Topic 3. Single factor ANOVA: Introduction [ST&D Chapter 7]

> "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. Statistical Tables for *F* (e.g. A6 in your book) show the cumulative probability distribution of *F* for several selected probability values. The values in the table represent $F_{\alpha[\nu 1,\, \nu 2]}$ where α is the proportion of the F-distribution to the right of the given- F-value (in one tail) and ν_1, ν_2 are the degrees of freedom pertaining to the numerator and denominator of the variance ratio, respectively.

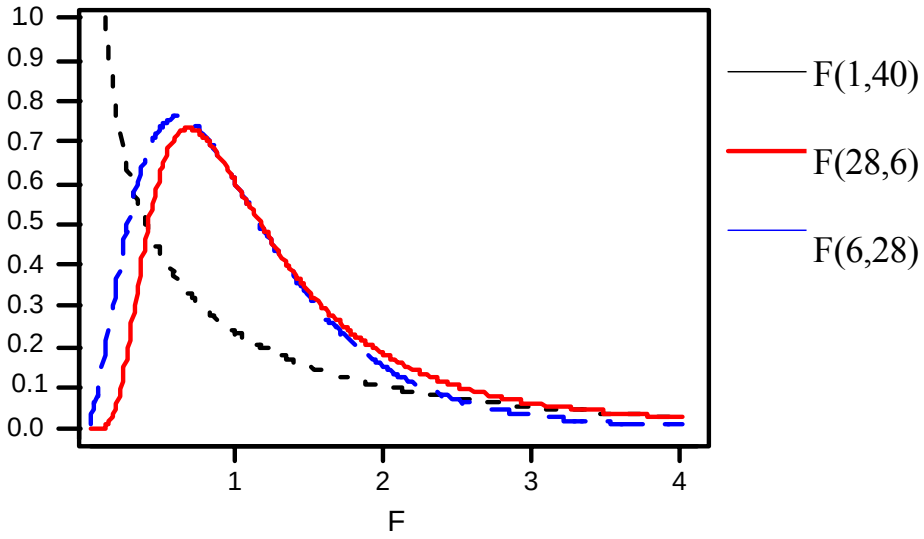

Figure 1 Three representative F-distributions (note similarity of $F_{(1,40)}$ to χ^2_1).

For example, a value $F_{\alpha/2=0.025,\, \nu 1=9,\, \nu 2=9]} = 4.03$ indicates that the ratio s^2_1 / s^2_2, from samples of ten individuals from normally distributed populations with equal variance, 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 tail 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.

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.** This can be calculated using SAS **PROC TTEST.**

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

The ratio between two estimates of σ^2 can be used to test differences between means, that is, a test of 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 from 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 variation *among* sample means. To obtain this estimate of σ^2 (variance among individuals) from the mean variance ($\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$), $\sigma^2_{\bar{Y}}$ needs to be multiplied by n.

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

When the two populations have different means (but same variance), the estimate of σ^2 based on sample means will include a contribution attributable to the difference between population means as well as any random difference (i.e. within-population variance). Thus, if there is a significant difference among means, the sample means are expected to be more variable than when chance alone operates and there are no significant differences among means.

Nomenclature: to be consistent with the book in the ANOVA we will use r for the number of replication and n for the total number of experimental units in the experiment

Example: We will explain the test using a data set of Little and Hills (p. 31).
Table 1. Yields (100 lb./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).

Each observation in the experiment has a unique "address" given by Y_{ij}, where i is the index for treatment (i = 1 or 2) and j is the index for replication (j = 1, 2, 3, 4, or 5). Thus $Y_{2,4} = 19$.

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 $Y_{.2} = 14 + 19$.

For this exercise, we will *assume* 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 sample variance σ^2 ***within samples***. To determine this variability called *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 samplevariances:

$$s_1^2 = \frac{\sum_j (Y_{1j} - \bar{Y}_{1.})^2}{r_1 - 1} \quad , \quad s_2^2 = \frac{\sum_j (Y_{2j} - \bar{Y}_{2.})^2}{r_2 - 1}$$

$$s^2_{pooled} = \frac{(r_1 - 1)s_1^2 + (r_2 - 1)s_2^2}{(r_1 - 1) + (r_2 - 1)} = 4*6.5 + 4*\ 4.0\ /\ (4 + 4) = 5.25 \equiv s^2_{within}$$

In this case, since $n_1 = n_2$, 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, we will designate it $\mathbf{s^2_w}$ (subscript w= **within**).

