STAT 3119

Week 15: 12/3/2019 @GWU

Outline

- Additional Design topics: Nest Design (Chapter 26)
- Part II: Final concept review

Nested Factors vs Crossed Factors (Ch 26.1)

- In the factorial studies considered so far, where every level of one factor appears with each level of every other factor, the factors are said to be crossed.
- A different situation occurs when a certain factor is **nested** within other factors.

Example 1

- A large manufacturing company operates three regional training schools for mechanics, one in each of its operating districts.
- The schools have **two** instructors each, who teach classes of about 15 mechanics in three-week sessions. The company was concerned about the effect of **school (factor A)** and **instructor (factor B)** on the learning achieved.
- To investigate these effects, **classes (experimental units)** in each district were formed in the usual way and then randomly assigned to one of the two instructors in the school. This was done for two sessions, and at the end of each session a suitable summary measure of learning for the class was obtained. (Here, *mechanics* is not experimental units that was to be individually randomized, measured and analyzed.)

Data

TABLE 26.1 **Sample Data for Nested Two-Factor Study—Training School Example (class learning scores, coded).**

Factor *A* (school) *i*		Factor *B* (instructor) *j* = 1	Factor *B* (instructor) *j* = 2	Average
Atlanta		25 29	14 11	
	Average	$\overline{Y}_{11\cdot} = 27$	$\overline{Y}_{12\cdot} = 12.5$	$\overline{Y}_{1\cdot\cdot} = 19.75$
Chicago		11 6	22 18	
	Average	$\overline{Y}_{21\cdot} = 8.5$	$\overline{Y}_{22\cdot} = 20$	$\overline{Y}_{2\cdot\cdot} = 14.25$
San Francisco		17 20	5 2	
	Average	$\overline{Y}_{31\cdot} = 18.5$	$\overline{Y}_{32\cdot} = 3.5$	$\overline{Y}_{3\cdot\cdot} = 11.00$
			Average	$\overline{Y}_{\cdot\cdot\cdot} = 15$

Note: The layout of Table 26.1 *appears* identical to an ordinary two-factor investigation, with two observations per cell. However, it is not a factorial design but a nested design because instructors in the Atlanta school did not also teach in the other two schools, thus six different instructors were involved (2 in each school).

FIGURE 26.1 **Illustration of Crossed and Nested Factors—Training School Example.**

(a) Crossed Factors

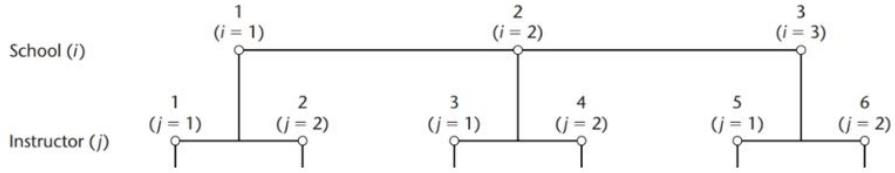
School (factor A)	Instructor (factor B)					
	1	2	3	4	5	6
Atlanta						
Chicago						
San Francisco						

An ordinary two-factor investigation with six different instructors would have consisted of 18 treatments. (Every factor level of B appears with every factor level of A, factors A and B are said to be crossed.)

(b) Nested Factors

School (factor A)	Instructor (factor B)					
	1	2	3	4	5	6
Atlanta						
Chicago						
San Francisco						

In the training school example, however, only six treatments were included. Factor B is nested within factor A.



In this example, if schools are randomly chosen from many schools or each school chose two instructors

randomly from a pool of many instructors, then school or instructor or both could be *random* factor(s) as we discussed in chapter 25.

Visualize the Example 1 data

1. read the data

```
Training  =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week-15/CH26TA01.txt"))
names(Training) =  c("Response", "School", "Instructors", "Units")

# make  categorical variables for factor A and B
Training$School =  as.factor(Training$School)
Training$Instructors =  as.factor(Training$Instructors)

str(Training)
```

```
## 'data.frame':    12 obs. of  4 variables:
##  $ Response   : int  25 29 14 11 11 6 22 18 17 20 ...
##  $ School     : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 2 2 2 3 3 ...
##  $ Instructors: Factor w/ 2 levels "1","2": 1 1 2 2 1 1 2 2 1 1 ...
##  $ Units      : int  1 2 1 2 1 2 1 2 1 2 ...
```

