Chapter 8
Comparing Means Across Multiple Groups: Analysis of Variance (ANOVA)

The independent samples t tests we discussed in Chap. 6 allow us to compare the means of two groups on some continuously-distributed outcome of interest (e.g., income, age, education, etc.). Often, however, we are interested in comparing the means of more than two groups. For example, we may be interested in comparing mean educational attainment across racial groups, with race measured as "White," "Black," and "Other." In cases such as these, we *could* compare each pair of means using a series of t tests, but this would require the computation of $\binom{g}{2}$ tests. This number becomes unreasonably large quickly as the number of groups, g, increases. For example, with 5 groups, we would have to perform $\binom{5}{2} = 10$ tests. Furthermore, it may be the case that you are uninterested in pairwise differences: Your theory may specifically ask simply whether there are racial differences in educational attainment and not specify which races differ from others. Thus, we need a test that simultaneously compares all means and tells us whether there is variation in the means across a number of groups. Analysis of Variance (ANOVA) is one approach to answering this type of question.

8.1 The Logic of ANOVA

The purpose of ANOVA is to determine whether differences *between* group means are large after accounting for differences in the variances *within* groups that may lead to highly variable group means from sample to sample. ANOVA is called "Analysis of Variance," because it accomplishes the comparison of differences between group means by decomposing the total variance in the outcome measure into within-group variance and between-group variance. If the between-group variance is sufficiently larger than the within-group variance, then the F test that stems from ANOVA calculations leads to the conclusion that there are differences between the means of the groups in the population.

As a demonstration of *how* ANOVA works, suppose we had data on weekly study hours for college students by class (excluding freshmen), and the data looked like

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_8, © Springer Science+Business Media New York 2013

Sophomore	Junior	Senior
4	6	8
4	6	8
4	6	8
4	6	8
4	6	8
4	6	8

Table 8.1. Hypothetical distribution of study hours by year in college

Sophomore	Junior	Senior
4	4	4
4	4	4
6	6	6
6	6	6
8	8	8
8	8	8

Table 8.2. Another hypothetical distribution of study hours by year in college

those presented in Table 8.1. The table shows the study hours for 18 students, six from each class (sophomore year through senior year). In this example, it is clear that there is a difference in studying habits by college year. Why? Because *every* sophomore spent 4 h studying, *every* junior spent 6 h studying, and *every* senior spent 8 h studying per day.

Consider, in contrast, the data presented in Table 8.2. In the data presented in that table, it seems that study hours do not depend on college class at all. Instead, of 6 students in each class, 2 spend 4 h per week studying, 2 spend 6 h per week studying, and 2 spend 8 h.

In both of these examples, the overall sample mean is the same:

$$\frac{4+4+4+4+6+6+6+6+8+8+8+8}{18} = 6 \qquad (8.1)$$

Furthermore, in both examples, the overall sample variance is the same:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 2.82. \qquad (8.2)$$

However, the two samples clearly exhibit different patterns. What differs between the two samples is the composition of study hours within versus between year of study. In the first example, it is clear that knowing class year perfectly predicts study hours, but in the second example, knowing a student's class tells us nothing about his/her study habits.

ANOVA helps us formally differentiate these two samples by decomposing the total sample variance into two components—the variance within groups versus the variance between groups—and then assessing their relative contributions to the total sample variance.

8.2 Some Basic ANOVA Computations

In order to assess the relative contributions of between vs. within group variance to the total sample variance, we must compute both types of variance, as well as the total sample variance. In Chap. 4, we discussed how to compute total sample variance: the computation involves (1) computing the overall sample mean, and (2) computing the average squared deviation of each individual value from the overall sample mean.

To decompose this total sample variance into between-group and within-group variance, we must consider how much individuals within groups deviate from their own group means and how much the group means differ from each other. In order to make this comparison, let's exclude the denominator of the variance calculations for a moment and just obtain the numerators of the variance calculations. These calculations are called the "sums of squares," and there are three such calculations. The first is the "total sum of squares" (abbreviated SST). This sum of squares is the numerator of the sample variance and is calculated as:

$$SST = \sum_{i=1}^{n} \left(x_i - \bar{\bar{x}}\right)^2 \tag{8.3}$$

$$= \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{\bar{x}})^2 \tag{8.4}$$

All of the terms in the first line should be familiar, except $\bar{\bar{x}}$. In ANOVA $\bar{\bar{x}}$ replaces $\bar{x}$ as the overall sample mean. The double-bar differentiates the "grand mean" from the group means that are used in the other sums of squares calculations.