The second estimate of the sample variance σ^2 is based on the variation ***between*** samples ($\mathbf{s^2_b}$). Assuming that these two samples are random samples drawn from the *same population* and that, therefore, $\bar{Y}_{1.}$ and $\bar{Y}_{2.}$ both estimate the same population mean, we estimate the variance of means using $s^2_{\bar{Y}}$. Recall from Topic 1 that the mean $\bar{Y}$ of a set of r 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/r.

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

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

and, from the central limit theorem, n times $s^2_{\bar{Y}}$ 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:

$$s^2_b = \mathbf{r}\, s^2_{\bar{Y}} = 5 * 4.5 = 22.5$$

These two variances are used in the F test as follows. If the null hypothesis is not true, then the **between** samples variance should be much larger than the **within** samples variance ("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 $F_{(t\text{-}1,\ t(r\text{-}1))}$ distribution. That is, we define:

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

and test whether this statistic is significantly greater than 1. The ***F*** statistics measures how many times larger is the variability **between** the samples compared with the variability **within** samples.

In this example, ***F*** = 22.5/5.25 = 4.29. The numerator s_b^2 is based on 1 df, since there are two sample means. The denominator, s_w^2, is based on pooling the df within each sample so df sw^2= t(r-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 (p.614 of ST&D), $F_{0.05,\ 1,\ 8} = 5.32$. 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:

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

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

In the example above, with 5 reps per treatment:

$$F_{(1,8),\ \mathbf{1}-\alpha} = (t_{5,\ \mathbf{1}-\alpha/2})^2$$ (be careful: *F* uses α and *t* α/2)

The total degrees of freedom for the t statistic is t(**r** - 1) = t**r**-t = **n**-t since there are n 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 $t > t_{\alpha/2,\ t(r-1)}$.

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. The value 2.306 is obtained from Table A3 (p.611) and df= 2(5-1)= 8. Note that $2.306^2 = 5.32 = F_{0.05,\ 1,\ 8}$.

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:

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

This model 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 **μ** plus a random element of variation $\boldsymbol{\varepsilon_i}$. The $\boldsymbol{\varepsilon_i}$'s are assumed to be from a population of **uncorrelated ε**'s with **mean zero**. Independence among **ε**'s is assured by random sampling.

3.4.2. Two populations:

This second model is more general than the previous model (3.4.1) because it permits us to describe two populations simultaneously:

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

For samples from **two** populations with possibly different means but a **common variance**, any given observation 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}.

As before, the ε's are assumed to be from a single population with normal distribution, mean μ= 0, and variance σ^2. The subindex i (= 1, 2) indicates the treatment number and the subindex j (= 1, ..., n) indicates the number of observations from each population (replications).

The τ_I are also referred as the treatment effects, measured as a deviation from each treatment means from an overall mean for the complete experiment. This overall mean is set as a middle reference point as:

$$\mu = (\mu_1 + \mu_2) / 2, \text{ which is estimated by } \bar{Y}_{..} = (\bar{Y}_1 + \bar{Y}_2)/2$$

Therefore $\boldsymbol{\tau_1 + \tau_2 = 0}$ or in a different way $\boldsymbol{-\tau_1 = \tau_2}$ (the difference between means is 2*|τ|).

If $\mathbf{r_1 \neq r_2}$ we set $\mathbf{r_1\tau_1 + r_2\tau_2 = 0}$.

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

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

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 $\boldsymbol{\varepsilon_{ij}}$ 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 $\boldsymbol{\tau}$'s that will differentiate the **Model I ANOVA** from the **Model II ANOVA**.

