STAT 423/523 Statistical Methods for Engineers and Scientists

Lecture 6: The Analysis of Variance

Chencheng Cai

Washington State University

Analysis of Variance

A **factor** is a qualitative variable that defines the groups to be compared.

The **levels** of a factor are the distinct values of the factor.

Analysis of Variance

A **factor** is a qualitative variable that defines the groups to be compared.

The **levels** of a factor are the distinct values of the factor.

Examples: (factors highlighted)

- An experiment to study the effects of five different **brands** of gasoline on automobile engine operating efficiency (mpg).
- An experiment to study the effects of the presence of four different **sugar solutions** (glucose, sucrose, fructose, and a mixture of the three) on bacterial growth.

Analysis of Variance

A **factor** is a qualitative variable that defines the groups to be compared.

The **levels** of a factor are the distinct values of the factor.

Examples: (factors highlighted)

- An experiment to study the effects of five different **brands** of gasoline on automobile engine operating efficiency (mpg).
- An experiment to study the effects of the presence of four different **sugar solutions** (glucose, sucrose, fructose, and a mixture of the three) on bacterial growth.
- An experiment to investigate whether **hardwood concentration in pulp** (%) at three different levels impacts tensile strength of bags made from the pulp.
- An experiment to decide whether the color density of fabric specimens depends on which of four different **dye amounts** is used

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method used to compare the subpopulations of a factor.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method used to compare the subpopulations of a factor.

- If there is one factor, it is called **one-way ANOVA** or **single-factor ANOVA**.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method used to compare the subpopulations of a factor.

- If there is one factor, it is called **one-way ANOVA** or **single-factor ANOVA**.
- If there is one factor with two levels, the ANOVA should be similar to a two-sample test.
- All examples in the previous slide are one-way ANOVA.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method used to compare the subpopulations of a factor.

- If there is one factor, it is called **one-way ANOVA** or **single-factor ANOVA**.
- If there is one factor with two levels, the ANOVA should be similar to a two-sample test.
- All examples in the previous slide are one-way ANOVA.
- If there are two (or more) factors, it is called **two-way ANOVA** (or **multi-factor ANOVA**).

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method used to compare the subpopulations of a factor.

- If there is one factor, it is called **one-way ANOVA** or **single-factor ANOVA**.
- If there is one factor with two levels, the ANOVA should be similar to a two-sample test.
- All examples in the previous slide are one-way ANOVA.
- If there are two (or more) factors, it is called **two-way ANOVA** (or **multi-factor ANOVA**).
- Example of two-way ANOVA:
  An experiment to study the effects of two factors, **temperature** and **humidity**, on the growth of a certain type of bacteria.

One-way ANOVA — Notations

- I: the number of levels of the factor.
- $\mu_i, i = 1, \ldots, I$: the population mean of the ith level of the factor.

One-way ANOVA — Notations

- I: the number of levels of the factor.
- $\mu_i, i = 1, \ldots, I$: the population mean of the ith level of the factor.
- The relevant hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$
$$H_a : \text{At least one of the means is different}$$

One-way ANOVA — Notations

- I: the number of levels of the factor.
- $\mu_i, i = 1, \ldots, I$: the population mean of the ith level of the factor.
- The relevant hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$
$$H_a : \text{At least one of the means is different}$$

- In experimental design, the ith level of the factor is often called a **treatment**.

One-way ANOVA — Notations

- I: the number of levels of the factor.
- $\mu_i, i = 1, \ldots, I$: the population mean of the ith level of the factor.
- The relevant hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$
$$H_a : \text{At least one of the means is different}$$

- In experimental design, the ith level of the factor is often called a **treatment**.
- X_{ij} is the jth observation in the ith treatment.
- x_{ij} is the value of X_{ij} when the experiment is conducted.

Example

Compress strength of different types of boxes.

Type of Box	Compression Strength (lb)						Sample Mean	Sample SD
1	655.5	788.3	734.3	721.4	679.1	699.4	713.00	46.55
2	789.2	772.5	786.9	686.1	732.1	774.8	756.93	40.34
3	737.1	639.0	696.3	671.7	717.2	727.1	698.07	37.20
4	535.1	628.7	542.4	559.0	586.9	520.0	562.02	39.87
				Grand mean =			682.50	

Different Means

Let X_{ij} be the j-th observation in the i-th treatment.

Suppose each treatment level has J observations. Then the total number of observations is $I \times J$.

Different Means

Let X_{ij} be the j-th observation in the i-th treatment.

Suppose each treatment level has J observations. Then the total number of observations is $I \times J$.

- The sample mean of the i-th treatment is

$$\bar{X}_{i\cdot} = \frac{1}{J} \sum_{j=1}^{J} X_{ij}$$

Different Means

Let X_{ij} be the j-th observation in the i-th treatment.

Suppose each treatment level has J observations. Then the total number of observations is $I \times J$.

- The sample mean of the i-th treatment is

$$\bar{X}_{i\cdot} = \frac{1}{J} \sum_{j=1}^{J} X_{ij}$$

- The sample mean of all observations (**grand mean**) is

$$\bar{X}_{\cdot\cdot} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}$$

Different Means

Let X_{ij} be the j-th observation in the i-th treatment.

