CHAPTER 28

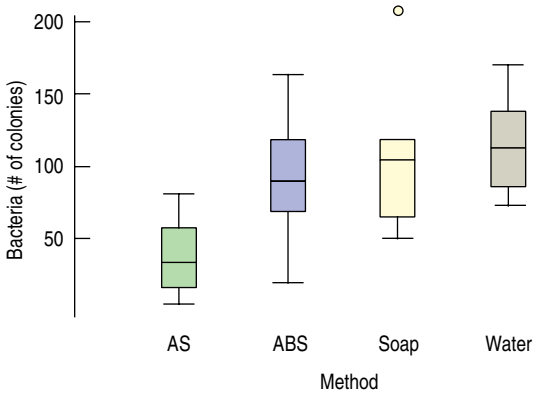
Analysis of Variance

WHO	Hand washings by four different methods, assigned randomly and replicated 8 times each
WHAT	Number of bacteria colonies
HOW	Sterile media plates incubated at 36 °C for 2 days

Did you wash your hands with soap before eating? You've undoubtedly been asked that question a few times in your life. Mom knows that washing with soap eliminates most of the germs you've managed to collect on your hands. Or does it? A student decided to investigate just how effective washing with soap is in eliminating bacteria. To do this she tested four different methods—washing with water only, washing with regular soap, washing with antibacterial soap (ABS), and spraying hands with antibacterial spray (AS) (containing 65% ethanol as an active ingredient). Her experiment consisted of one experimental factor, the washing method, at four levels.

She suspected that the number of bacteria on her hands before washing might vary considerably from day to day. To help even out the effects of those changes, she generated random numbers to determine the order of the four treatments. Each morning she washed her hands according to the treatment randomly chosen. Then she placed her right hand on a sterile media plate designed to encourage bacteria growth. She incubated each plate for 2 days at 36°C, after which she counted the bacteria colonies. She replicated this procedure 8 times for each of the four treatments.

A side-by-side boxplot of the numbers of colonies seems to show some differences among the treatments:



Boxplots of the bacteria colony counts for the four different washing methods suggest some differences between treatments. **Figure 28.1**

When we first looked at a quantitative variable measured for each of several groups in Chapter 5, we displayed the data this way with side-by-side boxplots. And when we compared the boxes, we asked whether the centers seemed to differ, using the spreads of the boxes to judge the size of the differences. Now we want to quantify this by asking whether the washing methods have the same mean bacteria count. We'll make the same kind of comparison, comparing the variability among the means with the spreads of the boxes. It looks like the alcohol spray has lower bacteria counts, but as always, we're skeptical. Could it be that the four methods really have the same mean counts and we just *happened* to get a difference like this because of natural sampling variability?

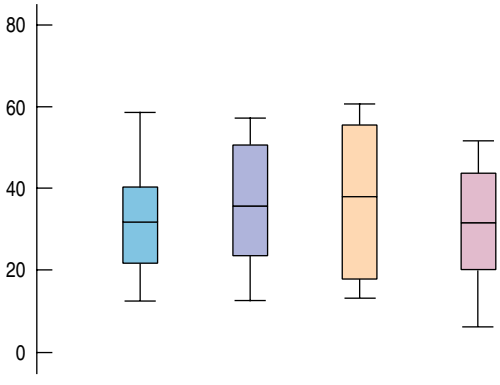
What is the null hypothesis here? It seems natural to start with the hypothesis that *all the group means are equal*. That would say it doesn't matter what method you use to wash your hands because the mean bacteria count will be the same. We know, however, that even if there were no differences at all in the *means* (for example, if someone replaced all the solutions with water) there would still be sample-to-sample differences. We want to see, statistically, whether differences as large as those observed in the experiment could naturally occur by chance in groups that have equal means. If we find that the differences are so large that they would occur only very infrequently in such groups, then, as we've done with other hypothesis tests, we'll reject the null hypothesis and conclude that the group means really are different.[1]

Are the Means of Several Groups Equal?

A S **The PERK Study of Radial Keratotomy.** Surgery on the eyeball to correct vision has developed over many years of research and practice. Here is a video story of an important medical trial that helped in this progress.

We already know how to use a *t*-test to see whether *two* groups have equal means. (If you don't remember, now is a good time to reread Chapter 24.) But those tests can't handle comparing the means of several groups at once. When we stepped up from comparing two proportions (with a *z*-test) to comparing several proportions, we found that a new sampling distribution model, the chi-square model, did the trick. For comparing several means, there is yet another sampling distribution model, called the *F*-model.

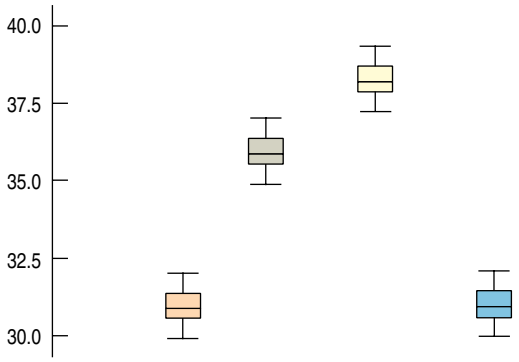
To get an idea of how it works, let's start by looking at the following two sets of boxplots:



It's hard to see the difference in the means in these boxplots because the spreads are large relative to the differences in the means.

Figure 28.2

[1] The alternative hypothesis is that "the means are not *all* equal." Be careful not to confuse that with "all the means are different." With 11 groups we could have 10 means equal to each other and 1 different. The null hypothesis would still be false.



In contrast with Figure 28.2, the smaller variation makes it much easier to see the differences among the group means. **Figure 28.3**

We're trying to decide if the means are different enough for us to reject the null hypothesis. If they're close, we'll attribute the differences to natural sampling variability. What do you think? It's easy to see that the means in the second set differ. It's hard to imagine that the means could be that far apart just from natural sampling variability alone. How about the first set? It looks like these observations *could* have occurred from treatments with the same means.[2] This much variation among groups does seem consistent with equal group means.

Believe it or not, the two sets of treatment means in both figures are the same. (They are 31, 36, 38, and 31, respectively.) Then why do the figures look so different? In the second figure, the variation *within* each group is so small that the differences *between* the means stand out. This is what we looked for when we compared boxplots by eye back in Chapter 5. And it's the central idea of the *F*-test. We compare the differences *between* the means of the groups with the variation *within* the groups. When the differences between means are large compared with the variation within the groups, we reject the null hypothesis and conclude that the means are (probably) not equal. In the first figure, the differences among the means look as though they could have arisen just from natural sampling variability from groups with equal means, so there's not enough evidence to reject H_0.

How can we make this comparison more precise statistically? All the tests we've seen have compared differences of some kind with a ruler based on an estimate of variation. And we've always done that by looking at the ratio of the statistic to that variation estimate. Here, the differences among the means will show up in the numerator, and the ruler we compare them with will be based on the underlying standard deviation—that is, on the variability *within* the treatment groups.

How Different Are They?

The challenge here is that we can't take a simple difference as we did when comparing two groups. In the hand-washing experiment, we have differences in mean bacteria counts across *four* treatments. How should we measure how different the four group means are? With only two groups, we naturally took the difference between their means as the numerator for the *t*-test. It's hard to imagine what else

[2] Of course, with a large enough sample, we can detect any differences that we like. For experiments with the same sample size, it's easier to detect the differences when the variation *within* each box is smaller.

Why variances? We've usually measured variability with standard deviations. Standard deviations have the advantage that they're in the same units as the data. Variances have the advantage that for independent variables, the variances add. Because we're talking about sums of variables, we'll stay with variances before we get back to standard deviations.

we could have done. How can we generalize that to more than two groups? When we've wanted to know how different many observations were, we measured how much they vary, and that's what we do here.

How much natural variation should we expect among the means if the null hypothesis were true? If the null hypothesis were *true*, then each of the treatment means would estimate the *same* underlying mean. If the washing methods are all the same, it's as if we're just estimating the mean bacteria count on hands that have been washed with plain water. And we have several (in our experiment, four) different, independent estimates of this mean. Here comes the clever part. We can treat these estimated means as if they were observations and simply calculate their (sample) variance. This variance is the measure we'll use to assess how different the group means are from each other. It's the generalization of the difference between means for only two groups.

The more the group means resemble each other, the smaller this variance will be. The more they differ (perhaps because the treatments actually have an effect), the larger this variance will be.

Level	*n*	Mean
Alcohol spray	8	37.5
Antibacterial soap	8	92.5
Soap	8	106.0
Water	8	117.0

For the bacteria counts, the four means are listed in the table to the left. If you took those four values, treated them as observations, and found their sample variance, you'd get 1245.08. Since the four values are *means*, this number should estimate the variance of a *mean*. With 8 observations in each group, we know that variance is $\sigma^2/8$. The estimate that we've just calculated, 1245.08, should estimate this quantity. If we want to get back to the variance of the *observations*, σ^2, we need to multiply it by 8. So $8 \times 1245.08 = 9960.64$ should estimate σ^2.

Of course, remember that when we computed the "sample" variance of these four means, we had to subtract their mean from each of these four "observations." So this estimate of σ^2 makes sense *only* if the treatments really do have the same mean. If they don't have the same overall mean—as they won't if the treatment means really are *different*—then the variance we just found won't really make sense. And in that case, the value we calculated should be *larger* than σ^2. How will we know? Can we get an independent estimate of σ^2 for comparison? (Would we ask, if the answer weren't "yes"?)

The Ruler Within

We need a suitable ruler for comparison—one based on the underlying variability in our measurements. That variability is due to the day-to-day differences in the bacteria count even when the same soap is used. Why would those counts be different? Maybe the experimenter's hands were not equally dirty, or she washed less well some days, or the plate incubation conditions varied. We randomized just so we could see past such things.

We need an independent estimate of σ^2, one that doesn't depend on the null hypothesis being true, one that won't change if the groups have different means. As in many quests, the secret is to look "within." We could look in *any* of the treatment groups and find its variance. But which one should we use? The answer is, *all* of them!

At the start of the experiment (when we randomly assigned experimental units to treatment groups), the units were drawn randomly from the same pool, so each treatment group had a sample variance that estimated the same σ^2. If the null hypothesis is true, then not much has happened to the experimental units—or at least, their means have not moved apart. It's not much of a stretch to believe that their variances haven't moved apart much either. (If the washing methods are

equivalent, then the choice of method would not affect the mean *or* the variability.) So each group variance still estimates a common σ^2.

As always, to test a null hypothesis model, we first assume that it's true. If the group variances are equal, then the common variance they all estimate is just what we've been looking for. Since all the group variances estimate the same σ^2, we can pool them to get an overall estimate of σ^2. Recall that we pooled to estimate variances when we tested the null hypothesis that two proportions were equal—and for the same reason. It's also exactly what we did in a pooled *t*-test. The variance estimate we get by pooling we'll denote, as before, by s_p^2.