***The Model I ANOVA or fixed model*:** In this model, the $\boldsymbol{\tau}$'s are fixed and

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

Setting the $\sum \boldsymbol{\tau}_i = 0$ is simply measuring treatment effects as deviations from an overall mean from the experiment. The null hypothesis is then

H_0: $\boldsymbol{\tau_1} = ... = \boldsymbol{\tau_i} = 0$

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. The τ's are assumed to constitute a finite population and are the parameters of interest, along with s^2. When the null hypothesis is false (and 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 added 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 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. However, we can estimate their variance, the added variance component among groups: σ^2_t. We test for its **presence** and estimate its **magnitude**, as well as its **percentage contribution to the variation** (calculated in SAS with PROC VARCOMP). The null hypothesis in the random model 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. The ***purpose* differs** between the two models, as do some of the supplementary tests and computations following the initial significance test. In the **fixed model**, we draw inferences about ***particular* treatments**; in the **random model**, we draw an inference about the ***population* of treatments**.

Until Topic 10, we will deal only with the **fixed model.**

Assumptions of the model [ST&D p.174]

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

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.

Variances are homogeneous

This assumption means that the variances of the different treatment groups are the same. This assumption means that the means and variances of treatments share no correlation, that is, that treatments with larger means do not have larger variances.

We need this assumption since we are calculating an overall sample variance, by averaging the variances of the different treatments.

There are alternative statistical analyses when the variances are not homogeneous (e.g. Welch's variance-weighted one-way ANOVA)

3.5. ANOVA: Single factor designs

3.5.1. The Completely Random Design CRD

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

Table 1. Inoculation of clover with *Rhizobium* strains [ST&D Table 7.1]

	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
$\sum(Y_{ij} - \bar{Y}_{i.})^2$	134.57	57.07	67.77	5.11	8.15	10.26	282.93
$\bar{Y}_{i.}$ = 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	

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 n. Treatments can have different number of replications, but the formulas become much more complicated and we will postpone them till later in the course.

Let the total sample size for the experiment be designated as **n** = **r** t. Let Y_{ij} denote the j^{th} measurement (replication) recorded from the i^{th} treatment. WARNING: some texts interchange the i and the j (i.e. the rows and columns of the table).

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 dot notation, we can write this as:

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

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

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

Here SST is the ***sum of squares for treatment*** (SAS refers to this as the Model SS).

Since the variance among treatment means estimates σ^2/**r**, the multiplication for **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**. In the example 3.3 above we also multiplied by **r** in order to estimate ***between samples*** variances ($s_b^2 = r\, s_{\bar{Y}}^2$).

In this notation we can write (remember rt=n):

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

We can then define:

The mean square for error: MSE = SSE/(**n**-t)) gives the average dispersion of the items around their respective group means. The df is t times (**r**-1), which is the df within each of the pooled treatments.

MSE is an estimate of a common σ^2, the experimental error (= within variation or 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 (because we are averaging the variance estimates from different treatments).

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 added component of variation due to treatment effects equal to $\mathbf{r}\sum\tau_i^2/(t-1)$ (Model I) or $\mathbf{r}\sigma^2_t$ (Model II) (see topic 3.4.3 and ST&D 155). The multiplication by **r** is required to express the variance per individual, not per mean (remember that $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{r}$ so $\sigma^2 = r\sigma^2_{\bar{Y}}$).

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

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:

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

It is clear from this formula, that the *F*-test is sensitive to the presence of the added component of variation due to treatment effects. In other words, the ANOVA permits us to test whether there are any added 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 number of degrees of freedom is the number of independent quantities in the statistic.

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

We can also use the following equation:

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

or

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 expand the quantity on the left-hand side of the above equation in our dot notation, there is a cross product terms of the form $2(\bar{Y}_{ij}\ \bar{Y}_{..})$ that should appear. It turns out, that all of these cross product terms cancel each other out. Quantities that satisfy this 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 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 "definition" formulas for of these quantities (TSS, SST, and SSE). But each also has a friendlier "calculation" form to compute them by hand.

