Purdue-NCKU program

Lecture 5
Single/Multiple-Factor Analysis

Dr. Qifan Song

Revisit Previous Lecture

- One factor analysis

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2 \ldots a \\ j = 1, 2, \ldots n_i \end{cases}$$

- F-test.

- Contrast test and multiple inferences

Pairwise Comparison

- A special case of contrast $\Gamma = \mu_i - \mu_j$
- All previous method (t test with Bonferroni or Scheffe) still work
- Unless one has specified which pairs to compare before analyzing the data, we want to make all possible pairwise comparisons ($m = (a(a-1)/2)$ pairs)

  We want to compare the largest group mean vs smallest group mean. Does it count as $m = 1$?
- Bonferroni or Scheffe may not be good choice, since both are conservative for this scenario.

Tukey's method

- Consider all possible pairwise comparisons: $\Gamma = \mu_i - \mu_j$. Estimate: $C = \bar{y}_{i.} - \bar{y}_{j.}$, St. Error: $\text{S.E.}_C = \sqrt{\text{MS}_E(1/n_i + 1/n_j)}$
- Need to find a critical value: Q
- Simultaneous Test: $|C/\text{S.E.}_C|$ vs Q
- This also implies that simultaneous CI: $C \pm Q$ S.E.$_C$
- To control the overall error rate, we need to study the distribution of

$$\max_{\text{all possible pairs}} |C/\text{S.E.}_C|$$

  under null (i.e., all means are the same), which can be obtained by a simulation. And Q will be the $1-\alpha$ quantile of this distribution

- When $n_i \equiv n$, this distribution is related to studentized range distribution (a groups, d.f. $= n - a$).

How to interpret multiple inferences

It is possible that there are contradictions between inferences. How do we explain it?

Rejection $\Leftrightarrow$ Scientific Discovery

α overall rate $\Leftrightarrow$ with high chance ,all claimed scientific discoveries are true

Dunnett's method

- Consider all possible pairwise comparisons against one specific control level (index c): $\Gamma = \mu_i - \mu_c$ for all $i \neq c$ Estimate: $C = \bar{y}_{i.} - \bar{y}_{c.}$, St. Error: S.E.$_C = \sqrt{\text{MS}_E(1/n_i + 1/n_c)}$
- The critical value: Q is the $1-\alpha$ quantile of this distribution of

$$\max_{i \neq c} |C/\text{S.E.}_C|$$

  under null (all means are same).
- Simultaneous Test and C.I.: $|C/\text{S.E.}_C|$ vs Q; $C \pm Q$ S.E.$_C$
- When $n_i \equiv n$, this table of Q is given by Dunnett (1964).

Hsu's MCB

Multiple Comparisons with the Best *sample mean* level: to find a subset of index $\{1, \ldots, a\}$ such that with $1 - \alpha$, it contains the index of the best treatment.

- Perform a hypotheses: H_0: The ith treatment is the best.
- MCB = index of fail-to-reject hypotheses
- We perform a one-side Dunnett's method
  - The one side critical value: Q is the $1 - \alpha$ quantile of this distribution of

    $$\max_{i \neq c} C/\text{S.E.}_C$$

    under null (all means are same).
- Since there is only one true null hypothesis, there is no need for multiplicity control,i.e., all a Dunnett's comparisons are conducted under α not α/a.

randomized complete block design (RCBD)

Penicillin Experiment
In this experiment, four penicillin manufacturing processes (A, B, C and D) were being investigated. Yield was the response. It was known that an important raw material, corn steep liquor, was quite variable. The experiment and its results were given below:

	blend 1	blend 2	blend 3	blend 4	blend 5
A	89_1	84_4	81_2	87_1	79_3
B	88_3	77_2	87_1	92_3	81_4
C	97_2	92_3	87_4	89_2	80_1
D	94_4	79_1	85_3	84_4	88_2

- Blend is a nuisance factor, treated as a block factor;
- (Complete) Blocking: all the treatments are applied within each block, and they are compared within blocks.
- Advantage: Eliminate blend-to-blend (between-block) variation from experimental error variance when comparing treatments.
- Cost: degree of freedom.

