Miscellaneous ANOVA issues

A. Balance

- We’ve already seen that ANOVA models can be parameterized as regression models. When there are equal sample sizes in all groups \((r_i \equiv r \quad \forall i)\), we build a regression model using only indicator variables to distinguish groups, and no other types of variables. This allows us to simplify the sums of squares calculations.
SSModel = \( Y' (H - \frac{J}{N}) Y \Rightarrow SST = r \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \)

SSError = \( Y' (I - H) Y \Rightarrow SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 \)

SSTotal = \( Y' (I - \frac{J}{N}) Y \Rightarrow SSTotal = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 \)

The calculation done with regression notation exactly equals the calculation done with ANOVA notation.
• For one-way ANOVA, when the $r_i$ are not all equal, this simplification still works. But with higher-way ANOVAs, it will no longer work.

For unbalanced higher-way ANOVAs, analyses are carried out using the regression framework. That is, group effects will be designated by indicator variables. Sub-group effects will be formed from interactions between indicator variables.
Thus, it is critical to understand how the coding of your indicator variables changes the interpretation of the estimated parameters. This tells you what the regression coefficients represent.
• Under higher-way ANOVAs, the order in which terms in the model are entered will affect the sum of squares calculations. Type I sums of squares (partial) will not equal Type III sums of squares (sequential), and you must understand which sums of squares give the appropriate F-tests to answer your questions of interest.
- For higher-way random effects and mixed effects ANOVAs, no exact F-tests exist. You cannot get the same test from both PROC GLM and PROC MIXED. Approximations must be used under PROC GLM (least squares estimation), or PROC MIXED (REML) must be used.

For full coverage of random and mixed effects models, take PubH 7430, Methods for Correlated Data.
B. When $t$ is huge

- A typical micro-array experiment may look at the expression of $> 10,000$ genes. Thus, rigorous multiple comparisons adjustment is called for.

- However, Bonferroni, Sidak, etc. are much too conservative, because they are based on assuming your hypothesis tests (e.g., all pairwise comparisons) are independent. That is most likely not true for genetic data, since some genes can be highly correlated.
Some newer multiple comparisons procedures have been developed that work by computing an unadjusted p-value (e.g., for each pairwise comparison) and then adjusting those p-values to control experimentwise error.

These can be implemented in PROC MULTTEST in SAS or using the S+ArrayAnalyzer in S-Plus.
• Examples:

  - Benjamini-Hochberg correction of the False Discovery Rate
  - Hochberg adjustment
  - Step-down procedures (Westfall-Young corrections)
  - Bootstrap procedures
  - Permutation procedures