Overview of Statistical Methods / ANOVA

BUS 735: Business Decision Making and Research

Goals

- Specific goals:
  - Re-familiarize ourselves with statistical tests.
  - Learn how to choose appropriate tests.
  - Learn how to compare means or medians among more than two populations.
- Learning objectives:
  - LO1: Be able to construct and test hypotheses using a variety of bivariate statistical methods to compare characteristics between two populations.
  - LO3: Be able to construct and use analysis of variance and analysis of covariance models to construct and test hypotheses considering complex relationships among multiple variables.
  - LO6: Be able to use standard computer packages such as SPSS and Excel to conduct the quantitative analyses described in the learning objectives above.

Goals

- Specific goals:
  - Re-familiarize ourselves with statistical tests.
  - Learn how to choose appropriate tests.
  - Learn how to compare means or medians among more than two populations.
- Learning objectives:
  - LO1: Be able to construct and test hypotheses using a variety of bivariate statistical methods to compare characteristics between two populations.
  - LO3: Be able to construct and use analysis of variance and analysis of covariance models to construct and test hypotheses considering complex relationships among multiple variables.
  - LO6: Be able to use standard computer packages such as SPSS and Excel to conduct the quantitative analyses described in the learning objectives above.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or..*
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or...*
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Selecting Right Method

- Parametric Methods:
  - Only for *interval or ratio data*.
  - Make sure assumptions of CLT hold:
    - Large sample size *or*..
    - Normal distributed *population*.
- Non-parametric methods using ranks
  - Ordinal data *and/or*...
  - Central limit theorem does not apply.
- Non-parametric Chi-squared test
  - Can be used for categorical data.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Single Population

- Examine a proportion
  - Parametric: treat data as 0s and 1s, T-test for a single mean.
  - Nonparametric: Binomial distribution.
- Examine the "average" (measure of center) of a single population.
  - Parametric method: T-test for a single mean.
  - Nonparametric methods: Test proportion of data at or below hypothesized median less than 50%.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Differences in Two Populations

- Independent Samples
  - Parametric: T-test for difference in means.
  - Nonparametric: Mann-Whitney U-Test - tests whether two populations are drawn from same distribution.
- Paired samples (Dependent Samples)
  - Parametric: Paired samples T-Test
  - Nonparametric: Wilcoxon signed rank test.

Relationships Between Two Variables

- Parametric method: Pearson linear correlation coefficient.
- Nonparametric method: Spearman correlation.
- One variable categorical: Chi-squared test of independence.

Relationships Between Two Variables

- Parametric method: Pearson linear correlation coefficient.
- Nonparametric method: Spearman correlation.
- One variable categorical: Chi-squared test of independence.

Relationships Between Two Variables

- Parametric method: Pearson linear correlation coefficient.
- Nonparametric method: Spearman correlation.
- One variable categorical: Chi-squared test of independence.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

Differences in More than Two Populations

- Parametric method: Analysis of Variance (ANOVA)
  - Compares the means of two or more populations.
  - Null hypothesis: all populations have the same mean.
  - Alternative hypothesis: at least one population has a mean different than the others.
- Nonparametric method:
  - Kruskal-Wallis test.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

One-Way ANOVA

- Method for testing for significant differences among means from two or more groups.
- Essentially an extension of the t-test for testing the differences between two means.
- Uses measures of *variance* to measure for differences in *means*.
- Total variation in your data is decomposed into two components:
  - **Among-group variation**: variability that is due to differences among groups, also called *explained* variation.
  - **Within-group variation**: total variability within each of the groups, this is unexplained variation.

Hypothesis Test

- Null hypothesis: $\mu_1 = \mu_2 = ... = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test compares whether among-group variation is greater than within-group variation.

Hypothesis Test

- Null hypothesis: $\mu_1 = \mu_2 = ... = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test compares whether among-group variation is greater than within-group variation.

Hypothesis Test

- Null hypothesis: $\mu_1 = \mu_2 = ... = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test compares whether among-group variation is greater than within-group variation.

Assumptions behind One-way ANOVA F-test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Sufficiently large (?) sample size, or else population must have a normal distribution.
- Homogeneity of variance: the variances of each of the K groups must be equal ($\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$).
  - Levene test for homogeneity of variance can be used to test for this.

Assumptions behind One-way ANOVA F-test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Sufficiently large (?) sample size, or else population must have a normal distribution.
- Homogeneity of variance: the variances of each of the K groups must be equal ($\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$).
  - Levene test for homogeneity of variance can be used to test for this.

