Topic 3: Fundamentals of analysis of variance

> "The analysis of variance is more than a technique for statistical analysis. Once it is understood, ANOVA is a tool that can provide an insight into the nature of variation of natural events"
> Sokal & Rohlf (1995), BIOMETRY.

3.1. The F distribution [ST&D p. 99]

Assume that you are sampling at random from a normally distributed population (or from two different populations with equal variance) by first sampling n_1 items and calculating their variance s^2_1 (df: n_1 - 1), followed by sampling n_2 items and calculating their variance s^2_2 (df: n_2 - 1). Now consider the ratio of these two sample variances:

$$\frac{s_1^2}{s_2^2}$$

This ratio will be close to 1, because these variances are estimates of the same quantity. The expected distribution of this statistic is called the **F-distribution.** The F-distribution is determined by **two** values for degrees of freedom, one for each sample variance. The values found within statistical Tables for F (e.g. Table A6) represent $F_{\alpha[df1,df2]}$ where α is the proportion of the F-distribution to the right of the given F-value and df_1, df_2 are the degrees of freedom pertaining to the numerator and denominator of the variance ratio, respectively.

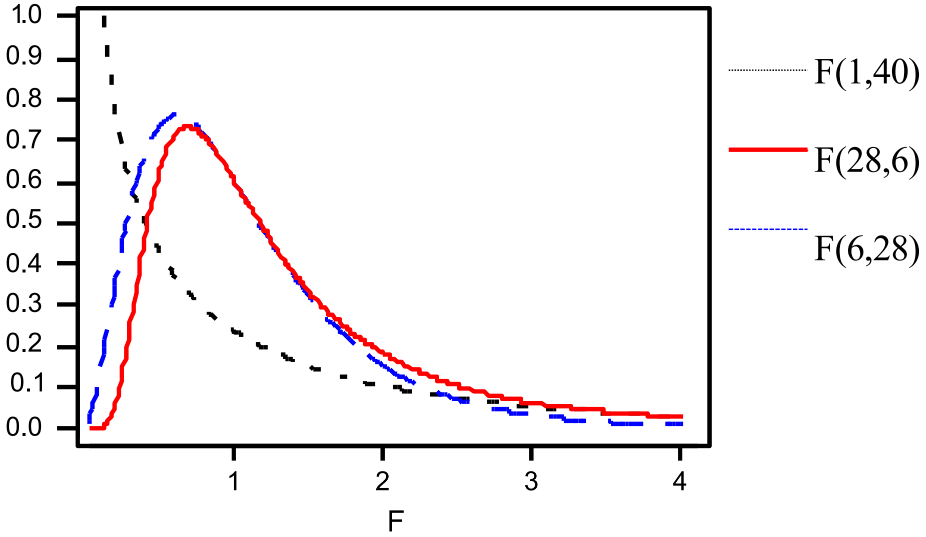

Figure 1 Three representative F-distributions (note similarity of $F_{(1,40)}$ to χ^2_1).

For example, a value $F_{\alpha/2=0.025,\ df1=9,\ df2=\ 9]}$ = 4.03 indicates that the ratio s^2_1 / s^2_2 , from samples of ten individuals from normally distributed populations with equal variances, is expected to be larger than 4.03 *by chance* in only **5%** of the experiments (the alternative hypothesis is $s^2_1 \neq s^2_2$ so it is a **two tailed test**).

3.2 Testing the hypothesis of equality of variances [ST&D 116-118]

Suppose X_1 ,..., X_m are observations drawn from a normal distribution with mean μ_X and variance σ_X^2; and Y_1, ..., Y_n are drawn from a normal distribution with mean μ_v and variance σ_Y^2. In theory, the F statistic can be used as a test for the hypothesis H_0: $\sigma_X^2 = \sigma_Y^2$ vs. the hypothesis H_1: $\sigma_X^2 \neq \sigma_Y^2$. H_0 is rejected at the α level of significance if the ratio $s_X^2 = s_Y^2$ is either $\geq F_{\alpha/2,\, dfX-1,\, dfY-1}$ *or* $\leq F_{1-\alpha/2,\, dfX-1,\, dfY-1}$. In practice, this test is rarely used because it is **very sensitive to departures from normality.**