Level	*n*	Mean	Std Dev	Variance
Alcohol spray	8	37.5	26.56	705.43
Antibacterial soap	8	92.5	41.96	1760.64
Soap	8	106.0	46.96	2205.24
Water	8	117.0	31.13	969.08

For the bacteria counts, the standard deviations and variances are listed to the left. If we pool the four variances (here we can just average them because all the sample sizes are equal), we'd get $s_p^2 = 1410.10$. In the pooled variance, each variance is taken around its *own* treatment mean, so the pooled estimate doesn't depend on the treatment means being equal. But the estimate from before—where we took the four means as observations and took their variance—does. That estimate gave 9960.64. That seems a lot bigger than 1410.10. Might this be evidence that the four means are not equal?

Let's see what we've got. We have an estimate of σ^2 from the variation *within* groups of 1410.10. That's traditionally called the **error mean square**[3] and written $\mathbf{MS_E}$**.** It's just the variance of the residuals. Because it's a pooled variance, we write it s_p^2. We've got a *separate* estimate of σ^2 from the variation *between* the groups of 9960.64 (by taking the variance of the four means and multiplying by *n*). At least we expect it to estimate σ^2 *if we assume the null hypothesis is true*. We call this quantity the **treatment mean square ($\mathbf{MS_T}$).**

The *F*-statistic

When the null hypothesis is true, the treatment means are equal, and both MS_E and MS_T estimate σ^2. Their ratio, then, should be close to 1.0.

When the null hypothesis is false, the MS_T will be *larger* because the treatment means are not equal. The MS_E is a pooled estimate in which the variation within each group is found around its own group mean, so differing means won't inflate it. That makes the ratio MS_T/MS_E suitable for testing the null hypothesis. When the null hypothesis is true, the ratio should be near 1. If the treatment means really are different, the numerator will tend to be larger than the denominator, and the ratio will tend to be bigger than 1.

Of course, even when the null hypothesis *is* true, the ratio will vary around 1 just due to natural sampling variability. How can we tell when it's big enough to reject the null hypothesis? To be able to tell, we need a sampling distribution model for the ratio. Sir Ronald Fisher found the sampling distribution model of the ratio in the early 20th century. In his honor we call the distribution of MS_T/MS_E the ***F*-distribution.** And we call the ratio MS_T/MS_E the ***F*-statistic.** By comparing this statistic with the appropriate *F*-distribution we (or the computer) can get a P-value.

NOTATION ALERT:

Capital F is used only for this distribution model and statistic. Fortunately, Fisher's name didn't start with a Z, a T, or an R.

[3] This terminology stretches back to the early 20th century when these methods were developed. If you think about it, that's just what variances are: means of squared differences.

The ***F*-test** is simple. It is one-tailed because any differences in the means make the *F*-statistic larger. Larger differences in the treatments' effects lead to the means being more variable, making the MS_T bigger. That makes the *F*-ratio grow. So the test is significant if the *F*-ratio is big enough. In practice, we find a P-value, and big *F*-statistic values go with small P-values.

The entire analysis is called the Analysis of Variance, commonly abbreviated **ANOVA** (and pronounced uh-NŌ-va). You might think that it should be called the analysis of means, since it's the equality of the means we're testing. But we use the *variances* within and between the groups for the test.

Animated Means Illustrate *F*-Tests. How does the *F*-test really work? This interactive activity lets you adjust the means of different groups to immediately see the effect on the *F*-statistic.

Like Student's *t*-models, the *F*-models are a family. *F*-models depend on not one, but two, degrees of freedom parameters. The degrees of freedom come from the two variance estimates and are sometimes called the *numerator df* and the *denominator df*. The *treatment mean square*, MS_T, is the sample variance of the observed treatment means. If you think of them as observations, then since there are k groups, this variance has $k - 1$ degrees of freedom. The *error mean square*, MS_E, is the pooled estimate of the variance within the groups. If there are n observations in each group, then we get $n - 1$ degrees of freedom from each for a total of $k(n - 1)$ degrees of freedom.

NOTATION ALERT:

What, first little *n* and now big *N*? In an experiment it's standard to use *N* for *all* the cases and *n* for the number in each treatment group.

A simpler way of tracking the degrees of freedom is to start with all the cases. We'll call that N. Each group has its own mean, costing us a degree of freedom—k in all. So we have $N - k$ degrees of freedom for the error. When the groups all have equal sample size, that's the same as $k(n - 1)$, but this way works even if the group sizes differ.

We say that the *F*-statistic, MS_T/MS_E, has $k - 1$ and $N - k$ degrees of freedom.

Back to Bacteria!

For the hand-washing experiment, the $MS_T = 9960.64$. The $MS_E = 1410.14$. If the treatment means were equal, the *treatment mean square* should be about the same size as the *error mean square*, about 1410. But it's 9960.64, which is 7.06 times bigger. In other words, $F = 7.06$. This *F*-statistic has $(4 - 1) = 3$ and $32 - 4 = 28$ degrees of freedom.

An *F*-value of 7.06 is bigger than 1, but we can't tell for sure whether it's big enough to reject the null hypothesis until we check the $F_{3,28}$ model to find its P-value. (Usually, that's most easily done with technology, but we can use printed tables.) It turns out the P-value is 0.0011. In other words, if the treatment means were actually equal, we would expect the ratio MS_T/MS_E to be 7.06 or larger about 11 times out of 10,000, just from natural sampling variability. That's not very likely, so we reject the null hypothesis and conclude that the means are different. We have strong evidence that the four different methods of hand washing are not equally effective at eliminating germs.

The ANOVA Table

A Simple ANOVA. There always seem to be lots of numbers in the ANOVA table, but they're really very organized. Take an animated tour of the ANOVA table.

You'll often see the mean squares and other information put into a table called the **ANOVA table.** Here's the table for the soaps:

Analysis of Variance Table

Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Soaps	29882	3	9960.64	7.0636	0.0011
Error	39484	28	1410.14		
Total	69366	31			

This table has a long tradition stretching back to when ANOVA calculations were done by hand. Major research labs had rooms full of mechanical calculators operated by women. (Yes, always women; women were thought—by the men in charge, at least—to be more careful at such an exacting task.) Three women would perform each calculation, and if any two of them agreed on the answer, it was taken as the correct value.

The ANOVA table was originally designed to hold the intermediate calculations so they could easily be repeated. We don't need to look at the Sum of Squares column, nor really at the Mean Square column. (It may be of interest that the mean squares are just the sum of squares divided by their respective degrees of freedom.) The *F*-ratio and the square root of the MS_E are the important quantities. You'll almost always see ANOVA results presented in a table like this, though. After nearly a century of writing the table this way, statisticians (and their technology) aren't going to change. Even though the table was designed to facilitate hand calculation, computer programs that compute ANOVAs still present the results in this form. Usually the P-value is found next to the *F*-ratio itself.[4]

Total? The ANOVA table includes a final line labeled "Total." It's easy to see that the total sum of squares is just the sum of the treatment and error SSs. You don't even need your calculator to see that the degrees of freedom add up. There is a surprise here, though. If you divide the total SS by its degrees of freedom, you get the variance of all the responses. (Of course, the square root of that is the response standard deviation.) Bet you didn't see that coming. You can prove that this always works out this way with a bit of algebra, but it doesn't really matter for our analyses, so we won't bother. When ANOVAs were found by hand, however, this was an important check on the calculations.

You'll sometimes see the two mean squares referred to as the *mean square between* and the *mean square within*—especially when we test data from observational studies rather than experiments. ANOVA is often used for such observational data, and as long as certain conditions are satisfied, there's no problem with using it in that context.

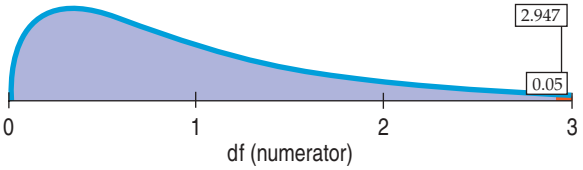
The *F*-table

Usually, you'll get the P-value for the *F*-statistic from technology. Any software program performing an ANOVA will automatically "look up" the appropriate one-sided P-value for the *F*-statistic. If you want to do it yourself, you'll need an *F*-table. (There's one on the CD called **Table F.**) *F*-tables are usually printed only for a few values of α, often 0.05, 0.01, and 0.001. They give the critical value of the *F*-statistic with the appropriate number of degrees of freedom determined by your data, for the α-level that you select. If your *F*-statistic is greater than that

[4] The P-value column may be labeled with a title such as "Prob > F," "sig," or "Prob." Don't let that confuse you. It's just the P-value.

value, you know that its P-value is less than that α level. So, you'll be able to tell whether the P-value is greater or less than 0.05, 0.01, or 0.001, but to be more precise, you'll need technology (or an interactive table like the one in *ActivStats*).

Here's an excerpt from an *F*-table for $\alpha = 0.05$:



df (denominator)	1	2	3	4	5	6	7
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313

Part of an *F*-table showing critical values for $\alpha = 0.05$ and highlighting the critical value, 2.947, for 3 and 28 degrees of freedom. We can see that only 5% of the values will be greater than 2.947 with this combination of degrees of freedom.

Figure 28.4

The *F*-Tables. View and interact with an animated exploration of the *F*-tables. On *ActivStats* you can watch the shape of the *F*-distribution change as you drag your mouse across and down the degrees of freedom. That's hard to see any other way.

Notice that the critical value for 3 and 28 degrees of freedom at $\alpha = 0.05$ is 2.947. Since our *F*-statistic of 7.06 is larger than this critical value, we know that the P-value is less than 0.05. We could also look up the critical value for $\alpha = 0.01$ and find that it's 4.568 and the critical value for $\alpha = 0.001$ is 7.193. So our *F*-statistic sits between the two critical values 0.01 and 0.0001, and our P-value is slightly *greater* than 0.001. Technology can find the value precisely. It turns out to be 0.011.

The ANOVA Model

We can write a simple model that describes the data with the components that we need to find the Analysis of Variance. Each observation can be found as the sum of two quantities: the mean of its treatment group, and a leftover residual.

We can write the i-th observation in the k-th group as

$$y_{ik} = \bar{y}_k + e_{ik}$$

where $\bar{y}_k$ is the mean of the k-th group, and e_{ik} is the "error" or residual for the i-th observation in group k: $e_{ik} = y_{ik} - \bar{y}_k$.

The MS_E is the variance of the errors. The MS_T comes from the variance of the group means.

As we did with regression, we can think of this as a model that gives a fitted or predicted value for each observation as

$$\hat{y}_{ik} = \bar{y}_k.$$

That is, we predict that each observation will be like its group mean—a reasonable, if somewhat simplistic, prediction. (The more complicated experimental designs we saw in Chapter 13 lead to more complicated ANOVA models and more complex tests. Those are beyond the scope of this book and, in fact, can easily fill entire courses.)

