Handout 9

Nested Designs

Nested design (both factors fixed):

Consider the data set given later. A company runs three schools for mechanics, one in each of the following three cities: Atlanta (i=1), Chicago (i=2) and San Francisco (i=3).· The instructors are:

Atlanta, David (j=1), Lisa (j=2);

Chicago, Jason (j=1), Mark (j=2),

San Francisco: Jennifer (j=1), Robert (j=2).

In each school n=2 students were trained in a particular session. A summary measure of learning was obtained for each student. Let Y_{ijk} be the learning score of the kth students trained by the jth instructor at the i^{th} school. Note that the instructors in different cities are entirely different. We can say that the factor (instructor) is nested in school (city). If the same two instructors taught in the three schools, then we will have what is called "crossed factors" which is what you have studied in two-factor ANOVA models..

A means model here can be written as

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, k = 1, .., n, j = 1, .., b, i = 1, ..., a,$$

where ε_{ijk}'s are iid $N(0, \sigma^2)$. We may rewrite μ_{ij} as

$$\mu_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{ij} - \mu_{i.}) := \mu_{..} + \alpha_i + \beta_{j(i)},$$

where $\mu_{..}$ is the overall mean, α_i is the school effect, and $\beta_{j(i)}$ is the instructor effect nested in school. Constraints are

$$\sum \alpha_i = 0, \ \sum_j \beta_{j(i)} = 0 \text{ for each } i.$$

So the model can be rewritten as

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}, \text{ with}$$
$$E(Y_{ijk}) = \mu_{..} + \alpha_i + \beta_{j(i)}, Var(Y_{ijk})) = \sigma^2,$$
$$Y_{ijk}\text{'s are independent.}$$

We may be interested in checking if there is an instructor effect,i.e. $H_0 : \beta_{j(i)} = 0$ for all j and i, vs. H_1 :not all $\beta_{j(i)}$ are equal to zero.

We may also be interested in testing $H_0 : \alpha_i = 0$ for all i, vs. H_1 : not all α_i are zero.

Estimates of $\mu_{..}$, α_i and $\beta_{j(i)}$ are

$$\hat{\mu}_{..} = \bar{Y}_{...}, \ \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \ \hat{\beta}_{j(i)} = \bar{Y}_{ij.} - \bar{Y}_{i..}$$

Fitted Y-values and the residuals are

$$\hat{Y}_{ijk} = \hat{\mu}_{..} + \hat{\alpha}_i + \ \hat{\beta}_{j(i)} = \bar{Y}_{ij.},\ e_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - \bar{Y}_{ij.} \tag{1}$$

Sums of squares, Mean squares

First note that we can write

$$Y_{ijk} - \bar{Y}_{...} = (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.}) = \hat{\alpha}_i + \ \hat{\beta}_{j(i)} + e_{ijk}.$$

As usual denote

$$\begin{aligned} SSTO &= \sum\sum\sum (Y_{ijk} - \bar{Y}_{...})^2\,,\ df = abn - 1 \\ SSA &= \sum\sum\sum \hat{\alpha}_i^2 = nb\sum \hat{\alpha}_i^2,\ df = a - 1, \\ SSE &= \sum\sum\sum e_{ijk}^2,\ df = (n-1)ab. \end{aligned}$$

Now note that for any fixed i,

$$SSB(A_i) = \sum_j \sum_k \hat{\beta}_{j(i)}^2 = n\sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..})^2,\ df = b - 1,$$

is the sum of squares due to instructor for a given school i. We can now write

$$\begin{aligned} SSB(A) &= \sum_i SSB(A_i) = \sum_i \sum_j \sum_k \hat{\beta}_{j(i)}^2 \\ &= n\sum_i \sum_j \hat{\beta}_{j(i)}^2,\ df = a(b-1), \end{aligned}$$

is the sum of squares due to instructors (nested in schools (factor A)). [Note that, if you run a two-factor ANOVA, ignoring that you have a nested design, then one still calculate $SSB(A)$ by adding SSB and $SSAB$, i.e., $SSB(A) = SSB + SSAB$.]