3.3 Testing the hypothesis of equality of two means [ST&D 98-112]

The ratio between two estimates of σ^2 can also be used to test differences between means; that is, it can be used to test H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$. In particular:

$$F = \frac{\text{estimate of } \sigma^2 \text{ from sample means}}{\text{estimate of } \sigma^2 \text{ from individuals}}$$

The denominator is an estimate of σ^2 provided by the individuals *within* each sample. That is, it is a ***weighted average*** of the sample variances.

The numerator is an estimate of σ^2 provided by the means *among* samples. The variance of a population of sample means is σ^2/n, where σ^2 is the variance of individuals in a parent population and all samples are of size n. This implies that means may be used to estimate σ^2 by multiplying the variance of sample means σ^2/n by n.

$$F = \frac{s^2_{among}}{s^2_{within}} = \frac{ns^2_{\bar{Y}}}{s^2}$$

When the two populations have different means (but the same variance), the estimate of σ^2 based on sample means will include a contribution attributable to the difference among population means as well as any random difference (i.e. within-population variance). Thus, in general, if the means differ, the sample means are expected to be more variable than predicted by chance alone.

Example: We will explain the test using a data set of Little and Hills (p. 31). Yields (100 lbs/acre) of wheat varieties 1 and 2 from plots to which the varieties were randomly assigned:

Varieties	Replications					$Y_{i.}$	$\bar{Y}_{i.}$	s^2_i
1	19	14	15	17	20	85	$\bar{Y}_{1.}= 17$	6.5
2	23	19	19	21	18	100	$\bar{Y}_{2.}= 20$	4.0
						$Y_{..}= 185$	$\bar{Y}_{..}= 18.5$	

In this experiment, there are two treatment levels (t = 2) and five replications (r = 5) (the symbol t stands for "treatments" and r stands for "replications"). Each observation in the experiment has a unique "address" given by Y_{ij}, where i is the index for treatment (i = 1,2) and j is the index for replication (j = 1,2,3,4,5). Thus $Y_{24} = 19$.

The dot notation is a shorthand alternative to using $\sum$. Summation is for all values of the subscript occupied by the dot. Thus $Y_{1.} = 19 + 14 + 15 + 17 + 20$ and $Y_{.2} = 14 + 19$.

We begin by *assuming* that the two populations have the same (unknown) variance σ^2 and then test H_0: $\mu_1 = \mu_2$. We do this by obtaining two estimates for σ^2 and comparing them.

First, we can compute the average variance of individuals ***within samples***, also known as the *experimental error*. To determine the experimental error, we compute the variance of each sample (s^2_1 and s^2_2), assume they both estimate a common variance, and then estimate that common variance by pooling the two estimates:

$$s_1^2 = \frac{\sum_j (Y_{1j} - \overline{Y}_{1.})^2}{n_1 - 1} \quad , \quad s_2^2 = \frac{\sum_j (Y_{2j} - \overline{Y}_{2.})^2}{n_2 - 1}$$

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = 4*6.5 + 4*\ 4.0 / (4 + 4) = 5.25 \equiv s^2_{within}$$

In this case, since r1 = r2, the pooled variance is simply the average of the two sample variances. Since pooling s^2_1 and s^2_2 gives an estimate of σ^2 based on the variability within samples, let's designate it s^2_w (subscript w = within).

The second estimate of σ^2 is based on the variation *between* or *among* samples. Assuming, by the null hypothesis, that these two samples are random samples drawn from the *same population* and that, therefore, $\overline{Y}_{1.}$ and $\overline{Y}_{2.}$ both estimate the same population mean, we can estimate the variance of means of that population by $s^2_{\overline{Y}}$. Recall from Topic 1 that the mean $\overline{Y}$ of a set of n random variables drawn from a normal distribution with mean μ and variance σ^2 is itself a normally distributed random variable with mean μ and variance σ^2/n.

