Module E: Advanced Inference

Slide Deck E5:

The Analysis of Variance Procedure

The section in which we cover ANOVA, the Analysis of Variance procedure. This procedure, developed by Fisher, allows us to do three equivalent things: Test if multiple means are equal; Test if the the inclusion of an additional variable aids in understanding the dependent variable; and Test if a categorical and a numeric variable are independent.

Today's Objectives

By the end of this slidedeck, you should

1. understand the theory behind testing. . .
   - equality means of more than two populations
   - whether a categorical variable helps explain a numberic
   - independence between a numeric and a categorical variable
2. better explain the p-value and how to test hypotheses

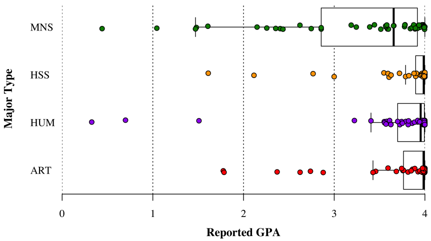
Framing Example

Example

I would like to determine if the average GPA is the same for the four types of majors: MNS, HSS, HUM, and ART. To test this, I asked 200 full-time students at Knox College, 50 of each major type, and asked two questions:

- What is your major type?
- What is your GPA?

Framing Example



The Theory

There are a few equivalent ways of looking at this question:

- Do the means in each group significantly differ?
- Are the group and the GPA independent?
- Does including the group identifier improve our ability to estimate?

The last gives some insight into the test statistic:

- Improving predictions implies we reduce uncertainty in those predictions

Think of this as the idea behind the Analysis of Variance procedure.

- Measure the variance of the original data
- Measure the variance unexplained in the model
- Calculate the ratio of the explained variance to the unexplained
- This last ratio is the test statistic

The ANOVA Table

With that background, let us calculate the test statistic (and p-value) using the following ANOVA Table:

Source	**SS**	**df**	**MS**	**F**	**p**
Model					
Error					
Total					

Let's fill it in the old-fashioned way…

The SS Column

The column marked "SS" contains the "sum of squares" for the three sources. The sum of squares is just the sum of the deviation between the observation and the mean. As such,

$$SS_{\text{Model}} = \sum_i \sum_j \left(\bar{y}_j - \bar{\bar{y}}\right)^2$$

$$SS_{\text{Error}} = \sum_i \sum_j \left(y_{ij} - \bar{y}_j\right)^2$$

$$SS_{\text{Total}} = \sum_i \sum_j \left(y_{ij} - \bar{\bar{y}}\right)^2$$

In each of these, the i represents a value within a group, and j represents a group. Also, $\bar{\bar{y}}$ is the average of all measurements (the grand mean) and $\bar{y}_j$ is the average of the measurements in group j.

The SS Column

Because each of the SS calculations require 50 sums, differences, and squares, calculation by hand is not realistic. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	7.83				
Error	92.46				
Total	100.29				

Note: $SS_M + SS_E = SS_T$. This is an interesting result. Check the formulas for SS_M and SS_E and marvel at this fact. For statisticians, this means that the Model and what remains (Error) are independent.

The df Column

The column marked "df" contains the "degrees of freedom" for the three sources. What are degrees of freedom? They are parameters that reflect the amount of information contributed by each source.*

$$\begin{aligned} df_{\text{Model}} &= k - 1 \\ df_{\text{Error}} &= k(n-1) = N - k \\ df_{\text{Total}} &= N - 1 \end{aligned}$$

In each of these, the k represents the number of groups, n represents the sample size *within each group*, and N represents the total sample size.

The df Column

Calculating the number of degrees of freedom is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	7.83	3			
Error	92.46	196			
Total	100.29	199			

Note: $df_M + df_E = df_T$.