RCBD

- b blocks each consisting of (partitioned into) a experimental units
- a treatments are randomly assigned to the experimental units within each block
- Typically after the runs in one block have been conducted, then move to another block. (Difference with completely randomized design)
- Typical blocking factors: day, batch of raw material etc.
- Results in restriction on randomization because randomization is only within blocks.

Statistical Model

- b blocks and a treatments
- Statistical model is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases}$$

  μ - grand mean

  τ_i - ith treatment effect

  β_j - jth block effect

  $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$

- Similarly to one factor analysis $\sum_{i=1}^{a} \tau_i = 0$;
- Random Block Effect (randomly selected blocks, farming fields): β_j are random variables following $N(0, \sigma_\beta^2)$
- Fixed Block Effect (fixed blocks, e.g. sex): β_j are unknown constants and $\sum_{j=1}^{b} \beta_j = 0$.

Sum of Squares (SS)

- $y_{ij} - \overline{y}_{..} = (\overline{y}_{i.} - \overline{y}_{..}) + (\overline{y}_{.j} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})$
- Can partition $\mathsf{SS}_\mathsf{T} = \sum\sum (y_{ij} - \overline{y}_{..})^2$ into

$$b\sum (\overline{y}_{i.} - \overline{y}_{..})^2 + a\sum (\overline{y}_{.j} - \overline{y}_{..})^2 + \sum\sum (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2$$

SS_Trt	$= b\sum (\overline{y}_{i.} - \overline{y}_{..})^2$	$= b\sum (\overline{\epsilon}_{i.} + \tau_i - \overline{\epsilon}_{..})^2$	df$= a-1$
SS_Blk	$= a\sum (\overline{y}_{.j} - \overline{y}_{..})^2$	$= a\sum (\overline{\epsilon}_{.j} + \beta_j - \overline{\epsilon}_{..} - \overline{\beta}_{.})^2$	df $= b-1$
SS_E	$= \sum\sum (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2$	$= \sum\sum (\epsilon_{ij} - \overline{\epsilon}_{i.} - \overline{\epsilon}_{.j} + \overline{\epsilon}_{..})^2$	df $= (a-1)$ $\times(b-1)$

How do we determine the d.f.?

Hence:

- $\mathsf{SS}_\mathsf{T} = \mathsf{SS}_\mathsf{Treatment} + \mathsf{SS}_\mathsf{Block} + \mathsf{SS}_\mathsf{E}$
- The Mean Squares are
$\mathsf{MS}_\mathsf{Treatment} = \mathsf{SS}_\mathsf{Treatment}/(a-1)$, $\mathsf{MS}_\mathsf{Block} = \mathsf{SS}_\mathsf{Block}/(b-1)$, and $\mathsf{MS}_\mathsf{E} = \mathsf{SS}_\mathsf{E}/(a-1)(b-1)$.

subheading Testing Basic Hypotheses

- $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$ vs H_1 : at least one is not
- Can show:

  $\text{E}(\text{MS}_\text{E}) = \sigma^2$

  $\text{E}(\text{MS}_\text{Treatment}) = \sigma^2 + b\sum_{i=1}^{a} \tau_i^2/(a-1)$

  $\text{E}(\text{MS}_\text{Block}) = \sigma^2 + a\sum_{j=1}^{b} \beta_j^2/(b-1)$ or $\sigma^2 + a\sigma_\beta^2$

- Use F-test to test H_0:

$$F = \frac{\text{MS}_\text{Treatment}}{\text{MS}_\text{E}} = \frac{\text{SS}_\text{Treatment}/(a-1)}{\text{SS}_\text{E}/((a-1)(b-1))}$$

- Under null, $F \sim F_{a-1,(a-1)(b-1)}$
- May perform test for block effects
  - Usually not of interest.
  - Randomization is restricted
  - Block effect may be confounded with other factor due to restricted randomization