The formula for $s^2_{\overline{Y}}$ is

$$s^2_{\overline{Y}} = \frac{\sum_{i=1}^{t} (\overline{Y}_{i.} - \overline{Y}_{..})^2}{t - 1} = [(17 - 18.5)^2 + (20 - 18.5)^2] / (2-1) = 4.5$$

and, from the central limit theorem, n times this quantity provides an estimate for σ^2 (n is the number of variates on which each sample mean is based).

Therefore, the ***between samples*** estimate is:

$$\mathbf{n}\, s_{\bar{Y}}^{2} = 5 * 4.5 = 22.5 \equiv s_{between}^{2}$$

These two variances are used in the F test as follows. If the null hypothesis is not true, we would expect the variance between samples to be much larger than the variance within samples ("much larger" means larger than one would expect by chance alone). Therefore, we look at the ratio of these variances and ask whether this ratio is significantly greater than 1. It turns out that under our assumptions (normality, equal variance, etc.), this ratio is distributed according to an $F_{(t-1,\ t(n-1))}$ distribution. That is, we define:

$$\boldsymbol{F} = \mathbf{s_b}^{\mathbf{2}} / \mathbf{s_w}^{\mathbf{2}}$$

and test whether this statistic is significantly greater than 1. The F statistic is a measure of how many times larger the variability **between** the samples is compared to the variability **within** samples. In this example, *F* = 22.5/5.25 = 4.29. The numerator $s_{\mathbf{b}}^2$ is based on 1 df, since there are only two sample means. The denominator, $s_{\mathbf{w}}^2$, is based on pooling the df within each sample so df_{den} = t(n-1) = 2(4) = 8. For these df, we would expect an F value of 4.29 or larger just by chance about 7% of the time. From Table A.6, $F_{0.05,\ 1,\ 8}$ = 5.32. Since 4.29 < 5.32, we fail to reject H_0 at the 0.05 significance level.

3.3.1 Relationship between F and t

In the case of only two treatments, the square-root of the F statistic is distributed according to a t distribution:

$$F_{1-\alpha, df=1, t(n-1)} = t^{2}_{1-\frac{\alpha}{2}, df=t(n-1)}$$

$$\text{meaning } t = \sqrt{\frac{s_b^2}{s_w^2}}$$

In the example above, with 5 reps per treatment:

$$F_{(1,8),\ \mathbf{1-\alpha}} = (t_{5,\ \mathbf{1-\alpha/2}})^2$$

$$5.32 = 2.306^2$$

The total degrees of freedom for the t statistic is t(n - 1) since there are nt total observations and they must satisfy t constraint equations, one for each treatment mean. Therefore, we reject the null hypothesis at the α significance level if $t > t_{a/2,t(n-1)}$.

Here are the computations for our data set:

$$t = \sqrt{\frac{s_b^2}{s_w^2}} = \sqrt{\frac{22.5}{5.25}} = 2.07$$

Since 2.07 < $t_{0.025,\,8}$ = 2.306, we fail to reject H_0 at the 0.05 significance level.

3.4 The linear additive model [ST&D p. 32, 103, 152]

3.4.1 One population: In statistics, a common model describing the makeup of an observation states that it consists of a mean plus an error. This is a linear additive model. A minimum assumption is that the errors are random, making the model probabilistic rather than deterministic.

The simplest linear additive model is this one:

$$\mathbf{Y_i = \mu + \varepsilon_i}$$

It is applicable to the problem of estimating or making inferences about population means and variances. This model attempts to explain an observation Y_i as a mean $\boldsymbol{\mu}$ plus a random element of variation $\boldsymbol{\varepsilon_i}$. The $\boldsymbol{\varepsilon}_i$'s are assumed to be from a population of uncorrelated $\boldsymbol{\varepsilon}$'s with mean zero. Independence among $\boldsymbol{\varepsilon}$'s is assured by random sampling.