The MS Column

The column marked "MS" contains the "mean squares" for the three sources. *These are the estimates* of the individual variances. They are the SS divided by the df for each source. Recall our Chapter 3 definition of sample variance. It is just the sum of squares divided by the degrees of freedom, $n - 1$.

$$MS_{\text{Model}} = SS_{\text{Model}}/df_{\text{Model}}$$
$$MS_{\text{Error}} = SS_{\text{Error}}/df_{\text{Error}}$$

Note that we *could* also calculate MS_{Total}. It is not used in ANOVA, so we do not. Its formula is

$$MS_{\text{Total}} = \frac{1}{N-1} \sum_i \sum_j \left(y_{ij} - \bar{\bar{y}}\right)^2$$

This is just the sample variance of the measurements.

The MS Column

Calculating the mean squares values is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	7.83	3	2.6104		
Error	92.46	196	0.4718		
Total	100.29	199			

Note: $MS_M + MS_E \neq MS_T$. That is, the total variance is *not* partitioned between the two sources.

The F Column

The column marked "F" contains the value of the "F-statistic" for the model.

$$F = \frac{MS_{\text{Model}}}{MS_{\text{Error}}}$$

As with all test statistics:

- it is a measure of how far the data are from the null hypothesis
- it has a distribution

As you should guess, the distribution of the F statistic is F — officially, it is "Snedecor's F" distribution. This distribution is named after George W. Snedecor, who used Fisher's statistical definition to calculate the probability density function. By the way, Snedecor also founded the first Statistics Department in the United States at Iowa State University in 1933.

The F Column

Calculating the F statistic is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	7.83	3	2.6104	5.533	
Error	92.46	196	0.4718		
Total	100.29	199			

The p Column

The column marked "p" contains the p-value for the model. It is interpreted as usual.

$$\text{p-value} = \mathbb{P}[\ F \geq f\ ] = 1 - \mathbb{P}[\ F \leq f\ ]$$

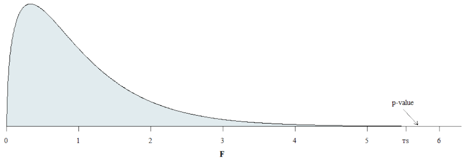
Here, f is the value of the test statistic you calculated above. As with all p-values, it is a measure of how far the data are from the null hypothesis. Compare it to your selected value of α. If the p-value is less than α, then you reject the null hypothesis.

So, what is the null hypothesis? These three are equivalent:

- All population means are the same.
- The numeric variable is independent of the categorical variable.
- The model does not significantly improve our prediction ability.

Framing Example

Here is the distribution of the test statistic and what we observed. Note that the p-value — the area to the right of the observed test statistic value — is extremely small. Thus, we would expect to reject the null hypothesis



The p-value Column

Given its definition, calculating the p-value is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	7.83	3	2.6104	5.533	0.00115
Error	92.46	196	0.4718		
Total	100.29	199			

The Brief Conclusion

Brief Conclusion:

Because the p-value of 0.00115 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. At least one of the population means differ. The two variables are not independent. The model helps in our prediction accuracy.

So, *which* population mean is different?

- ANOVA cannot tell us.
- We will need more statistics to tell us.

Example 1: Rice Yields

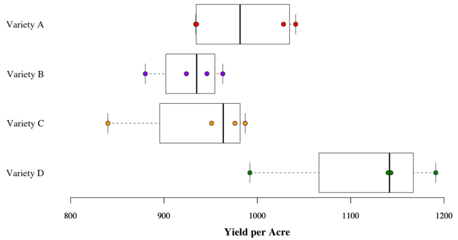
Example

Does rice variety influence average yield amongst these four varieties?

One of the typical examples for introducing ANOVA concerns comparing rice yields across four different varieties. Here are the raw data:

Variety A	934, 1041, 1028, 935
Variety B	880, 963, 924, 946
Variety C	987, 951, 976, 840
Variety D	992, 1143, 1140, 1191

