CHAPTER 11

ANOVA

ANALYSIS OF VARIANCE (ANOVA)

The idea behind the ANOVA is very simple. In words: Tabulate all of the variability in your experiment and divide into **variability ACROSS groups versus variability WITHIN groups.** If the variability ACROSS groups is much greater than the variability WITHIN groups, then at least one of the groups has a mean that is significantly different from the others.

ONE-WAY ANOVA (ONE FACTOR OR ONE TREATMENT)

"One way" means that there is a single outcome measure that is being compared across two or more groups simultaneously. Let us walk through the math for one-way ANOVA for an experiment that involves a total of k groups, each one having n datapoints in each group.

The variability of a group is simply its **variance**, which we know (from Chapter 3) is the **Sum of Squares** of the datapoints in the group divided by **df**, its **degrees of freedom**. If there are a total of n items x_i in group 1, and the mean of the datapoints is called m_1, then

$$\text{For group 1, Sum of Squares} = SS_1 = (x_1 - m_1)^2 + (x_2 - m_1)^2 + (x_3 - m_1)^2 + \ldots\ (x_n - m_1)^2$$

$$\text{Within-group SS added up across all groups} = SS_w = SS_1 + SS_2 + SS_3 + \ldots\ SS_n$$

$$\textbf{Within-group variance} = SS_w/df_w$$

where $df_w = nk - k$.

(Note that some textbooks call the variance the "mean square" when talking about ANOVA, and they refer to within-group variability as the "error." This terminology is unfortunate because it obscures the fact that ANOVA uses SS and df concepts, just as normal curves and correlations do.)

In contrast, the across-group variance looks at the deviation of each group's mean from the overall grand mean M of all datapoints in the experiment.

$$\text{Across-group SS added up across all groups} = SS_a = (m_1 - M)^2 + (m_2 - M)^2 + (m_3 - M)^2 + \ldots\ (m_k - M)^2$$

$$\textbf{Across-group variance} = SS_a/df_a$$

where, $df_a = k - 1$.

Data Literacy
http://dx.doi.org/10.1016/B978-0-12-811306-6.00011-7

Finally, we compute the ratio of across-group variability versus within-group variability:

$$\text{Across-group Var/Within-group Var} = (SS_a/df_a)\ /\ (SS_k/df_k)$$

If this value is much greater than 1, then at least one of the groups is significantly different from the others. A ratio of variances follows the **F-distribution** (Chapter 3). To calculate the threshold for significance, one looks up a table of F-values for a given value of df in the numerator (across-group Var, df = k − 1) and df in the denominator (within-group Var, df = nk − k), and for the desired type I error (usually 0.05). See Box 11.1 for an example worked out by hand.

Tip: When the one-way ANOVA is used to compare two groups, the test is mathematically equivalent to the t-test and the F-value is equal to the square of the t-value, or $F = t^2$. Thus, if a t-value of ~2 is at the threshold of significance at $P = .05$, then this corresponds to F-value of ~4.

Because across-group Var + within-group Var = Total Var, sometimes people will not compute the across-group variance directly, but instead will compute total variance and subtract the within-group variance, as follows:

$$\textbf{Total Variance} = SS_{tot} = (x_1 - M)^2 + (x_2 - M)^2 + (x_3 - M)^2 + \ldots (x_n - M)^2$$

$$df_{tot} = nk - 1$$

$$\text{Total Var} = SS_{tot}/df_{tot}$$

$$\text{Across-group Var} = \text{Total Var} - \text{Within-group Var}.$$

ANOVA IS A PARAMETRIC TEST

Like the t-test, the ANOVA is a parametric test, meaning that it makes certain assumptions about the distributions of its datapoints:

- each group is sampled from a normal distribution,
- each datapoint in the same group is independent of the others, randomly sampled, and follows the same underlying population distribution;
- the variance of each group is similar.

