

One-Way Independent Samples Analysis of Variance®

If we are interested in the relationship between a **categorical IV** and a **continuous DV**, the two categories analysis of variance (ANOVA) may be a suitable inferential technique. If the IV had only two levels (groups), we could just as well do a *t*-test, but the ANOVA allows us to have **2 or more categories**. The null hypothesis tested is that $\mu_1 = \mu_2 = \dots = \mu_k$, that is, all *k* treatment groups have identical population means on the DV. The alternative hypothesis is that at least two of the population means differ from one another.

We start out by making two **assumptions**:

- Each of the *k* populations is **normally distributed** and
- **Homogeneity of variance** - each of the populations has the same variance, the IV does not affect the variance in the DV. Thus, if the populations differ from one another they differ in location (central tendency, mean).

The **model** we employ here states that each score on the DV has two components:

- the effect of the treatment (the IV, Groups) and
- error, which is anything else that affects the DV scores, such as individual differences among subjects, errors in measurement, and other extraneous variables. That is, $Y_{ij} = \mu + \tau_j + e_{ij}$, or, $Y_{ij} - \mu = \tau_j + e_{ij}$

The difference between the grand mean (μ) and the DV score of subject number *i* in group number *j*, Y_{ij} , is equal to the effect of being in treatment group number *j*, τ_j , plus error, e_{ij}

[Note that I am using *i* as the subscript for subject # and *j* for group #]

Computing ANOVA Statistics From Group Means and Variances, Equal *n*.

Let us work with the following contrived data set. We have randomly assigned five students to each of four treatment groups, A, B, C, and D. Each group receives a different type of instruction in the logic of ANOVA. After instruction, each student is given a 10 item multiple-choice test. Test scores (# items correct) follow:

Group	Scores					Mean
A	1	2	2	2	3	2
B	2	3	3	3	4	3
C	6	7	7	7	8	7
D	7	8	8	8	9	8

Now, do these four samples differ enough from each other to reject the null hypothesis that type of instruction has no effect on mean test performance? First, we use the sample data to estimate the amount of error variance in the scores in the population from which the samples were randomly drawn. That is variance (differences among scores) that is due to anything other than the IV. One simple way to do this, assuming that you have an equal number of scores in each sample, is to compute the average within group variance,

$$MSE = \frac{s_1^2 + s_2^2 + \dots + s_k^2}{k}$$

s_j^2 is the sample variance in Group number j .

Thought exercise: Randomly chose any two scores that are in the same group. If they differ from each other, why? Is the difference because they got different treatments? Of course not, all subjects in the same group got the same treatment. It must be other things that caused them to have different scores. Those other things, collectively, are referred to as “error.”

MSE is the **mean square error** (aka mean square within groups): “Mean” because we divided by k , the number of groups, “square” because we are working with variances, and “error” because we are estimating variance due to things other than the IV. For our sample variances the **MSE** = $(.5 + .5 + .5 + .5) / 4 = 0.5$

MSE is not the only way to estimate the population error variance. If we assume that the null hypothesis is true, we can get a second estimate of population error variance that is independent of the first estimate. We do this by finding the sample variance of the k sample means and multiplying by n , where n = number of scores in each group (assuming equal sample sizes). That is,

$$MS_A = n * s_{means}^2$$

I am using **MS_A** to stand for the **estimated among groups or treatment variance** for Independent Variable A. Although you only have one IV now, you should later learn how to do ANOVA with more than one IV. For our sample data we compute the variance of the four sample means, $VAR(2,3,7,8) = 26 / 3$ and multiply by n , so $MS_A = 5 * 26 / 3 = 43.33$.

Now, our second estimate of error variance, the variance of the means, MS_A , assumes that the null hypothesis is true. Our first estimate, **MSE**, the mean of the variances, made no such assumption. If the null hypothesis is true, these two estimates should be approximately equal to one another. If not, then the MS_A will estimate not only error variance but also variance due to the IV, and $MS_A > MSE$. We shall determine whether the difference between MS_A and **MSE** is large enough to reject the null hypothesis by using the **F**-statistic. **F** is the ratio of two independent variance estimates. We shall compute **$F = MS_A / MSE$ which, in terms of estimated variances, is the effect of error and treatment divided by the effect of error alone. If the null hypothesis is true, the treatment has no effect, and $F = [error / error] =$ approximately one. If the null hypothesis is false, then $F = [(error + treatment) / error] > 1$. Large values of **F** cast doubt on the null hypothesis, small values of **F** do not.**

When first developed by Sir [Ronald A. Fisher](#), the **F** statistic was known as the variance ratio. Later, [George W. Snedecor](#) named it **F**, in honor of Fisher

For our data, $F = 43.33 / .5 = 86.66$. Is this **F** large enough to reject the null hypothesis or might it have happened to be this large due to chance? To find the probability of getting an **F** this large or larger, our exact significance level, p , we must work with the **sampling distribution of F**. This is the distribution that would be obtained if you repeatedly drew sets of k samples of n scores each all from identical populations and computed MS_A / MSE for each set. It is a positively skewed sampling distribution with a mean of about one. Using the **F**-table, we can approximate p . Like **t**-distributions, **F**-distributions have degrees of freedom, but unlike **t**, **F** has **df** for numerator (MS_A) and **df** for denominator (**MSE**). The **total df** in the k samples is $N - 1$ (where N = total # scores) because the total variance is computed using sums of squares for N scores about one point, the grand mean. The **treatment A df** is $k - 1$ because it is computed using sums of squares for k scores (group means) about one point, the grand mean. The **error df** is $k(n - 1)$ because **MSE** is computed using k within groups sums of squares each computed on n scores about one point, the group mean.

For our data, total $df = N - 1 = 20 - 1 = 19$. Treatment A $df = k - 1 = 4 - 1 = 3$. Error $df = k(n - 1) = N - k = 20 - 4 = 16$. Note that total $df =$ treatment $df +$ error df .

