Workshop:
Mixed Models for Analysis of Factorials in Plant Pathology

Larry Madden
(Ohio State University)

Alissa Kriss
(USDA-ARS, Ft. Pierce, FL)

APS Annual Meeting, 2012

Conceptual Background

- It is common in plant pathology (and in many other fields) to investigate the effects of *two* or more **factors** on a **response variable** of interest
- **Factor**:
  - An explanatory variable that may affect the response variable, which is "manipulated" by the investigator in an experiment
    - Sometimes a variable that is just measured in an observational study and not manipulated in a planned/replicated/randomized experiment
    - May be called a predictor variable or "independent variable"
    - Experiments with two or more factors are called **factorials**
  - Often the term "factor" is used for a ***classification* or *class* variable**, consisting of two or more discrete *levels* (*treatment 1, treatment 2, …;* or *group 1, group 2, …*)
    - Note: **"*treatment*"** or **"*group*"** can refer to a specific level of the factor
- **Covariable** (in contrast to a factor):
  - A continuous explanatory variable that is either manipulated or measured by the investigator

Conceptual Background, *continued*

- **Response variable** (dependent variable):
  - A random variable that is measured or observed
    - **Continuous** (normal, etc.) or discrete (binary [0,1], count, ordinal)
- *Investigations are carried out to determine if one or more factors or covariables (explanatory variables) affect the response variable*
- Measurements or observations are obtained from:
  - *Planned experiments* with randomization and replication; or from
  - *Observational studies* (where randomization is not possible)
- Data analysis (*equivalent to fitting a statistical model*) is key to determining if factors or covariables (explanatory variables, in general) affect the response variable
  - The analysis (modeling) must take into account both the *treatment design* and *experimental design*

Random effects are key to accounting for experimental designs

  - **Treatment design** includes the number of factors and/or covariables, and whether or not all treatment (factor level) combinations are present
  - **Experimental design** includes the manner in which treatment combinations are assigned to experimental units (*Examples: completely randomized, blocking, split-plots, repeated measures*)

Effects of density of *Verticillium dahliae* and *Pratylenchus penetrans* in the soil on potato tuber yield (Rowe et al. [*Phytopathology* 75:412-418]) – a randomized experiment

Two-way Factorial (example)

***Verticillium* (per 10 g)**	***Pratylenchus* (per 100 cc)** 0 Tuber weight (g)	9 Tuber weight (g)	30 Tuber weight (g)	106 Tuber weight (g)	**All** **Tuber weight (g)**
0	**320**	**249**	**325**	**323**	**302**
30	**294**	**162**	**107**	**156**	**183**
300	**114**	**143**	**78**	**70**	**99**
All	**245**	**188**	**160**	**177**	**192**

Does Verticillium (overall) affect yield?
Does Pratylenchus (overall) affect yield?

Does Verticillium (alone) affect yield?
Does Pratylenchus (alone) affect yield?

Does effect of Verticillium depend on Pratylenchus? (interaction?)
Does effect of Pratylenchus depend on Verticillium? (interaction?)

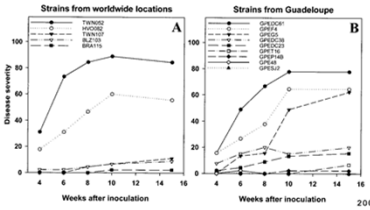
Effects of compost treatment (yes [2] or no [1]) on percent ground coverage after seeding with three turfgrass varieties [1,2,3] (Lochinkohl & Boehm. *HortScience* 36:790-794)

Two-way Factorial (example)

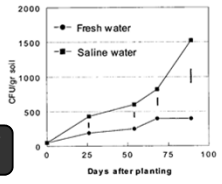
Turf variety	**Compost** 1 % coverage	2 % coverage	All % coverage
1	**58.3**	**78.3**	68.3
2	**71.7**	**85.0**	78.3
3	**63.3**	**80.0**	71.7
All	64.4	81.1	72.8

- **Does compost treatment (overall) affect turf ground cover?**
- **Is the compost effect consistent for all turf varieties? (interaction?)**
- **Does turfgrass variety (overall) affect ground cover?**
- **Does the variety effect depend on compost? (interaction?)**
- **How does one incorporate the experimental layout (split plot) into the analysis?**

Two or more factors, including Time



Xanthomonas albilineans on sugarcane (greenhouse study). Different **strains, strain sources,** and **times** on disease. (Champoiseau et al. (2006)



Fusarium density on roots of tomato (field study). Different **water sources** and **times**. Triky-Dotan et al. (2005)

Does the effect of treatment (or strain) depend on time? Does the change with time depend on treatment?

Multiple factors

- Fusarium head blight severity and DON toxin in wheat (Odenbach et al. (2008*). Proc. National Fusarium Head Blight Forum*).
  - **Planting date ×**
  - **Cultivars (with different levels of quantitative resistance) ×**
  - **Time of infection (inoculation) ×**
  - **Inoculum density**
- Effects of pre- and post-harvest treatments on gray mold of red raspberry (Ellis, Madden, Wright, Wilson. (2008). *Plant Health Progress*).
  - **In-field fungicide treatment (*pre-harvest*) ×**
  - **Harvest time [repeated measure] ×**
  - **Post-harvest incubation conditions (room temperature or cold) ×**
  - **Incubation time *post-harvest* [repeated measure]**

Outline

- Conceptual overview of factorial treatment structure, and examples
- Statistical linear models
  - Notation, especially for **means** (μ_{ij}) and **effects** [α_i, β_j, $(\alpha\beta)_{ij}$]
  - **Main effects, interactions, simple effects, slices**
  - Hypotheses, tests
- Example (two factors)
  - Use of SAS GLIMMIX procedure (with emphasis on features especially useful for factorials), graphs and tables
- Conceptual overview of random effects, experimental structure, and **mixed effect models**
  - Hypotheses and tests
- Example: 3 factors
- Contrasts for customized tests
- Examples revisited
- Experiments with a continuous explanatory variable (covariable)
- Conclusions

Model

- *Abstraction of a real phenomenon or process that emphasizes those aspects relevant to the objectives of the user*
  - **Used to describe, understand, predict, compare, and make inferences about the phenomenon**
  - Models consist of terms that are:
    - **deterministic** (**systematic, structural**), for the portion of the model that does not involve uncertainty; and/or
    - **stochastic** (random)
  - Often, stochastic terms can lead to a parsimonious abstraction of the phenomenon
- **Statistical model:**
  - ***Model with stochastic (random) components and deterministic components, containing unknown constants (i.e., parameters) to be estimated***
    - ANOVA and regression models are statistical models

Statistical Model:

Response = (systematic part) + (random part)

Response = structure + error

Outcome of interest, being measured or observed (***Y*****); a random variable.**

Mean or *expected* value of response (**μ**)

Function of variables (factors) and parameters; often of primary interest.

Residual: Difference between observed responses (i.e., the observations) and mean responses based on parameters ($e = Y - \mu$); a **random variable** (assume normal here). *Can be expanded for more random terms.*

Parameters (unknown constants to be estimated) given by Greek letters; variables given by Roman letters.