Finally, as always before we do inference, we must imagine the underlying "true" model for these data. Here, that's pretty easy. We don't even need new

Greek letters because we're just dealing with means. We write the underlying model as

$$y_{ik} = \mu_k + \varepsilon_{ik}.$$

That is, we can think of each response value as being like the underlying mean of its group plus a unique error. Thinking about the ANOVA model gives us predicted values and residuals. Those are helpful when we check assumptions and conditions.

Back to Standard Deviations

We've been using the variances because they're easier to work with. But when it's time to think about the data, we'd really rather have a standard deviation because it's in the units of the response variable. The natural standard deviation to think about is the standard deviation of the residuals.

The variance of the residuals is staring us in the face. It's the MS_E. All we have to do to get the **residual standard deviation** is take the square root of MS_E:

$$s_p = \sqrt{MS_E} = \sqrt{\frac{\Sigma e^2}{(N - k)}}.$$

The p subscript is to remind us that this is a *pooled* standard deviation, combining residuals across all k groups. The denominator in the fraction shows that each of the k groups cost us a degree of freedom.

This standard deviation should "feel" right. That is, it should reflect the kind of variation you expect to find in any of the experimental groups. For the hand-washing data, $s_p = \sqrt{1410.14} = 37.6$ bacteria colonies. Looking back at the boxplots of the groups, we see that 37.6 seems to be a reasonable compromise standard deviation for all four groups.

Assumptions and Conditions

When we checked assumptions and conditions for regression we had to take care to perform our checks in order. Here we have a similar concern. For regression we found that displays of the residuals were often a good way to check the corresponding conditions. That's true for ANOVA as well.

Plot the Data . . .

Just as you would never find a linear regression without looking at the scatterplot of y vs. x, you should never embark on an ANOVA without first examining side-by-side boxplots of the data comparing the responses for all of the groups. You already know what to look for—we talked about that back in Chapter 5. Check for outliers within any of the groups and correct them if there are errors in the data. Get an idea of whether the groups have similar spreads (as we'll need) and whether the centers seem to be alike (as the null hypothesis claims) or different. If the individual boxplots are all skewed in the same direction, you should consider re-expressing the response variable to make them more symmetric. Doing so is likely to make the analysis more powerful and more correct.

Don't ever carry out an Analysis of Variance without looking at the side-by-side boxplots first. The chance of missing an important pattern or violation is just too great.

Independence Assumptions

The groups must be independent of each other. No test can verify this assumption. You have to think about how the data were collected. The assumption would be violated, for example, if we measured subjects' performance before some treatment, again in the middle of the treatment period, and then again at the end.[5]

The data *within* each treatment group must be independent as well. The data must be drawn independently and at random from a homogeneous population, or generated by a randomized comparative experiment.

We check the **Randomization Condition:** Were the data collected with suitable randomization? For surveys, are the data drawn from each group a representative random sample of that group? For experiments, were the treatments assigned to the experimental units at random?

We were told that the hand-washing experiment was randomized.

Equal Variance Assumption

The ANOVA requires that the variances of the treatment groups be equal. After all, we need to find a pooled variance for the MS_E. To check this assumption, we can check that the groups have similar variances:

Similar Variance Condition: There are some ways to see whether the variation in the treatment groups seems roughly equal:

- Look at side-by-side boxplots of the groups to see whether they have roughly the same spread. It can be easier to compare spreads across groups when they have the same center, so consider making side-by-side boxplots of the residuals. If the groups have differing spreads, it can make the pooled variance—the MS_E—larger, reducing the *F*-statistic value and making it less likely that we can reject the null hypothesis. So the ANOVA will usually fail on the "safe side," rejecting H_0 less often than it should. Because of this, we usually require the spreads to be quite different from each other before we become concerned about the condition failing. If you've rejected the null hypothesis, this is especially true.
- Look at the original boxplots of the response values again. In general, do the spreads seem to change *systematically* with the centers? One common pattern is for the boxes with bigger centers to have bigger spreads. This kind of systematic trend in the variances is more of a problem than random differences in spread among the groups and should not be ignored. Fortunately, such systematic violations are often helped by re-expressing the data. (If, in addition to spreads that grow with the centers, the boxplots are skewed with the longer tail

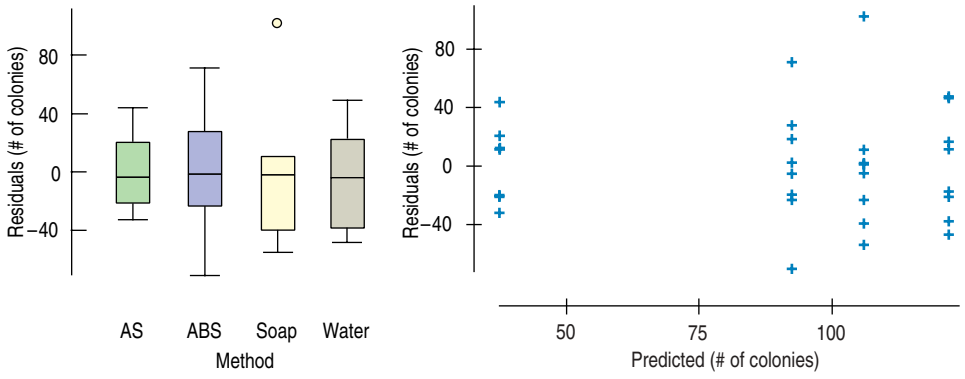
[5] There is a modification of ANOVA, called *repeated measures* ANOVA, that deals with such data. (If the design reminds you of a paired-*t* situation, you're on the right track, and the lack of independence is the same kind of issue we discussed in Chapter 25.)

stretching off to the high end, then the data are pleading for a re-expression. Try taking logs of the dependent variable for a start. You'll likely end up with a much cleaner analysis.)

- Look at the residuals plotted against the predicted values. Often, larger predicted values lead to larger magnitude residuals. This is another sign that the condition is violated. (This may remind you of the **Does the Plot Thicken? Condition** of regression. And it should.) When the plot thickens (to one side or the other), it's usually a good idea to consider re-expressing the response variable. Such a systematic change in the spread is a more serious violation of the equal variance assumption than slight variations of the spreads across groups.

Let's check the conditions for the hand-washing data. Here's a boxplot of residuals by group and residuals by predicted value:

Neither plot shows a violation of the condition. The IQRs (the box heights) are quite similar and the plot of residuals vs. predicted values does not show a pronounced widening to one end. The pooled estimate of 37.6 colonies for the error standard deviation seems reasonable for all four groups.



Boxplots of residuals for the four washing methods and a plot of residuals vs. predicted values. There's no evidence of a systematic change in variance from one group to the other or by predicted value.

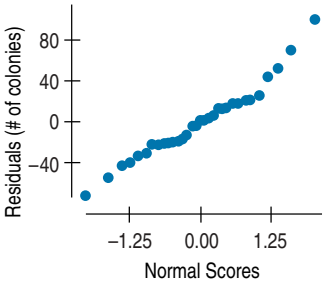
Figure 28.5

Normal Population Assumption

Like Student's *t*-tests, the *F*-test requires the underlying errors to follow a Normal model. As before when we've faced this assumption, we'll check a corresponding **Nearly Normal Condition.**

Technically, we need to assume that the Normal model is reasonable for the populations underlying *each* treatment group. We can (and should) look at the side-by-side boxplots for indications of skewness. Certainly, if they are all (or mostly) skewed in the same direction, the Nearly Normal Condition fails (and re-expression is likely to help).

In experiments, we often work with fairly small groups for each treatment, and it's nearly impossible to assess whether the distribution of only six or eight numbers is Normal (though sometimes it's so skewed or has such an extreme outlier that we can see that it's not). Here we are saved by the Equal Variance Assumption (which we've already checked). The residuals have their group means subtracted, so the mean residual for each group is 0. If their variances are equal, we can group all the residuals together for the purpose of checking the Nearly Normal Condition.



The hand-washing residuals look nearly Normal in this Normal probability plot. **Figure 28.6**

Check Normality with a histogram or a Normal probability plot of all the residuals together. The hand-washing residuals look nearly Normal in the Normal probability plot, although, as the boxplots showed, there's a possible outlier in the Soap group.

Because we really care about the Normal model *within each group*, the Normal population assumption is violated if there are outliers in any of the groups. Check for outliers in the boxplots of the values for each treatment group. The Soap group of the hand-washing data shows an outlier, so we might want to compute the analysis again without that observation. (For these data, it turns out to make little difference.)

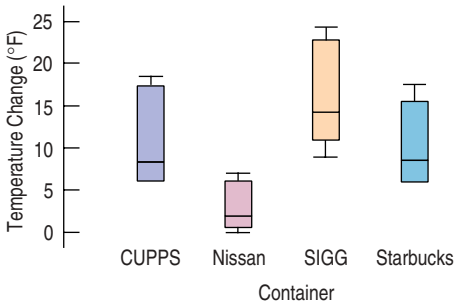
● **One-Way ANOVA *F*-test** We test the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ against the alternative that the group means are not all equal. We test the hypothesis with the *F*-statistic, $F = \frac{MS_T}{MS_E}$, where MS_T is the treatment mean square found from the variance of the means of the treatment groups and MS_E is the error mean square, found by pooling the variances within each of the treatment groups. If the *F*-statistic is large enough, we reject the null hypothesis. ●

Analysis of Variance Step-By-Step

In Chapter 5 we looked at side-by-side boxplots of four different containers for holding hot beverages. The experimenter wanted to know which type of container would keep his hot beverages hot longest. To test it, he heated water to a temperature of 180°F, placed it in the container, and then measured the temperature of the water again 30 minutes later. He randomized the order of the trials and tested each container 8 times. His response variable was the difference in temperature (in °F) between the initial water temperature and the temperature after 30 minutes. Let's test whether these containers really perform differently.

Think

Plot Plot the side-by-side boxplots of the data.



Hypotheses State what you want to know and the null hypothesis we wish to test. For ANOVA, the null hypothesis is that all the treatment groups have the same mean. The alternative is that at least one mean is different.

I want to know whether there is any difference among the four containers in their ability to maintain the temperature of a hot liquid for 30 minutes. Writing μ_k for the mean temperature difference for container *k*, then my null hypothesis is that these means are all the same:

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

Model Check the appropriate conditions.

The alternative is that the group means are not all equal.

✔ ***Randomization Condition:*** The experimenter performed the trials in random order.

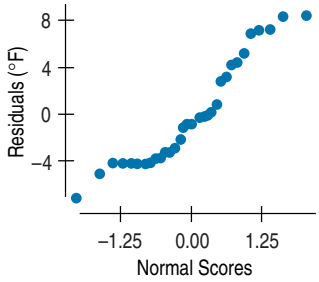
✔ ***Similar Variance Condition:*** The Nissan mug variation seems to be a bit smaller than the others. (I could also look later at the plot of residuals vs. predicted values to see if the plot thickens.)

