

Exercises

1. Fill in the blanks. The Analysis of Variance table breaks up the variation into two sources, that from _____ and from _____.
2. Why doesn't failing to reject the null hypothesis prove that there are no differences between treatment means?
3. An experiment to determine the effect of several methods of preparing for use in commercial yogurt was conducted by a food science research group. Three batches of yogurt were prepared using each of three methods: traditional, ultra filtration, and reverse osmosis. An expert then tasted each of the 9 samples, presented in random order, and judged them on a scale from 1 to 10. A partially complete analysis of variance table of the data is shown in table 4.8

<i>Source</i>	<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Mean Square</i>	<i>F - ratio</i>
Treatment	17.30			
Residual	0.460			
Total (corrected)	17.769			

Table 4.8 A partial ANOVA table for the yogurt data

- (a) Calculate the mean square of the treatments and the mean square of the error.
- (b) Form the F statistic by dividing the two mean squares.
- (c) The p-value of this F statistic turns out to be .000017. What does this say about the null hypothesis of equal means?
- (d) What assumptions have you made in order to answer part (c)?

Comparing more than two treatments

- (e) What would you like to see in order to justify the conclusions of the F test?
 - (f) What is the average size of the experimental standard deviation in the judges assessment?
4. Four smokestack filter designs are being tested to see which, if any, filter particulates more effectively. The amount of particulate matter escaping through the filter was captured at the end of the smokestack and is shown below in table 4.9 with data in parts per million of particulates.

A	B	C	D
2	4	8	3
1	7	6	2
3	1	7	1
4	4	7	1
3	5	8	3

Table 4.9 Particulates measured from 4 smokestacks in ppm.

- (a) Decompose the data into the grand average, the filter effect and the residual.
- (b) Calculate the sums of squares, and the mean squares for each effect.
- (c) Construct the ANOVA table and test the null hypothesis.
- (d) Plot the residuals and check the assumptions of the F test.
- (e) Summarize your conclusions.

A little insight from geometry -- Optional section

Let's look at the decomposition geometrically. What we want to do is to compare the size of the treatment effects column (or vector) **T**, to the size of the residuals column **R** (after adjusting them for their degrees of freedom). If you think of each effect column as a vector, the sums of squares is just the (squared)length of the vector.

Notice that the deviations vector **D** is just the sum of two other vectors: **T** and **R**. So they form a triangle.

$$\mathbf{D}=\mathbf{T}+\mathbf{R}.$$

But, by the sum of squares formula:

$S_D = S_T+S_R$, so the squares of their lengths add as well. When does this happen? When does the sum of squares of the length of one side of the triangle equal the sum of the squares of the lengths of the other two sides?

Yes -- a right triangle. Thank you, Pythagoras.

% ADD TRIANGLE PICTURE HERE

Figure 4.7 The vectors **T** and **R** add up to the vector **D**, and the sums of the squares of their lengths add up as well.

Now, what do these three vectors of the triangle mean, and why do we care about them? The hypotenuse, which contains the deviations from the average, shows how much the observations vary around the overall average. Its squared length represents the "total variation" in the experiment. We have broken this total variation into two components:

S_T the part due to the treatment levels, the signal, and S_R , the part *not* due to the change in the treatment levels -- the noise. When most of the variation in the experiment is due to noise, I can't distinguish the treatment levels. But, if most of the variation is due to the differences in the treatment levels as opposed to the noise, I'll be able to reject the null hypothesis.

Comparing more than two treatments

However, in order to compare these two lengths, to create a signal to noise ratio, I need to adjust them for their degrees of freedom. The test statistic becomes:

$$F = \frac{S_T / \mathbf{n}_T}{S_R / \mathbf{n}_R} = \frac{MS_T}{MS_R}$$

These adjusted sums of squares are known as mean squares, and if the mean squares are roughly the same size, this says the treatment effects are indistinguishable from noise. Under various assumptions (that we will talk about soon), and the assumption that the null hypothesis is true, this ratio of mean squares has an F distribution. So, if this ratio is too large (judged by looking at the reference F-distribution) this is evidence that the treatment means are different.

