Factorials, Blocking & Repeated Measures

Dewayne E Perry
ENS 623
Perry@ece.utexas.edu
Factorial Design

- **Factorial design**
  - 2 or more factors
  - Each with discrete values or levels
  - All possible combinations of the levels across all factors

- Enables the study of
  - The effect of each factor on the dependent variable
  - The effects of interactions between the factors on the dependent variable

- **Advantages**
  - Reduces the possibility of experimental error
  - Reduces the possibility of confounding variables

- **Disadvantages**
  - Difficulties when more than two factors, or many levels
### Comparing Multiple Conditions

#### Psychotherapy

<table>
<thead>
<tr>
<th>Drug</th>
<th>present</th>
<th>absent</th>
<th>mean</th>
</tr>
</thead>
</table>

#### Comparisons

- Column means: effect of psychotherapy
- Row means: effect of drug therapy
- Number of observations for mean has doubled
- Greater economy:
  - Each condition or group contributes data to several comparisons
Analysis of Variance

- Can decomposed 4 basis means into
  - **Grand mean**
    - \((8 + 4 + 4 + 2)/4 = 4.5\)
  - **Residual/interaction effects of group membership**
    - Group mean - grand mean = residual effect
    - PD: 8 - 4.5 = 3.5
    - D: 4 - 4.5 = -0.5
    - P: 4 - 4.5 = -0.5
    - O: 2 - 4.5 = -2.5
    - Sum of residual effects is always zero

© 2000-present, Dewayne E Perry
What Do We Learn

- **Group Mean** tells us general level of measurements
  - Usually not of great interest
- **Row Effects**
  - Better to receive drug therapy than not
- **Column Effects**
  - Better to receive psychotherapy than not
- **Interaction effects**
  - Better to receive both than either
  - Indication that it is better to receive neither than either is more than offset by row/column effects
Individual Differences

- Analysis so far does not tell quite the whole story
  - Does not take into account in various scores
  - Variability from mean - deviations
  - Call these deviations errors
    - Error = score - group mean
    - Large error: falls far from mean
    - Small error: falls close to the mean
  - Score = grand mean + row effect + column effect + interaction effect + error

- Variance
  - Drug therapy and psychotherapy
    - Large eta (.76), and significant (p = .012)
  - Interaction effect
    - Not trivial eta (.36), not close to statistically significant (p = .30)
    - Important in two way and higher order analyses of variance
    - Often misinterpreted
Interaction Effects

- \( \eta = \sqrt{\frac{SS_{between}}{SS_{between} + SS_{within}}} \)
- \( \eta \), like \( r \), represents square root of proportion of variance accounted for
  - But, \( \eta \) is a very non-specific index of effect size when it is based on a source of variance with \( df > 1 \)
  - Eg, \( \eta = .86 \) based on \( df = 3 \) is very large, but cannot say why it is large
  - When \( df = 1 \), \( \eta \) is identical with \( r \)
    - Drug/psychotherapy: \( \eta = r = .76 \)
    - Get all the ways of interpreting \( r \)
- While not significant (\( p = .30 \)), \( \eta \) is of promising magnitude (\( \eta = r = .36 \))
- We regard each effect size estimate as though it were the only one in the study
- Remember that when \( r^2 \) or \( \eta^2 \) exceeds 1.00
  - \( .574 + .574 + .130 = 1.278 \)
Testing Grand Mean

- Lack of interest in the magnitude of grand mean
  - In part due to arbitrary units of measurement often employed
- Sometimes, the constant of measurement may be of interest
  - When we failed to replicate a relationship obtained in an earlier experiment; compare our sample of subjects with an earlier sample
  - When dependent variable might estimate some skill that might or might not be better than chance
  - When our dependent variable might already be a difference score - e.g., the difference between pre and post test
    - GM is then equivalent to a matched pair t test
Unequal Sizes

- For one-way or omnibus analysis of variance it does not matter if we have the same number of units per condition or not.
- For two-way or higher order analysis must take special care when number varies from condition to condition.
  - One possible approach:
    - discard units til all conditions are equal
    - Almost never justified
- Multiple regression procedures available
  - Yield identical results when sample sizes equal
  - Vary substantially when samples sizes become increasingly unequal
  - procedure here represents yields closer to the “fully simultaneous multiple regression method” (FSMR) recommended by Overall et al 75
  - For factorial designs of any size, always having 2 levels per factor, yields results identical to FSMR
Higher Order Factorial Designs