Fit the ANOVA model.

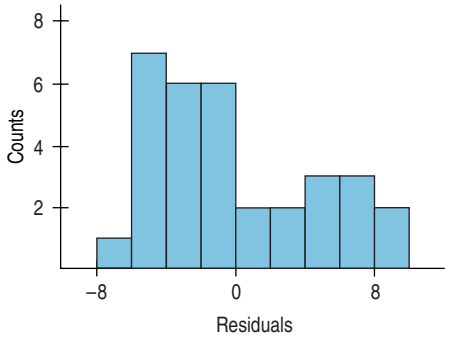
Analysis of Variance

Source	DF	Sum of Squares	Mean	F-ratio	P-value
Container	3	714.1875	238.063	10.713	<0.0001
Error	28	622.1875	22.221		
Total	31	1336.3750			

✔ ***Nearly Normal Condition:*** The Normal probability plot is not very straight, but there are no outliers. The histogram confirms that the data are skewed to the right, not symmetric.



The histogram shows that the distribution of the residuals is skewed to the right:



Looking at the table of means and SDs on the next page, I can see that the standard deviations grow along with the means. Possibly a re-expression of the data would improve matters.

Under these circumstances, I cautiously find the P-value for the *F*-statistic from the *F*-model with 3 and 28 degrees of freedom.

Show

Mechanics Show the table of means.

From the ANOVA table, the error mean square, MS_E, is 22.22, which means that the standard deviation of all the errors is estimated to be $\sqrt{22.22} = 4.71$ degrees.

This seems like a reasonable value for the error standard deviation in the four treatments (with the possible exception of the Nissan mug).

Level	*n*	Mean	Std Dev
CUPPS	8	10.1875	5.20259
Nissan	8	2.7500	2.50713
SIGG	8	16.0625	5.90059
Starbucks	8	10.2500	4.55129

The ratio of the mean squares gives an *F*-ratio of 10.7134 with a *P*-value <0.0001.

Tell

Conclusion Interpret the *F*-test.

An *F*-ratio this large would be very unlikely if the containers all had the same mean temperature difference.

State your conclusion in the proper context.

Conclusions: Even though some of the conditions are mildly violated, I am fairly confident that the means are not all equal. (I would have been more worried about the changing variance if I had failed to reject H_0.) More specific conclusions might require a re-expression of the data.

The Balancing Act

The two examples we've looked at so far share a special feature. Each treatment group has the same number of experimental units. For the hand-washing experiment, each washing method was tested 8 times. For the cups, there were also 8 trials for each cup. This feature (the equal numbers of cases in each group, not the number 8) is called **balance,** and experiments that have equal numbers of experimental units in each treatment are said to be balanced or to have balanced designs.

As usual, we give nice-sounding names like "regular," "simple," and "balanced" to things we like or hope to see. Balanced designs are a bit easier to analyze because the calculations are simpler. Usually, we try for balanced designs. But in the real world we often encounter unbalanced data. Participants drop out or become unsuitable, plants die, or maybe we just can't find enough experimental units to fit a particular criterion.

Everything we've done so far works just fine for unbalanced designs except that the calculations get a bit more complicated. Where once we could write *n* for

the number of experimental units in a treatment, now we have to write n_k and sum more carefully. Where once we could pool variances with a simple average, now we have to adjust for the different n's. Technology clears these hurdles easily, so you're safe thinking about the analysis in terms of the simpler balanced formulas and trusting that the technology will make the necessary adjustments.

Comparing Means

When we reject H_0, it's natural to ask which means are different. No one would be happy with an experiment to test 10 cancer treatments that concluded only with "We can reject H_0—the treatments are different!" We'd like to know more, but the F-statistic doesn't offer that information.

What can we do? If we can't reject the null, we've got to stop. There's no point in further testing. If we've rejected the simple null hypothesis, however, we *can* do more. In particular, we can test whether any pairs or combinations of group means differ. For example, we might want to compare treatments against a control or a placebo, or against the current standard treatment.

A S Boxplots and ANOVA. How should we understand how the group means differ? Make a picture. This activity discusses and illustrates what to do.

In the hand-washing experiment, we could consider plain water to be a control. Nobody would be impressed with (or want to pay for) a soap that did no better than water alone. A test of whether the antibacterial soap (for example) was different from plain water would be a simple test of the difference between two group means. To be able to perform an ANOVA, we first check the **Similar Variance Condition.** If things look OK we assume that the variances are equal. If the variances *are* equal then a pooled t-test is appropriate. Even better (this is the special part), we already have a pooled estimate of the standard deviation based on *all* of the tested washing methods. That's s_p, which, for the hand-washing experiment, was equal to 37.55 bacteria colonies.

The null hypothesis is that there is no difference between water and the antibacterial soap. As we did in Chapter 24, we'll write that as a hypothesis about the difference in the means:

$$H_0: \mu_W - \mu_{ABS} = 0. \text{ The alternative is}$$
$$H_A: \mu_W - \mu_{ABS} \neq 0.$$

The natural test statistic is $\bar{y}_W - \bar{y}_{ABS}$, and the (pooled) standard error is $SE(\mu_W - \mu_{ABS}) = s_p\sqrt{\frac{1}{n_W} + \frac{1}{n_{ABS}}}$.

Level	n	Mean	Std Dev
Alcohol spray	8	37.5	26.56
Antibacterial soap	8	92.5	41.96
Soap	8	106.0	46.96
Water	8	117.0	31.13

The difference in the observed means is $117.0 - 92.5 = 24.5$ colonies. The standard error comes out to 18.775. The t-statistic, then, is $t = \frac{24.5}{18.75} = 1.31$. To find the P-value we consult the Student's t-distribution on $N - k = 32 - 4 = 28$ degrees of freedom. The P-value is about 0.1—not small enough to impress us. So we can't discern a significant difference between washing with the antibacterial soap and just using water.

Our t-test asks about a simple difference. We could also ask a more complicated question about groups of differences. Does the average of the two soaps differ

from the average of three sprays, for example? Complex combinations like these are called *contrasts*. Finding the standard errors for contrasts is straightforward, but beyond the scope of this book. We'll restrict our attention to the common question of comparing pairs of treatments after H_0 has been rejected.

*Bonferroni Multiple Comparisons

Our hand-washing experimenter *was* pretty sure that alcohol would kill the germs even before she started the experiment. But alcohol dries the skin and leaves an unpleasant smell. She was hoping that one of the antibacterial soaps would work as well as alcohol so she could use that instead. That means she really wanted to compare *each* of the other treatments against the alcohol spray. We know how to compare two of the means with a *t*-test. But now we want to do several tests, and each test poses the risk of a Type I error. As we do more and more tests, the risk that we might make a Type I error grows bigger than the α level of each individual test. With each additional test, the risk of making an error grows. If we do enough tests, we're almost sure to reject one of the null hypotheses by mistake—and we'll never know which one.

There is a defense against this problem. In fact, there are several defenses. As a class, they are called **methods for multiple comparisons.** All multiple comparisons methods require that we first be able to reject the overall null hypothesis with the ANOVA's *F*-test. Once we've rejected the overall null, then we can think about comparing several—or even all—pairs of group means.

Let's look again at our test of the water treatment against the antibacterial soap treatment. This time we'll look at a confidence interval instead of the pooled *t*-test. We did a test at significance level $\alpha = 0.05$. The corresponding confidence level is $1 - \alpha = 95\%$. For *any* pair of means, a confidence interval for their difference is $(\bar{y}_1 - \bar{y}_2) \pm ME$, where the margin of error is

$$ME = t^* \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

As we did in the previous section, we get s_p as the pooled standard deviation found from *all* the groups in our analysis. We find the critical value t^* from the Student's *t*-model corresponding to the specified confidence level found with $N - k$ degrees of freedom, and the n_k's are the number of experimental units in each of the treatments.

To reject the null hypothesis that the two group means are equal, the difference between them must be larger than the ME. That way 0 won't be in the confidence interval for the difference. When we use it in this way, we call the margin of error the **least significant difference** (**LSD** for short). If two group means differ by more than this amount, then they are significantly different at level α *for each individual test*.

For our hand-washing experiment, each group has $n = 8$, and $s_p = 37.55$ colonies. From technology or Table T, we can find that t^* with 28 df (for a 95% confidence interval) is 2.048. So

$$\text{LSD} = 2.048 \times 37.55 \times \sqrt{\frac{1}{8} + \frac{1}{8}} = 38.45 \text{ colonies},$$

and we could use this margin of error to make a 95% confidence interval for any difference between group means. Any two washing methods whose means differ by more than 38.45 colonies could be said to differ at $\alpha = 0.05$ by this method.

Of course, we're still just examining individual pairs. If we want to examine *many* pairs simultaneously, there are several methods that adjust the critical t^*-value so that the resulting confidence intervals provide appropriate tests for all the pairs. And, in spite of making *many* such intervals, the overall Type I error rate stays at (or below) α.

Carlo Bonferroni (1892–1960) was a mathematician who taught in Florence. He wrote two papers in 1935 and 1936 setting forth the mathematics behind the method that bears his name.

One such method is called the **Bonferroni method.** This method adjusts the LSD to allow for making many comparisons. The result is a wider margin of error called the **minimum significant difference,** or **MSD.** The MSD is found by replacing t^* with a slightly larger number. That makes the confidence intervals wider for each contrast and the corresponding Type I error rates lower for *each* test. And it keeps the *overall* Type I error rate at or below α.

The Bonferroni method distributes the error rate equally among the confidence intervals. It divides the error rate among J confidence intervals, finding each interval at confidence level $1 - \frac{\alpha}{J}$ instead of the original $1 - \alpha$. To signal this adjustment, we label the critical value t^{**} rather than t^*. For example, to make the three confidence intervals comparing the alcohol spray with the other three washing methods, and preserve our overall α risk at 5%, we'd construct each with a confidence level of

$$1 - \frac{0.05}{3} = 1 - 0.01667 = 0.98333$$

The only problem with this is that t-tables don't have a column for 98.33% confidence (or, correspondingly, for $\alpha = 0.01667$). Fortunately, technology has no such constraints.[6] For the hand-washing data, if we want to examine the three confidence intervals comparing each of the other methods with the alcohol spray, the t^{**}-value (on 28 degrees of freedom) turns out to be 2.546. That's somewhat larger than the individual t^*-value of 2.048 that we would have used for a single confidence interval. And the corresponding ME is 47.80 colonies (rather than 38.45 for a single comparison). The larger critical value along with correspondingly wider intervals is the price we pay for making multiple comparisons.

Many statistics packages assume that you'd like to compare all pairs of means. Some will display the result of these comparisons in a table like this:

Level	*n*	Mean	Groups
Alcohol spray	8	37.5	A
Antibacterial soap	8	92.5	B
Soap	8	106.0	B
Water	8	117.0	B

This table shows that the alcohol spray is in a class by itself and that the other three hand-washing methods are indistinguishable from one another.