Block vs non-Block design

Estimating $\tau_i - \tau_j$ under random block effect:

- Block Design: $\overline{y}_{i.} - \overline{y}_{j.}$ has variance $2\sigma^2/b$
- Non-block Design: $y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(\sigma^2 + \sigma^2_\beta)$. $\overline{y}_{i.} - \overline{y}_{j.}$ has variance $2(\sigma^2 + \sigma^2_\beta)/b$
- Non-block Design has higher d.f. for SSE.
- Reducing SSE vs. Reducing degree of freedom;
- R.E. = Relative Efficiency of RCBD to CRD

$$\text{R.E.} = \frac{MSE_{CRD}}{MSE_{RCBD}} \approx \frac{SS_{Block} + b(a-1)MSE}{(ab-1)MSE}$$

Pairwise Treatments Comparison

- $\Gamma = \tau_i - \tau_j$. Estimate: $C = \bar{y}_{i.} - \bar{y}_{j.} = \bar{\epsilon}_{i.} - \bar{\epsilon}_{j.}$, and St. Error: $\text{S.E.}_C = \sqrt{\text{MS}_E(1/n_i + 1/n_j)}$
- Need to find a critical value: Q
- Simultaneous Test: $|C/\text{S.E.}_C|$ vs Q
- This also implies that simultaneous CI: $C \pm Q$ S.E.$_C$
- Under null hypothesis,

$$\max_{\text{all possible pairs}} |C/\text{S.E.}_C|$$

  has a tractable distribution (studentized range distribution with a groups and $(a-1)(b-1)$ d.f.). Q will be the $1-\alpha$ quantile of this distribution
- Other inference task, such as testing contrasts with Bonferroni adjustment can be defined as in one factor analysis.

Factorial analysis

Bottling Experiment

A soft drink bottler is interested in obtaining more uniform fill heights in the bottles produced by his manufacturing process. An experiment is conducted to study three factors of the process, which are

the percent carbonation (A):	10, 12, 14 percent
the operating pressure (B):	25, 30 psi
the line speed (C):	200, 250 bpm

The response is the deviation from the target fill height. Each combination of the three factors has two replicates and all 24 runs are performed in a random order. The experiment and data are shown below.

	pressure(B)			
	25 psi		30 psi	
	LineSpeed(C)		LineSpeed(C)	
Carbonation(A)	200	250	200	250
10	-3,-1	-1,0	-1,0	1, 1
12	0, 1	2,1	2,3	6,5
14	5,4	7,6	7,9	10,11

Factorial Design

- Structure
  - a number of factors: F_1, F_2, ..., F_r.
  - each with a number of levels: l_1, l_2, ..., l_r
  - number of all possible level combinations (treatments): $l_1 \times l_2 \ldots \times l_r$
  - interested in (main) effects, 2-factor interactions (2fi), 3-factor interactions (3fi), etc.
- Require complete randomization of the order of experiments
  - One needs to repeatly change experimental conditions

Statistical Model (Two Factors: A and B)

- Statistical model is

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

  μ - grand mean

  τ_i - ith level effect of factor A (ignores B) (main effects of A)

  β_j - jth level effect of factor B (ignores A) (main effects of B)

  $(\tau\beta)_{ij}$ - interaction effect of combination ij (Explain variation not explained by main effects)

  $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$

- Over-parameterized model: must include certain parameter constraints. Typically

$$\textstyle\sum_i \tau_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_i (\tau\beta)_{ij} = 0 \quad \sum_j (\tau\beta)_{ij} = 0$$

Estimates

- Rewrite observation as:

$$y_{ijk} = \overline{y}_{...} + (\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}) + (y_{ijk} - \overline{y}_{ij.})$$

- result in estimates

$$\widehat{\mu} = \overline{y}_{...}$$

$$\widehat{\tau}_i = \overline{y}_{i..} - \overline{y}_{...}$$

$$\widehat{\beta}_j = \overline{y}_{.j.} - \overline{y}_{...}$$

$$\widehat{(\tau\beta)}_{ij} = \overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}$$