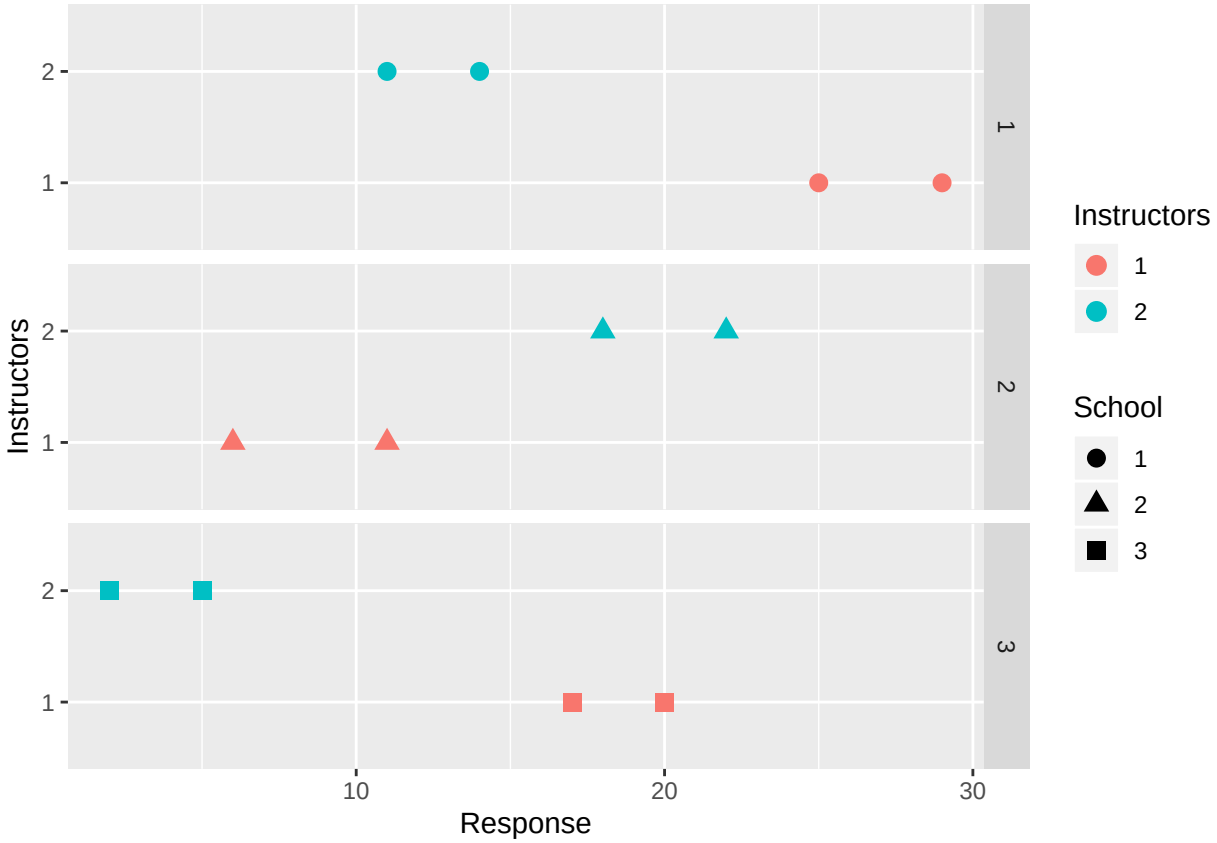
```
Training
```

```
##    Response School Instructors Units
## 1        25      1           1     1
## 2        29      1           1     2
## 3        14      1           2     1
## 4        11      1           2     2
## 5        11      2           1     1
## 6         6      2           1     2
## 7        22      2           2     1
## 8        18      2           2     2
## 9        17      3           1     1
## 10       20      3           1     2
## 11        5      3           2     1
## 12        2      3           2     2
```

Note: The teachers at 3 schools are different, although they are all labelled at "1" and "2".

2. plot the data

```
library(ggplot2)
ggplot(Training, aes(y = Instructors, x = Response )) +
     geom_point( aes(color= Instructors , shape=School), size=3) +
     facet_grid(School ~ .)
```



From the plot, it appears the responses (measurements of training) vary by both school and instructors.