Just as we said that the t-test can be applied to data which are quasinormal and can be applied when the variances are "not too different," so can the ANOVA be used to give relatively robust results when these assumptions are violated.

Tip: Never use an ANOVA when there are less than 5 observations or datapoints per group, and a minimum of at least 20 per group is generally preferred.

There are specialized statistical tests that can be used to measure how badly an experimental data set strays in terms of normality and equal variance although these are most useful when the number of datapoints per group is relatively large (>~30). (Equal variance goes by the infelicitous term **homoscedasticity**, which is in keeping with the idiosyncratic terminology used to describe ANOVAs!) Simply plotting and visualizing the data are the best ways to assess normality and equal variance, to decide if ANOVA is a suitable test. If not, nonparametric tests can be done instead (see Chapter 12).

BOX 11.1

AN EXAMPLE OF A ONE-WAY ANALYSIS OF VARIANCE TEST WORKED OUT BY HAND

Shown is a toy data set, with $n = 8$ samples per group, $k = 2$ groups, $Nk = 16$ total samples in the study.

Group 1	Group 2
1	1.5
2	4
3.5	3.5
2	3
8	4
3	2.5
4	1
2.5	3

Total variance:

Grand mean = M = sum of all x_i /N = 3.03125
$SS = \text{sum } (x_1 - 3.03125)^2$
$= (1 - 3.03125)^2 + (2 - 3.03125)^2 + (3.5 - 3.03125)^2 + (2 - 3.03125)^2 + (8 - 3.03125)^2 + (3 - 3.03125)^2 + (4 - 3.03125)^2 + (2.5 - 3.03125)^2 + (1.5 - 3.03125)^2 + (4 - 3.03125)^2 + (3.5 - 3.03125)^2 + (3 - 3.03125)^2 + (4 - 3.03125)^2 + (2.5 - 3.03125)^2 + (1 - 3.03125)^2 + (3 - 3.03125)^2$
$= 4.12 + 1.06 + 0.2197 + 1.06 + 24.688 + 0.00097 + 0.969 + 0.282 + 2.345 + 0.969 + 0.2197 + 0.00097 + 0.969 + 0.2822 + 4.12 + 0.00097$
$= 41.307$
$Df = Nk - 1 = 16 - 1 = 15$
Total variance = $SS/df = 2.7538$

Within-group variance:

Group 1 mean = $26.5/8 = 3.25$
$SS = (1 - 3.25)^2 + (2 - 3.25)^2 + (3.5 - 3.25)^2 + (2 - 3.25)^2 + (8 - 3.25)^2 + (3 - 3.25)^2 + (4 - 3.25)^2 + (2.5 - 3.25)^2$
$= 5.06 + 1.5625 + 0.0625 + 1.5625 + 22.5625 + 0.0625 + 0.5625 + 0.5625$
$= 31.9975$
$Df = nk - k = 16 - 2 = 14$, $Var1 = 31.9975/14 = 2.2855$
Group 2 mean = $22.5/8 = 2.8125$
$SS = (1.5 - 2.8125)^2 + (4 - 2.8125)^2 + (3.5 - 2.8125)^2 + (3 - 2.8125)^2 + (4 - 2.8125)^2 + (2.5 - 2.8125)^2 + (1 - 2.8125)^2 + (3 - 2.8125)^2$
$= 1.722 + 1.41 + 0.473 + 0.0351 + 1.41 + 0.098 + 3.29 + 0.0351$
$= 8.4732$
$Df = nk - k = 16 - 2 = 14$, $Var2 = SS/df = 0.605$
$Var1 + Var2 = 2.8905$

Across-group variance:

- Grand mean = 3.03,125, mean1 = 3.25, mean2 = 2.8125
- $SS = (3.03125 - 3.25)^2 + (3.03125 - 2.8125)^2 = 0.0478 + 0.0478 = 0.0957$
- $\mathbf{Df = k - 1} = 1$, $Var = SS/df = 0.0957$

Now, test significance:

- $\mathbf{Var_{across}/Var_{within}} = 0.0957/2.8905 \ll 1$
- Since the ratio is less than one, we immediately know that this will not be significant!
- The ratio will follow the F-distribution with numerator $k - 1 = 1$, denominator $nk - k = 14$ degrees of freedom
- Use an F-distribution table to look up the *P*-value.
- Or…let your statistical software do the work for you!

