

Lecture 14

Topic 10: ANOVA models for random and mixed effects

To this point, we have considered only the Fixed Model (Model I) ANOVA; now we will extend the method of ANOVA to other experimental objectives.

Fixed and Random Models in One-way Classification Experiments

A recap of the fixed-effects model

1. Treatment levels are selected by the researcher.
2. The interest is *only* in those specific treatment levels.
3. Conclusions are limited to those specific treatment levels.
4. The linear model:

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

where the treatment effects (τ_i) are *fixed* and $\sum \tau_i = \sum (\bar{Y}_i - \mu) = 0$.

5. $H_0: \tau_1 = \tau_2 = \dots = \tau_t = 0$
 $H_1: \text{some } \tau_i \neq 0$
6. When H_0 is false, there is an additional component of variance in the experiment:

$$r \sum_{i=1}^t \frac{\tau_i^2}{t-1}$$

7. The typical objective is to compare particular treatment means to one another in an attempt to detect differences.

A typical example: Measuring the effect of four different, specific fertilizers on yield.

Introducing the random-effects model

1. Treatment levels are *a random sample* drawn from a larger population of treatments with mean effect 0 and variance σ_τ^2 .
2. The interest is in this larger population of treatment levels.
3. Conclusions are meant to extend to the larger population of treatment levels.
4. The linear model:

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

where the treatment effect (τ_i) is *a random variable*. While $\sum \tau_i = \sum (\bar{Y}_i - \mu) = 0$ for the population of treatment effects, for any given random sample of treatment levels, $\sum \tau_i \neq 0$.

5. $H_0: \sigma_\tau^2 = 0$
 $H_1: \sigma_\tau^2 \neq 0$
6. When H_0 is false, there is an additional component of variance in the experiment:

$$r\sigma_\tau^2$$

7. The typical objective is to test for the presence of this additional component of variance and to estimate its magnitude.

Example: In a sheep breeding experiment conducted to characterize the breeding value of sires from a certain breed, four sires were randomly selected from a population and each sire was mated with six dams. The weights of all the newborn animals were recorded.

These four sires are of *no specific interest*; they are merely a random sample from a population of sires of the breed.

These four sires are interesting only to the extent that they represent their population. If the experiment were to be repeated, another set of sires very likely would be used.

For a one-way ANOVA, the computation is the same for the fixed and random models.
But the **objectives** and **conclusions** are different.
And the computations following the initial significance test are also different.

Sometimes the determination of whether an effect is fixed or random is not obvious.

Laboratories or technicians in a comparative study
Years in a multiple-year trial
Locations in a multiple-location study

These factors can be fixed or random depending on the objective of the study, the intended inferences to be made, and the process by which the levels of the factors were selected.

Differences between the fixed- and random-effects models

1. The objectives are different.

Fixed-effects (the fertilizer experiment): Each fertilizer is of specific interest, and the purpose is to compare these particular fertilizers to one another.

Random-effects (the sheep breeding experiment): Each sire is of no specific interest. The sires in the study are merely a random sample from which inferences are to be made concerning the population. The purpose is to estimate the component of variance σ_τ^2 .

2. The sampling procedures are different.

Fixed-effects (the fertilizer experiment): The fertilizers are selected purposefully (i.e. not randomly) by the investigator. If the experiment is repeated, the same four fertilizers will be used again and only the random errors change from experiment to experiment (i.e. the τ_i 's are assumed to be constants; only the ϵ_{ij} 's change).

Random-effects (the sheep breeding experiment): The sires are selected randomly, and the unknown variance of the population of sires contributes to the total sum of squares. If the experiment is repeated, the four sires most likely will differ each time (i.e. not only the errors vary but the sire effects (τ_i 's) vary as well).