3.4.2 Two populations: Now consider this model:

$$\mathbf{Y_{ij} = \mu + \tau_i + \varepsilon_{ij}}$$

It is more general than the previous model because it permits us to describe two populations simultaneously. For samples from **two** populations with possibly different means but a common variance, any given reading is composed of the grand mean μ of the population, a component τ_i for the population involved (i.e. $\mu + \tau_1 = \mu_1$ and $\mu + \tau_2 = \mu_2$), and a random deviation ε_{ij}. The subindex i (= 1,2) indicates the treatment number and the subindex j (= 1, ..., r) indicates the number of observations from each population (replications).

τ_i, the treatment effects, are measured as deviations from the overall mean [$\mu = (\mu_1 + \mu_2) / 2$] such that $\boldsymbol{\tau_1 + \tau_2 = 0}$ or $\boldsymbol{-\tau_1 = \tau_2}$. This does not affect the difference between means, 2τ. If $r_1 \neq r_2$ we may set $r_1\tau_1 + r_2\tau_2 = 0$.

The ε's are assumed to be from a single population with normal distribution, mean $\mu = 0$, and variance s^2.

Another way to express this model, using the dot notation from before, is:

$$Y_{ij} = \overline{Y}_{..} + (\overline{Y}_{i.} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{i.})$$

3.4.3 More than two populations. One-way classification ANOVA

As with the 2 sample t-test, the linear model is:

$$\mathbf{Y_{ij} = \mu + \tau_i + \varepsilon_{ij}}$$

where now i = 1,...,t and j = 1,...,r. Again, the ε_{ij} are assumed to be drawn from a normal distribution with mean 0 and variance σ^2. Two different kinds of assumptions can be made about the τ's that will differentiate the **Model I ANOVA** from the **Model II ANOVA**.

***The Model I ANOVA or fixed model*:** In this model, the τ's are fixed and

$$\mathbf{\sum \tau_i = 0}$$

The constraint $\sum \tau_i = 0$ is a consequence of defining treatment effects as deviations from an overall mean. The null hypothesis is then stated as H_0: $\tau_1 = ... = \tau_i = 0$ and the alternative as H_1: at least one $\tau_i \neq 0$. What a Model I ANOVA tests is the **differential** effects of treatments that are **fixed** and determined by the experimenter. The word "fixed" refers to the fact that each treatment is assumed to always have the same effect τ_i. Moreover, the set of τ's are assumed to constitute a finite population and are specific parameters of interest, along with s^2. In the case of a false H_0 (i.e. some $\tau_i \neq 0$), there will be an additional component of variation due to treatment effects equal to:

$$r\sum\frac{\tau_i^2}{t-1}$$

Since the τ_i are measured as deviations from a mean, this quantity is analogous to a variance but cannot be called such since it is not based on a random variable but rather on deliberately chosen treatments.

***The Model II ANOVA or random model*:** In this model, the additive effects for each group (τ's) are not fixed treatments but are *random* effects. In this case, we have not deliberately planned or fixed the treatment for any group, and the effects on each group are random and only partly under our control. The τ's themselves are *a random sample* from a population of τ's for which the mean is zero and the variance is σ^2_t. When the null hypothesis is false, there will be an additional component of variance equal to $r\sigma^2_t$. Since the effects are random, it is futile to estimate the magnitude of these random effects for any one group or the differences from group to group; but we can estimate their variance, the added variance component among groups: $r\sigma^2_t$. We test for its presence and estimate its magnitude, as well as its percentage contribution to the variation. In the fixed model, we draw inferences about *particular* treatments; in the random model, we draw an inference about the *population* of treatments. The null hypothesis in this latter case is stated as H_0: $\sigma^2_t = 0$ versus H_1: $\sigma^2_t \neq 0$.

An important point is that the basic setup of data, as well as the computation and significance test, in most cases is the same for both models. It is the *purpose* which differs between the two models, as do some of the supplementary tests and computations following the initial significance test. For now, we will deal only with the **fixed model.**

Assumptions of the model [ST&D 174]

1. Treatment and environmental effects are additive
2. Experimental errors are random, possess a common variance, and are independently and normally distributed about zero mean

Effects are additive

This means that all effects in the model (treatment effects, random error) cause deviations from the overall mean in an additive manner (rather than, for example, multiplicative).

