Lecture 5

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

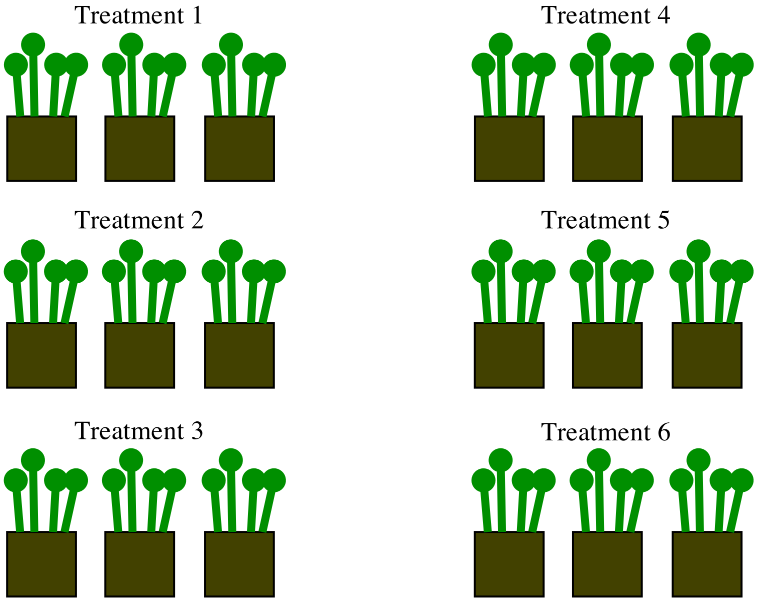
If observations made on the same experimental unit vary a lot, you may decide to make several observations on each experimental unit. Such observations are called **subsamples**.

Examples:

1. Individual plants within pots (where pots = experimental units)
2. Individual trees within an orchard plot (where treatments are assigned to the plots)
3. Individual sheep within a herd (where treatments are assigned to the herd)

Nested ANOVAs are not limited to two hierarchical levels. 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 ANOVA is to dissect the variation in a system into its components, thereby ascertaining the sources and magnitudes of error in an experiment or process.



Nesting within a CRD

The linear model

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

Two random elements contribute to each observation:

1. The variation among experimental units treated alike.

   $\varepsilon_{j(i)}$ are assumed normal with mean 0 and common variance

2. The variation among sampling units within an experimental unit.

   $\delta_{k(ij)}$ are also assumed normal with mean 0 and a common variance

We can rewrite the model this way:

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

Example: Mint plants are exposed to six different growing conditions and their one-week stem growth measured. The 6 treatments are assigned randomly across 18 pots. Within each pot are four plants.

	Cond 1			**Cond 2**			**Cond 3**			**Cond 4**			**Cond 5**			**Cond 6**		
Plant	**Pot**			**Pot**			**Pot**			**Pot**			**Pot**			**Pot**		
	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 Mean	3.8	4.4	2.9	4.5	3.5	4.4	4.5	5.4	5.5	8.0	7.0	7.0	5.5	6.6	7.3	8.3	6.8	8.8
Trt Mean	3.7			4.1			5.1			7.3			6.5			7.9		

In this example, t = 6, r = 3, s = number of subsamples = 4, and n = trs = 72. Overall mean = 5.8.

Recall that for a CRD:

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

For a nested CRD:

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

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 CRD, table of Expected Mean Squares (EMS)

Source of variation	**Expected MS**	**F**
Treatments (τ_i)	$\sigma_\delta^2+\mathbf{4}\sigma_\varepsilon^2+\mathbf{12}\Sigma\tau^2/5$	**MST / MSEE**
Exp. Error ($\varepsilon_{j(i)}$)	$\sigma_\delta^2+\mathbf{4}\sigma_\varepsilon^2$	MSEE / MSSE
Samp. Error ($\delta_{k(ij)}$)	σ_δ^2	
Total		

> 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 components of variance in the pot experiment

The main objective of a nested design is to estimate the components of variance.

Variance		Sum of	Mean	Variance	Percent of
Source	df	Squares	Squares	component	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 $\sigma_\varepsilon^2 = (\text{MSEE} - \sigma_\delta^2)/\mathbf{4} = (\mathbf{2.15} - 0.93)/\mathbf{4} = 0.30$

MST $= \sigma_\delta^2+4\sigma_\varepsilon^2+\mathbf{12}\Sigma\tau^2/5$, so $\Sigma\tau^2/5 = (\text{MST} - \text{MSEE})/\mathbf{12} = (\mathbf{35.92} - 2.15)/12 = 2.81$

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

The optimal allocation of resources

The main objective of a nested design is to investigate how variation is distributed among experimental units and among subsamples (i.e. quantifying the sources of error in the experiment).

For a two-level nested design, the total cost (C) is:

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

To find the number of subsamples (n_{sub}) per experimental unit that will simultaneously minimize cost *and* 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 ($s^2_{e.u.}$).

Example: One pot costs \$50. One plant costs \$18.

$$n_{sub} = \sqrt{\frac{C_{e.u.} * s^2_{sub}}{C_{sub} * s^2_{e.u.}}} = \sqrt{\frac{50 * 0.93}{18 * 0.30}} = 2.93 \approx 3$$

Given the components of variance and the relative costs of replications and subsamples in this experiment, an optimum allocation of resources would be 3 plants 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.