The actual calculations, when done by hand, use the formulas

$C = (Y_{..})^2 / n = (\sum_{ij} Y_{ij})^2 / n$	The correction term (C). Is the squared sum of all observations divided by their number.
$TSS = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^{\ 2} - C$	The total sum of squares that includes all sources of variation. This is the total SS.
$SST = \sum_{i=1}^{t} Y_{i.}^{\ 2} / r - C$	The sum of squares attributable to the variable of classification. This is the between SS, or among groups SS or treatment SS.
$SSE = TSS - SST$	The sum of squares among individuals treated alike. This is the within groups SS, or residual SS or error SS. It is easier to calculate as a difference

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. 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 corresponding mean squares.

Mean squares are essentially variances; and they are found by dividing SS by their respective df. Finally, the last column in an ANOVA table below presents the ***F*** statistic, which is a ratio of mean squares (i.e. a ratio of variances). Usually, a last column is added indicating the probability of finding that F values by chance.
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
Among cultures	5	847.05	169.41	14.37**
Within cultures	24	282.93	11.79	
Total	29	1129.98		

Notice that the MSE (11.79) is 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 average 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 and conclude that **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).

Note that the ANOVA requires that the residual errors have a normal distribution and not that the complete population of original values have a normal distribution. If there is a treatment effect, mixing the different treatments will result in a distribution with multiple peaks. Therefore, the treatment effects are first subtracted, the residuals ε_{ij} are calculated and then their normality is tested.

3.5.1.2.1. Normal distribution

Recall from the first lecture that the Shapiro-Wilk test statistic W (ST&D 567; produced by SAS via Proc UNIVARIATE NORMAL) provides a powerful test for normality for small to medium samples (n < 2000). Normality is rejected if W is sufficiently smaller than 1. W is similar to a correlation between the data and their normal scores (ST&D 566). In a perfectly normal population there is a perfect correlation W=1.

For large populations (n>2000), SAS recommends the use of the Kolmogorov-Smirnov statistics (ST&D 571; also produced by SAS via Proc UNIVARIATE NORMAL or via Analyst).

Both tests are applied to the residuals of the model, which are easy to calculate in SAS or R.

3.5.1.2.2. Homogeneity of 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 more robust to deviations from normality, and will be used in this class.

Levene's test is an ANOVA of the squares of the residuals of each observation from its treatment mean. An alternative form of the test, implemented in R, uses the absolute values of the deviation from the treatment median.

To perform Levene's test in SAS, you need to use the option HOVTEST (for Homogeneity of variance test) within the means statement in the PROC GLM procedure:

```
proc GLM;
        Class Treatment;
        Model Response = Treatment;
        Means Treatment / Hovtest = Levene;
```

If Levene's test rejects the hypothesis of homogeneity of variances there are three alternatives:

1. Transform the data (e.g. logarithm) so that the transformed values have uniform variances.
2. Use a non parametric statistical test.
3. Use the WELCH option which produces a **Welch's variance-weighted ANOVA** (Biometrika 1951 v38, 330) instead of the usual ANOVA. This alternative to the usual analysis of variance is more robust if variances are not equal.

```
proc GLM;
        Class Treatment;
        Model Response = Treatment;
        Means Treatment / Welch;
```

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 such as ST&D 606, or using PROC PLAN in SAS, etc.) select the plot numbers to be assigned to the six treatments (A, B, C, D, E, F). **Example**: On p. 607, 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 (without replication): Treatment A: 5, 19, 13, 20, 6; Treatment B: 14, 26, 1, 8, 4; etc. Or simply write 30 numbers, mixed and randomly assigned 5 numbers to each treatment…

B	2	3	**B**	**A**	**A**
7	**B**	9	10	11	12
A	**B**	15	16	17	18
A	**A**	21	22	23	24
25	**B**	27	28	29	30

3.5.1.4. Power and sample size

Pearson and Hartley (1953, Biometrika 38:112-130) provided power function charts that are easy to use to calculate the power of an ANOVA and the appropriate number of replications. The Tables are available in the class website at
http://www.plantsciences.ucdavis.edu/agr205/Lectures/2010%20Iago/Topic%203/PowerCharts.pdf

There are different charts for each different numerator degrees of freedom ν_1.

