

Analysis of Variance

(Chapter 10 in Zar, 2010)

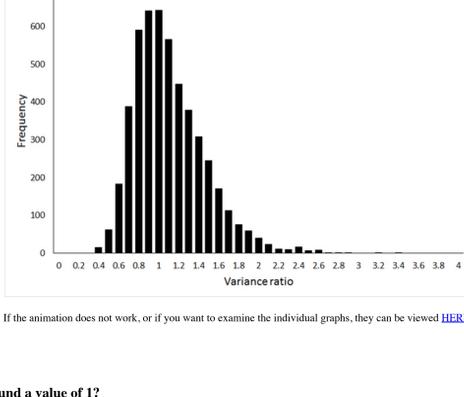
We now have a tool (the 2-sample t -test) that will allow us to compare two sample means. More often than not, however, we design experiments that require the comparison of more than two groups. While it might seem logical that we could use the t -test multiple times to accomplish such a comparison (e.g., compare among A, B, and C by comparing A to B, A to C, and B to C), this creates a serious issue with the type I error rate, α . Every time you conduct a comparison, you have a 1 in 20 chance of committing a type I error. Multiple analyses on a single data set compounds that error rate. In the example mentioned previously, the 3 comparisons would result in a type I error rate for the overall comparison of 0.15.

No problem...we'll just divide α by 3 to make our comparison, right? Actually, that creates a different problem. Decreasing α for each individual test increases the type II error rate for each of the individual comparisons. While there are corrections that operate on the principle of dividing the type I error rate, such as the Bonferroni correction, it is far better to utilize a single analysis to compare 2 or more means that is covered by a single type I error rate, such as the venerable analysis of variance (ANOVA).

The probability distribution that ANOVA relies upon is called the F -distribution (after Sir Ronald Fisher). The F -distribution is a distribution of variance ratios drawn from a single statistical population. In other words, each observation in the distribution is the result of drawing 2 samples from a single statistical population, calculating the variance (s^2) of each sample, and dividing one by the other:

$$F = \frac{S_1^2}{S_2^2}$$

For the example below, the 5000 observations (the ratio of 2 sample variances is one observation, so 10,000 samples were drawn in total) were drawn from a normally distributed population with $\mu = 10$ and $\sigma = 2$ (the R program can be viewed [HERE](#)):



If the animation does not work, or if you want to examine the individual graphs, they can be viewed [HERE](#).

Question 1: Why is the F -distribution centered around a value of 1?

You can see that the sample sizes used for the samples from which the variances (and thus the resulting variance ratios) were calculated influence the shape of the distribution. In this example, the same sample size was used for both the numerator and denominator variances, but (fortunately, as you will see) this is not a requirement. Unlike the previous theoretical distributions that we have employed (the t -distribution and the χ^2 -distribution), the shape of the F -distribution is influenced by both numerator and denominator degrees of freedom. In this example, these correspond to $n_1 - 1$ for the numerator degrees of freedom and $n_2 - 1$ for the denominator degrees of freedom.

Examine Table B.4 in your textbook (starting on p. 680 in the 5th Edition), which contains the critical values for the F -distribution. The first column of the table contains the denominator degrees of freedom (v_2), and there is a separate table for each value of the numerator degrees of freedom (v_1). It probably has occurred to you that this might have some utility in testing assumptions of homogeneity of variance, which is true, but unfortunately the assumptions required to apply the F -distribution to that question make that approach of limited utility, and so we will stick with Bartlett's test when it is important, and Hartley's F_{max} test for the examples you will analyze in this class. No, I am not implying that this class isn't important...

If you have been nodding and smiling up until this point, you have overlooked a very important paradox. The goal of ANOVA is the comparison of sample means, but it relies on the distribution of a variance ratio. Knowing that our most recently applied test for comparing means, the 2-sample t -test, requires the assumption of homogeneity of variance should indicate to us that it is quite possible for statistical populations with different means to have the same variance. So how is it possible to compare means with a variance ratio?

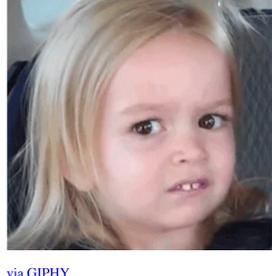
Examining the problem a different way, how do we test the null hypothesis:

$$\bar{Y}_1 = \bar{Y}_2 = \dots = \bar{Y}_a = \mu$$

Using a distribution that is the null expectation for:

$$S_1^2 = S_2^2 = \sigma^2$$

???



via GIPHY

The solution to the paradox involves breaking down the value of an observation into its various determinants. Every observation, Y belongs to a statistical population with a mean of μ , even though we likely have no idea what that population mean might be. That observation deviates from that mean as the result of individual variation, which we will denote as ϵ . If we apply some fixed treatment to an experimental unit, that treatment should add additional variation, which we will denote as α . For example, if you are shorter than the average height for the statistical population to which you belong, one basis for this could be genetic, and so (all else being equal) your height would be expected to deviate from the average by ϵ . However, this deviation could be altered by having smoked a lot of cigarettes and drank a lot of coffee in kindergarten. That alteration would be expressed as α .

