Stat 471/571, Fall 2022

Concepts of ANOVA tables,
with two ways to think about the sum-of-squares and associated degrees of freedom

ANOVA tables have three uses:

1. They summarize a study's design structure and treatment structure in way that is often more informative than a model equation. They provide more information than an R/SAS/JMP model specification.
2. They provide a way of organizing the computation of F statistics.
3. They can provide information about appropriate error terms for F statistics and follow-on comparisons of means.

Basic computations, leading to the F statistic, demonstrated using 1 way ANOVA:
Example is the motivation study from HW 2, 5 treatments, 10 subjects per treatment

Two sources of variability: treatment (fixed effect) and error (variability between subjects within a treatment)

Source	df	SS	MS	F	p-value
treatment	4	2189.3	547.33	2.55	0.052
error	45	9657.7	214.62		
c.total	49	11847.0			

- df = degrees of freedom, computed from the design, discussed below
- SS = sum-of-squares, computed from the data, discussed below
- MS = mean square = SS / df, for any row where it makes sense. Doesn't make sense for corrected total.
- F: for treatments = MS treatments / MS error
- p-value: upper tail probability for the F statistic. F statistics have two degrees of freedom, for the numerator MS and for the denominator MS. Here those are 4, 45.
- Note: error is often called Residual. Same thing.
- The MS error is one of the most useful numbers in the table. This is the pooled variance, i.e. the estimated variability among observations within a group. The pooled sd = $\sqrt{MS_{\text{error}}}$.

Once you have the df and the SS, the rest of the ANOVA table is just computations, with the last step requiring a computer or tables to get the tail probability.

Model for the data: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.
Subscripts: i for treatment, j for observation within treatment
Interpretation of terms in model:

- μ is a value common to all treatments.
  Often thought of as the overall average, but details depend on software.

- α_i are deviations of each treatment mean from that common value, μ. When all groups have exactly the same mean, all $\alpha_i = 0$. The mean for the i'th treatment is $\mu + \alpha_i$, which does not depend on how μ is defined.

- ε_{ij} error attached to the ij'th observation.
  Deviation of the ij'th observation from its treatment mean, $\mu + \alpha_i$.

Computing SS: Two views:

1. SS as formulae.
   When the data are balanced (equal sample sizes for all groups), there are formulae for the SS. These involve two design characteristics: k = # treatments and n = # observations per treatment, and two summaries of the data: $\overline{Y}_{i.}$ = treatment average for the i'th treatment and $\overline{Y}_{..}$ = average of all observations. A subscript of dot indicates averaging over that subscript.

   - SS treatment: variability between treatment means, on a per-observation basis
     $$SS_{\text{treatment}} = n \sum \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$$
   - SS error: variability of obs. (j) around their trt. mean, pooled over treatments (i)
     $$SS_{\text{error}} = \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i.}\right)^2$$
   - SS corrected total (c.total): variability of obs. ij around the overall mean $\overline{Y}_{..}$
     $$SS_{\text{c.total}} = \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{..}\right)^2$$
   - It is an algebraic identity that $SS_{\text{c.total}} = SS_{\text{treatment}} + SS_{\text{error}}$

2. SS as comparison between models. Consider fitting two models to the data:

   (a) The reduced model: all observations have same mean, i.e. no difference among population means. This is the null hypothesis for the F test. The model is $Y_{ij} = \mu + \varepsilon^*_{ij}$. Calculate the error SS for this model. That's $SS_{\text{c.total}}$.

   (b) The full model: each treatment has a different mean. That's the alternative hypothesis for the F test. The model is $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. Calculate the error SS for this model. That's SS_{error}.

3. Then compute the difference in error SS between the two models.
   $SS_{\text{treatment}} = SS_{\text{c.total}} - SS_{\text{error}}$

4. If the null hypothesis is wrong, and there is at least one difference among treatments, the different means model will fit much better. The difference, $SS_{\text{treatment}}$, will be large.

5. If the null hypothesis is reasonable, and the treatments really do have the same means, both models will fit equally well. The difference, $SS_{\text{treatment}}$, will be close to zero. How close to zero depends on the number of treatments, k, and the variability within each group.

6. The F statistic compares the observed difference to what would be expected given the number of treatments and error variability.

7. The model comparison approach works even when sample sizes are not equal. Most formulae only work for equal sample sizes, although there are some exceptions.

Degrees of freedom:
Each SS has an associated df that can be computed based on how the SS were computed. Since each SS is proportional to a variance, the same ideas apply. How many pieces of information minus the number of parameters that need to be estimated.

1. When SS are computed using formulae

   - df for treatment SS: There are k groups = treatments each with a group average. The SS treatment is computed from those k values, and you have to estimate the overall average. So, there are $k - 1$ df for treatments.
   - df for error SS: There are $k \times n$ observations, and you have to estimate k group means. So there are $k\,n - k = k(n - 1)$ df.
   - df for corrected total SS: There are $k\,n$ observations, and you have to estimate 1 overall mean. So there are $k\,n - 1$ df.
   - Note that the df add up: $(k - 1) + k(n - 1) = k\,n - 1$.

2. When SS are computed by model comparison. It is most straightforward to consider the error df for each model.

   - The null hypothesis model (the c.total line in the ANOVA table): There are $k\,n$ observations and the model has 1 parameter, the overall mean. So the error df has $k\,n - 1$ df.
   - The full model (the residual line in the ANOVA table): There are kn observations and the model has k parameters, one for each group mean. So the error df has $k\,n - k$ df.
   - The change in SS between the two models has df = change in error df.
     $(k\,n - 1) - (k\,n - k) = k - 1$.