- So far dealt only with 2 way
- Suppose current example had been done twice, once for females, once for males
  - Benefits: more subjects, more comparisons
  - $2 \times 2 \times 2$ factorial design = $2^3$ factorial
  - 3 factors: drug, psychotherapy, gender
  - $N = 2 \times 2 \times 2 \times 3 = 24$ if the same twice
  - $MS_{error} = 2.5$, adjustment factor $1/3 = .833$

- General strategy
  - Compute main effects first
  - Then two way interactions
    - Ie, residuals when two contributing main effects are subtracted from the variation in the two tables
  - Then the three way interactions
    - Ie, the residuals when the three main effects and the three two way interactions are subtracted from the total variation from the total variation among the 8 conditions
Higher Order Factorial Designs

- **Summary**
  - Effect sizes of .76, .76, and .36 are identical in earlier two-way analysis
  - F scores have all increases
  - p values are much smaller
    - As we would expect: study size increased
  - Tendency of gender to make some difference
    - gender and drug interactions significant

- **Generalized strategy**
  - Eg four way factorial
    - Construct all possible 2 and 3 way tables
    - Compute 4 main effects, 6 two way interactions, 4 three way interactions, and on four way interaction
Nature of Blocking

- Remember one of the ways to increase power:
  - Increase the size of the effect
- One way to increase effect size
  - Decrease the size of the within-group or error variation
- Blocking does this - increases precision
  - Stratifying or subdividing of subjects/samples
  - In such a way that those within a common block are more similar to each other on the dependent variable than they are to subjects/samples in another block/group
Nature of Blocking

Example
- Block according to anxiety level in study of new type of treatment
- Measure anxiety level prior to treatment
- Take top scores, randomly assign to T and C, iteratively
- Anxiety hst h m l lst sum
  T 8 6 3 1 1 19
  C 9 7 5 3 1 25
  sum 17 13 8 4 2 44
- Mean: T=3.8, C=5.0, GM=4.4
- Summary of sources
  - SS df MS F p eta
    - T 3.60 1 3.60 10.29 .04 .85
    - AB 77.40 4 19.35 55.29 .002 .99
    - R 1.40 4 0.35
- Comparison
  - Treatment effect is large and significant at p < .05
  - blocking variable effect even more so
Nature of Blocking

- Omitting blocking
  - SS  df  MS  F  p  eta
  - T  3.60  1  3.60  0.37  .56  .21
  - R  78.80  4  9.85

- Summary
  - Effect of treatment not very significant
  - But same mean squares
  - Residual variance decomposed into a large between blocks component compared with the small one in the blocked
  - Removes the large sources of variation known to be associated with the systematic pre-experimental differences among subjects
Benefits of Blocking

- Consider the size of the sample needed to achieve the same F ratio for blocked and unblocked analyses

\[ \text{reps} = \frac{MS_{\text{error}_{\text{unblocked}}}}{\text{(no. of blocks)}} \]

\[ \text{reps} = \frac{MS_{\text{error}_{\text{blocked}}}}{0.35 \times 5} = 140.7 \]

- Would need 140 pairs in the unblocked experiment to reach the same F we have in our blocked with 5 pairs
- But a difference in df and p: 4 and .04 vs 279.4 and .002
- To achieve the same significance level, would need 60.3 subjects/samples - a ratio of about 12 to 1 against blocked
- Example designed to show dramatic effects of blocking
- The larger the correlation between the blocking variable and the dependent variable
  - the greater the benefits
  - the greater the precision
Blocking and Covariance

- **Analysis of covariance** - a special case of ANOVA
  - observed scores adjusted for individual differences within conditions of
    - some predictor variable, or
    - some covariate known to correlate with the dependent variable
  - typical covariate is the pretest administration of the same (or similar) test that is to be employed as the dependent variable

- **Detecting interactions**
  - another benefit besides increasing precision
    - detection of interactions between experimental and blocking variable
    - usually in designs where each block has a number of replications for each treatment condition