Note that this is a different use for the same symbol that we used to denote type I error rate. They are not directly related. Later on, for regression analysis, we will use the same symbol to denote the Y-intercept. This means that you are going to have to pay careful attention to context!

If the influence of μ and the variance components, ϵ and α are additive, i.e., the combined effect is simply the sum of those effects, we can consider the value of an observation to be:

$$Y_{ij} = \mu + \epsilon_{ij} + \alpha_i$$

Where Y_{ij} is the j th observation in group i , and deviates from μ (the mean of the population from which all individuals in all groups were drawn) by the individual variance component ϵ_{ij} . The variation added by the treatment to which the group was exposed is denoted as α_i .

Variation added by random effects (treatment effects are considered fixed effects) is treated the same way, except that it is denoted as A_i :

$$Y_{ij} = \mu + \epsilon_{ij} + A_i$$

It is the variance components ϵ and α that are used to create a variance ratio that will be compared to the null expectation of the F -distribution as:

$$F_s = \frac{\epsilon_{ij} + \alpha_i}{\epsilon_{ij}}$$

The individual variation is included in both the numerator and denominator, but the treatment variation (or random effect variation) is included only in the numerator. If there is no variation added due to treatment effects, i.e., if $\alpha = 0$, then the numerator and denominator estimate the same thing, and the ratio should fit the null expectation. So, even though the null hypothesis we report for ANOVA is:

$$\bar{Y}_1 = \bar{Y}_2 = \dots = \bar{Y}_a = \mu$$

The null hypothesis that we really are testing is:

$$H_0: \alpha = 0$$

The good news is that this makes ANOVA a one-tailed test, regardless of the circumstances. The bad news is that you now have to explain why...

Question 2: Explain why the application of the variance ratio, F_s , as described above, is always a one-tailed test.

ANOVA with a fixed independent variable ($H_0: \alpha = 0$) is referred to as model I ANOVA, and ANOVA with a random independent variable ($H_0: A = 0$) is referred to as model II ANOVA. There is no difference between the 2 in terms of how F_s is calculated, only in how we structure the null hypothesis. The distinction between model I and model II ANOVA will, however, take on special significance next week when we address n-factor ANOVA, where there is more than 1 independent variable.

Speaking of F_s , how do we go about calculating it such that the numerator estimates $\epsilon + \alpha$, and the denominator only estimates ϵ ? Let's start with the denominator. The differences between the means of distributions of treatment groups should be the result of α , but individual variation (ϵ) should not vary among the groups if the experimental design has no bias. In other words, treatment effects should shift the central tendency, but not the dispersion around that central tendency. Thus, we should get a good estimate of ϵ within each group, and a better estimate by averaging among the groups. We can refer to this as **within group variation** because the components of the estimates come from within the groups.

Often, however, our sample sizes are unequal, and so we need a weighted average variance, because larger sample sizes produce better estimates. To do this, we could multiply each variance estimate by the sample size used to generate it, and divide the result by the sum of the sample sizes.

We are going to take a different approach that accomplishes the same thing. The reason for this is that it will allow us to construct an **ANOVA table** step by step, which is how most software packages display the results of ANOVA. We will approach the calculation of average within-group variance by first calculating the **within-group sum of squares**, also referred to as **error sum of squares** in your textbook. Within-group sum of squares is simply the sum of $\sum y_i^2$ across all groups:

$$SS_{within} = \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^a \sum y_i^2$$

Where a is the number of groups, i is a specific group (from 1 to a), and j is an individual observation, such that Y_{ij} is the j th observation in group i . It might look complicated, but it is a simple matter of calculating the sum of the squared deviates ($\sum y_i^2$) for each group and summing them together.

The denominator of F_s is referred to as **within-group mean squares** (MS_{within}) or **error mean squares** (we shall use the former), and is calculated by dividing SS_{within} by the appropriate degrees of freedom (the sum of the degrees of freedom for each group):

$$df_{within} = \sum_{i=1}^a (n_i - 1) = \left(\sum_{i=1}^a n_i \right) - a$$

Again, the calculation is simple: the equation on the left indicates subtracting 1 from each of the sample sizes and summing them together, and the one on the right indicates that the sample sizes are summed, and the number of groups being compared be subtracted from that sum. MS_{within} then, is simply:

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

Or, if you prefer the busier versions:

$$MS_{within} = \frac{\sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{\sum_{i=1}^a (n_i - 1)} = \frac{\sum_{i=1}^a \sum y_i^2}{\left(\sum_{i=1}^a n_i \right) - a}$$

