INTRODUCTION TO ONE-WAY ANALYSIS OF VARIANCE

Dale Berger, Claremont Graduate University http://wise.cgu.edu

The purpose of this paper is to explain the logic and vocabulary of one-way analysis of variance (ANOVA). The null hypothesis (H_0) tested by one-way ANOVA is that two or more population means are equal. A statistically significant test indicates that observed data sampled from each of the populations would be unlikely if the null hypothesis were true.

The logic used in ANOVA to compare means of multiple groups is similar to that used with the t-test to compare means of two independent groups. When one-way ANOVA is applied to the special case of two groups, one-way ANOVA gives identical results as the t-test.

Not surprisingly, the assumptions needed for the t-test are also needed for ANOVA. We need to assume:

1) random, independent sampling from the **k** populations;
2) normal population distributions for each of the **k** populations;
3) equal variances within the **k** populations.

Assumption 1 is crucial for any inferential statistic. As with the t-test, Assumptions 2 and 3 can be relaxed when large samples are used, and Assumption 3 can be relaxed when the sample sizes are roughly the same for each group even for small samples. (If there are extreme outliers or errors in the data, we need to deal with that problem first.) As a first step in this introduction to ANOVA, we will review the t-test for two independent groups, to prepare for an extension to ANOVA.

Review of the t-test to test the equality of means for two independent groups

Let us start with a small example. Suppose we wish to compare two training programs in terms of performance scores for people who have completed these training courses. The table below shows scores for six randomly selected graduates from each of two training programs. These (artificially) small samples show somewhat lower scores from the first program than from the second program. But, can fluctuations between means this large be expected from chance in the sampling process or is this compelling evidence of a real difference in the populations? The t-test for independent groups is designed to address just this question by testing the null hypothesis H_0: $\mu_1 = \mu_2$ where μ_1 and μ_2 are the two population means, respectively. In this short paper, we will conduct a standard t-test for two independent groups, but will develop the logic in a way that can be extended easily to more than two groups.

		Program 1		Program 2
		102		100
		90		108
		97		104
		94		111
		98		105
		101		102
Mean	$\bar{y}_1 =$	**97**	$\bar{y}_2 =$	**105**
Variance	$s_1^2 =$	20	$s_2^2 =$	16

The mean of all 12 scores = Grand mean = $\bar{y}_{..} =$ **101**

The first step is to check the data to make sure that the raw data are correctly assembled and that assumptions have not been violated in a way that makes the test inappropriate. A plot of our data shows that the sample distributions have roughly the same shape, and neither sample has extreme scores or extreme skew. The sample sizes are equal, so equality of population variances is of little concern. Note that in usual applications you would have much larger samples.

We assume that the variance is the same within the two populations (Assumption 3). An unbiased estimate of this common population variance can be calculated separately from each sample. The numerator of the formula for computing an estimate of the population variance from either sample is the sum of squared deviations around the sample mean, or simply the sum of squares for the sample. The sum of squares for sample j is abbreviated as SS_j. The denominator is the degrees of freedom for the population variance estimate from sample j (abbreviated as df_j).

$$\text{Unbiased estimate of } \sigma_j^2 = \frac{\sum_i \left(y_{ij} - \bar{y}_j\right)^2}{(n_j - 1)} = \frac{SS_j}{df_j} = s_j^2 \qquad \text{[Formula 1]}$$

For the first sample, $SS_1 = (102\text{-}97)^2 + ... + (101\text{-}97)^2 = 100$, and for the second sample, $SS_2 = 80$. This leads to two estimates of the population variance σ_y^2: $s_1^2 = 100/5 = 20$, and $s_2^2 = 80/5 = 16$.

To pool two or more sample estimates of a single population variance, each sample variance is weighted by its degrees of freedom. This is equivalent to adding together the sums of squares for the separate estimates, and dividing by the total of the degrees of freedom for the separate estimates.

$$\text{Pooled estimate of } \sigma_y^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} = \frac{SS_1 + SS_2}{df_1 + df_2} = s_y^2 \qquad \text{[Formula 2]}$$

Thus, for our example

$$s_y^2 = \frac{(6\text{-}1)(20) + (6\text{-}1)(16)}{(6 + 6 - 2)} = \frac{100 + 80}{5 + 5} = \frac{180}{10} = 18$$

A t-test can be conducted to assess the statistical significance of the difference between the sample means. The null hypothesis is that the population means are equal (H_0: $\mu_1 = \mu_2$).

