacking the notion of "analysis of variance"), but also in practical terms. This was already made clear by the guy who invented ANOVA, who was, of course, that statistical genius Ronald Fisher. Before he became famous, he got a job at an experimental research station funded by the British Ministry of Agriculture, studying how to raise crops more effectively. Being a genius, he quickly realized that many of the studies that had been run there had not been designed well enough to yield particularly useful answers. He summed up his crucial insight in a paper that helped make him famous:

> No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed. (Fisher, 1926, p. 511)

With his paper, Fisher invented the **factorial experiment**, which is now standard in many areas of science, including linguistics. As I just showed with my examples, many linguistic questions involve comparing multiple categories or looking for interactions. It might even be argued (as in Myers, 2009a, 2009b) that most linguistic hypotheses (at least in certain domains of linguistics) actually relate to interactions, not to main effects. This is because linguistic analyses usually aren't about individual units in isolation (e.g., just nouns), but rather about how they relate to other units (e.g., verb agreement with nouns), and that means that you want to know how the two units interact. Interactions are also important in cognitive science more

generally, where theorists often debate whether two processes (e.g., semantics and phonology) are truly separate. Well, if they don't interact, then they're probably not part of the same process (as argued by Sternberg, 1998; we'll expand on this point below).

Even more generally than all this, once Fisher opened up scientists' eyes to the notion of giving Nature a questionnaire rather than asking just one question at a time, they realized that ANOVA itself limited the types of questionnaires they could ask. Fortunately, just as the *t* test is a special case of ANOVA, it turns out that ANOVA is just a special case of regression, and regression lets you ask many, many questions at once, far beyond what can be addressed in a factorial experiment. I'll explain exactly how this works in a later chapter, but this fact may help explain why most of the rest of this book, after the two ANOVA chapters, focuses on regression. Still, it's not necessarily better to give Nature a complicated questionnaire than a simpler one: the more factors you have or the more levels they have or the more interactions you consider, the more complicated and confusing your results will be. ANOVA is a powerful technique, and regression is even more powerful, but remember Spiderman's motto: "With great power comes great responsibility!"

Wait a minute, did I just say that there are *two* chapters on ANOVA? Yes, that's how important ANOVA is. This first ANOVA chapter focuses on the basic concepts, and explains the kind of ANOVA that generalizes from the unpaired *t* test, that is, the kind of ANOVA used for independent samples, such as collected in a between-groups experimental design. The second ANOVA chapter will discuss the kind of ANOVA that generalizes from the paired *t* test, that is, the kind of ANOVA used for correlated samples, such as collected in a within-groups experimental design.

2. An overview of analysis of variance

Before we try out ANOVA for ourselves, let's get a few basic concepts clear: why we can't just use a bunch of t tests, why variance is so crucial to analysis of variance, and the different kinds of ANOVA and when we might use them.

2.1 The sin of multiple comparisons

Let's start with the most basic question. You've worked hard to see how *z* scores relate to the *z* test, and how the *z* test led to the one-sample *t* test, and how that leads to the unpaired and paired *t* test (and also saw how all of this is kind of related to the chi-squared test too, though that's for categorical data, not continuous data like all of the other things I just mentioned). So what if we have three or four samples now, instead of just two? Why can't we just do a bunch of *t* tests, testing them one pair at a time? For example, why not compare Chinese vs. English,

Chinese vs. Navajo, and English vs. Navajo? Why not compare Chinese men vs. Chinese women, Chinese men vs. English women, etc?

Because that would be very, very bad. Like any statistical test, each time you do a *t* test, its *p* value is calculated on the assumption that this is the only test you did: it doesn't "know" that you also did all the other *t* tests too. That is, the probability is something like P(chance), not the conditional probability P(chance | all those other tests). The *t* test math is carefully designed so that if your data set is large and normal enough, the probability of getting $p < .05$ if the null hypothesis is true really is pretty close to .05. But if you keep running *t* tests on overlapping subsets of the same data (e.g., Chinese vs. English, and then Chinese vs. Navajo), then it gets more likely than .05 that you'll get a result that says "$p < .05$", even though the null hypothesis is correct. So making these **multiple comparisons** increases the risk of a Type I error.

It's actually pretty easy to calculate exactly how unreliable the *p* value becomes in this kind of situation. For example, imagine that you have just one factor with five levels, and you want to compare all of the pairs. How many pairs is that? Well, each pair can include one of five levels, and the other level in the pair has to be different, so it can be one of four different levels, making 5×4 pairs. But that treats the pairs AB and BA as different, when actually order doesn't matter, so we divide the total by two: $5 \times 4/2 = 10$ pairs of levels to compare.

Now imagine that, sadly, the null hypothesis is actually true for all of these pairs: this factor just isn't significant at all. If we have set our alpha level $\alpha = .05$, that means (by definition) that the probability of getting a Type I error for comparison (one *t* test) is $p = .05$. That, in turn, means that the probability of *not* getting a Type I error for that one comparison is 1 - .05 = .95 (since getting a Type I error and not getting a Type I error are the only two possibilities, so their two probabilities must add up to 1, by the addition rule of probability).

But we're not just running this one *t* test on this one comparison, but 10 *t* tests on 10 comparisons. By the multiplication rule of probability, the total probability that *all* of these comparisons give *no* Type I error is $(.95)^{10}$. The opposite of this situation would be where *at least one* of these 10 comparisons gives a Type I error, and since these two possibilities are the only two possibilities, the probability of getting at least one Type I error among these 10 comparisons is $1 - (.95)^{10} = .4012631$.

This means that if we test this 5-level factor using *t* tests, we have a 40% chance of committing a Type I error. That's much higher than our alpha level of .05, which is what the Type I error risk should be for a proper statistical test. And this means that when you run a multiple comparison like this, you just can't trust the *p* values that any of the repeated tests give you. So don't do it!

What we need instead is a single test that looks at all of our data at the same time, and that's just what an ANOVA does. That is, the null hypothesis for an ANOVA looks like the following, with all of the *k* sample means compared all at once:

ANOVA null hypothesis: H_0: $\mu_1 = \mu_2 = ... = \mu_k$

2.2 Analyzing variance

With computers it's pointless to compute ANOVAs by hand (the way Fisher had to), but it's still useful to know basically how they work. Conceptually, this will help you feel confident that ANOVA is logical and justified (not just magic), and will also help you see its connection with regression, which will become crucial in later chapters. In practical terms, knowing how ANOVA works is also important for choosing the right kind of ANOVA for your particular situation, and to find mistakes in your work (and in other's work too, perhaps), and even for getting the syntax right when you run ANOVA in R.

The key idea is in the name: analysis of variance. How can analyzing variance help generalize the *t* test? Like many ideas given to us by geniuses, the core idea is so obvious that it's amazing nobody thought of it before Fisher. Here it is: when comparing multiple samples at the same time, the null hypothesis H_0: $\mu_1 = \mu_2 = ... = \mu_k$ should be rejected if there is "significantly" more variance *between* (組間) the samples (i.e., in the distribution of sample means) than *within* (組内) each of the samples. The variance across the abstract samples $\{\mu_1, \mu_2, ..., \mu_k\}$ is "interesting", but the variance within the samples S_1, S_2, ..., S_k is "boring", so we want to see if the former is bigger than the latter. How can we compare two variances? Hm, do we know any test invented by Fisher that relates to comparing variances...? Something starting with the letter "F" maybe...? Ah yes, the *F* test: we divide the "interesting" between-sample variance by the "boring" within-sample variance, and if the resulting *F* value is "big enough", we got ourselves a significant ANOVA result. This logic is illustrated pictorially in Figure 1 (C1, C2, ... = individual Chinese scores, and likewise for the other stuff).

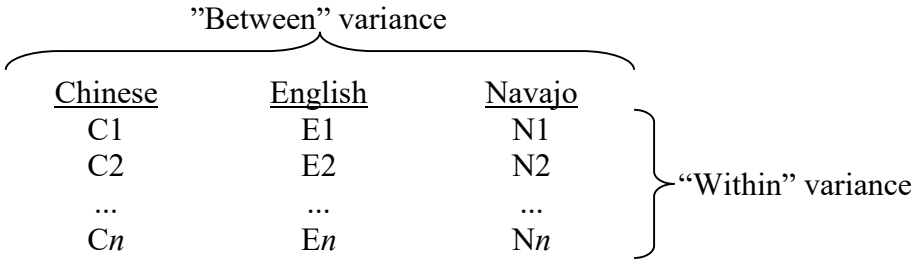

Figure 1. Variance between and within groups

Sheer genius, I say. Of course, although this key idea is simple, the math gets complicated because the precise mathematical definitions of "between", "within", and "significantly more variance" all depend on what your precise data situation is like. In particular, the "within" variance isn't the same as the variance of any particular sample, since there is more than one of them; we need some general way to pool the variance across the samples, the way we did

for the unpaired *t* test. As usual, though, I'll save the scary details until after we've practiced running some ANOVA ourselves.

2.3 Different types of ANOVA

We've already hinted at the crucial issues that you have to consider when choosing which type of ANOVA to run.

First, are you comparing samples that differ in terms one factor (e.g., just language)? Then do a **one-way** (單因子) ANOVA. If you're comparing samples that differ along two factors (e.g., language and sex), then do a **two-way** (雙因子) ANOVA. If you have three factors, do a three-way ANOVA, and so on (but remember Spiderman's words of wisdom).

Second, are you looking at a real-world situation where you can expect the samples to be independent, or are they probably correlated? If your samples of measurements are independent (e.g., because each data point comes from a separate speaker or linguistic form), then do an **independent-measures** ANOVA; that's what we'll look at in this chapter. If they are not independent, then you need to do a **repeated-measures** (重複測量) ANOVA; that's what we'll look at in the next chapter.