Error terms are independently and normally distributed

This means there is no correlation between experimental groupings of observations (e.g. by treatment level) and the sizes of the error terms. This could be violated if, for example, treatments are not assigned randomly.

This assumption essentially means that the means and variances of treatments share no correlation. For example, suppose yield is measured and the treatments cause yield to range from 1 g/plant up to 10 g/plant. A range of ± 1 gm would be much more "significant" at the low end than the high end but could not be considered any differently by this model.

Variances are homogeneous

The means the variances of the different treatment groups are the same.

3.5 ANOVA: Single factor designs

3.5.1 The Completely Randomized Design (CRD)

In single factor experiments, a single treatment (i.e. factor) is varied to form the different treatment levels. The experiment discussed below is taken from page 141 of ST&D. The experiment involves inoculating five different cultures of one legume, clover, with strains of nitrogen-fixing bacteria from another legume, alfalfa. As a sort of control, a sixth trial was run in which a composite of five clover cultures was inoculated. There are 6 treatments (t = 6) and each treatment is given 5 replications (r = 5).

	3DOK1	3DOK5	3DOK4	3DOK7	3DOK13	composite	Total
	19.4	17.7	17.0	20.7	14.3	17.3	
	32.6	24.8	19.4	21.0	14.4	19.4	
	27.0	27.9	9.1	20.5	11.8	19.1	
	32.1	25.2	11.9	18.8	11.6	16.9	
	33.0	24.3	15.8	18.6	14.2	20.8	
$\sum Y_{ij} = Y_{i.}$	144.1	119.9	73.2	99.6	66.3	93.5	$596.6 = Y_{..}$
$\sum Y_{ij}^2$	4287.53	2932.27	1139.42	1989.14	887.29	1758.71	12994.36
$Y_{i.}^2/r$	4152.96	2875.2	1071.65	1984.03	879.14	1748.45	12711.43
$\sum(Y_{ij} - \overline{Y}_{i.})^2$	134.57	57.07	67.77	5.11	8.15	10.26	282.93
$\overline{Y}_{i.}$ = mean	28.8	24.0	14.6	19.9	13.3	18.7	19.88
σ^2_{n-1} variance	33.64	14.27	16.94	1.28	2.04	2.56	

Inoculation of clover with *Rhizobium* strains [ST&D Table 7.1]

The completely randomized design (CRD) is the basic ANOVA design. It is used when there are t different treatment levels of a single factor (in this case, *Rhizobium* strain). These treatments are applied to t independent random samples of size r. Let the total sample size for the experiment be designated as n = rt. Let Y_{ij} denote the jth measurement (replication) recorded from the ith treatment. WARNING: Some texts interchange the i and the j (i.e. the rows and columns of the table), so be careful.

We wish to test the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_t$ against H_1: not all the μ_i's are equal. This is a straightforward extension of the two-sample t test of topic 3.3 since there was nothing special about the value t = 2. Recall that the test statistic was:

$$\boldsymbol{F = s_b^2 / s_w^2}$$

In our new dot notation, we can write:

$$s_w^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r} (Y_{ij} - \overline{Y}_{i.})^2}{t(r-1)} = \frac{SSE}{t(r-1)}, \text{ where } SSE \equiv \sum_{i=1}^{t}\sum_{j=1}^{r} (Y_{ij} - \overline{Y}_{i.})^2$$

Here SSE is the ***sum of squares for error***. Also:

$$s_b^2 = \frac{r\sum_{i=1}^{t}(\overline{Y}_{i.} - \overline{Y}_{..})^2}{t-1} = \frac{SST}{t-1}, \text{ where SST= } r\sum_{i=1}^{t}(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

Here SST is the ***sum of squares for treatment***.

Since the variance among treatment means estimates σ^2/r, the r in the definition formula for SST is required so that the ***mean square for treatment*** (MST) will be an estimate of σ^2 rather than σ^2/r. This is equivalent to the step we took in example 3.3 above when we multiplied by **n** in order to estimate the variance ***between samples*** (s_b^2 = **n** $s_{\overline{Y}}^2$).