Example 1: Rice Yields



Example 1: Rice Yields

Here is the blank ANOVA table. Let's perform the calculations by hand

Source	**SS**	**df**	**MS**	**F**	**p**
Model					
Error					
Total					

Example 1: The SS Column

The column marked "SS" contains the "sum of squares" for the three sources. The sum of squares is just the sum of the deviation between the observation and the mean. As such,

$$SS_{\text{Model}} = \sum_i \sum_j \left(\bar{y}_j - \bar{\bar{y}}\right)^2$$

$$SS_{\text{Error}} = \sum_i \sum_j \left(y_{i,j} - \bar{y}_j\right)^2$$

$$SS_{\text{Total}} = \sum_i \sum_j \left(y_{i,j} - \bar{\bar{y}}\right)^2$$

In each of these, the i represents a value within a group, and j represents a group. Also, $\bar{\bar{y}}$ is the average of all measurements (the grand mean) and $\bar{y}_j$ is the average of the measurements in group j.

Example 1: The SS Column

And so, let us calculate the sum of squares described by the model:

$$\begin{aligned} SS_{\text{Model}} &= \sum_i \sum_j \ (\bar{y}_j - \bar{\bar{y}})^2 \\ &= (\bar{y}_1 - \bar{\bar{y}})^2 + (\bar{y}_2 - \bar{\bar{y}})^2 + (\bar{y}_3 - \bar{\bar{y}})^2 + (\bar{y}_4 - \bar{\bar{y}})^2 \\ &+ (\bar{y}_1 - \bar{\bar{y}})^2 + (\bar{y}_2 - \bar{\bar{y}})^2 + (\bar{y}_3 - \bar{\bar{y}})^2 + (\bar{y}_4 - \bar{\bar{y}})^2 \\ &+ (\bar{y}_1 - \bar{\bar{y}})^2 + (\bar{y}_2 - \bar{\bar{y}})^2 + (\bar{y}_3 - \bar{\bar{y}})^2 + (\bar{y}_4 - \bar{\bar{y}})^2 \\ &+ (\bar{y}_1 - \bar{\bar{y}})^2 + (\bar{y}_2 - \bar{\bar{y}})^2 + (\bar{y}_3 - \bar{\bar{y}})^2 + (\bar{y}_4 - \bar{\bar{y}})^2 \end{aligned}$$

From the data:

$\bar{y}_1 = 984.50$	$\bar{y}_2 = 928.25$	$\bar{\bar{y}} = 991.9375$
$\bar{y}_3 = 938.50$	$\bar{y}_4 = 1116.50$	

Start of Lecture Material
A Framing Example
A Few Examples
End of Section Material
Example 1: Rice Yields
Example 2: Fisher 38

Example 1: The SS Column

and...

$$\begin{aligned} SS_{\text{Model}} &= (984.50 - 991.9375)^2 + (928.25 - 991.9375)^2 \\ &+ (938.50 - 991.9375)^2 + (1116.50 - 991.9375)^2 \\ &+ (984.50 - 991.9375)^2 + (928.25 - 991.9375)^2 \\ &+ (938.50 - 991.9375)^2 + (1116.50 - 991.9375)^2 \\ &+ (984.50 - 991.9375)^2 + (928.25 - 991.9375)^2 \\ &+ (938.50 - 991.9375)^2 + (1116.50 - 991.9375)^2 \\ &+ (984.50 - 991.9375)^2 + (928.25 - 991.9375)^2 \\ &+ (938.50 - 991.9375)^2 + (1116.50 - 991.9375)^2 \\ &= 89,931 \end{aligned}$$

Example 1: The SS Column

This should illustrate why calculation by hand is no longer realistic. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	89,931				
Error	49,876				
Total	139,807				

Note: Realize that $SSM + SSE = SST$.

Example 1: The df Column

The column marked "df" contains the "degrees of freedom" for the three sources. What are degrees of freedom? They are parameters that reflect the amount of information contributed by each source.*

$$
\begin{aligned}
df_{\text{Model}} &= k-1 && = 4-1 && = 3\\
df_{\text{Error}} &= k(n-1) && = 4(4-1) && = 12\\
df_{\text{Total}} &= kn-1 && = 16-1 && = 15
\end{aligned}
$$

In each of these, the k represents the number of groups, n represents the sample size *within each group*, and kn represents the total sample size, N.