Introductory statistics textbooks (e.g., Gravetter & Wallnau, 2004, just to take a random example from my collection) traditionally teach you how to do the three simplest types of ANOVA by hand. These three happen to be the same three types that Excel has built-in tools for. Of course, R (and other "real" stats programs) can do many more types of ANOVA, and once you learn about the ANOVA-regression connection, you can make Excel do some other types of ANOVA too.

These three simplest types of ANOVA are listed below in Table 3, along with the names for them commonly used in psychology (and thus also in experimental linguistics), then the names used in Excel (often quite different, since Excel is aimed more at business people than psychologists), and finally a note on their general purpose. All three types appear at the top of the Analysis ToolPak list, since "ANOVA" starts with "A", the first letter in the alphabet.

Note the difference between the terms "repeated-measures" (as used by psychologists) and "with replication" (as used by Excel). By "replication," Excel means that each of the *samples* has more than one item. This strange terminology will make a bit more sense when we see how Excel expects you to put the data into the spreadsheet, which in turn relates to the different types of math for the different types of ANOVA. A two-way independent-measures ANOVA involves an ANOVA table like Table 2 above, where each ANOVA cell represents a set of data points. But in Excel, each cell represents just one data point, right? So to mimic a two-way ANOVA table in Excel, each ANOVA cell actually contains a whole range of Excel cells, typically a column; the factor levels represented by this ANOVA cell are thus "replicated" within the column (e.g., Chinese-female data point 1, Chinese-female data point 2, etc). By

contrast, a one-way repeated-measures ANOVA has only one **fixed** (non-random) factor that our hypothesis is about (e.g., nouns vs. verbs), but it also has the purely **random** variable (e.g., speakers or words) that is grouping our measurements together. Entering this information into Excel thus again requires a two-dimensional table, with columns for the fixed variable and rows for the random variable. But since each row represents one level of the random variable (e.g., just one speaker or word), the Excel cells are not "replicated" across the ANOVA cells.

Table 3. The three simplest types of ANOVA

Name in psychology	Excel's (confusing) name	Purpose
One-way independent-measures ANOVA	ANOVA: One-Factor (單因子變異數分析)	One multi-level factor from a between-group experimental design
Two-way independent-measures ANOVA	ANOVA: Two-Factor with Replication (雙因子變異數分析：重複試驗)	Two multi-level factors (and interaction) from a between-group design
One-way repeated-measures ANOVA	ANOVA: Two-Factor Without Replication (雙因子變異數分析：無重複試驗)	One multi-level factor from a within-group design

Got it? No? Well, be patient; you'll get some concrete examples later in this chapter, and in the next chapter too. For now, just try to be aware of the terminology differences in Table 3.

3. One-way independent-measures ANOVA

Time to get our hands dirty. What is it like running an ANOVA in Excel or R? To find out, we'll look at a simple (fake) experiment that has one factor but three levels, each tested on a separate group of people, where what we care about is the means (not the frequencies). Since there are more than two levels, we can't do a *t* test, but a one-way independent-measures ANOVA is just the right tool for us. I'll first go through the example, and then explain how the math works. Since we might also be curious about comparing pairs of levels within this trio, we'll need to learn how to do that "safely" too, without falling into the Type I error trap of multiple comparisons.

3.1 Three colored rooms

Here's our fake experiment, with such a tiny data set that I can print it right in this chapter. It's tiny because I "borrowed" it from the equally fake example in Gravetter & Wallnau (2004) (Table 13.1, p. 401), which is one of those old textbooks that teach you how to do ANOVA by hand. I kept the numbers, but changed the factors to something a bit more linguistic. As we'll

see, their numbers make the statistics come out suspiciously "neat". In real life you shouldn't really do ANOVA with such a small sample, especially, as here, the measurements actually represent counts, which aren't completely normally distributed. But let's forget about these realities for now and get on with the story.

Once upon a time, a wise old Chinese teacher wanted to test whether room color affects how well foreign students learn new Chinese words (it's not impossible, right?). So she randomly assigned 15 students to red, blue, or yellow rooms (5 people per room) and had each one try to learn 10 new words. In this experiment, we have independent samples defining three levels of one factor, so it's a pretty good situation for this type of ANOVA (aside from the tiny sample size and the likelihood that the samples aren't very normal).

Building on something I mentioned in the previous section, this test still assumes that you only care about these three specific colors. In other words, color is a **fixed** variable here, not a **random** variable, so we cannot generalize to *all* colors; our conclusions only apply to red, blue, and yellow. In later chapters we'll learn other tests that *can* generalize like this, but not for a while yet.

Anyway, the (fake) results of this dumb experiment are as shown in Table 4 (also available in the file **ColoredRooms.txt**). Note that the rows don't represent anything; each room has five totally different people.

Red	Blue	Yellow
0	4	1
1	3	2
3	6	2
1	3	0
0	4	0

Table 4. The effect of room color on word learning (Excel-style table)

Hm, it seems there might be an effect of color here, since more words were learned in the blue room ($M_{Blue} = 4$) than in the other two rooms ($M_{Red} = 1$, $M_{Yellow} = 1$; I told you the values in this fake example are suspiciously neat). But is this *statistically significant*?

3.1.1 Three colored rooms in Excel

As usual, let's try Excel first. So enter that table into an Excel spreadsheet, fire up the Excel statistics toolbox, and choose 單因子變異數分析 (ANOVA: One-Factor), since Table 3 above shows that this is the right kind of Excel tool for this situation. Then we select the entire table of numbers including the labels (Excel lets you run a one-way independent-measures ANOVA even if the sample sizes aren't the same), do the rest of the usual stuff

(including clicking the check box to indicate that, yes, we do have labels on our columns), and we get our output. This includes a table summarizing the sizes, means, and standard deviations of our three levels, but also the ANOVA report table shown in Table 5.

ANOVA

變源	SS	自由度	MS	F	P-值	臨界值
組間	30	2	15	11.25	0.001771	3.88529
組內	16	12	1.333333			
總和	46	14				

Table 5. Excel's output for the colored room experiment

This table shows some statistical terms that you should already be able to figure out, even based on my informal discussion so far. So in the upper left we see 變源 (variance source; English Excel only has room for the one word "Sources"). As I told you recently, there are two main sources of variance in an independent-measures ANOVA, namely 組間 (the "interesting" variance between the groups, in this case, across three colored rooms) and 組內 (the "boring" variance that is pooled within the groups, i.e., each individual student). There's also a row for total variance (總和). You can also see the terms *SS* (sum of squares) and *MS* (mean sum of squares), which we've seen a few times so far, starting with the definition of standard deviation. So on the "between" row we have $SS_{between}$ and $MS_{between}$, and on the "within" row we have SS_{within} and MS_{within}. Another name for MS_{within} in ANOVA is **mean squared error**, which psychologists usually abbreviate as $\boldsymbol{MSE}$ (sometimes as $\boldsymbol{MS_e}$). And look: each of these two sources of variance also has its own *df* (自由度). In the total variance row, the *SS* value is just the sum of the other two *SS* values, and likewise for *df*.

And here's something else cool: in each row, $MS = SS/df$ (try it!). As we noted many chapters ago, *MS* is just another way to represent variance. So these *MS* values represent "how far away" a value is from the null hypothesis value (remember how we also divided by *df* in the *z* score formula, which eventually turned into the *t* test formula).

With that insight, where do you think that lonely *F* value comes from? Well, the *F* test is for comparing one variance to another, and it's computed simply by dividing one variance by the other. So what happens if we divide $MS_{between}$ by MS_{within}...? Wow! It's exactly the *F* value in the table! I'm so amazed!

That's a pretty large *F* value, much much higher than 1; I'll bet it's statistically significant. I wonder what the *p* value should be for this *F* value, in the *F* distribution defined by those two *df* values? How can I calculate this in Excel? Oh, right, I can use **=FDIST(F, df1, df2)**, and those three argument values are right there in the Excel table, so I can just click them, and... I can't believe it! I got the same *p* value shown in the table! The table also gives us the critical

value [臨界値], which is the *F* value for this distribution where we get $p = .05$, but that's mainly for computing confidence intervals, and we're not going to worry about ANOVA confidence intervals until the next chapter.

This whole cascade of calculations started with the *SS* values, but explaining those requires new formulas, so I'll save them for later too. For now, just rest assured that they are conceptually the same as the *SS* values we saw in earlier chapters: a measure of overall difference, with squaring to get rid of negative values. We'll need to explain the *df* values too, but that's also for later.

For now, let's return to the wise old Chinese teacher. What should she write in her report? Well, something like this would be good: "The number of words learned in each room varied across red (*M* 1, *SD* 1.22), blue (*M* 4, *SD* 1.22), and yellow (*M* 1, *SD* 1) rooms, a difference that proved significant by a one-way independent-measures ANOVA ($F(2, 12) = 11.25$, $MSE = 1.33$, $p < .01$)." Notice what she reports: the means (and standard deviations, for completeness), the full name of the test, the name of the distribution (*F*), the values that define it (the two *df* values), the actual *F* value, the *p* value, and (something new) the *MSE*. The *MSE* is not on the same scale as the means, since as a variance, it's squared, but it gives the readers some sense of the "noisiness" of the data, and can help them check your results (as we'll see, $MS_{between}$, the other *MS* value used to calculate *F*, can be estimated from the reported sample means).

Before leaving Excel, let me you show you something else that's useful to know. Remember how I said the *t* test is a special case of ANOVA? Specifically, the unpaired *t* test (not Welch's test, but the version assuming equal variance) is a special case of a one-way independent-samples ANOVA. We can see this by running Excel's ANOVA tool on just two columns of our colored room data set, and comparing its output with what we get with Excel's *t* test tool.

