MODEL BUILDING

Model building: The process of deciding what model to use for the context.

ANOVA models discussed so far:

- one-way
- cell-means
- two-way main-effect models
- two-way complete models

Recall from the reaction time data: The two-way and main-effects models give different test statistics.

Choosing models:
- Use existing well-supported theory or knowledge.
- Empirically: based on data.
- A combination.

Example of empirical choice of model: If we use a complete two-way model, test for interaction, find no evidence of interaction, we might decide that in the future, it makes sense to use a main effects model for that situation. (Better: Make this decision only after two experiments, each with adequate sample size, show no evidence of interaction.)

Considerations to be taken into account in empirical model building:

1. *Model building and inference to answer the original questions should be done using different data.*

Example: If we use the two-way complete model for the reaction time data, test for interaction, and find no evidence of interaction.

- It is *not* legitimate to switch to the main-effects model for further hypothesis testing and confidence intervals *with the same data*. This would be changing the model on the basis of data; significance levels and confidence levels produced by the new model in this instance would not be correct.
- Instead, either:
  - o continue with the complete two-way model to address our original questions (best), or
  - o collect new data and use the main effects model with the new data. (more iffy)
- It *would* be legitimate to use the finding that the data provide no evidence of interaction to consider a main-effects model for *future* experiments in the same context.

Sometimes data for use in model building and data for use in addressing the original questions are collected at the same time.

• To do it right: Randomly divide data into one data set for model building, one for model verification, and a third data set for addressing the original question (after deciding on a model).

• The randomization needs to respect experimental design constraints.

• Collecting this much data is not always possible because of time and budget constraints.

  o So do the best you can with the data available.

  o e.g., address the original question based on the full model, and then treat model building as "data snooping".

  o Point out limitations and uncertainties in the write-up.

2. Model building based on hypothesis tests raises legitimate questions of multiple comparisons.

Alternate approach ("prediction based approach"): Select model based on Mallow's C statistic (discussed in M 384G), or Akaike Information criterion, Bayesian information criterion, or ….

3. Although there is some difference of opinion, there seems to be fairly general agreement that ANOVA models should be *hierarchical*:

  If an interaction term is included in a model, then each factor involved in the interaction should also be included in the model.

In other words, if the interaction of factors A and B is included in the model, then both A and B should be included.