So, what is p ? Using the **F**-table in our text book, we see that there is a 5% probability of getting an $F(3, 16) >= 3.24$. Our $F > 3.24$, so our $p < .05$. The table also shows us that the upper 1% of an **F**-distribution on 3, 16 df is at and beyond $F = 5.29$, so our $p < .01$. We can reject the null

hypothesis even with an a priori alpha criterion of .01. Note that **we are using a one-tailed test with nondirectional hypotheses**, because regardless of the actual ordering of the population means, for example, $\mu_1 > \mu_2 > \mu_3 > \mu_4$ or $\mu_1 > \mu_4 > \mu_3 > \mu_2$, etc., etc., any deviation in any direction from the null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4$ will cause the value of F to increase. Thus we are only interested in the upper tail of the F -distribution.

Derivation of Deviation Formulae for Computing ANOVA Statistics

Let's do the ANOVA again using different formulae. Let's start by computing the **total sum-of-squares** (SS_{TOT}) and then partition it into treatment (SS_A) and error (SSE) components.

We shall derive formulae for the ANOVA from its model. If we assume that the error component is normally distributed and independent of the IV, we can derive formulas for the ANOVA from this model. First we substitute sample statistics for the parameters in the model:

$$Y_{ij} = GM + (M_j - GM) + (Y_{ij} - M_j)$$

Y_{ij} is the score of subject number i in group number j , GM is the grand mean, the mean of all scores in all groups, M_j is the mean of the scores in the group (j) in which Y_{ij} is.

Now we subtract GM from each side, obtaining:

$$(Y_{ij} - GM) = (M_j - GM) + (Y_{ij} - M_j)$$

Next, we square both sides of the expression, obtaining:

$$(Y_{ij} - GM)^2 = (M_j - GM)^2 + (Y_{ij} - M_j)^2 + 2(M_j - GM)(Y_{ij} - M_j)$$

Now, summing across subjects (i) and groups (j),

$$\sum_{ij}(Y_{ij} - GM)^2 = \sum_{ij}(M_j - GM)^2 + \sum_{ij}(Y_{ij} - M_j)^2 + 2 * \sum_{ij}(M_j - GM)(Y_{ij} - M_j)$$

Now, since the sum of the deviations of scores about their mean is always zero, $2 * \sum_{ij}(M_j - GM)(Y_{ij} - M_j)$ equals zero, and thus drops out, leaving us with:

$$\sum_{ij}(Y_{ij} - GM)^2 = \sum_{ij}(M_j - GM)^2 + \sum_{ij}(Y_{ij} - M_j)^2$$

Within each group ($M_j - GM$) is the same for every Y_{ij} , so

$\sum_{ij}(M_j - GM)^2$ equals $\sum_j [n_j * (M_j - GM)^2]$, leaving us with

$$\sum_{ij}(Y_{ij} - GM)^2 = \sum_j [n_j * (M_j - GM)^2] + \sum_{ij}(Y_{ij} - M_j)^2$$

Thus, we have partitioned the leftmost term (SS_{TOT}) into SS_A (the middle term) and SSE (the rightmost term).

$$SS_{TOT} = \sum (Y_{ij} - GM)^2.$$

For our data, $SS_{TOT} = (1 - 5)^2 + (2 - 5)^2 + \dots + (9 - 5)^2 = 138$.

To get the SS_A , the **among groups or treatment sum of squares**, for each score subtract the grand mean from the mean of the group in which the score is. Then square each of these deviations and sum them. Since the squared deviation for group mean minus grand mean is the same for every score within any one group, we can save time by computing SS_A as:

$$SS_A = \sum [n_j * (M_j - GM)^2]$$

[Note that each group's contribution to SS_A is weighted by its n , so groups with larger n 's have more influence. This is a weighted means ANOVA. If we wanted an "unweighted means" (equally weighted means) ANOVA we could use a harmonic mean $n_h = k \div \sum_j (1/n_j)$ in place of n (or just be sure we have equal n 's, in which case the weighted means analysis is an equally weighted means analysis). With unequal n 's and an equally weighted analysis,

$SS_{TOT} \neq SS_A + SSE$.

Given equal sample sizes (or use of harmonic mean n_h), the formula for SS_A simplifies to:

$$SS_A = n * \sum (M_j - GM)^2.$$

For our data, $SS_A = 5[(2 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (8 - 5)^2] = 130$.

The **error sum of squares**, $SSE = \sum (Y_{ij} - M_j)^2$.

These error deviations are all computed within treatment groups, so they reflect variance not due to the IV, that is, error. Since every subject within any one treatment group received the same treatment, variance within groups must be due to things other than the IV. For our data, $SSE = (1 - 2)^2 + (2 - 2)^2 + \dots + (9 - 8)^2 = 8$. Note that $SS_A + SSE = SS_{TOT}$. Also note that for each SS we summed across all N scores the squared deviations between either Y_{ij} or M_j and either M_j or GM . If we now divide SS_A by its df and SSE by its df we get the same mean squares we earlier obtained.

Computational Formulae for ANOVA

Unless group and grand means are nice small integers, as was the case with our contrived data, the above method (deviation formulae) is unwieldy. It is, however, easier to see what is going on in ANOVA with that method than with the computational method I am about to show you. Use the following computational formulae to do ANOVA on a more typical data set. In these formulae G stands for the total sum of scores for all N subjects and T_j stands for the sum of scores for treatment group number j .

$$SS_{TOT} = \sum Y^2 - \frac{G^2}{N}$$

$SS_A = \sum \frac{T_j^2}{n_j} - \frac{G^2}{N}$, which simplifies to: $SS_A = \frac{\sum T_j^2}{n} - \frac{G^2}{N}$ when sample size is constant across groups.