$\mu = f(X_1, X_2, \ldots; \theta, \tau, \alpha, \beta, \ldots)$

Model for one factor (*effects notation*):

We need at least **two** subscripts:

i (for factor [treatment or group] level designation)

k (observation or replicate designation, for each treatment or group)

Thus,

Y_{ik} : The *k*-th observation of the *i*-th treatment

Y_{31}: The 1st observation of the 3rd treatment (example)

For a situation with one factor, τ_i can represent the effect of the *i*-th factor level (treatment *i*) on the response variable

$$Y_{ik} = \theta + \tau_i + e_{ik}, \quad e_{ik} \sim N(0, \sigma_e^2)$$

θ: constant (intercept)

e.g., Effect of treatment 2:
τ_2 is the effect of treatment 2 (could even be zero). Indicates how much the mean for treatment 2 is above or below the overall constant (θ)

$$Y_{ik} = \theta + \tau_i + e_{ik}, \quad e_{ik} \sim N(0, \sigma_e^2)$$ ***Effects* notation**

Determining expected values (population means) for different factor levels:

$E(\bullet)$ *Expectation* operator: $E(Y_{ik}) = \mu_i$ (i.e., mean for treatment *i*)
(this notation is just giving the definition of an expected value)

$$E(Y_{ik}) = \mu_i = E(\theta + \tau_i + e_{ik}) = E(\theta) + E(\tau_i) + E(e_{ik}) = \theta + \tau_i + 0 = \theta + \tau_i$$

Thus, $\mu_i = \theta + \tau_i$, and $Y_{ik} = \mu_i + e_{ik}$, $e_{ik} \sim N(0, \sigma_e^2)$

***Means* notation**

Hypothesis testing:
Null hypothesis can be written in different, but equivalent, ways

$$H_0: \mu_1 = \mu_2 = \ldots . \mu_{last}$$
$$H_0: \tau_1 = \tau_2 = \ldots = 0$$

A classical one-way linear model

$Y_{ik} = \theta + \tau_i + e_{ik}$, $e_{ik} \sim N(0, \sigma_e^2)$ or

$\mu_i = \theta + \tau_i$, $Y_{ik} = \mu_i + e_{ik}$, $e_{ik} \sim N(0, \sigma_e^2)$

Y_{ik}: response (dependent variable) for the k-th observation in group (or treatment) i

θ: constant ("intercept")

τ_i: Group or treatment effect (effect of group or treatment i on response) - parameters

e_{ik}: Error associated with group (treatment) i and observation k (random variable, normal distribution). Residual.

Linear:
Sum of ***Variables x constants***, or just constants

This is considered a linear **fixed effects** model.

Definition: *a linear model with only parameters (constants) (e.g., effects of treatment) and one random variable (the error [residual], in this case).*

Single-factor example (SAS MIXED) output

$\hat{\theta}$ = 3.00
$\hat{\tau}_1$=2.67
$\hat{\tau}_2$=8.33
$\hat{\tau}_3$= 0

$\hat{\mu}_1 = \hat{\theta}+\hat{\tau}_1 =$ 3.00+2.67= 5.67
$\hat{\mu}_2 = \hat{\theta}+\hat{\tau}_2 =$ 3.00+8.33= 11.33
$\hat{\mu}_3 = \hat{\theta}+\hat{\tau}_3 =$ 3.00+0 = 3.00

The Mixed Procedure

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		3.0000	0.7201	6	4.17	0.0059
trt	biocon	2.6667	1.0184	6	2.62	0.0297
trt	control	8.3333	1.0184	6	8.18	0.0002
trt	fungic	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trt	2	6	34.93	0.0005

Least Squares Means

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
trt	biocon	5.6667	0.7201	6	7.87	0.0002
trt	control	11.3333	0.7201	6	15.74	<.0001
trt	fungic	3.0000	0.7201	6	4.17	0.0059

The *last* effect parameter is always 0 (what matters is the *sum* of the intercept and the τ_i).

Two or more fixed-effect factors

- An experiment may include two or more factors, such as:
  - Fungicide and plant cultivar
  - Irrigation (none, low, high) and crop rotation (corn-corn, corn-soybean)
  - Pathogen strain (or genotype), host genotype, tillage
  - Temperature and wetness duration
  - Fungicide treatment and time of observation (*repeated measure*)
  - Cultivar and distance from an inoculum source (*spatially repeated measure*)
- Note: the term "*treatment*" could be used as a label for *one* of the factors (e.g., "fungicide treatment"), but "*treatment*" can also be used more broadly as a general term to encompass the collection of factors
- With ≥ 2 factors, the **treatment structure or design** is called a ***factorial***
  - The factorial is completely ***crossed*** if each level of one factor is combined with each level of the other factors
    - E.g., if there are 2 cultivars and 2 cropping systems (two factors), then the design is crossed if all four combinations (2x2) are included in the study
  - If the levels of one factor are not identical at the levels of the other factors, then the design is ***nested*** (not explicitly considered here)
    - E.g., cultivars for the first cropping system are different from the cultivars in the second cropping system

Two fixed-effect factors

- The first factor is generically known as A (e.g., plant genotype), and the *effect of* A is given by **α**
- The second factor is generically known as B (e.g., fungicide treatment), and the *effect of* B is given by **β**
- Additional factors given by *C* (**γ**), etc. (Greek letters)
- With two factors, we need at least **three subscripts**:

  i (for factor A; i = 1, …, I). There are I levels of factor A

  j (for factor B; j = 1, …, J). There are J levels of factor B

  k (observation [or replicate] designation, for each combination of level i of A and level j of B)

  Thus,

  Y_{ijk} : The k-th observation of the i-th level of factor A and the j-th level of factor B

  Y_{321}: The 1st observation of the 3rd level of A and 2nd level of B (e.g., first replicate of cultivar 3 and fungicide 2)

Two fixed-effect factors: Expected values (means): μ_{ij}, $\mu_{i\bullet}$, $\mu_{\bullet j}$, $\mu_{\bullet\bullet}$

μ_{ij} (Level i of A, j of B)

Effect:	Factor B:		
Factor A:	1	2	mean
1	μ_{11}	μ_{12}	$\mu_{1\bullet}$
2	μ_{21}	μ_{22}	$\mu_{2\bullet}$
3	μ_{31}	μ_{32}	$\mu_{3\bullet}$
mean	$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\mu_{\bullet\bullet}$

Effect:	B:		
A:	1	2	mean
1	10	20	15
2	15	30	22.5
3	20	40	30
mean	15	30	22.5

Interaction mean

Main effect means.
"Main Effect" can also describe a contrast (e.g., difference) of the main-effect means.

Main effect means.
"Main Effect" can also describe a contrast (e.g., difference) of the main-effect means

Grand mean

Expected values (= **LSMEANS**) are not necessarily simple arithmetic averages of observations.

Two fixed-effect factors

- With two factors, a standard linear statistical model is:

  $Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$, $e_{ijk} \sim N(0, \sigma_e^2)$

- α_i : *Main effect of* factor A on the expected value (analogous to τ_i)
- β_j : *Main effect of* factor B on the expected value
- $(\alpha\beta)_{ij}$: The *interaction* effect (combined or joint effect of both factors)
  - **An interaction occurs when the effect of A depends on the level of B, or the effect of B depends on the level of A (other meanings of interaction will follow)**
  - The $(\alpha\beta)_{ij}$ term adjusts the main effects up or down, depending on the level of the *other* factor
- Usually, interest is on the means (expected values; μ_{ij}), and the "effects" terms (e.g., α_i, β_j, $(\alpha\beta)_{ij}$) reside somewhat in the *background* (although they are used for all tests and for determining the expected values)
- Alternative formulation:

  $\mu_{ij} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} \Rightarrow Y_{ijk} = \mu_{ij} + e_{ijk}$, $e_{ijk} \sim N(0, \sigma_e^2)$

Two fixed-effect factors

- With two factors, a standard linear statistical model is:

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \quad , \quad e_{ijk} \sim N(0, \sigma_e^2) \quad \text{or}$$

$$\mu_{ij} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad \Rightarrow \quad Y_{ijk} = \mu_{ij} + e_{ijk} \quad , \quad e_{ijk} \sim N(0, \sigma_e^2)$$

μ_{ij} : sometimes known as an interaction mean (not necessarily a simple arithmetic average from observations; rather, an estimate from a model)

$\mu_{i\bullet}$: main-effect mean for A (average over the levels of B).
If two levels of factor B, then main effect for level 2 of A:

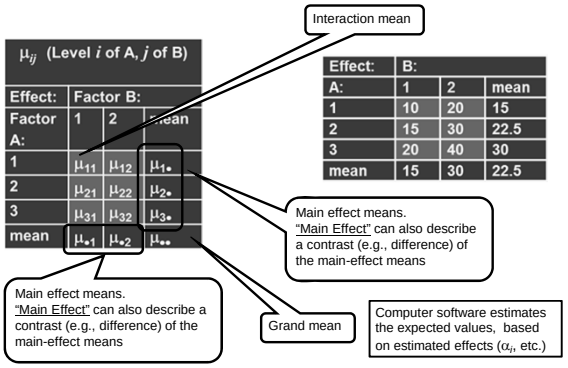
$$\mu_{2\bullet} = (\mu_{21} + \mu_{22})/2$$