Using our new notation, we can write:

$$F = \frac{SST/(t-1)}{SSE/t(r-1)} = \frac{SST/(t-1)}{SSE/(n-1)}$$

We can then define:

The mean square for error: MSE = SSE/(n-t). This is the average dispersion of the observations around their respective group means. It is an estimate of a common σ^2, the experimental error (i.e. the variation among observations treated alike). MSE is a valid estimate of the common σ^2 ***if*** the assumption of equal variances among treatments is true.

The mean square for treatment: MST = SST/(t-1). (MS Model in SAS) This is an independent estimate of σ^2, when the null hypothesis is true (H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_t$). If there are differences among treatment means, there will be an additional source of variation in the experiment due to treatment effects equal to $r\sum\tau_i^2/(t-1)$ (Model I) or $r\sigma^2_t$ (Model II) (see topic 3.4.3 and ST&D 155).

$$\boldsymbol{F = \text{MST/MSE}}$$

The *F* value is obtained by dividing the treatment mean square by the error mean square. We expect to find *F* approximately equal to 1. In fact, however, the expected ratio is:

$$\frac{MST}{MSE} = \frac{\sigma^2 + r\sum\tau_i^2/(t-1)}{\sigma^2}$$

As is clear from this formula, the *F*-test is sensitive to the presence of the added component of variation due to treatment effects. In other words, ANOVA permits us to test whether there are any nonzero treatment effects. That is, to test whether a group of means can be considered random samples from the same population or whether we have sufficient evidence to conclude that the treatments that have affected each group separately have resulted in shifting these means sufficiently so that they can no longer be considered samples from the same population.

Recall that the degrees of freedom is the number of
independent, unconstrained quantities underlying a statistic.

Underlying SST are t quantities ($\overline{Y}_{i.}$ - $\overline{Y}_{..}$) which have one constraint (they must sum to 0); so df_{trt} = t-1. Underlying SSE are n quantities Y_{ij}, which have t constraints for the t sample means; so df_e = n-t.

Consider the following equation:

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\overline{Y}_{..})^2 = r\sum_{i=1}^{t}(\overline{Y}_{i.}-\overline{Y}_{..})^2+\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\overline{Y}_{i.})^2$$

If you take the time to deal with the messy algebra, you will find that this equality is true. The reason this is relevant is because this is just our dot notation for:

TSS = SST + SSE

where TSS is the ***total sum of squares*** of the experiment. In other words, sums of squares are perfectly additive. If you were to fully expand the quantity on the left-hand side of the equation, you find lots of cross product terms of the form $2(\overline{Y}_{ij}\ \overline{Y}_{..})$. It turns out, upon simplification, that all of those cross product terms cancel each other out (note that none appear on the right-hand side of the equation). Quantities that satisfy this criterion are said to be ***orthogonal***. Another way of saying this is that we can decompose the total SS into a portion due to variation among groups and another, completely independent portion due to variation within groups. The degrees of freedom are also additive (i.e. $df_{Tot} = df_{Trt} + df_e$).

The dot notation above provides the "definitional" forms of these quantities (TSS, SST, and SSE). But each also has a friendlier "calculational" form, for when you compute them by hand. The two expressions are mathematically equivalent. The calculational forms make use of a quantity called the "correction factor":

$$C = \frac{1}{n}(Y_{..})^2 = \frac{1}{n}(\sum_{ij} Y_{ij})^2$$

This term is just the squared sum of all observations in the experiment, divided by their total number. Once you calculate C, you can tackle the **total SS (TSS)**:

$$TSS = \left(\sum_{i=1}^{t} \sum_{j=1}^{r} Y_{ij}^{\ 2} \right) - C$$

The total SS is the sum of squares that includes all sources of variation. In the dot notation, you see that it is the sum of the squares of the deviation of each observation from the overall mean.

Next, you can tackle the **treatment sum of squares (SST)**:

$$SST = \frac{1}{r}\left(\sum_{i=1}^{t} Y_{i.}^{\ 2} \right) - C$$

The SST is the sum of squares attributable to the variable of classification. This is the SS due to differences among treatment groups and is referred to as the within-or-among groups SS.