$$SSE = SS_{TOT} - SS_A.$$

For our sample data,

$$SS_{TOT} = (1 + 4 + 4 + \dots + 81) - [(1 + 2 + 2 + \dots + 9)^2] \div N = 638 - (100)^2 \div 20 = 138$$

$$SS_A = [(1+2+2+2+3)^2 + (2+3+3+3+4)^2 + (6+7+7+7+8)^2 + (7+8+8+8+9)^2] \div 5 - (100)^2 \div 20 = 130$$

$$SSE = 138 - 130 = 8.$$

ANOVA Source Table and APA-Style Summary Statement

Summarizing the ANOVA in a **source table**:

Source	SS	df	MS	F
Teaching Method	130	3	43.33	86.66
Error	8	16	0.50	
Total	138	19		

In an APA journal the results of this analysis would be summarized this way: "Teaching method significantly affected test scores, $F(3, 16) = 86.66$, $MSE = 0.50$, $p < .001$, $\omega^2 = .93$." If the researcher had a means of computing the exact significance level, that would be reported. For example, one might report " $p = .036$ " rather than " $p < .05$ " or " $.01 < p < .05$." One would also typically refer to a table or figure with basic descriptive statistics (group means, sample sizes, and standard deviations) and would conduct some additional analyses (like the pairwise comparisons we shall study in our next lesson). If you are confident that the population variances are homogeneous, and

have reported the *MSE* (which is an estimate of the population variances), then reporting group standard deviations is optional.

Violations of Assumptions

You should use boxplots, histograms, comparisons of mean to median, and/or measures of skewness and kurtosis (available in SAS, the Statistical Analysis System, a delightful computer package) on the scores within each group to evaluate the normality assumption and to identify outliers that should be investigated (and maybe deleted, if you are willing to revise the population to which you will generalize your results, or if they represent errors in data entry, measurement, etc.). If the normality assumption is not tenable, you may want to transform scores or use a nonparametric analysis. If the sample data indicate that the populations are symmetric, or, slightly skewed but all in the same direction, the ANOVA should be sufficiently robust to handle the departure from normality.

You should also compute F_{max} , the ratio of the largest within-group variance to the smallest within-group variance. If F_{max} is less than 4 or 5 (especially with equal or nearly equal sample sizes and normal or nearly normal within-group distributions), then the ANOVA should be sufficiently robust to handle the departure from homogeneity of variance. If not, you may wish to try data transformations or a nonparametric test, keeping in mind that if the populations cannot be assumed to have identical shapes and dispersions, rejection of the nonparametric null hypothesis cannot be interpreted as meaning the populations differ in location.

The origin of the " $F_{max} < 4$ or 5 " rule of thumb has eluded me, but [Wilcox et al.](#) hinted at it in 1986. Their article also points out that the ANOVA F is not nearly as robust to its homogeneity of variance assumption as many would like to believe. There are [modern robust statistical procedures](#) which do not assume homogeneity of variance and normality, but they are not often used, probably in large part because they are not very easy to obtain with popular software like SPSS and SAS.

David Howell advised "In general, if the populations can be assumed to be symmetric, or at least similar in shape (e.g., all negatively skewed), and if the largest variance is no more than four times the smallest, the analysis of variance is most likely to be valid. It is important to note, however, that heterogeneity of variance and unequal sample sizes do not mix. If you have reason to anticipate unequal variances, make every effort to keep sample sizes as <nearly> equal as possible. This is a serious issue and people tend to forget that noticeably unequal sample sizes make the test appreciably less robust to heterogeneity of variance." (Howell, D. C. (2013). *Statistical methods for psychology*, 8th Ed. Belmont, CA: Wadsworth).

It is possible to **adjust the df to correct for heterogeneity of variance**, as we did with the separate variances t -test. **Box** has shown that the true critical F under heterogeneity of variance is somewhere between the critical F on $1, (n - 1)$ df and the unadjusted critical F on $(k - 1), k(n - 1)$ df , where n = the number of scores in each group (equal sample sizes). It might be appropriate to use a harmonic mean $n_h = k \div \sum_j (1/n_j)$, with unequal sample sizes (consult Box - the reference is in Howell). If your F is significant on $1, (n - 1)$ df , it is significant at whatever the actual adjusted df are. If it is not significant on $(k - 1), k(n - 1)$ df , it is not significant at the actual adjusted df . If it is significant on $(k - 1), k(n - 1)$ df but not on $1, (k - 1)$ df , you don't know whether or not it is significant with the true adjusted critical F .

If you cannot reach an unambiguous conclusion using Box's range for adjusted critical F , you may need to resort to **Welch's test**, explained in our textbook (and I have an example below). You must compute for each group W , the ratio of sample size to sample variance. Then you compute an adjusted grand mean, an adjusted F , and adjusted denominator df .

You may prefer to try to meet the assumptions by employing nonlinear transformations of the data prior to analysis. Here are some suggestions:

When the group standard deviations appear to be a linear function of the group means (try correlating the means with the standard deviations or plotting one against the other), a **logarithmic**

transformation should reduce the resulting heterogeneity of variance. Such a transformation will also reduce positive skewness, since the log transformation reduces large scores more than small scores. If you have negative scores or scores near zero, you will need to add a constant (so that all scores are 1 or more) before taking the log, since logs of numbers of zero or less are undefined.

If group means are a linear function of group variances (plot one against the other or correlate them), a **square root transformation** might do the trick. This will also reduce positive skewness, since large scores are reduced more than small scores. Again, you may need to first add a constant, c , or use $\sqrt{X + c}$ to avoid imaginary numbers like $\sqrt{-1}$.

A **reciprocal transformation**, $T = 1/Y$ or $T = -1/Y$, will very greatly reduce large positive outliers, a common problem with some types of data, such as running times in mazes or reaction times.

If you have negative skewness in your data, you may first **reflect** the variable to convert negative skewness to positive skewness and then apply one of the transformations that reduce positive skewness. For example, suppose you have a variable on a scale of 1 to 9 which is negatively skewed. Reflect the variable by subtracting each score from 10 (so that 9's become 1's, 8's become 2's, 7's become 3's, 6's become 4's, 4's become 6's, 3's become 7's, 2's become 8's, and 1's become 9's). Then see which of the above transformations does the best job of normalizing the data. Do be careful when it comes time to interpret your results—if the original scale was 1 = complete agreement with a statement and 9 = complete disagreement with the statement, after reflection high scores indicate agreement and low scores indicate disagreement. This is also true of the reciprocal transformation $1/Y$ (but not $-1/Y$). For more information on the use of data transformation to reduce skewness, see my documents [Using SAS to Screen Data](#) and [Using SPSS to Screen Data](#).

