Cheat Sheet: Analysis of Variance

Essential Concepts

- The purpose of a one-way ANOVA (Analysis of Variance) test is to determine the existence of a statistically significant difference among several group means.
- The null hypothesis for a one-way ANOVA states that all the group/population means are the same. This can be written as: $H_0 : \mu_1 = \mu_2 = ... = \mu_k$where k is the number of independent groups or samples.
- The alternative hypothesis for a one-way ANOVA should be written as: H_A : At least two of the group means are different.
- The statistic measuring the variation within the groups is the error sum of squares. This is calculated by summing the variation within each of the groups. The variation within each of the groups is visualized in the boxplot by the size of the box and in the dotplot as the spread of the dots within each group.
- A statistic measuring the variation between the groups is the group sum of squares. This is calculated by summing the variation between each of the group means and the grand mean (i.e., the mean of all the data values).
- In order to perform a one-way ANOVA test, there are four basic assumptions to be fulfilled:
    1. All samples are randomly selected and independent.
    2. An ANOVA also requires that the data within each group be normally distributed, but testing for that is outside the scope of this course.
    3. The populations are assumed to have equal standard deviations (or variances).
    4. Data for ANOVA
        - The factor is a categorical variable.
        - The response is a numerical variable.
        - The mean of the response variable is the parameter of interest.

- Steps to conduct a one-way ANOVA hypothesis test

    1. Set up the null and alternative hypothesis.
    2. Check the conditions/assumptions for the ANOVA hypothesis test.
        - The right types of data—factor of interest should be categorical and response variable should be numeric and continuous
        - Similar levels of variability
        - Randomly assigned, independent groups
    3. Calculate the F-statistic (See ANOVA Table below).
    4. Calculate the P-value.
    5. Compare the P-value to the significance level, α, to make a decision.
    6. Write a conclusion in context (e.g., we do/do not have convincing evidence...)

- ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Statistic
Group	$k-1$ (The number of groups minus 1)	SSGroup	$\frac{\text{SSGroup}}{k-1}$	$\frac{\text{MSGroup}}{\text{MSError}}$
Error	$N-k$ (The total number of data points minus the number of groups)	SSError	$\frac{\text{SSError}}{N-k}$	
Total	$N-1$ (The total number of data points minus 1)	SSGroup + SSError		

- The pair-wise comparison for ANOVA is a process of analyzing groups/populations by comparing them against each other in pairs. When conducting pair-wise comparison for ANOVA, we will be conducting multiple two-sample tests in order to find the significant difference(s) among the means.

- The probability of committing a type I error is equal to the significance level: $P(\text{Type I Error}) = \alpha$.

- We need a method to maintain an overall level of significance even when several tests are performed. We call this the family-wise error rate. The family-wise error rate is defined as the probability of rejecting at least one of the true null hypotheses.Suppose we perform m independent hypothesis tests.The probability of making a type I error (at least one false rejection) is: $1-(1-\alpha)^m$.
- One method for controlling for a family-wise error rate is the Tukey method for all pair-wise comparisons (formally Tukey-Kramer method). This method adjusts the length of the confidence interval (to ensure an overall level of confidence) and the P-value (to ensure an overall significance level for all pair-wise comparisons).

Key Equations

F-Statistic

$$\frac{\text{MSGroup}}{\text{MSError}} = \frac{\text{Variation BETWEEN groups}}{\text{Variation WITHIN groups}}$$

Mean Square for Error (MSError)

$$\frac{\text{Error sum of squares}}{\text{degrees of freedom (Error)}} = \frac{SSE}{N-k}$$

Mean Square for Group (MSGroup)

$$\frac{\text{Group sum of squares}}{\text{degrees of freedom (Group)}} = \frac{SSG}{k-1}$$

Total Sum of Squares (SSTotal)

$$\text{SSTotal} = \text{SSGroup} + \text{SSError}$$

Glossary

data fishing/data snooping

only showing the comparisons you want to show based on the boxplot

error sum of squares

the statistic measuring the variation within the groups

family-wise error rate

the probability of rejecting at least one of the true null hypotheses

group sum of squares

a statistic measuring the variation between the groups

one-way ANOVA

a statistical test for comparing and making inferences about means associated with two or more groups