Finally, the **error sum of squares (SSE)**:

$$SSE = TSS - SST$$

The SSE is that part of the total sum of squares that cannot be explained by difference among groups. It is the sum of squares among individuals treated alike. It is referred to as the within groups SS, residual SS, or error SS.

An ANOVA table provides a systematic presentation of everything we've covered until now. The first column of the ANOVA table specifies the components of the linear model. In a single factor CRD, remember, the linear model is just:

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_{ij}$$

By this model, we have two named sources of variation: Treatments and Error. The next column indicates the df associated with each of these components. Next is a column with the SS associated with each, followed by a column with the mean squares associated with each. Mean squares, mathematically, are essentially variances; and they are found by dividing SS by their respective df. Finally, the last column in an ANOVA table present the F statistic, which is a ratio of mean squares (i.e. a ratio of variances).

An ANOVA table (including an additional column of the SS definitional forms):

Source	df	Definition	SS	MS	F
Treatments	t - 1	$r\sum_{i}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	SST	SST/(t-1)	MST/MSE
Error	t(r-1) = n - t	$\sum_{i,j}(Y_{ij} - \overline{Y}_{i.})^2$	TSS - SST	SSE/(n-t)	
Total	n - 1	$\sum_{i,j}(Y_{ij} - \overline{Y}_{..})^2$	TSS		

The ANOVA table for our Rhizobium experiment would look like this:

Source	df	SS	MS	F
Treatment	5	847.05	169.41	14.37**
Error	24	282.93	11.79	
Total	29	1129.98		

Notice that the mean square error (MSE = 11.79) is just the pooled variance or the average of variances within each treatment (i.e. MSE = $\Sigma\ s_i^2 / t$; where s_i^2 is the variance estimated from the ith treatment). The *F* value of 14 indicates that the variation among treatments is over 14 times larger than the mean variation within treatments. This value far exceeds the critical F value for such an experimental design at $\alpha = 0.05$ ($F_{crit} = F_{(5,24),0.05} = 2.62$), so we reject H_0. At least one of the treatments has a nonzero effect on the response variable, at the specified significance level.

3.5.1.2 Assumptions associated with ANOVA

The assumptions associated with ANOVA can be expressed in terms of the following statistical model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

First, ε_{ij} (the residuals) are assumed to be **normally distributed** with mean 0 and possess a **common variance** σ^2, independent of treatment level i *and* sample number j).

3.5.1.2.1 Normal distribution

Recall from the first lecture that the Shapiro-Wilk test statistic W provides a powerful test for normality for small to medium samples (n < 2000).

For large populations (n>2000), the use of the Kolmogorov-Smirnov statistic is recommended.

3.5.1.2.2 Homogeneity of treatment variances

Tests for homogeneity of variance (i.e. homoscedasticity) attempt to determine if the variance is the same within each of the groups defined by the independent variable. Bartlett's test (ST&D 481) can be very inaccurate if the underlying distribution is even slightly nonnormal, and it is not recommended for routine use. Levene's test is much more robust to deviations from normality.

Levene’s test is an ANOVA of the absolute values of the residuals of the observation from their respective treatment means. If Levene's test leads you reject the null hypothesis (H_0: the mean residual absolute value is the same for all treatments; i.e. the within-treatment variance is the same across all treatment groups; i.e. variances are homogeneous), one option is to perform a Welch's variance-weighted ANOVA (Biometrika 1951 v38, 330) instead of the usual ANOVA to test for differences between group means in a CRD. This alternative to the usual analysis of variance is more robust if variances are not equal.

3.5.1.3 Experimental Procedure: Randomization

Here is how the clover plots might look if this experiment were conducted in the field:

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30

The experimental procedure would be: First, randomly (e.g. from a random number table, etc.) select the plot numbers to be assigned to the six treatments (A,B,C,D,E,F). **Example**: On p. 607 [ST&D], starting from Row 02, columns 88-89 (a random starting point), move downward. Take for treatment A the first 5 random numbers under 30, and so forth: Treatment A: 05, 19, 13, 20, 6; Treatment B: 14, 26, 1, 8, 4; etc.

