2.11 The Random Effects Model

- So far we have assumed the factor levels were fixed. That is, the factor levels were set at fixed levels by the experimenter.
- In many one-factor CRDs, the a factor levels are *randomly selected from a population* of levels. For this type of experiment, the factor is called a **random factor**, and the associated effects are called **random effects**.
- In theory, for random effects, we assume the population is infinite. In practice, it is acceptable if the number of randomly selected factor levels (a) is small relative to the number of levels in the population (N). In general, we want $a/N < .10$ (or, $< 10\%$).
- The **random effects model** for a one-factor CRD is:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{3}$$

  where both τ_i and ϵ_{ij} are random variables. The model assumptions are

  - The ϵ_{ij}'s are IID $N(0, \sigma^2)$ and the τ_i's are IID $N(0, \sigma_\tau^2)$.
  - τ_i and ϵ_{ij} are independent for all i, j.

  Note: IID means 'independent and identically distributed'.

- τ_i is a random variable with $\text{Var}(\tau_i) = \sigma_\tau^2$. σ_τ^2 is the variance associated with the distribution or population of all τ_i's. We assume τ_i is independent of the random error ϵ_{ij} which has variance σ^2.
- The variances σ_τ^2 and σ^2 are called **variance components**.
- The hypotheses of interest involve the variance component σ_τ^2:

  and

- If $\sigma_\tau^2 = 0$, then the random τ_i effects are identically 0. In this case, the variability of the $\widehat{\tau}_i$ estimates ($i = 1, 2, \ldots, a$) should be close to 0 relative to the MSE.
- If $\sigma_\tau^2 > 0$, then the random τ_i effects are not identically 0. In this case, the variability of the $\widehat{\tau}_i$ estimates ($i = 1, 2, \ldots, a$) should be large relative to the MSE.
- Testing hypotheses about the equality of means is meaningless in the random effects case. Therefore, we do not perform a multiple comparison procedure to compare means.
- The ANOVA table for a random factor is the same as the ANOVA table for a fixed factor with $SS_T = SS_{trt} + SS_E$.
- To see this we need to look at the expected mean squares for the random effects model in (3).

2.11.1 Expected Mean Squares

- Theoretically, the expected values of MS_E and MS_{trt} are

$$E(MS_E) = \sigma^2 \tag{4}$$
$$E(MS_{trt}) = \sigma^2 + \tag{5}$$

  where $n_0 = \dfrac{1}{a-1}\left[N - \dfrac{\sum_{i=1}^{a} n_i^2}{N}\right]$.

- If all of the sample sizes are equal ($n_i = n$ for $i = 1, 2, \ldots, a$), this reduces to $n_0 = n$.

- If $H_0 : \sigma_\tau^2 = 0$ is true, then $\quad E(MS_{trt}) \;=\; \sigma^2 + 0 \;=\; \sigma^2$.
- If $H_0 : \sigma_\tau^2 = 0$ is false, then $\sigma_\tau^2 > 0$. This implies

$$E(MS_{trt}) \;=\; \sigma^2 + (\textit{positive quantity}) \;\longrightarrow$$

- As the variability among the τ_i's increases, the F-ratio

$$F = \frac{E(MS_{trt})}{E(MS_E)} \;=$$

  also increases.
- Again we use $\quad F_0 = \dfrac{MS_{trt}}{MS_E}$ $\quad$ to estimate F. This implies that $H_0 : \sigma_\tau^2 = 0$ is rejected for large values of F_0.
- The ANOVA table computations for the random effects model are identical to the ANOVA table computations for the fixed effects model. **The hypotheses being tested are different.**
- Remember: The hypotheses for the random effects model apply to a distribution or population (the variance component σ_τ^2) while the hypotheses for a fixed effects model apply to equality of fixed treatment effects (τ_i's) or means (μ_i's).

2.11.2 Estimation of Variance Components

- For the random effects model, we are usually interested in estimates $\widehat{\sigma}^2$ and $\widehat{\sigma}_\tau^2$ of the variance components σ^2 and σ_τ^2.