Play along! If you run an independent-measures ANOVA on just Red vs. Blue, you get $F(1, 8) = 15$, $p = .004721383...$. If you do an unpaired *t* test (assuming equal variance) on these same two samples, you get $t(8) = -3.872983346...$, $p = .004721383...$ (two-tailed). The exact same *p* value comes out, but do you see that the *df* and *t* values are also related? Namely, the *df* for this *t* test is the same as the df_{within} used by ANOVA, and $t^2 = F$: $(-3.872983346...)^2 = 15$. Not only that, but *MSE* = pooled $s^2 = 1.5$. So a *t* test really is a special case of ANOVA!

3.1.2 Three colored rooms in R

How do you do all of this stuff in R, you ask? Well, R does have a function called **anova()**, but it turns out it's not designed for running ANOVA tests per se, but rather for creating ANOVA tables in general. Since almost everything in statistics is related to everything else, ANOVA pops up in other situations too; in fact, we've already seen that when you run a

regression analysis in Excel or R, you get an ANOVA table, along with the coefficients table. That's what R's **anova()** function is for, so we don't want that now (but we'll come back to it again when we return to regression in later chapters).

Instead, what we want is a specialized R function called **aov()** (which also stands for Analysis Of Variance, of course). This is designed to reflect the logic of ANOVA in its syntax, so it lets us run just the kind of ANOVA that we want.

The first step to using it is to give R our data. As usual, R expects the data to be arranged with separate columns for the independent variable (Color) and dependent variable (Learning). In our case, we have to rewrite Table 4 as shown in Table 6 below.

Color	Learning
Red	0
Red	1
Red	3
Red	1
Red	0
Blue	4
Blue	3
Blue	6
Blue	3
Blue	4
Yellow	1
Yellow	2
Yellow	2
Yellow	0
Yellow	0

Table 6. The effect of room color on word learning (R-style table)

To save you copy/pasting, typing, and file-loading trouble, here's some code that puts these values into a data frame called **exp1** (for Experiment 1, since we'll get another one later):

```
exp1 = data.frame(Color = c(rep("Red",5), rep("Blue",5), rep("Yellow",5)),
  Learning=c(c(0,1,3,1,0),c(4,3,6,3,4),c(1,2,2,0,0))) # To keep track of the 3 samples
head(exp1) # See what it looks like
```

```
  Color Learning
1   Red        0
2   Red        1
3   Red        3
4   Red        1
5   Red        0
6  Blue        4
```

As usual for R, **aov()** takes a formula as an argument (unlike **t.test()**, it has no option to list the samples separately). Also as usual for R (though again, different from **t.test()**), running **aov()** merely creates a model; to show the actual ANOVA table we need to put this model inside the **summary()** function.

Let's try it out. I'll do it in two steps first, just to show that **aov()** merely creates a so-called **aov** object. Given the link between ANOVA and regression, I hope you're not surprised to see that R mentions "residuals" here; remember that those are the differences between the actual values and the values predicted by a model, and thus gives a measure of the noisiness of your data (relative to your model). Given that this fake data set was designed to give a significant result despite being extremely tiny, it's also not surprising the R warns us that something may be "unbalanced" here (which we'll ignore anyway).

```
colors.aov = aov(Learning~Color, data=exp1) # Creates an "aov" object
colors.aov # Basic descriptive information about this object
```

```
Call:
   aov(formula = Learning ~ Color, data = exp1)

Terms:
                    Color Residuals
Sum of Squares         30        16
Deg. of Freedom         2        12

Residual standard error: 1.154701
Estimated effects may be unbalanced
```

But now if we put this **aov** object inside **summary()**, we can get basically the same ANOVA report table that Excel gives us (except without the critical value or the row for total variance). R's terms are slightly different, though; so *SS* is "Sum sq", *MS* is "Mean Sq", and "Within" is "Residuals" (again, representing the noisiness of our data relative to our model).

```
summary(colors.aov)
```

```
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
Color       2   30      15.000   11.25     0.00177   **
Residuals   12  16      1.333
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Unless you want to use the **aov** object for something else (e.g., checking the residuals or balance, or running some follow-up that we'll discuss in a later section), you can get the same ANOVA report table in just one step, like so (try it!):

summary(aov(Learning~Color, data=exp1))

3.2 More about the math of ANOVA

Let's flesh out how this works in a bit more detail. As a review, recall that the ultimate goal of ANOVA is to get a *p* value for the null hypothesis that the between-sample variance is the same as the within-sample variance, which we do with *F* tests. In an *F* test, you divide one sample variance by another: $F = s_1^2 / s_2^2$, and to figure out the *p* value from the *F* value, you need the *df* for both the numerator (分子 on top) and the denominator (分母 on the bottom).

To see how this relates to ANOVA, you have to recall two more things. First, what we want to know is whether between-sample variance is bigger than within-sample variance, so we want to make between-sample variance the numerator (top) and within-sample variance the denominator (bottom). Second, recall what variance (s^2) really means. It means the mean of all the squared differences between the data points and the mean. When you calculate this mean, you first calculate the **sum of squares** (*SS*) and then divide by a number related to the total (adjusted to *df*, by subtracting something, in order to "punish" our arrogant attempt to learn about the abstract, infinite population from our finite sample). This gives you a **mean sum of squares** (*MS*). As we have seen, *SS*, *MS*, and *df* all appear in the ANOVA reports generated by Excel and R.

Let's put these concepts together. It's easiest to follow the logic if we work backwards from the last step.

Our final goal is to compute the *p* value. This is the area in an *F* distribution to the right of the *F* value generated by the *F* test. The specific *F* distribution is determined by $df_{between}$ (numerator on top) and df_{within} (denominator on the bottom), which are determined by different formulas depending on the particular ANOVA. In the case of the one-way independent-measures ANOVA, which is the simplest type of ANOVA, the two *df* formulas are as follows:

One-way independent-measures ANOVA *df*:

$df_{between} = k - 1$, where k = the number of levels of the factor
$df_{within} = \Sigma_i(n_i - 1)$ for each sample i of size n_i

Do these formulas work for the colored room experiment? Well, our one factor (Color) has three levels, so $k = 3$, so $df_{between} = 2$, and indeed, that's what the Excel and R reports both show. Each room had five people in it, so $df_{within} = (5\text{-}1) + (5\text{-}1) + (5\text{-}1) = 3 \times 4 = 12$, and that's just what's shown in our ANOVA report tables as well.

As with the unpaired *t* test, the math is the simplest if all of the samples have the same size (i.e., $n = n_1 = n_2 = \ldots n_k$), and so that's what I'll explain below, just to give you the core concepts. Equal cell sizes aren't obligatory to run an ANOVA, but again, as with the unpaired

t test, the greater the difference in sample sizes, the riskier it is to violate the other usual assumptions of this kind of test, namely equal variance across all samples, and normally distributed populations.

Various tricks have been suggested to make cell sizes equal by estimating the missing data (see, for example, https://www.r-bloggers.com/2018/06/dealing-with-missing-data-in-anova-models/). Sadly, the simplest method you might think of, namely replacing all missing values with the mean of the cell, won't work, since this changes that cell's variance, as you can see if think about it: putting a distribution's mean into the distribution makes the distribution "taller" in the middle, which will shrink the standard deviation. Instead, the valid methods all involve using regression to predict what the missing values are most likely to be, but I find that ironic, since as I keep saying, ANOVA is itself just a special case of regression anyway, so you may as well use regression for the whole analysis in the first place.

In any case, no matter what kind of ANOVA you do, the *F* test formula is always as follows:

General ANOVA *F* formula: $$F = \frac{MS_{between}}{MS_{within}}$$

If the null hypothesis is false, the two *MS* values will not only be different, but $MS_{between}$ will be bigger than MS_{within} (also known as *MSE*). Thus the bigger the "interesting" difference, the bigger *F* will be, and the smaller *p* will be. By contrast, if the null hypothesis is false, *F* will be close to or even below 1.

At an abstract level, $MS_{between}$ and MS_{within} are both calculated in the same general way, as shown below. Remember that by dividing the sum of squares (*SS*) by the degrees of freedom (*df*), you get the variance.

General ANOVA MS formulas:

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

The formula for $MS_{within} = MSE$ is also related to something we saw earlier with the unpaired *t* test. You probably don't remember this, but for that test, the **pooled variance** was calculated as follows:

Unpaired *t* test pooled variance formula: $$s_p{}^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

The *MSE* used in ANOVA is just a generalization of this to multiple samples:

General ANOVA *MSE* formula: $MSE = \frac{SS_1 + \cdots + SS_k}{df_1 + \cdots + df_k}$

Since the bottom part of the *MSE* formula is just the sum of all the sample *df* values, and we already know that this gives us df_{within}, and since $MS = SS/df$, it must therefore be the case that SS_{within} is just what's at the top of the *MSE* formula, namely the sum of all the samples' *SS* values:

One-way independent-measures ANOVA SS_{within}: $SS_{within} = \Sigma_i SS_i$ (i across samples)

Does this SS_{within} formula work for the colored room experiments? Let me look up how to calculate *SS* for samples.... Oh, right, it's just the step right before dividing by *df*, on the way to calculating variance. Namely, you subtract the sample mean *M* from each data point *x*, then square the difference, then sum up all of these squares:

Sample *SS*: $\Sigma_i (x_i - M)^2$ (i across data points)

You can try computing this in Excel, but it's easier to show you using R (try it!):

```
# Pull out our three samples
red = exp1$Learning[exp1$Color=="Red"]
blue = exp1$Learning[exp1$Color=="Blue"]
yellow = exp1$Learning[exp1$Color=="Yellow"]

# Compute sample SS values (so glad R uses vector logic!)
SS.red = sum((red-mean(red))^2)
SS.blue = sum((blue -mean(blue))^2)
SS.yellow = sum((yellow -mean(yellow))^2)

# Compute SSwithin:
SS.red + SS.blue + SS.yellow
```

This gives you 16. Is there a 16 somewhere in those ANOVA report tables? Oh, there it is: it's SS_{within}! Perfect.

The other *MS* value, $MS_{between}$, is trickier, but I'll just to give its formula for the simplest case, where all of the samples (ANOVA cells) have the same size *n*. In that case, the formula is like so, where SS_M represents the sum of squares across the sample means (i.e., treating the set of sample means itself as a set, over which we compute *SS*). Conceptually, this value is capturing the differences across the samples (SS_M) and enhancing this value by the sample size

(n), since as usual, with larger samples, even a slight difference across sample means can become statistically significant.

One-way independent-measures ANOVA $SS_{between}$ (same n): $SS_{between} = n \cdot SS_M$

where $SS_M = \sum(M_i - M_{grand})^2$ (for i across samples)

Does this formula work for our colored rooms experiment? Remember that our three means are $M_{Red} = 1$, $M_{Blue} = 4$, $M_{Yellow} = 1$, so $SS_M = \Sigma_i (M_i - M_{grand})^2$, where M_{grand} is the **grand** mean, across all three of these means. Then $SS_{between}$ is that number times 5 (the size of each same-sized sample). Let's just let R do it for us, and if you try it, you'll see that it gives us 30, which is indeed the same as the $SS_{between}$ value reported in the ANOVA report tables.