If we square both sides of (1) and sum over i, j, k, we have

$$SSTO = SSA + SSB(A) + SSE.$$

The mean squares are

$$MSA = SSA/(a-1),\ MSB(A) = SSB(A)/[a(b-1)],\ MSE = SSE/[(n-1)ab].$$

Here is an important fact.

Fact 1. (a) $E(MSE) = \sigma^2$, (b) $E(MSB(A)) = \sigma^2 + n\sum_i \sum_j \beta_{j(i)}^2/[a(b-1)]$, (c) $E(MSA) = nb\sum \alpha_i^2/(a-1)$.

Remark 1. (i) It is clear that the F-statistic $F^* = MSB(A)/MSE$ can be used to test $H_0 : \beta_{j(i)} = 0$ for all j and i, vs. H_1 :not all $\beta_{j(i)}$ are equal to zero. Similarly, the statistic $F^* = MSA/MSE$ can be used to test $H_0 : \alpha_i = 0$ for all i, vs. H_1 : not all α_i are zero.

(ii) If we want to test that there is no instructor effect at a particular school i, say at Chicago. Then we may use the statistic $F^* = MSB(A_i)/MSE$, where $MSB(A_i) = SSB(A_i)/(b-1)$. The df's are $(b-1, (n-a)ab)$.

Here is the ANOVA table for Example 1.

Source	df	SS	MS	F	p-val
School (A)	$a-1=2$	156.5	78.25	11.2	0.009
Instructor, B(A)	$a(b-1)=3$	567.5	189.17	27.0	0.001
Error	$(n-1)ab=6$	42.0	7.00		
Total	$nab-1=11$	766.0			
Decomposition of SSB(A)					
Source	df	SSB(A_i)	MSB(A_i)	F	p-val
i=1, Atlanta	1	210.25	210.25	30.036	0.0015
i=2, Chicago	1	132.25	132.25	18.893	0.0048
i=3, San Francisco	1	225.00	225.00	32.143	0.0013
Total	3	567.5			

From this table it seems that both instructor effects and school effect are significant. Decomposition of $SSB(A)$ also shows that for each school, instructor effect is significant. In order to understand the table where decomposition of SSB(A) is given, let us assume that we want to test if the instructor effect for Atlanta (i=1) is significant. This is equivalent to testing $H_0 : \beta_{j(1)} = 0$ for all j vs. H_1 :not all $\beta_{j(1)}$ are zero. The test statistic is $F^* = MSB(A_1)/MSE$, where $MSB(A_1) = SSB(A_1)/(b-1)$. The df's for this test are $(b-1, (n-1)ab) = (1, 6)$.

Estimation: Estimates of $\mu_{..}$, α_i and $\beta_{j(i)}$ are given above. Note here $\mu_{i.} = \mu_{..} + \alpha_i$. The following table summarizes various estimates, variances of estimates and their standard errors.

Parameter (θ)	$\hat{\theta}$	$Var(\hat{\theta})$	$s^2(\hat{\theta})$
$\theta = \mu_{..}$	$\bar{Y}_{...}$	$\sigma^2/(nab)$	$MSE/(nab)$
$\theta = \mu_{ij}$	$\bar{Y}_{ij.}$	σ^2/n	MSE/n
$\theta = \mu_{i.}$	$\bar{Y}_{i..}$	$\sigma^2/(nb)$	$MSE/(nb)$
$\theta = \alpha_i$	$\bar{Y}_{i..} - \bar{Y}_{...}$	$[(a-1)/(nab)]\sigma^2$	$[(a-1)/(nab)]MSE$
$\theta = \beta_{j(i)}$	$\bar{Y}_{ij.} - \bar{Y}_{i..}$	$[(b-1)/(nb)]\sigma^2$	$[(b-1)/(nb)]MSE$
$\theta = \sum c_i \mu_{i.}$	$\sum c_i \bar{Y}_{i..}$	$[1/(nb)]\sum c_i^2 \sigma^2$	$[1/(nb)]\sum c_i^2 MSE$
$\theta = \sum c_j \beta_{j(i)}$, i fixed, $\sum c_j = 0$	$\sum c_j \hat{\beta}_{j(i)}$	$(1/n)\sum c_j^2 \sigma^2$	$(1/n)\sum c_j^2 MSE$