TYPES OF ANOVAs

When samples can be paired across groups, this reduces the within-group variability, and this increases the power of the ANOVA to detect small differences as being significant.

To see how this occurs, consider the following two situations:

In the first, we examine systolic blood pressure measured in two independent groups of subjects. One group (subjects 1—3) is measured under baseline conditions, and the other (subjects 4—6) is exposed to a drug that might alter blood pressure. Note that the subjects in the no-drug condition will vary among themselves according to their baseline blood pressures, but the subjects given drug will vary BOTH because they have different baseline blood pressures (not explicitly measured here) AND ALSO because they will vary in their response to the drug.

	Subject 1	**Subject 2**	**Subject 3**
No drug	100	110	120
	Subject 4	**Subject 5**	**Subject 6**
Given drug	140	100	130

In contrast, if the same subjects are measured before versus after drug treatment, the values are paired across the two groups:

	Subject 1	**Subject 2**	**Subject 3**
Before drug	100	110	120
After drug	110	120	135

Here, subjects 1—3 still vary in their baseline blood pressures, but the drug-treated group will ONLY vary in how each subject responds to the drug. The variability within the drug-treated group will be less with paired than with nonpaired designs.

This pairing is a simple example of **repeated-measures** ANOVA. The same subject might be measured at many different times, e.g., weekly, resulting in many paired measurements per subject.

Another related way to reduce within-group variability is **blocking**:

In an unblocked one-way ANOVA, all datapoints in a group are considered together:

	Group 1	**Group 2**
No drug	100, 110, 120	
Given drug		140, 100, 130

In a blocked **two-way** ANOVA, different subgroups, say males and females, are assessed separately

	Group 1		Group 2	
	Males	Females	Males	Females
No drug	100	110, 120		
Given drug			100	140, 130

Blocking reduces overall within-group variability and hence increases the power of the ANOVA, because when computing the Sum of Squares, males are only compared against males and females against females.

A **factorial** ANOVA is similar to a two-way or blocked ANOVA, but it explicitly looks for interactions among the factors when the experiment employs factorial designs (Chapter 5). In the example just given, the ANOVA can detect whether subjects respond differently to the drug, as well as whether different genders respond differently to the drug.

In a **three-way** ANOVA, one outcome is related to three different independent variables. **Multivariate** ANOVAs are performed when there are two or more outcome variables measured for each subject or datapoint. And this is only a partial list—seems that every type of experimental design has a corresponding ANOVA designed to handle it. Generally, you will be using statistical software to perform ANOVA and simply need to choose which type of ANOVA is desired, as well as the desired threshold level of significance (type I error).

THE ANOVA SHOWS SIGNIFICANCE; WHAT NEXT?

Obtaining a significant result in an ANOVA test, say at $P = .05$ or better, merely says that SOME groups or factors are different than others, but does not pinpoint WHICH groups are different. So, after all that effort, you still need to carry out multiple comparisons anyway! Usually these comparisons are done pairwise among each of the groups or subgroups in the experiment. But there is a subtle difference between carrying out pairwise comparisons alone versus doing so after performing ANOVA. Namely, in the former case, each comparison has a false-positive rate of (say) 5% and there is no assurance that any comparisons should be significant. In contrast, the ANOVA provides evidence that a significant effect does exist, with an overall risk of a false positive of 5% (or less, depending on the F-value that was obtained).

