Overview of Statistical Methods / ANOVA

BUS 735: Business Decision Making and Research

1

1.1 Goals

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.

2 Selecting the Right Method

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.

2.1 Single Population

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%.

2.2 Differences in Two Populations

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.

2.3 Relationships Between Two Variables

Relationships Between Two Variables

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

2.4 Differences in More than Two Populations

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.

3 Analysis of Variance

3.1 Variance Decomposition

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.

3.2 Parametric Test

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.

3.3 Example Using SPSS

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?

FYI: Explanation of all the variables

- R: Crime rate: # of offenses reported to police per million population
- Age: The number of males of age 14-24 per 1000 population
- S: Indicator variable for Southern states (0 = No, 1 = Yes)
- Ed: Mean # of years of schooling x 10 for persons of age 25 or older
- Ex0: 1960 per capita expenditure on police by state and local government
- Ex1: 1959 per capita expenditure on police by state and local government
- LF: Labor force participation rate per 1000 civilian urban males age 14-24
- M: The number of males per 1000 females
- N: State population size in hundred thousands

- NW: The number of non-whites per 1000 population
- U1: Unemployment rate of urban males per 1000 of age 14-24
- U2: Unemployment rate of urban males per 1000 of age 35-39
- W: Median value of transferable goods and assets or family income in tens of $
- X: The number of families per 1000 earning below 1/2 the median income

Using SPSS to Conduct One-way ANOVA Tests

1. Download and open the dataset `crime.sav` in SPSS.
2. Click on `Analyze` menu, then `Compare Means`, then select `One-Way ANOVA`.
3. Move `Crime rate` to the `Dependent List`.
4. Move `Unemployment` to `Factor`.
5. For extra tests:
   - Click on `Post-hoc` button for tests to compare pair-wise differences in the means.
   - Click on `Options` button for descriptive statistics for for homogeneity of variance test.

One-way ANOVA output

1. Descriptive Statistics: shows the mean unemployment rate for each of the three groups, also includes standard deviation, standard error, and confidence intervals. It's nice to present such statistics in your papers.
2. Levene's Test of Homogeneity of Variances. The null hypothesis is that the variances are equal.
3. ANOVA Table: presents the sum of squares, the mean sum of squares, the F-statistic, and the p-value.
4. Tukey Tests for all pairwise comparisons.

4 Kruskal-Wallis Test: Nonparametric Test

4.1 Nonparametric "ANOVA"

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.

4.2 Assumptions

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.

4.3 Example Using SPSS

Using SPSS to Conduct Kruskal-Wallis Test

1. Click on `Analyze` menu, then `Nonparametric Tests`, then select `K-Independent Samples`.
2. Move `Crime rate` to `Test Variable List`.
3. Move `Unemployment` to `Grouping Variable`.
4. Make sure Kruskal-Wallis H text box is selected.
5. Click on `Exact` button if you need exact p-values.
6. Click OK!
7. Results show average ranks for each group and χ^2 test statistic and p-values.