Nested Design: Example 2

Pastes data set in *lme4* package. The strength of a chemical paste product was measured for a total of 60 samples coming from 10 randomly selected delivery batches (factor A) each containing 3 randomly selected casks (factor B). Hence, two samples were taken from each cask. We want to check what part of the variability of strength is due to batch and cask.

```
data("Pastes", package = "lme4")
str(Pastes)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ strength: num  62.8 62.6 60.1 62.3 62.7 63.1 60 61.4 57.5 56.9 ...
##  $ batch   : Factor w/ 10 levels "A","B","C","D",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ cask    : Factor w/ 3 levels "a","b","c": 1 1 2 2 3 3 1 1 2 2 ...
##  $ sample  : Factor w/ 30 levels "A:a","A:b","A:c",..: 1 1 2 2 3 3 4 4 5 5 ...
```

```
xtabs(~ batch + cask, data=Pastes)
```

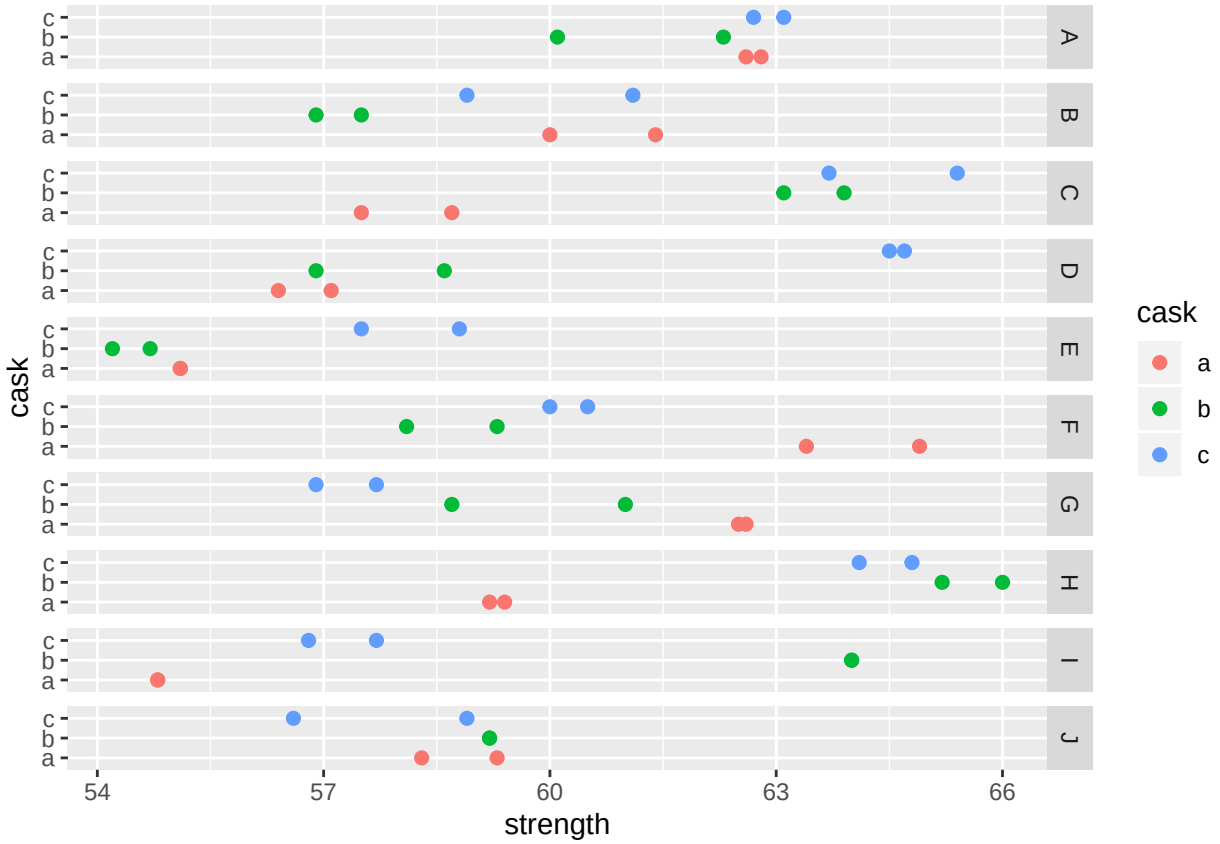
```
##      cask
## batch a b c
##     A 2 2 2
##     B 2 2 2
```

```
##      C 2 2 2
##      D 2 2 2
##      E 2 2 2
##      F 2 2 2
##      G 2 2 2
##      H 2 2 2
##      I 2 2 2
##      J 2 2 2
```

Note: If we carefully think about the data structure: For each batch, different cask were used. Cask 1 in batch 1 has nothing to do with cask 1 in batch 2 and so on. The "1" of cask has a different meaning for every batch. Hence, cask and batch are not crossed. We say cask is **nested** in batch.