- **Example (Treatment, Control, Mean, Residual)**
  - above 60: TM=6, CM=7 TR=.33, CR= -.33
  - 40-59: TM=3, CM=6 TR= -.67, CR=.67
  - below 40: TM=6, CM=7 TR=.33, CR= -.33
  - Middle age tend to benefit more than younger or older
Blocking and Covariance

- **Increasing precision**
  - **Might have used ANOCVA instead of blocking**
    - Sometimes better at increasing precision
    - Especially when pre and post test scores are highly correlated
    - Special case of ANOVA when perfectly correlated
  - **Useful rules of thumb (Cox 57)**
    - Blocking better when correlation is .6 or less
    - ANOCVA better when correlation .8 or more
    - Neither clearly better when correlation is between .6 and .8
  - **Blocking equally efficient for both linear and curvilinear; ANCOVA only when linear**
  - **Blocking also useful when the blocks differ in qualitative rather than quantitative ways**
  - **Blocking always imposes some cost in terms of loss of df for error.**
    - Cost usually small in relation to decreased MS error
    - If little reduction in MS error, can always unblock and recapture the lost df
Blocking and Repeated Measures

- Remember matched pairs t test?
  - Example of blocking: each pair of observations is a block
  - simplest form of repeated measures design

- All repeated measures designs are examples of blocked designs
  - the more positively correlated the successive observations on the same sampling units, the more we benefit from increased precision
    - versus between subjects design
    - typically get greatest precision when block on sampling unit itself
Blocking within Blocks

- **Example:** determining accuracy of decoding non-verbal clues from face, body, voice
  - Repeated measures design
  - 30 students, 60 clips, 20 of each
    - might have face items over-represented in last half, eg
  - Alternative: divide into 20 blocks of 3 each
    - randomly present one of each
    - doesn't necessarily increase precision
    - does eliminate possible confounding effects of the order of presentation
    - does allow us to learn from blocks x channels interaction
      - extent to which differences change over time
Use of Repeated Measures

- **Between subjects designs**
  - Sampling units only observed once
  - Variation based on individual differences between subjects
  - Subjects nested within their treatment conditions

- **Within Subjects Designs**
  - Very efficient to administer two or more treatments to the same sampling units
  - Sampling units serve as their own control
  - Subjects crossed by treatment conditions
  - The more correlated, the more advantageous this approach

- **Intrinsic nature of experiment might call for repeated measures type of design**
  - Effect of practice on learning a task
  - Effects in a longitudinal study of development
  - Series of tests or subtests for a variety of reasons

- **Simplest type: subjects measured twice**
  - Compare scores under each condition
  - Use non-independent t test to compare correlation
Fixed and Random Effects

- Distinction to help us employ the appropriate error term
  - **Fixed**
    - Selected particular levels of the factor
    - Cannot generalize to other levels
    - Includes most factors involving experimental manipulations, various organismic variables and repeated measures factors
  - **Random**
    - Randomly sampled from population of levels
      - Most common is that of sampling units, especially people

- In previous example
  - **If we regard between subjects as random**
    - We can test its significance only very conservatively
  - **If we regard it as fixed**
    - Restrict inferences to this four subjects
    - Can test subject factors against sessions x subjects interactions
  - **Will consider all combinations**
    - Fixed and random
    - For between and within subjects factors
Fixed and Random Effects

- **Examples**
  - 4 countries as our between sampling units factor
    - If only interested in these 4, fixed
    - If view as a sample from which we want to generalize, random
  - Longitudinal design with a summary score for each country for each of 3 decades
    - Scores are repeated measures, or within sampling units factors
    - Regard as fixed if we have chosen them specifically
    - Regard as random if we view as samples from which to generalize

- **General principle that helps in determining the appropriateness of the error term**
  - The effect (fixed or random) we to test are properly tested by dividing MS for that effect by the MS for a random source of variation
Latin Squares

- Consider three drugs and 4 patients
  - Suppose each subject given three drugs in the same sequence
    - Confound drug and order
    - Suppose A is best
    - Rival hypothesis is the first is best
  - Use counterbalancing to avoid confounding
    - Sequence is systematically varied
    - Essential in organization and sequencing presentation
      - Primacy: opinions influenced by arguments presented 1st
      - Recency: opinions influenced by what is presented last
Latin Squares

- Latin squares has counterbalancing built in
  - Nr of rows equals the nr of columns
  - The letter presenting treatments appears in each column and row only once
  - Effects of treatment, order and sequence are isolated - systematic counterbalancing