[6] The electronic *t*-tables provided on the CD-ROM in *ActivStats* let you add new columns to the *t*-table at any alpha level, so you can do the Bonferroni calculation easily.

ANOVA on Observational Data

So far we've applied ANOVA only to data from designed experiments. That's natural for several reasons. The primary one is that, as we saw in Chapter 13, randomized comparative experiments are specifically designed to *compare* the results for different treatments. The overall null hypothesis, and the subsequent tests on pairs of treatments in ANOVA, address such comparisons directly. In addition, as we discussed earlier, the Equal Variance Assumption (which we need for all of the ANOVA analyses) is often plausible in a randomized experiment because the treatment groups start out with sample variances that all estimate the same underlying variance of the collection of experimental units.

Sometimes, though, we just can't perform an experiment. When ANOVA is used to test equality of group means from observational data, there's no *a priori* reason to think the group variances might be equal at all. Even if the null hypothesis of equal means were true, the groups might easily have different variances. But if the side-by-side boxplots of responses for each group show roughly equal spreads and symmetric, outlier-free distributions, you can use ANOVA on observational data.

Observational data tend to be messier than experimental data. They are much more likely to be unbalanced. If you aren't assigning subjects to treatment groups, it's harder to guarantee the same number of subjects in each group. And because you are not controlling conditions as you would in an experiment, things tend to be, well, less controlled. The only way we know to avoid the effects of possible lurking variables is with control and randomized assignment to treatment groups, and for observational data, we have neither.

ANOVA is often applied to observational data when an experiment would be impossible or unethical. (We can't randomly break some subjects' legs, but we *can* compare pain perception among those with broken legs, those with sprained ankles, and those with stubbed toes by collecting data on subjects who have already suffered those injuries.) In such data, subjects are already in groups, but not by random assignment.

Be careful; if you have not assigned subjects to treatments randomly, you can't draw *causal* conclusions even when the F-test is significant. You have no way to control for lurking variables or confounding, so you can't be sure whether any differences you see among groups are due to the grouping variable or to some other unobserved variable that may be related to the grouping variable.

Because observational studies often are intended to estimate parameters, there is a temptation to use pooled confidence intervals for the group means for this purpose. Although these confidence intervals are statistically correct, be sure to think carefully about the population that the inference is about. The relatively few subjects that you happen to have in a group may not be a simple random sample of any interesting population, so their "true" mean may have only limited meaning.

One More Example Step-By-Step

Here's an example that exhibits many of the features we've been discussing. It gives a fair idea of the kinds of challenges often raised by real data.

A study at a liberal arts college attempted to find out who watches more TV at college: men or women? Varsity athletes or non-athletes? Student researchers asked 200 randomly selected students questions about their backgrounds and about their television-viewing habits. The researchers

found that men watch, on average, about 2.5 hours per week more TV than women, and that varsity athletes watch about 3.5 hours per week more than those who are not varsity athletes. But is this the whole story? To investigate further, they divided the students into four groups: male athletes (MA), male non-athletes (MNA), female athletes (FA), and female non-athletes (FNA). Let's do the ANOVA step-by-step.

Think

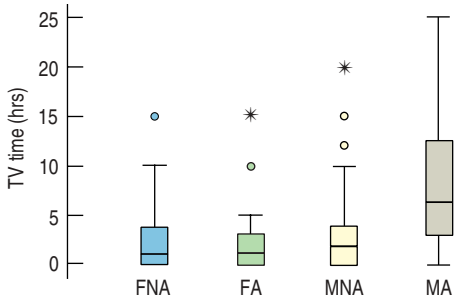
Plan Name the variables, report the W's, and specify the questions of interest.

I have the number of hours spent watching TV in a week for 200 randomly selected students. I know their sex and whether they are varsity athletes or not. I wonder whether TV watching differs according to sex and athletic status.

Make a picture. Always start an ANOVA with side-by-side boxplots of the responses in each of the groups. Always.

These data offer a good example why.

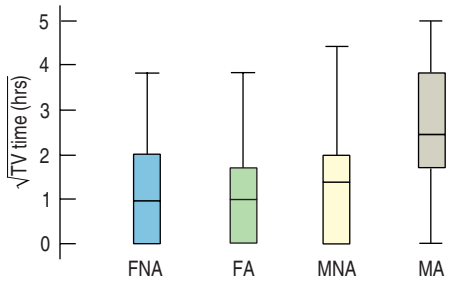
Here are the side-by-side boxplots of the data:



This plot suggests problems with the data. Each box shows a distribution skewed to the high end, and outliers pepper the display, including some extreme outliers. The box with the highest center (MA) also has the largest spread. These data just don't pass the first screening for suitability. This sort of pattern calls for a re-expression.

The responses are counts—numbers of TV hours. You may recall from Chapter 10 that a good re-expression to try first for counts is the square root.

Here are the boxplots for the square root of TV hours.



The spreads in the four groups are now more similar and the individual distributions more symmetric. And now there are no outliers.

Model and Mechanics Check the appropriate conditions.

✔ **Randomization Condition:** The data come from a random sample of students.

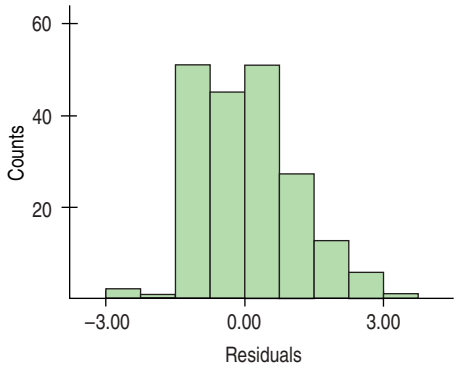
✔ **Similar Variance Condition:** The boxplots show similar spreads. (I could also check the residuals later.)

Show Fit The ANOVA model.

The ANOVA table looks like this:

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Group	3	47.24733	15.7491	12.8111	<0.0001
Error	193	237.26114	1.2293		
Total	196	284.50847			

✔ **Nearly Normal Condition:** A histogram of the residuals looks reasonably Normal:



Interestingly, the few cases that seem to stick out on the low end are male athletes who watched no TV, making them different from all the other male athletes.

Under these conditions, it's appropriate to use Analysis of Variance.

Tell **Conclusion** Interpret the results in the proper context.

The F-statistic is large and the corresponding P-value small. I am confident that the TV-watching behavior is not the same among these groups.

*So Do Male Athletes Watch More TV?

Here's a Bonferroni comparison of all pairs of groups:

In case you were wondering . . . The standard errors are different because this isn't a balanced design. Differing numbers of experimental units in the groups generate differing standard errors.

	Difference	Std. Err.	P-Value
FA–FNA	0.049	0.270	0.9999
MNA–FNA	0.205	0.182	0.8383
MNA–FA	0.156	0.268	0.9929
MA–FNA	1.497	0.250	<0.0001
MA–FA	1.449	0.318	<0.0001
MA–MNA	1.292	0.248	<0.0001

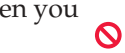
Three of the differences are very significant. It seems that among women there's little difference in TV watching between varsity athletes and others. Among men, though, the corresponding difference is large. And among varsity athletes, men watch significantly more TV than women.

But wait. How far can we extend the inference that male athletes watch more TV than other groups? The data came from a random sample of students made during the week of March 21. If the students carried out the survey correctly using a simple random sample, we should be able to make the inference that the generalization is true for the entire student body during that week.

Is it true for other colleges? Is it true throughout the year? The students conducting the survey followed up the survey by collecting anecdotal information about TV watching of male athletes. It turned out that during the week of the survey, the NCAA men's basketball tournament was televised. This could explain the increase in TV watching for the male athletes. It could be that the increase extends to other students at other times, but we don't know that. Always be cautious in drawing conclusions too broadly. Don't generalize from one population to another.

What Can Go Wrong?

- ***Watch out for outliers.*** One outlier in a group can change both the mean and the spread of that group. It will also inflate the error mean square, which can influence the *F*-test. The good news is that ANOVA fails on the safe side by losing power when there are outliers. That is, you are less likely to reject the overall null hypothesis if you have (and leave) outliers in your data. But they are not likely to cause you to make a Type I error.
- ***Watch out for changing variances.*** The conclusions of the ANOVA depend crucially on the assumptions of independence and constant variance, and (somewhat less seriously as n increases) on Normality. If the conditions on the residuals are violated, it may be necessary to re-express the response variable to approximate these conditions more closely. ANOVA benefits so greatly from a judiciously chosen re-expression that the choice of a re-expression might be considered a standard part of the analysis.
- ***Be wary of drawing conclusions about causality from observational studies.*** ANOVA is often applied to data from randomized experiments for which causal conclusions are appropriate. If the data are not from a designed experiment, however, the Analysis of Variance provides no more evidence for causality than any other method we have studied. Don't get into the habit of assuming that ANOVA results have causal interpretations.
- ***Be wary of generalizing*** to situations other than the one at hand. Think hard about how the data were generated to understand the breadth of conclusions you are entitled to draw.
- ***Watch for multiple comparisons.*** When rejecting the null hypothesis, you can conclude that the means are not *all* equal. But you can't start comparing every pair of treatments in your study with a *t*-test. You'll run the risk of inflating your Type I error rate. Use a multiple comparisons method when you want to test many pairs.



CONNECTIONS

We first learned about side-by-side boxplots in Chapter 5. There we made general statements about the shape, center, and spread of each group. When we compared groups, we asked whether their centers looked different compared with how spread out the distributions were. Now we've made that kind of thinking precise. We've added confidence intervals for the difference and tests of whether the means are the same.

We pooled data to find a standard deviation when we tested the hypothesis of equal proportions. For that test, the assumption of equal variances was a consequence of the null hypothesis that the proportions were equal, so it didn't require an extra assumption. Means don't have a linkage with their corresponding variances, so to use pooled methods we must make the additional assumption of equal variances. But in a randomized experiment, that's a plausible assumption.

Chapter 13 offered a variety of designs for randomized comparative experiments. Each of those designs can be analyzed with a variant or extension of the ANOVA methods discussed in this chapter. Entire books and courses deal with these extensions, but all follow the same fundamental ideas presented here.

ANOVA is closely related to the regression analyses we saw in Chapter 27. (In fact, most statistics packages offer an ANOVA table as part of their regression output.) The assumptions are similar—and for good reason. The analyses are, in fact, related at a deep conceptual (and computational) level, but those details are beyond the scope of this book.

The pooled two-sample *t*-test for means is a special case of the ANOVA *F*-test. If you perform an ANOVA comparing only two groups, you'll find that the P-value of the *F*-statistic is exactly the same as the P-value of the corresponding pooled *t*-statistic. That's because in this special case the *F*-statistic is just the square of the *t*-statistic. The *F*-test is more general. It can test the hypothesis that several group means are equal.

