Ch 11 – Analysis of Variance

Def'n: ANalysis Of VAriance (ANOVA) is a procedure to test the equality of three or more pop'n means. NOTE: the name of the test refers to comparing different sources of variability; it WILL test differences among means.

Assumptions:

1. The populations are all normally distributed.
2. The populations all have the same standard deviation.
3. The samples from different populations are random and independent.

Checking Assumptions:

- Assumption #1 is checked with histograms/boxplots for each group.
- Assumption #2 is more critical but harder to assess. Still, we can use side-by-side boxplots (or the informal rule from Ch. 9). If samples sizes (approx.) equal, the condition can be relaxed.
- Assumption #3 by analyzing the experiment design.

Notation:

- x_{ij} = observation for j^{th} subject in i^{th} group (a.k.a. treatment)
- $i = 1, \ldots, k$ indexes groups
- $j = 1, \ldots, n_i$ indexes subjects within groups
- n_i = # of observations in i^{th} group; $n_T = \sum n_i$ = total # of observations
- $\bar{x}_i.$ and s_i^2 are sample mean and variance for the i^{th} group
- $\bar{x}..$= grand mean = mean for combined sample:

$$\bar{x}.. = \frac{1}{n_T}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n_T}\sum_{i=1}^{k} n_i \bar{x}_i .$$

Statistical model, parameters, hypotheses:

Using a linear statistical model, each observation can be represented by

$$x_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad i = 1, \ldots, k; j = 1, \ldots, n_i$$

where x_{ij} are independent random observations, μ is the *overall mean*, τ_i is a parameter associated with the i^{th} group called the i^{th} *treatment effect*, and ϵ_{ij} is a random error.

- random order of obs'ns & uniform environment→ "completely randomized design"
- consider that only k groups exist → fixed-effects model
- parameters: σ, $\tau_1, \ldots, \tau_k$; the τ_i are defined such that $\sum \tau_i = 0$.
- H_0: $\mu_1 = \ldots = \mu_k$
  H_A: $\mu_i \neq \mu_j$ for some i and j

As in Section 12.6, a model with different sources of variability suggests an identity:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}..)^2 = \sum_{i=1}^{k} n_i(\bar{x}_i. - \bar{x}..)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i.)^2$$

$$SST = SSTr + SSE$$

Also, the presence of "squares" suggests a ratio test. Thus, for the above H_0, we have

$$F_0 = \frac{SSTr/(k-1)}{SSE/(n_T-k)} = \frac{MSTr}{MSE} \sim F_{k-1,n_T-k}$$

Reject H_0 when F_0 is large (greater than 10 works for most values of α)

- Rationale:
    - If H_0 is true, then both types of variability are identical. → $F_0 \approx 1$
    - If H_A is true, then *MSTr* should be larger. → $F_0 > 1$

ANOVA Table:

Source	**SS**	**df**	**MS**	***F***	***p*-value**
Treatments	*SSTr*	$k-1$	*MSTr*	*MSTr*/*MSE*	?
Error	*SSE*	n_T-k	*MSE*		
Total	**SST**	n_T-1			

Calculating SS is tedious! More important to understand values and how they relate to other values in the ANOVA table. If you received an incomplete table, you should be able to fill it in.

Ex11.1) Consider a *k*-mean problem. Five observations are gathered for each group. Use the given information in the table below to answer the questions. Assume all populations are normal with some common variance.
(a) What is *k*?
(b) What are the test statistic and *p*-value for the test to determine if any of the *k* groups are different?

Source	**SS**	**df**	**MS**	***F***	
Treatments					
Error			6		
Total	**360**	**54**			

(filled out in class)

(Additional Excel example seen in class with full hypothesis test.)

<u>Summary for Single Factor ANOVA</u>
- check assumptions.
- rejecting H_0 does NOT mean all means are different, AT LEAST ONE is.
- not rejecting H_0 "finishes" the analysis, rejecting H_0 requires subsequent determination of which means are significantly different from the rest.