BST 203 Lab 1: ANOVA
July 25th 2022

Review: Analysis of Variance (Chapter 12)

- Analysis of variance (ANOVA) is an extension of the two-sample t-test to k > 2 groups.

- The basic idea:
  - Two sources of variation:
    1. "Within group variation" – variation of individual values around their group mean.

$$s_W^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

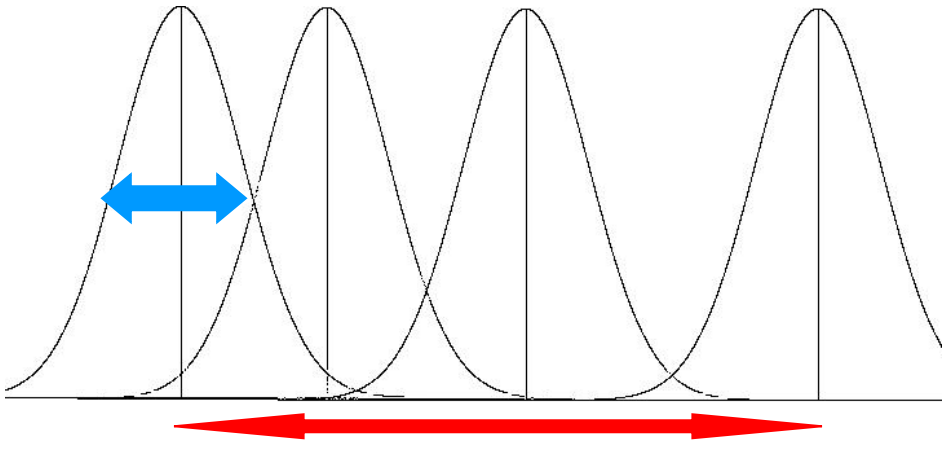
    2. "Between group variation" – variation of the group means around the overall mean.

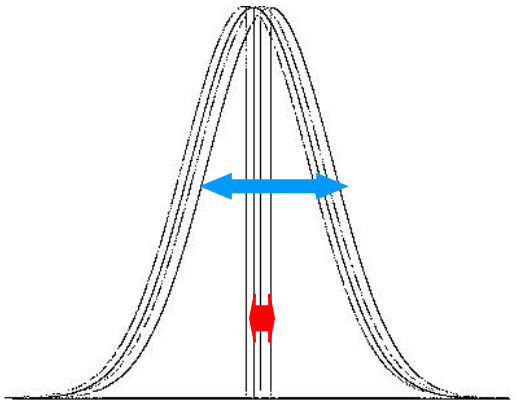
$$s_B^2 = \frac{(\bar{x}_1 - \bar{x})^2 n_1 + \cdots + (\bar{x}_k - \bar{x})^2 n_k}{k-1}$$

    where

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}$$

  - If s_B^2 is big relative to s_W^2, then at least one of the means is different from the others.





- Assumptions
  - Samples from the k populations are independent.
  - Samples from the k populations are normally distributed.
  - Variances in the k populations are equal. i.e., $\sigma_1 = \sigma_2 = \ldots = \sigma_k$
    (We do not test this assumption, it is assumed true. You can do Bartlett's test with `STATA`)

- Test Statistic

$$F = \frac{s_B^2}{s_W^2}$$

- F distribution:
  - Two types of degrees of freedom:
    Numerator: k-1 (corresponds to the df for variation between groups)
    Denominator: n-k (corresponds to the df for variation within groups)
  - The F-statistic cannot assume negative values (do NOT double the p-value)

- Multiple comparisons:
  - If null hypothesis is rejected, conclude that at least one of the means is different from the others.
  - To determine which means are different, perform all $\binom{k}{2}$ pair-wise comparisons of the means with two-sample t-tests and use a Bonferroni correction $\alpha^* = \alpha / \#\ tests\ performed = \alpha / \binom{k}{2}$ because the tests are not independent.
  - In each pair-wise t-test, use s_W^2 as the pooled standard deviation and *n-k* degrees of freedom (don't forget to double the p-value for 2-sided tests when using a t-test).
  - The Bonferroni correction is highly conservative (low power) and may fail to reject a difference between means when one actually exists.

Example
We will look at a dataset collected from the 2010 World Cup. The data contains variables regarding a player's name, country, position, minutes played, and number of passes. We're interested in seeing if different positions pass more than others. The three positions listed are forward, midfielder, and defender. We will use a transformed version of passes called logpasses (the log of passes). The data, `lab1.dta`, can be found on the course website in the data section of the "Labs" tab.

- What is the overall mean logpasses? What is the sample size/ mean logpasses / variance in each group?

  Helpful `STATA` code:

```
summarize logpasses
sort position
by position : summarize logpasses
```

- Are the assumptions met for ANOVA?

  Helpful STATA code:

```
histogram logpasses, by(position) freq
graph box logpasses, over(position)
```

- State the null and alternative hypotheses.

- What is the estimate of the within-groups variance (by hand)?

- What is the estimate of the between-groups variance?

- What is the value of the test statistic and what distribution does the test statistic follow?

- What is the p-value for the test?

- What do you conclude?

- Perform ANOVA with STATA:

REFERENCE: oneway [variable of interest] [group variable], [correction]

```
. oneway logpasses position, bonferroni

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      20.7070961      2   10.3535481     16.78     0.0000
 Within groups      214.083692    347   .616955885
------------------------------------------------------------------------
    Total           234.790788    349   .672752974

Bartlett's test for equal variances:  chi2(2) =   0.6581  Prob>chi2 = 0.720

                     Comparison of logpasses by position
                                (Bonferroni)
Row Mean-|
Col Mean |   Defender    Forward
---------+----------------------
 Forward |   -.580234
         |      0.000
         |
Midfield |    .013295    .593529
         |      1.000      0.000
```

- Should any other tests be conducted? Explain.

- How many pair-wise tests do you have had to perform to determine where the difference is?

- What is the new Type I error rate after adjusting for the multiple tests?

- Note that when presenting the p-values for the Bonferroni tests, STATA scales the values such that you should compare them to_____ and not _____. What do we conclude?

- Generate the table of contrasts (i.e. differences between pairs of means) and Bonferroni corrected p-values and 95% confidence intervals. Present and interpret the 95% Bonferroni corrected confidence interval for the difference between the mean logpasses for Forward vs Defender positions.

```
REFERENCE:
encode position, gen(posgp)   <- note this command ensures the variable 'position' is a
                                 categorical, not a string variable

pwmean [variable of interest], over(group variable) mcompare(bonferroni) effects
```

Stata Tips for ANOVA

Unlike the one- and two-sample hypothesis tests for the mean, STATA requires you to have the entire dataset and not just the summary statistics to perform ANOVA.

From the *Statistics* menu, choose *linear models and related*/*ANOVA* then choose *One-way ANOVA*. In the resulting window, choose the response variable and the factor (group) variable. If you're doing a multiple comparison procedure, check the appropriate box.