3. The expected sums of the effects are different.

Because the effects are constants, in the fixed-effects model:

$$\sum \tau_i = \sum (\bar{Y}_i - \mu) = 0$$

But for any given sample in the random-effects model:

$$\sum \tau_i = \sum (\bar{Y}_i - \mu) \neq 0$$

4. And, therefore, the expected variances are different.

For Model I:

$$\begin{aligned}\text{Var of } Y_{ij} &= \text{variance of } \mu + \text{variance of } \tau_i + \text{variance of } \varepsilon_{ij} \\ &= \text{variance of } \varepsilon_{ij} \\ &= \sigma_\varepsilon^2\end{aligned}$$

For Model II:

$$\begin{aligned}\text{Var of } Y_{ij} &= \text{variance of } \mu + \text{variance of } \tau_i + \text{variance of } \varepsilon_{ij} \\ &= \text{variance of } \tau_i + \text{variance of } \varepsilon_{ij} \\ &= r\sigma_\tau^2 + \sigma_\varepsilon^2 \quad (\sigma_\tau^2 \text{ and } \sigma_\varepsilon^2 \text{ are the variance components})\end{aligned}$$

The **expected mean squares** for these models:

Source	df	MS	EMS	
			Fixed Model	Random Model
Treatment	t-1	MST	$\sigma_\varepsilon^2 + r \sum \frac{\tau_i^2}{t-1}$	$\sigma_\varepsilon^2 + r\sigma_\tau^2$
Error	t(r-1)	MSE	σ_ε^2	σ_ε^2

Example: Suppose five different cuts of meat are taken from each of three pigs, all from the same breed, and the fat content is measured for each cut.

1. If the goal were to compare the fat content among these five cuts of meat:

Fixed-effects model

Five treatment levels (cuts)

Three replications (pigs)

2. If the goal were to assess animal-to-animal and within-animal variation for this breed:

Random-effects model

Three levels to estimate among-animal variation (pigs; like treatment)

Five levels to estimate within-animal variation (cuts; like replications)

Sticking with the pigs...let's assume the goal is to assess the components of variation (among-animals and within-animals):

$$\begin{aligned} \text{MST} &= 80 \quad (\text{df} = 2) \\ \text{MSE} &= 20 \quad (\text{df} = 12) \\ F &= 80/20 = 4 \quad (p < 0.05) \end{aligned}$$

Our conclusion? $\sigma_{\tau}^2 > 0$

Source	df	MS	EMS
			Random Model
Treatment	t-1	MST	$r\sigma_{\tau}^2 + \sigma_{\varepsilon}^2$
Error	t(r-1)	MSE	σ_{ε}^2

$$80 = r\sigma_{\tau}^2 + \sigma_{\varepsilon}^2$$

Since $r = 5$ and $\sigma_{\varepsilon}^2 = 20$, we can solve:

$$\begin{aligned} 80 &= 5\sigma_{\tau}^2 + 20 \\ \sigma_{\tau}^2 &= 12 \end{aligned}$$

Fat content is nearly twice as variable among cuts within this breed of pig (σ_{ε}^2) as it is among pigs of this breed (σ_{τ}^2).

Two-way classification experiments

$$Y_{ijk} = \mu + \tau_{Ai} + \tau_{Bj} + (\tau_A \tau_B)_{ij} + \varepsilon_{ijk}$$

Consider: A field study of several varieties of sunflower (Factor A) tested across a number of locations (Factor B). In each location, a CRD is used, with each variety replicated over r plots. Let Y_{ijk} represent the plot yield of the k^{th} plot of the i^{th} variety at the j^{th} location.

Fixed-effects model: Interested in a particular set of varieties grown in a particular set of locations.

Random-effects model: Interested in a population of varieties grown in some random set of locations which represent a particular region.

Mixed-effects model: Interested in a particular set of varieties grown in some random set of locations which represent a particular region.

