

# Analysis of Variance

Mgmt 230: Introductory Statistics

## Goals of this section

- Learn how to detect differences among two or more means.
- Learn how to use measures of variability to detect these differences.

# One-Way ANOVA

3 / 10

- 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

3 / 10

- 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

3 / 10

- 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.

## Variance Decomposition: Explained Sum of Squares 4 / 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Explained Sum of Squares 4 / 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Explained Sum of Squares 4 / 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Explained Sum of Squares 4 / 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Explained Sum of Squares 4 / 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Explained Sum of Squares 4/ 10

**Sum of squares groups (SSG):** Measure of the variability of the variable that is explained by being in a particular group.

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

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

## Variance Decomposition: Unexplained Sum Squares 5 / 10

**Sum of squares within-groups (SSW):** Measure of the variability within the groups.

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

- $x_{i,k}$ :  $i$ th observation in group  $k$ .

## Variance Decomposition: Unexplained Sum Squares 5 / 10

**Sum of squares within-groups (SSW):** Measure of the variability within the groups.

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

- $x_{i,k}$ :  $i$ th observation in group  $k$ .

## Variance Decomposition: Unexplained Sum Squares 5 / 10

**Sum of squares within-groups (SSW):** Measure of the variability within the groups.

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

- $x_{i,k}$ :  $i$ th observation in group  $k$ .

## Variance Decomposition: Total Sum of Squares

**Sum of squares total (SST):** Total measure of variability. Does the formula look somewhat familiar?

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

## Variance Decomposition: Total Sum of Squares

**Sum of squares total (SST):** Total measure of variability. Does the formula look somewhat familiar?

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

## Variance Decomposition: Mean Measures

- **Mean Squares Groups (MSG)**: Measure of the average variability that is explained by being in a particular group.

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2}{K - 1}$$

- **Mean Squares Within-Groups (MSW)**: Measure of the average variability of the data within the groups.

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

- No textbook ever mentions a **Mean Squares Total**. What would it be?

## Variance Decomposition: Mean Measures

- **Mean Squares Groups (MSG)**: Measure of the average variability that is explained by being in a particular group.

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2}{K - 1}$$

- **Mean Squares Within-Groups (MSW)**: Measure of the average variability of the data within the groups.

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

- No textbook ever mentions a **Mean Squares Total**. What would it be?

## Variance Decomposition: Mean Measures

- **Mean Squares Groups (MSG)**: Measure of the average variability that is explained by being in a particular group.

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2}{K - 1}$$

- **Mean Squares Within-Groups (MSW)**: Measure of the average variability of the data within the groups.

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

- No textbook ever mentions a **Mean Squares Total**. What would it be?

## Variance Decomposition: Mean Measures

- **Mean Squares Groups (MSG)**: Measure of the average variability that is explained by being in a particular group.

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2}{K - 1}$$

- **Mean Squares Within-Groups (MSW)**: Measure of the average variability of the data within the groups.

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

- No textbook ever mentions a **Mean Squares Total**. What would it be?

## Variance Decomposition: Mean Measures

- **Mean Squares Groups (MSG)**: Measure of the average variability that is explained by being in a particular group.

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2}{K - 1}$$

- **Mean Squares Within-Groups (MSW)**: Measure of the average variability of the data within the groups.

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

- No textbook ever mentions a **Mean Squares Total**. What would it be?

# Hypothesis Test

- Null hypothesis:  $\mu_1 = \mu_2 = \dots = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test (has an F-distribution with degrees of freedom  $K - 1$ ,  $n - 1$ ):

$$F = \frac{MSG}{MSW}$$

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

# Hypothesis Test

- Null hypothesis:  $\mu_1 = \mu_2 = \dots = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test (has an F-distribution with degrees of freedom  $K - 1$ ,  $n - 1$ ):

$$F = \frac{MSG}{MSW}$$

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

# Hypothesis Test

- Null hypothesis:  $\mu_1 = \mu_2 = \dots = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test (has an F-distribution with degrees of freedom  $K - 1$ ,  $n - 1$ ):

$$F = \frac{MSG}{MSW}$$

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

# Hypothesis Test

- Null hypothesis:  $\mu_1 = \mu_2 = \dots = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test (has an F-distribution with degrees of freedom  $K - 1$ ,  $n - 1$ ):

$$F = \frac{MSG}{MSW}$$

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

# Hypothesis Test

- Null hypothesis:  $\mu_1 = \mu_2 = \dots = \mu_K$
- Alternative hypothesis: At least one of the means are different from the others.
- F-test (has an F-distribution with degrees of freedom  $K - 1$ ,  $n - 1$ ):

$$F = \frac{MSG}{MSW}$$

- 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 ( $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$ ).

## 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 ( $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$ ).

## 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 ( $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$ ).

## 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 ( $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$ ).

## 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?