Lecture 18

Part 4 Analysis of Variance

Analysis of Variance

Which Tutor is the Best?

- Suppose there are four tutors for a course: Mandy, Thomas, Jasmine and Bob.
- Is there a difference in the teaching abilities of the four tutors?
- Let μ_j denote the population mean grade for students in the jth tutor's class.

Which Tutor is the Best?

- What we are trying to test is the following hypotheses:

$$H_0 : \mu_M = \mu_T = \mu_J = \mu_B$$

$$H_1 : \begin{cases} \text{Not all population means are equal.} \\ \text{At least two population means differ.} \\ H_0 \text{ is false.} \end{cases}$$

- The above three statements for H_1 are *equivalent*.
- Examine the average grade in each class and compare them.

Which Tutor is the Best?

- What if we find that:

Tutor/Class	Average Grade
Mandy	$\bar{Y}_M = 70.0\%$
Thomas	$\bar{Y}_T = 70.2\%$
Jasmine	$\bar{Y}_J = 69.9\%$
Bob	$\bar{Y}_B = 70.4\%$

Which Tutor is the Best?

- What if we find that:

Tutor/Class	Average Grade
Mandy	$\bar{Y}_M = 60.0\%$
Thomas	$\bar{Y}_T = 74.0\%$
Jasmine	$\bar{Y}_J = 67.0\%$
Bob	$\bar{Y}_B = 80.0\%$

Which Tutor is the Best?

- It's the amount of variation in the average grades *between* the classes that will help us determine whether or not we should reject H_0.
- But how much variation in average grades between classes should lead to a rejection of H_0?

Which Tutor is the Best?

- What if there is much variation in grades *within* classes?

Tutor	Grades										$\bar{Y}$
Mandy	60	70	40	50	40	69	51	80	45	95	60
Thomas	64	80	70	98	49	96	95	54	74	60	74
Jasmine	71	45	99	50	40	79	61	85	80	60	67
Bob	90	89	86	95	79	46	81	96	84	54	80

Which Tutor is the Best?

- What if there is very little variation in grades *within* classes?

Tutor	Grades										$\bar{Y}$
Mandy	58	62	60	60	60	59	61	60	58	62	60
Thomas	74	75	70	78	71	76	75	72	74	75	74
Jasmine	67	65	69	70	68	69	63	65	70	64	67
Bob	80	79	78	80	79	78	81	84	83	78	80

Which Tutor is the Best?

- Whether or not we reject H_0 depends on the variation *between* classes and the variation *within* classes.
- The variation between classes represents variation arising from the different tutors and the variation within classes represents underlying variation that is not due to the different tutors.

Analysis of Variance

- **Analysis of variance (ANOVA)** are methods that are used to test the hypothesis that the means of *two or more* populations are equal.
- The methods involve taking independent samples from each population and analysing the *amounts of variation* and the *sources of the variation*.

One-way ANOVA

- In a one-way ANOVA, we have:
  - One continuous response variable Y.
  - One categorical variable with k categories.
    - The categorical variable is also called a **factor**.
    - Each category is called a **level** and there must be at least two levels ($k \geq 2$).
- For our example:
  - The student's grade is the continuous response variable.
  - The variable *tutor* is a factor with four levels (Mandy, Thomas, Jasmine and Bob).

One-way ANOVA

- For a one-way ANOVA, each level of the factor is also called a **treatment** and represents a population.
  - For example, Mandy is a treatment and all possible students who have ever been in Mandy's tutorial define a population.
- A one-way ANOVA tests for differences in the means of k populations, where each population corresponds to a treatment (i.e., a level of the factor).

One-way ANOVA

- Said another way, a one-way ANOVA tests the importance of the factor in *explaining the variation in the response variable*, by comparing the means of the response variable between the different treatments (levels of the factor).
- For our example:
  - We want to test whether tutor is important in explaining the variation in grades, by comparing the sample mean grades of students in each tutor's class.