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, it is necessary to calculate first a critical value ϕ. This critical value depends on the number of treatments (t), the number of replications (n), 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,** $\mathbf{y_{ij} = \mu + \tau_i + \varepsilon_{ij}}$,

- **i** = 1,2,…,t; **j** = 1,2,...,r;
- **μ** is the overall mean;
- $\boldsymbol{\tau_i}$ 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. The exact formula is:

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

This general formula can be simplified using an **approximation** that assumes 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 approximate ϕ formula simplifies:

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

First selecting the chart for $\nu_1 = df_1 = df_{numerator} = df_{treatment}$ = t-1 and then choose the x-axis scale for the appropriate α (0.05 or 0.01). Draw an imaginary vertical line at the calculated ϕ and look for the interception with the curve for $\nu_2 = df_2 = df_{denominator} = df_{error}$ = t(r-1). The corresponding value at the y-axis gives the power of the test.

Example: Suppose that one experiment had t = 6 treatments with **r** = 2 replications each. The difference between the extreme means was 10 units, MSE= 5.46, and the required α = 5%. To calculate the power using the approximate formula:

$$\phi = \sqrt{\frac{d^2 * r}{2t * MSE}} = \sqrt{\frac{10^2 * 2}{2(6) * 5.46}} = 1.75$$

Use Chart ν_1 = t-1 = 5 and the set of curves to the left (α = 5%). Select curve ν_2 = t(n-1) = 6. The height of this curve corresponding to the abscissa of ϕ = 1.75 is the power of the test. In this

case, the power is slightly greater than 0.55. Experiments should be designed to have a power of at least 80% (i.e. $\beta \leq 0.20$).

To calculate the power using *Analyst:* Statistics →ANOVA → One-Way ANOVA → Tests → Power analysis. Or *Analyst* → Sample Size → One-Way ANOVA → Complete the number of treatments, the corrected sum of squares CSS (= SST = between SS = among groups SS = treatment SS), and the standard deviation, which is the square root of the mean squared error (MSE). You must also specify the significance level of the test; the default is 0.05.

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

To calculate the number of replications for a given α and desired power:

- Specify the constants.
- Start with an arbitrary **r** to compute ϕ.
- Use the appropriate Pearson and Hartley chart to find the power.
- Iterate the process until a minimum **r** value which satisfies the required power for a given α level is found.

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, $t = 6$, $\alpha = 1\%$, $\beta = 10\%$, $d = 15$, and MSE = $\sigma^2 = 12$).

r	df	ϕ	$(1-\beta)$ 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.

3.5.2. Subsampling: the nested design [ST&D p. 157 - 167]

It may happen that the experimenter wishes to make several observations within each ***experimental unit***, the unit to which the treatment is applied. Such observations are made on subsamples or ***sampling units***.

The classical example of this is given in Steel and Torrie: sampling individual plants within pots where the pots are the experimental units randomly assigned to treatments. Other examples would be individual trees within an orchard plot (where the treatment is assigned to the plot), individual sheep within a herd (where the treatment is assigned to the herd), etc. We call the analysis of this kind of data organized in a hierarchical way ***nested analysis of variance.*** Nested ANOVAs are not limited to two hierarchical levels (e.g. pots, and then plants within pots). We can divide the subgroups into sub-subgroups, and even further, as long as the sampling units

within each level (e.g. pots, then plants within pots, then flowers within plants, etc.) are chosen randomly.

The essential objective of a nested ANOVAs is to dissect the MSE of a system into its components, thereby ascertaining the sources and magnitudes of error in an experiment or process.

Examples of applications of nested ANOVA are:

- To ascertain the magnitude of error at various stages of an experiment or process.
- To estimate the magnitude of the variance attributable to various levels of variation in a study of quantitative genetics
- To discover sources of variation in natural population in systematic studies, etc.

If you are confused in a nested design, **you can always average the subsamples and perform a simpler ANOVA.** This is also a good strategy to test if your nested design analysis is correct. The final *P* value will be the same with a correct nested design analysis and a non-nested design using the averages of the subsamples. The advantage of doing the more complex analysis including the subsamples is to calculate the different component of variance.