```
three.means = c(mean(red),mean(blue),mean(yellow))
SSM = sum((three.means - mean(three.means))^2)
5 * SSM
```

If the sample sizes are different, the formula for $SS_{between}$ gets more complicated, but it remains conceptually the same, as a measure of cross-sample means enhanced by the sample sizes.

Since we're mainly discussing these formulas for conceptual purposes, let's add one more conceptually important formula. Remember how the last line of Excel's ANOVA report table gave "total" *SS* and *df* values? SS_{total} literally represents the total sums of squares, as reflected in the following formula, where x represents any data point in any set and M is the grand mean across all of the data:

$SS_{total} = \Sigma(x - M_{grand})^2$

But as we saw from Excel's ANOVA report table, it's also related to the other *SS*, just as the total *df* is:

General ANOVA SS_{total} formula: $SS_{total} = SS_{within} + SS_{between}$

General ANOVA df_{total} formula: $df_{total} = df_{within} + df_{between}$

The conceptual value of these formulas is that they express the idea that SS_{within} and $SS_{between}$ **partition** (divide) the total variability expressed by SS_{total}. This is a great illustration of how statistics goes beyond the clockwork logic of Newton's day: instead of being scared of variability, statistics sees *everything* as variable, yet manages to bring it under control by systematically dividing the "interesting" variability (associated with the factors in our model) from the "boring" variability (the residual noise that our model can't explain). The notions of

partitioning the variance, residuals, and the ratio of "explained" to "total" variability will come up repeatedly in the rest of this book.

3.3 Post-hoc tests

All right, the stupid color experiment shows that color matters in word learning, but can we also say that blue is the best color? After all, that was the color associated with the best learning (mean of 4, compared to 1 for each of the other two colors). But this is actually a new hypothesis, not the same as the one tested by the ANOVA itself. The analysis we just did for the stupid color experiment merely tested whether the three colors are different overall, but it didn't test which color is different from what.

Another situation where this kind of problem arises in linguistics as in a **priming** experiment, where you ask people to respond to target items preceded by primes that are related to the targets in various ways, so you can see which kind of prime affects the target response speed. For example, if the target is *CAT*, maybe one type of prime is related to the meaning (*dog*) and another is related to the sound (*hat*). In such experiments, you also need a **control** condition with no relation (e.g., *pen*). So the factor PrimeType has three levels (Semantic, Phonological, Control), but what you actually care about are the pairs of levels (Semantic vs. Control and Phonological vs. Control to see if there is any priming at all, and Semantic vs. Phonology to see if these properties differ in processing). The problem is that a one-way ANOVA, by itself, won't test these individual comparisons.

Personally, I think the best way to deal with this problem is to avoid it in the first place. Why not design your experiment in terms of binary factors? For example, if the Chinese teacher suspected that blue was the best color, why not just test it against what she thought would be the second-best color? If a multi-level factor is unavoidable, you could use a regression analysis (generalizing beyond ANOVA), which lets you compare each level in a multi-level factor against a single **baseline** level (we'll learn how to do this when we get to the multiple regression chapter).

But for our Chinese teacher, it's too late: she already ran her study, and she wants to compare her colors. What are our options?

3.3.1 The dumbest way: Planned comparisons

The most obvious way to compare pairs of levels within a multi-level factor seems like cheating at first, and in fact, the more you think about it, the more like cheating it seems to be. Yet unfortunately it remains common, just because it's so simple to do. This method is called **planned comparisons**. This is where you claim that even before you ran your experiment, you planned on comparing just two specific levels, say Blue vs. Red. So in a sense you did two

experiments at the same time: not just the Red-vs.-Blue-vs.-Yellow experiment, but also the Red-vs.-Blue experiment.

But didn't I say at the start of this chapter that doing multiple comparisons like this raises the risk of Type I error so high that the *p* values become reliable? Yes, I did say that, and it's still true. For this reason, many experts (e.g., Gelman & Loken, 2013) strongly advise against so-called planned comparisons.

The reason why some people say planned comparisons are OK anyway is that by claiming that you were planning to do these two-sample tests all along, the conditional probabilities are technically still "innocent" of all the other tests you might also do. That is, you don't run the Red-vs.-Blue experiment *because* of what you got in the Red-vs.-Blue-vs.-Yellow experiment; you supposedly planned to do both all along, so the *p* values for the two-sample experiment can still be seen as testing the basic two-sample null hypothesis, independently of all other possible tests that you might also do on the data.

This kind of game drives Bayesian statisticians (like Gelman) crazy, since traditional statisticians always complain that Bayesian statistics is too "subjective", and yet here's one of the many clear cases in traditional statistics where we're being asked to trust the researcher's word about his or her psychological state. After all, maybe the Chinese teacher actually expected that Red and Blue would be the same, so when the means implied otherwise, this caused her to look at Red and Blue more carefully. But in that case, the *p* values of the *t* test are suddenly invalid: the focus on Red and Blue was based on the main ANOVA results, so they're not randomly chosen samples anymore. The *p* value from an ANOVA doesn't represent this kind of conditional probability, and as we noted earlier, as the number of two-level comparisons increases, the Type I error risk goes through the roof.

3.3.2 The simplest safe way: Bonferroni adjustment

One way to make planned comparisons less suspicious-looking is to lower the alpha level, to make it harder to get a Type I error (since a lower alpha means that you need to get a lower *p* value for the result to count as statistically significant). The simplest way to do this is to use something called the **Bonferroni adjustment** or **Bonferroni correction**, named after the Italian mathematician Carlo Emilio Bonferroni (1892-1960). Due to the addition rule of probability theory, the probability of getting a combination of independent, non-overlapping events is the sum of the individual event probabilities. So if you're going to do some number of multiple comparisons, you can simply divide the alpha level by that number, and only count any given *p* value as significant if it goes below this lowered alpha level.

For example, since we want to do three planned comparisons in the colored room experiment (Red-Blue, Red-Yellow, Blue-Yellow), we should lower α to (.05)/3 = .0167. That is, only if a *t* test on any of these pairs gets a *p* value below .0167 (not .05) should we consider

it statistically significant. In the case of this fake experiment, we win: the *p* value for comparing Red with Blue is $p = .0047$, which is still below our adjusted alpha level of .0167 (likewise for Blue vs. Yellow, but not for Red vs. Yellow, since their means were identical).

Though it is extremely simple to use, the Bonferroni adjustment is very conservative. That is, it avoids Type I errors by greatly increasing the risk of Type II errors, so it's easy to give you $p > \alpha$ even though the null hypothesis is actually false. It also ignores the ANOVA itself; you can you use the Bonferroni adjustment in any situation where you're doing multiple comparisons. Both of these features can be good things. Regarding the first, if your "planned comparisons" are still significant after Bonferroni adjustment, you can convince skeptics that there really are pairwise differences. Regarding the second, you can use the Bonferroni adjustment beyond ANOVA, for example, when testing a whole bunch of correlations on the same data set (though a multiple regression would be a better thing to do there).

One shortcoming of this procedure is that it's not very powerful: if you want to avoid *any* Type I error in *all* of your comparisons, you may reduce your alpha value too much to catch some genuinely significantly different comparisons. To deal with this, Benjamini & Hochberg (1995) proposed instead trying to minimize what they called the **false discovery rate** just in the subset of comparisons with *p* values below the usual alpha level. They proved that you can manage this computing *p* values for all of your comparisons in the usual way (for example, using ordinary *t* tests), ranking them from smallest to largest and numbering them (1 = smallest *p* value, 2 = next-smallest *p* value, and so on), and then finding the largest rank *i* where the following formula is true:

$$p_i \leq \frac{i}{m}\alpha$$

where m = total number of comparisons, i = ranking number, α = alpha level, and p_i = the ith p value. Then reject the null hypothesis for all of comparisons up to that largest rank.

For example, if you have three comparisons with the *p*-values .01, .02, .03, and an alpha level of .05, only the first *p* value would be significant by the Bonferroni adjustment, since .05/3 = .0167. But using the false discovery rate procedure would allow us to conclude that all of the comparisons are statistically significant:

$.01 \leq \frac{1}{3}(.05) = .0167$ True

$.02 \leq \frac{2}{3}(.05) = .0333$ Also true!

$.03 \leq \frac{3}{3}(.05) = .05$ True again!

3.3.3 The most sophisticated way: Post-hoc tests

But here we have in fact run an ANOVA, and all we want to do is look a bit closer at the results, without having to claim that we planned to do all of these extra analyses from the very beginning. The best approach here would thus be to do a so-called a **post hoc test** ("post hoc" means "after that"). Post-hoc tests are pretty widely used, and much less controversial and problematic than so-called planned comparisons, and less conservative than Bonferonni corrections.

There are many types of post-hoc tests, since there's no one right answer to the question "How best to balance Type I and Type II errors?" So I'll just mention three commonly used post-hoc tests, and only explain the third one in detail.

The oldest post hoc test is **Fisher's (P)LSD** ("[protected] least significant difference"), invented by you-know-who. Like planned comparisons, this basically involves doing a bunch of *t* tests, and like the Bonferroni adjustment, it makes it harder to commit a Type I error, though instead of adjusting alpha, it computes pooled variance across all of the samples, not just the two being compared. The problem with this is the opposite of the Bonferonni adjustment, in that it's an overly "generous" test, giving a too-high risk of Type I errors.

The second oldest of the post hoc tests is **Scheffé's test**, named after American statistician Henry Scheffé (1907-1977), which can look for significant effects in subsets of all your samples, not just in two levels at a time (for example, in a factor with levels A, B, C, D, it could test if the subset A, B, C shows a significant difference in means). One advantage of this test is that it's very robust to violations of the usual assumptions: it doesn't matter if the sample sizes are different or if they have different variances. However, sort of like the ranked tests we've discussed in other chapters, this robustness also means that it is very "conservative" (i.e., it has a high risk of Type II errors).

The third type of post-hoc test seems to be the most commonly used post hoc test today, perhaps because it strikes a better balance between Type I and Type II errors: the **Tukey** (or Tukey-Kramer) **HSD test** (for "honestly significant difference"). You've seen the name Tukey already; he was the guy who invented the box-and-whiskers plot. Luckily for us, R also has built-in base functions for computing it (to run the other two post-hoc tests, you'd need to install a special package; e.g., the **agricolae** package has the functions **scheffe.test()** and **LSD.test()**).

Tukey's HSD is just a modification of the oldest test, in that it's "really" a bunch of *t* tests with a modified way of calculating variability. The new thing is that instead of using the ordinary *t* distribution ("Student's *t*"), Tukey's HSD uses a distribution called the **Studentized range statistic**, symbolized *q* (I guess they were running out of letters), which represents the range (maximum minus minimum) of a sample divided by the standard deviation. The output

of this process is a set of adjusted *p* values, one for each cross-level comparison, that you can test for significance against your alpha level (e.g., $\alpha = .05$).

The *HSD* value is calculated using the formula below (again, this is the version for the simplest case, where every sample has the same size *n*). This number represents how big the difference must be between two factor level means to count as significant.

Tukey's HSD value (for equal sample sizes): $HSD = q_{crit} \cdot \sqrt{\frac{MS_{within}}{n}}$

where q_{crit} is the critical value of q that makes $p = \alpha$

The three crucial values are all available in R: you can get *n* using **length()**, MS_{within} (*MSE*) is what **summary(aov())** reports as the Mean Sq for Residuals, and q_{crit} can be calculated using R's built-in function for quantiles in the Studentized range distribution, called **qtukey(1-alpha, k, df)**, where *k* = number of factor levels in the full ANOVA analysis and $df = df_{within}$. Putting this all together, the computation of the HSD value would be like so: **HSD = qtukey(1-alpha, k, df)*sqrt(MSE/n)**. Let's try it on our color experiment values:

```
qtukey(1-0.05, 3, 12)*sqrt(1.333/5)
[1] 1.948089
```

Since the difference in Red vs. Blue means is 3 (4-1), which is larger than the HSD value (1.95), this post-hoc comparison must be significant ($p < .05$); likewise for Blue vs. Yellow, but not for Red vs. Yellow.

However, R makes the job even easier than this: it has a specialized function to give you the actual Tukey-test *p* values directly from your ANOVA model. So here's a case where you do indeed need your **aov** object, not just the summary of it. The function is called **TukeyHSD()**, and you have to remember to type the letters properly in uppercase and lowercase.

For example, with our color experiment, we could use the following R code and get the following results:

```
TukeyHSD(aov(Learning~Color,data=exp1))
```

	diff	lwr	upr	p adj
Red-Blue	-3.000000e+00	-4.948332	-1.051668	0.003832
Yellow-Blue	-3.000000e+00	-4.948332	-1.051668	0.003832
Yellow-Red	9.992007e-16	-1.948332	1.948332	1.000000

In this output table, "diff" stands for the difference between levels (here, 3, 3, and 0), "lwr" and "upr" represent (respectively) the lower and upper bounds of the confidence interval for

the Tukey test (R knows that statisticians don't like to rely just on point estimates), and "p adj" stands for the adjusted *p* value (i.e., adjusted according to Tukey's test). This last shows that $p < .05$ for both comparisons with Blue, so we can conclude that it is indeed a better color than either Red or Yellow. By contrast, the Yellow and Blue means are identical (9.992007e-16 = zero), the adjusted $p = 1$. Notice the upper bound on this particular confidence interval? It's practically the same value (1.948332) that we got when we used **qtukey()** to compute the critical value (1.948089).

Just for completeness, we can further confirm that my "manual" calculation of HSD was right by plugging in the Red-vs.-Blue *p* value into my calculation above. And yes, the output is practically 3, the actual difference in sample means.

```
qtukey(1-0.003832, 3, 12)*sqrt(1.333/5) # See what I'm doing here?
[1] 2.999621
```

To report this result, the wise old Chinese teacher would add, after her main ANOVA report, that learning in the blue room was significantly better than in the red and yellow rooms by a two-tailed Tukey post-hoc test ($ps < .01$). There are two different significant *p* values here; they just happen to be identical in this particular analysis because the means for the red and yellow rooms are identical.

4. Two-way independent-measures ANOVA

Even if everybody listens to me (ha!) and designs all of their studies so that all of their factors are always binary (so that they never have to worry about multiple comparisons or post-hoc tests), Fisher is still right: it's best to give Nature a questionnaire, not just ask one question at a time. In particular, it's often crucial to test for an interaction between your factors, and that means that we still need to do something like an ANOVA. If we have two factors, it's a two-way ANOVA, and if both factors are between-groups, it's a two-way independent-measures ANOVA.

In this section I'll first pound home the importance of testing for interactions, and then we'll see how to actually do it in Excel and ANOVA. Only at the last minute will I update our ANOVA math to handle this new kind of situation.

4.1 Why we might want to test for interactions

I've already mentioned some reasons to test for interactions, but let's make the discussion a bit more concrete.

If two variables interact, then they aren't independent (of course). This kind of information is often crucial for distinguishing between competing scientific hypotheses. For

example, linguists assume that semantics and phonology are totally separate things, processed in the mind/brain by totally separate systems. So we predict that they will *not* show an interaction in an ANOVA, if we do the right kind of experiment.

For example, let's say you run a version of the priming experiment I mentioned above, but you **cross** semantics and phonology in the primes (avoiding shared morphemes). This is a so-called **factorial experiment**, testing all possible combinations of our factors. We might summarize our experiment as in Table 7.

Table 7. The design of a two-factor factorial priming experiment

Semantics	Phonology (same first sound)	Prime example	Target example
Related	Related	cod (kind of fish)	CAT
Related	Unrelated	dog	CAT
Unrelated	Related	cot (kind of bed)	CAT
Unrelated	Unrelated	dig	CAT

Figure 2 shows three possible outcomes of this experiment, where the y-axis represents priming (i.e., RT for control condition minus RT for primed condition, so that higher values means faster responses for the primed condition). I faked the data with the following R code (you can try faking it in Excel too, just for practice). Note the use of R's base function **interaction.plot()**; check **?interaction.plot** for more information on how it works. Actually, the code below only creates the leftmost plot; the comments at the end of the code give you a hint about how to plot the other two.

```
par(mfrow=c(1,3)) # I'll plot three graphs side by side
par(cex=1.1) # Makes the font bigger (since the graphs will be shrunk)

# Leftmost plot
semphon = data.frame(SemRel=c(rep("Not related",2),rep("Related",2)),
  PhonRel=rep(c("Not related","Related"),2),Priming = c(0,20,40,60))

interaction.plot(semphon$SemRel, # Variable on x-axis
  semphon$PhonRel, # Variable in the legend
  semphon$Priming, # Variable on y-axis
  xlab="Semantics", ylab="Priming ms", # Default labels are ugly
  ylim=c(0,100), # This gives room at the top for the legend
  legend=F, # Default legend style for interaction.plot is ugly, so I turned it off
  lwd=2) # Makes the lines thicker (since the graphs will be shrunk)
  legend("topright",lty=c(1,2),legend=c("Phon not rel","Phon rel")) # Prettier legend

