# LINEAR MIXED-EFFECT MODELS

- Studies / data / models seen previously in 511 assumed a single source of "error" variation
- $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$.
  $\beta$ are fixed constants (in the frequentist approach to inference)
  $\epsilon$ is the only random effect
- What if there are multiple sources of "error" variation?

# Examples: Subsampling

- Seedling weight in 2 genotype study from Aitken model section.
- Seedling weight measured on each seedling.
- Two (potential) sources of variation:
  among flats and among seedlings within a flat.

$$\begin{aligned} Y_{ijk} &= \mu + \gamma_i + T_{ij} + \epsilon_{ijk} \\ T_{ij} &\sim N(0, \sigma_F^2) \\ \epsilon_{ijk} &\sim N(0, \sigma_e^2) \end{aligned}$$

  where $i$ indexes genotype, $j$ indexes flat within genotype, and $k$ indexes seedling within flat

- $\sigma_F^2$ quantifies variation among flats,
  if have perfect knowledge of the seedlings in them
- $\sigma_e^2$ quantifies variation among seedlings

# Examples: Split plot experimental design

- Influence of two factors: temperature and a catalyst on fermentation of dry distillers grain (byproduct of EtOH production from corn)
  - Response is $CO_2$ production, measured in a tube
  - Catalyst (+/-) randomly assigned to tubes
  - Temperature randomly assigned to growth chambers
    6 tubes per growth chamber, 3 +catalyst, 3 -catalyst
    all 6 at same temperature
  - Use 6 growth chambers, 2 at each temperature
- The o.u. is a tube. What is the experimental unit?

# Examples: Split plot experimental design - 2

- Two sizes of experimental unit: tube and temperature

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \gamma_{ik} + \beta_j + \alpha\beta_{ij} + \epsilon_{ijkl} \\ \gamma_{ik} &\sim N(0, \sigma^2_G) \text{ Var among growth chambers} \\ \epsilon_{ijkl} &\sim N(0, \sigma^2) \text{ Var among tubes} \end{aligned}$$

  where $i$ indexes temperature, $j$ indexes g.c. within temp., $k$ indexes catalyst, and $l$ indexes tube within temp., g.c., and catalyst

- $\sigma^2_G$ quantifies variation among g.c.,
  if have perfect knowledge of the tubes in them
- $\sigma^2$ quantifies variation among tubes

# Examples: Gauge R&R study

- Want to quantify repeatability and reproducibility of a measurement process
  - Repeatability: variation in meas. taken by a single person or instrument on the same item.
  - Reproducibility: variation in meas. made by different people (or different labs), if perfect repeatability
- One possible design:
  10 parts, each measured twice by 10 operators. 200 obs.

$$
\begin{aligned}
Y_{ijk} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \\
\alpha_i &\sim N(0, \sigma_P^2) \text{ Var among parts} \\
\beta_j &\sim N(0, \sigma_R^2) \text{ Reproducibility variance} \\
\alpha\beta_{ij} &\sim N(0, \sigma_I^2) \text{ Interaction variance} \\
\epsilon_{ijk} &\sim N(0, \sigma^2) \text{ Repeatability variance}
\end{aligned}
$$

# Linear Mixed Effects model

- $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon$
- $\boldsymbol{X}$ $n \times p$ matrix of known constants $\beta \in I\!R^p$ an unknown parameter vector
- $\boldsymbol{Z}$ $n \times q$ matrix of known constants $\boldsymbol{u}$ $q \times 1$ random vector
- $\epsilon$ $n \times 1$ vector of random errors
- The elements of $\beta$ are considered to be non-random and are called ”fixed effects.”
- The elements of $\boldsymbol{u}$ are called ”random effects”
- The errors are always a random effect

# Review of crossing and nesting

- Seedling weight study:
  seedling effects nested in flats
- DDG study:
  temperature effects crossed with catalyst effects
  tube effects nested within growth chamber effects
- Gauge R&R study:
  part effects crossed with operator effects
  measurement effects nested within part$\times$operator
- fixed effects are usually crossed, rarely nested
- random effects are commonly nested, sometimes crossed

- Because the model includes both fixed and random effects (in addition to the residual error), it is called a "mixed-effects" model or, more simply, a "mixed" model.
- The model is called a "linear" mixed-effects model because (as we will soon see) $E(\boldsymbol{y}|\boldsymbol{u}) = X\beta + Z\boldsymbol{u}$, a linear function of fixed and random effects.
- The usual LME assumes that

$$\begin{aligned} E(\epsilon) &= \mathbf{0} \qquad Var(\epsilon) = \boldsymbol{R} \\ E(\boldsymbol{u}) &= \mathbf{0} \qquad Var(\boldsymbol{u}) = \boldsymbol{G} \\ Cov(\epsilon,\ \boldsymbol{u}) &= 0. \end{aligned}$$

- It follows that:

$$\begin{aligned} \mathrm{E}\boldsymbol{y} &= \boldsymbol{X}\beta \\ \mathrm{Var}\boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} \end{aligned}$$

- The details:

$$\begin{aligned}
\mathrm{E}\boldsymbol{y} &= \mathrm{E}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon) \\
&= \boldsymbol{X}\beta + \boldsymbol{Z}\mathrm{E}(\boldsymbol{u}) + \mathrm{E}(\epsilon) \\
&= \boldsymbol{X}\beta \text{ and} \\
\mathrm{Var}\boldsymbol{y} &= \mathrm{Var}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon) \\
&= \mathrm{Var}(\boldsymbol{Z}\boldsymbol{u} + \epsilon) \\
&= \mathrm{Var}(\boldsymbol{Z}\boldsymbol{u}) + \mathrm{Var}(\epsilon) \\
&= \boldsymbol{Z}\mathrm{Var}(\boldsymbol{u})\boldsymbol{Z}' + \boldsymbol{R} \\
&= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} \equiv \Sigma
\end{aligned}$$

- The conditional moments, given the random effects, are:

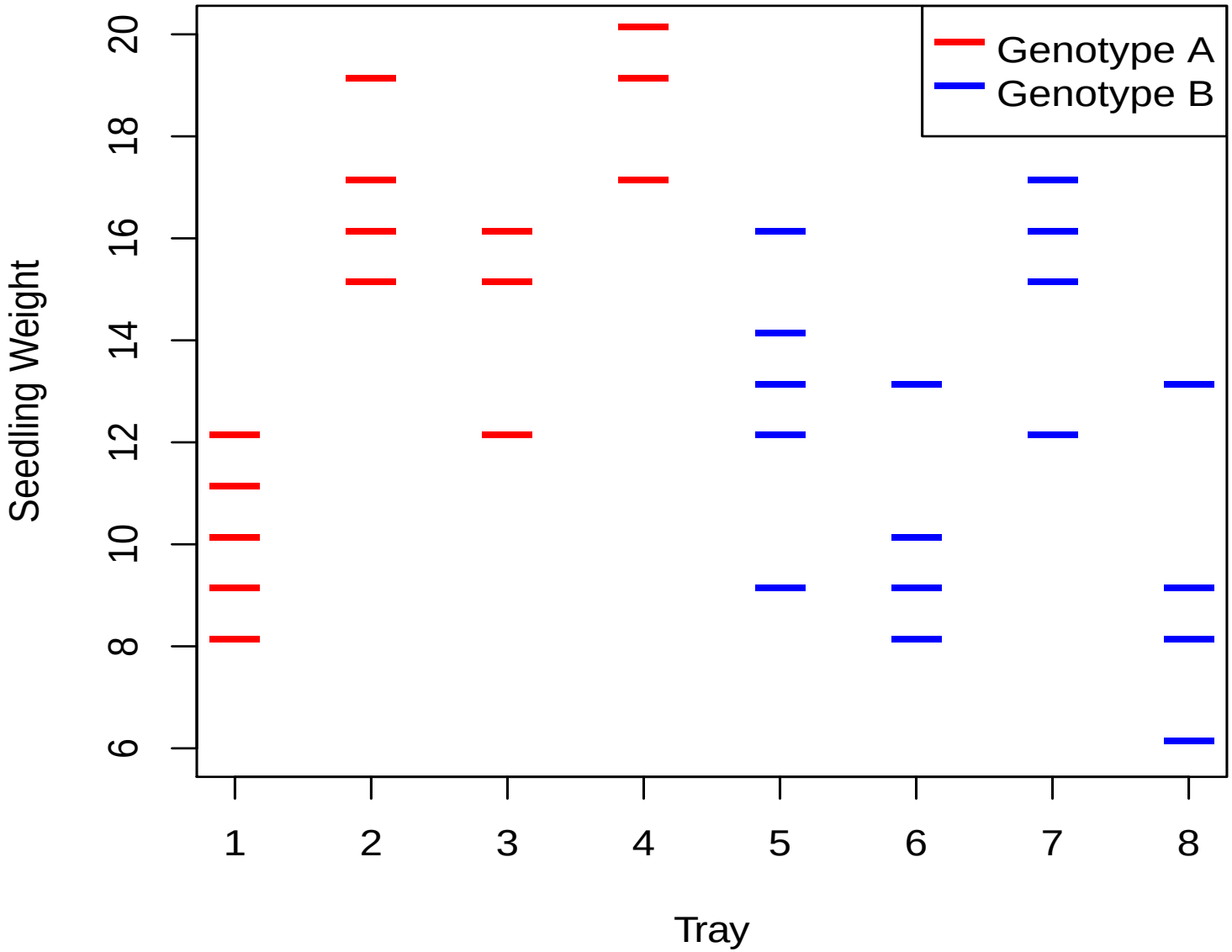
$$\begin{aligned}
\mathrm{E}\boldsymbol{y}|\boldsymbol{u} &= \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} \text{ and} \\
\mathrm{Var}\boldsymbol{y}|\boldsymbol{u} &= \boldsymbol{R}
\end{aligned}$$

- We usually consider the special case in which

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} \right)$$

$$\implies \boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{ZGZ}' + \boldsymbol{R}).$$

- Example: Recall the seedling metabolite study. Previously, we looked at metabolite concentration. Each individual seedling was weighed
- Let $y_{ijk}$ denote the weight of the $k^{th}$ seedling from the $j^{th}$ flat of genotype $i$ ($i = 1, 2, j = 1, 2, 3, 4, k = 1, ..., n_{ij}$; where $n_{ij} = \#$ of seedlings for genotype $i$ flat $j$.)


Genotype A
Genotype B
Seedling Weight
6
8
10
12
14
16
18
20
1
2
3
4
5
6
7
8
Tray

- Consider the model

$$y_{ijk} = \mu + \gamma_i + F_{ij} + \epsilon_{ijk}$$

$$F_{ij}\,'s \overset{i.i.d.}{\sim} N(0, \sigma_F^2) \text{ independent of } \epsilon_{ijk}\,'s \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$$

This model can be written in the form

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{114} \\ y_{115} \\ y_{121} \\ y_{122} \\ . \\ . \\ . \\ y_{247} \\ y_{248} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} \quad \mu = \begin{bmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{14} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{24} \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} e_{111} \\ e_{112} \\ e_{113} \\ e_{114} \\ e_{115} \\ e_{121} \\ e_{122} \\ . \\ . \\ . \\ e_{247} \\ e_{248} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 0 & 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & . & & & & & \\ & & . & & & & & \\ & & . & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- What is $\mathrm{Var}\boldsymbol{y}$?

$$\boldsymbol{G} = \mathrm{Var}(\mu) = \mathrm{Var}([\mathrm{F}_{11}, ..., \mathrm{F}_{24}]') = \sigma^2_{\mathrm{F}}\, \boldsymbol{I}_{8\times 8}$$

$$\boldsymbol{R} = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2_{\mathrm{e}}\, \boldsymbol{I}_{56\times 56}$$

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} = \boldsymbol{Z}\sigma^2_{\mathrm{F}}\, \boldsymbol{I}\boldsymbol{Z}' + \sigma^2_{\mathrm{e}}\, \boldsymbol{I} = \sigma^2_{\mathrm{F}}\boldsymbol{Z}\boldsymbol{Z}' + \sigma^2_{\mathrm{e}}\boldsymbol{I}$$

$$\boldsymbol{Z}\boldsymbol{Z}' = \begin{bmatrix} \mathbf{1}_{n_{11}}\mathbf{1}'_{n_{11}} & 0 & 0 & . & . & . & 0 \\ 0 & \mathbf{1}_{n_{12}}\mathbf{1}'_{n_{12}} & 0 & . & . & . & 0 \\ 0 & 0 & \mathbf{1}_{n_{13}}\mathbf{1}'_{n_{13}} & & & & \\ . & . & & . & & & \\ . & . & & & . & & \\ . & . & & & & . & \\ 0 & 0 & & & & & \mathbf{1}_{n_{24}}\mathbf{1}'_{n_{24}} \end{bmatrix}$$

- This is a block diagonal matrix with blocks of size $n_{ij} \times n_{ij},\ i = 1, 2,\ j = 1, 2, 3, 4.$

- Thus, $\mathrm{Var}(\boldsymbol{y}) = \sigma^2_{\mathrm{F}}\boldsymbol{Z}\boldsymbol{Z}' + \sigma^2_{\mathrm{e}}\boldsymbol{I}$ is also block diagonal. The first block is

$$\mathrm{Var}\begin{bmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{114} \\ y_{115} \end{bmatrix} = \begin{bmatrix} \sigma^2_F + \sigma^2_e & \sigma^2_F & \sigma^2_F & \sigma^2_F & \sigma^2_F \\ \sigma^2_F & \sigma^2_F + \sigma^2_e & \sigma^2_F & \sigma^2_F & \sigma^2_F \\ \sigma^2_F & \sigma^2_F & \sigma^2_F + \sigma^2_e & \sigma^2_F & \sigma^2_F \\ \sigma^2_F & \sigma^2_F & \sigma^2_F & \sigma^2_F + \sigma^2_e & \sigma^2_F \\ \sigma^2_F & \sigma^2_F & \sigma^2_F & \sigma^2_F & \sigma^2_F + \sigma^2_e \end{bmatrix}$$

- Other blocks are the same except that the dimensions are $(\#$ of seedlings per tray $)^2$.
- A $\Sigma$ matrix with this form is called 'Compound Symmetric'

- Two different ways to write the same model
  1. As a mixed model:
     $Y_{ijk} = \mu + \gamma_i + T_{ij} + \epsilon_{ijk}, \qquad T_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_F^2), \qquad \epsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$
  2. As a fixed effects model with correlated errors:
     $Y_{ijk} = \mu + \gamma_i + \epsilon_{ijk}, \qquad \boldsymbol{\epsilon} \sim N(0, \Sigma)$
- In either model:

$$\begin{aligned} \mathrm{Var}(\mathrm{y_{ijk}}) &= \sigma_F^2 + \sigma_e^2 \qquad \forall\ i, j, k. \\ \mathrm{Cov}(\mathrm{y_{ijk}},\ \mathrm{y_{ijk'}}) &= \sigma_F^2 \qquad \forall\ i, j, \text{ and } k \neq k'. \\ Cov(y_{ijk},\ y_{i'j'k'}) &= 0 \qquad \text{if } i \neq i' \text{ or } j \neq j'. \end{aligned}$$

- Any two observations from the same flat have covariance $\sigma_F^2$.
- Any two observations from different flats are uncorrelated.