We have found that the school effect is significant. Let us now compare the three schools by looking at $\mu_{i.} - \mu_{i'.}, i \neq i'$. We will construct simultaneous 90% confidence intervals for pairwise differences of $\mu_{i.}$'s

using Tukey's method. Estimate of $\mu_{i\cdot} - \mu_{i'\cdot}$ is $\bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\cdot\cdot}$. Note that

$$Var(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\cdot\cdot}) = \frac{1}{nb}(2)\sigma^2,\ s^2(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\ \cdot\cdot}) = \frac{1}{(2)(b)}(2)(7.00) = 3.5, s(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\ \cdot\cdot}) = 1.87.$$

The Tukey multiplier is

$$T = \frac{1}{\sqrt{2}}q[1 - 0.10; a.(n-1)ab] = \frac{1}{\sqrt{2}}q(0.90; 3, 6) = 2.52, \text{ and}$$
$$Ts(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\ \cdot\cdot}) = (2.52)(1.87) = 4.71.$$

We have

$$\bar{Y}_{1\cdot\cdot} = 19.75,\ \bar{Y}_{2\cdot\cdot} = 14.25,\ \bar{Y}_{3\cdot\cdot} = 11.00.$$

Simultaneous 90% confidence intervals for pairwise differences of $\mu_{i\cdot}$'s are

$$\mu_{1\cdot} - \mu_{2\cdot} : (19.75 - 14.25) \pm 4.71, \text{ i.e., } (0.8, 10.2),$$
$$\mu_{1\cdot} - \mu_{3\cdot} : (19.75 - 11.00) \pm 4.71, \text{ i.e., } (4.0, 13.5),$$
$$\mu_{2\cdot} - \mu_{3\cdot} : (14.25 - 11.00) \pm 4.71, \text{ i.e., } (-1.25, 8.0).$$

These intervals suggest that the mean learning scores for Chicago and San Francisco may not be all that different.

Nested Design (factor A fixed, factor B (nested in A) random)·

A data set is given in Example 2 for this case, where technician effect (A) is fixed, but rat effect (B, nested in A) is random. The appropriate model here is:

$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},\ k = 1, \ldots, n = 2,\ j = 1, \ldots, b = 3, i = 1, \ldots, a = 2,$$

where $\mu_{\cdot\cdot}$ is the overall mean, α_i is the technician effect (factor A) and $\beta_{j(i)}$ is the rat (factor B, random) effect nested in technician (factor A). It is understood that

(i) $\sum \alpha_i = 0$, (ii) $\beta_{j(i)}$'s are iid $N(0, \sigma^2_{B(A)})$, (iii) ε_{ijk}'s are iid $N(0, \sigma^2)$,

(iv) $\{\beta_{j(i)}\}$ and $\{\varepsilon_{ijk}\}$ are independent.

Here

$$E(Y_{ijk}) = \mu_{\cdot\cdot} + \alpha_i,\ Var(Y_{ijk}) = \sigma^2_{B(A)} + \sigma^2.$$

So the components of variance are $\sigma^2_{B(A)}$ and σ^2. Note that

$$Cov(Y_{ijk}, Y_{ijk'}) = \sigma^2_{B(A)} \text{ if } k \neq k',$$
$$Cov(Y_{ijk}, Y_{ijk'}) = 0 \text{ if } (i, j) \neq (i', j').$$

It is clear that Y_{ijk}'s are no longer independent.