Visualize the Example 2 data

```
ggplot(Pastes, aes(y = cask, x = strength)) +
   geom_point(aes(color= cask), size=2) + facet_grid(batch ~ .)
```



From the plot, it appears there are lots of variation due to cask within batch.

Two-factor ANOVA Models for nested design (Ch 26.2):

Fixed effect model

- Let Y_{ijk} denote the response for the kth trial when factor A is at the ith level and factor B is at the jth level.

- We assume that there are n replications for each factor level combination, i.e., $k = 1, ..., n$, and that $i = 1, ..., a$ and $j = 1, ..., b$ (balanced design: the same number of factor B levels is nested within each factor A level and the number of replications is the same throughout).

When both factors A and B have fixed effects, the nested design model is:

$$Y_{ijk} = \mu.. + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

where:

$\mu..$ is a constant
α_i are constants subject to the restriction $\sum \alpha_i = 0$
$\beta_{j(i)}$ are constants subject to the restrictions $\sum_j \beta_{j(i)} = 0$ for all i
ε_{ijk} are independent $N(0, \sigma^2)$
$i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n$

The expected value and variance of observation Y_{ijk} for nested design model (26.7) with fixed factor effects are:

$$E\{Y_{ijk}\} = \mu.. + \alpha_i + \beta_{j(i)} \quad \textbf{(26.8a)}$$

$$\sigma^2\{Y_{ijk}\} = \sigma^2 \quad \textbf{(26.8b)}$$

Thus, all observations have a constant variance. Further, the observations Y_{ijk} are independent and normally distributed for this model.

Note: Since the different levels of factor B are nested within factor A. It would be it would be meaningless to consider the effect of the jth level, averaged over all levels of factor A. Instead, the individual effects of each level of factor B within factor A need to be considered. We therefore denote the individual effects by $\beta_{j(i)}$ as he jth factor level of B is nested within the ith factor level of A. There is no need to include interaction term in the model since factor B is nested within factor A, not crossed with it.

Random Factor Effects

If both factors A and B have random factor levels, nested design model (26.7) is modified with α_i, $\beta_{j(i)}$ and ϵ_{ijk} being independent normal random variables with mean 0 and variances σ^2_α , σ^2_β and σ^2, respectively.

ANOVA analysis for two-factor nested design: Fixed effects model (Ch 26.3)

Model fitting and estimation of parameters

Parameter	Estimator	
$\mu..$	$\hat{\mu}.. = \overline{Y}...$	(26.9a)
α_i	$\hat{\alpha}_i = \overline{Y}_{i}.. - \overline{Y}...$	(26.9b)
$\beta_{j(i)}$	$\hat{\beta}_{j(i)} = \overline{Y}_{ij}. - \overline{Y}_{i}..$	(26.9c)

The fitted values therefore are:

$$\hat{Y}_{ijk} = \overline{Y}... + (\overline{Y}_{i}.. - \overline{Y}...) + (\overline{Y}_{ij}. - \overline{Y}_{i}..) = \overline{Y}_{ij}. \quad (26.10)$$

and the residuals are:

$$e_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - \overline{Y}_{ij}. \quad (26.11)$$

ANOVA table

TABLE 26.3 ANOVA Table for Nested Balanced Two-Factor Fixed Effects Model (26.7) (B nested within A).

Source of Variation	SS	df	MS	$E\{MS\}$	Test
Factor A	$SSA = bn\sum(\overline{Y}_{i}.. - \overline{Y}...)^2$	$a-1$	MSA	$\sigma^2 + bn\frac{\sum\alpha_i^2}{a-1}$	$\frac{MSA}{MSE}$ ~ $F(a\text{-}1, ab(n\text{-}1))$
Factor B (within A)	$SSB(A) = n\sum\sum(\overline{Y}_{ij}. - \overline{Y}_{i}..)^2$	$a(b-1)$	$MSB(A)$	$\sigma^2 + n\frac{\sum\sum\beta_{j(i)}^2}{a(b-1)}$	$\frac{MSB(A)}{MSE}$ ~ $F(a(b\text{-}1), ab(n\text{-}1))$
Error	$SSE = \sum\sum\sum(Y_{ijk} - \overline{Y}_{ij}.)^2$	$ab(n-1)$	MSE	σ^2	
Total	$SSTO = \sum\sum\sum(Y_{ijk} - \overline{Y}...)^2$	$abn-1$			