Assumptions behind One-way ANOVA F-test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Sufficiently large (?) sample size, or else population must have a normal distribution.
- Homogeneity of variance: the variances of each of the K groups must be equal ($\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$).
  - Levene test for homogeneity of variance can be used to test for this.

Assumptions behind One-way ANOVA F-test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Sufficiently large (?) sample size, or else population must have a normal distribution.
- Homogeneity of variance: the variances of each of the K groups must be equal ($\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$).
  - Levene test for homogeneity of variance can be used to test for this.

Assumptions behind One-way ANOVA F-test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Sufficiently large (?) sample size, or else population must have a normal distribution.
- Homogeneity of variance: the variances of each of the K groups must be equal ($\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$).
    - Levene test for homogeneity of variance can be used to test for this.

Example: Crime Rates

- Data on 47 states from 1960 (I know its old) on the crime rate and a number of factors that may influence the crime rate.
- In particular, I made a variable that put unemployment into categories:
  - Unemployment = 1 if unemployment rate was less than 8%.
  - Unemployment = 2 if unemployment rate was between 8 and 10%.
  - Unemployment = 3 if unemployment rate was greater than 10%.
- I also made a variable that categorized schooling:
  - Schooling = 1 if mean years of schooling for given state was less than 10 years.
  - Schooling = 2 otherwise.
- Is there statistical evidence that the mean crime rate is different among the different categories for the level of unemployment?

Example: Crime Rates

- Data on 47 states from 1960 (I know its old) on the crime rate and a number of factors that may influence the crime rate.
- In particular, I made a variable that put unemployment into categories:
  - Unemployment = 1 if unemployment rate was less than 8%.
  - Unemployment = 2 if unemployment rate was between 8 and 10%.
  - Unemployment = 3 if unemployment rate was greater than 10%.
- I also made a variable that categorized schooling:
  - Schooling = 1 if mean years of schooling for given state was less than 10 years.
  - Schooling = 2 otherwise.
- Is there statistical evidence that the mean crime rate is different among the different categories for the level of unemployment?

Example: Crime Rates

- Data on 47 states from 1960 (I know its old) on the crime rate and a number of factors that may influence the crime rate.
- In particular, I made a variable that put unemployment into categories:
  - Unemployment = 1 if unemployment rate was less than 8%.
  - Unemployment = 2 if unemployment rate was between 8 and 10%.
  - Unemployment = 3 if unemployment rate was greater than 10%.
- I also made a variable that categorized schooling:
  - Schooling = 1 if mean years of schooling for given state was less than 10 years.
  - Schooling = 2 otherwise.
- Is there statistical evidence that the mean crime rate is different among the different categories for the level of unemployment?

Example: Crime Rates

- Data on 47 states from 1960 (I know its old) on the crime rate and a number of factors that may influence the crime rate.
- In particular, I made a variable that put unemployment into categories:
  - Unemployment = 1 if unemployment rate was less than 8%.
  - Unemployment = 2 if unemployment rate was between 8 and 10%.
  - Unemployment = 3 if unemployment rate was greater than 10%.
- I also made a variable that categorized schooling:
  - Schooling = 1 if mean years of schooling for given state was less than 10 years.
  - Schooling = 2 otherwise.
- Is there statistical evidence that the mean crime rate is different among the different categories for the level of unemployment?

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Nonparametric One-way ANOVA

- Kruskal-Wallis Rank Test: non-parametric technique for testing for differences in the *medians* among two or more groups.
- Like the Mann-Whitney U-test, uses information about the ranks of the observations, instead of the actual sizes.
- Null hypothesis: $\theta_1 = \theta_2 = ... = \theta_K$ (i.e. all groups have the same median).
- Alternative hypothesis: at least one of the medians differ.
- As the sample size gets large (over 5 per group some say!), the Kruskal-Wallis test statistic approaches a χ^2 distribution with $K - 1$ degrees of freedom.
- For small sample sizes: possible to compute exact p-values without depending on asymptotic distributions.

Assumptions for Kruskal-Wallis Test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Only the location (i.e. the center) of the distributions differ among the groups. The populations otherwise have the same distribution.

Assumptions for Kruskal-Wallis Test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Only the location (i.e. the center) of the distributions differ among the groups. The populations otherwise have the same distribution.

Assumptions for Kruskal-Wallis Test

- Randomness: individual observations are assigned to groups *randomly*.
- Independence: individuals in each group are independent from individuals in another group.
- Only the location (i.e. the center) of the distributions differ among the groups. The populations otherwise have the same distribution.