What have we learned?

We've learned how to compare the means of more than two independent groups based on samples drawn from those groups.

We've learned that using a *t*-test to test the equality of each pair of groups means would lead to a higher Type I error rate than we want. We can correct for that by lowering the alpha level of each test by using a Bonferroni correction.

We've learned that we can also test the hypothesis that all the means are equal using the Analysis of Variance (ANOVA). The main idea is to compare overall differences *between* the treatment means with the variation *within* each group. If the null hypothesis of equal means is true, the ratio we form will follow an *F*-distribution. But if that ratio is large compared with the *F*-distribution, that provides evidence against the null hypothesis.

We've learned that ANOVA has assumptions and conditions that need to be checked before we infer anything from the *F*-ratio:

- The data values must be independent (think about how they were collected).
- The spread within each group must be equal (check the side-by-side boxplots and a plot of residuals against predicted values).
- The errors must be Normal (check a histogram or Normal probability plot of the residuals).

We've learned that when these conditions are satisfied, we can be confident in our inference from the *F*-test. A large *F*-ratio gives a small P-value and provides evidence *against* the null hypothesis of equal means. If we do reject the null hypothesis, then we need to use multiple comparisons methods to determine which means are different.

TERMS

Error mean square (MS_E)
The error mean square (MS_E) is the estimate of the error variance obtained by *pooling* the variances of each treatment group. The square root of the MS_E is the estimate of the error standard deviation, s_p.

Treatment mean square (MS_T)
The treatment mean square (MS_T) is the estimate of the error variance under the assumption that the treatment means are all equal. If the (Null) Assumption is not true, the MS_T will be larger than the error variance.

***F*-distribution**
The *F*-distribution is the sampling distribution of the *F*-statistic when the null hypothesis that the treatment means are equal is true. It has two degrees of freedom, one for the numerator, $(k - 1)$, and one for the denominator, $N - k$, where N is the total number of observations and k is the number of groups.

***F*-statistic**
The *F*-statistic is the ratio MS_T/MS_E. When the *F*-statistic is sufficiently large, we reject the null hypothesis that the group means are equal.

***F*-test**
The *F*-test tests the null hypothesis that all the group means are equal against the one-sided alternative that they are not all equal. We reject the hypothesis of equal means if the *F*-statistic exceeds the critical value from the *F*-distribution corresponding to the specified significance level and degrees of freedom.

ANOVA
An analysis method for testing equality of means across treatment groups.

ANOVA table
The ANOVA table is convenient for showing the degrees of freedom, the treatment mean square, the error mean square, their ratio, the *F*-statistic, and its P-value. There are usually other quantities of lesser interest included as well.

ANOVA model
The model for a one-way (one response, one factor) ANOVA is

$$y_{ik} = \mu_k + \varepsilon_{ik}.$$

Estimating with $y_{ik} = \bar{y}_k + e_{ik}$ gives predicted values $\hat{y}_{ik} = \bar{y}_k$ and residuals $e_{ik} = y_{ik} - \bar{y}_k$.

Assumptions for ANOVA (and conditions to check)
- Independence Assumption. (Think about the design of the experiment or, if an observational study, how the data were collected.)
- Equal Variance Assumption. (Similar Variance Condition. Look at side-by-side boxplots to check for similar spreads, or look at residuals vs. predicted to see if the plot thickens.)
- Normal Population Assumption. (Nearly Normal Condition. Check a histogram or Normal probability plot of the residuals.)

Residual standard deviation
The residual standard deviation,

$$s_p = \sqrt{MS_E} = \sqrt{\frac{\sum e^2}{N - k}},$$

gives an idea of the underlying variability of the response values.

Balance
An experiment's design is balanced if each treatment level has the same number of experimental units. Balanced designs make calculations simpler and are generally more powerful.

Multiple comparisons If we reject the null hypothesis of equal means, we often then want to investigate further and compare pairs of treatment group means to see if they differ. If we want to test several such pairs, we must adjust for performing several tests to keep the overall risk of a Type I error from growing too large. Such adjustments are called methods for multiple comparisons.

Least significant difference (LSD) The standard margin of error in the confidence interval for the difference of two means is called the least significant difference. It has the correct Type I error rate for a single test, but not when performing more than one comparison.

Bonferroni method One of many methods for adjusting the length of the ME when testing the differences between several group means.

Minimum significant difference (MSD) The Bonferroni method's ME for the confidence interval for the difference of two group means is called the minimum significant difference. This can be used to test differences of several pairs of group means. If their difference exceeds the MSD, they are different at the overall α rate.

SKILLS *When you complete this lesson you should:*

- Recognize situations for which ANOVA is the appropriate analysis.
- Know how to examine your data for violations of conditions that would make ANOVA unwise or invalid.
- Recognize when a further analysis of differences between group means would be appropriate.

- Be able to perform an ANOVA using a statistics package or calculator for one response variable and one factor with any number of levels.
- Be able to perform several subsequent tests using a multiple comparisons procedure.

- Be able to explain the contents of an ANOVA table, in particular the role of the MS_T, MS_E, and the standard deviation of the error, s_p.
- Be able to interpret a test of the null hypothesis that the true means of several independent groups are equal. (Your interpretation should include a defense of your Assumption of Equal Variances.)
- Be able to interpret the results of tests that use multiple comparison methods.

ANOVA on the Computer

Most analyses of variance are found with computers. And all statistics packages present the results in an ANOVA table much like the one we discussed. Technology also makes it easy to examine the side-by-side boxplots and check the residuals for violations of the assumptions and conditions.

Statistics packages offer different choices among possible multiple comparisons methods (although Bonferroni is quite common). This is a specialized area. Get advice or read further if you need to choose a multiple comparisons method.

As we saw in Chapter 24, there are two ways to organize data recorded for several groups. We can put all the response values in a single variable and use a second, "factor," variable to hold the group identities. This is sometimes called *stacked format*. The alternative is to place the data for each group in its own column or variable. Then the variable identities become the group identifiers.

Most statistics packages expect the data to be in stacked format because this form also works for more complicated experimental designs. Some packages can work with either form, and some use one form for some things and the other for others. (Be careful, for example, when you make side-by-side boxplots; be sure to give the appropriate version of the command to correspond to the structure of your data.)

Most packages offer to save residuals and predicted values and make them available for further tests of conditions. In some packages you may have to request them specifically.

DATA DESK

- Select the response variable as Y and the factor variable as X.
- From the **Calc** menu, choose **ANOVA.**
- Data Desk displays the ANOVA table.
- Select plots of residuals from the ANOVA table's HyperView menu.

Comments

Data Desk expects data in "stacked" format. You can change the ANOVA by dragging the icon of another variable over either the Y or X variable name in the table and dropping it there. The analysis will recompute automatically.

EXCEL

- From the tools menu, select **Data Analysis.**
- Select **Anova Single Factor** from the list of analysis tools.
- Click the **OK** button.
- Enter the data range in the box provided.
- Check the **Labels in First Row** box, if applicable.
- Enter an alpha level for the F-test in the box provided.
- Click the **OK** button.

Comments

The data range should include two or more columns of data to compare. Unlike all other statistics packages, Excel expects each column of the data to represent a different level of the factor. However, it offers no way to label these levels. The columns need not have the same number of data values, but the selected cells must make up a rectangle large enough to hold the column with the most data values.

JMP

- From the **Analyze** menu select **Fit Y by X.**
- Select variables: a quantitative Y, Response variable, and a categorical X, Factor variable.
- JMP opens the **Oneway** window.
- Click on the red triangle beside the heading, select **Display Options,** and choose **Boxplots.**
- From the same menu choose the **Means/ANOVA.t-test** command.
- JMP opens the oneway ANOVA output.

Comments

JMP expects data in "stacked" format with one response and one factor variable.

MINITAB

- Choose **ANOVA** from the Stat menu.
- Choose **One-way...** from the **ANOVA** submenu.
- In the One-way Anova dialog, assign a quantitative Y variable to the Response box and assign a categorical X variable to the Factor box.
- Check the **Store Residuals** check box.
- Click the **Graphs** button.
- In the ANOVA-Graphs dialog, select **Standardized residuals,** and check **Normal plot of residuals** and **Residuals versus fits.**
- Click the **OK** button to return to the Regression dialog.
- Click the **OK** button to compute the regression.

Comments

If your data are in unstacked format, with separate columns for each treatment level, choose **One-way (unstacked)** from the **ANOVA** submenu.

SPSS

- Choose **Compare Means** from the Analyze menu.
- Choose **One-way ANOVA** from the **Compare Means** submenu.
- In the One-Way ANOVA dialog, select the Y-variable and move it to the dependent target. Then move the X-variable to the independent target.
- Click the **OK** button.

Comments

SPSS expects data in stacked format. The **Contrasts** and **Post Hoc** buttons offer ways to test contrasts and perform multiple comparisons. See your SPSS manual for details.

TI-89

Under **STAT Tests** choose **C:ANOVA**

- Specify the input method (Data or Stats) according to whether you have data entered as one list for each group or summary statistics for each group, and specify the number of groups. Press ÷.
- If Data, you will then be asked to supply the name of each list.
- If Stats, you will be asked for the stats for each group. Enter n, ì, and s for each group separated by commas and within curly braces ({and}).
- Press ÷ to perform the calculations.

Comments

In addition to the ANOVA table output, the calculator creates three new lists—the means for each group (in the order specified) and *individual* 95% confidence interval upper and lower bounds.

EXERCISES

1. Popcorn. A student runs an experiment to test four different popcorn brands, recording the number of kernels left unpopped. She pops measured batches of each brand 4 times, using the same popcorn popper and randomizing the order of the brands. After collecting her data and analyzing the results, she reports that the *F*-ratio is 13.56.

a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) Assuming that the conditions required for ANOVA are satisfied, what is the P-value? What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?

2. Skating. A figure skater tried various approaches to her Salchow jump in a designed experiment using 5 different places for her focus (arms, free leg, midsection, take-off leg, and free). She tried each jump 6 times in random order, using two of her skating partners to judge the jumps on a scale from 0 to 6. After collecting the data and analyzing the results, she reports that the *F*-ratio is 7.43.

a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) Assuming that the conditions are satisfied, what is the P-value? What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?

3. Gas mileage. A student runs an experiment to study the effect of three different mufflers on gas mileage. He devises a system so that his Jeep Wagoneer uses gasoline from a one-liter container. He tests each muffler 8 times, carefully recording the number of miles he can go in his Jeep Wagoneer on one liter of gas. After analyzing his data, he reports that the *F*-ratio is 2.35 with a P-value of 0.1199.