- What about the variance of the flat average:

$$\begin{aligned}
\mathrm{Var}\bar{y}_{ij.} &= \mathrm{Var}\left(\frac{1}{n_{ij}}\,\mathbf{1}'(y_{ij},\ldots,y_{ijn_{ij}})'\right) = \frac{1}{n_{ij}^2}\,\mathbf{1}'(\sigma_F^2\mathbf{1}\mathbf{1}'+\sigma_e^2\boldsymbol{I})\,\mathbf{1} \\
&= \frac{1}{n_{ij}^2}\left(\sigma_F^2\mathbf{1}'\mathbf{1}\mathbf{1}'\mathbf{1}+\sigma_e^2\mathbf{1}'\boldsymbol{I}\mathbf{1}\right) = \frac{1}{n_{ij}^2}\left(\sigma_F^2 n_{ij}n_{ij}+\sigma_e^2 n_{ij}\right) \\
&= \sigma_F^2+\frac{\sigma_e^2}{n_{ij}}\;\forall\; i,j.
\end{aligned}$$

- If we analyzed seedling dry weight as we did average metabolite concentration, that analysis assumes $\sigma_F^2 = 0$.
- Because that analysis models $\mathrm{Var}(\bar{y}_{ij})$ as $\frac{\sigma^2}{n_{ij}}\;\forall\; i,j$.

- The variance of the genotype average, using $f_i$ flats per genotype:

$$\begin{aligned} \mathrm{Var}\bar{y}_{i..} &= \frac{1}{f_i^2}\sum \mathrm{Var}\bar{y}_{ij.} \\ &= \frac{\sigma_F^2}{f_i} + \frac{1}{f_i^2}\sum \sigma_e^2/n_{ij} \;\forall\; i \end{aligned}$$

- When all flats have same number of seedlings per flat, i.e. $n_{ij} = n \;\forall\; i, j$

$$\mathrm{Var}\bar{y}_{i..} = \frac{\sigma_F^2}{f_i} + \frac{\sigma_e^2}{f_i\, n} \;\forall\; i$$

- Var among flats divided by # flats
  + Var among sdl. divided by **total** # seedlings
- When balanced, mixed model analysis same as computing flat averages, then using OLS model on $\bar{y}_{ij.}$
- except that mixed model analysis provides estimate of $\sigma_F^2$

- Note that $Var(\boldsymbol{y})$ may be written as $\sigma_e^2 V$ where V is a block diagonal matrix with blocks of the form

$$\begin{bmatrix} 1+\frac{\sigma_F^2}{\sigma_e^2} & \frac{\sigma_F^2}{\sigma_e^2} & \cdots & \frac{\sigma_F^2}{\sigma_e^2} \\ \frac{\sigma_F^2}{\sigma_e^2} & 1+\frac{\sigma_F^2}{\sigma_e^2} & \cdots & \frac{\sigma_F^2}{\sigma_e^2} \\ \vdots & \vdots & \ddots & \\ \frac{\sigma_F^2}{\sigma_e^2} & \frac{\sigma_F^2}{\sigma_e^2} & \cdots & 1+\frac{\sigma_F^2}{\sigma_e^2} \end{bmatrix}$$

- Thus, if $\frac{\sigma_F^2}{\sigma_e^2}$ were known, we would have the Aitken Model.

$$\boldsymbol{y} = X\beta + Z\boldsymbol{u} + \boldsymbol{\epsilon} = X\beta + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0},\ \sigma^2 V),\ \sigma^2 \equiv \sigma_e^2$$

- We would use GLS to estimate any estimable $C\beta$ by $C\hat{\beta}_V = C(X'V^{-1}X)^- X'V^{-1}\boldsymbol{y}$

- However, we seldom know $\frac{\sigma_F^2}{\sigma_e^2}$ or, more generally, $\boldsymbol{\Sigma}$ or $\boldsymbol{V}$.
- Thus, our strategy usually involves estimating the unknown parameters in $\boldsymbol{\Sigma}$ to obtain $\hat{\boldsymbol{\Sigma}}$.
- Then inference proceeds based on $C\hat{\beta}_{\hat{V}} = C(X'\hat{\boldsymbol{V}}^{-1}X)^{-}X'\hat{V}^{-1}\boldsymbol{y}$ or $C\hat{\beta}_{\hat{\Sigma}} = C(X'\hat{\boldsymbol{\Sigma}}^{-1}X)^{-}X'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}$.

- Remaining Questions:
  1. How do we estimate the unknown parameters in $\boldsymbol{V}$ (or $\boldsymbol{\Sigma}$) ?
  2. What is the distribution of $C\hat{\beta}_{\hat{V}} = C(X'\hat{\boldsymbol{V}}^{-1}X)^{-}X'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$ ?
  3. How should we conduct inference regarding $C\beta$ ?
  4. Can we test Ho: $\sigma_F^2 = 0$?
  5. Can we use the data to help us choose a model for $\boldsymbol{\Sigma}$?

# R code for LME's

```
d <- read.table('SeedlingDryWeight2.txt',as.is=T,
  header=T)
names(d)
with(d, table(Genotype, Tray))
with(d, table(Tray,Seedling))

# for demonstration, construct a balanced data set
#  5 seedlings for each tray
d2 <- d[d$Seedling <=5,]

# create factors
d2$g <- as.factor(d2$Genotype)
d2$t <- as.factor(d2$Tray)
d2$s <- as.factor(d2$Seedling)
```

```
# fit the ANOVA
temp <- lm(SeedlingWeight ~ g+t, data=d2)
# get the ANOVA table
anova(temp)

# note that all F tests use MSE = seedlings
#   within tray as denominator
# will need to hand-calculate test for genotypes

# better to fit LME.  2 functions:
# lmer() in lme4 or lme() in nlme
#   for most models lmer() is easier to use
library(lme4)
```

```
temp <- lmer(SeedlingWeight~g+(1|t),data=d2)
summary(temp)
anova(temp)

# various extractor functions:
fixef(temp)     # estimates of fixed effects
vcov(temp)      # VC matrix for the fixed effects
VarCorr(temp)   # estimated variance(-covariance) for r
ranef(temp)     # predictions of random effects
coef(temp)      # fixed effects + pred's of random effe

fitted(temp)    # conditional means for each obs (X bha
resid(temp)     # conditional residuals (Y - fitted)
```

```
# REML is the default method for estimating
#  variance components.
# if want to use ML, can specify that
lmer(SeedlingWeight~g+(1|t),REML=F, data=d2)

# but how do we get p-values for tests
#  or decide what df to use to construct a ci?

# we'll talk about inference in R soon
```

# Experimental Designs and LME's

- LME models provide one way to model correlations among observations
- Very useful for experimental designs where there is more than one size of experimental unit
- Or designs where the observation unit is not the same as the experimental unit.
- My philosophy (widespread at ISU and elsewhere) is that the way an experiment is conducted specifies the random effect structure for the analysis.
- Observational studies are very different
  No randomization scheme, so no clearcut random effect structure
  Often use model selection methods to help chose the random effect structure
- NOT SO for a randomized experiment.

- One example:
  - study designed as an RCBD.
  - treatments are randomly assigned within a block
  - analyze data, find out block effects are small
  - should you delete block from the model and reanalyze?
  - above philosophy says “tough”. Block effects stay in the analysis
  - for the **next** study, seriously consider:
    blocking in a different way
    or using a CRD

- Following pages work through details of LME’s for some common study designs

# Mouse muscle study

- Mice grouped into blocks by litter. 4 mice per block, 4 blocks.
- Four treatments randomly assigned within each block
- Collect two replicate muscle samples from each mouse
- Measure response on each muscle sample
- 16 eu., 32 ou.
- Questions concern differences in mean response among the four treatments
- The model usually used for a study like this is:
- $y_{ijk} = \mu + \tau_i + l_j + m_{ij} + e_{ijk}$, where
  $y_{ijk}$ is the $k^{th}$ measurement of the response for the mouse from litter $j$ that received treatment $i$,
  $(i = 1, 2, 3, 4;\ j = 1, 2, 3, 4;\ k = 1, 2)$.

- The fixed effects:

$$\beta = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} \in I\!R^5 \text{ is an unknown vector of fixed parameters}$$

- $\boldsymbol{u} = [l_1, l_2, l_3, l_4, m_{11}, m_{21}, m_{31}, m_{41}, m_{12}, \ \ldots\ , m_{34}, m_{44}]'$ is a vector of random effects describing litters and mice.
- $\epsilon = [e_{111}, e_{112}, e_{211}, e_{212}, \ \ldots\ , e_{441}, e_{442}, \ \ldots\ , e_{441}, e_{442}]'$ is a vector of random errors.
- with $\boldsymbol{y} = [y_{111}, y_{112}, y_{211}, y_{212}, \ \ldots\ , y_{411}, y_{412}, \ \ldots\ , y_{441}, y_{442}]'$, the model can be written as a linear mixed effects model

$$\boldsymbol{y} = X\beta + Z\boldsymbol{u} + \epsilon, \text{ where}$$

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & & & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & & & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & & & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & & & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & . & . & . & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & 0 \\ & & & & & . & & . & & & \\ & & & & & . & & & . & & \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & & 1 \end{bmatrix}$$

- We can write less and be more precise using Kronecker product notation.

$$X = \underbrace{\mathbf{1}}_{4\times 1} \otimes [\underbrace{\mathbf{1}}_{8\times 1}, \underbrace{I}_{4\times 4} \otimes \underbrace{\mathbf{1}}_{2\times 1}] \qquad Z = [\underbrace{I}_{4\times 4} \otimes \underbrace{\mathbf{1}}_{8\times 1}, \underbrace{I}_{16\times 16} \otimes \underbrace{\mathbf{1}}_{2\times 1}]$$

- In this experiment, we have two random factors: Litter and Mouse.
- We can partition our random effects vector $\boldsymbol{u}$ into a vector of litter effects and a vector of mouse effects:

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{m} \end{bmatrix}, \boldsymbol{l} = \begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \end{bmatrix}, \boldsymbol{m} = \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \\ m_{41} \\ m_{12} \\ . \\ . \\ . \\ m_{44} \end{bmatrix}$$

- We make the usual assumption that

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{m} \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_l^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_m^2 \boldsymbol{I} \end{bmatrix} \right)$$

  Where $\sigma_l^2, \sigma_m^2 \in I\!R^+$ are unknown parameters.

- We can partition:

$$Z = [\underbrace{\boldsymbol{I}}_{4\times 4} \otimes \underbrace{\boldsymbol{1}}_{8\times 1}, \underbrace{\boldsymbol{I}}_{16\times 16} \otimes \underbrace{\boldsymbol{1}}_{2\times 1}] = [\boldsymbol{Z}_l, \boldsymbol{Z}_m]$$

  We have

$$\boldsymbol{Zu} = \begin{bmatrix} \boldsymbol{Z}_l, & \boldsymbol{Z}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{m} \end{bmatrix} = \boldsymbol{Z}_l \boldsymbol{l} + \boldsymbol{Z}_m \boldsymbol{m} \qquad \text{and}$$

$$\text{Var}(\boldsymbol{Zu}) = \boldsymbol{ZGZ}' = \left[ \begin{array}{cc} \boldsymbol{Z}_I, & \boldsymbol{Z}_m \end{array} \right] \left[ \begin{array}{cc} \sigma_I^2 \boldsymbol{I} & 0 \\ 0 & \sigma_m^2 \boldsymbol{I} \end{array} \right] \left[ \begin{array}{c} \boldsymbol{Z}_I' \\ \boldsymbol{Z}_m' \end{array} \right]$$

$$= \boldsymbol{Z}_I(\sigma_I^2 \boldsymbol{I})\boldsymbol{Z}_I' + \boldsymbol{Z}_m(\sigma_m^2 \boldsymbol{I})\boldsymbol{Z}_m'$$

$$= \sigma_I^2 \boldsymbol{Z}_I \boldsymbol{Z}_I' + \sigma_m^2 \boldsymbol{Z}_m \boldsymbol{Z}_m'$$

$$= \sigma_I^2 \underbrace{\boldsymbol{I}}_{4\times 4} \otimes \underbrace{\boldsymbol{11}'}_{8\times 8} + \sigma_m^2 \underbrace{\boldsymbol{I}}_{16\times 16} \otimes \underbrace{\boldsymbol{11}'}_{2\times 2}$$

- We usually assume that all random effects and random errors are mutually independent and that the errors (like the effects within each factor) are identically distributed:

$$\begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{m} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_I^2 \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_m^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma_e^2 \boldsymbol{I} \end{bmatrix} \right)$$