$\mu_{\bullet j}$: main-effect mean for B (average over the levels of A).
If three levels of factor A, main effect for level 2 of B:

$$\mu_{\bullet 2} = (\mu_{12} + \mu_{22} + \mu_{32})/3$$

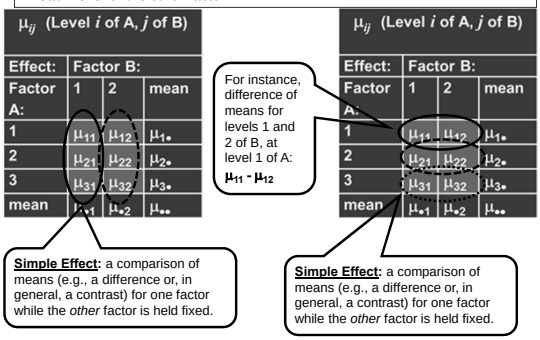
If there is no interaction [$(\alpha\beta)_{ij} \equiv 0$], one can base all analyses on the main-effect means, greatly increasing the power to detect true differences of the expected values

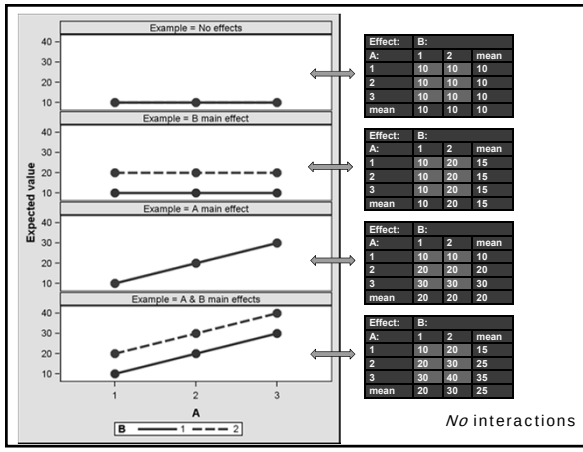
Two (or more) fixed-effect factors



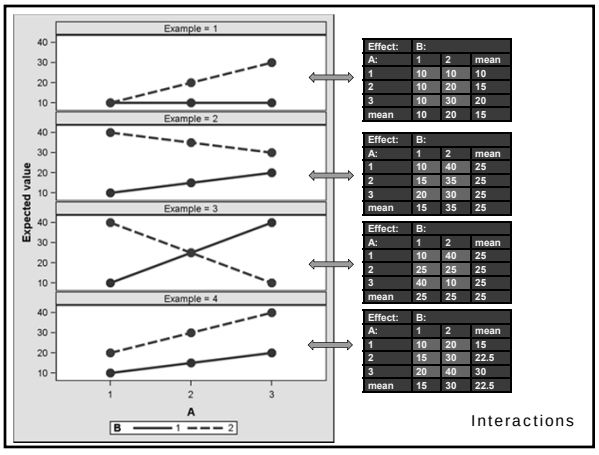
Two (or more) fixed-effect factors

- A **Slice** is a test of a simple effect.
- An **interaction** is present when the simple effects are not the same for each level of the *other* factor.





No interactions



Interactions

Two fixed-effect factors: Linear model

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \qquad e_{ijk} \sim N(0, \sigma_e^2)$$

Y_{ijk}: response (dependent variable) for the *i-th* level of factor A, *j-th* level of factor B, and *k-th* replicate

θ: constant ("intercept")

α_i: Effect of the *i*-th level of factor A on *Y*

β_j: Effect of the *j*-th level of factor B on *Y*

$(\alpha\beta)_{ij}$: Interaction effect (effect of *i*-th level of A and *j*-th level of B on *Y*)

e_{ijk}: Residual (error) (a random variable, with variance σ_e^2)

Example:
Effect of density of the fungus *Verticillium dahliae* (factor A; "Vert") and the lesion nematode *Pratylenchus penetrans* (factor B; "Prat") on yield of potatoes (*Y*). Experimental unit was a plant (in a micro-plot), and all units were randomized.

Protocol for analysis: Preliminary

- Fit model to data (based on the type of data (assume normal here) and the methods used to collect the data [experimental structure])
  - Use statistical procedure appropriate for one or more random effects, with a wide range of post-modeling methods (contrasts, graphs)
  - We use GLIMMIX procedure in SAS (although other procedures could be used for the first example, this PROC has a very wide range of post-model-fitting methods (graphs, mean comparisons, contrasts)
- Evaluate goodness of fit (graphically) to determine if the selected model is reasonable for the data
  - If not reasonable, consider other models or transformations of the data. Could use generalized linear mixed models (for non-normal) (not here)
    - Example: log or square-root transformation
- Determine if factors (or covariables), and their interactions, have significant effects on response variable
  - Typically, use Type 3 (III) tests (*F* tests)
  - If significant, determine various contrasts (such as pairwise mean differences, **slices**, etc.) to elucidate the nature of the significant effects
  - Always consider the *interactions first*, and consider main effects mostly when there is no interaction (graphs are usually a good idea)

Potato response to the pathogens causing early dying. Input variables (as columns), including disease rating, and weight of foliage, roots, and tubers

Transform, as needed.

FA 1.sas

```
data ped80;
input  treat  Vert  Prat  obs   rating tops roots tubers;
       stubers=sqrt(tubers); *<--transformations;
       sroots=sqrt(roots);
datalines;
0 0  0     1          0          490        2.9        752
0 0  0     2          0          440        2.8        130
0 0  0     3          2          70         0.3        132
…
;
```

To get graphs in 9.2 or 9.22, one must invoke ODS HTML and ODS GRAPHICS (automatic in 9.3). Plot options in red are used for model diagnostics and for looking at means

```
ods html;
ods graphics on;
proc glimmix data=ped80 plots=studentpanel ;
CLASS Vert Prat;
MODEL tubers = Vert Prat Vert*Prat / solution ;
LSMEANS Vert Prat Vert*Prat
          /diff  plots=(mean(join cl) diff(center));
run;
ods graphics off;
ods html close;
```

Identify factors with CLASS

MODEL gives the model, /solution option to see estimated effects (α, etc.) . The intercept (θ) and residual (ε) are automatically part of the MODEL

LSMEANS gives estimated expected values and ancillary statistics

Protocol for analysis of fixed-effects: *Synopsis*

- Use *plots=studentpanel* option on the GLIMMIX procedure statement to assess appropriateness of the model fit (adjust if necessary)
  - Look for random plot of residuals (versus predicted *Y* values), straight-line for normal plot, and when there are many observations, bell-shaped histogram of residuals – *Transform data if necessary*
- Look at *Type III Test* results for fixed effects
  - If interaction is significant, simple effects are not the same across the levels of the *other* factor
  - Look at the simple effects (*slices*), graphically and in tabular form (*slice* and *plots* options with the *LSMEANS* statement)
- If interaction is not significant, focus analysis on main effects (no slices)
- If relevant or of interest, do *mean separations* for main-effect means (*lines* option in *LSMEANS*), or for the interaction means (*sliceby* and *lines* options in the *SLICE* statement
  - **As a general rule, do <u>not</u> consider mean separations (or other tests) if the overall test for a factor or interaction is not significant**
  - Many statisticians do not like these all-possible pairwise mean separations, and there are more refined alternatives. However, the pairwise comparisons do provide a quick way to '*see what is going on*'

Protocol for analysis of fixed-effects: *Synopsis (continued)*

- With quantitative levels of a factor (0, 30, 300, etc.) there are better approaches for analysis of the means than pairwise mean separations
  - Mean comparisons do not account for the monotonic (increasing or decreasing) levels of the factor or interaction (discussed later)
- Increased sophistication in hypothesis testing of expected values is achieved with **contrasts**. Can be done using ESTIMATE, CONTRAST, or LSMESTIMATE statements (to be discussed later)
- Use judgment in interpreting *P* value for overall tests and for individual pairwise tests of expected values
  - Without going into the precise details, *P* gives evidence for or against the null hypothesis (but it is <u>not</u> the probability of H_0 being true or not)
  - Should interpretation be different if $P = 0.045$ versus $P = 0.055$?
    - A *Neyman-Pearson* (pure) hypothesis tester (decision maker) would say *yes!*
    - A more contemporary data analyst – who wishes to estimate the magnitude of the effects of factors on responses – would say *no!*
  - There are strong arguments that one should adjust (correct) individual *P* values when multiple tests (as with mean separations) are being performed, in order to 'control' the overall *P* value (for the collection of tests) – *discussed later with a three-factor example (see FA 3.sas)*