Where Y is a proportion, p , for example, proportion of items correct on a test, variances (npq , binomial) will be smaller in groups where mean p is low or high than in groups where mean p is close to .5. An **arcsine transformation**, $T = 2 * \text{ARCSINE}(Y)$, may help. It may also normalize by stretching out both tails relative to the middle of the distribution.

Another option is to **trim the samples**. That is, throw out the extreme $X\%$ of the scores in each tail of each group. This may stabilize variances and reduce kurtosis in heavy-tailed distributions. A related approach is to use **Winsorized samples**, where all of the scores in the extreme $X\%$ of each tail of each sample are replaced with the value of the most extreme score remaining after trimming. The modified scores are used in computing means, variances, and test statistics (such as F or t), but should not be counted in n when finding error df for F , t , s^2 , etc. Howell suggests using the scores from trimmed samples for calculating sample means and MS_A , but Winsorized samples for calculating sample variances and MSE .

If you have used a nonlinear transformation such as log or square-root, it is usually best to report sample means and standard deviations this way: find the sample means and standard deviations on the transformed data and then reverse the transformation to obtain the statistics you report. For example, if you used a log transformation, find the mean and sd of log-transformed data and then the antilog (INV LOG on most calculators) of those statistics. For square-root-transformed data, find the square of the mean and sd , etc. These will generally not be the same as the mean and sd of the untransformed data.

How do you choose a transformation? I usually try several transformations and then evaluate the resulting distributions to determine which best normalizes the data and stabilizes the variances. It is not, however, proper to try many transformations and choose the one that gives you the lowest significance level - to do so inflates alpha. Choose your transformation prior to computing F or t . Do check for adverse effects of transformation. For example, a transformation that normalizes the data may produce heterogeneity of variance, in which case you might need to conduct a Welch test on transformed data. If the sample distributions have different shapes a transformation that normalizes

the data in one group may change those in another group from nearly normal to negatively skewed or otherwise nonnormal.

Some people get very upset about using nonlinear transformations. If they think that their untransformed measurements are interval scale data, linear transformations of the true scores, they delight in knowing that their computed t 's or F 's are exactly the same that would be obtained if they had computed them on God-given true scores. But if a nonlinear transformation is applied, the transformed data are only ordinal scale. Well, keep in mind that Fechner and Stevens (psychophysical laws) have shown us that our senses also provide only ordinal data, positive monotonic (but usually not linear) transformation of the physical magnitudes that constitute one reality. Can we expect more of our statistics than of our senses? I prefer to simply generalize my findings to that abstract reality which is a linear transformation of my (sensory or statistical) data, and I shall continue to do so until I get a hot-line to God from whom I can obtain the truth with no distortion.

Do keep in mind that one additional nonlinear transformation available is to rank the data and then conduct the analysis on the ranks. This is what is done in most nonparametric procedures, and they typically have simplified formulas (using the fact that the sum of the integers from 1 to n equals $n(n + 1) \div 2$) with which one can calculate the test statistic.

Computing ANOVA Statistics From Group Means and Variances with Unequal Sample Sizes and Heterogeneity of Variance

Wilbur Castellow (while he was chairman of our department) wanted to evaluate the effect of a series of changes he made in his introductory psychology class upon student ratings of instructional excellence. Institutional Research would not provide the raw data, so all we had were the following statistics:

Semester	Mean	SD	N	p_j
Spring 89	4.85	.360	34	34/133 = .2556
Fall 88	4.61	.715	31	31/133 = .2331
Fall 87	4.61	.688	36	36/133 = .2707
Spring 87	4.38	.793	32	32/133 = .2406

1. Compute a weighted mean of the K sample variances. For each sample the weight is

$$p_j = \frac{n_j}{N}.$$

$$MSE = \sum p_j s_j^2 = .2556(.360)^2 + .2331(.715)^2 + .2707(.688)^2 + .2406(.793)^2 = .4317.$$

2. Obtain the Among Groups SS, $\sum n_j (M_j - GM)^2$.

$$\text{The } GM = \sum p_j M_j = .2556(4.85) + .2331(4.61) + .2707(4.61) + .2406(4.38) = 4.616.$$

Among Groups SS =

$$34(4.85 - 4.616)^2 + 31(4.61 - 4.616)^2 + 36(4.61 - 4.616)^2 + 32(4.38 - 4.616)^2 = 3.646.$$

With 3 df , $MSA = 1.215$, and $F(3, 129) = 2.814$, $p = .042$.

3. Before you get excited about this significant result, notice that the sample variances are not homogeneous. There is a negative correlation between sample mean and sample variance, due to a ceiling effect as the mean approaches its upper limit, 5. The ratio of the largest to the smallest variance is $.793^2/.360^2 = 4.852$, which is significant beyond the .01 level with Hartley's maximum F -ratio statistic (a method for testing the null hypothesis that the variances are homogeneous). Although the sample sizes are close enough to equal that we might not worry about violating the homogeneity of variance assumption, for instructional purposes let us make some corrections for the heterogeneity of variance.

4. Box (1954, see our textbook) tells us the critical (.05) value for our F on this problem is somewhere between $F(1, 30) = 4.17$ and $F(3, 129) = 2.675$. Unfortunately our F falls in that range, so we don't know whether or not it is significant.