Sample Data

	Factor Level or Treatment				
	1	$\cdots$	j	$\cdots$	k
	Y_{11}	$\cdots$	Y_{1j}	$\cdots$	Y_{1k}
	$\vdots$		$\vdots$		$\vdots$
Sample Values	Y_{i1}	$\cdots$	Y_{ij}	$\cdots$	Y_{ik}
	$\vdots$		$\vdots$		$\vdots$
	$Y_{n_1 1}$	$\cdots$	$Y_{n_j j}$	$\cdots$	$Y_{n_k k}$
Sample Size	n_1	$\cdots$	n_j	$\cdots$	n_k
Sample Mean	$\bar{Y}_1$	$\cdots$	$\bar{Y}_j$	$\cdots$	$\bar{Y}_k$

Assumptions

- When performing a one-way ANOVA, we must make the following assumptions:

1. The levels of the factor are fixed beforehand.
    - Assumed to be true based on experimental design.
2. The response variable is normally distributed with constant variance in each treatment.
    - Can construct histograms and calculate sample variances within each sample.
3. The samples are independent.
    - Can construct certain plots, but again assumed to be true based on experimental design.

Hypotheses

- For a one-way ANOVA, we are testing:

$$
\begin{aligned}
H_0 &: \text{The population means at different levels of} \\
&\quad \text{the factor are all equal.} \\
H_1 &: \text{At least two of the population means differ.}
\end{aligned}
$$

- If H_0 is true, the sample means from each level should be similar, i.e., little variation between the sample means.
- If H_1 is true, the sample means from each level should be quite different, i.e., lots of variation between the sample means.

Sum of Squares for Treatment

- The variation between the sample means (or the variation *between the treatments*) is measured by the **sum of squares for treatment** (SST):

$$SST = \sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2$$

where n_j and $\overline{Y}_j$ are the sample size and sample mean of the jth sample, respectively, and $\overline{Y}$ is the overall sample mean of **all** observations across all samples.

Sum of Squares for Treatment

- The SST measures the variation that is *caused by* the factor.
- If the sample means are similar to each other, they will all be close to $\overline{Y}$, so the SST will be small.
- If some of the sample means differ from each other, they will be far from $\overline{Y}$, so the SST will be large.
- How large does the SST have to be before we reject the null hypothesis that all population means are equal?

Sum of Squares for Error

- To answer this question, we need to also know how much variation there is *within the treatments*, which is measured by the **sum of squares for error** (SSE):

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2$$

  where Y_{ij} is the ith observation in the jth sample.
- The SSE measures the variation that is *not* caused by the factor and can be thought of as the underlying, unexplained variation.

Total Sum of Squares

- There is one more sum of squares term called the **total sum of squares** ($SS(Total)$), which is defined to be:

$$SS(Total) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y})^2$$

- The $SS(Total)$ measures the *total* amount of variation that exists in the data.

SST, SSE and $SS(Total)$

- There is a special relationship between the sum of squares for treatment, the sum of squares for error and the total sum of squares, given by the following identity:

$$SS(Total) = SST + SSE$$

- That is, the total variation in the response variable Y in the sample is equal to the variation that is explained by the factor *plus* the left-over, unexplained variation.

Test Statistic

- Now that we have quantified these different sources of variation, we can construct a test statistic.
- We want to see how big the SST is compared to the SSE.
- But we first have to convert these sums of squares to *mean squares*, by scaling them by their appropriate *degrees of freedom*:

$$MST = \frac{SST}{k-1} \quad \text{and} \quad MSE = \frac{SSE}{n-k}$$

Test Statistic

- These mean squares are essentially sample variances and they are now more directly comparable to each other.
- So finally, the test statistic that we use is the ratio of the MST to the MSE:

$$F = \frac{MST}{MSE}$$

- It is called the F-statistic or the F-ratio.

Test Statistic

- If H_0 is true, then:
  - $\Rightarrow$ The population means are equal,
  - $\Rightarrow$ the sample means should be similar to each other,
  - $\Rightarrow$ the MST should be close to or less than the MSE,
  - $\Rightarrow$ the F-statistic should be close to or less than 1.
- If H_1 is true, then:
  - $\Rightarrow$ The population means are different,
  - $\Rightarrow$ the sample means should differ from each other,
  - $\Rightarrow$ the MST should be greater than the MSE,
  - $\Rightarrow$ the F-statistic should be greater than 1.
- So we reject H_0 when the F-statistic is too large, so this is a one-tailed test.