Fixed *versus* Random Effects

- As relevant for the experimental structure and effects in the model, make sure to properly incorporate random effects in the model, before performing any assessment of fixed effects
- **<u>Fixed-effects variable (or factor)</u>**
  - Levels in the study (i.e., the particular groups or treatments) represent all possible levels of the factor, or all levels of interest by the investigator (levels deliberately chosen by the investigator)
    - e.g., fungicide treatment, biocontrol treatment, inoculum dose, cultivar,…
- **<u>Random-effects variable (or factor)</u>**
  - Levels in the study represent only a *random sample* of a larger set of potential levels, or one is not interested in the specific result for each level in the study, or the effects on *Y* are stochastic
    - **e.g., block, location, plot (experimental unit), host or pathogen genotype (sometimes), etc.**
    - **Inference is on the population of possible levels, not just on those in the data set**

We mostly consider here random effects that are a **consequence of the experimental design** (i.e., from clustering of data, such as splitting and repeated measures).

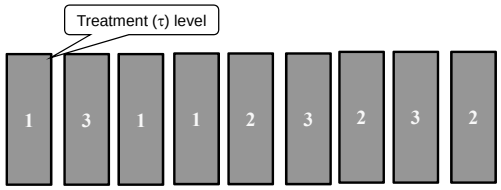
Random Effects – *elaboration…*

- A random effect is a random variable
  - Part of the stochastic component of models
    - e_{ijk} is partitioned into two or more terms: ($b_k + d_{ik} + e_{ijk}$)
- Random effects arise from
  1. **Random selection** of the levels of factor studied, or from:
  2. **Clustering of data**
     - **Cluster:** *collection of observations that are somehow stochastically related* (*correlated)*.
     - Experimental design and type of data collection "create" (induce) the clustering
  - **Mechanisms for clustering**:
    1. **Splitting: randomly assigning levels of one factor *within* experimental units of another factor**
    2. **Sampling and sub-sampling – multiple observations within the same experimental unit (*nesting*)**
    3. **Repeated observations of same experimental unit over time (or space)**

One Factor (τ [3])

No Clustering

Completely Randomized (3 replicates)



No clustering. Random variation among plots (σ_e^2), characterized through e_{ik} in model

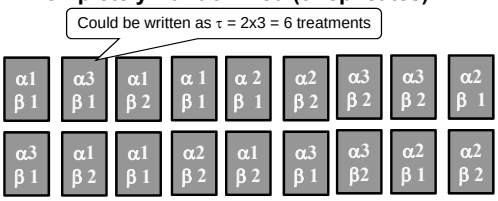
$$Y_{ik} = \theta + \tau_i + e_{ik}, \quad e_{ik} \sim N(0, \sigma_e^2)$$

```
proc glimmix data=datafile;
CLASS A;
MODEL Y = A / solution ;
LSMEANS A / diff;
run;
```

Two Factors (α [3], β [2])

No Clustering

Completely Randomized (3 replicates)



No clustering. Random variation among plots (σ_e^2), characterized through e_{ijk} in model

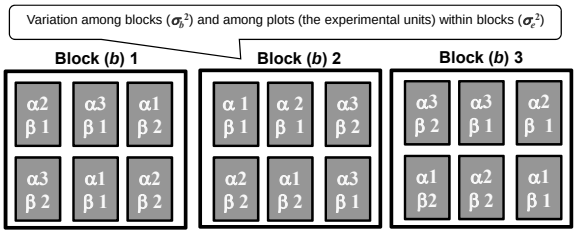
$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad e_{ijk} \sim N(0, \sigma_e^2)$$

```
proc glimmix data=datafile;
CLASS A B;
MODEL Y = A B A*B / s ;
LSMEANS A B A*B / diff;
run;
```

Splitting: Two Factors (α [3], β [2])

Clustering

Randomized Complete Block



Each block is a cluster (randomly assigning combinations of levels of A (α) and B (β) within each cluster). Observations within a block are correlated.

Two factors (with blocking)

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + e_{ijk}, \quad b_k \sim N(0, \sigma_b^2), \; e_{ijk} \sim N(0, \sigma_e^2)$$

- Y_{ijk}: response (dependent variable) for the *i-th* level of factor A, *j-th* level of factor B, and *k-th* block
- θ: constant ("intercept")
- α_i: Effect of the *i*-th level of factor A on *Y*
- β_j: Effect of the *j*-th level of factor B on *Y*
- $(\alpha\beta)_{ij}$: Interaction effect (effect of *i*-th level of A and *j*-th level of B on *Y*)
- b_k: Effect of the *k*-th level of block on *Y*, a random variable
- e_{ijk}: Error associated with experimental unit in block *k* that received level *i* of A and level *j* of B [**residual**]

Note: two random-effect terms (b_k and e_{ijk})

This is considered a **linear mixed effects model** ("Mixed Model" for short).

Definition: *Linear model with at least <u>two</u> random variables (including the residual error, e), plus fixed-effects parameters, and possibly an intercept constant.*

Note: with random effects, at least some observations are <u>correlated</u>!

Two factors (with blocking)

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + e_{ijk}, \quad b_k \sim N(0, \sigma_b^2), \; e_{ijk} \sim N(0, \sigma_e^2)$$

- Y_{ijk}: response (dependent variable) for the *i-th* level of factor A, *j-th* level of factor B, and *k-th* block
- θ: constant ("intercept")
- α_i: Effect of the *i*-th level of factor A on *Y*
- β_j: Effect of the *j*-th level of factor B on *Y*
- $(\alpha\beta)_{ij}$: Interaction effect (effect of *i*-th level of A and *j*-th level of B on *Y*)
- b_k: Effect of the *k*-th level of block on *Y*, a random variable
- e_{ijk}: Error associated with experimental unit in block *k* that received level *i* of A and level *j* of B [**residual**]

Options for graphs can be added.

<u>Note</u>:
A B A*B
Can be written as:
A|B

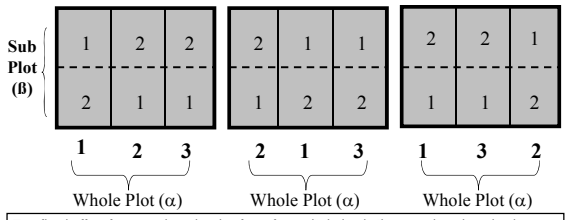
```
proc glimmix data=datafile;
CLASS A B block;
MODEL Y = A B A*B / s ;
RANDOM block;
LSMEANS A B A*B / diff;
run;
```

Note: incorporating random effects (experimental structure) is done separately from the incorporation of fixed effects in the model.

Splitting of experimental units: Clustering

Split Plot (with blocks)

Variation among blocks (σ_b^2), among "whole plots" (large experimental units) (σ_d^2), and among sub-plots within whole plots (σ_e^2)



Two fixed-effect factors, where levels of one factor ("whole plot") are assigned randomly to (large) experimental units, these units are split into sections, and levels of the second factor ("sub-plot") are randomly assigned to these (small) units.
Each block is a cluster and each whole plot within a block is a cluster.