5. The Welch procedure ([see the formulae in our textbook](#)) is now our last resort, since we cannot transform the raw data (which we do not have).

$$W_1 = 34 / .360^2 = 262.35,$$

$$W_2 = 31 / .715^2 = 60.638, W_3 = 36 / .688^2 = 76.055, \text{ and } W_4 = 32 / .793^2 = 50.887.$$

$$X! = \frac{262.35(4.85) + 60.638(4.61) + 76.055(4.61) + 50.887(4.38)}{262.35 + 60.638 + 76.055 + 50.887} = \frac{2125.44}{449.93} = 4.724.$$

The numerator of $F'' =$

$$\frac{262.35(4.85 - 4.724)^2 + 60.638(4.61 - 4.724)^2 + 76.055(4.61 - 4.724)^2 + 50.887(4.38 - 4.724)^2}{3} =$$

3.988. The denominator of F'' equals

$$1 + \frac{4}{15} \left[\left(\frac{1}{33} \right) \left(\frac{1 - 262.35}{449.93} \right)^2 + \left(\frac{1}{30} \right) \left(\frac{1 - 60.638}{449.93} \right)^2 + \left(\frac{1}{35} \right) \left(\frac{1 - 76.055}{449.93} \right)^2 + \left(\frac{1}{31} \right) \left(\frac{1 - 50.887}{449.93} \right)^2 \right] =$$

$1 + 4 / 15(.07532) = 1.020$. Thus, $F'' = 3.988 / 1.020 = 3.910$. Note that this F'' is greater than our standard F . Why? Well, notice that each group's contribution to the numerator is inversely related to its variance, thus increasing the contribution of Group 1, which had a mean far from the Grand Mean and a small variance.

We are not done yet, we still need to compute adjusted error degrees of freedom; $df' = (15) / [3(.07532)] = 66.38$. Thus, $F(3, 66) = 3.910, p = .012$.

Directional Hypotheses

I have never seen published research where the authors used ANOVA and employed a directional test, but it is possible. Suppose you were testing the following directional hypotheses:

H_0 : The classification variable is not related to the outcome variable in the way specified in the alternative hypothesis

$$H_1: \mu_1 > \mu_2 > \mu_3$$

The one-tailed p value that you obtain with the traditional F test tells you the probability of getting sample means as (or more) different from one another, in any order, as were those you obtained, were the truth that the population means are identical. Were the null true, the probability of your correctly predicting the order of the differences in the sample means is $k!$, where k is the number of groups. By application of the multiplication rule of probability, the probability of your getting sample means as different from one another as they were, and in the order you predicted, is the one-tailed p times $k!$. If k is three, you take the one-tailed p and divide by $3 \times 2 = 6$ – a one-sixth tailed test. I know, that sounds strange. Lots of luck convincing the reviewers of your manuscript that you actually PREDICTED the order of the means. They will think that you POSTdicted them.

Fixed vs. Random vs. Mixed Effects ANOVA

As in correlation/regression analysis, the IV in ANOVA may be fixed or random. If it is fixed, the researcher has arbitrarily (based on e's opinion, judgement, or prejudice) chosen k values of the IV. E will restrict e's generalization of the results to those k values of the IV. E has defined the population of IV values in which e is interested as consisting of only those values e actually used, thus, e has used the entire population of IV values. For example, I give subjects 0, 1, or 3 beers and measure reaction time. I can draw conclusions about the effects of 0, 1, or 3 beers, but not about 2 beers, 4 beers, 10 beers, etc.

With a random effects IV, one randomly obtains levels of the IV, so the actual levels used would not be the same if you repeated the experiment. For example, I decide to study the effect of dose of phenylpropanolamine upon reaction time. I have my computer randomly select ten dosages from a uniform distribution of dosages from zero to 100 units of the drug. I then administer those 10 dosages to my subjects, collect the data, and do the analyses. I may generalize across the entire range of values (doses) from which I randomly selected my 10 values, even (by interpolation or extrapolation) to values other than the 10 I actually employed.

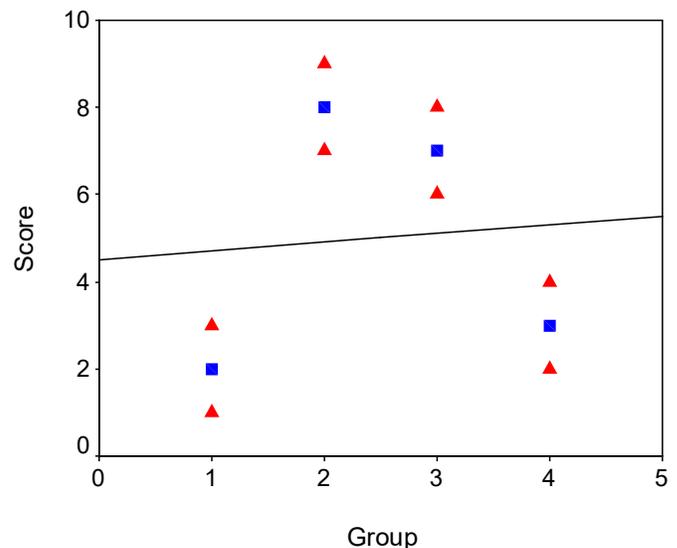
In a factorial ANOVA, one with more than one IV, you may have a mixed effects ANOVA - one where one or more IV's is fixed and one or more is random.

Statistically, our one-way ANOVA does actually have two IV's, but one is sort of hidden. The hidden IV is SUBJECTS. Does who the subject is affect the score on the DV? Of course it does, but we count such effects as error variance in the one-way independent samples ANOVA. Subjects is a random effects variable, or at least we pretend it is, since we randomly selected subjects from the population of persons (or other things) to which we wish to generalize our results. In fact, if there is not at least one random effects IV in your research, you don't need ANOVA or any other inferential statistic. If all of your IV's are fixed, your data represent the entire population, not a random sample therefrom, so your descriptive statistics are parameters and you need not infer what you already know for sure.