So, given a positive ANOVA test, one does further testing using t-tests to identify which groups or factors are different from the others. It is necessary to correct post hoc comparisons for the number of t-tests that are performed, to obtain a more realistic estimate of the false-positive rate.

CORRECTION FOR MULTIPLE TESTING

How to correct for multiple testing most appropriately is an entire subject in itself. Corrections are not specific for ANOVA tests but need to be applied whenever multiple statistical tests are carried out.

Colquhoun's Correction

Perhaps the simplest procedure is to focus only on tests that achieve a P-value of .001 or less. This takes care of most sins that might have been committed in experimental design as well as multiple testing and has the virtue of restricting the investigator's attention to findings that are large and most likely to be reproducible. However, this is a very **conservative** procedure, that is, it reduces the power of each test by bending over backward to avoid false positives. Most scientifically meaningful findings will not satisfy this criterion.

Bonferroni Correction

The most famous method of correcting for multiple tests is the Bonferroni procedure. If the desired type 1 error for each test is 0.05, and you carry out 10 tests, then the corrected threshold for significance is $P = .05/10 = .005$. That is, none of the 10 tests will be deemed significant unless they achieve a P-value of .005 or less. For 100 tests, the corrected threshold is $P = .05/100 = .0005$. This is, again, a very conservative procedure.

The underlying assumption of the Bonferroni procedure is that each test is independent of the others, which is often *not* the case at all. For example, suppose I am comparing two authors, A and B, to see how frequently they use pronouns (say, "he" vs. "she") in the body of their published books. This results in two tests: author A versus author B for "he," and author A versus author B for "she." But is the frequency of the word "he" in a person's writings independent of the word "she"? Maybe or maybe not. If mentions of "he" and "she" are correlated across one or both person's writings (see Chapter 13), then the Bonferroni correction (here, correcting for only two tests) will actually overcorrect and result in a P-value threshold that is too stringent, i.e., too low. In sociology, suppose I am studying the elderly and testing whether two subgroups tend to apply for retirement benefits at different ages. Test A compares low-income versus high-income people, and test B compares people who identify as belonging to one race or another. If there is a correlation between race and income, these two tests are not independent, and again, the Bonferroni procedure will be too conservative.

Other Correction Procedures

A variety of correction methods have been proposed and popularly applied for post hoc analysis of ANOVA test results. These go by names such as the Newman–Keuls method, Tukey's honest significance test, Scheffé's method, and others. They differ in their details and how conservative they are, and none is clearly preferred for all experiments. Choosing one of these goes beyond the scope of this book. With statistical software, it is easy to examine and compare how these different methods will alter the P-values of a finding in your own experiment.

Benjamini–Hochberg Procedure

This is very popular in fields such as bioinformatics where many thousands of statistical tests may be carried out simultaneously. The idea is to specify a desired False Discovery Rate (FDR), which is **the fraction of positive tests that are false positives**. The Benjamini–Hochberg method at a specified FDR of 5% identifies the threshold level of significance, such that only 5% of the tests that achieve that P-value or less will be false positives. This generally gives a corrected P-value threshold that is less conservative than applying the Bonferroni correction. (Bonferroni adjusts $P = .05$ to a corrected threshold of .05/n, where n is the total number of statistical tests carried out, whereas the Benjamini—Hochberg procedure adjusts the threshold to i × 0.05/n, where i is the number of positive tests.)

Permutation Testing

Permutation testing is a nonparametric method of carrying out statistical testing, which has a built-in correction for multiple testing. Permutation tests are discussed in detail in Chapter 12.

You should be aware that if you carry out 20 different t-tests during the course of analyzing your data, on average, one of the tests will achieve significance at $P = .05$ just by chance. So it is important to compare significance levels before versus after correction (by any one of these methods) to judge for yourself whether to take the finding seriously or not. Just as we said there is no magical value of $P = .05$ that determines whether a finding is truly significant or not, there is no magical value of correction that determines whether the finding is worth studying and reporting, either.