For the data set given in Example 2, we see that neither technician effect nor rat effect seem to be statistically significant. We may still try to estimate the variance component $\sigma^2_{B(A)}$. Note that

$$s^2_{B(A)} = \frac{MSB(A) - MSE}{n} = \frac{0.10725 - 0.06294}{4} = 0.01108.$$

Thus the proportion of variability $\sigma^2_{B(A)}/(\sigma^2_{B(A)} + \sigma^2)$ explained by factor B (rats, nested in A) is

$$\frac{s^2_{B(A)}}{s^2_{B(A)} + MSE} = \frac{0.01108}{0.01108 + .06294} = 0.1495.$$

This proportion is rather small and thus it is consistent with the result from the ANOVA table that the null hypothesis $H_0 : \sigma^2_{B(A)} = 0$ cannot be rejected.

Estimation: The following table summarizes various estimates, variances of estimates and their standard errors.

Parameter (θ)	$\hat{\theta}$	$Var(\hat{\theta})$	$s^2(\hat{\theta})$
$\theta = \mu_{..}$	$\bar{Y}_{...}$	$(n\sigma^2_{B(A)} + \sigma^2)/(nab)$	$MSB(A)/(nab)$
$\theta = \mu_{i.}$	$\bar{Y}_{i..}$	$(n\sigma^2_{B(A)} + \sigma^2)/(nb)$	$MSB(A)/(nb)$
$\theta = \alpha_i$	$\bar{Y}_{i..} - \bar{Y}_{...}$	$[(a-1)/(nab)](n\sigma^2_{B(A)} + \sigma^2)$	$[(a-1)/(nab)]MSB(A)$
$\theta = \sum c_i\mu_{i.}$	$\sum c_i\bar{Y}_{i..}$	$[1/(nb)]\sum c_i^2\sigma^2$	$[1/(nb)]\sum c_i^2 MSB(A)$

Nested Design (factor A random, factor B (nested in A) random).

A data set is given Example 3 where the plant effect (A) is is random, and the leaf effect (B, nested in A) is also random. The appropriate model here is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},\ k = 1, \ldots, n = 2,\ j = 1, \ldots, b = 3, i = 1, \ldots, a = 4,$$

where $\mu_{..}$ is the overall mean, α_i is the plant effect (factor A, random) and $\beta_{j(i)}$ is the leaf effect (factor B, random) nested in plant (factor A). Here $\mu_{..}$ overall mean calcium concentration. It is understood that

(i) α_i's iid $N(0, \alpha^2_\alpha)$, (ii) $\beta_{j(i)}$'s are iid $N(0, \sigma^2_{B(A)})$, (iii) ε_{ijk}'s are iid $N(0, \sigma^2)$,

(iv) $\{\alpha_i\}$, $\{\beta_{j(i)}\}$ and $\{\varepsilon_{ijk}\}$ are independent.

For this model,

$$E(Y_{ijk}) = \mu_{..},\ Var(Y_{ijk}) = \sigma^2_\alpha + \sigma^2_{B(A)} + \sigma^2.$$

Thus $\sigma^2_\alpha, \sigma^2_{B(A)}$ and σ^2 are components of $Var(Y_{ijk})$. Note that

$$Cov(Y_{ijk}, Y_{ijk'}) = \sigma^2_\alpha + \sigma^2_{B(A)} \text{ if } k \neq k',$$
$$Cov(Y_{ijk}, Y_{i'jk'}) = \sigma^2_{B(A)} \text{ if } i \neq i',$$
$$Cov(Y_{ijk}, Y_{i'j'k'}) = 0 \text{ if } i \neq i' \text{ and } j \neq j'.$$

It is clear that Y_{ijk}'s are not independent here.