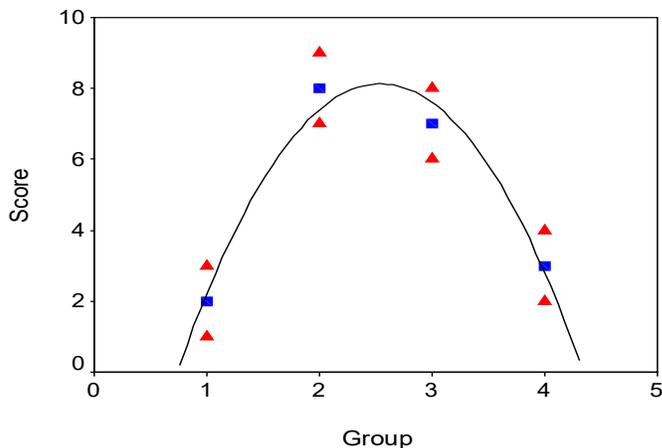
ANOVA as a Regression Analysis

The ANOVA is really just a special case of a regression analysis. It can be represented as a multiple regression analysis, with one dichotomous "dummy variable" for each treatment degree of freedom (more on this in another lesson). It can also be represented as a bivariate, curvilinear regression.

Here is a scatter plot for our ANOVA data. Since the numbers used to code our groups are arbitrary (the independent variable being qualitative), I elected to use the number 1 for Group A, 2 for Group D, 3 for Group C and 4 for Group B. Note that I have used blue squares to plot the points with a frequency of three and red triangles to plot those with a frequency of one. The blue squares are also the group means. I have placed on the plot the linear regression line predicting score from group. The regression falls far short of significance, with the $SS_{Regression}$ being only 1, for an r^2 of $1/138 = .007$.



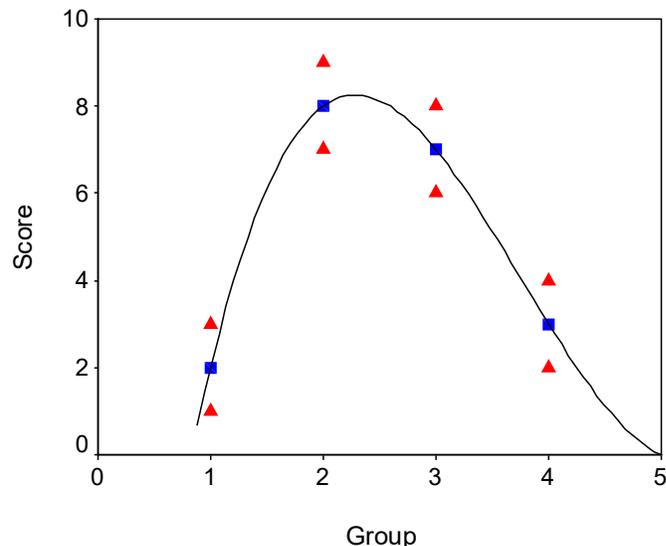
We could improve the fit of our regression line to the data by removing the restriction that it be a straight line, that is, by doing a curvilinear regression. A quadratic regression line is based on a polynomial where the independent variables are Group and Group-squared — that is, $\hat{Y} = a + b_1X + b_2X^2$ — more on this when we cover trend analysis. A quadratic function allows us one bend in the curve. Here is a plot of our data with a quadratic regression line.



Eta-squared (η^2) is a curvilinear correlation coefficient. To compute it, one first uses a curvilinear equation to predict values of $Y|X$. You then compute the SS_{Error} as the sum of squared residuals between actual Y and predicted Y , that is, $SSE = \sum (Y - \hat{Y})^2$. As usual,

$SS_{Total} = \sum (Y - GM)^2$, where GM is the grand mean, the mean of all scores in all groups. The $SS_{Regression}$ is then $SS_{Total} - SS_{Error}$. Eta squared is then $SS_{Regression} / SS_{Total}$, the proportion of the SS_{Total} that is due to the curvilinear association with X . For our quadratic regression (which is highly significant), $SS_{Regression} = 126$, $\eta^2 = .913$.

We could improve the fit a bit more by going to a cubic polynomial model (which adds Group-cubed to the quadratic model, allowing a second bending of the curve). Here is our scatter plot with the cubic regression line. Note that the regression line runs through all of the group means. This will always be the case when we have used a polynomial model of order = $K - 1$, where K = the number of levels of our independent variable. A cubic model has order = 3, since it includes three powers of the independent variable (Group, Group-squared, and Group-cubed). The $SS_{Regression}$ for the cubic model is 130, $\eta^2 = .942$. Please note that this $SS_{Regression}$ is exactly the same as that we computed earlier as the ANOVA $SS_{Among\ Groups}$. We have demonstrated that a polynomial regression with order = $K - 1$ is identical to the traditional one-way ANOVA.



Take a look at my document [T = ANOVA = Regression](#).

Strength of Effect Estimates – Proportions of Variance Explained

We can employ η^2 as a measure of the magnitude of the effect of our ANOVA independent variable without doing the polynomial regression. We simply find $\frac{SS_{AmongGroups}}{SS_{Total}}$ from our ANOVA source table. This provides a fine measure of the strength of effect of our independent variable in our sample data, but it generally overestimates the population η^2 . My programs [Conf-Interval-R2-Regr.sas](#) and [CI-R2-SPSS.zip](#) will compute a confidence interval for population η^2 . For our data $\eta^2 = 130/138 = .94$. A 95% confidence interval for the population parameter extends from .84 to .96. It might be better to report a 90% confidence interval here, more on that soon.

One well-known alternative is **omega-squared**, ω^2 , which estimates the proportion of the variance in Y in the population which is due to variance in X. $\omega^2 = \frac{SS_{Among} - (K - 1)MS_{Error}}{SS_{Total} + MS_{Error}}$. For our data, $\omega^2 = \frac{130 - (3).5}{138 + .5} = .93$.

Benchmarks for η^2 .