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_y^2\left(1/n_1 + 1/n_2\right)}} = \frac{97 - 105}{\sqrt{18\left(1/6 + 1/6\right)}} = \frac{-8}{\sqrt{6}} = -3.266$$

$df = (n_1 + n_2 - 2) = (6 + 6 - 2) = 10$.

For a two-tailed *t*-test with alpha set at .01 and $df = 10$, the tabled critical value is 3.169. Because the absolute value of the observed *t* exceeds the critical value we can reject the null hypothesis (H_o: $\mu_1 = \mu_2$) at the .01 level of significance. (The exact $p = .0085$.) If our assumptions are valid, then the probability that we will find a *t*-value so far from zero if the two population means are equal is less than 1%. We can conclude that the mean for the population of people represented by Sample 2 probably is larger than the mean for the population of people represented by Sample 1.

An equivalent test of the null hypothesis can be calculated with the F distribution, because t^2 with df = ν is exactly equal to F (df = 1, ν). (The Greek letter ν "nu" is often used to represent df for the *t*-test.) For our example, $t^2 = (-3.266)^2 = 10.67$. From the F table, F(1, 10; .01) = 10.04, so we find that the null hypothesis can just be rejected at the .01 level of significance (actual $p = .0085$). This test result is identical to the result of the t test.

ANOVA as a comparison of two estimates of the population variance

In this section we examine a conceptual approach that can be extended directly to one-way analysis of variance with **k** groups. We can use our data to calculate two independent estimates of the population variance: one is the pooled variance of scores within groups, and the other is derived from the observed variance of the group means. These two estimates of the population variance are expected to be equal if the population means are equal for all **k** groups (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$), but the estimates are expected to differ if the population means are not all the same.

Within-groups estimate. Our single best estimate of the population variance is the pooled within groups variance, s_y^2 from Formula 2. In our example $s_y^2 = 18$, with df = 10. In ANOVA terminology, the numerator of Formula 2 is called the Sum of Squares Within Groups, or SS_{WG}, and the denominator is called the degrees of freedom Within Groups, or df_{WG}. The estimate of the population variance from Formula 2, SS_{WG}/df_{WG}, is called the Mean Square Within Groups, or MS_{WG}. Formula 3 is an equivalent way to express and compute MS_{WG}.

$$\text{Within-groups estimate of } \sigma_y^2 = \frac{\sum_{ij}\left(y_{ij} - \bar{y}_j\right)^2}{\sum_{j}\left(n_j - 1\right)} = \frac{SS_{WG}}{df_{WG}} = MS_{WG} \qquad \text{[Formula 3]}$$

$$= \frac{100 + 80}{5 + 5} = \frac{180}{10} = 18.00$$

Between-groups estimate. If the null hypothesis ($\mu_1 = \mu_2$) is true and the assumptions are valid (random, independent sampling from normally distributed populations with equal variances), then a second independent estimate of the population variance can be calculated. As is stated by the Central Limit Theorem, if independent samples of size **n** are drawn from a population with variance = σ_y^2, then the variance of all possible such sample means $\sigma_{\bar{y}}^2$ is σ_y^2/n. We can use our observed sample means to calculate an unbiased estimate of the variance for the distribution of all possible sample means (for samples of size **n**). Our estimate of the variance of means is not very stable because it is based on only two scores, $\bar{y}_1$=97 and $\bar{y}_2$=105, but nonetheless it is an unbiased estimate of $\sigma_{\bar{y}}^2$. With our data, $est\,\sigma_{\bar{y}}^2 = s_{\bar{y}}^2 = 32$ and df = 1, as calculated with Formula 4.

$$est\,\sigma_{\bar{y}}^2 = s_{\bar{y}}^2 = \frac{\sum_{j}\left(\bar{y}_j - \bar{y}..\right)^2}{k - 1} \qquad \text{[Formula 4]}$$

$$= \frac{(97 - 101)^2 + (105 - 101)^2}{(2 - 1)} = \frac{(-4)^2 + (4)^2}{1} = 16 + 16 = 32.$$

Note: This is the variance of **means**, which would be smaller than the variance within groups if the population means were all equal.