# And likewise for the other two plots:
# Middle plot: Priming = c(0,40,40,20)
# Rightmost plot: Priming = c(20,20,40,60)
```

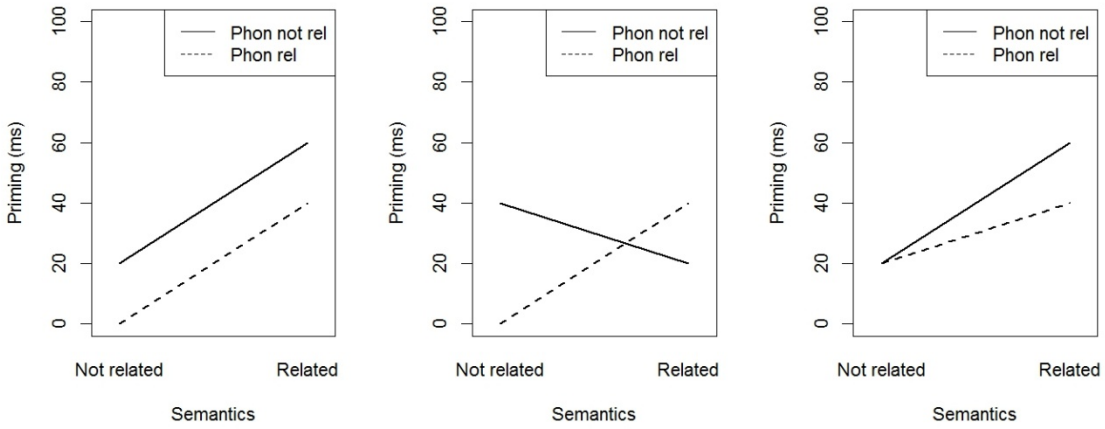

Figure 2. Three possible outcomes from our two-factor factorial experiment

In the first possible outcome (leftmost plot in Figure 2), there is no interaction: semantic relatedness increases priming by 40 ms, whether or not there is also phonological relatedness, which in turn increases priming by 20 ms, whether or not there is also semantic relatedness. In this situation we know the amount of priming when both are related just by adding the two effects: 40 + 20 = 60 ms. Thus this kind of no-interaction model is also called an **additive model**. Only if we get a result like this can we safely preserve the assumption that semantics and phonology are processed independently. When this kind of cognitive logic is applied to reaction times, as it is here, it is called **the method of additive factors** (Sternberg, 1998).

Technically, of course, a non-significant interaction is still a null result, and as we know, in traditional (non-Bayesian) statistics it's risky to try to interpret a null result, but if the power of our test seems pretty strong (if there are large effect sizes for the main effects, and if our sample size is pretty big, it's not crazy to say that the lack of a significant interaction here is at least consistent with a linguistic theory where semantics and phonology are processed separately).

In the other two possible outcomes, however, simply adding up the two factors doesn't give the right results. In order to know what the effect of semantic relatedness is in any specific condition, we also have to know whether the prime was also phonologically related. In the middle graph the lines actually cross, which is clearly an interaction, but there also seems to be an interaction in the rightmost graph, where the two lines have different slopes instead of being parallel.

While there is only way *not* to have an interaction (parallel lines in an interaction plot), there are many ways to *have* an interaction, as these latter two graphs show. This is why it's always crucial to make a plot when your statistical analysis reveals a significant interaction, so

you can try to figure out what the interaction actually means for your real-life research question: the numbers alone don't give your monkey brain enough information.

Note that I used line plots for what are actually categorical data points: in reality there is no data "between" semantically unrelated and related, since this is a categorical variable, not a continuous variable. To reflect this reality more directly, you could also use a bar plot (a separate bar for each of the four values), and this is indeed often done in published papers. Nevertheless, researchers (and R's programmers, who wrote the **interaction.plot()** function) also approve of using lines in interaction plots, since they make the absence vs. presence of interactions much easier to see (i.e., parallel vs. not-parallel lines) than bar plots do (where you have to look closely at the relative differences in the heights of each subgroup of bars).

While users of the method of additive factors hope to *maintain* the null hypothesis (no interaction between supposedly independent factors), you can also test for interactions that your scientific hypothesis *does* predict. This is what Cowart (1997) does in a test of the *that*-trace effect (Chomsky & Lasnik, 1977), where English sentences with the complementizer *that* and subject extraction are claimed to be less acceptable. The factorial design, and sample sentences, are shown in Table 8.

Table 8. Design of the factorial syntax experiment of Cowart (1997)

Extraction	*that*	Example sentences (from Cowart, 1997, p. 165)	Prediction
Subject	Present	Who was the nurse imagining *that* _ would find her?	Worse
Subject	Absent	Who was the nurse imagining ∅ _ would find her?	Better
Object	Present	Who was the nurse imagining *that* she would find _?	Better
Object	Absent	Who was the nurse imagining ∅ she would find _?	Better

The hypothesis thus crucially predicts an interaction between the two factors; whether they also show main effects is less important. This kind of design is quite common in theoretical syntax, where researchers are actually quite familiar with factorial designs, though they don't always know that this is what they're using (see Myers, 2009a, 2009b, for discussion, including Chinese examples).

Another reason why we should care about testing for interactions is to avoid making a stupid (but sadly common) mistake.

Suppose you have a sociolinguistic theory that says that Venusian culture is so egalitarian that men and women process language exactly the same way, unlike sexist Martian culture, where men and women process language differently. You run two carefully matched experiments, one in Venusian culture and one in Martian culture, and the results clearly show that there's a significant difference in the language processing by men and women on Mars ($p < .05$, using an unpaired t test, let's say, or a one-way independent-measures ANOVA, which

is the same thing), but there's no significant difference ($p > .05$) between men and women on Venus. Does this support your linguistic theory?

Nope, sorry. The problem here is that, as Gelman and Stern (2006) put it in the title of their paper, "the difference between 'significant' and 'not significant' is not itself statistically significant". This mistake is made in the research literature all the time. Nieuwenhuis et al. (2011) found that it was particularly common in neuroscience, sad to say. They also point out a simple intuitive reason why the logic is flawed. Suppose the "significant" p value for the Martian experiment was $p = .049$ and the "non-significant" p value for the Venusian experiment was $p = .051$: would that convince you that there's a meaningful difference between the languages or the cultures? I hope not!

Solution: do a two-way independent-measures ANOVA. Take the data from your Martian and Venusian experiments, and combine them, making four cells in one two-way ANOVA table, instead of two independent two-cell one-way ANOVA tables. Now we can test for a main effect of language/culture, a main effect of gender, and crucially, also an interaction between the two. This interaction addresses the question that we really care about: does language/culture modulate how gender affects language processing?

4.2 Three colored rooms and two genders

Interactions sound great! Let's try a two-way ANOVA, starting, as usual, with Excel.

4.2.1 Three colored rooms and two genders, in Excel

As I tried to explain at the start of this chapter, to do a two-way independent-measures ANOVA in Excel, the tool you need is confusingly called **Anova: Two-Factor with Replication" (雙因子變異數分析：重複試驗)**. Remember that by "replication", Excel means that the levels for the "row" factor each have more than one data point (forget about Excel's third type of ANOVA for now, since we'll save that for the next chapter, when we discuss repeated-measures ANOVA).

We need some fake data to illustrate this. Hm, where can we get some? Let's look in Gravetter and Wallnau (2004) again. Ah, here's the example they use to introduce this type of ANOVA (on p. 490 in that old edition). Again I've kept their numbers (tiny samples, overly neat calculations), but I've changed the factors to seem more linguisticky.

So here's the wise old Chinese teacher again. While her first experiment taught her something about the importance of learning Chinese words in blue rooms, insatiable curiosity still burns deeply within her, because it seems that maybe student gender matters too. For example, maybe female students learn words better than male students (another main effect),

and/or maybe female students learn words best in blue rooms, but for male students, room color doesn't matter at all (an interaction).

So the wise old Chinese teacher decides to run a new factorial experiment that crosses room color with student gender. Her exciting (i.e., fake and pirated) results are shown in Table 9, arranged the way Excel likes them (also available in the file **ColoredRooms.txt**). Note that the thick cell borders mark out the four ANOVA cells, defined by this two-way ANOVA crossing two binary factors; the thin-bordered cells are Excel cells. Note how the room color factor defines the columns, while the gender factor defines the rows, even though the gender levels are labeled in just one cell each, while the actual measurements within each ANOVA cell appear in ranges of multiple rows (which is why Excel calls this type of ANOVA "with replication").

Table 9. The effects of room color and student gender on word learning (Excel style)

	Red	Blue	Yellow
Female	3	2	9
	1	5	9
	1	9	13
	6	7	6
	4	7	8
Male	0	3	0
	2	8	0
	0	3	0
	0	3	5
	3	3	0

To analyze this in Excel, you have to tell the Analysis ToolPak tool how many data points there are in each cell. Note that this is only a limitation of Excel, since this kind of ANOVA doesn't actually require the cells to have the same number of observations, though like *t* tests and one-way ANOVA, unequal sample sizes make the other test assumptions more important (normality and homoscedicity).

Select the whole range, including the row and column labels, though for some reason these labels only appear in Excel's descriptive statistics tables (counts, means, and so on for each cell), but not in the ANOVA table. In the case of this new experiment, the ANOVA table appears as in Table 10.

Table 10. Excel's results for two-way independent-measures ANOVA for new experiment

ANOVA

變源	SS	自由度	MS	F	P-值	臨界值
樣本	120	1	120	24	5.37E-05	4.259675
欄	60	2	30	6	0.007707	3.402832
交互作用	60	2	30	6	0.007707	3.402832
組內	120	24	5			
總和	360	29				

Since Excel doesn't show the labels in this table, you have to remember that 樣本 (sample) represents the "row" factor (Female vs. Male) and 欄 represents the "column" factor (Red vs. Blue vs. Yellow). The interaction is labeled "交互作用".

Based on these results, the Chinese teacher has three things to report, so she could write something like this: "There was a significant main effect of gender ($F(1, 24) = 24$, $MSE = 5$, $p < .05$), a significant main effect of room color ($F(2, 24) = 6$, $MSE = 5$, $p < .05$), and a significant interaction ($F(2, 24) = 6$, $MSE = 5$, $p < .05$)." Do you see where I got all these values from? If not, keep looking; it's pretty easy. She should also report the six cell means and standard deviations, which you can compute from the raw data (the means are also in Excel's descriptive statistics table).

As usual with interactions, a plot would help make the results a lot more clear. Excel's descriptive statistics tables give you the cell means that you need to plot, though they're not arranged in a convenient way. So let's move them around to make a little table as shown in Table 11, and then use them to make the bar plot in Figure 3.

Table 11. Cell means arranged for plotting in Excel

	Red	Blue	Yellow
Female	3	6	9
Male	1	4	1

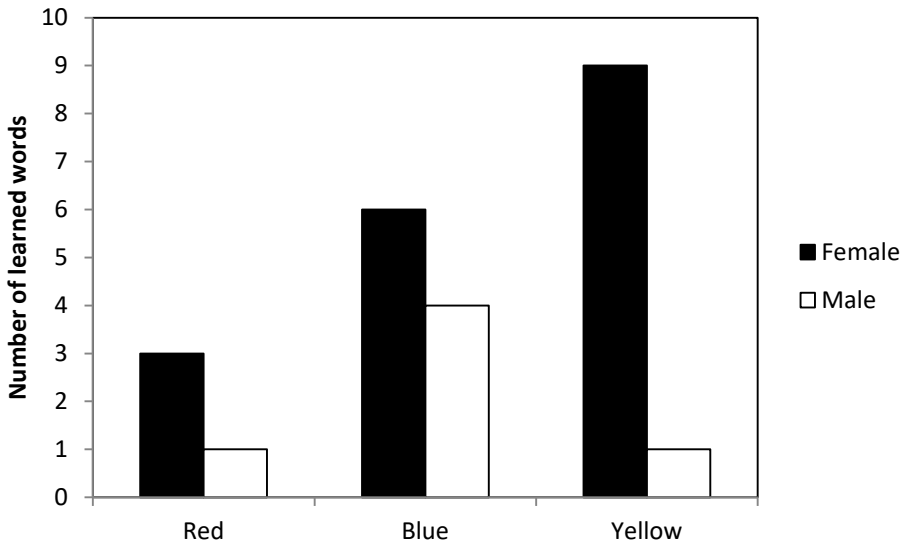

Figure 3. Results of new experiment (Excel style bar plot)

Now it's a bit clearer what each of the ANOVA results mean. To see the main effect of color, you have to imagine the average score for each color, averaging across the black and white pair of bars (roughly at the midpoint between the tops of both bars, which will work if the distributions aren't too skewed). This shows that red is overall the worst color, whereas blue and yellow are about the same. In fact, they're exactly the same, as shown by Excel's Total (總和) summary statistics table; the overall color means (ignoring gender) are 2, 5, and 5, for Red, Blue, and Yellow, respectively.

Similarly, to see the main effect of gender, you have to imagine averaging across the white bars for males and across the black bars for females; the mean female score seems to be higher. You can get the actual numbers from Excel's other descriptive statistics tables, in cell in the total (總和) column of the mean (平均) row, which shows that the overall Female mean is 6, while the overall Male mean is only 2.

As for the interaction, the pattern seems to be that yellow is the best color for women, whereas for the men the best color is still blue (maybe Experiment 1 only had men in it?).

4.2.2 Three colored rooms and two genders, in R

Of course, we can also do the same ANOVA in R, using the **aov()** function. As usual, R wants the data arranged in columns, with separate columns for the dependent variable (the Learning scores) and for the two independent variables (Gender, with levels Female and Male; Color, with levels Red, Blue, and Yellow). To save you typing and loading, you can just run the following code to create the data frame **exp2** (for Experiment 2).