	Fixed	Random	Mixed
μ	Mean yield in the study	Mean yield of all varieties at all possible locations	Mean yield of specific varieties at all possible locations
A (variety)	<i>True effect</i> of the i^{th} variety $\sum \tau_{Ai} = 0$	<i>Random effect</i> from $N(0, \sigma_\alpha^2)$ $\sum \tau_{Ai} \neq 0$	<i>True effect</i> of the i^{th} variety $\sum \tau_{Ai} = 0$
B (location)	<i>True effect</i> of the j^{th} location $\sum \tau_{Bj} = 0$	<i>Random effect</i> from $N(0, \sigma_\beta^2)$ $\sum \tau_{Bj} \neq 0$	<i>Random effect</i> from $N(0, \sigma_\beta^2)$ $\sum \tau_{Bj} \neq 0$
A:B	<i>Specific interaction effect</i> of the i^{th} variety and the j^{th} location	<i>Random effect</i> from $N(0, \sigma_{AB}^2)$	<i>Random effect</i> from $N(0, \sigma_{AB}^2)$
Variance Y	σ_ε^2	$\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2$	$\sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2$
Objective	Estimate and test hypotheses about τ_{Ai} , τ_{Bj} , and $(\tau_A \tau_B)_{ij}$	Estimate and test hypotheses about σ_A^2 , σ_B^2 , and σ_{AB}^2	Estimate and test hypotheses about τ_{Ai} , σ_B^2 , and σ_{AB}^2

Expected mean squares and F tests

EMS: Algebraic expressions which specify the underlying model parameters that are estimated by the calculated mean squares. Used to determine the appropriate error variances for F tests.

An EMS is composed of:

1. The error variances
2. Functions of variances of random effects
3. Functions of sums of squares of fixed effects

EMS table for three two-way classification experiments, featuring **a** varieties, **b** locations, and **r** replications:

Source	df	MS	Fixed Model	Random Model	Mixed Model Fixed Var, Random Loc
Var	a-1	MSV	$\sigma_\varepsilon^2 + br \sum \frac{\tau_A^2}{a-1}$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2 + br\sigma_A^2$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2 + br \sum \frac{\tau_A^2}{a-1}$
Loc	b-1	MSL	$\sigma_\varepsilon^2 + ar \sum \frac{\tau_B^2}{b-1}$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2 + ar\sigma_B^2$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2 + ar\sigma_B^2$
V:L	(b-1)(a-1)	MSVL	$\sigma_\varepsilon^2 + r \sum \frac{(\tau_A \tau_B)^2}{(a-1)(b-1)}$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2$	$\sigma_\varepsilon^2 + r\sigma_{AB}^2$
Error	ba(r-1)	MSE	σ_ε^2	σ_ε^2	σ_ε^2

Based on the expected mean squares, correct denominators can be identified to perform the appropriate F tests on locations, varieties, and their interaction:

The appropriate test statistic (F) is a ratio of mean squares that is chosen such that the expected value of the *numerator* differs from the expected value of the *denominator* only by the specific variance component or fixed factor being tested.

Source	Fixed	Random	Mixed
Variety	MSV/MSE	MSV/MSVL	MSV/MSVL
Location	MSL/MSE	MSL/MSVL	MSL/MSVL
V:L	MSVL/MSE	MSVL/MSE	MSVL/MSE

Note that the appropriate F tests change, depending on the model.

Expected Mean Squares (EMS) and Custom F Tests

Assume we are testing 8 varieties (fixed) across 4 locations (random), with 3 replications in each location. To generate the correct sums of squares, the proper R code would be:

```
lm( Yield ~ Variety + Location + Variety:Location )
```

But to carry out the proper F tests, we need to consult the table of expected mean squares:

Source	Mixed Model Fixed Var, Random Loc	Mixed Model Fixed Var, Random Loc
Var	$\sigma_{\varepsilon}^2 + r\sigma_{AB}^2 + br \sum \frac{\tau_A^2}{a-1}$	$\sigma_{\varepsilon}^2 + 3\sigma_{AB}^2 + 12 \sum \frac{\tau_A^2}{a-1}$
Loc	$\sigma_{\varepsilon}^2 + r\sigma_{AB}^2 + ar\sigma_B^2$	$\sigma_{\varepsilon}^2 + 3\sigma_{AB}^2 + 24\sigma_B^2$
V:L	$\sigma_{\varepsilon}^2 + r\sigma_{AB}^2$	$\sigma_{\varepsilon}^2 + 3\sigma_{AB}^2$
Error	σ_{ε}^2	σ_{ε}^2