ANOVA -- Take Two: Some Mathematics ¹

What exactly is the F-test doing? To get some more insight into the F-ratio, let's look at both the numerator and denominator of it in more detail. In order to start, let's estimate the variance of the experimental error.

But what exactly do I mean by *the* variance of the experimental error? There are k different treatment levels, and the experimental error may be different at each level. To get around this, we will *assume* that the error has the same variance at each treatment level. Then, it makes sense to average these in some

Let the number of observations getting treatment level t be denoted by n_t . Now, to estimate the error variance, we'll first estimate the error variance in each group (*i.e.* at each treatment level). For treatment level t , we'd get:

¹ Note: This section is a bit more technical and can be skipped

$$s_t^2 = \frac{\sum (y_{it} - \bar{y}_t)^2}{n_t - 1}$$

In other words, it's the sum of squares of the residuals in treatment level t divided by the degrees of freedom, $n_t - 1$.

To get an estimate of the overall experimental error variance, we pool all these individual estimates, just as we did for the (equal variance) t test in the last chapter. That is, we average them, weighted by their degrees of freedom ($n_t - 1$).

So, the pooled estimate of the error variance is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + \dots + (n_k - 1)s_k^2}{n_1 + \dots + n_k - k}$$

where k is the number of treatment levels. Now, look closely at the numerator. Each term, $(n_t - 1)s_t^2$ is just the sum of squares of the residuals at that treatment level. So the numerator of this expression is just the total sum of squares of the residuals, S_R . The denominator is just $N - k$, where N is the total number of observations in the experiment. So, s_p^2 is the same as $S_R / (N - k)$ which is just MS_R , the mean square of the residuals.

This quantity, the mean square of the residuals is a very important quantity. Its square root is the estimate of the experiment wide error standard deviation. In our example above it is found in the second line of table 4.6 where it is seen to be equal to 5.6. As we saw, this, in turn, means that the error standard deviation is estimated to be $\sqrt{5.6} = 2.37$. An important question to ask before proceeding is whether this is a reasonable estimate of the error that one would expect by replicating an individual observation at some treatment level. Moreover, is this a reasonable estimate of the error for all the groups? If not, the analysis may be invalid.

Now, to the numerator of the F statistic. What is S_T / \mathbf{n}_T ? Let's look at it more closely:

$$S_T / \mathbf{n}_T = \frac{\sum n_t (\bar{y}_t - \bar{\bar{y}})^2}{k - 1}$$

Comparing more than two treatments

This is a little more easily understood if we assume for a moment that all the treatment levels have the same number of observations, so that n_t is just n .

$$S_T / \mathbf{n}_T = \frac{n \sum (\bar{y}_t - \bar{\bar{y}})^2}{k - 1}$$

Now, let's assume the null hypothesis is true. Then why do the treatment averages vary at all? Only because of chance variation. How much should they vary? Well, remember that the variance of an average is σ^2/n . So, if I treat each of the k treatment averages \bar{y}_t as an observation, and take their sample variance, this new sample variance then the sample variance of these k "observations" will estimate σ^2/n . But the sample variance of these k treatment averages is

$$\frac{\sum (\bar{y}_t - \bar{\bar{y}})^2}{k - 1} \text{ which estimates } \sigma^2/n \text{ when the null hypothesis is true.}$$

So, n times this estimates σ^2 when all the treatment means are actually the same. So, under H_0 , $S_T / \mathbf{n}_T = MS_T$ estimates σ^2 too!

So what, you ask? Well, let's see what we've got. We know that MS_R estimates the error variance. And now, if H_0 is true, MS_T estimates the same quantity. One's based on averaging the true error variance in each group; the other's based on the assumption that the groups all have the same mean, the null hypothesis. So we can compare the ratio of the two to see whether the null hypothesis is reasonable. If it's true, the ratio should be near 1, and in fact, it's distributed as an F-statistic with the associated degrees of freedom. So if the ratio is a reasonable value from the F-distribution, then this would mean that the results are consistent with the null hypothesis. On the other hand, if the treatment averages are too far apart from each other, then S_T will be big, making MS_T too big and thus the F ratio too large. And this then provides evidence against the null hypothesis.