3.5.2.1. Linear model for subsampling

Before we compute a nested ANOVA, we should examine the linear model upon which it is based:

$$Y_{ijk} = \mu + \tau_i + \varepsilon_{j(i)} + \delta_{k(ij)}$$

The interpretations of μ, τ, and ε are as before. But now *two* random elements are obtained with each observation:

The $\boldsymbol{\varepsilon_{j(i)}}$ are assumed normal with mean 0 and variance σ_ε^2. The subscript $\varepsilon_{j(i)}$ indicates that the j^{th} level of replication is nested within the i^{th} level of treatment. Note: this is a different notation from ST&D. The $\varepsilon_{j(i)}$ measures, as before, the variation among real replications within treatment groups, and the subindex indicates that there are subsamples within the replications. In the experiment where pots are randomized among treatments, and each pot includes 4 plants, $\varepsilon_{j(i)}$ measures the variation among pots means (averages of 4 plants) within a treatment.

The $\boldsymbol{\delta_{k(ij)}}$ represents the errors associated with the variation among subsamples within an experimental unit. In the pot experiment $\delta_{k(ij)}$ measures the variation among the 4 plants within each pot. The $\delta_{k(ij)}$ are also assumed normal with mean 0 and variance σ^2. This is represented in the sample data as:

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

Remember that in this notation the dot replaces a subscript and indicates that all values covered by that subscript have been added

Applying this formula to the pot experiment:

τ_i measures the difference between a treatment mean and the overall mean
$\varepsilon_{j(i)}$ measures the difference between a pot mean and the mean of its assigned treatment.
$\delta_{k(ij)}$ measures the difference between a plant and the mean of its pot.

3.5.2.2. Nested ANOVA with equal subsample numbers: computation

In this experiment, mint plants are exposed to six different combinations of temperature and daylight and stem growth was measured at 1 week. The 6 treatments are assigned randomly across 18 pots (i.e. 3 replications per treatment combination). Within each pot are four plants (i.e. subsamples).

Sometimes we may be uncertain as to whether a factor is crossed or nested. If the levels of the factor are just for identification (is not a classification criteria) and can be renumbered arbitrarily without affecting the analysis, then the factor is nested.

For example, pots 1, 2, 3 within treatment level 1 could be relabeled 2, 3, 1 without causing any problems. That is because pot number is simply an ID, not a classification variable. Pot 1 in treatment 1 has nothing to do with Pot 1 in treatment 2.

The data (from ST&D page 159):

Treatment	Low T, 8 hs			Low T, 12 hs			Low T, 16 hs			High T, 8 hs			High T, 12 hs			High T, 16 hs		
	Pot number			Pot number			Pot number			Pot number			Pot number			Pot number		
Plant N_o	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	3.5	2.5	3.0	5.0	3.5	4.5	5.0	5.5	5.5	8.5	6.5	7.0	6.0	6.0	6.5	7.0	6.0	11.0
2	4.0	4.5	3.0	5.5	3.5	4.0	4.5	6.0	4.5	6.0	7.0	7.0	5.5	8.5	6.5	9.0	7.0	7.0
3	3.0	5.5	2.5	4.0	3.0	4.0	5.0	5.0	6.5	9.0	8.0	7.0	3.5	4.5	8.5	8.5	7.0	9.0
4	4.5	5.0	3.0	3.5	4.0	5.0	4.5	5.0	5.5	8.5	6.5	7.0	7.0	7.5	7.5	8.5	7.0	8.0
Pot totals = $Y_{ii.}$	15	17.5	11.5	18	14	17.5	19	21.5	22	32	28	28	22	26.5	29	33	27	35
Trt. totals = $Y_{i..}$	44.0			49.5			62.5			88.0			77.5			95.0		
Trt. means= $\bar{Y}_{i..}$	3.7			4.1			5.2			7.3			6.5			7.9		

In this example, t = 6, r = 3, s = number of subsamples = 4, and n = trs = 72.