The second line in the SST calculation is an alternative way to write this sum of squares. In this expression, G refers to the total number of groups, n_g is the number of persons in group g. x_{ig} is therefore the value of x for the ith person in group g. This expression explicitly shows that the total sum of squares involves summing the squared deviations of each member (i) of each group (g) from the grand mean.

The second sum of squares calculation is the within-group sum of squares (abbreviated SSW). It measures the amount of variation that exists within each group and is calculated as:

$$SSW = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(x_{ig} - \bar{x}_g\right)^2 . \tag{8.5}$$

Sophomore	Junior	Senior
4	6	8
4	6	8
4	6	8
4	6	8
4	6	8
4	6	8
$\bar{x}_1 = 4$	$\bar{x}_2 = 6$	$\bar{x}_3 = 8$
$\sum_{i=1}^{6}(x_{i1} - \bar{x}_1)^2 = 0$	$\sum_{i=1}^{6}(x_{i2} - \bar{x}_2)^2 = 0$	$\sum_{i=1}^{6}(x_{i3} - \bar{x}_3)^2 = 0$
	↓	
	$SSW = 0 + 0 + 0$	

Table 8.3. Sums of squares for the first hypothetical set of data. Note: $\bar{\bar{x}} = 6$; $SST = \sum_{g=1}^{G=3} \sum_{i=1}^{n_g} (x_{ig} - \bar{\bar{x}})^2 = 48$.

This equation looks similar to the second version of the SST equation, however, $\bar{\bar{x}}$ has been replaced with $\bar{x}_g$. This is the mean of x for group g. Thus, the key difference between SSW and SST is that each individual's value of x is deviated from its *group* mean rather than the grand mean.

The third sum of squares calculation is the between-group sum of squares (abbreviated SSB). It measures the extent of variation between the group means and is calculated as:

$$\sum_{g=1}^{G} n_g (\bar{x}_g - \bar{\bar{x}})^2. \tag{8.6}$$

The calculation simply involves computing the squared deviation of each group's mean ($\bar{x}_g$) from the grand mean ($\bar{\bar{x}}$), weighting this quantity by the group size (n_g), and summing across groups.

To make the idea of sums of squares concrete, consider again the data presented in Table 8.1. Table 8.3 shows the original data, as well as the group means, the overall sample mean, and the sums of squares calculations. From this table, we can see that the sums of squares within each group is 0 (for all groups), while the total sums of squares is 48. In other words, while there is variation across the sample, there is no variation of study hours within classes.

It can be shown that:

$$SST = SSB + SSW. \tag{8.7}$$

In these hypothetical data, then, given that SSW is 0 and SST is 48, SSB must be 48.

Sophomore	Junior	Senior
4	4	4
4	4	4
6	6	6
6	6	6
8	8	8
8	8	8
$\bar{x}_1 = 6$	$\bar{x}_2 = 6$	$\bar{x}_3 = 6$
$\sum_{i=1}^{6}(x_{i1} - \bar{x}_1)^2 = 16$	$\sum_{i=1}^{6}(x_{i2} - \bar{x}_2)^2 = 16$	$\sum_{i=1}^{6}(x_{i3} - \bar{x}_3)^2 = 16$

$$\downarrow$$

$$SSW = 16 + 16 + 16$$

Table 8.4. Sums of squares for the second hypothetical set of data. Note: $\bar{\bar{x}} = 6$; $SST = \sum_{g=1}^{G=3} \sum_{i=1}^{n_g}(x_{ig} - \bar{\bar{x}})^2 = 48$.

In contrast, what happens if we apply these calculations to the second set of hypothetical data? Table 8.4 shows these results. In this case, the SSW is 48, the SSB is 0, and the SST is (still) 48.

Comparing these two sets of results is revealing. In the first case, SSB = SST; in that second case, SSW = SST. In other words, in the first scenario, all of the variation in study hours in the overall sample was attributable to between group variation. All of the sophomores looked the same, all of the juniors looked the same, and all of the seniors looked the same, but the classes differed from each other. The only variation across the sample was attributable to year of study. In the latter case, on the other hand, there was considerable variation in study habits between members of the same class, but, overall, the classes looked identical.