The appropriate F tests in this case would be:

Source	DF	SS	Mean Square	F Value	Pr > F
Var	7	16.187979	2.312568	2.36	0.0606
Loc	3	0.739818	0.246606	0.25	0.8594
Error: MS (Var:Loc)	21	20.596432	0.980782		

Source	DF	SS	Mean Square	F Value	Pr > F
Var:Loc	21	20.596432	0.980782	1.08	0.3929
Error: MS (Error)	64	58.218486	0.909664		

An aside about RCBD's...

Source	Mixed Model	
	Fixed Var,	Random Loc
Trtmt	$\sigma_{\varepsilon}^2 + r\sigma_{\tau\beta}^2 + br \sum \frac{\tau^2}{t-1}$	
Block	$\sigma_{\varepsilon}^2 + r\sigma_{\tau\beta}^2 + tr\sigma_{\beta}^2$	
T:B	$\sigma_{\varepsilon}^2 + r\sigma_{\tau\beta}^2$	
Error	σ_{ε}^2	

Now we see the reasoning behind the standard practice of having just one replication per block-treatment combination. Merging the Trtmt:Block interaction into the error makes some sense because the correct denominator for Treatment *is* the Trt:Block interaction (i.e. we want to generalize across the entire population of blocks represented by the blocks actually used in the study).

Expected Mean Squares for a three-way ANOVA

Now consider a three-factor factorial experiment: A is fixed; B and C are both random. The linear model:

$$Y_{ijkl} = \mu + \tau_{Ai} + \tau_{Bj} + \tau_{Ck} + (\tau_A \tau_B)_{ij} + (\tau_A \tau_C)_{ik} + (\tau_B \tau_C)_{jk} + (\tau_A \tau_B \tau_C)_{ijk} + \varepsilon_{ijkl}$$

Source	Expected Mean Squares	F
A	$\tau\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + br\sigma_{AC}^2 + cr\sigma_{AB}^2 + bcr \sum \frac{\tau_A^2}{a-1}$?
B	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + ar\sigma_{BC}^2 + cr\sigma_{AB}^2 + acr\sigma_B^2$?
C	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + ar\sigma_{BC}^2 + br\sigma_{AC}^2 + abr\sigma_C^2$?
AB	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + cr\sigma_{AB}^2$	$\frac{MS(AB)}{MS(ABC)}$
AC	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + br\sigma_{AC}^2$	$\frac{MS(AC)}{MS(ABC)}$
BC	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2 + ar\sigma_{BC}^2$	$\frac{MS(BC)}{MS(ABC)}$
ABC	$\sigma_\varepsilon^2 + r\sigma_{ABC}^2$	$\frac{MS(ABC)}{MSE}$
Error	σ_ε^2	

No exact tests exist for the main effects of A, B, or C!

Approximate F tests: synthetic errors

Satterthwaite's method (1946) utilizes **linear combinations of extant mean squares** to build appropriate error terms.

$$MS' = MS_A + MS_B + \dots + MS_N$$

and

$$MS'' = MS_a + MS_b + \dots + MS_n$$

where:

- a. No MS appears simultaneously in MS' and MS''; and
- b. EMS' – EMS'' is equal to the effect considered in the null hypothesis

Under these conditions, the test statistic:

$$F = MS' / MS''$$

is distributed approximately as $F_{p,q}$ where p and q are the **effective degrees of freedom**:

$$p = \frac{(MS_A + \dots + MS_X)^2}{\frac{MS_A^2}{df_A} + \dots + \frac{MS_X^2}{df_X}} \quad \text{and} \quad q = \frac{(MS_a + \dots + MS_n)^2}{\frac{MS_a^2}{df_a} + \dots + \frac{MS_n^2}{df_n}}$$

In these expressions for p and q , df_i is the number of degrees of freedom associated with the mean square MS_i .