Recall that for a CRD the sum of squares satisfies:

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{..})^2 = r\sum_{i=1}^{t}(\bar{Y}_{i.}-\bar{Y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i.})^2 \text{ or } \mathbf{TSS = SST + SSE}.$$

The degrees of freedom associated with these sums of square are n-1, t-1, and n-t, respectively. In the nested design, TSS and SST are unchanged but the SSE is partitioned into two components. The resulting equation can be written as:

$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{...})^2 = rs\sum_{i=1}^{t}(\bar{Y}_{i..}-\bar{Y}_{...})^2 + s\sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{Y}_{ij.}-\bar{Y}_{i..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{ij.})^2$$

or: **TSS = SST + SSEE + SSSE**

The two error terms represent the sum of squares due to ***experimental error*** and the sum of squares due to ***sampling error***. In the pot experiment, **SSEE** represents the variation among pots within treatments and **SSSE** represents the variation among plants within pots.

Nested ANOVA table:

Source of variation	df	SS	MS	F	Expected MS
Treatments (τ_i)	t - 1 = 5	SST	SST / 5	**MST / MSEE**	$\sigma_\delta^2+\mathbf{4}\sigma_\varepsilon^2+\mathbf{12}\Sigma\tau^2/5$
Exp. Error ($\varepsilon_{j(i)}$)	t (r - 1) = 12	SSEE	SSEE / 12	MSEE / **MSSE**	$\sigma_\delta^2+\mathbf{4}\sigma_\varepsilon^2$
Samp. Error ($\delta_{k(ij)}$)	nt (s - 1) = 54	SSSE	SSSE / 54		σ_δ^2
Total	tns - 1 = 71	TSS			

In each case, the number of degrees of freedom is the product of the number of levels associated with each subscript between brackets and the number of levels minus one associated with the subscript outside the brackets.

The ***expected mean squares*** are the theoretical models of the variance components included in each MSE. The MSSE estimates σ_δ^2(variation among plants), and the MSEE estimates both the variation between plants (σ_δ^2) and the variation between pots (σ_ε^2). The last one is multiplied by 4 because pots are **means** of 4 plants ($\sigma_\varepsilon^2= \sigma_\delta^2/4$) and to put everything in the same scale (σ_δ^2) it needs to be multiplied. The treatment effects are based on treatment means calculated from 12 plants (4*3), and that is why it is multiplied by 12.

The most important part of this table:

> In testing a hypothesis about treatment means, the appropriate divisor for *F* is the mean square experimental error (**MSEE**) since it includes the variation from *all sources* (pot and plant) that contribute to the variability of treatment means except the treatment effects themselves.

If you do not inform the statistical program the plants are subsamples, the program will automatically divide by the **MSSE**, and the *P* value will answer the question:

Is there a significant difference between <u>treatments or pots</u>? MST/MSSE->EMS= $\mathbf{4}\sigma_\varepsilon^2+\mathbf{12}\Sigma\tau^2/5$

instead of the one you thing you are answering which is:

Is there a different between <u>treatments</u>? MST/**MSEE**->EMS= $\mathbf{12}\Sigma\tau^2/5$

In a nested design the most critical part is the selection of the correct error term

Estimation of the different variance components in the pot experiment

The main objective in a nested design is to estimate the variance components. To do this, we deconstruct the calculated mean squares according to their underlying theoretical models or ***expected mean squares*** (EMS, last column in the table) for each component of the linear model, as shown below:

Variance Source	df	Sum of Squares	Mean Squares	Variance component	Percent of total
Total	71	255.91	3.60	4.05	100.0 %
trtmt	5	179.64	**35.92**	2.81	69.4 %
pot	12	25.83	**2.15**	0.30	7.5 %
plant	54	40.43	**0.93**	0.93	23.0 %

MSSE = σ_δ^2 , so σ_δ^2= **0.93**

MSEE = σ_δ^2+**4**σ_ε^2 , so σ_ε^2= (MSEE - σ_δ^2)/**4** = (**2.15** - 0.93)/**4** = 0.30

MST = σ_δ^2+4σ_ε^2+**12**$\Sigma\tau^2$/5 , so $\Sigma\tau^2$/5= (MST - MSEE)/**12** = (**35.92** - 2.15)/12 = 2.81

In this example, the variation among plants within a pot is three times larger than the variation among pots within a treatment.