```
exp2 = data.frame(Gender = c(rep("Female",15),rep("Male",15)),    # F+M
  Color = rep(c(rep("Red",5), rep("Blue",5), rep("Yellow",5)),2),    # RBY+RBY
  Learning=c(c(3,1,1,6,4), c(2,5,9,7,7), c(9,9,13,6,8),    # F: RBY
                c(0,2,0,0,3), c(3,8,3,3,3), c(0,0,0,5,0)))      # M: RBY

head(exp2) # See what it looks like
```

	Gender	Color	Learning
1	Female	Red	3
2	Female	Red	1
3	Female	Red	1
4	Female	Red	6
5	Female	Red	4
6	Female	Blue	2

Then we use the following commands to replicate Excel's results. As before, I create and named the **aov** object first, since we'll be doing a couple further things with it. The * symbol means that this model tests not just the main effects of Gender and Color, but also the interaction between these two factors. As we'll see shortly, it's not a coincidence that this symbol is the same as R's symbol for multiplication. If we had used + instead of *, the analysis would only give the two main effects, because we'd be telling R that we wanted an **additive** model, ignoring any possible interaction. Again, it's not a coincidence that + looks like an arithmetical symbol.

```
colorgender.aov = aov(Learning ~ Gender * Color, data = exp2)
summary(colorgender.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Gender	1	120	120	24	5.37e-05	***
Color	1	60	30	6	0.00771	**
Gender:Color	2	69	30	6	0.00771	**
Residuals	24	120	5			

Go back and compare this with Excel's ANOVA table, shown earlier in Table 10. It's the same table, isn't it? It is, except for the usual differences in terminology, and the fact that Excel also gives values for estimating effect size. Specifically, Excel also gives the critical values for computing confidence intervals, and the totals for computing how much of the data variance is captured by the model; R can give you this information as well, but you have to ask for it (as I'll explain in the next chapter).

Note also that R indicates the interaction with a colon (:) rather than a star (*). That's because a star in a model says that we want to test not just the interaction, but also the main effects; the colon symbolizes the actual interaction itself.

Am I lying? There's an easy way to find out: if what I say is true, then in R's model formula notation, A*B must be synonymous with A+B+A:B. So if we run the following command, we should get exactly the same results as above. Try it and see!

summary(aov(Learning ~ Gender + Color + Gender:Color, data = exp2))

Because of R's formula notation, we have a great deal of power to create all sorts of other types of ANOVA models, far beyond anything that Excel can do. Below, notice the use of the parentheses and various arithmetical symbols: not just * and +, but - and the power symbol ^ as well. You can try out the first two yourself, using Gender and Color for the factors A and B; the other ones you'll just have to imagine for now (or make up your own fake data).

aov(Y~A+B) # Test only main effects in a two-way ANOVA
aov(Y~A*B-A:B) # Test only main effects in a two-way ANOVA (same as above)
aov(Y~A*B*C) # Three-way ANOVA
aov(Y~(A+B+C)^2) # Only test two-way interactions in a three-way ANOVA
aov(Y~A*(B+C)) # Test A, B, C, A:B and A:C, but not B:C or A:B:C

But regarding this power, remember Spiderman's words of wisdom again. There are two competing forces to consider when choosing a statistical model. On the one hand, the most objective approach is to test all of the factors and all interactions implied by your research design (a so-called **maximal model**), because if you drop out some factors arbitrarily, you might be accused of **cherry-picking** just certain effects. But on the other hand, you should try to keep your analysis as simple as possible: two-way interactions may have a reasonable real-life interpretation, but three-way and higher interactions quickly get very confusing. This balance between try-everything vs. keep-it-simple will become really important when we get to multiple regression.

Anyway, the R analysis confirms that we have a significant interaction, so we need to plot it to get a sense of what it means. We could make one by hand, computing the means and then plugging them into **barplot()** or **interaction.plot()**, but there's a much easier way to do this, if we install another package. We'll be using this package in later chapters anyway.

The package is called **effects** (Fox, 2003; Fox & Hong, 2009; that's the same Fox who created the user-friendly **Rcmdr** package mentioned earlier in this book). Please install it, and then run the following little bit of code. This will create the colorful plot in Figure 4 (the colors come by default). This isn't necessarily something you'd want to put into a public report (a bar plot might be more familiar-looking to your readers), but an interaction plot like this is certainly useful for you, as a researcher, to get a feeling for your own results.

library(effects) # You have to install it first
plot(allEffects(aov(Learning ~ Gender * Color, data = exp2)))

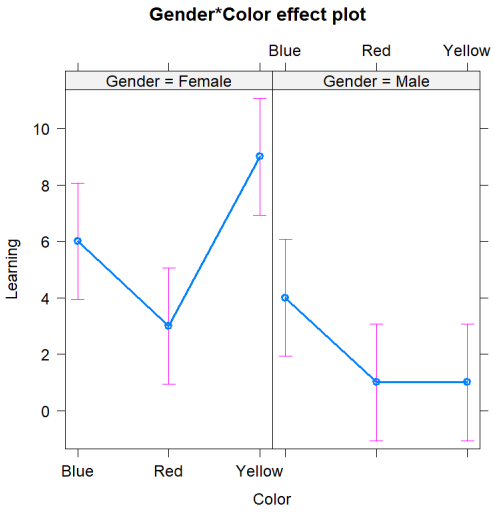

Figure 4. Interaction plot for Learning and Gender

What are those vertical red lines going through each black dot? Why, they're 95% confidence intervals for the ANOVA model, which give you a general impression of the noisiness of the data and the statistical significance of the model (as I just said, we'll explain a bit more about how they're calculated in the next chapter).

Since Color is still a multi-level factor in this new experiment, we might want to know if its levels are significantly different from each other. Just as with the one-way ANOVA in the first experiment, we can run a Tukey test to find out:

TukeyHSD(colorgender.aov)

The text output for this is very large, so I'll just let you look at it yourself. Notice that it consists of three tables: one looking at the difference between the two genders (p = 5.37e-05), one looking at the differences between pairs of colors (p = 0.0164705 for Red-Blue and for Yellow-Red, so this time Yellow stands out as best for learning), and one comparing each of the six ANOVA cells (defined by the 2 × 3 design) with the other cells in its row or column (e.g., Male:Blue is compared with Female:Blue, Male:Red and Male:Yellow, but with no other cell). The significant comparisons in this third table are Male:Red-Female:Blue, Male:Yellow-Female:Blue, and Female:Yellow-Male:Blue (all p = 0.0185503), Female:Yellow-Female:Red (p = 0.0034418), and Female:Yellow-Male:Red andMale:Yellow-Female:Yellow (both p = 0.0001055).

That's a lot of complicated data, and it's not clear what practical linguistic information we really learn from all of it. Probably for this reason, post-hoc tests aren't much used with anything but one-way ANOVA models.

4.3 More about the math behind independent-measures two-way ANOVA

A two-way ANOVA gives us three results for the price of one: it tests the significance of each of the two main effects, and also tests their interaction. How does it manage to do this?

A hint comes from R's formula syntax, and the notion of additive model (i.e., a model without an interaction). Because of the deep connection between ANOVA and regression, an ANOVA model actually represents a kind of **linear equation**. Schematically, an additive model looks something like this:

A two-factor additive ANOVA:

Dependent variable = Factor A + Factor B + Error

If the mode includes an interaction, then the linear equation also includes the **product** (乘積) of the two factors:

A full two-way ANOVA:

Dependent variable = Factor A + Factor B + Factor A × Factor B + Error

That × symbol isn't a metaphor: computing the interaction literally involves multiplying the two factors values together. I'll explain exactly how this works when we get to the multiple regression chapter.