This gives us the denominator for F_s , which you may recall is our estimate of ϵ . For the numerator, which estimates $\epsilon + \alpha$ (where α is variance added due to treatment in a model I ANOVA, not type I error rate), or $\epsilon + A$ for model II ANOVA, we need an independent estimate of variance that includes variance added due to treatment. We already have used the individual observations for MS_{within} , and so employing them again would not give us an independent estimate. Using the sample means, however, would give us an independent measure, and one that included treatment effects (ϵ) or random effects (A) if these effects contributed to the differences among the means. The first step is to calculate a **grand mean**, which we will denote as a \bar{Y} with 2 bars over it (the use of a subscript in your textbook to distinguish sample means from grand means seems likely to lead to confusion):

$$\bar{\bar{Y}} = \frac{\sum Y_{ij}}{\sum n_i}$$

As the name implies, this is the overall mean, determined by summing all of the observations across all of the groups, and dividing by the total number of observations. It may occur to you that you also could take the average of the sample means, which is true, **but only when the sample sizes are all equal**. To get our estimate of variance using the means, we calculate the **among-group sum of squares** (SS_{among}) as the squared deviations of sample means from the grand mean (essentially treating the sample means as observations), multiplying each deviation by the sample size used to generate the sample mean:

$$SS_{among} = \sum n_i (\bar{Y}_i - \bar{\bar{Y}})^2$$

Because there are a sample means, the degrees of freedom (df_{among}) for the **among-group mean squares** (MS_{among}) are $a - 1$, and so the numerator of F_s is calculated as:

$$MS_{among} = \frac{SS_{among}}{df_{among}} = \frac{\sum n_i (\bar{Y}_i - \bar{\bar{Y}})^2}{a - 1}$$

Thus, our variance ratio F_s is calculated as:

$$F_s = \frac{MS_{among}}{MS_{within}}$$

And the resultant value compared to F_{crit} at $\alpha(1) = 0.05$, at df_{among} numerator degrees of freedom (v_1 in Table B.4) and df_{within} denominator degrees of freedom (v_2 in Table B.4). Remember that because the other α , variation added due to treatment, or A (variation added due to random effects) always is added to the numerator only, ANOVA is always a one-tailed test. **It is important to note that rejecting the null hypothesis will not tell us which means differ from each other; it will only tell us that there is a difference in at least one of the means.** Finding which means differ from each other involves analyses referred to as **multiple comparisons**, which we will get to in due time.

As mentioned previously, the calculations are the same for both model I (fixed effects) and model II (random effects) ANOVA, and so we can apply the above calculations to our examples without worrying about distinguishing between the two for the time being.

In order that you may become familiar with the presentation of an ANOVA table, **you will be presenting your results in this format for the remaining questions**. The following table serves the dual purpose of showing a, a buried marble covered by a log, a buried truffle, and a buried truffle covered by a log.

Source of Variation	df	SS	MS	F_s
$(\bar{Y} - \bar{Y})$ among groups	$a - 1$	$\sum n_i (\bar{Y}_i - \bar{\bar{Y}})^2$	$\frac{SS_{among}}{df_{among}}$	$\frac{MS_{among}}{MS_{within}}$
$(Y - \bar{Y})$ within groups (error)	$\left(\sum n \right) - a$	$\sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$\frac{SS_{within}}{df_{within}}$	
$(Y - \bar{Y})$ total	$\left(\sum n \right) - 1$	$SS_{among} + SS_{within}$		

Of course, you will be filling in the values, for df , SS , MS , and F_s , and not recopying the formulas...

Download this week's Excel workbook [HERE](#).

The first worksheet ("Example10.2"), not surprisingly, contains the data from Example 10.2 (page 201 in the 5th edition). Work it through to make certain that you understand the calculations. **Do not use the formulas that they use in the example!** Use the formulas that have been provided above. The methods in the text are shortcuts that hide the actual basis for the calculations. You should, however, arrive at the same result, which will tell you whether you have applied our formulas correctly or not, which is the whole point of starting with an example from the text.

Question 3: Present the ANOVA table for the data in Example 10.2.

The data in the second worksheet ("squirrel") come from an experiment testing whether northern flying squirrels hunt by scent. The time spent digging was recorded for individual squirrels in pens with a buried marble, a buried marble covered by a log, a buried truffle, and a buried truffle covered by a log.

Question 4: Present the ANOVA table for the squirrel data. Is there a significant difference among treatments?

The third worksheet contains data from a gut content analysis of alpine newts, comparing the number of *Daphnia longispina* found in each digestive tract for sample of newts collected in June, August, and October of the same year.

Question 5: Present the ANOVA table for the newt data. Is there a significant difference in the mean number of *D. longispina* among the months?

Not surprisingly, ANOVA is a parametric test, which means that we need to address [the assumptions of ANOVA](#).