  If we replace $E(MS_E)$ with MS_E in equation (4), we get

$$\widehat{\sigma}^2 =$$

  If we replace σ^2 with $\widehat{\sigma}^2 = MS_E$ and $E(MS_{trt})$ with MS_{trt} in equation (5), we get

$$MS_{trt} \;\approx$$

  Solving for σ_τ^2 gives

$$\widehat{\sigma}_\tau^2 \;=$$

- This estimation approach can give a negative estimate of σ_τ^2 $\;(\widehat{\sigma}_\tau^2 < 0)$. But we know that a variance component cannot be negative. The following are 3 possible ways to handle this situation:
  1. Assume $\sigma_\tau^2 = 0$ and the negative estimate occurred due to random sampling. The problem is that using zero instead of a negative number can affect other estimates.
  2. Estimate σ_τ^2 using the REML (restricted maximum likelihood) method because it always yields a nonnegative estimate. This method will adjust other variance component estimates. REML methods are included in *SAS*.
  3. Assume the model is incorrect, and examine the problem in another way. For example, add or remove an effect from the model, and then analyze the new model.

2.11.3 A Random Effects Example

A company supplies a customer with a large number of batches of raw materials used in a chemical production process. The customer wants a high percentage of usable chemical to be produced from the raw material. The customer is concerned that there may be significant variation among the batches (which is not good for the production process). An experiment was run. Five random batches are selected, and three random samples are taken from each batch. The response is 'percent usable chemical'. The experimental data are

Batch				
1	2	3	4	5
74	68	75	72	79
76	71	77	74	81
75	72	77	73	79

- The model is $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ with

$$\epsilon_{ij} \sim IIDN(0, \sigma^2) \qquad \tau_i \sim IIDN(0, \sigma_\tau^2)$$

- The ANOVA table on the next page indicates the p-value $< .0001$ for testing

$$H_0 : \sigma_\tau^2 = 0 \qquad \text{and} \qquad H_1 : \sigma_\tau^2 > 0.$$

  Therefore we **reject** $H_0 : \sigma_\tau^2 = 0$, and conclude that there is significant variability in percent usable chemical in the population of batches.

- The estimates of the variance components are

$$\widehat{\sigma}^2 \; = \; MSE \; = \qquad\qquad \widehat{\sigma}_\tau^2 \; = \; \frac{MS_{batch} - MS_E}{n} \; =$$

- In the SAS code, include a RANDOM statement to perform the F-test and output the expected mean squares. Use the VARCOMP procedure to generate estimates of the variance components.

ANOVA WITH RANDOM BATCH EFFECTS

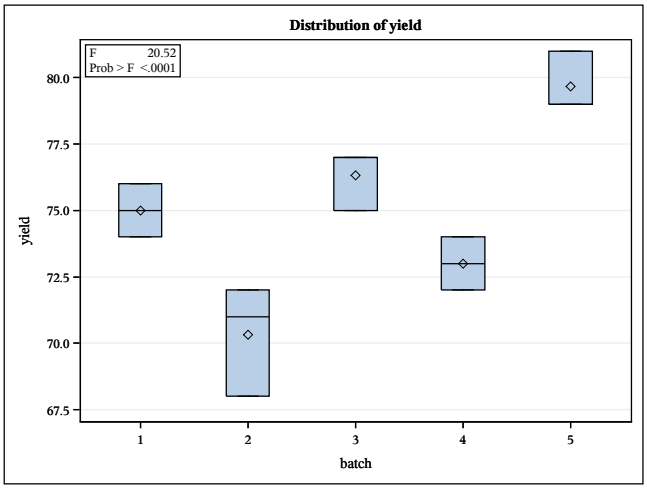
The GLM Procedure

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	147.7333333	36.9333333	20.52	<.0001
Error	10	18.0000000	1.8000000		
Corrected Total	14	165.7333333			

R-Square	Coeff Var	Root MSE	yield Mean
0.891392	1.792040	1.341641	74.86667

Source	DF	Type III SS	Mean Square	F Value	Pr > F
batch	4	147.7333333	36.9333333	20.52	<.0001



Level of batch	N	yield	
		Mean	Std Dev
1	3	75.0000000	1.00000000
2	3	70.3333333	2.08166600
3	3	76.3333333	1.15470054
4	3	73.0000000	1.00000000
5	3	79.6666667	1.15470054

ANOVA WITH RANDOM BATCH EFFECTS

The GLM Procedure

Source	Type III Expected Mean Square
batch	Var(Error) + 3 Var(batch)

The GLM Procedure
Tests of Hypotheses for Random Model Analysis of Variance

Dependent Variable: yield

Source	DF	Type III SS	Mean Square	F Value	Pr > F
batch	4	147.733333	36.933333	20.52	<.0001
Error: MS(Error)	10	18.000000	1.800000		

Variance Components Estimation Procedure

Class Level Information		
Class	Levels	Values
batch	5	1 2 3 4 5

Number of Observations Read	15
Number of Observations Used	15

Dependent Variable:	yield

REML Iterations			
Iteration	Objective	Var(batch)	Var(Error)
0	20.3143245916	11.7111111111	1.8000000000
1	20.3143245916	11.7111111111	1.8000000000

Convergence criteria met.