3.5.1.4 Power and sample size

3.5.1.4.1 Power

The power of a test is the probability of detecting a nonzero treatment effect. To calculate the power of the F test in an ANOVA using Pearson and Hartley's power function charts (1953, Biometrika 38:112-130), it is necessary first to calculate a critical value ϕ. This critical value depends on the number of treatments (t), the number of replications (r), the magnitude of the treatment effects that the investigator wishes to detect (d), an estimate of the population variance (σ^2 = MSE), and the probability of rejecting a true null hypothesis (α).

In a CRD, $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$,

where i = 1,2,…,t; j = 1,2,...,r; $\boldsymbol{\mu}$ is the overall mean; and $\boldsymbol{\tau_i}$ is the treatment effect ($\tau_i = \mu_i - \mu$). To calculate the power, you first need to calculate ϕ, a standardized measure (in σ units) of the expected differences among means which can be used to determine sample size from the power charts. Its general form:

$$\phi = \sqrt{\frac{r}{MSE}\sum\frac{\tau_i^2}{t}}$$

With this value, you enter the chart for $\nu_1 = df_1 = df_{numerator} = df_{treatment}$ = t-1 and choose the x-axis scale for the appropriate α (0.05 or 0.01). The interception of the calculated ϕ with the curve for $\nu_2 = df_2 = df_{denominator} = df_{error}$ = t(n-1) gives the power of the test (the y-axis on both sides of the chart).

Example: Suppose an experiment has t = 6 treatments with r = 2 replications each. Given the MSE and the required $\alpha = 5\%$, you calculate $\phi = 1.75$. To find the power associated with this value of ϕ, use Chart v_1= t-1 = 5 and the set of curves to the left ($\alpha = 5\%$). Select curve v_2 = t(r-1) = 6. The height of this curve corresponding to the abscissa of $\phi = 1.75$ is the power of the test. In this case, the power is slightly greater than 0.55. As a rule of thumb, experiments should be designed with a power of at least 80% (i.e. $\beta \leq 0.20$).

3. 5. 1. 4. 2. Sample size

To calculate the number of replications required for a given α and desired power, a simplification of the general power formula above can be used. The general power formula can be **simplified** if we assume all $\boldsymbol{\tau_i}$ are zero except the two extreme treatment effects (let's call them τ_K and τ_L , so that **d** = $|\mu_K - \mu_L|$. You can think of **d** as the difference between the extreme treatment means. Taking μ to be in the middle of μ_K and μ_L , $\boldsymbol{\tau_i}$ = **d/2**:

$$\sum \frac{\tau_i^2}{t} = \frac{(d/2)^2 + (d/2)^2}{t} = \frac{d^2/4 + d^2/4}{t} = \frac{d^2/2}{t} = \frac{d^2}{2t}$$

And the $\boldsymbol{\phi}$ formula simplifies:

$$\phi = \sqrt{\frac{d^2 * r}{2t * MSE}}$$

With this simplified expression, one can estimate the required number of replications for a given α and desired power by: 1) Specifying the constants, 2) Starting with an arbitrary r to compute ϕ, 3) Using the appropriate Pearson and Hartley chart to find the power; and 4) Iterating the process until a minimum r value is found which satisfies the required power for a given α level.

Example: Suppose that 6 treatments will be involved in a study and the anticipated difference between the extreme means is 15 units. What is the required sample size so that this difference will be detected at $\alpha = 1\%$ and power = 90%, knowing that $\sigma^2 = 12$? (note, t = 6, $\alpha = 0.01$, $\beta = 0.10$, d = 15, and MSE = 12).

r	df	ϕ	(1-β) for α=1%
2	6(2-1)= 6	1.77	0.22
3	6(3-1)= 12	2.17	0.71
4	6(4-1)= 18	2.50	0.93

Thus 4 replications are required for each treatment to satisfy the required conditions.