Analysis of Variance (ANOVA)

MGMT 662: Integrative Research Project

August 7, 2008.

1

1.1 Goals

Goals of this class meeting

- Learn how to test for significant differences in means from two or more groups.
- Learn how to account for an additional factor.
- Learn how to test for significant differences in medians from two or more groups. Why?

2 One-Way ANOVA

2.1 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.

Variance Decomposition

- **Sum of squares groups (SSG)**:

$$SSG = \sum_{k=1}^{K} n_k \left(\bar{x}_k - \bar{x}\right)^2$$

  - K: number of groups.
  - $\bar{x}_k$ is the mean of group k.
  - $\bar{x}$ is the mean of all the data.

- **Sum of squares within-group (SSW)**:

$$SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_{i,k} - \bar{x}_k\right)^2$$

- **Sum of squares total (SST)**:

$$SST = SSG + SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_{i,k} - \bar{x}\right)^2$$

2.2 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 (has an F-distribution with degrees of freedom $K-1$, $n-1$):

$$F = \frac{SSG/(K-1)}{SSW/(n-1)}$$

- Intuitively, what is implied when the F-statistic is large?

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?

Using SPSS to Conduct One-way ANOVA Tests

1. Download and open the dataset `crime.xls` 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.

2.3 Nonparametric Test

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.

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.

3 Two-way ANOVA

3.1 Variance Decomposition

Two-way ANOVA

- One-way ANOVA, the effects of one factor where examined.
- **Two-way ANOVA**, also called **two-factor factorial design**: two factors are simultaneously evaluated.
- Total variance is decomposed into:
  - variability explained by being in different groups of factor A.
  - variability explained by being in different groups of factor B.
  - variability explained by the interaction of factors A and B.
  - unexplained variability.

ANOVA Descriptive Statistics

	Unemployment Level		
Schooling Level	Less than 8%	8% to 10%	More than 10%
Less than 10 years	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{13}$
10 years or more	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{23}$

- Goals of ANOVA:
  - Determine if schooling level (factor A) leads to different levels for crime rate.
  - Determine if unemployment level (factor B) leads to different levels for crime rate.
  - Determine if schooling and unemployment have a joint effect on crime rate (i.e. does the unemployment level effect the impact of schooling on the crime rate, or vice versa).

3.2 Parametric Tests

Hypothesis Test for Factor A

- $H_0 : \mu_{1.} = \mu_{2.} = ... = \mu_{r.}$
- H_a : At least on of the means of the groups in factor A are different from the others.

$$F = \frac{SSA/(r-1)}{SSE/(N-rc)}$$

- F-statistic has degrees of freedom $r-1$, $N-rc$.
- $N = \sum_{i=1}^{r} n_{i.}$.
- r (number of rows) is the number of groups for factor A.
- c (number of columns) is the number of groups for factor B.
- $\mu_{i.}$ is the mean of group i of factor A.
- SSA is the sum of squares from factor A.

Hypothesis Test for Factor B

- $H_0 : \mu_{.1} = \mu_{.2} = ... = \mu_{.c}$
- H_a : At least on of the means of the groups in factor B are different from the others.

$$F = \frac{SSB/(c-1)}{SSE/(N-rc)}$$

- F-statistic has degrees of freedom $c-1$, $N-rc$.
- $\mu_{j.}$ is the mean of group j of factor B.
- SSB is the sum of squares from factor B.

Hypothesis Test for Interaction of Factors A and B

- H_0 : there is no interaction effect.
- H_a : there is an interaction effect.

$$F = \frac{SSAB/(r-1)(c-1)}{SSE/(N-rc)}$$

- F-statistic has degrees of freedom $(r-1)(c-1)$, $N-rc$.
- SSAB is the sum of squares from factors A and B.

Using SPSS to Conduct Two-Way ANOVA

1. Click on `Analyze`, then `General Linear Model`, then `Univariate`.
2. Move `Crime Rate` to the dependent variable.
3. Move `Unemployment` and `Smoothing` to the fixed factor list.
4. Click on `Options` and select `Descriptive Statistics` check box. Click Continue.

5. Click OK!

Results show:

- Sample sizes for each pairwise group.
- Descriptive statistics for each pairwise group.
- Analysis of Variance Table:
  - Variance decompositions.
  - F-test for factor unemployment.
  - F-test for factor schooling.
  - F-test for interaction effects.

3.3 Nonparametric Two-way ANOVA

Nonparametric Two-Way ANOVA

- Advise: run two, one-way ANOVA's using Kruskal-Wallis test.
- Or.. run a Kruskal Wallis test with r x c different groups:
  - Group 1: Unemployment=1, Schooling=1
  - Group 2: Unemployment=2, Schooling=1
  - Group 3: Unemployment=3, Schooling=1
  - Group 4: Unemployment=1, Schooling=2
  - Group 5: Unemployment=2, Schooling=2
  - Group 6: Unemployment=3, Schooling=2
- The Tukey pairwise tests for these combinations will indicate if there are interaction effects.