From the ANOVA table, both plant and leaf effects seem to be statistically significant (especially leaf effect). Let us estimate σ_α^2 and $\sigma_{B(A)}^2$. Estimates are

$$s_\alpha^2 = \frac{MSA - MSB(A)}{nb} = \frac{2.3318 - 0.3273}{(2)(3)} = 0.3341.$$
$$s_{B(A)}^2 = \frac{MSB(A) - MSE}{n} = \frac{0.3273 - 0.0163}{2} = 0.1555.$$

Thus of the total variability of calcium contents, the proportions due to plant and leaf are respectively,

$$\frac{s_\alpha^2}{s_\alpha^2 + s_{B(A)}^2 + MSE} = \frac{0.3341}{0.3341 + 0.1555 + 0.0163} = \frac{0.3341}{0.5049} = 0.662,$$
$$\frac{s_{B(A)}^2}{s_\alpha^2 + s_{B(A)}^2 + MSE} = \frac{0.1555}{0.3341 + 0.1555 + 0.0163} = \frac{0.1555}{0.5049} = 0.308.$$

Finally, let now estimate $\mu_{...}$. An estimate is $\bar{Y}_{...}$. Note that

$$E(\bar{Y}_{...}) = \mu_{..},\ Var(\bar{Y}_{...}) = \frac{nb\sigma_\alpha^2 + n\sigma_{B(A)}^2 + \sigma^2}{nab},\ s^2(\bar{Y}_{...}) = \frac{MSA}{nab}.$$

So a $(1-\alpha)100\%$ confidence interval for $\mu_{..}$ is $\bar{Y}_{...} \pm t(1-\alpha/2; a-1)s(\bar{Y}_{...})$. For the "Calcium Concentration data, we have

$$\bar{Y}_{...} = 2.9954,\ s(\bar{Y}_{...}) = \sqrt{\frac{MSA}{nab}} = \sqrt{\frac{2.3318}{24}} = 0.3117.$$

Hence a 95% confidence interval for $\mu_{..}$ is

$$\bar{Y}_{...} \pm t(0.975; 3)(0.3117), \text{i.e., } 2.9954 \pm (3.1824)((0.3117),$$
$$\text{i.e., } 2.9954 \pm 1.0088, \text{ i.e., } (1.987, 4.003).$$

Mean Square Errors

	A fixed, B fixed	A fixed, B random	A random, B random
$E(MSA)$	$[nb/(a-1)]\sum \alpha_i^2 + \sigma^2$	$[nb/(a-1)]\sum \alpha_i^2 + n\sigma^2_{B(A)} + \sigma^2$	$nb\sigma^2_\alpha + n\sigma^2_{B(A)} + \sigma^2$
$E(MSB(A))$	$[n/\{a(b-1)\}]\sum\sum \beta^2_{j(i)} + \sigma^2$	$n\sigma^2_{B(A)} + \sigma^2$	$n\sigma^2_{B(A)} + \sigma^2$
$E(MSE)$	σ^2	σ^2	σ^2

F-tests

	A fixed, B fixed	A fixed, B random	A random, B random
$E(MSA)$	$H_0 : \alpha_i = 0$ for all i, $F^* = MSA/MSE$	$H_0 : \alpha_i = 0$ for all i, $F^* = MSA/MSB(A)$	$H_0 : \sigma^2_\alpha = 0$, $F^* = MSA/MSB(A)$
$E(MSB(A))$	$H_0 : \beta_{j(i)} = 0$ for all i, j, $F^* = MSB(A)/MSE$	$H_0 : \sigma^2_{B(A)} = 0$, $F^* = MSB(A)/MSE$	$H_0 : \sigma^2_{B(A)} = 0$, $F^* = MSB(A)/MSE$

Estimation of Variance Components

Parameter	A fixed, B fixed	A fixed, B random	A random, B random
σ^2	MSE	MSE	MSE
$\sigma^2_{B(A)}$		$\max(s^2_{B(A)}, 0)$, with $s^2_{B(A)} = [MSB(A) - MSE]/n$	$\max(s^2_{B(A)}, 0)$, with $s^2_{B(A)} = [MSB(A) - MSE]/n$
σ^2_α			$\max(s^2_\alpha, 0)$ with $s^2_\alpha = [MSB(A) - MSA]/(nb)$

Nested Design for unbalanced data.