R: ANOVA analysis of Example 1

```
# note to use A/B to specify the nesting
Nestfit = aov(Response ~ School/Instructors, data=Training)
summary(Nestfit)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## School               2  156.5   78.25   11.18 0.009473 **
## School:Instructors   3  567.5  189.17   27.02 0.000697 ***
## Residuals            6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results (as p.1098):

1. For to determine whether or not main school effects exist, $F^* = 11.18$ with $p = 0.0094$, therefore we reject H_0 and conclude that the three schools differ in mean learning effects.
2. For the test for differences in mean learning effects between instructors within each school, $F^* = 27.02$ with $p = 0.007$, we reject H_0 and conclude that instructors within at least one school differ in terms of mean learning effects.

ANOVA analysis for two-factor nested design: random/mixed effects model (Ch 26.3)

ANOVA Table

TABLE 26.5 Expected Mean Squares for Nested Balanced Two-Factor Designs with Random Factor Effects (*B* nested within *A*).

	Expected Mean Square	
Mean Square	***A* Fixed, *B* Random**	***A* Random, *B* Random**
MSA	$\sigma^2 + bn\frac{\sum\alpha_i^2}{a-1} + n\sigma_\beta^2$	$\sigma^2 + bn\sigma_\alpha^2 + n\sigma_\beta^2$
MSB(A)	$\sigma^2 + n\sigma_\beta^2$	$\sigma^2 + n\sigma_\beta^2$
MSE	σ^2	σ^2
	Appropriate Test Statistic	
Test for	***A* Fixed, *B* Random**	***A* Random, *B* Random**
Factor *A*	*MSA/MSB(A)*	*MSA/MSB(A)*
Factor *B(A)*	*MSB(A)/MSE*	*MSB(A)/MSE*

Test statistic (26.17b) for factor A main effects is not appropriate if either or both factor effects are random. From the EMS, the appropriate F test uses $MSB(A)$ as the denominator.

Model fitting and estimation

We will use software such as R functions **lmer** to estimate the model parameter and variance component based on maximum likelihood method. The procedure to set up the fixed or random factors are similar as we discussed in the example for Chapter 25.

R: ANOVA analysis of Example 2:

Both batch and cask are random , and cask is nested within batch.

Step1: testing effect of A

```
Nestfit2 = aov(strength ~ batch/cask, data=Pastes)
summary(Nestfit2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## batch        9  247.4  27.489   40.55 2.28e-14 ***
## batch:cask  20  350.9  17.545   25.88 9.79e-14 ***
## Residuals   30   20.3   0.678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore

```
MSA = 40.55
MSBnA =  25.88
(Fstat= MSA/MSBnA)
```

```
## [1] 1.566847
```

```
(Pv= 1- pf(Fstat, 9,20))
```

```
## [1] 0.1925247
```

Results: From the ANOVA table

1. To test factor A: we use $F = MSA/MSB(A) = 1.57$, $p = 0.19$ and we fail to reject H_0 (no batch differences).
2. To test factor B(A): the $F = MSB(A)/MSE = 25.88$ with $p < 0.0001$ and we reject H_0 (no cask effects).

Step 2: estimate model parameters

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
# use batch/cask to specify the nested factor
fit.paste <- lmer(strength ~ (1 | batch/cask), data = Pastes)
summary(fit.paste)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: strength ~ (1 | batch/cask)
##    Data: Pastes
##
## REML criterion at convergence: 247
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4798 -0.5156  0.0095  0.4720  1.3897
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  cask:batch (Intercept) 8.434    2.9041
##  batch      (Intercept) 1.657    1.2874
##  Residual               0.678    0.8234
## Number of obs: 60, groups:  cask:batch, 30; batch, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  60.0533     0.6769   88.72
```

Results: From model output, the variance components are 1.657 due to batch, 8.434 due to cask, and 0.678 due to residual errors, respectively. Therefore most variation is due to cask within batch.

Summary

- Reading: Briefly for Chapter 26.1-26.3
- Reminder: Homework 11 and Project due 12/5
- Final exam is Dec 12 here: 7:40 -9:40pm