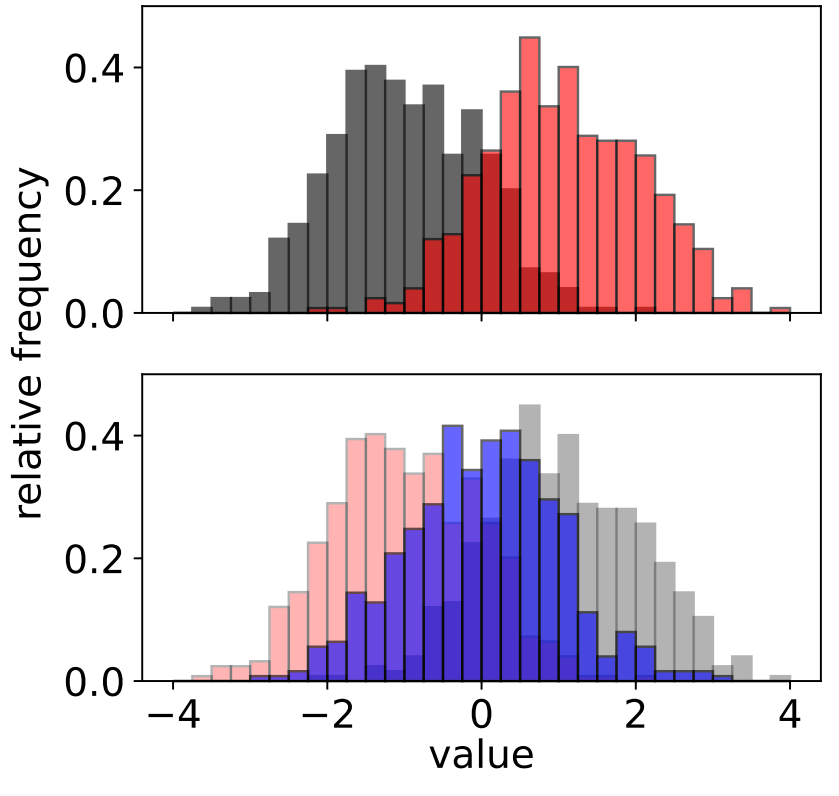
EGM101 – Skills Toolbox

Week 8, Part 4: ANOVA

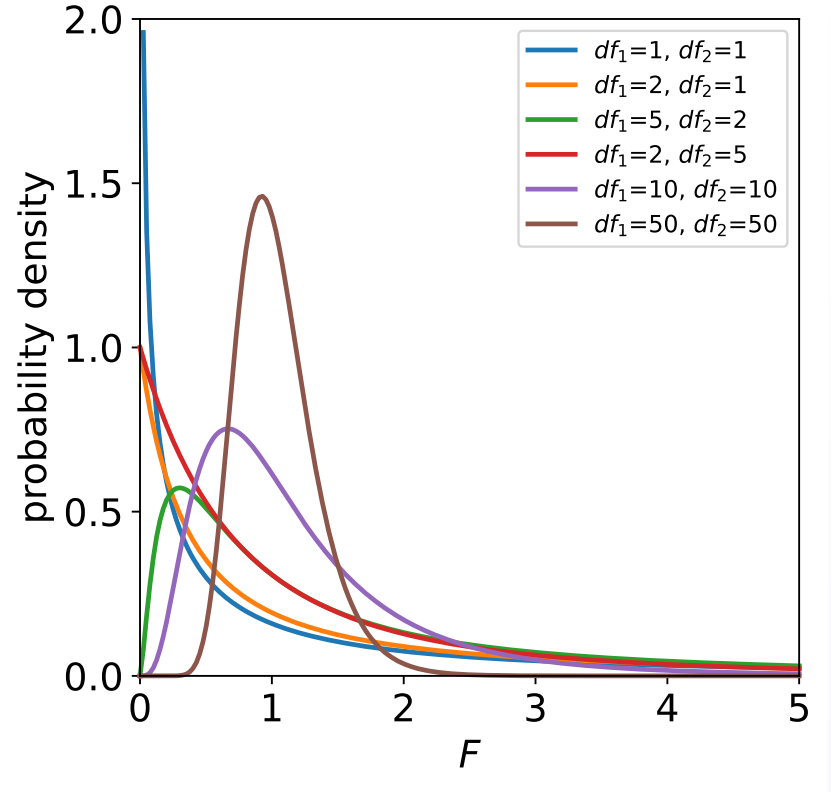
What happens if we have >2 samples?