- predicted value at level combination ij is

$$\widehat{y}_{ijk} = \overline{y}_{ij.}$$

- Residuals are

$$\widehat{\epsilon}_{ijk} = y_{ijk} - \overline{y}_{ij.}$$

Partitioning the Sum of Squares

- Based on

$$y_{ijk} = \overline{y}_{...} + (\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}) + (y_{ijk} - \overline{y}_{ij.})$$

- Calculate $SS_T = \sum (y_{ijk} - \overline{y}_{...})^2$

- Right hand side simplifies to

$$\begin{array}{rcl} SS_A: & bn\sum_i (\overline{y}_{i..} - \overline{y}_{...})^2 + & df = a-1 \\ SS_B: & an\sum_j (\overline{y}_{.j.} - \overline{y}_{...})^2 + & df = b-1 \\ SS_{AB}: & n\sum_i\sum_j (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 + & df = (a-1)(b-1) \\ SS_E: & \sum_i\sum_j\sum_k (y_{ijk} - \overline{y}_{ij.})^2 & df = ab(n-1) \end{array}$$

- SS_T=SS_A + SS_B + SS_{AB} + SS_E
- Using SS/df leads to MS_A,MS_B, MS_{AB} and MS_E.

Testing Hypotheses

1 Interaction effects of AB:

$$H_0 : (\tau\beta)_{ij} = 0 \text{ for all } i, j \text{ vs } H_1 : \text{at least one } (\tau\beta)_{ij} \neq 0.$$

2 Main effects of A: H_0 : All $\tau_i = 0$ vs H_1 : at least one $\tau_i \neq 0$.

3 Main effects of B: H_0 : All $\beta_j = 0$ vs H_1 : at least one $\beta_j \neq 0$.

- $\mathrm{E(MS_E)}=\sigma^2$

  $\mathrm{E(MS_A)} = \sigma^2 + bn\sum \tau_i^2/(a-1)$

  $\mathrm{E(MS_B)} = \sigma^2 + an\sum \beta_j^2/(b-1)$

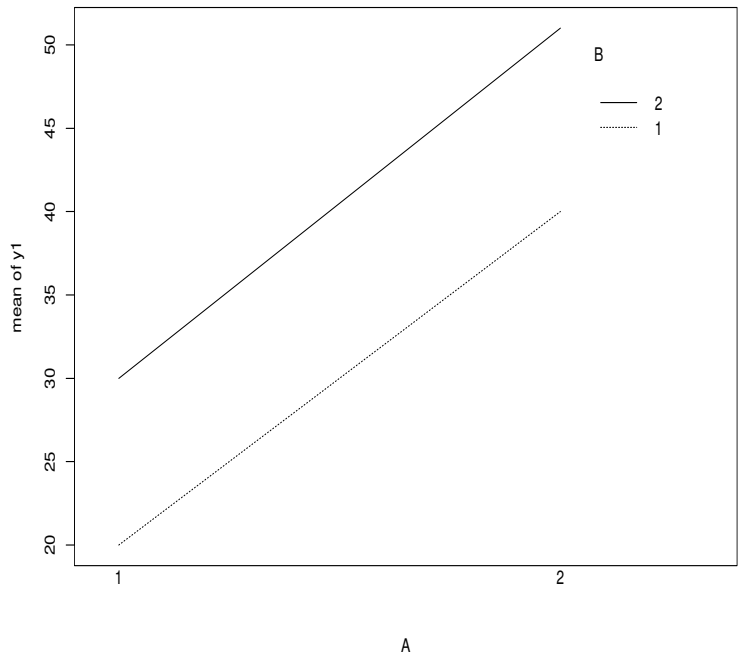
  $\mathrm{E(MS_{AB})} = \sigma^2 + n\sum (\tau\beta)_{ij}^2/(a-1)(b-1)$

- Use F-statistics for testing the hypotheses above:

  $$1:\ F = \frac{\mathrm{MS_{AB}}}{\mathrm{MS_E}} \quad 2:\ F = \frac{\mathrm{MS_A}}{\mathrm{MS_E}} \quad 3:\ F = \frac{\mathrm{MS_B}}{\mathrm{MS_E}} \sim F_{df_1, df_2}$$

  for respective d.f.s under null hypotheses.