A linear equation like this involves the same kind of **partitioning of variance** that we saw in the simpler ANOVA formulas:

Total variability = Variability A + Variability B + Interaction variability + Error

Mathematically, the partitioning means that for two-way independent-measures ANOVA, the *SS* values are as follows:

Two-way ANOVA SS_{total}: $SS_{total} = SS_{between} + SS_{error} = SS_A + SS_B + SS_{A\times B} + SS_{error}$

SS_{total} and $SS_{between}$ are computed exactly the same way as for the one-way ANOVA, and SS_A and SS_B are computed the same way as $SS_{between}$, but only relative to each factor (A and B). This leaves $SS_{A\times B}$ as what's left over in $SS_{between}$ after you subtract away SS_A and SS_B.

As with any ANOVA, the ultimate goal is to compute the *F* ratios for *A*, *B*, and *A* × *B*, each relative to SS_{error}, which, in any independent-measures ANOVA, is the same as SS_{within}.

And that's good enough for now! We'll come back to these concepts in a bit more computational detail in the other ANOVA chapter and the multiple regression chapters, since they'll have practical uses when we estimate effect size and compute regression interactions.

5. Non-parametric alternatives to ANOVA

Did you notice anything missing in this chapter? I'll give you hint: look at the title of this section. Well, this section is going to be very short, because even though ranked correlation (Spearman correlation) and heteroscedastic *t* tests (Welch's test) are relatively widely used, in the world of ANOVA (or ANOVA-like data sets), people don't worry so much about violations of the usual parametric assumptions. (This relaxed attitude is unfortunately not the case with repeated-measures ANOVA, so we'll need to spend more time in the next chapter dealing with a special assumption of this kind of ANOVA.)

Just like the ordinary (homoscedastic) unpaired *t* test, independent-measures ANOVA assumes that all of the cell samples come from population(s) with the same variance. Like all parametric tests, ANOVA is robust to violations of assumptions like this, especially if the overall sample size is not too tiny and each cell has (almost) the same number of data points. A rule of thumb that I've seen (I can't remember where!) is that you don't really have to worry about heteroscedasticity in ANOVA unless maximum cell variance is no more than four times larger than the minimum cell variance. In mostly normal data, which is what you get from most types of linguistics studies, this problem just isn't going to arise, though you might worry a bit more if you're running an ANOVA for a lexical analysis (e.g., of a corpus or dictionary), since Zipf's law can create some pretty extreme skew (but you should be lognorming or using categorical tests wherever you can anyway).

How could we find out if our multi-level factor shows heteroscedacity? One method is to something called **Levene's test** (proposed by an otherwise unexciting guy named Howard Levene in 1960). This tests the null hypothesis that the populations have the same variance (i.e., are homoscedastic). You can run this test in R, although to do so, you first have to install the **car** package; the name stands for "companion to applied regression", part of the title of the book by Fox and Weisberg (2011) (yes, that's the same Fox again). As with the **effects** package, we'll be using the **car** package later in this book anyway.

The **car** package function is called **leveneTest(formula, data)**, and as shown, it expects you to enter a model formula and the data frame that the model is applied to. So in the case of our one-way independent-measures ANOVA for the first experiment, we'd do this:

```
library(car) # Don't forget to install it first!
leveneTest(Learning~Color, data=exp1)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df  F value    Pr(>F)
group  2          0          1
      12
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
```

So $p = 1$: can't get any more non-significant than that. I guess we don't have to worry about our data being too heteroscedastic here. (By the way, the warning is just complaining that **exp1$Color** is technically a character vector, not a factor; **aov()** silently converts it to a factor first, but **leveneTest()** tells us what it's doing. R's distinction between ordinary vectors and factors will become crucial in the next chapter.)

Since an unpaired *t* test is just a special case of a one-way independent-measures ANOVA, you might be curious how the results of Levene's test works on the **BoysGirls.txt** data set we played with in the *t* test chapter. Try it!

```
# Reload data
bg = read.table("BoysGirls.txt",T)
study1 = subset(bg, bg$Study==1)
boys1 = study1$Measure[study1$Gender=="Boy"]
girls1 = study1$Measure[study1$Gender=="Girl"]

# Compare the two tests
var.test(boys1,girls1) # p = 0.339
leveneTest(Measure~Gender, data=bg) # 0.7533
```

Even though the variance test and Levene's test both rely on the *F* distribution, they don't work exactly the same way: the variance test gets the *F* value by dividing one sample variance by the other, while Levene's test does so by running a one-way ANOVA on the absolute values of the difference between each data point and the mean of the cell it comes from. That's why the two tests give different *p* values here, even though here both show $p > .05$, so we can assume that this data set doesn't violate the homoscedacity assumption.

The fact that Levene's test *p* value is larger for the boys and girls than the variance test *p* value suggests that it's hard to violate heteroscedacity in ANOVA. To demonstrate this further, let's try faking some data where the cells vary a lot in variance, by modifying our Experiment 1 data. Even with the ridiculously different (and tiny) samples below, we still don't get a significant result in Levene's test:

```
exp1.het = data.frame(Color = c(rep("Red",5), rep("Blue",5), rep("Yellow",5)),
  Learning=c(c(1,1,1,1,1),c(1,1,1000,1,1),c(1,1,1,1,1)))
leveneTest(Learning~Color, data = exp1.het) # p = 0.3966
```

Supposing we did come across a significant violation of ANOVA's homoscedasticity assumption anyway. What are our options (besides not worrying about it, which most people do)? For a one-way independent-measures ANOVA, we could use a **Welch ANOVA** (**one-way analysis of means not assuming equal variances**), a generalization of Welch's *t* test, which you can run using R's base function **oneway.test()**, with the **var.equal** argument left at its default setting of **FALSE**. For example, compare the two ANOVA analyses of the homoscedastic **exp1** data (try it!). The *p* value for the Welch ANOVA is higher than that for the ordinary ANOVA, since it makes fewer assumptions, its power is less, and thus its Type II error rate is higher.

```
summary(aov(Learning~Color, data = exp1)) # p = 0.00177
oneway.test(Learning~Color, data = exp1) # p = 0.007043
```

Another option would be used the ranking-based **Kruskal-Wallis test** (named after American statisticians William Kruskal [1919-2005] and W. Allen Wallis [1912-1998]). As a generalization of the Mann-Whitney *U* test that we saw in the *t* test chapter, it ignores everything about the data except the ranks, so violations of homoscedasticity and normality don't matter. But as you can see by running the code below, its Type II error rate is even higher than the Welch ANOVA (note the higher *p* value):

```
kruskal.test(Learning~Color, data = exp1) # p = 0.01133
```

Perhaps the most sophisticated option would be to run a regression with something called **White's adjustment** (proposed by an American economist named Halbert White [1950-2012]). This is a general trick that allows you to partially undo the bad effects of heteroscedacity, not just in an ANOVA, but also in a regression or related models. You can run it using the **Anova()** function in the **car** package (so remember to load this package first, if it's not already running), with the argument **white.adjust** set to **TRUE**. Note how close the *p* value is to that given by the ordinary ANOVA, showing that this method avoids Type II errors better than the others I mentioned.

```
library(car) # Only if it's not already running
Anova(aov(Learning~Color, data = exp1), white.adjust=TRUE) # p = 0.004874
```

6. Conclusions

You've always wondered what ANOVA was, and now you know! It stands for "analysis of variance", because the genius idea at the heart of it is to see all data in terms of variation, and to seek patterns in it by partitioning the variance into "interesting" (associated with the

fixed factors in your model) and "boring" (the residuals, or the noise that your model can't explain). In this chapter we focused on the simplest kind of ANOVA, namely independent-measures ANOVA, where all of your data points are independent of all others (rather than being grouped by units, like speakers or words, as in a repeated-measures ANOVA, which we'll discuss in the next chapter). In other words, the independent-measures ANOVA is a generalization of the unpaired *t* test. So like *t* tests, the goal of an ANOVA is to test the null hypothesis that the sample means (i.e., the means of the cells defined by your factors and their levels) are the same (i.e., are sampled from populations with the same means). What's new is that we can now test any number of samples at the same time. Excel comes with built-in tools for running one-way independent-measures ANOVA (with samples of any size), where there's just one factor with two or more levels, and for running two-way independent-measures ANOVA (with samples of the same size), where there are two factors, each with two or more levels, and we cross them. In the latter type of model, the ANOVA gives us three results: an analysis of the first factor as a main effect, the same for the second factor, and an analysis of their interaction (which can be especially useful in linguistics). R is much more powerful than Excel, allowing us to run ANOVA with any cell sizes, any number of factors, whether crossed completely or only partially (including additive models, which don't bother testing for an interaction). Special R functions can also help us compare levels within a multi-level factor (e.g., Tukey's post-hoc test) or analyze factorial data where assumptions about equal variance or normality are false.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, *57*(1), 289-300.

Chomsky, N., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry, 8*, 425-504.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage.

Fisher, R. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 503-513.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1-27.

Fox, J., & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software, 32*(1), 1-24.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (second edition). Sage.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research

hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University, New York, NY.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.

Gravetter, Frederick J., & Wallnau, Larry B. (2004). *Statistics for the behavioral sciences* (6th edition). Wadsworth. [Newer editions have the same examples on different pages.]

Myers, J. (2009a). Syntactic judgment experiments. *Language & Linguistics Compass, 3* (1), 406-423.

Myers, J. (2009b). The design and analysis of small-scale syntactic judgment experiments. *Lingua, 119*, 425-444.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience, 14*(9), 1105-1107.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *An invitation to cognitive science, vol. 4: Methods, models, and conceptual issues* (pp. 703-863). MIT Press.