REML Estimates	
Variance Component	Estimate
Var(batch)	11.71111
Var(Error)	1.80000

SAS code for random effects analysis

```
DM 'LOG; CLEAR; OUT; CLEAR;';

ODS GRAPHICS ON;
ODS PRINTER PDF file='C:\COURSES\ST541\RANDOM.PDF';
OPTIONS NODATE NONUMBER;

***********************************************;
*** A COMPLETELY RANDOMIZED ONE-FACTOR DESIGN ***;
***     WITH RANDOM BATCH EFFECTS             ***;
***********************************************;

TITLE 'ANOVA WITH RANDOM BATCH EFFECTS';

DATA in;
  DO batch = 1 TO 5;
  DO rep = 1 TO 3;
     INPUT yield @@; OUTPUT;
  END; END;
LINES;
 74 76 75   68 71 72   75 77 77    72 74 73    79 81 79
 ;
PROC GLM DATA=IN;
     CLASS batch;
     MODEL yield = batch / SS3;
     MEANS batch;
     RANDOM batch / TEST;
     OUTPUT OUT=diag P=pred R=resid;

PROC VARCOMP DATA=in METHOD=REML;
     CLASS batch;
     MODEL yield = batch / FIXED=0;
TITLE2 'VARIANCE COMPONENT ESTIMATION';
RUN;

ODS GRAPHICS OFF;
RUN;
```

R code for random effects analysis

```
library(lme4)
batch <- c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5)
yield <- c(74,76,75,68,71,72,75,77,77,72,74,73,79,81,79)

# Version 1
randaov <- aov(yield~factor(batch))
summary(randaov)
raneff <- lmer(yield~1+(1|factor(batch)))
summary(raneff)

# Version 2
library(nlme)
raneff2 <- lme(yield~1,random=~1|factor(batch))
summary(raneff2)
```

R output for random effects analysis

```
> # Version 1

            Df Sum Sq Mean Sq F value    Pr(>F)
batch        4 147.73  36.933  20.518 8.246e-05 ***
Residuals   10  18.00   1.800
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Linear mixed model fit by REML
Formula: yield ~ 1 + (1 | batch)

Random effects:
 Groups   Name        Variance Std.Dev.
 batch    (Intercept) 11.711   3.4221  <-- Variance component estimates and
 Residual              1.800   1.3416  <-- their square roots

Number of obs: 15, groups: batch, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   74.867      1.569   47.71

--------------------------------------------------------------------------
--

# Version 2

Linear mixed-effects model fit by REML

 Data: aovdata
       AIC      BIC    logLik
  68.75265 70.66983 -31.37633

Random effects:

 Formula: ~1 | batch
        (Intercept) Residual
StdDev:     3.42215 1.341641     <-- Square roots of variance component
                                     estimates
Fixed effects: yield ~ 1
               Value Std.Error DF t-value p-value
(Intercept) 74.86667  1.569147 10 47.7117       0    <-- for mean only

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-1.903841911 -0.531525823  0.004843469  0.613864789  1.168172858

Number of Observations: 15
Number of Groups: 5
```

2.11.4 Confidence Intervals for Variance Components

Given the normality and independence assumptions of the random effects model, we can generate various confidence intervals related to the variance components.

1. For σ^2: Because $\dfrac{SS_E}{\sigma^2} = \dfrac{(N-a)MS_E}{\sigma^2} \sim \chi^2_{N-a}$, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\leq$$

2. For $\dfrac{\sigma^2_\tau}{\sigma^2}$: Because MS_E and MS_{trt} are independent,

$$\frac{MS_{trt}/(\sigma^2 + n_0\sigma^2_\tau)}{MS_E/\sigma^2} \;=\; F_0\left(\frac{\sigma^2}{\sigma^2 + n_0\sigma^2_\tau}\right) \;\sim\; F(a-1, N-a).$$