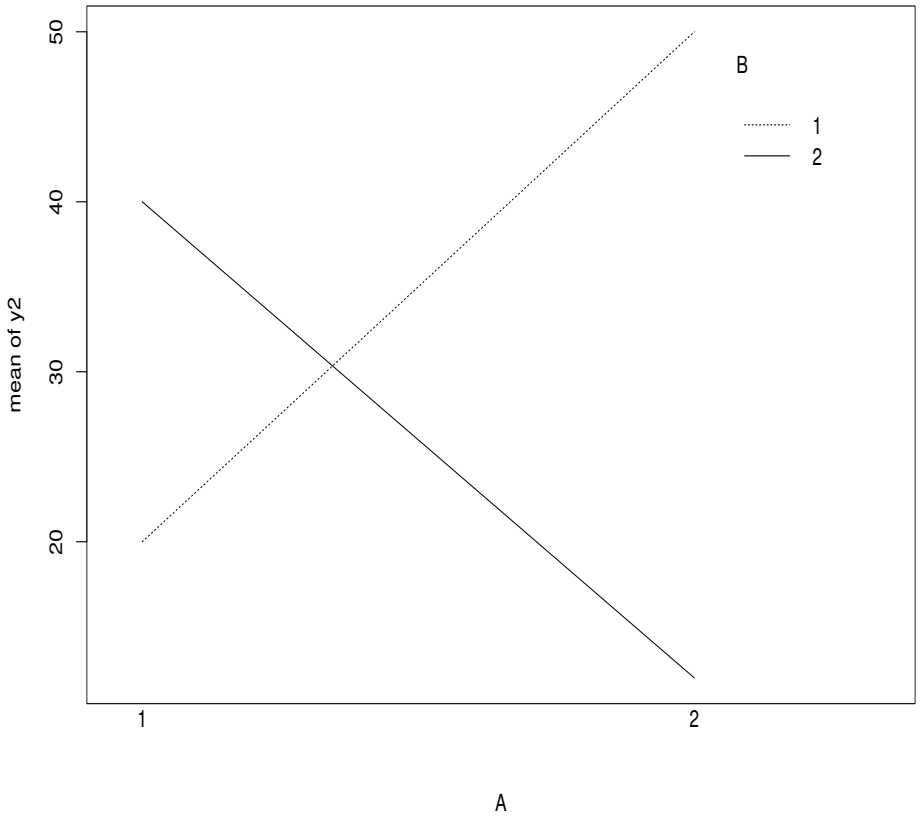
Interaction Effect (No interaction)



Difference between level means of B (with A fixed at a level) does not depend on the level of A; demonstrated by two parallel lines.

Optimal level of A plus optimal level of B leads to optimal combination.

Interaction Effect



Difference between level means of B (with A fixed at a level) depends on the level of A. Making any argument over main effects will be misleading.

Multiple comparison when factors don't interact

When factors don't interact, i.e., the F test for interaction is not significant in the ANOVA, factor level means can be compared to draw conclusions regarding their effects on response, if the main effect F test is significant.

Pairwise comparison for factor A:

- Consider all possible pairwise comparisons: $\Gamma = \mu + \tau_i - \mu - \tau_j$. Estimate: $C = \bar{y}_{i..} - \bar{y}_{j..}$, St. Error: S.E.$_C = \sqrt{\text{MS}_E(1/bn + 1/bn)}$
- Need to find a critical value: Q
- Simultaneous Test: $|C/\text{S.E.}_C|$ vs Q
- This also implies that simultaneous CI: $C \pm Q$ S.E.$_C$
- Q will be the $1-\alpha$ quantile of a distribution that is related to studentized range distribution (a groups, $ab(n-1)$ d.f.).
- Bonferroni/Scheffe are alternative methods

Similar approach for the comparison of factor B.