Split Plot Design (with blocking)

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + d_{ik} + e_{ijk},$$

$$b_k \sim N(0, \sigma_b^2),\ d_{ik} \sim N(0, \sigma_d^2),\ e_{ijk} \sim N(0, \sigma_e^2)$$

Y_{ijk}: response (dependent variable) for the *i-th* level of whole-plot factor, *j-th* level of sub-plot factor, and *k-th* block

θ: constant ("intercept")

α_i: Effect of the *i*-th level of whole-plot factor on *Y*

β_j: Effect of the *j*-th level of sub-plot factor on *Y*

$(\alpha\beta)_{ij}$: Interaction effect (effect of *i*-th whole plot and *j*-th subplot on *Y*)

b_k: Effect of the *k*-th level of block on *Y* (random effect)

d_{ik}: Whole-plot error (effect of *ik*-th experimental unit on *Y* **[could be written as interaction effect of block *k* and whole-plot *i* on *Y*) (random effect)**

e_{ijk}: Sub-plot error associated with experimental unit in block *k* that received whole-plot *i* and sub-plot *j* [**residual**]

```
proc glimmix data=datafile;
CLASS A B block;
MODEL Y = A B A*B / s ;
RANDOM block A*block;
LSMEANS A B A*B / diff;
run;
```

FA 2.sas

Contemporary linear mixed model analysis – A likelihood-based framework

- Fit model with:
  - Restricted (residual) Maximum Likelihood (**REML**) or with maximum likelihood (**ML**)
    - REML is the contemporary standard (default in GLIMMIX and other procedures)
- There are no sums of squares or mean squares
- In general, REML is superior to traditional ANOVA (based on mean squares) for mixed-model analysis, especially for unbalanced data sets, missing values, and complex correlation structures
- *Iterative approach*, requires:
  - sophisticated computer algorithms
  - fast computer processing speed, and ample computer memory
- In SAS, one can conduct the analysis with MIXED, GLIMMIX, or HPMIXED (for a subset of possible models, but with immense numbers of factor levels)
  - Do *not* use GLM procedure (inappropriate with unbalanced data sets, missing values, correlated responses, and in general, when there are multiple levels of splitting/clustering)

Tests

$H_0: \mu_{1\bullet} = \mu_{2\bullet} = \ldots.$

$H_0: \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots.$

$H_0: \mu_{ij} + \mu_{\bullet\bullet} = \mu_{i\bullet} + \mu_{\bullet j}$, for all i and j

Hypotheses: Main effects for first and second factor

Hypothesis: Interaction

- Note: If there *is* an interaction, the effect of the *one* factor depends on the level of the *other* factor
  - **Equivalently, slices (for the simple effects) for *one* factor are not the same for each level of the *other* factor (when there is an interaction)**
- F statistics are used to test hypotheses of the fixed effects
  - Known as Type 3 (Type III) tests
- With unbalanced data, especially with correlated data, missing values, and empty cells, one should use the Kenward-Roger degrees of freedom adjustment (**model Y = A|B / ddfm = KR;**) to improve accuracy of the tests
- Random effects are characterized by variances (e.g., σ_b^2) or covariances (correlations). There are tests for the random effects (likelihood-ratio tests are best) – not covered here.
  - **When a variance estimate is 0, one must make a follow-up decision for model fitting and post-fit analysis (*discussed later for a 3-factor case*)**

Model Fitting with PROC GLIMMIX

- For distributions in the exponential family (including Poisson, binomial, gamma, and *normal*)
  - PROC MIXED can also be used for normal distributions, but GLIMMIX has some very nice options not available in MIXED
- The **MODEL** statement is used for specifying all fixed effects
  - Options for denominator degrees of freedom (very important with mixed models, especially for unbalanced data sets, missing values, and correlated observations). (**/ddfm=KR**). Other options also.
- One or *more* **RANDOM** statements are used for specifying random effects (e.g., a random block effect) and correlations of the observations
  - Only consider simple situations in this workshop. Here we focus on random effects that are a consequence of the experimental structure.
  - In general, one may need to spend considerable effort in deciding on the proper random effects. One must get the random-effects portion of the model right before considering the significance of the fixed effects or in interpreting means and SEs.
    - *See notes from previous (more general) Mixed Model Workshop*
- Use **LSMEANS** statement to look at estimated expected values (*least squares means*) and differences, in tables and graphs

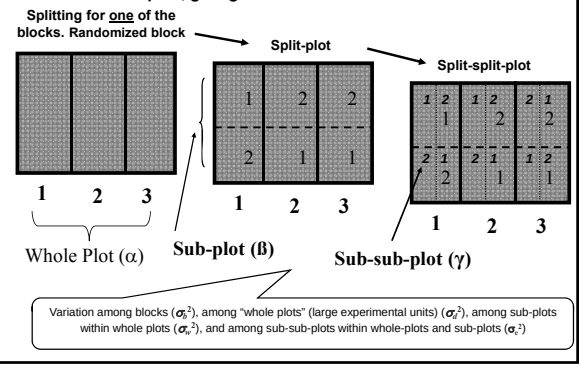
Model Adjustments

- One can evaluate model fits for residual normality, constant variance, independence (between subjects), and linearity (for continuous predictor variables)
  - Focus on possible nonconstant variability here
- The classical approach is to transform the observations when there is evidence (or theory) that variability is not constant
  - Y^*: log or square-root for counts (or others), when var. increases with *Y*
  - Y^*: Angular (arcsine square-root) for proportions (or others)
  - In general, Y^* becomes the new response variable
- Concept: New response variable may have constant variance
  - Should check residuals *after* fitting the model to Y^*
  - But, analysis is then on the scale of Y^*.
  - The null and alternative hypotheses are in terms of mean of Y^* (not the same meaning as the mean of *Y*)
  - This remains a generally useful approach with mixed models
- Alternatives available with contemporary mixed models:
  - Other statistical distributions, especially for discrete data: **Generalized linear mixed models (GLMMs)**

Factorial: Split-split plot (with blocks)

Clustering

Three (or more) fixed-effect factors, where there are multiple levels of splits, giving four levels of variation



Variation among blocks (σ_b^2), among "whole plots" (large experimental units) (σ_d^2), among sub-plots within whole plots (σ_w^2), and among sub-sub-plots within whole-plots and sub-plots (σ_e^2)

Split-Split Plot (with Blocks) Design

$$Y_{ijkl} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + b_l + d_{il} + w_{ijl} + e_{ijkl}$$

$$b_l \sim N(0,\sigma_b^2)\ ,\ d_{il} \sim N(0,\sigma_d^2)\ ,\ w_{ijl} \sim N(0,\sigma_w^2)\ e_{ijkl} \sim N(0,\sigma_e^2)$$

Y_{ijkl}: response (dependent variable) for the *i-th* level of whole-plot factor, *j-th* level of sub-plot factor, *k-th* level of the sub-sub-plot, and *l-th* block

θ: constant ("intercept")

α_i: Effect of the *i*-th level of whole-plot factor on *Y*

β_j: Effect of the *j*-th level of sub-plot factor on *Y*

γ_k: Effect of the *k*-th level of the sub-sub-plot on *Y*

$(\alpha\beta)$, $(\alpha\gamma)$, $(\beta\gamma)$, $(\alpha\beta\gamma)$: *Interaction* effects

b_l: Effect of the *l*-th level of block on *Y*

d_{il}: Whole-plot error (same as *block*whole*)

w_{ijl}: Sub-plot error [same as *subplot(block whole)*]

e_{ijkl}: Sub-sub-plot error [**residual**]