In this case, S_T / \mathbf{n}_T actually is estimating σ^2 PLUS a term that contains the true treatment mean differences as well. More precisely, the expected value of the Mean Square of Treatment is:

$$E(MS_T) = \sigma^2 + \frac{\sum n_t \tau_t^2}{k - 1}$$

where k is the number of treatment levels, and τ_t is the size of the true effect of treatment level t . This equation is *always* true. So, when the treatment effects are all 0, and H_0 is true, then MS_T estimates *only* σ^2 and the ratio of MS_T/MS_R is really distributed as an F statistic. But if the ratio is too big, then it is probably due to the fact that the treatment effects are *not* all zero and thus we conclude that H_0 is false.

This is what the F-test tests. When we look at the F statistic, we ask: can the size of this ratio be explained by chance, or is it so big as to cast doubt on the null hypothesis? What we have done is to convert a hypothesis about means being equal to a test of whether two variances are the same. This is why a test of differences in means is called an Analysis of Variance (and why the procedure wasn't discovered until the 20th century). We have mathematically tested what our eyes test in the box plots: whether the differences in the treatments (the signal) is large compared to the noise.

And by how much

Up to now we have concentrated on seeing if there are any differences at all among the treatment means. Suppose we've been assigned to see which among 10 vendors is the best. We test all 10, run an analysis of variance and find a large F statistic with a very small p-value.

Let's suppose that the diagnostics plots reveal no problems with the model assumptions. So, we reject the null hypothesis and proudly announce to the executive committee that the vendors' performances are not all equal. Are they surprised? I doubt the executive committee will be very impressed. They just might want to know which one is best, and how much better it is. Unfortunately, the F-test tells us nothing about that, so we have some more work to do.

To start things simply, suppose we want to know if two specific treatment means are the same:

Q: Is $\mu_i = \mu_j$?

We could use a t-test as in Chapter 3. What could be wrong with that? The answer is, if this is the *only* pair we want to test, everything is fine.

Comparing more than two treatments

But, imagine for a minute that we had lots of treatment groups, say 15. To compare all pairs of vendors against each other, I'd need to do more than 100 t-tests. Now, as much fun as doing 100 t-test is, there's a problem. Even if there are absolutely no differences in the means of these fifteen groups, if I do 100 t-tests, each with an α level of say .05, I will get a false positive (a type I error) 5% of the time on each test. So in 100 tests, I'll get about 5 false positive results. Five of the pairs will pass the statistical test, even when their means are identical. Since people are very good at coming up with explanations for nearly anything, someone will probably come up with a not unreasonable explanation for each of these results.

And, since about 5 of them will have p-values less than .05, they'll be statistically validated! No one will consider that we've run all these tests and on the average we should get 5 type I errors. To make matters worse, there's over a 99% chance that we'll make at least 1 type I error. So the real α level has increased from .05 to over .99.

For 6 or even 4 groups, the situation is not nearly as terrible, but the *overall* or experiment-wide type I error is still increased if we run a bunch of separate t-tests between pairs. In fact, if you do a lot of pairwise t-tests whenever you have several groups, you will see differences sometimes even when the F-test says that there's no evidence that all the group means are different.

There are several ways out of this dilemma, and, in fact, this field of study, called **multiple comparisons** is huge. We are just going to mention one of two of the most common practices. The interested reader should check the bibliography at the end of the chapter.

A very simple method, called a *Bonferroni* adjustment, is to reduce α on each individual test enough so that the *overall* type I error is back to what we want. The trouble here is that if you plan on making a lot of comparisons, it makes seeing each individual pairwise difference very difficult and increases the chance of failing to detect any real differences (type II errors).

For our example, with $k=4$ groups, if we want overall error rate, $\alpha^* = .05$, then we'd need $\alpha = .0085$ for each test, which means we would have to replace the value $t_{20,.025} = 2.09$ by the value $t_{20,.0042} = 2.92$, resulting in confidence intervals over 30% wider.