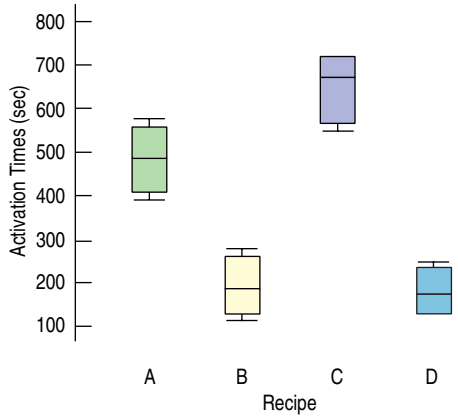
a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?
e) If your conclusion in part c is wrong, what type of error have you made?

4. Darts. A student interested in improving her dart-throwing technique designs an experiment to test 4 different stances to see whether they affect her accuracy. After warming up for several minutes, she randomizes the order of the 4 stances, throws a dart at a target using each stance and, measures the distance of the dart in centimeters from the center of the bull's-eye. She replicates this procedure 10 times. After analyzing the data she reports that the *F*-ratio is 1.41.

a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?
e) If your conclusion in part c is wrong, what type of error have you made?

T **5. Activating baking yeast.** To shorten the time it takes him to make his favorite pizza, a student designed an experiment to test the effect of sugar and milk on the activation times for baking yeast. Specifically, he tested four different recipes and measured how many seconds it took for the same amount of dough to rise to the top of a bowl. He randomized the order of the recipes and replicated each treatment 4 times.

Here are the boxplots of activation times from the four recipes:



The ANOVA table follows:

Analysis of Variance

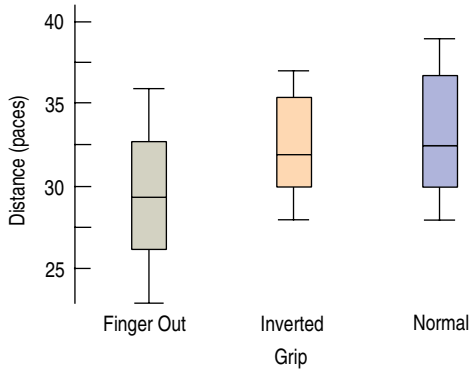
Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Recipe	3	638967.69	212989	44.7392	<0.0001
Error	12	57128.25	4761		
Total	15	696095.94			

a) State the hypotheses about the recipes (both numerically and in words).

b) Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion. Be sure to state it in terms of activation times and recipes.

c) Would it be appropriate to follow up this study with multiple comparisons to see which recipes differ in their mean activation times? Explain.

T **6. Frisbee throws.** A student performed an experiment with three different grips to see what effect it might have on the distance of a backhanded Frisbee throw. She tried it with her normal grip, with one finger out, and with the Frisbee inverted. She measured in paces how far her throw went. The boxplots and the ANOVA table for the three grips are shown below:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Grip	2	58.58333	29.2917	2.0453	0.1543
Error	21	300.75000	14.3214		
Total	23	359.33333			

a) State the hypotheses about the grips.

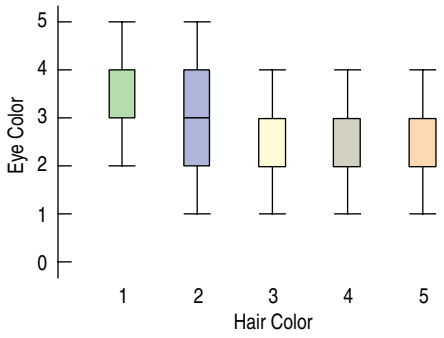
b) Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion. Be sure to state it in terms of Frisbee grips and distance thrown.

c) Would it be appropriate to follow up this study with multiple comparisons to see which grips differ in their mean distance thrown? Explain.

7. Eye and hair color. In Chapter 5, Exercise 47, we saw a survey of 1021 school-age children conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey, the following codes were used:

Hair color	Eye color
1 = Blond	1 = Blue
2 = Brown	2 = Green
3 = Black	3 = Brown
4 = Red	4 = Grey
5 = Other	5 = Other

The students analyzing the data were asked to study the relationship between eye and hair color. They produced this plot:



and ran an Analysis of Variance with *Eye color* as the response and *Hair color* as the predictor. The ANOVA table they produced follows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Hair color	4	1.46946	0.367365	0.4024	0.8070
Error	1016	927.45317	0.912848		
Total	1020	928.92263			

What suggestions do you have for the Statistics students? What alternative analysis might you suggest?

8. Zip codes revisited. The intern from the marketing department at the Holes R Us online piercing salon has recently finished a study of the company's 500 customers. He wanted to know whether people's zip codes vary by the last product they bought. They have 16 different products, and the ANOVA table of zip code by product showed the following:

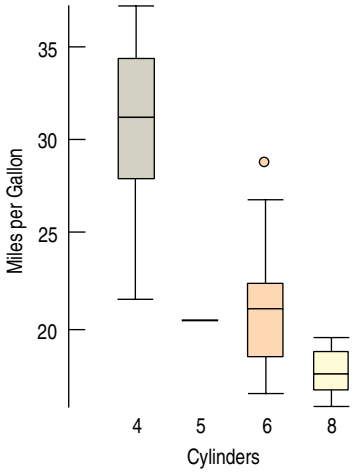
ANOVA table

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Product	15	3.836e10	2.55734e9	4.9422	<0.0001
Error	475	2.45787e11	517445573		
Total	490	2.84147e11			

(Nine customers were not included because of missing zip code or product information.)

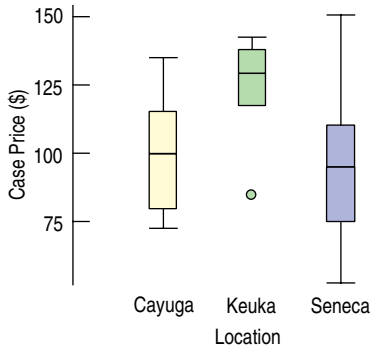
What criticisms of the analysis might you make? What alternative analysis might you suggest?

T **9. Fuel economy revisited.** In Chapter 5, we looked at what these boxplots told us about the relationship between the number of cylinders a car's engine has and the car's fuel economy.



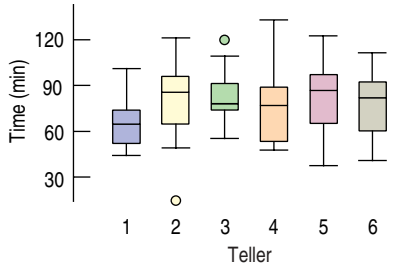
a) State the null and alternative hypotheses.
b) Do the conditions for an ANOVA seem to be met here? Why or why not?

T 10. Wines revisited. The boxplots we saw in Chapter 5, Exercise 23, display case prices (in dollars) of wines produced by wineries along three of the Finger Lakes.



a) What are the null and alternative hypotheses? Talk about prices and location, not symbols.
b) Do the conditions for an ANOVA seem to be met here? Why or why not?

T 11. Tellers. A bank is studying the time that it takes 6 of its tellers to serve an average customer. Customers line up in the queue and then go to the next available teller. Here is a boxplot of the last 200 customers and the times it took each teller:

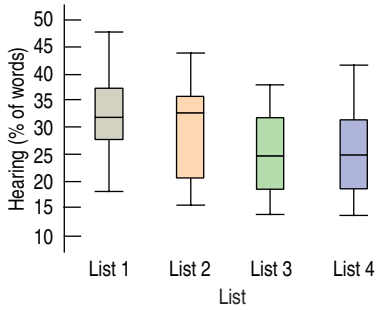


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Teller	5	3315.32	663.064	1.508	0.1914
Error	134	58919.1	439.695		
Total	139	62234.4			

a) What are the null and alternative hypotheses?
b) What do you conclude?
c) Would it be appropriate to run a multiple comparisons test (for example, a Bonferroni test) to see which tellers differ from each other? Explain.

T 12. Hearing. A researcher investigated four different word lists for use in hearing assessment (Loven, 1981). She wanted to know whether the lists were equally difficult to understand in the presence of a noisy background. To find out, she tested 24 subjects with normal hearing and measured the number of words perceived correctly in the presence of background noise. Here are the boxplots of the four lists:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
List	3	920.4583	306.819	4.9192	0.0033
Error	92	5738.1667	62.371		
Total	95	6658.6250			

a) What are the null and alternative hypotheses?
b) What do you conclude?
c) Would it be appropriate to run a multiple comparisons test (for example, a Bonferroni test) to see which lists differ from each other in terms of mean percent correct? Explain.

13. Yogurt. An experiment to determine the effect of several methods of preparing cultures 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, ultrafiltration, and reverse osmosis. A trained 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 on the following page.

An incomplete ANOVA Table for the Yogurt Data

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Treatment	17.300			
Residual	0.460			
Total	17.769			

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 0.000017. What does this say about the null hypothesis of equal means?
d) What assumptions have you made in order to answer part c?
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 error standard deviation in the judge's assessment?

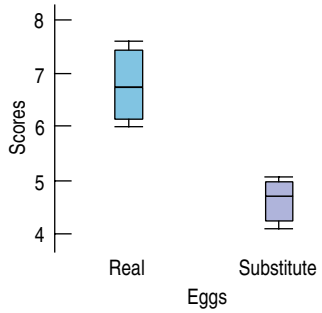
14. Smokestack scrubbers. Particulate matter is a serious form of air pollution often arising from industrial production. One way to reduce the pollution is to put a filter, or scrubber, at the end of the smokestack to trap the particulates. An experiment to determine which smokestack scrubber design is best was run by placing four scrubbers of different designs on an industrial stack in random order. Each scrubber was tested 5 times. For each run, the same material was produced, and the particulate emissions coming out of the scrubber were measured (in parts per billion). A partially complete Analysis of Variance table of the data is shown below.

An incomplete ANOVA Table for the Smokestack Data

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Treatment	81.2			
Residual	30.8			
Total	112.0			

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 0.00000949. What does this say about the null hypothesis of equal means?
d) What assumptions have you made in order to answer part c?
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 error standard deviation in particulate emissions?

T **15. Eggs.** A student wants to investigate the effects of real vs. substitute eggs on his favorite brownie recipe. He enlists the help of 10 friends and asks them to rank each of 8 batches on a scale from 1 to 10. Four of the batches were made with real eggs, four with substitute eggs. The judges tasted the brownies in random order. Here is a boxplot of the data:

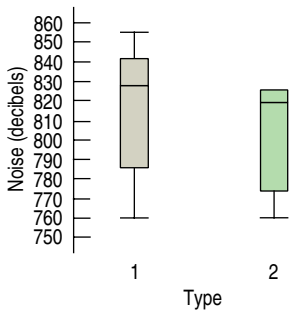


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Eggs	1	9.010013	9.01001	31.0712	0.0014
Error	6	1.739875	0.28998		
Total	7	10.749883			

The mean score for the real eggs was 6.78 with a standard deviation of 0.651. The mean score for the substitute eggs was 4.66 with a standard deviation of 0.395.
a) What are the null and alternative hypotheses?
b) What do you conclude from the ANOVA table?
c) Do the assumptions for the test seem to be reasonable?
d) Perform a two-sample pooled *t*-test of the difference. What P-value do you get? Show that the square of the *t*-statistic is the same (to rounding error) as the *F*-ratio.