Suppose each treatment level has J observations. Then the total number of observations is $I \times J$.

- The sample mean of the i-th treatment is

$$\bar{X}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$$

- The sample mean of all observations (**grand mean**) is

$$\bar{X}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$$

- Q: what if we have unequal number of observations in each treatment?

Sum of Squares

- The **Sum of Squares Error** (SSE) is

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2$$

Sum of Squares

- The **Sum of Squares Error** (SSE) is

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2$$

- The **Sum of Squares Treatment** (SSTr) is

$$SSTr = J \sum_{i=1}^{I} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

Sum of Squares

- The **Sum of Squares Error** (SSE) is

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i.})^2$$

- The **Sum of Squares Treatment** (SSTr) is

$$SSTr = J\sum_{i=1}^{I}(\bar{X}_{i.} - \bar{X}_{..})^2$$

- The **Sum of Squares Total** (SST or SSTo) is

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{..})^2$$

Sum of Squares

The relationship between SST, SSTr, and SSE:

$$SST = SSTr + SSE$$

Sum of Squares

The relationship between SST, SSTr, and SSE:

$$SST = SSTr + SSE$$

Q: what if we have unequal number of observations in each treatment?

Mean Sqaures

The mean sqaures are the sum of squares divided by the degrees of freedom.

$$MSX = \frac{SSX}{\text{degrees of freedom}}$$

Mean Sqaures

The mean sqaures are the sum of squares divided by the degrees of freedom.

$$MSX = \frac{SSX}{\text{degrees of freedom}}$$

The degrees of freedom (df) can be calculated as

$$\text{df} = \text{number of observations} - \text{number of parameters}$$

Mean Sqaures

- The **Mean Square Error** (MSE) is

$$MSE = \frac{SSE}{IJ - I} = \frac{1}{IJ - I} \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2$$

Mean Sqaures

- The **Mean Square Error** (MSE) is

$$MSE = \frac{SSE}{IJ - I} = \frac{1}{IJ - I} \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2$$

- The **Mean Square Treatment** (MSTr) is

$$MSTr = \frac{SSTr}{I - 1} = \frac{J}{I - 1} \sum_{i=1}^{I} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

Nested Models

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{v.s.} \quad H_a : \text{not all equal}$$

- The **full model** is the model with all the treatment means different. (i.e. $H_0 \cup H_a$)
  The estimators are
  $$\hat{\mu}_i = \bar{X}_{i\cdot} \quad \text{for } i = 1, \ldots, I.$$

Nested Models

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{v.s.} \quad H_a : \text{not all equal}$$

- The **full model** is the model with all the treatment means different. (i.e. $H_0 \cup H_a$)
  The estimators are
  $$\hat{\mu}_i = \bar{X}_{i\cdot} \quad \text{for } i = 1, \ldots, I.$$
- The **reduced model** is the model with all the treatment means equal. (i.e. H_0)
  The estimator is
  $$\hat{\mu}_i = \hat{\mu} = \bar{X}_{\cdot\cdot} \quad \text{for } i = 1, \ldots, I.$$

Nested Models

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{v.s.} \quad H_a : \text{not all equal}$$

- The **full model** is the model with all the treatment means different. (i.e. $H_0 \cup H_a$)
  The estimators are
  $$\hat{\mu}_i = \bar{X}_{i\cdot} \quad \text{for } i = 1, \ldots, I.$$
- The **reduced model** is the model with all the treatment means equal. (i.e. H_0)
  The estimator is
  $$\hat{\mu}_i = \hat{\mu} = \bar{X}_{\cdot\cdot} \quad \text{for } i = 1, \ldots, I.$$
- The two models are **nested** because the reduced model is a special case of the full model.

Nested Models

The sum of squared error for the full model is

$$SSE_{full} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot})^2 = SSE$$

Nested Models

The sum of squared error for the full model is

$$SSE_{full} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot})^2 = SSE$$

The sum of squared error for the reduced model is

$$SSE_{reduced} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = SST$$

Nested Models

The sum of squared error for the full model is

$$SSE_{full} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot})^2 = SSE$$

The sum of squared error for the reduced model is

$$SSE_{reduced} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = SST$$

- The **extra sum of squares** is

$$SSE_{reduced} - SSE_{full} = SST - SSE = SSTr$$

Nested Models

The sum of squared error for the full model is

$$SSE_{full} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot})^2 = SSE$$

The sum of squared error for the reduced model is

$$SSE_{reduced} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = SST$$

- The **extra sum of squares** is

$$SSE_{reduced} - SSE_{full} = SST - SSE = SSTr$$

- The full model uses $I - 1$ more parameters than the reduced model.
- The full model improves the fit by $SSTr$.

Nested Models

For the full model:

- Sum of squared error is SSE with $IJ - I$ degrees of freedom.
- Sum of squares fitted is $SSTr$ with $I - 1$ degrees of freedom.
- Sum of squares total is SST with $IJ - 1$ degrees of freedom.

For the reduced model:

- Sum of squared error is SST with $IJ - 1$ degrees of freedom.
- Sum of squares fitted is 0 with 0 degrees of freedom.
- Sum of squares total is SST with $IJ - 1$ degrees of freedom.