- Question: do our samples come from populations with different means?
  - One sample: one-sample *t*-test
  - Two samples: two-sample *t*-test
  - Three samples: ¯_(ツ)_/¯
- Problem: multiple tests increases likelihood of Type-I error
- Solution: Analysis of Variance (ANOVA)



The F-Distribution

- Ratio of two “chi-square” distributions
  - More on this later
- Two different degrees of freedom:
  - $df_1 = k - 1$, where k is the number of groups
  - $df_2 = n - \mathrm{k}$, where n is the sum of the individual sample sizes



One-way ANOVA: Assumptions

- Dependent variable is continuous
- Independent variable is categorical
- Populations are normally-distributed
- Independent, random, samples
- Populations have “equal” variances
  - i.e., ratio between 0.5, 2
  - cf. unpaired samples *t*-test

Mean squares

- $MS_{treatment}$ (treatment mean square):
  - Depends on the differences *between* groups
  - Essentially, the (sample) variance of the sample means
  - Has $k - 1$ degrees of freedom
- MS_{error} (error mean square):
  - Depends on the differences *within* groups
  - Essentially, the pooled estimate of population variance
  - Has $n - k$ degrees of freedom, where $n = \sum_i n_i$

$$SS_{\text{treatment}} = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x} \right)^2$$

$$MS_{\text{treatment}} = \frac{SS_{\text{treatment}}}{k-1}$$

$$SS_{\text{error}} = \sum_{i=1}^{k} \left(n_i - 1 \right) s_i^2$$

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{n-k}$$

The One-way ANOVA Identity

- Total sum of squares, SS_{total}
  - Total variation among all sample data
- Equals sum of treatment, error sums of squares
  - In effect, can partition SS_{total} into treatment, error sums of squares
- With this, we don't actually have to calculate all 3 sums of squares

$$SS_{\text{total}} = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{error}}$$

The F-statistic

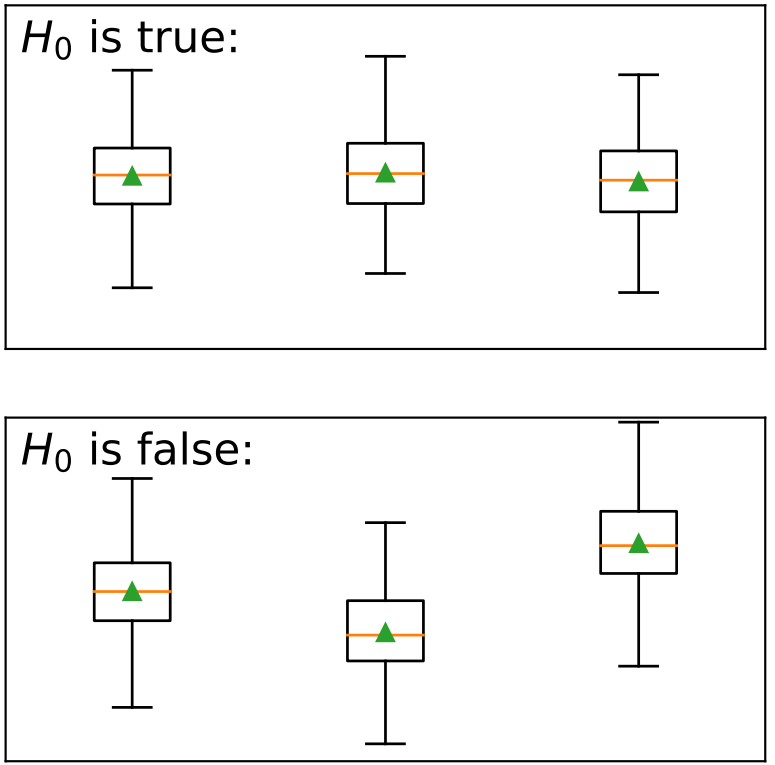
- Compares variation between sample means to variation within samples:

$$F = \frac{MS_{\text{treatment}}}{MS_{\text{error}}}$$

- Larger values of $F \rightarrow$ more variation *between* than *within*
  - In other words, more likely that samples have different population means
- Degrees of freedom:
  - $k - 1$: numerator
  - $n - k$: denominator
- cf. coefficient of determination, R^2:
  - R^2: Ratio of explained variability to total variability
  - F-statistic: ratio of explained variance to unexplained variance

One-way ANOVA

- Hypotheses:
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - H_a: $\mu_i \neq \mu_j$ for some $i \neq j$
- All ANOVA will tell us is there is a difference
  - Need to do post hoc tests to learn more
- Steps:
  - Calculate F-statistic
  - Compare to F-distribution for df_1, df_2
- Test is always right-tailed



Summary

- ANOVA helps us determine if there is a difference in means between multiple (> 2) samples/groups
- Compares variance between groups to variance within groups
  - If variance between groups is greater than variance within groups, likely they are populations with different means
- Remember: ANOVA only tells us if there is a difference, not what the difference is!

Additional Resources

- Illowsky and Dean, Chapter 13
- Weiss, Chapter 13
- How F-tests work in ANOVA [Jim Frost]
- Using Post Hoc Tests with ANOVA [Jim Frost]
- Calculating total sum of squares [Khan Academy]
- Calculating SS_{error} and $SS_{\text{treatment}}$ [Khan Academy]