Multiple comparisons when factors interact

When factors interact, multiple comparison is usually directly applied to treatment means

$$\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} \text{ vs } \mu_{i'j'} = \mu + \tau_{i'} + \beta_{j'} + (\tau\beta)_{i'j'}$$

This is equivalent to viewing a two-factor problem as a one-factor (with ab levels) problem

- $\widehat{\mu}_{ij} = \bar{y}_{ij.}$ and $\widehat{\mu}_{i'j'} = \bar{y}_{i'j'.}$
- Var$(\bar{y}_{ij.} - \bar{y}_{i'j'.})$=$\frac{2\sigma^2}{n}$ and standard error is $\sqrt{2\text{MSE}/n}$
- there are ab treatment means and $m = \frac{ab(ab-1)}{2}$ pairs.
- Tukey's/Bonferroni's/Scheffe's method.

Combine the interaction plot with field knowledge to better understand the underlying mechanism.

Test simple effects within an interaction

- Test for the effect of A within each level of B (as A*B is significant)
- This yields b F test, testing: given a fixed j, does $\mu_{1j} = \mu_{2j} = \ldots = \mu_{aj}$?
- Each F test is an ANOVA task for a column of data, but we use the MSE of the whole data rather than the MSE obtained from data column. Thus the d.f. of the F test is $(a-1,\ ab(n-1))$, not $(a-1,\ a(n-1))$.
- Reason: larger degree of freedom means a better σ^2 estimation

Pooling Sums of Squares in Two-Factor ANOVA

- Some argue that an insignificant interaction should be dropped from the model (i.e., pooled with error)

  Model: $y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}$

  $\text{SSE}^* = \text{SSE} + \text{SSAB}$

  $\text{df}_\text{E}^* = ab(n-1) + (a-1)(b-1)$

- Increases DF but takes a risk of inappropriate modeling

- p-value under the pooled model is different to interpret. It becomes a "conditional" p-value

- Should be considered when

  - The test statistic $MSAB/MSE$ falls substantially below the action limit of the decision rule (say $MSAB/MSE < 2$ for $\alpha = 0.05$, or reporting large p-value like > 0.25);

  - The degrees of freedom associated with MSE are small, perhaps 5 or less.

General Plan for Two-Factor ANOVA

- Construct scatterplot / interaction plot
- Run full model
- Check assumptions
  - Residual plots
  - Histogram / QQplot
  - Ordered residuals plot
- Check significance of interaction

Similar strategy for high order factor ANOVA

If Interactions Are Not Significant

- Determine whether pooling is beneficial
  - If yes, rerun analysis without interaction
- Check significance of main effects
  - If factor insignificant, determine whether pooling is beneficial
    * If yes, rerun analysis as one-way ANOVA
  - If statistically significant factor has more than two levels, use multiple comparison procedure to assess differences
    * Contrasts can also be used

If Interactions Are Significant but not Important

- Plots and a careful examination of the cell means may indicate that the interaction is not very important even though it is statistically significant.
  - The interaction effects may be much smaller in magnitude than the main effects;
  - The interaction effects may only be apparent in a small number of treatments as in Plot #2.
  - The subject area specialist (researcher) needs to be consulted in deciding whether an interaction is important or unimportant
- Use the marginal means for each significant main effect to describe the important results for the main effects, but carefully interpret the marginal means as averages over the levels of the other factor and not a main effect.
- Keep the interaction in the model.

If Interactions Are Significant and Important

The interaction effect is so large and/or pervasive that main effects cannot be interpreted on their own. Options include the following:

- Can take the approach of one-way ANOVA with ab levels to compare factor level means. Use linear combinations to compare various means (e.g., levels of factor A for each level of factor B).
- Use the interaction plots for discussion purposes on interactions between factors.
- Performs one-way ANOVA on one factor for a fixed level of the other factor.

Chapter Review

- Pairwise Comparison
- RCBD
- Factorial Design