- The unknown variance parameters $\sigma_l^2, \sigma_m^2, \sigma_e^2 \in I\!R^+$ are called **variance components**.
- In this case, we have $\boldsymbol{R} = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_{\mathrm{e}}^2 \boldsymbol{I}$.
- Thus, $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{ZGZ}' + \boldsymbol{R} = \sigma_{\mathrm{l}}^2 \boldsymbol{Z}_{\mathrm{l}} \boldsymbol{Z}_{\mathrm{l}}' + \sigma_{\mathrm{m}}^2 \boldsymbol{Z}_{\mathrm{m}} \boldsymbol{Z}_{\mathrm{m}}' + \sigma_{\mathrm{e}}^2 \boldsymbol{I}$.
- This is a block diagonal matrix.
  Each litter is one of the following blocks.

$$\begin{bmatrix}
a+b+c & a+b & a & a & a & a & a & a \\
a+b & a+b+c & a & a & a & a & a & a \\
a & a & a+b+c & a+b & a & a & a & a \\
a & a & a+b & a+b+c & a & a & a & a \\
a & a & a & a & a+b+c & a+b & a & a \\
a & a & a & a & a+b & a+b+c & a & a \\
a & a & a & a & a & a & a+b+c & a+b \\
a & a & a & a & a & a & a+b & a+b+c
\end{bmatrix}$$

- (To save space, $a = \sigma_l^2,\ b = \sigma_m^2,\ c = \sigma_e^2$)

- The random effects specify the correlation structure of the observations
- This is determined (in my philosophy) by the study design
- Changing the random effects changes the implied study design
- Delete the mouse random effects:
  The study is now an RCBD with 8 mice (one per obs.) per block
- Also delete the litter random effects:
  The study is now a CRD with 8 mice per treatment

# Should blocks be fixed or random?

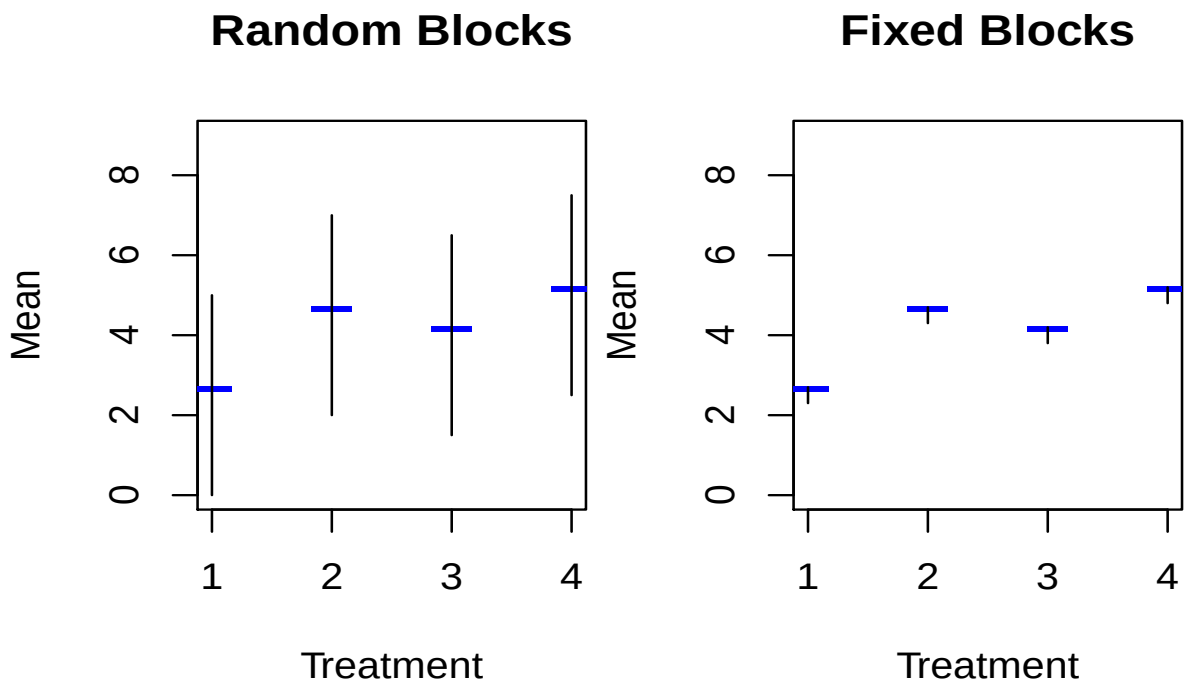
- Some views:
- The design is called an RCBD, so of course blocks are random. But, R in the name is Randomized (describing treatment assignment).
- Introductory ANOVA: blocks are fixed, because that's all we know about.
- Some things that influence my answer
  - If blocks are complete and balanced (all blocks have equal #'s of all treatments), inferences about treatment differences (or contrasts) are same either way.
  - If blocks are incomplete, an analysis with fixed blocks evaluates treatment effects within each block, then averages effects over blocks.
    An analysis with random blocks, "recovers" the additional information about treatment differences provided by the block means.
    Will talk about this later, if time.

- However, the biggest and most important difference concerns the relationship between the se's of treatment means and the se's of treatment differences (or contrasts).
- Simpler version of mouse study:
  no subsampling, 4 blocks, 4 mice/block, 4 treatments
- $Y_{ij} = \mu + \tau_i + l_j + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma_e^2)$
- If blocks are random, $l_j \sim N(0, \sigma_l^2)$

| | Blocks are: | |
|---|---|---|
| | Fixed | Random |
| Var $\hat{\tau}_1 - \hat{\tau}_2$ | $2\,\sigma_e^2/4$ | $2\,\sigma_e^2/4$ |
| Var $\hat{\mu} + \hat{\tau}_1$ | $\sigma_e^2/4$ | $(\sigma_l^2 + \sigma_e^2)/4$ |

- When blocks are random, Var trt mean includes the block-block variability.
- Matches implied inference space:
  repeating study on new litters and new mice

- Common to add $\pm$ se bars to a plot of means
- If blocking effective, $\sigma_I^2$ is large, so you get:



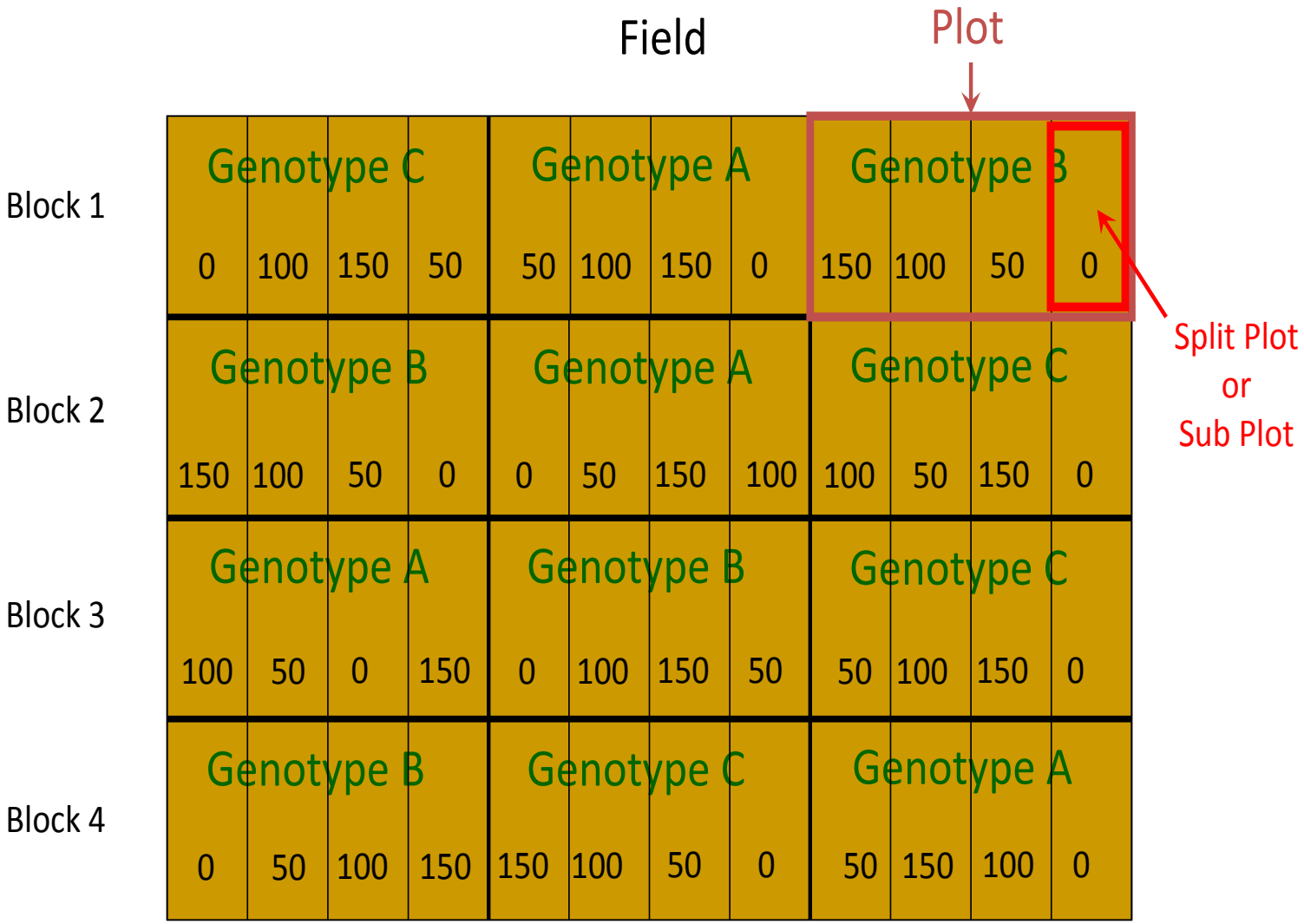

- The random blocks picture does not support claims about trt. diff.

# So should blocks be Fixed or Random?

- My approach:
- When the goal is to summarize difference among treatments, use fixed blocks
- When the goal is to predict means for new blocks, use random blocks
- If blocks are incomplete, think carefully about goals of study
- Many (most) have different views.

# Split plot experimental design

- Field experiment comparing fertilizer response of 3 corn genotypes
- Large plots planted with one genotype (A, B, C)
- Each large plot subdivided into 4 small plots.
- Small plots fertilized with 0, 50, 100, or 150 lbs Nitrogen / acre.
- Large plots grouped into blocks; Genotype randomly assigned to large plot,
- Small plots randomly assigned to N level within each large plot.
- Key feature: Two sizes of experimental units


Field
Plot
Block 1
Genotype C
0 100 150 50
Genotype A
50 100 150 0
Genotype B
150 100 50 0
Block 2
Genotype B
150 100 50 0
Genotype A
0 50 150 100
Genotype C
100 50 150 0
Block 3
Genotype A
100 50 0 150
Genotype B
0 100 150 50
Genotype C
50 100 150 0
Block 4
Genotype B
0 50 100 150
Genotype C
150 100 50 0
Genotype A
50 150 100 0
Split Plot
or
Sub Plot

- Names:
  - The large plots are called main plots or whole plots
  - Genotype is the main plot factor
  - The small plots are split plots
  - Fertilizer is the split plot factor
- Two sizes of eu's
  - Main plot is the eu. for genotype
  - Split plot is the eu. for fertilizer
  - Split plots are nested in Main plots
- Many different variations, this is the most common
- RCB for main plots; CRD for split plots within main plots
- Can extend to three or more sizes of eu.
  split-split plot or split-split-split plot designs

- Split plot studies commonly mis-analyzed as RCBD with one eu.
- Treat Genotype$\times$Fertilizer combination (12 treatments) as if they were randomly assigned to each small plot

Field

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Block 1 | B 100 | B 0 | A 0 | C 100 | B 150 | C 50 | A 50 | A 150 | C 150 | B 50 | C 0 | A 100 |
| Block 2 | A 150 | A 0 | C 50 | A 50 | B 100 | B 50 | C 100 | C 0 | A 100 | C 150 | B 150 | B 0 |
| Block 3 | C 0 | A 0 | A 100 | B 100 | B 50 | B 0 | A 150 | C 50 | A 50 | C 150 | C 100 | B 150 |
| Block 4 | B 0 | C 150 | B 50 | A 150 | C 100 | A 0 | B 150 | C 50 | B 100 | C 0 | A 100 | A 50 |

- And if you ignore blocks, you have a very different set of randomizations

Field

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B 50 | B 0 | A 150 | B 100 | A 100 | C 150 | A 50 | B 0 | A 50 | C 100 | C 0 | C 100 |
| A 50 | A 0 | C 50 | B 50 | B 150 | B 50 | A 0 | C 0 | A 100 | C 50 | B 150 | B 0 |
| C 0 | A 0 | A 100 | A 150 | A 0 | B 0 | A 150 | B 150 | A 50 | B 150 | C 100 | A 100 |
| B 50 | B 100 | B 100 | C 150 | C 100 | C 50 | A 150 | C 50 | C 150 | C 0 | C 150 | B 100 |

- Confusion is (somewhat understandable)
  Same treatment structure (2 way complete factorial)
- Different experimental design because different way of randomizing treatments to eu.s
- So different model than the usual RCBD
- Use random effects to account for the two sizes of eu.

