Topic 3: Fundamentals of analysis of variance (continued)
Subsampling, nesting, and components of variance

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 called subsamples. 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 a ***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 are chosen randomly (e.g. pots, then plants within pots, then flowers within plants, then anthers within flowers, etc. etc.).

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. One example of this objective would be to discover and characterize sources of variation in systematic studies of natural populations.

3.5.2.1 Linear model for subsampling

Before we perform a nested ANOVA, let us 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 $\varepsilon_{j(i)}$ are assumed normal with mean 0 and variance σ_ε^2, and the subscript $\varepsilon_{j(i)}$ indicates that the j^{th} level of replication (pot) is nested within the i^{th} level of treatment. These terms (the treatment residuals) measure the variation within treatment groups. The new term $\delta_{k(ij)}$ are the errors associated with each subsample (the pot residuals). It is convenient to think of the subsamples as being nested within each unique combination of treatment and replication. The $\delta_{k(ij)}$ are also assumed normal with mean 0 and variance σ^2. We can rewrite the model this way:

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

To get an intuitive sense of what this equation says, consider the plants within pots idea. τ_i measures the difference between a treatment mean and the overall mean (i.e. the treatment effect). $\varepsilon_{j(i)}$ measures the difference between a pot mean and the mean of its assigned treatment (i.e. the experimental error, the variation among replications treated alike). $\delta_{k(ij)}$ measures the difference between a plant and the mean of its pot (i.e. the subsampling error).

3.5.2.2 Nested ANOVA with equal subsample numbers: computation

Following the example in ST&D, page 159.

In this experiment, mint plants are exposed to combinations of temperature and daylight and their one-week stem growth measured. The 6 treatment combinations (2 temperature levels by 3 light levels) 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 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 page 159):

	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 = $Y_{ij.}$	15	17.5	11.5	18	14	17.5	19	21.5	22	32	28	28	22	26.5	29	33	27	35
Treatment totals = $Y_{i..}$	44.0			49.5			62.5			88.0			77.5			95.0		
Treatment means= $\overline{Y}_{i..}$	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 sums of squares satisfies:

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

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 variation among pots within a treatment (experimental error) and the variation among plants within a pot (subsample error). The resulting equation can be written

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

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 *subsampling error*.

Nested ANOVA table:

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

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 among treatment means except the treatment effects themselves.

Estimation of the different components of variance in the pot experiment

Again, the main objective of a nested design is to estimate the components of variance. To do this, we deconstruct the calculated mean squares according to their underlying theoretical models, called their ***expected mean squares*** (EMS, see last column in the above table) for each component of the linear model, as shown in the table 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 $= \sigma_\delta^2$, so $\sigma_\delta^2 =$ **0.93**

MSEE $= \sigma_\delta^2 + \mathbf{4}\sigma_\varepsilon^2$, so $\sigma_\varepsilon^2 =$ (MSEE - σ_δ^2)/**4** = (**2.15** - 0.93)/**4** = 0.30

MST $= \sigma_\delta^2 + 4\sigma_\varepsilon^2 + \mathbf{12}\Sigma\tau^2/5$, 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.

3.5.2.3 The optimal allocation of resources

(See Biometry Sokal & Rohlf page 309 for a detailed description). The main objective of a nested design is to investigate how the variation is distributed between 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 variance of the means can be calculated:

$$s^2_{\bar{Y}} = \frac{s^2_{e.u.}}{n_{sub} * r} + \frac{s^2_{sub}}{r}$$

Where n_{sub} is the number of subsamples per experimental unit and r is the number of replications per treatment. You can use this formula to test the effect of the different numbers of subsamples and replications on $s^2_{\bar{Y}}$ (and thus the total information in the experiment) and use the different values to calculate relative efficiencies among designs.

However, the relative efficiency of one design with respect to another is not very meaningful unless the relative costs of the two designs are also taken into consideration. Clearly, if one design is twice as efficient as another but at the same time is ten times as expensive, we might not choose it. 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 plus the cost of each experimental unit multiplied by the number of experimental units:

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

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 number of subsamples will increase when the relative cost of the subsamples is low and the variance within experimental units is high ($s^2_{e.u.}$).

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

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

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, that means be put only one plant per pot). In terms of efficiency, 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.