Nested Design (A fixed, B fixed): So far we have only discussed the balanced case. For the unbalanced case, a regression approach is needed. One needs to create $a-1$ X-variables for school, and then for each school, one needs to create $b-1$ X-variables. Thus the number of X variables is $a-1+a(b-1)=ab-1$. The following examples will describe what are to be done.

(i) When $a=3$ and $b=2$, the following table describes how the variables are to be created.

$X_1=$	1 (school 1)	0 (school 2)	-1 (school 3)
$X_2=$	0 (school 1)	1 (school 2)	-1 (school 3)
$X_3=$	1 (instructor 1, school1)	-1 (instructor 2, school 1)	
$X_4=$	1 (instructor 1, school 2)	-1 (instructor 2, school 2)	
$X_5=$	1 (instructor 1, school 3)	-1 (instructor 2, school 3)	

Thus the model for Example 1 can be rewritten as

$$Y_{ijk}=\mu_{..}+\alpha_1X_{ijk1}+\alpha_2X_{ijk2}+\beta_{1(1)}X_{ijk3}+\beta_{1(2)}X_{ijk4}+\beta_{1(3)}X_{ijk5}+\varepsilon_{ijk}.$$

(ii) When $a=3$ and $b=3$, i.e, there are 3 instructors in each school, the following table describes how the X-variables are to be created. Note that for each school you now need two X-variables.

School	$X_1=$	1 (school 1)	0 (school 2)	-1 (school 3)
School	$X_2=$	0 (school 1)	1 (school 2)	-1 (school 3)
Instructor variables	$X_3=$	1 (instructor 1, school1)	0 (instructor 2, school 1)	-1 (instructor 3, school 1)
for school1	$X_4=$	1 (instructor 2, school 1)	0 (instructor 1, school 1)	-1 (instructor 3, school 1)
Instructor variables	$X_5=$	1 (instructor 1, school 2)	0 (instructor 2, school 2)	-1 (instructor 3, school 2)
for school 2	$X_6=$	1 (instructor 2, school 2)	0 (instructor 1, school 2)	-1 (instructor 3, school 2)
Instructor variables	$X_7=$	1 (instructor 2, school 3)	0 (instructor 2, school 3)	-1 (instructor 3, school 3)
for school 3	$X_8=$	1 (instructor 2, school 3)	0 (instructor 1, school 3)	-1 (instructor 3, school 3)

Thus the model can be written as

$$\begin{aligned}Y_{ijk}=\mu_{..}&+[\alpha_1X_{ijk1}+\alpha_2X_{ijk2}]+[\beta_{1(1)}X_{ijk3}+\beta_{2(1)}X_{ijk4}]\\&+[\beta_{1(2)}X_{ijk5}+\beta_{2(2)}X_{ijk6}]+[\beta_{1(3)}X_{ijk7}+\beta_{2(3)}X_{ijk8}]+\varepsilon_{ijk}.\end{aligned}$$

Nested Design (A fixed, B random): As discussed in the previous handouts, we will need to take a regression approach for the unbalanced case. The usual approach (also called Henderson's approach, pages 4-5 in Handout 8), we create $a-1$ X-variables for factor A and, for each i, create $b-1$ X-variables for factor B. Thus there are $a-1+a(b-1)=ab-1$ X-variables and then one can run a regression model that is equivalent to the ANOVA model.

Henderson's approach suffers from the deficiency that it implicitly assumes that the factor effects sum to zero even when the factor is random (pages 4-5 in Handout 8). For this reason, a more careful and proper approach requires one to create $a-1$ X-variables and then, for each i, create b X-variables each taking 0-1 values. Thus there will be a total of $a-1+ab = a(b+1)-1$ X-variables and a mixed linear model along the lines discussed in Handout 8 (page 5) needed to be employed.