- Rarely done on purpose. Usually forced by treatment constraints
  - Cleaning out planters is time consuming, so easier to plant large areas
  - Growth chambers can only be set to one temperature, have room for many flats
- In the engineering literature, split plot designs are often called “hard-to-change” factor designs

# Diet and Drug effects on mice

- Split plot designs may not involve a traditional plot
- Study of diet and drug effect on mice
- Have 2 mice from each of 8 litters; cages hold 2 mice each.
- Put the two mice from a litter into a cage.
- Randomly assign diet (1 or 2) to the cage/litter
- Randomly assign drug (A or B) to a mouse within a cage
- Main plots = cage/litter, in a CRD
  Diet is the main plot treatment factor
- Split plots = mouse, in a CRD within cage
  Drug is the split plot treatment factor
- Let's construct a model for this mouse study, 16 obs.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + l_{ik} + e_{ijk} \ (i = 1, 2;\ j = 1, 2;\ k = 1, ..., 4)$$

$$\boldsymbol{y} = \begin{bmatrix} y_{111} \\ y_{121} \\ y_{112} \\ y_{122} \\ y_{113} \\ y_{123} \\ y_{114} \\ y_{124} \\ y_{211} \\ y_{221} \\ y_{212} \\ y_{222} \\ y_{213} \\ y_{223} \\ y_{214} \\ \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} \mu = \begin{bmatrix} l_{11} \\ l_{12} \\ l_{13} \\ l_{14} \\ l_{21} \\ l_{22} \\ l_{23} \\ l_{24} \end{bmatrix} \epsilon = \begin{bmatrix} e_{111} \\ e_{112} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_{224} \end{bmatrix}$$

$$\boldsymbol{X} = [\underbrace{\mathbf{1}}_{16\times 1}, \underbrace{\boldsymbol{I}}_{2\times 2} \otimes \underbrace{\mathbf{1}}_{8\times 1}, \underbrace{\mathbf{1}}_{8\times 1} \otimes \underbrace{\boldsymbol{I}}_{2\times 2}, \underbrace{\boldsymbol{I}}_{2\times 2} \otimes \underbrace{\mathbf{1}}_{4\times 1} \otimes \underbrace{\boldsymbol{I}}_{2\times 2}]$$

$$\boldsymbol{Z} = \underbrace{\boldsymbol{I}}_{8\times 8} \otimes \underbrace{\mathbf{1}}_{2\times 1}$$

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon$$

$$\begin{bmatrix} \boldsymbol{u} \\ \epsilon \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_I^2 \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{I} \end{bmatrix}\right)$$

$$\text{Var}(\boldsymbol{Z}\boldsymbol{u}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' = \sigma_1^2 \boldsymbol{Z}\boldsymbol{Z}' = \sigma_1^2 [\underbrace{\boldsymbol{I}}_{8\times 8} \otimes \underbrace{\mathbf{1}}_{2\times 1}][\underbrace{\boldsymbol{I}}_{8\times 8} \otimes \underbrace{\mathbf{1}}_{2\times 1}]'$$

$$= \sigma_I^2 \underbrace{\boldsymbol{I}}_{8\times 8} \otimes \underbrace{\mathbf{1}\mathbf{1}'}_{2\times 1} = \text{Block Diagonal with blocks } \begin{bmatrix} \sigma_I^2 & \sigma_I^2 \\ \sigma_I^2 & \sigma_I^2 \end{bmatrix}$$

$$\text{Var}(\boldsymbol{\epsilon}) = \text{R} = \sigma^2_{\text{e}} \boldsymbol{I}$$

$$\text{Var}(\boldsymbol{y}) = \sigma^2_{\text{I}} \underbrace{\boldsymbol{I}}_{8 \times 8} \otimes \underbrace{\mathbf{1}\mathbf{1}'}_{2 \times 2} + \sigma^2_{\text{e}} \boldsymbol{I}$$

$$= \text{Block Diagonal with blocks} \begin{bmatrix} \sigma^2_I + \sigma^2_e & \sigma^2_I \\ \sigma^2_I & \sigma^2_I + \sigma^2_e \end{bmatrix}$$

- Thus, the covariance between two observations from the same litter is $\sigma_l^2$ and the correlation is $\frac{\sigma_l^2}{\sigma_l^2+\sigma_e^2}$.
- This is also easy to compute from the non-matrix expressinon of the model.

$$\forall i,j\ \mathrm{Var}(\mathrm{y_{ijk}}) = \mathrm{Var}(\mu+\alpha_\mathrm{i}+\beta_\mathrm{j}+\gamma_\mathrm{ij}+\mathrm{l_{ik}}+\mathrm{e_{ijk}}) = \mathrm{Var}(\mathrm{l_{ik}}+\mathrm{e_{ijk}}) = \sigma_\mathrm{l}^2+\sigma_\mathrm{e}^2$$

$$\begin{aligned}
\mathrm{Cov}(\mathrm{y_{i1k}}, \mathrm{y_{i2k}}) &= \mathrm{Cov}(\mu + \alpha_\mathrm{i} + \beta_1 + \gamma_\mathrm{i1} + \mathrm{l_{ik}} + \mathrm{e_{i1k}}, \\
&\qquad \mu + \alpha_i + \beta_2 + \gamma_{i2} + l_{ik} + e_{i2k}) \\
&= \mathrm{Cov}(\mathrm{l_{ik}} + \mathrm{e_{i1k}},\ \mathrm{l_{ik}} + \mathrm{e_{i2k}}) \\
&= \mathrm{Cov}(\mathrm{l_{ik}}, \mathrm{l_{ik}}) + \mathrm{Cov}(\mathrm{l_{ik}} + \mathrm{e_{i2k}}) + \mathrm{Cov}(\mathrm{e_{i1k}}, \mathrm{l_{ik}}) + \mathrm{Cov}(\mathrm{e_{i1k}} + \mathrm{e_{i2k}}) \\
&= \mathrm{Cov}(\mathrm{l_{ik}} + \mathrm{l_{ik}} + 0 + 0 + 0 = \mathrm{Var}(\mathrm{l_{ik}}) \\
&= \sigma_l^2
\end{aligned}$$

$$\mathrm{Cor}(y_{i1k}, y_{i2k}) = \frac{\mathrm{Cov}(\mathrm{y_{i1k}}, \mathrm{y_{i2k}})}{\sqrt{\mathrm{Var}(\mathrm{y_{i1k}})\mathrm{Var}(\mathrm{y_{i2k}})}} = \frac{\sigma_l^2}{\sigma_l^2 + \sigma_e^2}$$

# THE ANOVA APPROACH TO THE ANALYSIS OF LINEAR MIXED EFFECTS MODELS

- A model for expt. data with subsampling

$$y_{ijk} = \mu + \tau_i + u_{ij} + e_{ijk},\ (i = 1, ..., t;\ j = 1, ..., n;\ k = 1, ..., m)$$

$$\beta = (\mu, \tau_i, ..., \tau_t)',\ u = (u_{11}, u_{12}, ..., u_{tn})', \epsilon = (e_{111}, e_{112}, ..., e_{tnm})',$$

$$\beta \in I\!R^{t+1},\ \text{an unknown parameter vector,}$$

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{I} \end{bmatrix} \right)$$

$$\sigma_u^2, \sigma_e^2 \in I\!R^{+},\ \text{unknown variance components}$$

- This is the commonly-used model for a CRD with t treatments, n experimental units per treatment, and m observations per experimental unit.
- We can write the model as $\boldsymbol{y} = X\beta + Z\boldsymbol{u} + \epsilon$, where

$$X = [\underbrace{\mathbf{1}}_{tnm\times 1}, \underbrace{\boldsymbol{I}}_{t\times t} \otimes \underbrace{\mathbf{1}}_{nm\times 1}] \quad \text{and} \quad Z = [\underbrace{\boldsymbol{I}}_{tn\times tn} \otimes \underbrace{\mathbf{1}}_{m\times 1}]$$

- Consider the sequence of three column spaces:
  $X_1 = \mathbf{1}_{tnm\times 1},\ X_2 = [\mathbf{1}_{tnm\times 1},\ \boldsymbol{I}_{t\times t} \otimes \mathbf{1}_{nm\times 1}]$,
  $X_3 = [\mathbf{1}_{tnm\times 1},\ \boldsymbol{I}_{t\times t} \otimes \mathbf{1}_{nm\times 1}, \boldsymbol{I}_{tn\times\ tn} \otimes \mathbf{1}_{m\times\ 1}]$
- These correspond to the models:

$$\begin{aligned} X_1\text{:} \quad & Y_{ijk} = \mu + \epsilon_{ijk} \\ X_2\text{:} \quad & Y_{ijk} = \mu + \tau_i + \epsilon_{ijk} \\ X_3\text{:} \quad & Y_{ijk} = \mu + \tau_i + \boldsymbol{u}_{ij} + \epsilon_{ijk} \end{aligned}$$

# ANOVA table for subsampling

| Source | SS | df | df |
|---|---|---|---|
| *treatments* | $\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$ | $\text{rank}(\boldsymbol{X}_2) - \text{rank}(\boldsymbol{X}_1)$ | $t-1$ |
| *eu*(*treatments*) | $\boldsymbol{y}'(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$ | $\text{rank}(\boldsymbol{X}_3) - \text{rank}(\boldsymbol{X}_2)$ | $t(n-1)$ |
| *ou*(*eu*, *treatments*) | $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_3)\boldsymbol{y}$ | $tnm - \text{rank}(\boldsymbol{X}_3)$ | $tn(m-1)$ |
| *C.total* | | $tnm - 1$ | |

- In terms of squared differences:

| Source | df | Sum of Squares |
|---|---|---|
| *trt* | $t-1$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2$ |
| *eu*(*trt*) | $t(n-1)$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot})^2$ |
| *ou*(*eu*, *trt*) | $tn(m-1)$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(\overline{y}_{ijk} - \overline{y}_{ij\cdot})^2$ |
| *C.total* | $tnm-1$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(y_{ijk} - \overline{y}_{\cdot\cdot\cdot})^2$ |

- Which simplifies to:

| Source | df | Sum of Squares |
|---|---|---|
| $trt$ | $t-1$ | $nm\sum_{i=1}^{t}(\overline{y}_{i\cdot\cdot}-\overline{y}_{\cdots})^2$ |
| $eu(trt)$ | $tn-t$ | $m\sum_{i=1}^{t}\sum_{j=1}^{n}(\overline{y}_{ij\cdot}-\overline{y}_{i\cdot\cdot})^2$ |
| $ou(eu,trt)$ | $tnm-tn$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(\overline{y}_{ijk}-\overline{y}_{ij\cdots})^2$ |
| $C.total$ | $tnm-1$ | $\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{m}(y_{ijk}-\overline{y}_{\cdots})^2$ |

- Each line has a MS = SS / df
- MS's are random variables. Their expectations are informative.

# Expected Mean Squares, EMS

$$
\begin{aligned}
E(MS_{trt}) &= \frac{nm}{t-1}\sum_{i=1}^{t} E(\overline{y}_{i..} - \overline{y}_{...})^2 \\
&= \frac{nm}{t-1}\sum_{i=1}^{t} E(\mu + \tau_i + \overline{u}_{i.} + \overline{e}_{i..} - \mu - \overline{\tau}_{.} - \overline{u}_{..} - \overline{e}_{...})^2 \\
&= \frac{nm}{t-1}\sum_{i=1}^{t} E(\tau_i - \overline{\tau}_{.} + \overline{u}_{i} - \overline{u}_{..} + \overline{e}_{i..} - \overline{e}_{...})^2 \\
&= \frac{nm}{t-1}\sum_{i=1}^{t} [(\tau_i - \overline{\tau}_{.})^2 + E(\overline{u}_{i} - \overline{u}_{..})^2 + E(\overline{e}_{i..} - \overline{e}_{...})^2] \\
&= \frac{nm}{t-1}[\sum_{i=1}^{t}(\tau_i - \overline{\tau})^2 + E(\sum_{i=1}^{t}(\overline{u}_{i} - \overline{u}_{..})^2) \\
&\quad + E(\sum_{i=1}^{t}\overline{e}_{i.} - \overline{e})^2)]
\end{aligned}
$$

- $\overline{u}_1,...,\overline{u}_t. \overset{i.i.d.}{\sim} N(0, \frac{\sigma_u^2}{n})$. Thus, $E(\sum_{i=1}^{t}(\overline{u}_{i} - \overline{u}_{..})^2) = (t-1)\frac{\sigma_u^2}{n}$
- $\overline{e}_1...,...,\overline{e}_t.. \overset{i.i.d.}{\sim} N(0, \frac{\sigma_u^2}{nm})$. Thus, $E(\sum_{i=1}^{t}(\overline{e}_{i} - \overline{e}_{..})^2) = (t-1)\frac{\sigma_e^2}{mn}$

- It follows that $E(MS_{trt}) = \frac{nm}{t-1}\sum_{i=1}^{t}(\tau_i - \overline{\tau})^2 + m\sigma_u^2 + \sigma_e^2$

- Similar calculations allow us to add Expected Mean Squares (EMS) to our Anova table.

| Source | EMS |
|---|---|
| $trt$ | $\sigma_e^2 + m\sigma_u^2 + \frac{nm}{t-1}\sum_{i=1}^{t}(\tau_i - \overline{\tau})^2$ |
| $eu(trt)$ | $\sigma_e^2 + m\sigma_u^2$ |
| $ou(eu, trt)$ | $\sigma_e^2$ |

- Mean Squares could also be computed using $E(\boldsymbol{y}'A\boldsymbol{y}) = tr(A\epsilon) + [E(\boldsymbol{y})]'AE(\boldsymbol{y})$,
- Where $\sum = Var(y)=ZGZ' + R = \sigma_u^2 \underbrace{I}_{tn\times\ tn} \otimes \underbrace{\mathbf{11}'}_{m\times\ m} + \sigma_e^2 \underbrace{I}_{tnm\times\ tnm}$ and

$$E(\boldsymbol{y}) = [\mu + \tau_1, \mu + \tau_2, \ldots, \mu + \tau_t]' \otimes \underbrace{I}_{nm\times\ 1}$$

- Futhermore, it can be shown that
  - $\boldsymbol{y}' \frac{(P_2 - P_1)}{(\sigma_e^2 + m\sigma_u^2)} \boldsymbol{y} \sim \chi^2_{t-1} \left( [\frac{nm}{\sigma_e^2 + m\sigma_u^2}] \sum_{i=1}^{t} (\tau_i - \overline{\tau}.)^2/(t-1) \right)$
  - $\boldsymbol{y}' \frac{(P_3 - P_2)}{(\sigma_e^2 + m\sigma_u^2)} \boldsymbol{y} \sim \chi^2_{tn-t}$
  - $\boldsymbol{y}' \frac{(I - P_3)}{\sigma_e^2} \boldsymbol{y} \sim \chi^2_{tnm-tn}$
- and these three $\chi^2$ random variables are independent.
- It follows that
  - $F_1 = \frac{MS_{trt}}{MS_{eu(trt)}} \sim F^{[\frac{nm}{\sigma_e^2 + m\sigma_u^2} \sum_{i=1}^{t} (\tau_i - \overline{\tau}.)^2/(t-1)]}_{t-1,\ tn-t}$
  - use $F_1$ to test $H_0 : \tau_1 = ... = \tau_t$
  - $F_2 = \frac{MS_{eu(trt)}}{MS_{ou(eu,trt)}} \sim (\frac{\sigma_e^2 + m\sigma_u^2}{\sigma_e^2}) F_{tn-t,\ tnm-tn}$
  - use $F_2$ to test $H_0 : \sigma_u^2 = 0$

- Notice an important difference between the distributions of the F statistics testing Ho: $\tau_1 = \tau_2 = \ldots = \tau_T$ and testing Ho: $\sigma_u^2 = 0$:
- The F statistic for a test of fixed effects, e.g. $F_1$, has a non-central F distribution under Ha
- The F statistic for a test of a variance component, e.g. $F_2$, has a scaled central F distribution under Ha
- Both have same central F distribution under Ho
- If change a factor from "fixed" to "random":
  - critical value for $\alpha$-level test is same, but
  - power/sample size calculations are not the same!