Example: In the three-factor mixed effects model discussed above, an appropriate test statistic for $H_0: \tau_{A1} = \dots = \tau_{At} = 0$ would be:

$$F = \frac{MSA + MS(ABC)}{MS(AB) + MS(AC)}$$

$$F = \frac{\sigma_\epsilon^2 + r\sigma_{ABC}^2 + br\sigma_{AC}^2 + cr\sigma_{AB}^2 + bcr \sum \frac{\tau_A^2}{a-1} + \sigma_\epsilon^2 + r\sigma_{ABC}^2}{\sigma_\epsilon^2 + r\sigma_{ABC}^2 + br\sigma_{AC}^2 + cr\sigma_{AB}^2 + \dots + \sigma_\epsilon^2 + r\sigma_{ABC}^2}$$

Nested effects (subsamples)

Recall that subsamples force the experimental units into the linear model, and experimental units (i.e. replications) are random, by definition. So even when all factors are fixed effects, nesting transforms the model into a mixed model.

Consider a two-way factorial (fixed A, random B) with C (random) replications in each A-B combination and D subsamples measured on each replication.

Source	Mixed Model EMS
	Fixed A, Random B, Reps C, Subsamples D
A	$\sigma_\varepsilon^2 + d\sigma_{(AB)C}^2 + cd\sigma_{AB}^2 + bcd \sum \frac{\tau_A^2}{a-1}$
B	$\sigma_\varepsilon^2 + d\sigma_{(AB)C}^2 + cd\sigma_{AB}^2 + acd\sigma_B^2$
(A:B)	$\sigma_\varepsilon^2 + d\sigma_{(AB)C}^2 + cd\sigma_{AB}^2$
(A:B):C	$\sigma_\varepsilon^2 + d\sigma_{(AB)C}^2$
Error	σ_ε^2

1. To test the interaction, we must use the MS of the factor nested within the interaction (i.e. the EU), as with any previous nested design.
2. To test A or B, we must use the interaction ($MS_{A:B}$), also as before.

Blocks nested within random locations

Four eel grass restoration methods (Factor A) are tested in five locations (Factor B), selected at random within the Great Bay. At each location, the experiment is organized as an RCBD with 3 blocks.

Source	Mixed Model Fixed Var, Random Loc
Method	$\sigma_{\varepsilon}^2 + 3\sigma_{AB}^2 + 15 \sum \frac{\tau_A^2}{a-1}$
Loc	$\sigma_{\varepsilon}^2 + 4\sigma_{(B)Block}^2 + 3\sigma_{AB}^2 + 12\sigma_B^2$
Method:Loc	$\sigma_{\varepsilon}^2 + 3\sigma_{AB}^2$
(Loc):Block	$\sigma_{\varepsilon}^2 + 4\sigma_{(B)Block}^2$

The methods need to be tested using the Method:Location interaction variance because we are trying to make generalizations which are valid across all locations.

$$F_{Location} = \frac{MS_{Loc} + MS_{Error}}{MS_{Interaction} + MS_{Loc/Block}}$$

Some typical designs with blocks nested in random "environments"

Model 1

Fixed Genotypes

Environments = Random Locations

Blocks within Locations

Source	df	MS	F	VC
G	g-1	M_G	M_G/M_{GL}	$\sigma_G = (M_G - M_{GL})/rl$
L	l-1	M_L	synthetic	$\sigma_L = (M_L - M_{B(L)} - M_{GL} + M_e)/rg$
Block(Loc)	(r-1)l	$M_{B(L)}$	$M_{B(L)}/M_e$	$\sigma_{B(L)} = (M_{B(L)} - M_e)/g$
GxL	(g-1)(l-1)	M_{GL}	M_{GL}/M_e	$\sigma_{GL} = (M_{GL} - M_e)/r$
Error	(r-1)(g-1)l	M_e	-	$\sigma_e = M_e$