- .01 (1%) is small but not trivial
- .06 is medium
- .14 is large

A Word of Caution. Rosenthal has found that most psychologists misinterpret strength of effect estimates such as r^2 and ω^2 . Rosenthal (1990, *American Psychologist*, 45, 775-777.) used an example where a treatment (a small daily dose of aspirin) lowered patients' death rate so much that the researchers conducting this research the research prematurely and told the participants who were in the control condition to start taking a baby aspirin every day. So, how large was the effect of the baby aspirin? As an odds ratio it was 1.83 – that is, the odds of a heart attack were 1.83 times higher in the placebo group than in the aspirin group. As a proportion of variance explained the effect size was .0011 (about one tenth of one percent).

One solution that has been proposed for dealing with r^2 -like statistics is to report their square root instead. For the aspirin study, we would report $r = .033$ (but that still sounds small to me).

Also, keep in mind that anything that artificially lowers “error” variance, such as using homogeneous subjects and highly controlled laboratory conditions, artificially inflates r^2 , ω^2 , etc. Thus, under highly controlled conditions, one can obtain a very high ω^2 even if outside the laboratory the IV accounts for almost none of the variance in the DV. In the field those variables held constant in the lab may account for almost all of the variance in the DV.

What Confidence Coefficient Should I Employ for η^2 and *RMSSE*?

If you want the confidence interval to be equivalent to the ANOVA *F* test of the effect (which employs a one-tailed, upper tailed, probability) you should employ a confidence coefficient of $(1 - 2\alpha)$. For example, for the usual .05 criterion of statistical significance, use a 90% confidence interval, not 95%. Please see my document [Confidence Intervals for Squared Effect Size Estimates in ANOVA: What Confidence Coefficient Should be Employed?](#).

Strength of Effect Estimates – Standardized Differences Among Means

When dealing with differences between or among group means, I generally prefer strength of effect estimators that rely on the standardized difference between means (rather than proportions of

variance explained). We have already seen such estimators when we studied two group designs (estimated Cohen's d) – but how can we apply this approach when we have more than two groups?

My favorite answer to this question is that you should just report estimates of Cohen's d for those contrasts (differences between means or sets of means) that are of most interest – that is, which are most relevant to the research questions you wish to address. Of course, I am also of the opinion that we would often be better served by dispensing with the ANOVA in the first place and proceeding directly to making those contrasts of interest without doing the ANOVA.

There is, however, another interesting suggestion. We could estimate the average value of Cohen's d for the groups in our research. There are several ways we could do this. We could, for example, estimate d for every pair of means, take the absolute values of those estimates, and then average them.

James H. Steiger (2004: *Psychological Methods*, 9, 164-182) has proposed the use of *RMSSE* (root mean square standardized effect) in situations like this. Here is how the *RMSSE* is calculated:

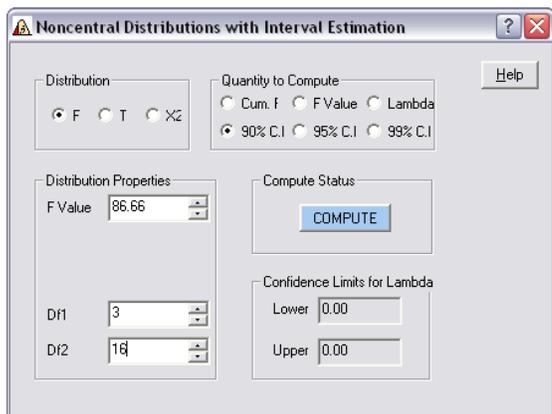
$$RMSSE = \sqrt{\left(\frac{1}{k-1}\right) \sum_1^k \left(\frac{M_j - GM}{\sqrt{MSE}}\right)^2}$$
, where k is the number of groups, M_j is a group mean, GM is the

overall (grand) mean, and the standardizer is the pooled standard deviation, the square root of the within groups mean square, MSE (note that we are assuming homogeneity of variances). Basically what we are doing here is averaging the values of $(M_j - GM)/SD$, having squared them first (to avoid them summing to zero), dividing by among groups degrees of freedom ($k - 1$) rather than k , and then taking the square root to get back to un-squared (standard deviation) units.

Since the standardizer (sqrt of MSE) is constant across groups, we can simplify the expression above to $RMSSE = \sqrt{\left(\frac{1}{k-1}\right) \frac{\sum(M_j - GM)^2}{MSE}}$.

For our original set of data, the sum of the squared deviations between group means and grand mean is $(2-5)^2 + (3-5)^2 + (7-5)^2 + (8-5)^2 = 26$. Notice that this is simply the among groups sum of squares (130) divided by n (5). Accordingly, $RMSSE = \sqrt{\left(\frac{1}{4-1}\right) \frac{26}{.5}} = 4.16$, a Godzilla-sized average standardized difference between group means.

We can place a confidence interval about our estimate of the average standardized difference between group means. To do so we can use the NDC program from Steiger's page at <http://www.statpower.net/Content/NDC/NDC.exe>. Download and run that exe. Ask for a 90% CI and give the values of F and df .



Click "COMPUTE."

You are given the *CI* for lambda, the noncentrality parameter:

Now we transform this confidence interval to a confidence interval for *RMSSE* by with the following transformation (applied to each end of the *CI*): $RMSSE = \sqrt{\frac{\lambda}{(k-1)n}}$. For the lower boundary, this yields $\sqrt{\frac{120.6998}{(3)5}} = 2.837$, and for the upper boundary $\sqrt{\frac{436.3431}{(3)5}} = 5.393$. That is, our estimate of the effect size is between King Kong-sized and beyond Godzilla-sized.

Steiger noted that a test of the null hypothesis that Ψ (the parameter estimated by *RMSSE*) = 0 is equivalent to the standard ANOVA *F* test if the confidence interval is constructed with $100(1-2\alpha)\%$ confidence. For example, if the ANOVA were conducted with .05 as the criterion of statistical significance, then an equivalent confidence interval for Ψ should be at 90% confidence -- Ψ cannot be negative, after all. If the 90% confidence interval for Ψ includes 0, then the ANOVA *F* falls short of significance, if it excludes 0, then the ANOVA *F* is significant.