Because $\sigma_{\bar{y}}^2 = \sigma_y^2/n$, it follows that $\sigma_y^2 = n\,\sigma_{\bar{y}}^2$. Now we can estimate the variance of the population σ_y^2 based on the observed variance of 32 for the sample means. With our data, where n = 6 for each sample, we find $s_y^2 = (n)(s_{\bar{y}}^2) = (6)(32) = 192$. This tells us that if we draw samples of size $n_j = 6$ from a population where $\sigma_y^2 = 192$, the expected variance of sample means is $\sigma_{\bar{y}}^2 = \sigma_y^2/n = 192/6 = 32$. Thus, if the groups in the population have equal means and equal within group variances, we can estimate this common within group variance to be 192 in the population, because that would account for the observed variance of 32 between our sample means. This new estimate of the population variance within groups comes from observed variance between group means.

Calculation of this second estimate of the population variance using ANOVA notation is shown in Formula 5. The MS_{BG} is our best estimate of the population variance based only on knowledge of the variance among the sample means. Formula 5 allows for unequal sample sizes.

$$\text{Between-groups estimate of } \sigma_y^2 = \frac{\sum_j n_j\left(\bar{y}_j - \bar{y}_{..}\right)^2}{(k-1)} = \frac{SS_{BG}}{df_{BG}} = MS_{BG} \qquad \text{[Formula 5]}$$

$$\frac{\sum_j n_j\left(\bar{y}_j - \bar{y}_{..}\right)^2}{(k-1)} = \frac{6(97-101)^2 + 6(105-101)^2}{2-1} = \frac{6(16)+6(16)}{1} = 192.$$

Comparing the two estimates of population variance. The estimate of the population variance based on the variability between sample means ($MS_{BG} = 192$) is considerably larger than the estimate of population variance based on variability within samples ($MS_{WG} = 18$). We should like to know how likely it is that two estimates of the same population variance would differ so widely if all of our assumptions are valid and ($\mu_1 = \mu_2$). The F ratio is designed to test this question. (H_o: $\sigma_1^2 = \sigma_2^2$)

$$F(df_{BG}, df_{WG}) = \frac{\textit{Between Groups estimate of } \sigma_y^2}{\textit{Within Groups estimate of } \sigma_y^2} = \frac{MS_{BG}}{MS_{WG}} \qquad \text{[Formula 6]}$$

$$F(1, 10) = \frac{192}{18} = 10.67 \quad (p = .0085)$$

The degrees of freedom for the two estimates of variance in Formula 6 are $df_{BG} = k - 1 = 2 - 1 = 1$, and $df_{WG} = (n_1 + n_2 - k) = (6 + 6 - 2) = 10$. Notice that these values are exactly the same *F* ratio and degrees of freedom that we calculated earlier when we converted the t-test to an *F*-test.

If the null hypothesis and assumptions were true, such that independent random samples were drawn from two normally distributed populations with equal means and equal variances, then it would be very surprising indeed ($p < .01$; actual $p = .0085$) to find that these two estimates of the common population variance within each group (σ_y^2) would differ so widely.

We conclude that it is not likely that the null hypothesis and all assumptions are true. If we are confident that our assumptions are OK, then we reject the null hypothesis (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$).

More than two groups: One-way ANOVA

The extension to more than two groups is easy. Formula 5 and Formula 3 can be used directly to calculate the between-groups and within-groups estimates of the population variance for any number of groups **k**, and Formula 6 can be used to test the two estimates for equality. The *p*-value from the F test is the probability of observing an F ratio this large or larger if the population means are equal for all groups and all assumptions are met.

Assumptions of the test. We can expect our calculated level of statistical significance (the *p*-value from the *F* distribution) to be accurate only if the assumptions required for the test procedure have been satisfied. Recall the assumptions:

1) the observations were randomly and independently chosen from each of **k** populations;
2) population distributions are normal for each of the **k** populations that was sampled; and
3) population variances are equal for all **k** populations.