Let $F_L = F_{1-\alpha/2}(a-1, N-a)$ and $F_U = F_{\alpha/2}(a-1, N-a)$. Then,

$$\begin{aligned}
1-\alpha &= P\left[F_L \le F_0\left(\frac{\sigma^2}{\sigma^2 + n_0\sigma^2_\tau}\right) \le F_U\right] \\
&= P\left[\frac{1}{F_L} \ge \frac{1}{F_0}\left(\frac{\sigma^2 + n_0\sigma^2_\tau}{\sigma^2}\right) \ge \frac{1}{F_U}\right] \;=\; P\left[\frac{1}{F_U} \le \frac{1}{F_0}\left(\frac{\sigma^2 + n_0\sigma^2_\tau}{\sigma^2}\right) \le \frac{1}{F_L}\right] \\
&= P\left[\frac{1}{F_U} \le \frac{1}{F_0}\left(1 + \frac{n_0\sigma^2_\tau}{\sigma^2}\right) \le \frac{1}{F_L}\right] \\
&= P\left[\frac{1}{n_0}\left(\frac{F_0}{F_U} - 1\right) \le \frac{\sigma^2_\tau}{\sigma^2} \le \frac{1}{n_0}\left(\frac{F_0}{F_L} - 1\right)\right]
\end{aligned}$$

Thus, a $100(1-\alpha)\%$ confidence interval for σ^2_τ/σ^2 is (L, U) where

$$L \;=\; \qquad\qquad \text{and} \qquad U \;=$$

3. For $\dfrac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2}$: Note that

$$\begin{aligned}
1-\alpha &= P[\,L \le \sigma^2_\tau/\sigma^2 \le U\,] \;=\; P[\,1+L \le 1 + \sigma^2_\tau/\sigma^2 \le 1+U\,] \\
&= P[\,1+L \le \frac{\sigma^2 + \sigma^2_\tau}{\sigma^2} \le 1+U\,] \;=\; P\left[\frac{1}{1+L} \ge \frac{\sigma^2}{\sigma^2 + \sigma^2_\tau} \ge \frac{1}{1+U}\right] \\
&= P\left[1 - \frac{1}{1+L} \le 1 - \frac{\sigma^2}{\sigma^2 + \sigma^2_\tau} \le 1 - \frac{1}{1+U}\right] \\
&= P\left[\frac{L}{1+L} \le \frac{\sigma^2_\tau}{\sigma^2 + \sigma^2_\tau} \le \frac{U}{1+U}\right]
\end{aligned}$$

Thus, $\left(\dfrac{L}{1+L}, \dfrac{U}{1+U}\right)$ is a $100(1-\alpha)\%$ confidence interval for $\dfrac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2}$ which represents the proportion of the total variability attributable to the variability among the treatments.

4. For σ^2_τ: There is no closed form for an exact confidence interval for σ^2_τ. The following formula produces an approximate confidence interval.

   (L_τ, U_τ) is an approximate $100(1-\alpha)\%$ confidence interval for σ^2_τ where

$$L_\tau = \frac{SS_{trt}\left(1 - \frac{F_U}{F_0}\right)}{n_0 \chi^2_{\alpha/2,a-1}} \qquad \text{and} \qquad U_\tau = \frac{SS_{trt}\left(1 - \frac{F_L}{F_0}\right)}{n_0 \chi^2_{1-\alpha/2,a-1}}$$

Example of 95% Confidence Intervals for Variance Components

Dependent Variable:	yield

Type 1 Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**
batch	4	147.733333	36.933333	Var(Error) + 3 Var(batch)
Error	10	18.000000	1.800000	Var(Error)
Corrected Total	14	165.733333		

Type 1 Estimates			
Variance Component	**Estimate**	**95% Confidence Limits**	
Var(batch)	11.71111	3.77313	101.00296
Var(Error)	1.80000	0.87877	5.54363

- **We are 95% confident that the observed variability attributable to random (replication) error is between .8788 and 5.5436.**
- **We are 95% confident that the ratio of variance attributable to differences in batches to the variance attributable to random error is between 1.197 and 60.156.**
- **We are 95% confident that the proportion of variability of an observation attributable to differences in batches is between .5448 and .9945.**
- **We are 95% confident that the observed variability attributable to random batch-to-batch variability is between 3.773 and 101.003.**

In this example, $SS_E = 18.0$, $SS_{trt} = 147.7\overline{3}$, $F_0 = 20.51852$, and $n_0 = n = 3$, with

- $\chi^2(.975, 10) = 3.2470$ and $\chi^2(.025, 10) = 20.4832$
- $F_L = F(.975, 4, 10) = 0.11307$ and $F_U = F(.075, 4, 10) = 4.46834$
- $\chi^2(.975, 4) = 0.4844$ and $\chi^2(.025, 4) = 11.1433$