Analysis of Messy Design

“Does it Matter?”

Messy Data

Agenda

- “Messy Data”
- Why do Analysts Care?
- Example
  - Single Treatment ANOVA w/ Heterogeneous Variance Terms
- Learning Points
- References: Analysis of Messy Data: Volume I, Design of Experiments, George A. Milliken and Dallas E. Johnson. Chapman and Hall, New York, 1992.

What is Messy Data?

- Failure of statistical assumptions that influence underlying distribution theory:
  - Homogeneity of variance across treatments
  - Severe outliers in data sets
  - Missing treatments and/or treatment combinations in an experiment
  - Unequal number of observations in experiments

Why Do Analysts Care?

- Experiments influence action: decisions on the influence of different treatments on a measured response
- Sometimes, we don’t.
- Appropriate to study the differences.

The Model of Interest

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

$$\text{Where,} \quad \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

The variance terms may be different for each population or treatment applied.

Try to address three questions:

- Are the treatment means equal?
- Can the analyst compare the different means using a pair-wise approach?
- Is the analyst able to use linear contrasts to test hypotheses of interest?

Distribution Results

For large n_i, let $Z = \dfrac{\left(\left(\sum_i c_i \hat{\mu}_i - \sum_i c_i \mu_i \right) \Big/ \sqrt{\sum_i c_i \sigma_i^2 \Big/ n_i} \right)}{\sqrt{\sum_i c_i \hat{\sigma}_i^2 \Big/ n_i} \Big/ \sqrt{\sum_i c_i \sigma_i^2 \Big/ n_i}}$

The above statistic has an approximate N(0,1) distribution.

Satterthwaite's Approximation

- If sample sizes are not large, then:

$$\hat{\upsilon} = \left(\sum_i c_i^2 \hat{\sigma}_i^4 \Big/ n_i \right)^2 \Bigg/ \sum_i \left[c_i^4 \sigma_i^4 \Big/ n_i^2 (n_i - 1) \right]$$

$$\text{Reject } H_0 : \sum c_i \mu_i = a \text{ with } |t_c| = \frac{\left| \sum_i c_i \hat{\mu}_i - a \right|}{\sqrt{\sum_i c_i \hat{\sigma}_i^2 \Big/ n_i}} > t_{\alpha/2, \hat{\upsilon}}$$

Learning Points

- The data will guide analysis approach; assumptions are key
- Software applications lend themselves to analysis of the treatment model: SAS is a good starting point with GLM; discourage Minitab.