T **16. Auto noise filters.** In a statement to a Senate Public Works Committee, a senior executive of Texaco, Inc., cited a study on the effectiveness of auto filters on reducing noise. Because of concerns about performance, two types of filters were studied, a standard silencer and a new device developed by the Associated Octel Company. Here are the boxplots from the data on noise reduction (in decibels) of the two filters. Type 1 = standard; Type 2 = Octel.



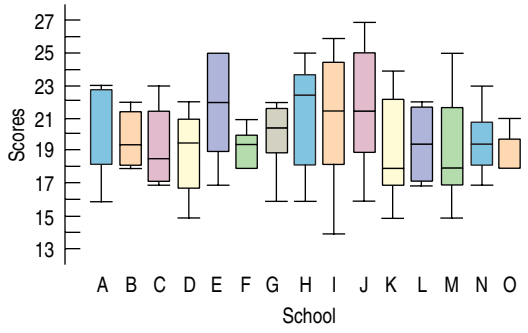
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Type	1	631.186	631.186	0.7673	0.3874
Error	33	27147.386	822.648		
Total	34	27778.572			

Level	*n*	Mean	StdDev
Standard	18	815.556	32.2166
Octel	17	807.059	24.3708

a) What are the null and alternative hypotheses?
b) What do you conclude from the ANOVA table?
c) Do the assumptions for the test seem to be reasonable?
d) Perform a two-sample pooled *t*-test of the difference. What P-value do you get? Show that the square of the *t*-statistic is the same (to rounding error) as the *F*-ratio.

17. School system. A school district superintendent wants to test a new method of teaching arithmetic in the fourth grade at his 15 schools. He plans to select 8 students from each school to take part in the experiment, but to make sure they are roughly of the same ability, he first gives a test to all 120 students. Here are the scores of the test by school:



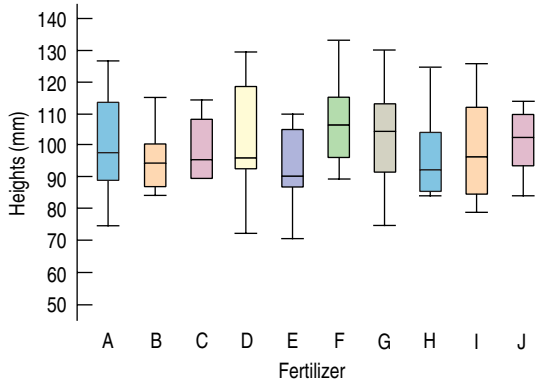
The ANOVA table shows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
School	14	108.800	7.7714	1.0735	0.3899
Error	105	760.125	7.2392		
Total	119	868.925			

a) What are the null and alternative hypotheses?
b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of scores and schools.)
c) An intern reports that he has done *t*-tests of every school against every other school and finds that several of the schools seem to differ in mean score. Does this match your finding in part b? Give an explanation for the difference, if any, of the two results.

18. Fertilizers. A biology student is studying the effect of 10 different fertilizers on the growth of mung bean sprouts. She sprouts 12 beans in each of 10 different petri dishes, and adds the same amount of fertilizer to each dish. After one week she measures the heights of the 120 sprouts in millimeters. Here are boxplots and an ANOVA table of the data:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Fertilizer	9	2073.708	230.412	1.1882	0.3097
Error	110	21331.083	193.919		
Total	119	23404.791			

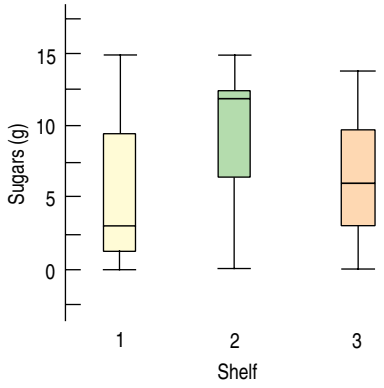
a) What are the null and alternative hypotheses?
b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of heights and fertilizers).
c) Her lab partner looks at the same data and says that he did *t*-tests of every fertilizer against every other fertilizer and finds that several of the fertilizers seem to have significantly higher mean heights. Does this match your finding in part b? Give an explanation for the difference, if any, between the two results.

19. Cereals. Supermarkets often place similar types of cereal on the same supermarket shelf. The same data set we met in Chapter 8 keeps track of the shelf as well as the sugar, sodium, and calorie content of 77 cereals. Does sugar content vary by shelf? Here's a boxplot and an ANOVA table for the 77 cereals:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Shelf	2	248.4079	124.204	7.3345	0.0012
Error	74	1253.1246	16.934		
Total	76	1501.5325			

Level	*n*	Mean	StdDev
1	20	4.80000	4.57223
2	21	9.61905	4.12888
3	36	6.52778	3.83582



a) What are the null and alternative hypotheses?
b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of sugar and shelves.)
c) Can we conclude that cereals on shelf 2 have a higher mean sugar content than cereals on shelf 3? Can we conclude that cereals on shelf 2 have a higher mean sugar content than cereals on shelf 1? What *can* we conclude?
d) To check for significant differences between the shelf means, we can use a Bonferroni test, whose results are shown below. For each pair of shelves, the difference is shown along with its standard error and significance level. What does it say about the questions in part c?

Dependent Variable: SUGARS

	(I) SHELF	(J) SHELF	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
Bonferroni						Lower Bound	Upper Bound
	1	2	−4.819(*)	1.2857	0.001	−7.969	−1.670
		3	−1.728	1.1476	0.409	−4.539	1.084
	2	1	4.819(*)	1.2857	0.001	1.670	7.969
		3	3.091(*)	1.1299	0.023	0.323	5.859
	3	1	1.728	1.1476	0.409	−1.084	4.539
		2	−3.091(*)	1.1299	0.023	−5.859	−0.323

* The mean difference is significant at the .05 level.

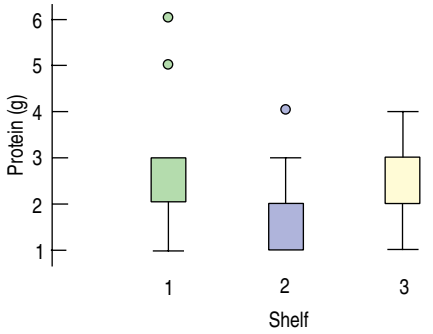
20. Cereals redux. We also have data on the protein content of cereals by their shelf number. Here is the boxplot and ANOVA table:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Shelf	2	12.4258	6.2129	5.8445	0.0044
Error	74	78.6650	1.0630		
Total	76	91.0909			

Means and Std Deviations

Level	*n*	Mean	StdDev
1	20	2.65000	1.46089
2	21	1.90476	0.99523
3	36	2.86111	0.72320



a) What are the null and alternative hypotheses?
b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of protein content and shelves.)
c) Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 3? Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 1? What *can* we conclude?
d) To check for significant differences between the shelf means we can use a Bonferroni test, whose results are shown below. For each pair of shelves, the difference is shown along with its standard error and significance level. What does it say about the questions in part c?

Dependent Variable: PROTEIN
Bonferroni

(I) SHELF	(J) SHELF	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	0.75	0.322	0.070	−0.04	1.53
	3	−0.21	0.288	1.000	−0.92	0.49
2	1	−0.75	0.322	0.070	−1.53	0.04
	3	−0.96(*)	0.283	0.004	−1.65	−0.26
3	1	0.21	0.288	1.000	−0.49	0.92
	2	0.96(*)	0.283	0.004	0.26	1.65

*The mean difference is significant at the .05 level.

T 21. Downloading. To see how much of a difference time of day made on the speed at which he could download files, a college sophomore performed an experiment. He placed a file on a remote server and then proceeded to download it at three different time periods of the day. He downloaded the file 48 times in all, 16 times in each time period.

Time of day	Time (sec)	Time of day	Time (sec)	Time of day	Time (sec)
Early (7 a.m.)	68	Evening (5 p.m.)	299	Late night (12 a.m.)	216
Early (7 a.m.)	138	Evening (5 p.m.)	367	Late night (12 a.m.)	175
Early (7 a.m.)	75	Evening (5 p.m.)	331	Late night (12 a.m.)	274
Early (7 a.m.)	186	Evening (5 p.m.)	257	Late night (12 a.m.)	171
Early (7 a.m.)	68	Evening (5 p.m.)	260	Late night (12 a.m.)	187
Early (7 a.m.)	217	Evening (5 p.m.)	269	Late night (12 a.m.)	213
Early (7 a.m.)	93	Evening (5 p.m.)	252	Late night (12 a.m.)	221
Early (7 a.m.)	90	Evening (5 p.m.)	200	Late night (12 a.m.)	139
Early (7 a.m.)	71	Evening (5 p.m.)	296	Late night (12 a.m.)	226
Early (7 a.m.)	154	Evening (5 p.m.)	204	Late night (12 a.m.)	128
Early (7 a.m.)	166	Evening (5 p.m.)	190	Late night (12 a.m.)	236
Early (7 a.m.)	130	Evening (5 p.m.)	240	Late night (12 a.m.)	128
Early (7 a.m.)	72	Evening (5 p.m.)	350	Late night (12 a.m.)	217
Early (7 a.m.)	81	Evening (5 p.m.)	256	Late night (12 a.m.)	196
Early (7 a.m.)	76	Evening (5 p.m.)	282	Late night (12 a.m.)	201
Early (7 a.m.)	129	Evening (5 p.m.)	320	Late night (12 a.m.)	161

a) State the null and alternative hypotheses, being careful to talk about download times and time of day as well as parameters.
b) Perform an ANOVA on these data. What can you conclude?
c) Check the assumptions and conditions for an ANOVA. Do you have any concerns about the experimental design or the analysis?
d) (Optional) Perform a multiple comparisons test to determine which times of day differ in terms of mean download time.

T 22. Analgesics. A pharmaceutical company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 = extreme pain 30 minutes after taking the drug.

Drug	Pain	Drug	Pain	Drug	Pain
A	4	B	6	C	6
A	5	B	8	C	7
A	4	B	4	C	6
A	3	B	5	C	6
A	2	B	4	C	7
A	4	B	6	C	5
A	3	B	5	C	6
A	4	B	8	C	5
A	4	B	6	C	5

a) State the null and alternative hypotheses, being careful to talk about drugs and pain levels as well as parameters.
b) Perform an ANOVA on these data. What can you conclude?
c) Check the assumptions and conditions for an ANOVA. Do you have any concerns about the experimental design or the analysis?
d) (Optional) Perform a multiple comparisons test to determine which drugs differ in terms of mean pain level reported.