# Shortcut to identify Ho

- The EMS tell you what Ho associated with any F test
- Ho for F = $MS_A/MS_B$ is whatever is necessary for $EMS_A/EMS_B$ to equal 1.

| F statistic | EMS ratio | Ho |
|---|---|---|
| $MS_{trt}/MS_{eu(trt)}$ | $\frac{\sigma_e^2+m\sigma_u^2+\frac{nm}{t-1}\sum_{i=1}^{t}(\tau_i-\overline{\tau})^2}{\sigma_e^2+m\sigma_u^2}$ | $\sum_{i=1}^{t}(\tau_i-\overline{\tau})^2=0$ |
| $MS_{eu(trt)}/MS_{ou(eu,trt)}$ | $\frac{\sigma_e^2+m\sigma_u^2}{\sigma_e^2}$ | $\sigma_u^2=0$ |
| $MS_{trt}/MS_{ou(eu,trt)}$ | $\frac{\sigma_e^2+m\sigma_u^2+\frac{nm}{t-1}\sum_{i=1}^{t}(\tau_i-\overline{\tau})^2}{\sigma_e^2}$ | $\sigma_u^2=0$, and $\sum_{i=1}^{t}(\tau_i-\overline{\tau})^2=0$ |

# Estimating estimable functions of $\beta$

- Consider the GLS estimator

$$\hat{\beta}_{\mathbf{\Sigma}} = (X'\mathbf{\Sigma}^{-1}X)^{-}X'\mathbf{\Sigma}^{-1}\boldsymbol{y}$$

- When $m$, number of subsamples per eu. constant,

$$\text{Var}(\boldsymbol{y}) = \sigma_u^2 \underbrace{I}_{tn\times\, tn} \otimes \underbrace{\mathbf{1}\mathbf{1}'}_{m\times\, m} + \sigma_e^2 \underbrace{I}_{tnm\times\, tnm} \equiv \mathbf{\Sigma} \qquad (1)$$

- This is a very special $\mathbf{\Sigma}$.
- It turns out that $\hat{\beta}_{\mathbf{\Sigma}} = (X'\mathbf{\Sigma}^{-1}X)^{-}X'\mathbf{\Sigma}^{-1}\boldsymbol{y} = (X'X)^{-}X'\boldsymbol{y} = \hat{\beta}$
- Thus, when $\mathbf{\Sigma}$ has the form of (1),
  the OLS estimator of any estimable $C\beta$ is the GLS
- Estimates of $\boldsymbol{C}\beta$ do not depend on $\sigma_u^2$ or $\sigma_e^2$
- Estimates of $\boldsymbol{C}\beta$ are BLUE

# ANOVA estimate of a variance component

- Given data, how can you estimate $\sigma_u^2$?
- $E\left(\frac{MS_{eu(trt)} - MS_{ou(eu,trt)}}{m}\right) = \frac{(\sigma_e^2 + m\sigma_u^2) - \sigma_e^2}{m} = \sigma_u^2$
- Thus, $\frac{MS_{eu(trt)} - MS_{ou(eu,trt)}}{m}$ is an unbiased estimator of $\sigma_u^2$
- Method of Moments estimator because it is obtained by equating observed statistics with their moments (expected values in this case) and solving the resulting set of equations for unknown parameters in terms of observed statistics.
- Notice $\hat{\sigma}_u^2$ is smaller than Var $\bar{y}_{ij.}$ = sample var among e.u.'s.
  - $\bar{y}_{ij.}$ includes variability among eu's, $\sigma_u^2$, and variability among observations, $\sigma_e^2$.
  - Hence my "if perfect knowledge" ($\sigma_e^2 = 0$) qualifications when I introduced variance components.
- Although $\frac{MS_{eu(trt)} - MS_{ou(eu,trt)}}{m}$ is an unbiased estimator of $\sigma_u^2$, it can take negative values.
- $\sigma_u^2$, the population variance of the u random effects, cannot be negative!

# Why $\hat{\sigma}_u^2$ might be $< 0$

- Sampling variation in $MS_{eu(trt)}$ and $MS_{ou(eu,trt)}$
- especially if $\sigma_u^2 = 0$ or $\approx 0$.
- Incorrect estimate of $\sigma_e^2$
  - One outlier can inflate $MS_{ou(eu,trt)}$
  - Consequence is $\hat{\sigma}_u^2$ might be $< 0$
- Model not correct
  - Var $Y_{ijk}|u_{ij}$ may be incorrectly specified
  - e.g. $R = \sigma_e^2 \boldsymbol{I}$, but Var $Y_{ijk}|u_{ij}$ not constant
  - or obs not independent
- My advice: don't blindly force $\hat{\sigma}_u^2$ to be $\geq 0$.
- Check data, think carefully about model.

# Two ways to analyze data with subsampling

- Mixed model with $\sigma_u^2$ and $\sigma_e^2$

$$y_{ijk} = \mu + \tau_i + u_{ij} + e_{ijk}, \ (i = 1, ..., t; j = 1, ..., n; k = 1, ..., m)$$

- Analyses of e.u. averages
  - The average of observations for experimental unit $ij$ is $\overline{y}_{ij\cdot} = \mu + \tau_i + u_{ij} + \overline{e}_{ij\cdot}$
  - $\overline{y}_{ij} = \mu + \tau_i + \epsilon_{ij}$, where $\epsilon_{ij} = u_{ij} + \overline{e}_{ij\cdot}$. $\forall\ i, j$.
  - $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ where $\sigma^2 = \sigma_u^2 + \frac{\sigma_e^2}{m}$
  - This is a nGM model for the averages $\left\{\overline{y}_{ij\cdot} : i = 1, ..., t; j = 1, ..., n\right\}$
  - When $m_{ij} = m$ for all e.u.'s,
    - inferences about estimable functions of $\boldsymbol{\beta}$ are exactly the same
    - $\hat{\sigma}^2$ is an estimate of $\sigma_u^2 + \frac{\sigma_e^2}{m}$.
    - We can't separately estimate $\sigma_u^2$ and $\sigma_e^2$, but not an issue if focus is on inference for estimable functions of $\boldsymbol{\beta}$.

# Support for these claims

- What can we estimate? For both models,

$$E(y) = [\mu + \tau_1,\ \mu + \tau_2,\ \ldots,\ \mu + \tau_t] \otimes \underbrace{I}_{nm\times\ 1}$$

- The only estimable quantities are linear combinations of the treatment means $\mu + \tau_1, \mu + \tau_2, ..., \mu + \tau_t$
- Best Linear Unbiased Estimators are $\overline{y}_1.., \overline{y}_2.., .., \overline{y}_t..,$, respectively.

- What about Var $y_{ij.}$?
- Two ways to compute:

$$\begin{aligned}
Var(\overline{y}_{i..}) &= Var(\mu + \tau_i + \overline{u}_{i.} + \overline{e}_{i..}) \\
&= Var(\overline{u}_{i.} + \overline{e}_{i..}) \\
&= Var(\overline{u}_{i.}) + Var(\overline{e}_{i..}) \\
&= \frac{\sigma_u^2}{n} + \frac{\sigma_e^2}{nm} \\
&= \frac{1}{n}(\sigma_u^2 + \frac{\sigma_e^2}{m}) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

- Or using matrix representation,

$$
\begin{aligned}
Var(\overline{y}_{i\cdot\cdot}) &= Var(\frac{1}{n}\sum_{j=1}^{n}\overline{y}_{ij\cdot}) \\
&= \frac{1}{n}Var(\overline{y}_{i1\cdot}) \\
&= \frac{1}{n}Var(\frac{1}{m}\underbrace{\mathbf{1}'}_{m\times 1}(y_{i11},\ldots,y_{i1m})') \\
&= \frac{1}{n}\frac{1}{m^2}\underbrace{\mathbf{1}'}_{m\times 1}(\sigma_e^2 I+\sigma_u^2\mathbf{1}\mathbf{1}')\mathbf{1} \\
&= \frac{1}{n}\frac{1}{m^2}(\sigma_e^2 m+\sigma_u^2 m^2) \\
&= \frac{\sigma_e^2+m\sigma_u^2}{nm} \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

- Thus,

$$Var\left(\begin{bmatrix} \overline{y}_{1..} \\ \overline{y}_{2..} \\ . \\ . \\ . \\ \overline{y}_{t..} \end{bmatrix}\right) = \frac{\sigma^2}{n} \underbrace{I}_{t \times t} \text{ and}$$

$$Var\left(C \begin{bmatrix} \overline{y}_{1..} \\ . \\ . \\ . \\ \overline{y}_{t..} \end{bmatrix}\right) = \frac{\sigma^2}{n} CC'.$$

- Notice Var $\boldsymbol{C}\beta$ depends only on $\sigma^2 = \sigma_u^2 + \sigma_e^2/m$
- Don't need separate estimates of $\sigma_u^2$ and $\sigma_e^2$ to carry out inference for estimable $C\beta$.
- Do need to estimate $\sigma^2 = \sigma_u^2 + \frac{\sigma_e^2}{m}$.
  - Mixed model: $\hat{\sigma}^2 = \frac{MS_{eu(trt)}}{m}$
  - Analysis of averages: $\hat{\sigma}^2 = MSE$
- For example:

$$\begin{aligned} Var(\overline{y}_{1..} - \overline{y}_{2..}) &= Var(\overline{y}_{1..}) + Var(\overline{y}_{2..}) \\ &= 2\frac{\sigma^2}{n} = 2(\frac{\sigma_u^2}{n} + \frac{\sigma_e^2}{mn}) \\ &= \frac{2}{mn}(\sigma_e^2 + m\sigma_u^2) \\ &= \frac{2}{mn}E(MS_{eu(trt)}) \end{aligned}$$

- Thus,

$$\widehat{Var}(\overline{y}_{1..} - \overline{y}_{2..}) = \frac{2MS_{eu(trt)}}{mn}$$

- error d.f. are the same in both analyses:
  - Mixed model: $MS_{eu(trt)}$ has $t(n-1)$ df
  - Analysis of averages: $MSE$ has $t(n-1)$ df
- A $100(1-\alpha)\%$ Confidence interval for $\tau_1 - \tau_2$ is $\overline{y}_{1..} - \overline{y}_{2..} \pm t^{\frac{\alpha}{2}}_{t(n-1)} \sqrt{\frac{2MS_{eu(trt)}}{mn}}$.
- A test of $H_0 : \tau_1 = \tau_2$ can be based on $t = \frac{\overline{y}_{1..} - \overline{y}_{2..}}{\sqrt{\frac{2MS_{eu(trt)}}{mn}}} \sim t_{t(n-1)} \left( \frac{\tau_1 - \tau_2}{\sqrt{\frac{2(\sigma^2_e + m\sigma^2_u)}{mn}}} \right)$
- All assumes same number of observations per experimental unit

- What if the number of observations per experimental unit is not the same for all experimental units?
  Let's look at two miniature examples.
- Treatment 1: two eu's, one obs each
  Treatment 2: one eu. but two obs on that eu

| Treatment 1 | Treatment 2 |
|---|---|
| $y_{111}$ $y_{121}$ | $y_{211}$ $y_{212}$ |

$$\boldsymbol{y} = \begin{bmatrix} y_{111} \\ y_{121} \\ y_{211} \\ y_{212} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$X_1 = 1, \qquad X_2 = X, \qquad X_3 = Z$$

$$MS_{TRT} = \boldsymbol{y}'(P_2 - P_1)\boldsymbol{y} = 2(\overline{y}_{1..} - \overline{y}_{...})^2 + 2(\overline{y}_{2..} - \overline{y}_{...})^2 = (\overline{y}_{1..} - \overline{y}_{2..})^2$$

$$MS_{eu(TRT)} = y'(P_3 - P_2)y = (y_{111} - \overline{y}_{1..})^2 + (y_{121} - \overline{y}_{1..})^2 = \frac{1}{2}(y_{111} - y_{121})^2$$

$$MS_{ou(eu.TRT)} = y'(I - P_3)y = (y_{211} - \overline{y}_{2..})^2 + (y_{212} - \overline{y}_{2..})^2 = \frac{1}{2}(y_{211} - y_{212})^2$$

$$
\begin{aligned}
E(MS_{TRT}) &= E(\overline{y}_{1\cdot\cdot} - \overline{y}_{2\cdot\cdot})^2 \\
&= E(\tau_1 - \tau_2 + \overline{u}_{1\cdot} - u_{21} + \overline{e}_{1\cdot\cdot} - \overline{e}_{2\cdot\cdot})^2 \\
&= (\tau_1 - \tau_2)^2 + Var(\overline{u}_{1\cdot}) + Var(u_{21}) + Var(\overline{e}_{1\cdot\cdot}) + Var(\overline{e}_{2\cdot\cdot}) \\
&= (\tau_1 - \tau_2)^2 + \frac{\sigma_u^2}{2} + \sigma_u^2 + \frac{\sigma_e^2}{2} + \frac{\sigma_e^2}{2} \\
&= (\tau_1 - \tau_2)^2 + 1.5\sigma_u^2 + \sigma_e^2
\end{aligned}
$$

$$
\begin{aligned}
E(MS_{eu(TRT)}) &= \tfrac{1}{2}E(y_{111} - y_{121})^2 \\
&= \tfrac{1}{2}E(u_{11} - u_{12} + e_{111} - e_{121})^2 \\
&= \tfrac{1}{2}(2\sigma_u^2 + 2\sigma_e^2) \\
&= \sigma_u^2 + \sigma_e^2 \\
E(MS_{ou(eu,TRT)}) &= \tfrac{1}{2}E(y_{211} - y_{212})^2 \\
&= \tfrac{1}{2}E(e_{211} - e_{212})^2 \\
&= \sigma_e^2
\end{aligned}
$$

| SOURCE | EMS |
|---|---|
| *trt* | $(\tau_1 - \tau_2)^2 + 1.5\sigma_u^2 + \sigma_e^2$ |
| *eu*(*TRT*) | $\sigma_u^2 + \sigma_e^2$ |
| *ou*(*eu*, *TRT*) | $\sigma_e^2$ |

$$F = \frac{\frac{MS_{TRT}}{1.5\sigma_u^2+\sigma_e^2}}{\frac{MS_{eu(TRT)}}{\sigma_u^2+\sigma_e^2}} \sim F_{1,1}^{\left(\frac{(\tau_1-\tau_2)^2}{1.5\sigma_u^2+\sigma_e^2}\right)}$$