In SAS, **PROC VARCOMP** computes these variance components for different models. For our example experiment here:

```
Proc GLM;
      Class Trtmt Pot;
      Model Growth = Trtmt Pot(Trtmt);
      Random Pot(Trtmt);
      Test h = Trtmt e = Pot(Trtmt);
Proc Varcomp;
      Class Trtmt Pot;
      Model Growth = Trtmt Pot(Trtmt);
```

Pot(Trtmt): indicates that pot is a nested factor in treatment. Pot 1 in treatment 1 is not more similar to pot 1 in treatment 2 than to pots 2 and 3.

Random Pot(Trtmt): This statement tells SAS that the pots are a random factor (i.e. pots 1, 2 and 3 are just a random sample, not a classification based on a common property).

Test h = Trtmt **e** = Pot(Trtmt): This statement tells SAS which error term to use to test a particular hypothesis. For the hypothesis about treatments (**h** = Trtmt), the appropriate error term is the MSEE (i.e. Pot(trtmt)), so **e** = Pot(Trtmt). This specifies the test **MST/MSEE**.

Note that you do not include a class variable for the last level of sub-sampling (in this case, plant). By default, SAS will use this last level of variation (among plants within a pot) as the error term for the experiment. This is why the **test** statement is so important in a nested design:

SAS's default error term (MSSE) is correct for testing difference among pots within a treatment (MSEE), but **not** for testing the differences among treatments (MST).

If you have **two levels of nesting** (e.g. you measure two leaves from each plant), then you include plant as a class variable (but not leaf) and you indicate that a plant is nested in (pot trtmt).

```
Proc GLM;
      Class Trtmt Pot Plant;
      Model Growth = Trtmt Pot(trtmt) Plant(Pot Trtmt);
      Random Pot(Trtmt) Plant(Pot Trtmt);
      Test h = Trtmt e = Pot(Trtmt);

Proc Varcomp;
      Class Trtmt Pot Plant;
      Model Growth = Trtmt Pot(trtmt) Plant(Pot Trtmt);
```

3.5.2.3. Optimal allocation of resources

Additional information in Biometry Sokal & Rohlf page 309.

One of the main reasons to use a nested design is to investigate how the variation is distributed among experimental units and among subsamples (i.e. where are the sources of error in the experiment). Once the variance component of the experimental units ($s^2_{e.u.}$) and the variance component of the subsamples (s^2_{sub}) are known, the optimum number of samples and subsamples can be calculated using the relative cost of experimental units and subsamples and the formulas below:

To introduce the idea of cost, we write a cost function. For a two-level nested design, the total cost (C) will be the cost of the subsamples multiplied by the total number of subsamples (N_{sub}) plus the cost of each experimental unit multiplied by the number of experimental units (N_{eu}):

$$C = N_{sub} * N_{eu}(C_{sub}) + N_{eu}(C_{eu})$$

To find the number of subsamples (N_{sub}) per experimental unit that will result in simultaneous **minimal cost** and **minimal variance**, the following formula may be used:

$$N_{sub} = \sqrt{\frac{C_{e.u.} * s^2_{sub}}{C_{sub} * s^2_{e.u.}}}$$

The optimum the number of subsamples will increase when the experimental units are more expensive than the subsamples, and when the subsamples are more variable than the experimental units.

If the cost of samples and subsamples is the same, the optimum number of subsamples in our example can be calculated as:

$$N_{sub} = \sqrt{\frac{s^2_{sub}}{s^2_{e.u.}}} = \sqrt{\frac{0.93}{0.30}} = 1.76 \text{ or } \approx 2 \text{ plants per pot}$$

In the case of equal costs the number of subsamples is the square root of the ratio between subsample and sample variance. If the subsamples are 4 times more variable, the optimum is two subsamples.

If the cost is the same and $s_{sub} < s_{e.u.}$, it is better to allocate all the resources to experimental units (in this example, would be to use one plant per pot). Therefore, subsampling is only useful when the variation among subsamples is larger than the variation among experimental units and/or the cost of the subsamples is smaller than the cost of the experimental units.