Source	Expected Mean Squares
Geno	$\sigma_e + r\sigma_{GL} + lr \sum \frac{\gamma^2}{g-1}$
Loc	$\sigma_e + g\sigma_{B(L)} + r\sigma_{GL} + gr\sigma_L$
Geno:Loc	$\sigma_e + r\sigma_{GL}$
Block(Loc)	$\sigma_e + g\sigma_{B(L)}$

The varieties need to be tested using the Geno:Loc interaction because we are trying to make generalizations which are valid across all locations.

Model 2

Fixed Genotypes

Environments = Random Locations

Random Years (shared across locations)

Blocks within (Locations x Years)

Source	df	MS	F	VC
G	g-1	M _G	synthetic	
L	l-1	M _L	synthetic	
Y	y-1	M _Y	synthetic	
Block(LxY)	(r-1)l	M _{B(LY)}	-	
GxL	(g-1)(l-1)	M _{GL}	M _{GL} /M _{GLY}	(M _{GL} -M _{GLY})/ry
GxY	(g-1)(l-1)	M _{GY}	M _{GY} /M _{GLY}	(M _{GY} -M _{GLY})/rl
GxLxY	(g-1)(l-1)	M _{GLY}	M _{GLY} /M _e	(M _{GLY} -M _e)/r
Error	(r-1)(g-1)l	M _e	-	M _e

Source	Expected Mean Squares
Geno	$\sigma_e + r\sigma_{GLY} + rl\sigma_{GY} + ry\sigma_{GL} + lyr \sum \frac{\gamma^2}{g-1}$
Loc	$\sigma_e + r\sigma_{GLY} + ry\sigma_{GL} + g\sigma_{B(LY)} + gr\sigma_{LY} + gry\sigma_L$
Year	$\sigma_e + r\sigma_{GLY} + rl\sigma_{GY} + g\sigma_{B(LY)} + gr\sigma_{LY} + grl\sigma_Y$
Block(Loc:Year)	$\sigma_e + g\sigma_{B(LY)}$
Geno:Loc	$\sigma_e + r\sigma_{GLY} + ry\sigma_{GL}$
Geno:Year	$\sigma_e + r\sigma_{GLY} + rl\sigma_{GY}$
Geno:Loc:Year	$\sigma_e + r\sigma_{GLY}$

Model 3

Fixed Genotypes

Environments = Random Locations

Random Years within Locations (years not shared across locations)

Blocks within Years within Locations

Source	df	MS	F	VC
G	g-1	M _G	M _G /M _{GL}	
L	l-1	M _L	synthetic	
Y(L)	y-1	M _{Y(L)}	synthetic	
Block(YxL)	(r-1)l	M _{B(YL)}	-	
GxL	(g-1)(l-1)	M _{GL}	M _{GL} /M _{GLY}	(M _{GL} -M _{GY(L)})/ry
GxY(L)	(g-1)(l-1)	M _{GY(L)}	M _{GY(L)} /M _e	(M _{GY(L)} -M _e)/r
Error	(r-1)(g-1)l	M _e	-	M _e

Source	Expected Mean Squares
Geno	$\sigma_e + r\sigma_{GLY} + ry\sigma_{GL} + lry \sum \frac{\gamma^2}{g-1}$
Loc	$\sigma_e + r\sigma_{GLY} + ry\sigma_{GL} + g\sigma_{B(LY)} + gr\sigma_{Y(L)} + gry\sigma_L$
Year(Loc)	$\sigma_e + r\sigma_{GLY} + g\sigma_{B(LY)} + grl\sigma_{Y(L)}$
Block(Year:Loc)	$\sigma_e + g\sigma_{B(LY)}$
Geno:Loc	$\sigma_e + r\sigma_{GLY} + ry\sigma_{GL}$
Geno:Year(Loc)	$\sigma_e + r\sigma_{GY(L)}$