<table>
<thead>
<tr>
<th>Order</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq 1</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>seq 2</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>seq 3</td>
<td>C</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
Latin Squares

Analysis

★ Sequence effects tell how sequencings differ
★ Order effects tell how orders differ
★ Where is the treatment effect in latin squares?
  ➢ In a 2 x 2 latin square, treatment effect is the sequence x order interaction effect
  ➢ Compares the two diagonals
  ➢ Sources of variation:
    ✓ Sequences: 1 df
    ✓ Orders: 1 df
    ✓ Treatments (s x o): 1 df
★ Significance testing is a problem
  ➢ No df available for error terms for the 3 sources of variation
  ➢ Can use the mean of the MS 's of the other two in computing very conservative F 's
★ As the size of the latin square increases, the df for an error term increases
  ➢ Get less conservative F 's
  ➢ More accurate in the sense of less Type II errors
  ➢ On average F 's will be too small
Other Counterbalancing Designs

- What if we have an unequal number of subjects and treatments?
- Useful strategies
  - Multiple squares
  - Rectangular arrays
- Rectangular arrays
  - Eg, 3 treatments and 6 subjects
    - Could do two 3x3 squares
    - Could assign each subject a unique sequence of treatments since 3! = 6
  - If 4 treatments, then would need 4! = 24 subjects
  - Call such designs \( t \times t! \) designs (\( t = 2 \) is latin squares)
- Fewer than \( t! \) sample units?
  - Multiple latin squares
  - Random assignment from \( t! \) Sequences
    - Constraint: ensure maximum degree of balancing in the resulting samples
- More than \( t! \) Sample units?
  - Multiple rectangular arrays
  - Subjects-within sequences designs
Other Counterbalancing Designs

- **Subjects-within sequences Designs**
  - **Suppose 2 × t! Subjects**
    - Could randomly assign half the subjects to each of two rectangular arrays
      - Treat each array as a different and replicated experiment
    - Could randomly assign from t! Sequences
      - Constraint: ensure maximum counterbalancing
  - **Suppose 18 subjects, 3 treatments**
    - 3! = 6, assign 3 subjects at random to each sequence
    - Subjects are not confounded with sequence as in latin squares
    - Subjects are nested within sequences so sequences can be tested
Other Counterbalancing Designs

★ Noteworthy features of this design

➢ More than a single error term in the design
  ✓ Previously only one error term
  ✓ Usually associate with differences among subjects
  ✓ Have that: subjects within sequences
    • Used to test whether sequences differ from each other
    • Note: error is within conditions but between subjects
  ✓ Other error term: orders x subjects within-sequences interaction
    • Test all within-subjects sources of variation
    • Is itself a within-subjects source of variation
    • Formed by crossing repeated measures factors by the random factor of sampling units
Other Counterbalancing Designs

- To test for treatments, we must reach into the order x sequences interactions and pull out the variation of the treatment means around the grand mean

- Analysis – pretty much same as already discussed
  - Between-subjects SS is broken down into a sequences SS and a subjects-within-sequences SS
  - Later is the difference between the between-subjects SS and the sequences SS
  - etc
3 or more Factors

- So far considered only two factors
  - Between-subjects factor
  - Within-subjects or repeated measures factor
  - Often have two or more of each

- 2 or more between-subjects factors
  - Does not increase complexity of the design as much increasing the number of within-subjects factors
  - Eg, 4 subtests of personality test of three age levels and two genders
Fixed vs Random Factors

- Examples so far have assumed that all factors other than subjects within conditions have been fixed
  - The most common situation

- Example:
  - 5 f and 5 m teachers, 4 schools, teach lesson to 3 pupils, one designated as bright
  - If fixed, 2 error terms
  - If school is random factor, 5 error terms
Do repeated measures help?

- Basic utility: subjects are their “own control”
- High correlation yields advantages
- If low correlations, little advantage
- Note on Assumptions
  - For $F$ and $t$ tests
    - Independence of errors
    - Homogeneity of variance
    - Normality
  - Additional needed for repeated measures
    - Homogeneity of correlation coefficients among the various levels of the repeated measures factors
    - Patterns of inter-correlations is consistent among various levels is consistent from level to level