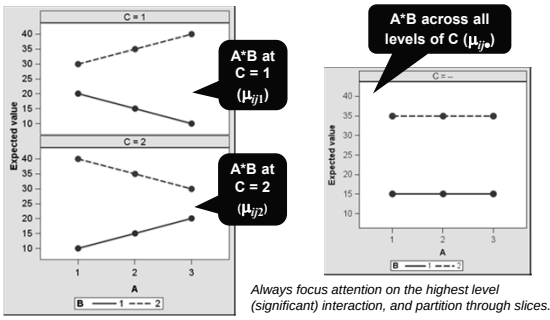
FA 3 split_split_n.sas

FA 3.sas

Three Factors: *Interpreting Interactions*

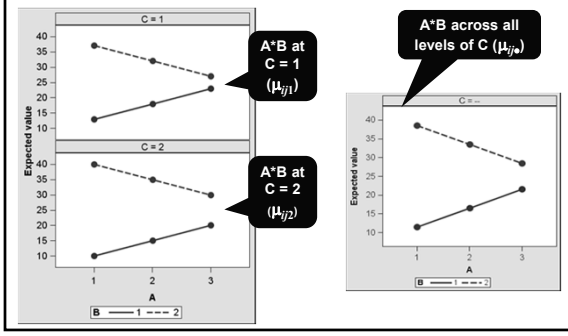
- Recall that with two factors, an interaction indicates that **simple effects** (say, differences of the means of A, at each level of B) varies with the level of the *other* factor
  - **Slices** can help quantify the specific interpretation of the interaction
  - With a true interaction, one *may* not be able to interpret main effects (need to look at graphs and contrasts of means [slices] to resolve what is going on)
- With three factors (say, A, B, and C), there are four possible interactions:
  - *Two-way*: A*B, A*C, B*C (or $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$)
  - *Three-way*: A*B*C (or $\alpha\beta\gamma$)
- **If the three-way interaction is real, one or more of the two-way interactions varies with (depends on) the level of the *other* factor**
  - *Example*: Suppose that there is a A*B interaction. If the 'form' of the interaction (say, direction of the mean differences for simple effects) varies with the level of C, then there is a (true, real) A*B*C interaction.
  - With a three-way interaction, the overall two-way interactions *may* not mean very much, in the same way that a main effect *may* not mean very much when there is a two-way interaction
- General rule: start with the highest-level interaction, and work your way down (through *slices*) in order to interpret the results

Example: **A*B*C three-way interaction**, with the A*B two-way interaction being depended on the level of C (the other factor). Here, the A*B interaction at C=1 and C=2 "cancel" each other, so that there is no manifested overall A*B interaction (but one main effect remains). In other cases, the A*B interactions at the different C levels may not cancel.



Always focus attention on the highest level (significant) interaction, and partition through slices.

Example: **no A*B*C three-way interaction**, but with an A*B two-way interaction. Here, the A*B interaction is the same at C=1 and C=2 (because of the lack of 3-way interaction); thus, the overall A*B 2-way interaction has the same shape as found for A*B at each level of C.



Contrasts

- The **LSMEANS** statement, with various options (**/ diff lines**) and graphs (**plots=mean(….**), can provide considerable insight about main effects and interactions
- However, more complex comparisons of means may often be of value
  - Example: Is the control mean different from the average of all the other treatments
    - $H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$ or $H_0 : \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3 = 0$
- Contrasts (linear combination of expected values) are used to test individual hypotheses of interests
  - Contrast example: $\Psi = 1\mu_1 - (1/3)\mu_2 - (1/3)\mu_3 - (1/3)\mu_4$
  - More generally: $\Psi = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$
    - Where (in the example): $c_1 = 1$, $c_2 = -1/3$, $c_3 = -1/3$, $c_4 = -1/3$
  - Fuller definition of contrast: *Linear combination of expected values, where the coefficients (c_i) sum to 0*
  - General hypothesis test: $H_0: \Psi = 0$
    - That is, $H_0: 1\mu_1 - (1/3)\mu_2 - (1/3)\mu_3 - (1/3)\mu_4 = 0$

Contrasts

- The **LSMEANS** statement (with the **/diff lines** options) can be used to easily give pairwise differences, which are contrasts (with $c_1 = 1$ and $c_2 = -1$; or more generally, $c_i = 1$ and $c_{i'} = -1$)
  - Example: $\mu_1 - \mu_2$ or $1\mu_1 - 1\mu_2$
- The **CONTRAST** and **ESTIMATE** statements have long been available in PROC GLM, MIXED, and now GLIMMIX
  - However, use of these statements can be quite tricky (with two or more factors), and one can easily get confused, because the statements are based on the effects (e.g., α) and not on the means
  - For example, to obtain $\mu_{11} - \mu_{21}$, one must write out (in procedure syntax):

    $|\alpha_1 + \beta_1 + (\alpha\beta)_{11}| - |\alpha_2 + \beta_1 + (\alpha\beta)_{21}|$ or: $\alpha_1 - \alpha_2 + (\alpha\beta)_{11} - (\alpha\beta)_{21}$
- Recently, the **LSMESTIMATE** statement has been added to make the testing of contrasts (much) easier
  - Contrasts are constructed in terms of the least-squares means, not in terms of the effects
  - When used with the new so-called *non-positional syntax*, testing of contrasts is even easier

Contrasts, alternative coding

- **LSMESTIMATE** statement
  - A "combination" of **LSMEANS** and **ESTIMATE**
  - One can essentially ignore the *effects*, and work with the means (which is what you want [most of the time])
  - One can use to get simple means or any contrast involving means for the factor level
- Think of the example, with factors A, B, and interaction A*B
  - One could obtain three separate tables of means with **LSMEANS A B A*B;**
  - LSMESTIMATE can operate on the contents of each one of these tables
    - There would be separate LSMESTIMATE statements for A, B, and A*B (if needed). There can be any number of statements.

```
lsmestimate A 'label' ... ;
lsmestimate B 'label' ... ;
lsmestimate A*B 'label' ... ;
```

One puts the factor *before* the label, then the coefficients (c_i). The red boxes refer to the position of the LSMEANS in the table.

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

Example: Potato early dying--(square-root of root weight) in relation to density of two pathogens

A (α): *Verticillium* density (**Vert**)
B (β): *Pratylenchus* density (**Prat**)

Obtain selected contrasts of the means

Prat.					
Vert.	1	2	3	4	Mean
1	μ_{11}	μ_{12}	μ_{13}	μ_{14}	$\mu_{1\bullet}$
2	μ_{21}	μ_{22}	μ_{23}	μ_{24}	$\mu_{2\bullet}$
3	μ_{31}	μ_{32}	μ_{33}	μ_{34}	$\mu_{3\bullet}$
Mean	$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\mu_{\bullet 3}$	$\mu_{\bullet 4}$	$\mu_{\bullet\bullet}$

FA 1_contrasts.sas

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

Example: Potato early dying--(square-root of root weight) in relation to density of two pathogens

A (α): *Verticillium* density (**Vert**)
B (β): *Pratylenchus* density (**Prat**)

```
lsmeans A B A*B;
```

Vert Least Squares Means

Vert	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1.6499	0.07042	146	23.43	<.0001
2	1.2463	0.06857	146	18.17	<.0001
3	1.0520	0.06718	146	15.66	<.0001

Prat Least Squares Means

Prat	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1.5213	0.08114	146	18.75	<.0001
2	1.2960	0.08086	146	16.03	<.0001
3	1.2370	0.07779	146	15.90	<.0001
4	1.2099	0.07763	146	15.59	<.0001

Vert*Prat Least Squares Means

Vert	Prat	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1	1.5568	0.1495	146	10.41	<.0001
1	2	1.7042	0.1325	146	12.86	<.0001
1	3	1.8471	0.1432	146	12.90	<.0001
1	4	1.4913	0.1375	146	10.84	<.0001
2	1	1.6578	0.1280	146	12.95	<.0001
2	2	1.1932	0.1495	146	7.98	<.0001
2	3	0.9609	0.1325	146	7.25	<.0001
2	4	1.1733	0.1375	146	8.53	<.0001
3	1	1.3494	0.1432	146	9.43	<.0001
3	2	0.9907	0.1375	146	7.20	<.0001
3	3	0.9029	0.1280	146	7.05	<.0001
3	4	0.9651	0.1280	146	7.54	<.0001

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

The means that are displayed <u>vertically</u> in LSMEANS tables are referenced <u>horizontally</u> by position in the LSMESTIMATE statement

```
lsmestimate Vert 'label' ... ;
```

$\mu_{1\bullet}$ $\mu_{2\bullet}$ $\mu_{3\bullet}$

Vert Least Squares Means

Vert	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1.6499	0.07042	146	23.43	<.0001
2	1.2463	0.06857	146	18.17	<.0001
3	1.0520	0.06718	146	15.66	<.0001

```
lsmestimate Prat 'label' ... ;
```

$\mu_{\bullet 1}$ $\mu_{\bullet 2}$ $\mu_{\bullet 3}$ $\mu_{\bullet 4}$

Prat Least Squares Means

Prat	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1.5213	0.08114	146	18.75	<.0001
2	1.2960	0.08086	146	16.03	<.0001
3	1.2370	0.07779	146	15.90	<.0001
4	1.2099	0.07763	146	15.59	<.0001

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

lsmestimate A 'label' ■■■ ;

lsmestimate B 'label' ■■■■ ;

Use "constants" (c_i) to obtain means or combinations of means. *Trailing zeros can be omitted.* Leading zeros must be used. The second mean of Prat would be identified with 0 1 0 0; or 0 1; but the first Prat mean could be identified with just 1; or with 1 0 0 0;

```
lsmestimate Vert 'mu_2.' 0 1;
lsmestimate Prat 'mu_.2' 0 1 0 0;
```

$\mu_{\bullet 2}$ or $0\cdot\mu_{\bullet 1}+1\cdot\mu_{\bullet 2}+0\cdot\mu_{\bullet 3}+0\cdot\mu_{\bullet 4}$

```
lsmestimate Vert 'mu_2.-mu_1.' -1 1;
lsmestimate Prat 'mu_.4-mu_.2' 0 -1 0 1;
```

A difference of two means is specified with 1 and -1, based on position.