Nested Design (A random, B random): Henderson's approach is the same as the approach where factors A and B are assumed to be fixed requiring creation of $ab-1$ X-variables. A more appropriate and correct approach requires creating a dummy X-variabIes (0-1 valued) and then creating, for each i, b dummy X-variables (0-1 values). Thus there will be a total of $a+ab$ X-variables and a mixed linear model can be used (page 5, Handout 8).

Nested design with three or more factors:

Conceptually, there is no problem if there are three or more factors. Thy could all be nested: B nested in A and C nested in B. It may also be B and C could be crossed but both are nested in A. There are many other possibilities. The textbook provides a few examples.

Nested design with covariates.

It is possible to have an analysis of covarince type scenario where there may be one or more quantitative independent variables (covariates). This by itself does not pose any special problem. However, one always needs to be careful about various sums of squares, hypotheses tests and confidence intervals.

Data Sets.

Example 1 (Training schools for mechanics).

A company runs three schools: one each in Atlanta (i=1), Chicago (i=2) and San Francisco (i=3). There are two instructors in each school. In Atlanta, they are David (j=1), Lisa (j=2);

Chicago, Jason (j=1), Mark (j=2); San Francisco: Jennifer (j=1), Robert (j=2). Two mechanics are assigned to each instructor and after a specified time, a summary measure (Y) of learning is obtained for each student. Note that instructor (factor B) is nested in factor A, since the instructors in a school only work in that school. The goal is to see if there are differences among schools and also if there are differences among instructors in each school.

	Instructor (B)			
School (A)	j=1		j=2	
i=1	25	29	14	11
i=2	11	6	22	18
i=3	17	20	5	2

The model here is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},\ k = 1, \ldots, n = 2,\ j = 1 \ldots, b = 2, i = 1, \ldots, a = 3,$$

where $\mu_{..}$ is the overall mean, α_i is the school effect (factor A) and $\beta_{j(i)}$ is the instructor (factor B) effect nested in school (factor A). It is understood that ε_{ijk}'s are iid $N(0, \sigma^2)$, and α's and $\beta_{j(i)}$'s satisfy the constraints:

(i) $\sum \alpha_i = 0$, (ii) $\sum_j \beta_{j(i)} = 0$ for each i.

Table of estimated means ($\bar{Y}_{ij.}$).

	j=1	j=2	$\bar{Y}_{i..}$	$\hat{\alpha}_i$
i=1	27	12.5	19.75	4.75
i=2	8.5	20	14.25	-0.75
i=3	18.5	3.5	11.00	-4
			$\bar{Y}_{...} = 15$	

Table of estimated $\hat{\beta}_{j(i)}$

	j=1	j=2
i=1	7.25	-7.25
i=2	-5.75	5.75
i=3	7.5	-7.5

Example 2. (Protein uptake of fluorescently labeled protein in rat kidneys).

A researcher wanted to know if his two technicians, Brad (i=1) and Janet (i=2) were performing the procedure consistently. Brad chose three rats randomly: Arnold (j=1), Ben (j=2), Charlie (j=3) , and Janet chose three rants randomly: Dave (j=1), Eddy (j=2) and Frank (j=3).

Here factor A (technician) is fixed, but factor B (rat) is random and is nested in factor A. In your text it has been called two-stage subsampling.

Technician	i=1			i=2		
Rat	j=1	j=2	j=3	j=1	j=2	j=3
	1.119	1.045	0.987	1.388	1.395	1.257
	1.300	1.142	0.987	1.104	0.971	1.030
	1.541	1.260	0.871	1.158	1.397	1.941
	1.518	0.619	0.945	1.319	1.547	1.076

The goal here is to know if the technician effect exists.