Power Analysis

One-way ANOVA power analysis can be conducted by hand using the noncentral *F* table in

our textbook. The effect size is specified in terms of $\Sigma\tau^2$: $\phi' = \sqrt{\frac{\sum_{j=1}^k \sum (\mu_j - \mu)^2}{k\sigma_{error}^2}}$. Cohen used the

symbol *f* for this same statistic, and considered an *f* of .10 to represent a small effect, .25 a medium effect, and .40 a large effect. In terms of percentage of variance explained η^2 , small is 1%, medium is 6%, and large is 14%.

$\sum_{j=1}^k \sum (\mu_j - \mu)^2$ is the sum of the squared deviations between the group means and the overall (grand) mean. When you divide that by *k*, the number of groups, you get the population variance of

the group means. When you divide by σ_{error}^2 , and then take the square root, you get the standardized difference among the group means.

Suppose that I wish to test the null hypothesis that for GRE-Q, the population means for undergraduates intending to major in social psychology, clinical psychology, and experimental psychology are all equal. I decide that the minimum nontrivial effect size is if each mean differs from the next by 12.2 points (about $.12 \sigma$). For example, population means of 487.8, 500, and 512.2. The $\Sigma \tau^2$ is then $12.2^2 + 0^2 + 12.2^2 = 297.68$. Next we compute ϕ' . Assuming that the σ is about 100, $\phi' = \sqrt{297.68/3/10000} = 0.10$.

To use the noncentral F table in Howell, we compute $\phi = \phi' * \sqrt{n}$

Suppose we have 25 subjects in each group. $\phi = .10 * \sqrt{25} = .50$.

Treatment $df = 2$, error $df = 3(25 - 1) = 72$. From the **noncentral F table** in our text book, for $\phi = .50$, $df_t = 2$, $df_e = \text{infinite}$, $\alpha = .05$, $\beta = 89\%$, thus power = 11%.

How many subjects would be needed to raise power to 80%? $\beta = .80$. Go to the table, assuming that you will need enough subjects so that $df_e = \text{infinity}$. For $\beta = .80$, $\phi = 1.8$. Now, $n = (\phi^2)(k)(\sigma_e^2) / \Sigma \tau^2 = (1.8)^2(3)(100)^2 / 297.68 = 327$. Total required $N = 3(327) = 981$.

Make it easy on yourself. Use [G*Power](#) to do the power analysis.

The screenshot shows the G*Power software interface for an ANOVA analysis. The 'Test family' is set to 'F tests' and the 'Statistical test' is 'ANOVA: Fixed effects, omnibus, one-way'. The 'Type of power analysis' is 'Post hoc: Compute achieved power - given alpha, sample size, and effect size'. The 'Input Parameters' section includes: Effect size f (0.0996126), alpha error prob (0.05), Total sample size (75), and Number of groups (3). The 'Output Parameters' section includes: Noncentrality parameter lambda (0.7442003), Critical F (3.1239074), Numerator df (2), Denominator df (72), and Power (1-beta error prob) (0.1076733). On the right, a table shows the group means and sizes: Group 1 (Mean: -12.2, Size: 25), Group 2 (Mean: 0, Size: 25), and Group 3 (Mean: 12.2, Size: 25). The 'Calculate' button is highlighted in blue.

F tests – ANOVA: Fixed effects, omnibus, one-way

Analysis: A priori: Compute required sample size

Input: Effect size f = 0.0996126
 alpha error prob = 0.05
 Power (1-beta error prob) = .80
 Number of groups = 3

Output: Noncentrality parameter lambda = 9.6746033
 Critical F = 3.0049842
 Numerator df = 2

Denominator df	= 972
Total sample size	= 975
Actual power	= 0.8004415

One can define an effect size in terms of η^2 . For example, if $\eta^2 = 10\%$, then

$$\phi' = \sqrt{\frac{\eta^2}{1-\eta^2}} = \sqrt{\frac{.10}{1-.10}} = .33.$$

Suppose I had 6 subjects in each of four groups. If I employed an alpha-criterion of .05, how large [in terms of % variance in the DV accounted for by variance in the IV] would the effect need be for me to have a 90% chance of rejecting the null hypothesis? From the table, for $df_t = 3$, $df_e = 20$, $\phi = 2.0$ for $\beta = .13$, and $\phi = 2.2$ for $\beta = .07$. By linear interpolation, for $\beta = .10$, $\phi = 2.0 + (3/6)(.2) = 2.1$. $\phi' = \frac{\phi}{\sqrt{n}} = \frac{2.1}{\sqrt{6}} = 0.857$.

$\eta^2 = \phi'^2 / (1 + \phi'^2) = .857^2 / (1 + .857^2) = 0.42$, a very large effect!

Do note that this method of power analysis does not ignore the effect of error df , as did the methods employed in Chapter 8. If you were doing small sample power analyses by hand for independent t -tests, you should use the methods shown here (with $k = 2$), which will give the correct power figures (since $t = \sqrt{F}$, t 's power must be the same as F 's).

For a nice primer on power for ANOVA, see [Lakens](#).

APA-Style Summary Statement

Teaching method significantly affected the students' test scores, $F(3, 16) = 86.66$, $MSE = 0.50$, $p < .001$, $\eta^2 = .942$, 95% CI [.858, .956]. As shown in Table 1,

SAS Code to Do the Analysis

```
data Cliett;
input Group; Do I=1 to 5; Input Score @@; output; end; cards;
1
  1 2 2 2 3
2
  2 3 3 3 4
3
  6 7 7 7 8
4
  7 8 8 8 9
proc GLM; class Group;
  model Score = Group / ssl EFFECTSIZE alpha=0.1; means Group / regwq; run;
```

This code will do the analysis. The output will include a confidence intervals for the proportion of variance explained and the noncentrality parameter, and pairwise comparisons.

References

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591–601. doi: 10.1037/0003-066X.63.7.591

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics - Simulation and Computation*, 15, 933-943. doi: 10.1080/03610918608812553

Copyright 2017, [Karl L. Wuensch](#) - All rights reserved.