- The test statistic that we used to test $H_0 : \tau_1 = ... = \tau_t$ in the balanced case is not $F$ distributed in this unbalanced case:

$$\frac{MS_{TRT}}{MS_{eu(TRT)}} \sim \frac{1.5\sigma_u^2 + \sigma_e^2}{\sigma_u^2 + \sigma_e^2} F_{1,1}^{\left(\frac{(\tau_1-\tau_2)^2}{1.5\sigma_u^2+\sigma_e^2}\right)}$$

- We'd like our denominator to be an unbiased estimator of $1.5\sigma_u^2 + \sigma_e^2$ in this case.
- No MS with this expectation, so construct one
- Consider $1.5MS_{eu(TRT)} - 0.5MS_{ou(eu,TRT)}$
  The expectation is $1.5(\sigma_u^2 + \sigma_e^2) - 0.5\sigma_e^2 = 1.5\sigma_u^2 + \sigma_e^2$
- Use $F = \frac{MS_{TRT}}{1.5MS_{eu(TRT)} - 0.5MS_{ou(eu,TRT)}}$
- What is its distribution under Ho?

# COCHRAN-SATTERTHWAITE APPROXIMATION FOR LINEAR COMBINATIONS OF MEAN SQUARES

- Cochran-Satterthwaite method provides an approximation to the distribution of linear combinations of Chi-square random variables
- Suppose $MS_1, ..., MS_k$ are independent random variables and that $\frac{df_i MS_i}{E(MS_i)} \sim \chi^2_{df_i} \qquad \forall\ i = 1, ..., k.$
- Consider the random variable $S^2 = a_1 MS_1 + a_2 MS_2 + ... + a_k MS_k$, where $a_1, a_2, ..., a_k$ are known positive constants in $\mathbb{R}^+$
- C-S says $\frac{\nu_{cs} S^2}{E(S^2)} \sim \chi^2_{\nu_{cs}}$, where d.f., $\nu_{cs}$, can be computed
- Often used when some $a_i < 0$.
  Theoretical support in this case much weaker. Support of the random variable should be $\mathbb{R}^+$ but $S^2$ may be $< 0$.

- Deriving the approximate df for $S^2 = a_1 MS_1 + a_2 MS_2 + ... + a_k MS_k$:

$$\begin{aligned} E(S^2) &= a_1 E(MS_1) + ... + a_k E(MS_k) \\ Var(S^2) &= a_1^2 Var(MS_1) + ... + a_k^2 Var(MS_K) \\ &= a_1^2 [\frac{E(MS_1)}{df_1}]^2 2df_1 + ... + a_k^2 [\frac{E(MS_k)}{df_k}]^2 2df_k \\ &= 2 \sum \frac{a_i^2 [E(MS_i)]^2}{df_i} \end{aligned}$$

- A natural estimator of $Var(S^2)$ is $\hat{Var}(S^2) \equiv 2 \sum_{i=1}^{k} a_i^2 MS_i^2 / df_i$
- Recall that $\frac{df_i MS_i}{E(MS_i)} \sim X^2_{df_i} \; \forall \qquad i = 1, ..., k.$
- If $S^2 = a_1 MS_1 + ... + a_k MS_k$ is distributed like each of the random variables in the linear combination, $\frac{\nu_{cs} S^2}{E(S^2)} \approx X^2_{\nu_{cs}}$
- May be a stretch when some $\boldsymbol{a}_i < 0$

- Note that

$$E[\frac{\nu_{cs}S^2}{E(S^2)}] = \nu_{cs} = E(X^2_{\nu_{cs}})$$
$$Var[\frac{\nu_{cs}S^2}{E(S^2)}] = \frac{\nu^2_{cs}}{[E(S^2)]^2} Var(S^2)$$

- Equating this expression to $Var(X^2_{\nu_{cs}}) = 2\nu_{cs}$ and solving for $\nu_{cs}$ yields $\nu_{cs} = \frac{2[E(S^2)]^2}{Var(S^2)}$.
- Replacing $E(S^2)$ with $S^2$ and $Var(S^2)$ with $\hat{Var}(S^2)$ gives the Cochran-Satterthwaite formula for the approximate degrees of freedom of a linear combination of Mean Squares

$$\hat{\nu}_{cs} = \frac{(S^2)^2}{\sum_{i=1}^k a_i^2 MS_i^2/df_i} = \frac{(\sum_{i=1}^k a_i MS_i)^2}{\sum_{i=1}^k a_i^2 MS_i^2/df_i}$$

- $1/\nu_{cs}$ is linear combination of $1/df_i$
- We want approximate df for $1.5MS_{eu(trt)} - 0.5MS_{ou(eu,trt)}$
- That is:

$$\hat{\nu}_{cs} = \frac{\left(1.5MS_{eu(trt)} - 0.5MS_{ou(eu,trt)}\right)^2}{(1.5MS_{eu(trt)})^2/df_{eu(trt)} + (0.5MS_{ou(eu,trt)})^2/df_{ou(eu,trt)}}$$

- So our F statistic $\sim F_{1,\nu_{cs}}$
- Examples:

| eu(trt) | | ou(eu,trt) | | $1.5MS_1 + 0.5MS_2$ | $1.5MS_1 - 0.5MS_2$ |
|---|---|---|---|---|---|
| MS | df | MS | df | $\hat{\nu}_{cs}$ | $\hat{\nu}_{cs}$ |
| 1 | 5 | 0 | 20 | 5 | 5 |
| 0 | 5 | 0 | 20 | 20 | 20 |
| 1 | 5 | 1 | 20 | 8.65 | 2.16 |
| 1 | 5 | 1 | 200 | 8.86 | 2.21 |
| 4 | 5 | 1 | 20 | 5.86 | 4.19 |
| 1 | 5 | 4 | 20 | 18.86 | 0.38 |

- What does the BLUE of the treatment means look like in this case?

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$Var(y) = \sum = ZGZ' + R = \sigma_u^2 ZZ' + \sigma_e^2 I$$

$$= \sigma_u^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} + \sigma_e^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- It follows that

$$\hat{\beta}_{\boldsymbol{\Sigma}} = (X'\boldsymbol{\Sigma}^{-1}X)^{-}X'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

$$= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} y = \begin{bmatrix} \overline{y}_{1..} \\ \overline{y}_{2..} \end{bmatrix}$$

- Fortunately, this is a linear estimator that does not depend on unknown variance components.

- Consider a slightly different scenario:
  Treatment 1: eu #1 with 2 obs, eu #2 with 1 obs
  Treatment 2: 1 eu with 1 obs

$$\boldsymbol{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{211} \end{bmatrix} X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- In this case, it can be shown that $\hat{\beta}_{\mathbf{\Sigma}} = (X'\mathbf{\Sigma}^{-1}X)^{-}X'\mathbf{\Sigma}^{-1}\boldsymbol{y}$

$$= \begin{bmatrix} \frac{\sigma_e^2+\sigma_u^2}{3\sigma_e^2+4\sigma_u^2} & \frac{\sigma_e^2+\sigma_u^2}{3\sigma_e^2+4\sigma_u^2} & \frac{\sigma_e^2+2\sigma_u^2}{3\sigma_e^2+4\sigma_u^2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{211} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2\sigma_e^2+2\sigma_u^2}{3\sigma_e^2+4\sigma_u^2}\,\overline{y}_{11\cdot} + \frac{\sigma_e^2+2\sigma_u^2}{3\sigma_e^2+4\sigma_u^2}\,\overline{y}_{121} \\ \overline{y}_{211} \end{bmatrix}$$

- It is straightforward to show that the weights on $\overline{y}_{11.}$ and $y_{121}$ are

$$\frac{\frac{1}{Var(\overline{y}_{11.})}}{\frac{1}{Var(\overline{y}_{11.})}+\frac{1}{Var(y_{121})}} \text{ and } \frac{\frac{1}{Var(y_{121})}}{\frac{1}{Var(\overline{y}_{11.})}+\frac{1}{Var(y_{121})}} \text{ respectively}$$

- Of course,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \begin{bmatrix} \frac{2\sigma_e^2+2\sigma_u^2}{3\sigma_e^2+4\sigma_u^2}\,\overline{y}_{11.} & + & \frac{\sigma_e^2+2\sigma_u^2}{3\sigma_e^2+4\sigma_u^2}\,y_{121} \\ \overline{y}_{211} & & \end{bmatrix}$$

  is not an estimator because it is a function of unknown parameters.

- Thus we use $\hat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$ as our estimator (replace $\sigma_e^2$ and $\sigma_u^2$ by estimates in the expression above).

- $\hat{\beta}_{\widehat{\mathbf{\Sigma}}}$ is an approximation to the BLUE.
- $\hat{\beta}_{\widehat{\mathbf{\Sigma}}}$ is not even a linear estimator in this case.
- Its exact distribution is unknown.
- When sample sizes are large, it is reasonable to assume that the distribution of $\hat{\beta}_{\widehat{\mathbf{\Sigma}}}$ is approx. the same as the distribution of $\hat{\beta}_{\mathbf{\Sigma}}$.

$$
\begin{aligned}
Var(\hat{\beta}_{\mathbf{\Sigma}}) &= Var[(X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}\boldsymbol{y}] \\
&= (X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}\,Var(\boldsymbol{y})[(X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}]' \\
&= (X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1})\mathbf{\Sigma}\mathbf{\Sigma}^{-1}X(X'\mathbf{\Sigma}^{-1}X)^{-1} \\
&= (X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}X(X'\mathbf{\Sigma}^{-1}X)^{-1} \\
&= (X'\mathbf{\Sigma}^{-1}X)^{-1} \\
Var(\hat{\beta}_{\widehat{\mathbf{\Sigma}}}) &= Var[(X'\widehat{\mathbf{\Sigma}}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}y] =? \approx (X'\widehat{\mathbf{\Sigma}}^{-1}X)^{-1}
\end{aligned}
$$

# Summary of Main Points

- Many of the concepts we have seen by examining special cases hold in greater generality.
- For many of the linear mixed models commonly used in practice, balanced data are nice because

  1. It is relatively easy to determine degrees of freedom, sums of Squares and expected mean squares in an ANOVA table.
  2. Ratios of appropriate mean squares can be used to obtain exact F-tests.
  3. For estimable $C\beta$, $C\hat{\beta}_{\widehat{\boldsymbol{\Sigma}}} = C\hat{\beta}$.(The OLS estimator equals the GLS estimator).
  4. When $Var(C\hat{\beta}) = \text{constant} \times E(MS)$, exact inferences about $\boldsymbol{C}\beta$ can be obtained by constructing $t$ tests or confidence intervals

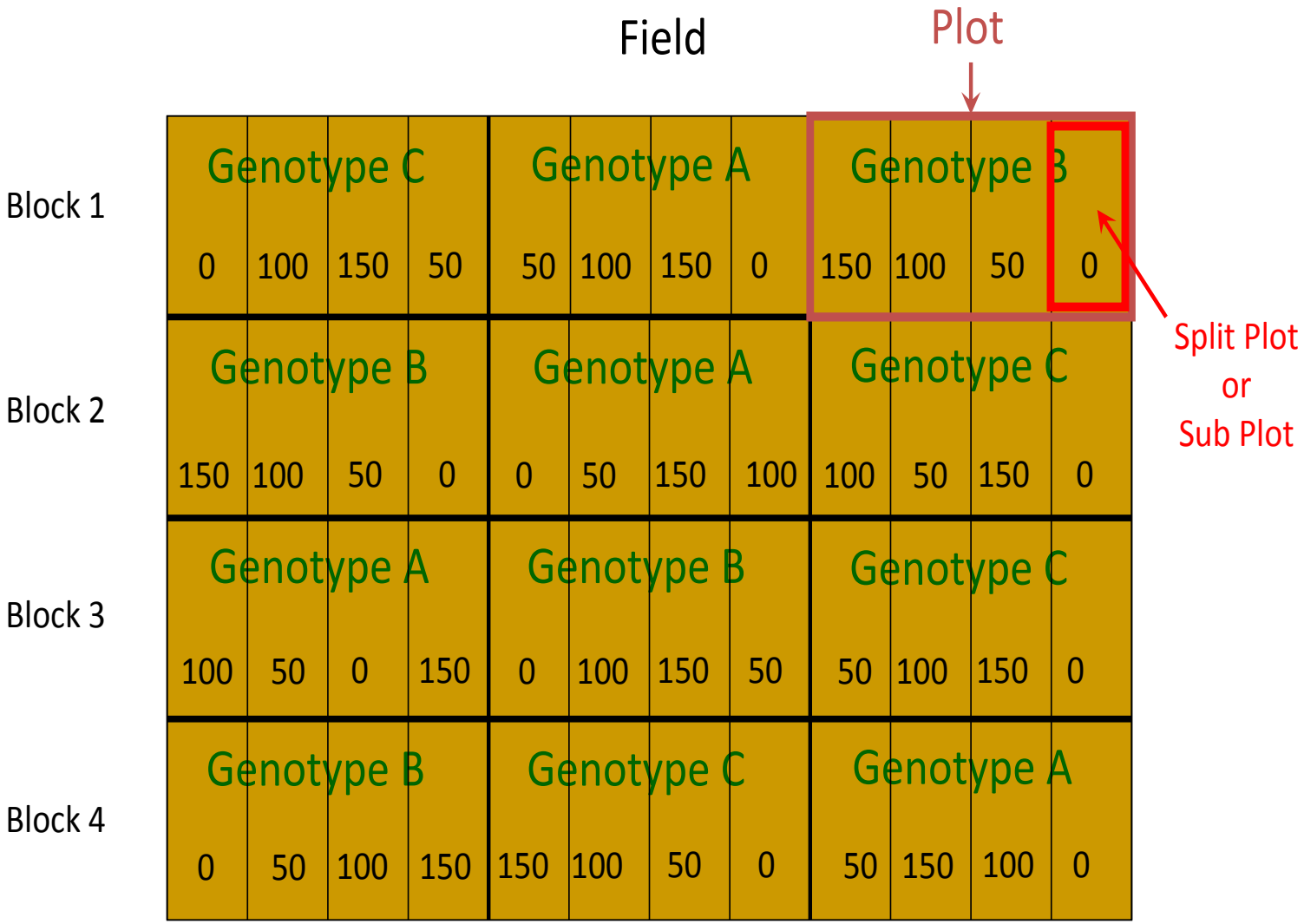
  $$t = \frac{C\hat{\beta} - C\beta}{\sqrt{\text{constant} \times (MS)}} \sim t_{\nu_{cs}(ms)}$$

  5. Simple analysis based on experimental unit averages gives the same results as those obtained by linear mixed model analysis of the full data set.