These results suggest a simple statistic to differentiate these samples, namely, the proportion of the total sample variance (or sums of squares) that is between-group variance. This statistic is called R^2 and is computed as:

$$R^2 = \frac{SSB}{SST} \equiv 1 - \frac{SSW}{SST}. \tag{8.8}$$

In the first example, $R^2 = 1$; in the latter, $R^2 = 0$. In other words, 100 % of the variance in the first sample was explained by the groupings we chose (year of college), but 0 % of the variance in the second sample was explained by our groupings. R^2 is an important quantity in statistics, and we will see it again when we discuss the correlation coefficient and regression modeling. However, the R^2 statistic we have derived leaves several things to be desired. Importantly, R^2 by itself tells us nothing about whether the observed between-group differences we've observed could be attributable to sampling fluctuation. Second, how high of an R^2 do we need before we can conclude that, in the population, the outcome of interest (differences in sample means) in fact varies across the groups we're interested in differentiating? The F statistic from an ANOVA table can help us with this determination.

8.3 A Real ANOVA Example

In the examples presented in the previous section, either all of the variation in the sample was explained by college class or none of it was. In reality, we can never fully explain variation in an outcome on the basis of a unidimensional classification of people. Instead, real data tend to produce R^2 values that are substantially less than 1: some of the variance in the sample is due to between group differences, while some is due to within group variance. An ANOVA table is constructed in order to show these relative contributions to the overall sample variance and to present an F statistic, which can be assessed to determine whether the ratio of between group variance to within group variance is sufficiently large to reject a null hypothesis that the means of all groups are equal.

A generic ANOVA table follows the format shown in Table 8.5. As the table shows, if one knows (1) the sample size, (2) the number of groups, (3) the total sums of squares, and (4) one other of the two types of sums of squares, the remainder of the table can be computed with no additional information. The first column of the table simply labels the source of the sums of squares. The second column presents the three sums of squares discussed previously. The third column presents the degrees of freedom associated with the sums of squares. Notice that the degrees of freedom for the between and within sums of squares sum to the total degrees of freedom, which is simply the denominator of the total sample variance formula. The fourth column presents the mean sums of squares, which are simply the various sums of squares divided by their degrees of freedom. Finally, the fifth column presents the F statistic, which is simply the ratio of the between and within mean squares.

Source	Sum of Squares	DF	Mean Squares	F
Between	$\sum_{g=1}^{G} n_g \times (\bar{x}_g - \bar{\bar{x}})^2$	$G-1$	$MSB = \frac{SSB}{df(B)}$	$\frac{MSB}{MSE}$
Within	$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)^2$	$n-G$	$MSE = \frac{SSW}{df(W)}$ *	
Total	$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{\bar{x}})^2$	$n-1$	(Sample Variance**)	

* The notation for the within sum of squares divided by its degrees of freedom changes: MSE is called the mean squared error—it is the portion of the variance unexplained by the grouping variable.

** Generally not included in the actual table, but is implicit.

Table 8.5. Format for a generic one-way ANOVA table.

Source	Sum of Squares	DF	Mean Squares	F
Between	483.91	2	241.95	26.57 ($p < .001$)
Within	12903.73	1417	9.11	
Total	13387.64	1419	9.43	$R^2 = .036$

Table 8.6. ANOVA example: Race differences in educational attainment (2006 GSS data).

The F statistic is a measure of the extent to which the total variance in the variable of interest is accounted for by the grouping variable. The null hypothesis in the F test is that the group means are equal; therefore, none of the variance is from group differences. In repeated sampling, of course, differences between group means will sometimes be evident, just as we have discussed throughout earlier chapters. Thus, the F statistic can help us assess, under the assumption that there are no differences in group means, how likely it is that we could obtain the difference in means we observed in a random sample. As with other statistical tests, the rarer our observed sample would be under this null hypothesis, the more confident we can be in rejecting it.

The F statistic can be assessed using an F table, just as you have assessed z scores using a z (normal distribution) table, t scores using a t table, and χ^2 statistics using a χ^2 table—the F distribution is simply another distribution. Like the t and χ^2 distributions, the F distribution has degrees of freedom associated with it; however, the F distribution has a pair of degrees of freedom, generally called "numerator" and "denominator" degrees of freedom. These degrees of freedom for ANOVA are $df(B)$ and $df(W)$ respectively.