The appropriate model here is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},\ k = 1, \ldots, n = 4,\ j = 1, \ldots, b = 3, i = 1, \ldots, a = 2,$$

where $\mu_{..}$ is the overall mean, α_i is the technician effect (factor A) and $\beta_{j(i)}$ is the rat (factor B, random) effect nested in technician (factor A). It is understood that

(i) $\sum \alpha_i = 0$, (ii) $\beta_{j(i)}$'s are iid $N(0, \sigma^2_{B(A)})$, (iii) ε_{ijk}'s are iid $N(0, \sigma^2)$, (iv) $\{\beta_{j(i)}\}$ and $\{\varepsilon_{ijk}\}$ are independent.

Here

$$\bar{Y}_{...} = 1.2049, \bar{Y}_{1..} = 1.1112, \bar{Y}_{2..} = 1.2986,$$
$$\hat{\alpha}_1 = \bar{Y}_{1..} - \bar{Y}_{...} = -0.0937,\ \hat{\alpha}_2 = \bar{Y}_{2..} - \bar{Y}_{...} = 0.0937.$$

ANOVA Table

Source	df	SS	MS	F	p-val
Technician(A)	$a-1=1$	0.2108	0.21075	$MSA/MSB(A) = 1.965$	0.234
Rat (B), nested in A	$a(b-1)=4$	0.4290	0.10725	$MSB(A) = 1.704$	0.192
Error	$(n-1)ab = 18$	1.1330	0.06294		
Total	$nab-1=23$	1.7728			

An important issue on notations: In your textbook, the $\beta_{j(i)}$'s have been denoted by $\varepsilon_{j(i)}$ and ε_{ijk}'s have been denoted by η_{ijk}. The textbook notations are somewhat nonstandard and, for that reason, we will continue with the notations given above.

Example 3: (Calcium concentration in turnip greens, Snedecor and Cochran(1976))

An experiment is conducted to study the variability of calcium concentration in turnip greens. Four plants are selected at random, then three leaves are selected from each plant. Two 100-mg samples are taken from each leaf. The amount of calcium is determined by microchemical methods.

Note that plant (factor A) is random, leaf (factor B) is also random but is nested in plant. This has been called "three-stage" subsampling in the text. The goal here is to estimate the mean calcium concentration per leaf, and the variability in concentration across plants and among leaves in plants.

	Leaf (B)					
Plant (A)	j=1		j=2		j=3	
i=1	3.28	3.09	3.52	3.48	2.88	2.80
i=2	2.46	2.44	1.87	1.92	2.19	2.19
i=3	2.77	2.66	3.74	3.44	2.55	2.55
i=4	3.38	3.87	4.07	4.12	3.31	3.31

The appropriate model here is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},\ k = 1, \ldots, n = 2,\ j = 1, \ldots, b = 3, i = 1, \ldots, a = 4,$$

where $\mu_{..}$ is the overall mean, α_i is the plant effect (factor A, random) and $\beta_{j(i)}$ is the leaf effect (factor B, random) nested in plant (factor A). Here $\mu_{..}$ overall mean calcium concentration. It is understood that

(i) α_i's iid $N(0, \alpha^2_\alpha)$, (ii) $\beta_{j(i)}$'s are iid $N(0, \sigma^2_{B(A)})$, (iii) ε_{ijk}'s are iid $N(0, \sigma^2)$, (iv) $\{\alpha_i\}$, $\{\beta_{j(i)}\}$ and $\{\varepsilon_{ijk}\}$ are independent.

Here $\bar{Y}_{...} = 2.9954$.

ANOVA Table

Source	df	SS	MS	F	p-val
Plant(A)	$a-1=3$	6.995	2.3318	$MSA/MSB(A)=7.124$	0.012
Leaf(B), nested in A	$a(b-1)=8$	2.618	0.3273	$MSB(A)=20.080$	0.000
Error	$(n-1)ab=12$	0.196	0.0163		
Total	$nab-1=23$	1.7728			

An important issue on notations: In your textbook, as in Example 2, $\beta_{j(i)}$'s have been denoted by $\varepsilon_{j(i)}$ and ε_{ijk}'s have been denoted by η_{ijk}. The textbook notations are somewhat nonstandard and, for that reason, we will continue with the notations given above.