- When data are unbalanced, the analysis of linear mixed may be considerably more complicated.
  1. Approximate $F$ tests can be obtained by forming linear combinations of Mean Squares to obtain denominators for test statistics.
  2. The estimator $C\hat{\beta}_{\boldsymbol{\Sigma}}$ may be a nonlinear estimator of $C\beta$ whose exact distribution is unknown.
  3. Approximate inference for $C\beta$ is often obtained by using the distribution of $C\hat{\beta}_{\widehat{\boldsymbol{\Sigma}}}$, with unknowns in that distribution replaced by estimates.
- Whether data are balanced or unbalanced, unbiased estimators of variance components can be obtained by the method of moments.

# ANOVA ANALYSIS OF A BALANCED SPLIT-PLOT EXPERIMENT

- Example: the corn genotype and fertilization response study
- Main plots: genotypes, in blocks
- Split plots: fertilization
- 2 way factorial treatment structure
- split plot variability nested in main plot variability nested in blocks.


Field
Plot
Block 1
Genotype C
0 100 150 50
Genotype A
50 100 150 0
Genotype B
150 100 50 0
Block 2
Genotype B
150 100 50 0
Genotype A
0 50 150 100
Genotype C
100 50 150 0
Block 3
Genotype A
100 50 0 150
Genotype B
0 100 150 50
Genotype C
50 100 150 0
Block 4
Genotype B
0 50 100 150
Genotype C
150 100 50 0
Genotype A
50 150 100 0
Split Plot
or
Sub Plot

- A model: $y_{ijk} = \mu_{ij} + b_k + w_{ik} + e_{ijk}$
  $\mu_{ij}$=Mean for Genotype $i$, Fertilizer $j$
  Genotype $i = 1, 2, 3$, Fertilizer $j = 1, 2, 3, 4$, Block $k = 1, 2, 3, 4$

$$\boldsymbol{u} = \begin{bmatrix} b_1 \\ \vdots \\ b_4 \\ w_{11} \\ w_{21} \\ \vdots \\ w_{34} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{w} \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_b^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_w^2 \boldsymbol{I} \end{bmatrix} \right)$$

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_e^2 \boldsymbol{I} \end{bmatrix} \right)$$

| Source | DF | | |
|---|---|---|---|
| *Blocks* | $4-1$ | $=3$ | |
| *Genotypes* | $3-1$ | $=2$ | |
| *Blocks* $\times$ *Geno* | $(4-1)(3-1)$ | $=6$ | whole plot $\times$ *u* |
| *Fert* | $4-1$ | $=3$ | |
| *Geno* $\times$ *Fert* | $(3-1)(4-1)$ | $=6$ | |
| *Block* $\times$ *Fert* | $(4-1)(4-1)$ | | |
| $+$*Blocks* $\times$ *Geno* $\times$ *Fert* | $+(4-1)(3-1)(4-1)$ | 27 | split plot $\times$ *u* |
| *C.Total* | $48-1$ | 47 | |

| Source | Sum of Squares |
|---|---|
| *Blocks* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{..k}-\bar{y}_{...})^2$ |
| *Geno* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{i..}-\bar{y}_{...})^2$ |
| *Blocks* $\times$ *Geno* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{i.k}-\bar{y}_{i..}-\bar{y}_{..k}+\bar{y}_{...})^2$ |
| *Fert* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{.j.}-\bar{y}_{...})^2$ |
| *Geno* $\times$ *Fert* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2$ |
| *error* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{ijk}-\bar{y}_{i.k}-\bar{y}_{ij.}+\bar{y}_{i..})^2$ |
| *C.Total* | $\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{4}(\bar{y}_{ijk}-\bar{y}_{...})^2$ |

$$
\begin{aligned}
E(MS_{GENO}) &= \frac{FB}{G-1}\sum_{i=1}^{G} E(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \\
&= \frac{FB}{G-1}\sum_{i=1}^{G} E(\bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot} + \bar{w}_{i\cdot} - \bar{w}_{\cdot\cdot} + \bar{e}_{i\cdot} - \bar{e}_{\cdot\cdot\cdot})^2 \\
&= FB\frac{\sum_{i=1}^{G}(\bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot})^2}{G-1} + FB\,E\frac{\sum_{i=1}^{G}(\bar{w}_{i\cdot} - \bar{w}_{\cdot\cdot})^2}{G-1} + FB\,E\frac{\sum_{i=1}^{G}(\bar{e}_{i\cdot} - \bar{e}_{\cdot\cdot})}{G-1} \\
&= FB\frac{\sum_{i=1}^{G}(\bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot})^2}{G-1} + FB\frac{\sigma_w^2}{B} + FB\frac{\sigma_e^2}{FB} \\
&= FB\frac{\sum_{i=1}^{G}(\bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot})^2}{G-1} + F\sigma_w^2 + \sigma_e^2
\end{aligned}
$$

$$E(MS_{Block\times GENO}) = \frac{F}{(B-1)(G-1)}\sum_{i=1}^{G}\sum_{k=1}^{B} E(\bar{y}_{i.k} - \bar{y}_{i..} - \bar{y}_{..k}\bar{y}...)^2$$

$$= \frac{F}{(B-1)(G-1)}\sum_{i=1}^{G}\sum_{k=1}^{B} E(w_{ik} - \bar{w}_{i.} - \bar{w}_{.k} + \bar{w}_{..} + \bar{e}_{i.k} - \bar{e}_{i..} + \bar{e}_{..k} + \bar{e}...)^2$$

$$= \frac{F}{(B-1)(G-1)} E[\sum_{i=1}^{G}\sum_{k=1}^{B}(w_{ik} - \bar{w}_{i.})^2 - 2\sum_{i=1}^{G}\sum_{k=1}^{B}(w_{ik} - \bar{w}_{i.})(\bar{w}_{.k} - \bar{w}_{..})$$
$$+ \sum_{i=1}^{G}\sum_{k=1}^{B}(\bar{w}_{.k} - \bar{w}_{..})^2 + e \text{ terms}]$$

$$= \frac{F}{(B-1)(G-1)} E[\sum_{i=1}^{G}\sum_{k=1}^{B}(w_{ik} - \bar{w}_{i.})^2 - G\sum_{k=1}^{B}(w_{.k} - \bar{w}_{..})^2 + e \text{ terms}]$$

$$= \frac{F}{(B-1)(G-1)}[G(B-1)\sigma_w^2 - G(B-1)\sigma_w^2/G + e \text{ terms}]$$

- To test for genotype main effects, i.e.,

$$H_0 : \bar{\mu}_{1.} = \bar{\mu}_{2.} = \bar{\mu}_{3.} \iff H_0 : \frac{FB}{G-1} \sum_{i=1}^{G} (\bar{\mu}_{i.} - \bar{\mu}_{..})^2 = 0,$$

  compare $\frac{MS_{GENO}}{MS_{Blcok \times Geno}}$ to a central F distribution with $G-1$ and $(B-1)(G-1)$ *df*.
- Reject $H_0$ at level $\alpha$ iff $\frac{MS_{GENO}}{MS_{Blcok \times Geno}} \geq F^{\alpha}_{G-1,\,(B-1)(G-1)}$
- N.B. notation: $F^{\alpha}$ is the $1-\alpha$ quantile.

- Contrasts among genotype means:

$$\begin{aligned} Var(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}) &= Var(\bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot} + \bar{w}_{1\cdot} + \bar{w}_{2\cdot} + \bar{e}_{1\cdot\cdot} - \bar{e}_{2\cdot\cdot}) \\ &= \frac{2\sigma_w^2}{B} + \frac{2\sigma_e^2}{FB} \\ &= \frac{2}{FB}(F\sigma_w^2 + \sigma_e^2) = \frac{2}{FB}E(MS_{Block\times GENO}) \end{aligned}$$

$$\hat{Var}(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}) = \frac{2}{FB}MS_{Block\times GENO}$$

- Can use

$$t = \frac{\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} - (\bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot})}{\sqrt{\frac{2}{FB}MS_{Block\times GENO}}} \sim t_{(B-1)(G-1)}$$

to get tests of $H_0 : \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot}$
or construct confidence intervals for $\bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot}$

- Furthermore, suppose $C$ is a matrix whose rows are contrast vectors so that $\boldsymbol{C}\mathbf{1} = \mathbf{0}$. Then

$$Var\left(\boldsymbol{C}\begin{bmatrix}\bar{y}_{1\cdot}\\ \vdots \\ \bar{y}_{G\cdot}\end{bmatrix}\right) = Var\left(\boldsymbol{C}\begin{bmatrix}\bar{b}_{\cdot} + \bar{w}_{1\cdot} + \bar{e}_{1\cdot}\\ \vdots \\ \bar{b}_{\cdot} + \bar{w}_{G\cdot} + \bar{e}_{G\cdot}\end{bmatrix}\right)$$

$$= Var\left(\boldsymbol{C}\mathbf{1}\bar{b}_{\cdot} + \boldsymbol{C}\begin{bmatrix}\bar{w}_{1\cdot} + \bar{e}_{1\cdot}\\ \vdots \\ \bar{w}_{G\cdot} + \bar{e}_{G\cdot}\end{bmatrix}\right) = \boldsymbol{C}\, Var\left(\begin{bmatrix}\bar{w}_{1\cdot} + \bar{e}_{1\cdot}\\ \vdots \\ \bar{w}_{G\cdot} + \bar{e}_{G\cdot}\end{bmatrix}\right) \boldsymbol{C}'$$

$$= \boldsymbol{C}\left(\frac{\sigma^2_w}{B} + \frac{\sigma^2_e}{FB}\right)\boldsymbol{I}\boldsymbol{C}' = \left(\frac{\sigma^2_w}{B} + \frac{\sigma^2_e}{FB}\right)\boldsymbol{C}\boldsymbol{C}' = \frac{E(MS_{MS_{Block\times GENO}})}{FB}\boldsymbol{C}\boldsymbol{C}'$$

- An F statistic for testing

$$H_0 : \boldsymbol{C} \begin{bmatrix} \bar{\mu}_{1.} \\ \vdots \\ \bar{\mu}_{G.} \end{bmatrix} = \mathbf{0}, \text{ with } df = q, \ (B-1)(G-1), \text{ is}$$

$$F = \frac{\left( \boldsymbol{C} \begin{bmatrix} \bar{y}_{1.} \\ \vdots \\ \bar{y}_{G.} \end{bmatrix} \right)' \left[ \frac{MS_{Block \times GENO}}{FB} \boldsymbol{C}\boldsymbol{C}' \right]^{-1} \left( \boldsymbol{C} \begin{bmatrix} \bar{y}_{1..} \\ \vdots \\ \bar{y}_{G..} \end{bmatrix} \right)}{q}$$

- where q is the number of rows of $\boldsymbol{C}$ (which must have full row rank to ensure that the hypothesis is testable)

- Inference on fertilizer main effects, $\mu_{.j}$:

$$\begin{aligned} E(MS_{FERT}) &= \frac{GB}{F-1}\sum_{i=1} FE(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &= \frac{GB}{F-1}\sum_{i=1} FE(\bar{\mu}_{.j} - \bar{\mu}_{..} + \bar{e}_{.j} - \bar{e}_{..})^2 \\ &= \frac{GB}{F-1}\sum_{i=1} F(\bar{\mu}_{.j} - \bar{\mu}_{..})^2 + \sigma_e^2 \end{aligned}$$

- To test for fertilizer main effects, i.e.

$$Ho : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4} \Longleftrightarrow Ho : \frac{GB}{F-1}\sum_{i=1} F(\mu_{.i} - \mu_{..})^2 = 0,$$

compare $\frac{MS_{FERT}}{MS_Error}$ to a central F distribution with $F-1$ and $G(B-1)(F-1)$ df