$\mu_{\bullet 4}-\mu_{\bullet 2}$ or $0\cdot\mu_{\bullet 1}-1\cdot\mu_{\bullet 2}+0\cdot\mu_{\bullet 3}+1\cdot\mu_{\bullet 4}$

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

```
lsmestimate Vert 'mu_2.' 0 1;
lsmestimate Prat 'mu_.2' 0 1;
```

```
lsmestimate Vert 'mu_2.-mu_1.' -1 1;
lsmestimate Prat 'mu_.4-mu_.2'  0 -1 0 1;
```

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert	mu_2.	1.2463	0.06857	146	18.17	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.2	1.2960	0.08086	146	16.03	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert	mu_2.-mu_1.	-0.4036	0.09829	146	-4.11	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.4-mu_.2	-0.08613	0.1121	146	-0.77	0.4435

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

lsmestimate A 'label' ■■■ ;

lsmestimate B 'label' ■■■■ ;

The real value of lsmestimate is in obtaining more complex contrasts than just differences. For example, with Vert, the difference between the control (level 1) and the rest would be obtained using 1 -.5 -.5 for the coefficients. With Prat, the difference between the control and the rest would be specified with 1 -0.3333 -0.3333 -0.3333;

$$\mu_{1\bullet}-(\mu_{2\bullet}+\mu_{3\bullet})/2$$
$$\mu_{\bullet 1}-(\mu_{\bullet 2}+\mu_{\bullet 3}+\mu_{\bullet 4})/3$$
$$(\mu_{11}+\mu_{12})/2-(\mu_{13}+\mu_{14})/2$$

```
lsmestimate Vert 'mu_1.vs rest' 1 -.5 -.5;
lsmestimate Prat 'mu_.1 vs rest' 1 -.33333 -.33333 -.33333;

/* Or, convert to integers and rescale with divisor */
lsmestimate Prat 'mu_.1 vs rest' 3 -1 -1 -1 / divisor=3;
lsmestimate Prat 'mu_.1&2 vs mu_.3&4' 1 1 -1 -1 / divisor=2;

/*with the same effect, one can put several together (to get one table)*/
lsmestimate Prat
        'mu_.1 vs rest' 3 -1 -1 -1 ,
        'mu_.1&2 vs mu_.3&4' 1 1 -1 -1 / divisor=3,2;
```

```
lsmestimate Vert 'mu_1.vs rest' 1 -.5 -.5;
lsmestimate Prat 'mu_.1 vs rest' 1 -.33333 -.33333 -.33333;
/* Or, convert to integers and rescale with divisor */
lsmestimate Prat 'mu_.1 vs rest' 3 -1 -1 -1 / divisor=3;
lsmestimate Prat 'mu_.1&2 vs mu_.3&4' 1 1 -1 -1 / divisor=2;
/*with the same effect, one can put several together (to get one table)*/
lsmestimate Prat
        'mu_.1 vs rest' 3 -1 -1 -1 ,
        'mu_.1&2 vs mu_.3&4' 1 1 -1 -1 / divisor=3,2;
```

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert	mu_1.vs rest	0.5007	0.08522	146	5.88	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.1 vs rest	0.2737	0.09301	146	2.94	0.0038

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.1 vs rest	0.2737	0.09301	146	2.94	0.0038

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.1&2 vs mu_.3&4	0.1852	0.07937	146	2.33	0.0210

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Prat	mu_.1 vs rest	0.2737	0.09301	146	2.94	0.0038
Prat	mu_.1&2 vs mu_.3&4	0.1852	0.07937	146	2.33	0.0210

Prat.					
Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

With a significant interaction, focus should really be on contrasts of the *interaction means* (μ_{ij}).

Must keep track of the order of the interaction means (look at table from LSMEANS A*B;)

lsmestimate A*B 'label' ■■■■■■■■■■■■ ;

μ11	μ12	μ13	μ14	μ21	μ22	μ23	μ24	μ31	μ32	μ33	μ34

$$\mu_{24}$$
$$(\mu_{21}+\mu_{31})/2-\mu_{11}$$
$$(\mu_{12}+\mu_{13}+\mu_{14})/3-\mu_{11}$$

```
lsmestimate Vert*Prat 'mu24' 0 0 0 0 0 0 0 1 0 0 0 0;
lsmestimate Vert*Prat '(mu21+mu31)/2 - mu11'
        -2 0 0 0  1 0 0 0  1 / divisor=2;
lsmestimate Vert*Prat '(mu12+mu13+mu14)/3 - mu11'
        -3 1 1 1 0 0 0 0 0 0 0 0/ divisor=3;
lsmestimate Vert*Prat
        '(mu21+mu31)/2 - mu11' -2 0 0 0  1 0 0 0  1,
        '(mu12+mu13+mu14)/3 - mu11'
        -3 1 1 1 0 0 0 0 0 0 0 0/ divisor=2,3;
```

Vert*Prat Least Squares Means

Vert	Prat	Estimate	Standard Error
1	1	1.5568	0.1495
1	2	1.7042	0.1325
1	3	1.8471	0.1432
1	4	1.4913	0.1375
2	1	1.6578	0.1280
2	2	1.1932	0.1495
2	3	0.9609	0.1325
2	4	1.1733	0.1375
3	1	1.3494	0.1432
3	2	0.9907	0.1375
3	3	0.9029	0.1280
3	4	0.9651	0.1280

```
lsmestimate Vert*Prat 'mu24' 0 0 0 0 0 0 0 1 0 0 0 0;
lsmestimate Vert*Prat '(mu21+mu31)/2 - mu11'
        -2 0 0 0  1 0 0 0  1 / divisor=2;
lsmestimate Vert*Prat '(mu12+mu13+mu14)/3 - mu11'
        -3 1 1 1  0 0 0 0  0 0 0 0/ divisor=3;
lsmestimate Vert*Prat
        '(mu21+mu31)/2 - mu11' -2 0 0 0  1 0 0 0  1,
        '(mu12+mu13+mu14)/3 - mu11'
        -3 1 1 1  0 0 0 0  0 0 0 0/ divisor=2,3;
```

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	mu24	1.1733	0.1375	146	8.53	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	(mu21+mu31)/2 - mu11	-0.05320	0.1777	146	-0.30	0.7651

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	(mu12+mu13+mu14)/3 - mu11	0.1241	0.1694	146	0.73	0.4650

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	(mu21+mu31)/2 - mu11	-0.05320	0.1777	146	-0.30	0.7651
Vert*Prat	(mu12+mu13+mu14)/3 - mu11	0.1241	0.1694	146	0.73	0.4650

Prat.

Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

There is now an alternative to the so-called positional syntax for LSMESTIMATE and ESTIMATE. With nonpositional syntax, one simply identifies the coefficients for the means that matter, using square brackets, [ ] (order does not matter).

```
lsmestimate A 'label' [# , i ] [# , i ] …;
lsmestimate B 'label' [# , j ] [# , j ] … ;
lsmestimate A*B  'label' [# , i j ] [# , i j ] … ;
```

Here, # is the coefficient (e.g., c_i), and i and j refer to the level of the factor in the main effect or interaction.
The advantage is for complicated contrasts.