Decision Rule

- To determine whether or not the F-statistic is too large, we need to know its sampling distribution under H_0 (null distribution).
- Since $F = \frac{MST}{MSE}$ is a ratio of two sample variances, we compare it to an F-distribution (recall testing the equality of two population variances).
- Specifically, an F-distribution with $k - 1$ numerator degrees of freedom and $n - k$ denominator degrees of freedom.

Decision Rule

- So we need to use the F-tables.
- At a significance level of α, we reject H_0 if $F > F_{\alpha,k-1,n-k}$, where $F_{\alpha,k-1,n-k}$ is the critical value that cuts off $100\alpha\%$ in the upper tail of an F-distribution with the appropriate degrees of freedom.
- Note that if we reject H_0, we can only conclude that at least two population means are different.

ANOVA Table

Source	Sum of squares	Deg. of freedom	Mean squares	F-statistic
Factor (Treatments)	SST	$k-1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$	
Total	$SS(Total)$	$n-1$		

ANOVA Table

- If we add up all the sum of squares for each source of variation, they will equal the total sum of squares, i.e., $SST + SSE = SS(Total)$.
- The degrees of freedom for the total sum of squares is *always* equal to $n - 1$.
- If we add up all the degrees of freedom for each source of variation, they will sum to $n - 1$.
- Note that $n = \sum_{j=1}^{k} n_j$ is the total sample size.

Airline Ratings Example

- There are three airlines that fly a particular route.
- Twenty passengers from each airline were randomly selected and asked about their flight experience.
- Particularly, they were asked to rate their experience on a 0 to 100 scale using various criteria such as meal service, comfort, friendliness, etc.
- The data is displayed in the following table.

Airline Ratings Example

Airline					
1		2		3	
52	61	36	70	50	74
33	57	36	74	80	65
36	46	50	63	50	62
54	47	46	81	45	67
40	53	48	54	76	81
39	55	66	61	61	83
41	78	41	71	51	61
36	65	43	76	44	97
22	64	37	76	57	92
26	64	45	53	75	68

Airline Ratings Example

- Given this sample data, suppose you have calculated that $SS(Total) = 16011.25$ and $SST = 3446.8$.
- Using these figures, determine whether there is sufficient evidence to conclude that the population mean ratings differ between airlines.

Airline Ratings Example

- To answer this question we need to:

1. Write down H_0 and H_1.
2. Complete a one-way ANOVA table to calculate the test statistic.
3. Determine the appropriate decision rule.
4. Reach a conclusion regarding whether the population mean ratings differ between airlines.

1. Hypotheses

- Let μ_1, μ_2 and μ_3 denote the population mean ratings for the three airlines.
- The null and alternative hypotheses are:

$$H_0 : \begin{cases} \mu_1 = \mu_2 = \mu_3 \\ \text{That is, the population mean rating for} \\ \text{each airline is the same.} \end{cases}$$

H_1 : At least two population mean ratings differ.

2. ANOVA Table and Test Statistic

Source	Sum of squares	Deg. of freedom	Mean squares	F-statistic
Airline	3446.8	2	1723.4	7.8184
Error	12564.45	57	220.4289	
Total	16011.25	59		

3. Decision Rule

- We compare our test statistic to an F-distribution with $k-1=2$ numerator degrees of freedom and $n-k=57$ denominator degrees of freedom.
- At a 5% significance level, we will reject H_0 if the test statistic is larger than
$$F_{0.05,2,57} \approx F_{0.05,2,60} = 3.15.$$
- At a 1% significance level, we will reject H_0 if the test statistic is larger than
$$F_{0.01,2,57} \approx F_{0.01,2,60} = 4.98.$$

4. Conclusion

- Since our test statistic, $F = 7.8184$ is larger than 4.98, we reject H_0 at a significance level of $\alpha = 0.01$.
- That is, there is sufficient evidence to conclude that the population mean ratings differ between airlines.
- Note that we would also reject H_0 at a significance level of $\alpha = 0.05$.