- Contrasts among fertilizer means:

$$\text{Note: } \bar{y}_{.1.} - \bar{y}_{.2.} = (\mu_{.1} + \bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.1.}) - (\mu_{.2} + \bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.2.})$$

$$\begin{aligned} Var(\bar{y}_{.1.} - \bar{y}_{.2.}) &= Var(\bar{\mu}_{.1} - \bar{\mu}_{.2} + \bar{e}_{.1.} - \bar{e}_{.2.}) \\ &= \frac{2\sigma_e^2}{GB} = \frac{2}{GB} E(MS_{Error}) \end{aligned}$$

$$\hat{Var}(\bar{y}_{.1.} - \bar{y}_{.2.}) = \frac{2}{GB} MS_{Error}$$

- Can use

$$t = \frac{\bar{y}_{.1.} - \bar{y}_{.2.} - (\bar{\mu}_{.1} - \bar{\mu}_{.2})}{\sqrt{\frac{2}{GB} MS_{Error}}} \sim t_{G(B-1)(F-1)}$$

to get tests of $H_0 : \bar{\mu}_{.1} = \bar{\mu}_{.2}$
or construct confidence intervals for $\bar{\mu}_{.1} - \bar{\mu}_{.2}$

- Furthermore, suppose $C$ is a matrix whose rows are contrast vectors so that $\boldsymbol{C}\mathbf{1} = \mathbf{0}$. Then

$$Var\left(\boldsymbol{C}\begin{bmatrix}\bar{y}_{.1.}\\ \vdots \\ \bar{y}_{.F.}\end{bmatrix}\right) = Var\left(\boldsymbol{C}\begin{bmatrix}\bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.1.}\\ \vdots \\ \bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.F.}\end{bmatrix}\right)$$

$$= Var\left(\boldsymbol{C}\mathbf{1}\bar{b}_{.} + \boldsymbol{C}\mathbf{1}\bar{w}_{..} + \begin{bmatrix}\bar{e}_{.1.}\\ \vdots \\ \bar{e}_{.F.}\end{bmatrix}\right) = \boldsymbol{C}\, Var\left(\begin{bmatrix}\bar{e}_{.1.}\\ \vdots \\ \bar{e}_{.F.}\end{bmatrix}\right)\boldsymbol{C}'$$

$$= \boldsymbol{C}\left(\frac{\sigma_e^2}{GB}\right)\boldsymbol{I}\boldsymbol{C}' = \frac{E(MS_{Error})}{GB}\boldsymbol{C}\boldsymbol{C}'$$

- An F statistic for testing

$$H_0 : \boldsymbol{C} \begin{bmatrix} \bar{\mu}_{.1} \\ \vdots \\ \bar{\mu}_{.F} \end{bmatrix} = \mathbf{0}, \text{ with } df = q, \; G(B-1)(F-1), \text{ is}$$

$$F = \frac{\left( \boldsymbol{C} \begin{bmatrix} \bar{y}_{.1} \\ \vdots \\ \bar{y}_{.F} \end{bmatrix} \right)' \left[ \frac{MS_{Error}}{FB} \boldsymbol{C}\boldsymbol{C}' \right]^{-1} \left( \boldsymbol{C} \begin{bmatrix} \bar{y}_{.1.} \\ \vdots \\ \bar{y}_{.F.} \end{bmatrix} \right)}{q}$$

- where q is the number of rows of $\boldsymbol{C}$ (which must have full row rank to ensure that the hypothesis is testable)

- Inferences on the interactions: same as fertilizer main effects
  - Use $MS_{Error}$
- Inferences on simple effects: Two types, details different
- Diff. between two fertilizers within a genotype, e.g. $\mu_{11} - \mu_{12}$

$$\begin{aligned} Var(\bar{y}_{11.} - \bar{y}_{12.}) &= Var(\mu_{11} - \mu_{11} + \bar{b}_{.} - \bar{b}_{.} + \bar{w}_{..} - \bar{w}_{..} + \bar{e}_{11.} - \bar{e}_{12} \\ &= \frac{2\sigma_e^2}{B} \end{aligned}$$

$$\hat{Var}(\bar{y}_{11.} - \bar{y}_{12.}) = \frac{2}{B} MS_{Error}$$

- Diff. between two genotypes within a fertilizer, e.g. $\mu_{11} - \mu_{21}$

$$\begin{aligned} Var(\bar{y}_{11.} - \bar{y}_{21.}) &= Var(\mu_{11} - \mu_{21} + \bar{w}_{1.} - \bar{w}_{2.} + \bar{e}_{11.} - \bar{e}_{21.}) \\ &= \frac{2\sigma_w^2}{B} + \frac{2\sigma_e^2}{B} \\ &= \frac{2}{B}(\sigma_w^2 + \sigma_e^2) \end{aligned}$$

- Need an estimator of $\sigma_w^2 + \sigma_e^2$
- From ANOVA table:
  E $MS_{Block \times GENO} = F\sigma_w^2 + \sigma_e^2$
  E $MS_{Error} = \sigma_e^2$
- Use a linear combination of these to estimate $\sigma_w^2 + \sigma_e^2$
- $E\left(\frac{MS_{Block \times GENO}}{F} + \frac{F-1}{F} MS_{ERROR}\right) = \sigma_w^2 + \frac{\sigma_e^2}{F} + \frac{(F-1)\sigma_e^2}{F} = \sigma_w^2 + \sigma_e^2$
- $\hat{Var}(\bar{y}_{11.} - \bar{y}_{21.}) \equiv \frac{2}{BF} MS_{Block \times GENO} + \frac{2(F-1)}{BF} MS_{error}$ is an unbiased estimate of $Var(\bar{y}_{11.} - \bar{y}_{21.})$
- $\frac{\bar{y}_{11.} - \bar{y}_{21.} - (\mu_{11} - \mu_{21})}{\sqrt{\hat{Var}(\bar{y}_{11.} - \bar{y}_{21.})}} \approx t$ with $d.f.$ determined by the Cochran-Satterthwaite approximation.
- For simple effects in a split plot, required linear combination is usually a sum of MS. CS works well for sums.

# Inferences on means:

- E MS for random effect terms in the ANOVA table
  E $MS_{Block} = \sigma_e^2 + F\sigma_w^2 + GF\sigma_b^2$
  E $MS_{Block \times GENO} = \sigma_e^2 + F\sigma_w^2$
  E $MS_{Error} = \sigma_e^2$
- Cell mean, $\mu_{ij}$

$$\begin{aligned} Var\ \bar{y}_{ij.} &= Var(\mu_{ij} + \bar{b}_{.} + \bar{w}_{i.} + \bar{e}_{ij.}) \\ &= \frac{\sigma_b^2}{B} + \frac{\sigma_w^2}{B} + \frac{\sigma_e^2}{B} \end{aligned}$$

  - Need to construct an estimator:

$$\hat{Var}(\bar{y}_{ij.}) = \frac{1}{BGF}\left[MS_{Block} + (G-1)\, MS_{Block \times GENO} + F(G-1)\, MS_{Error}\right]$$

  - with approximate df from Cochran-Satterthwaite

- Main plot marginal mean, $\mu_{i.}$

$$\begin{aligned} Var\ \bar{y}_{i..} &= Var(\mu_{i.} + \bar{b}_{.} + \bar{w}_{i.} + \bar{e}_{i..}) \\ &= \frac{\sigma_b^2}{B} + \frac{\sigma_w^2}{B} + \frac{\sigma_e^2}{FB} \end{aligned}$$

  requires MoM estimate

- If blocks considered fixed:

$$\begin{aligned} Var\ \bar{y}_{i..} &= Var(\mu_{i.} + \bar{b}_{.} + \bar{w}_{i.} + \bar{e}_{i..}) \\ &= \frac{\sigma_w^2}{B} + \frac{\sigma_e^2}{FB} \\ &= \frac{1}{FB}\left(F\sigma_w^2 + \sigma_e^2\right) \end{aligned}$$

- Can estimate this by $\frac{MS_{Block \times GENO}}{FB}$ with $(B-1)(G-1)$ df

- Split plot marginal mean, $\mu_{.j}$

$$\begin{aligned} Var\ \bar{y}_{.i.} &= Var(\mu_{.j} + \bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.j.}) \\ &= \frac{\sigma_b^2}{B} + \frac{\sigma_w^2}{BG} + \frac{\sigma_e^2}{BG} \end{aligned}$$

- If blocks considered fixed:

$$\begin{aligned} Var\ \bar{y}_{.i.} &= Var(\mu_{.j} + \bar{b}_{.} + \bar{w}_{..} + \bar{e}_{.j.}) \\ &= \frac{\sigma_w^2}{BG} + \frac{\sigma_e^2}{BG} \end{aligned}$$

- Both require MoM estimates.

# Summary of ANOVA for a balanced split plot study

- Use main plot error MS for inferences on
  - main plot marginal means (e.g. genotype)
- Use split plot error MS for inferences on
  - split plot marginal means (e.g. fertilizer)
  - main*split interactions (e.g. geno $\times$ fert)
  - a simple effect within a main plot treatment
- Construct a method of moments estimator for inferences on
  - a simple effect within a split plot treatment
  - a simple effect between two split trt and two main trt, e.g. $\mu_{11} - \mu_{22}$
  - most means

# Unbalanced split plots

- Unequal numbers of observations per treatment
- Details get complicated fast
- In general:
  - coefficients for random effect variances in E MS do not match
  - so $MS_{Block \times GENO}$ is not the appropriate denominator for $MS_{GENO}$
  - need to construct MoM estimators for all F denominators
- All inference based on approximate F tests or approximate t tests

# Two approaches for E MS

- RCBD with random blocks and multiple obs. per block

$$Y_{ijk} = \mu + \beta_i + \tau_j + \beta\tau_{ij} + \epsilon_{ijk}$$

  where $i \in \{1, \ldots, B\}$, $j \in \{1, \ldots, T\}$, $k \in \{1, \ldots, N\}$.

- with ANOVA table:

| | Source | df |
|---|---|---|
| R | Blocks | B-1 |
| | Treatments | T-1 |
| R | Block$\times$Trt | (B-1)(T-1) |
| R | Error | BT(N-1) |
| | C. total | BTN - 1 |

- Expected Mean Squares from two different sources

| Source | 1: Searle (1971) | 2: Graybill (1976) |
|---|---|---|
| Blocks | $\sigma_e^2 + N\sigma_p^2 + NT\sigma_B^2$ | $\eta_e^2 + NT\eta_B^2$ |
| Treatments | $\sigma_e^2 + N\sigma_p^2 + Q(\tau)$ | $\eta_e^2 + N\eta_p^2 + Q(\tau)$ |
| Block$\times$Trt | $\sigma_e^2 + N\sigma_p^2$ | $\eta_e^2 + N\eta_p^2$ |
| Error | $\sigma_e^2$ | $\eta_e^2$ |

  - 1: Searle, S (1971) *Linear Models*
  - 2: Graybill, F (1976) *Theory and Application of the Linear Model*.
- Same except for Blocks
- Different F tests for Blocks
  - EMS 1: MS Blocks / MS Block*Trt
  - EMS 2: MS Blocks / MS Error

- Can find “rules” for computing coefficients in EMS
- Rules generally give EMS 2, e.g. Schulz (1954, Biometrics), Cornfield-Tukey (1956, Ann. Math. Stat.)
- Long-standing controversy / disagreement
  - EMS 1: dates from Mood (1950) *Introduction to the Theory of Statistics*, p 344
  - EMS 2: dates from Anderson and Bancroft (1952) *Statistical Theory in Research*
    Bancroft was first dept. head of ISU Statistics

# What's going on?

- Important point: Your model is important!!!
- Ignore "rules". Focus on the choice of model
- EMS 1: follows the model

$$\begin{aligned} \beta_i &\overset{i.i.d.}{\sim} N(0, \sigma_B^2) \\ \beta\tau_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma_p^2) \\ \epsilon_{ijk} &\overset{i.i.d.}{\sim} N(0, \sigma_e^2) \end{aligned}$$

- EMS 2: follows model 1, with addition that interactions sum to 0 within each level of the fixed effect.
- Forces a negative covariance between interactions within same treatment ($j$)
- Because if $\sum_i \alpha\beta_{ij} = 0$, then $\text{Var}\left(\sum_{\text{i}} \alpha\beta_{\text{ij}}\right) = 0$

$$
\begin{aligned}
\beta_i &\overset{i.i.d.}{\sim} N(0, \eta_B^2) \\
\beta\tau_{ij} &\sim N(0, (T-1)\eta_p^2/T) \\
\text{Cov}(\alpha\beta_{\text{ij}}, \alpha\beta_{\text{i}'\text{j}}) &= -\eta_p^2/T \\
\text{Cov}(\alpha\beta_{\text{ij}}, \alpha\beta_{\text{ij}'}) &= 0 \\
\text{Cov}(\alpha\beta_{\text{ij}}, \alpha\beta_{\text{i}'\text{j}'}) &= 0 \\
\epsilon_{ijk} &\overset{i.i.d.}{\sim} N(0, \eta_e^2)
\end{aligned}
$$

- Resolution: different parameterizations
- Hocking 1985, *The Analysis of Linear Models*, section 10.4 has an especially clear explanation.
- Easy conversion between parameterizations

$$\begin{array}{rcl} \text{EMS 1} & & \text{EMS 2} \\ \sigma_e^2 & = & \eta_e^2 \\ \sigma_p^2 & = & \eta_p^2 \\ \sigma_B^2 & = & \eta_B^2 - \eta_p^2/T \end{array}$$

- Equivalent models

$$\begin{array}{rcl} \text{(EMS 1) } \sigma_e^2 + N\sigma_p^2 + NT\sigma_B^2 & = & \eta_e^2 + N\eta_p^2 + NT(\eta_B^2 - \eta_p^2/T) \\ & = & \eta_e^2 + N\eta_p^2 + NT\eta_B^2 - N\eta_p^2 \\ & = & \eta_e^2 + NT\eta_B^2 \text{ (EMS 2)} \end{array}$$

- F tests for “Blocks” are tests of different quantities

- Different software implements different algorithms:
  - SAS: Searle, EMS 1
  - Stata (I'm told): Anderson and Bancroft, EMS 2
  - R: Searle model (but doesn't use ANOVA)
- Difference really matters for unbalanced data.
  - There, EMS for treatments not the same

- I believe there is no reason for the sum-to-zero assumption
  - A legacy of ANOVA for fixed effects that is forced to be full-rank
  - Hocking (1985) has another technical reason
- Much more natural to assume independent random effects
- N.B. More natural doesn't mean must be correct!