$\mu_{3\bullet} - \mu_{1\bullet}$

μ_{24}

$(\mu_{21} + \mu_{31})/2 - \mu_{11}$

$(\mu_{12} + \mu_{13} + \mu_{14})/3 - \mu_{11}$

```
lsmestimate Vert*Prat  'mu24' [1,2 4];
```

$1 \cdot \mu_{24}$

```
lsmestimate Vert  '3-1' [1 , 3 ]  [-1 , 1 ]  …;
```

$1 \cdot \mu_{3\bullet} - 1 \cdot \mu_{1\bullet}$

Prat.

Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

$\mu_{3\bullet} - \mu_{1\bullet}$

μ_{24}

$(\mu_{21} + \mu_{31})/2 - \mu_{11}$

$(\mu_{12} + \mu_{13} + \mu_{14})/3 - \mu_{11}$

Positional (first) and nonpositional (second) syntax for contrasts. Results are identical (as required).

```
lsmestimate Vert '3-1' -1 0 1;
lsmestimate Vert*Prat 'mu24' 0 0 0 0 0 0 0 1 0 0 0 0;
lsmestimate Vert*Prat '(mu21+mu31)/2 - mu11' -2 0 0 0  1 0 0 0  1    /divisor=2;
lsmestimate Vert*Prat '(mu12+mu13+mu14)/3-mu11' -3 1 1 1 0 0 0 0 0 0 0 0
                                                                     /divisor=3;

lsmestimate Vert '3-1 n' [1, 3] [-1,1];
lsmestimate Vert*Prat 'mu24 n' [1,2 4];
lsmestimate Vert*Prat '(mu21+mu31)/2-mu11 n' [-2,1 1] [1,2 1] [1,3 1] /divisor=2;
lsmestimate Vert*Prat '(mu12+mu13+mu14)/3-mu11 n' [-3,1 1] [1,1 2] [1,1 3] [1,1 4]
                                                                     /divisor=3;
```

```
lsmestimate Vert '3-1 n' [1, 3] [-1,1];
lsmestimate Vert*Prat 'mu24 n' [1,2 4];
lsmestimate Vert*Prat '(mu21+mu31)/2-mu11 n' [-2,1 1] [1,2 1] [1,3 1] /divisor=2;
lsmestimate Vert*Prat '(mu12+mu13+mu14)/3-mu11 n' [-3,1 1] [1,1 2] [1,1 3] [1,1 4]
                                                                     /divisor=3;
```

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert	3-1 n	-0.5978	0.09732	146	-6.14	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	mu24 n	1.1733	0.1375	146	8.53	<.0001

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	(mu21+mu31)/2 - mu11 n	-0.05320	0.1777	146	-0.30	0.7651

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Vert*Prat	(mu12+mu13+mu14)/3 - mu11 n	0.1241	0.1694	146	0.73	0.4650

Prat.

Vert.	1	2	3	4	Mean
1	1.56	1.70	1.85	1.49	1.65
2	1.66	1.19	0.96	1.17	1.25
3	1.35	0.99	0.90	0.97	1.05
Mean	1.52	1.30	1.24	1.21	1.32

$H_0 : (\mu_{32} - \mu_{12}) = (\mu_{31} - \mu_{11})$

$H_a : (\mu_{32} - \mu_{12}) \neq (\mu_{31} - \mu_{11})$

- Use the **lsmestimate** statement, with **positional** *and* **nonpositional** syntax, and estimate:

μ_{32} and

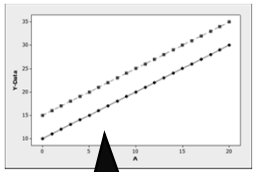
$(\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11})$ or

$\mu_{32} - \mu_{12} - \mu_{31} + \mu_{11}$ or

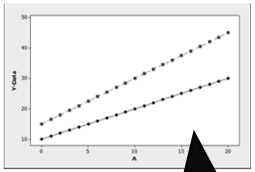
$1 \cdot \mu_{32} - 1 \cdot \mu_{12} - 1 \cdot \mu_{31} + 1 \cdot \mu_{11}$

Known as a tetrad contrast (a component of the interaction)

"Analysis of Covariance": Combining continuous variables and factors in the same model



No interaction (main effects of A and B): Same slope for all levels of B



Interaction (plus main effects of A and B): *Different* slopes for the different levels of B

Continuous B variable (X) in a Split Plot Design (with blocking)—example

$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + d_{ik} + e_{ijk}$

$Y_{ijk} = \theta + \alpha_i + \beta X_j + \delta_i X_j + b_k + d_{ik} + e_{ijk}$

$Y_{ijk} = (\theta + \alpha_i) + (\beta + \delta_i) X_j + b_k + d_{ik} + e_{ijk}$

$b_k \sim N(0, \sigma_b^2)$, $d_{ik} \sim N(0, \sigma_d^2)$, $e_{ijk} \sim N(0, \sigma_e^2)$

The intercept of the *Y*:*X* line is $\theta + \alpha_i$ where α_i is the effect of the *i*-th level of A on the intercept (height of the line)

The slope of the *Y*:*X* line is $\beta + \delta_i$ where δ_i is the effect of the *i*-th level of A on the slope

FA 4.sas

```
proc glimmix data=datafile;
/* B is _not_ in CLASS */
CLASS A  block;
MODEL Y = A B A*B / s ;
RANDOM block A*block;
run;
```

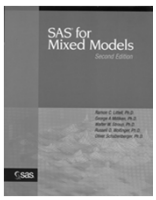
Conclusions

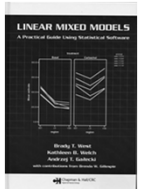
- There has been incredible changes in statistics over the past two decades, where likelihood-based mixed-model analyses have largely supplanted traditional ANOVA-type ('sum of squares') analyses, the latter not fully or properly accounting for the random effects
  - Stimulated by advances in statistical theory and computational algorithms, coupled with the drastic increases in speed and memory of modern computers
  - In general, true mixed-model analysis is superior to other approaches
- Unfortunately, researchers outside of statistics have been slow to adapt to the changes in statistics, or are confused by the advances in statistical methodology
  - This is partly because the advances are not even mentioned in the first two or three courses of statistics, even though real-world data analysis depends heavily on the use of mixed models
  - Hence the need for workshops, and new textbooks (which are coming)
- Contemporary mixed-model data analysis requires careful consideration of both the fixed and random effects in the model (often equivalent to consideration of the treatment structure and the experimental structure)
  - This workshop has focused on the fixed effects in mixed models, dealing explicitly with two or more fixed-effect factors (i.e., crossed factorials)

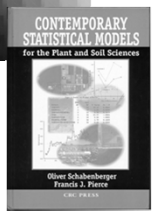
Conclusions, *continued*

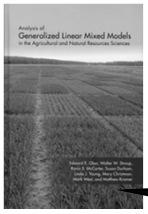
- For mixed-model analysis of factorials, one must make sure the proper random-effect terms are in the model
  - For repeated measures and for situations with unequal variances, this is a major step (not considered in this workshop)
  - We have emphasized random effects that are a consequence of the experimental structure
    - Even here, there is always the question regarding variance estimates equal to 0
  - With GLIMMIX or MIXED in SAS, one uses RANDOM statements for the random effects
- For the fixed effects (specified with the MODEL statement):
  - One should always make sure that the interactions are included in the model
  - Interpretation of the results (such as the Type III tests) should always start with the highest-level (significant) interaction (e.g., A*B*C), then lower-level interactions (e.g., A*B), and then the main effects
  - Use slices of simple effects and graphs/tables to explore the nature of the interactions (options on LSMEANS statement and the SLICE statement)
- Mixed-model analysis can accommodate continuous explanatory variables as well as factors, and preserve the experimental structure with the random effects

Some good references









2012. Published by American Society of Agronomy.

SAS GLOBAL FORUM Paper 351-2011
CONTRAST and ESTIMATE Statements Made Easy: The LSMESTIMATE Statement .
Kathleen Kiernan, Randy Tobias, Phil Gibbs, and Jill Tao.
http://support.sas.com/resources/papers/proceedings11/351-2011.pdf