Assumptions and their importance: mostly like T-test
- independence: crucial.
  - if violated, e.g., by cluster effects, se wrong. p-values and ci's wrong
- equal variance: more important than in T-test
  - overall F test: robust to unequal variance when equal sample sizes
  - Comparisons between specific means: equal variance matters a lot
- normality: not very important

Assessment:
- My primary tool: residual vs predicted value plot
  - Good: "flat fat sausage"
  - Bad: trumpet-shaped.
- Often see higher mean is more variable
  - Log transformation often fixes this pattern
- Plots of residuals for other issues:
  - X = time, Y = residual: detect serial dependence
  - X and Y = spatial coordinates, symbols for > 0 or < 0. look for clusters

After the ANOVA:
- F test should be just the start of the analysis
- Linear contrasts for *a-priori* questions
- Multiple comparisons adjustments for large number of vague questions

Contrasts:
- Coefficients, l_i, from structure of the question
- $g = \Sigma l_i \overline{Y}_i$ estimates $\gamma = \Sigma l_i \mu_i$
- se $g = s_p \sqrt{\Sigma l_i^2 / n_i}$
- df are df of s_p
- tests and ci's using t-based inference
- can ask Q about more than 1 contrast simultaneously
  - F test based on SS for the contrast, no details

Quantitative treatments:
- Treatment is an amount of something,
  - e.g. fertilizer amount or years of exposure, $\cdots$
- Best analysis takes advantage of relationship between amounts
- common Q: is there a linear effect of the treatment
- Answer using a contrast with linear contrast coefficients
- $l_i = X_i - \overline{X}$, where X_i is the amount for trt i
  - $\overline{X}$ is mean of the amounts, ignoring # replicates per group
- value of g is not immediately interpretable
  - value related to regression slope, but not equal to it.
- So approach most useful as a test

Multiple testing / multiple comparisons: the issue
- Remember the concept of a p-value: how unusual is some observed result?
  - probability of 0.05 = 1/20 is unusual when look at one test
  - what if do 100 tests?
    - 1/20 is no longer unusual. Expect 5 events when try 100 times
- Lots of tests when compare all pairs of groups
  - 15 groups: 105 tests, 20 groups: 190 tests
- Or when have many responses
- Focusing on the test with the smallest p-value is almost certainly misleading

Multiple testing: approaches
- Ignore it, report usual (unadjusted) p-values
  - More common in exploratory (hypothesis generating) studies
  - Many of the "significant" results probably are not
  - Not recommended!
- Ignore it after first checking for a difference somewhere
  - "Fisher's protected LSD": do the overall ANOVA
    - If $p \leq 0.05$, analyze all pairwise differences without adjustment
    - If $p > 0.05$, stop - don't even think about pairwise differences
- Use a stricter criterion for "significant": family-wise error rate
- Change the criterion: false discovery rate

Multiple testing: family-wise error rate methods
- Comparison-wise error rate: P[declare one comparison significant when no difference]
- Consider a family of tests:
  - all pairwise differences, all comparisons to control, all linear contrasts
- Family-wise error rate: P[declare any test significant when no differences anywhere]
- Adjustment method depends on the statistical properties of the family
- We consider two and mention the third
  - All pairwise differences: Tukey honestly-significant difference
  - k tests, no better structure: Bonferroni: $p_{adj} = k * p_{unadj}$
  - Any linear combination: Scheffe (rarely used)
- There are a huge number of other methods
  - Some for specific circumstances (Dunnett: many groups to a single control)
  - Some are different approaches to adjustment
- All methods make it "harder" to declare a difference "significant"
  - Reduces number of false differences declared "significant" (good)
  - Also reduces number of true differences declared "significant" (bad, lower power)

Multiple testing: false discovery rate (FDR) methods
- Motivation: when many tests, FWER methods are very conservative
  - Very hard to detect any difference
- False discovery rate:
  - Very common in genomics, measure expression of 10000 genes, which changed?
  - Given a list of "discoveries" (e.g., "significant" effects), what fraction are wrong?

Examples:
10000 tests, no true difference anywhere,
unadjusted p=0.05 ⇒ 500 "significant", all false. FDR = 100%
10000 tests, 1000 have a true difference, assume all detected
unadjusted p=0.05 ⇒ 1000 true "significant" + 450 false, FDR = 450/1450 = 31%
adjustment, Benjamini-Hochberg: Specify desired FDR (e.g., 10%)
manipulate the "usual" p-value to produce a list of "significant" tests
with the property that (on average) % of that list that are false is ≤ FDR

What is the difference between targeted questions, the "overall" F test, and all pairs?
Targeted questions are comparisons of specific groups (or specific linear combinations)
more likely to find a difference when that effect is present
usual F test looks for any difference; all pairs and FDR look for any difference
less likely to find a difference when you aren't sure which effects to look for

PMD approach to "after the ANOVA":
Specific questions always better
answer those questions!
When small number (≈ or ≤ # groups), no adjustment needed
Doing an F test first is not necessary - concern is with the specific questions
When no specific questions
use some sort of multiple comparisons / multiple testing adjustment
details depend on the family of comparisons
Compromise: (e.g. Am. J. Clin. Nutrition advice to authors)
Identify primary comparisons prior to looking at data
no adjustment required
Explore further questions
with some sort of adjustment and identify as exploratory
Goal of the compromise is to retain power for primary effects
while reducing number of incorrect assertions about other effects