If the sampling was not independent and random, the results of the *F*-test may be completely spurious. No statistical procedure will allow strong generalizations to a population if random sampling is not used. Fortunately, the sampling procedure is generally under the control of the researcher, so faulty sampling as an explanation for a surprisingly large *F* usually can be ruled out.

Perhaps the best approach to identify serious departures from normality in the shape of the population distributions is to plot the sample distributions and apply the "intraocular trauma test." Extreme departures from normality, especially strong skew or outliers, will be apparent. Admittedly, some practice is needed to calibrate your eyeballs, but a plot is likely to be more useful than summary statistics alone for identifying problems in your data. Distributions with isolated extreme scores typically cause more serious problems than smoothly skewed distributions.

There are several ways to deal with extreme scores. Transformations may be useful to reduce the effects of extreme scores (and reduce skew). Sometimes an outlier is caused by an error in coding that can be corrected. Be especially alert for missing data codes that accidentally are used as legitimate data. Sometimes outliers are legitimate scores from cases that are qualitatively different from the population of interest. Such cases should be removed and treated separately. They may be very interesting and important cases, so they should not automatically be ignored. "Robust" methods are less sensitive to extreme scores. With Winsorized data, some number (**g**) of scores in each tail of the distribution are set equal to the next most extreme score (the **g+1**st score from the end). With trimmed data, some proportion of the scores from each tail are discarded. A popular level of trimming is 15% from each end. Hampel and biweight procedures retain all data but give less weight to scores farther from the mean. Resampling and bootstrapping may also be appropriate.

Equality of variance can be tested, but there are compelling arguments against using such a test to decide whether or not to use ANOVA. First, ANOVA is little affected by small to moderate departures from homogeneity of variance, especially if the sample sizes are equal or nearly equal. Second, the tests of homogeneity are more powerful for larger samples than for smaller samples, but ANOVA is less affected by heterogeneity when the samples are larger. This leads to the awkward situation where the tests of homogeneity are most likely to detect a violation of homogeneity when it least matters. Third, several of the most commonly used tests of homogeneity are inaccurate for

non-normal distributions. This includes Bartlett's test, F max, and Cochran's C (see Kirk, 1994, for a discussion of these tests). Levene's test of homogeneity of variance is less sensitive to departures from normality. Box (1953) characterized testing for homogeneity of variance before using ANOVA as sending a rowboat out into the ocean to see if it is calm enough for an ocean liner.

Unequal within-group variances for the different populations is a problem for ANOVA (and the t-test) only when three conditions exist simultaneously – I call this the "triple whammy": 1) the variances are quite unequal (say a 2:1 ratio or greater); 2) the samples are quite unequal in size (say a 2:1 ratio or greater), and 3) at least one sample is small (say 10 or fewer cases). In this situation, ANOVA is too liberal (gives false significance) when the smallest samples are taken from the populations with the largest variance. Conversely, ANOVA is too conservative (fails to detect real differences among means) when the smallest samples are taken from the populations with the smallest variance (see Boneau, 1960). Many statistical packages, including SPSS, provide tests of equality of variance in ANOVA and alternative ANOVA tests that do not assume equal variance.

If you suspect that an assumption of ANOVA has been violated in a way that compromises the test, it is prudent to supplement the regular analyses with procedures that are robust to the suspected violation of the assumption. If both approaches yield the same conclusions, report the results from the standard test and note that the results were confirmed with the robust procedure. If the results differ, considerable caution is warranted, and the more conservative test is probably appropriate.

Statistical Significance, Effect size, Practical Significance, and Non-Significance

The practice of significance testing of null hypotheses of no difference has come under severe criticism (e.g., see an excellent paper by Nickerson, 2000). Statistical significance depends heavily on sample size. A *p*-value does not tell us about effect size or importance of the finding. Failure to reject the null hypothesis does not mean that the null hypothesis is true (it almost certainly is false). On the other hand, if we reject the null hypothesis, we have not necessarily found a large, practical, or important effect. It is important to examine plots of the data, and to report the effect size (perhaps means and variances or confidence intervals) and practical implications, not just *p*-values.

Sources

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57,* 49-64.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt-Brace.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth.

Kirk, R. E. (1994). *Experimental design* (3rd ed.). Belmont, CA: Wadsworth.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

CHAN1509