Table 8.6 an ANOVA table for education by race obtained from Stata using the 2006 GSS data. Race is measured as a three category variable (black, white, other), and educational attainment is measured in years. Mean attainment for blacks was 12.66 (s.d. = 3.15); the mean for whites was 13.72 (s.d. = 2.79); the mean for others was 12.14 (s.d. = 4.03). The F test in the ANOVA table tests the hypothesis that there is no difference in these means. As the results show, we can reject this null hypothesis, with the implication being that there is at least one mean that is different from the others. Realize, however, the test does not tell us which mean(s) differ(s) from which.

8.4 Conclusions

In this chapter, we have developed a statistical approach to examining differences in means between multiple groups. The method—ANOVA—extends the independent samples t test to handle more than two groups simultaneously. The null hypothesis of interest in ANOVA is that all group means are equal. Rejecting this null hypothesis tells us only that at least one mean differs from the other; it does not tell us which mean or means differ from which.

ANOVA modeling can be extended to simultaneously examine more than one grouping variable (called "two-way" ANOVA). We did not discuss such approaches; instead, we will discuss regression modeling approaches to incorporating additional variables. Regression modeling and ANOVA modeling are similar in many ways, but regression modeling is generally more flexible and more commonly used across the social sciences.

8.5 Items for Review

- Between, within, and total sums of squares
- R^2 (Explained variance)
- The F test and F distribution
- Numerator and denominator degrees of freedom
- The ANOVA table

8.6 Homework

1. Below is a partially-completed ANOVA table examining racial differences in family income from the 2006 GSS. Complete the table, state the null hypothesis, and draw a conclusion.

Source	Sum of Squares	DF	Mean Squares	F
Between	78169.41	2		
Within				
Total	2328739.13	1419		$R^2 =$

2. Below are data from a hypothetical clinical trial testing the effectiveness of a new drink, the manufacturers of which claim the drink boosts energy levels. The clinical trial involved three groups: a treatment group (the members of which received the beverage), a placebo group (the members of which received a "fake" beverage), and a control group (the members of which received no beverage at all). Assume that participants in the trial were randomly selected

from the population and were randomly assigned to treatment/placebo/control groups. Also assume that participants' energy levels were measured before and after receiving the beverage, and the reported values in the table below represent the increase (or decrease) in energy level after receiving the drink (or, in the case of the control group, after NOT receiving anything). Is there evidence that the energy drink works?

Person	Treatment Group		
	Treatment	Placebo	Control
1	0	1	2
2	3	1	0
3	1	2	0
4	5	3	3
5	2	0	1

3. Below is a table with descriptive statistics for health by level of education, where education is defined by three groups: those with less than a high school diploma, those with a high school diploma, and those with education beyond high school. Using these data, construct the appropriate ANOVA table and test the null hypothesis that mean health is constant across levels of education.

Education	n	Mean Health	s.d. for Health
Less than H.S.	223	1.60	.88
High School	372	1.93	.79
More than H.S.	825	2.12	.78
All	1420	1.99	.82

4. Below is a partially-completed ANOVA table examining political party (Democrat, Independent, Republican) differences in happiness (assume happiness is continuous). Complete the table, state the null hypothesis, and draw a conclusion.

Source	Sum of Squares	DF	Mean Squares	F
Between		2		
Within	553.11			
Total	565.76	1419		$R^2 =$

5. Below is a table with descriptive statistics for life satisfaction by first term of three US Presidents: Reagan (1981–1984), Bush Sr. (1989–1992); and Clinton (1993–1996). Is there variation in mean life satisfaction by presidency? (Hint: you will have to determine the overall mean for life satisfaction using the subsample means).

President	n	Mean satisfaction	s.d. for satisfaction
Reagan	2356	22.73	4.61
Bush Sr.	1085	23.06	4.70
Clinton	580	22.80	4.69

6. Are their regional differences in health, based on the region in which a respondent was raised? Below are data on health (0 = poor. . . 3 = excellent) by region at age 16 from the 2010 GSS (persons age 40).

Person	Region at 16			
	Northeast	Midwest	South	West
1	2	1	2	1
2	2	1	2	2
3	3	2	2	3
4	3	2	2	3
5		3	2	
6			2	
7			2	
8			3	