Example 1: The df Column

Calculating the number of degrees of freedom is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	89,931	3			
Error	49,876	12			
Total	139,807	15			

Note: $df_M + df_E = df_T$.

Example 1: The MS Column

The column marked "MS" contains the "mean squares" for the three sources. These are the estimates of the individual variances. They are the SS divided by the df for each source.

$$MS_{\text{Model}} = \frac{SS_{\text{Model}}}{df_{\text{Model}}} = \frac{89,931}{3} = 29,977$$

$$MS_{\text{Error}} = \frac{SS_{\text{Model}}}{df_{\text{Model}}} = \frac{49,876}{12} = 24,156$$

Example 1: The MS Column

Calculating the number of mean squares values is rather easy. Here are the results:

Source	SS	df	MS	F	p
Model	89,931	3	29,977		
Error	49,876	12	4,156		
Total	139,807	15			

Note: $MSM + MSE \neq MST$. Also note that MST is the variance of the data, s^2.

Example 1: The F Column

The column marked “F” contains the “F-statistic” for the model.

$$F = \frac{MS_{\text{Model}}}{MS_{\text{Error}}} = \frac{29,977}{4,156} = 7.212$$

As with all test statistics, it is a measure of how far the data are from the null hypothesis. It has a distribution. As you can/should guess, the distribution of the F statistic is F.

Example 1: The F Column

Calculating the F statistic is rather easy. Here are the results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	89,931	3	29,977	7.212	
Error	49,876	12	4,156		
Total	139,807	15			

Example 1: The p-value Column

The column marked “p” contains the p-value for the model. It is interpreted in the usual way.

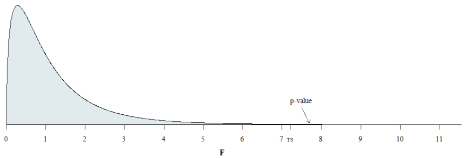
$$\text{p-value} = \mathbb{P}[\ F \geq f\ ] = 1 - \mathbb{P}[\ F \leq f\ ]$$

Here, f is the value of the test statistic you calculated above.

As with all p-values, it is a measure of how far the data are from the null hypothesis. Compare it to your selected value of α. If the p-value is less than α, then you reject the null hypothesis.

Framing Example

Here is the distribution of the test statistic and what we observed. Note that the p-value — the area to the right of the observed test statistic value — is extremely small. Thus, we would expect to reject the null hypothesis.



Example 1: The Results

With that, here are the final results:

Source	**SS**	**df**	**MS**	**F**	**p**
Model	89,931	3	29,977	7.212	0.00503
Error	49,876	12	4,156		
Total	139,807	15			

Brief conclusion: Because the p-value of 0.00503 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. At least one of the population means differ. The two variables are not independent. The model helps in our prediction accuracy.

Example 1: Rice Yields (with R)

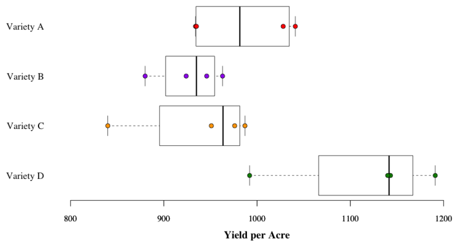
Example

Does rice variety influence average yield amongst these four varieties?

The data can be downloaded from the expected place in the usual manner:

```
dt = read.csv("http://rfs.kvasaheim.com/data/rice.csv")
summary(dt)
attach(dt)
```


Example 1: Rice Yields (with R)



Example 1a: Rice Yields (with R)

Here is the code to run the ANOVA in R:

```
ricemod = aov(yield ~ variety)
summary(ricemod)
```

Here are the results:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
variety      3  89931   29977   7.212 0.00503 **
Residuals   12  49876    4156
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Example 1: Rice Yields (with R)

Brief Conclusion:

Because the p-value of 0.00503 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. The average yield per acre for the four varieties is not the same. The yield and variety variables are *dependent*. Our ability to predict the yield for a plot depends on knowing the rice yield planted.

Example 2: Fisher 38

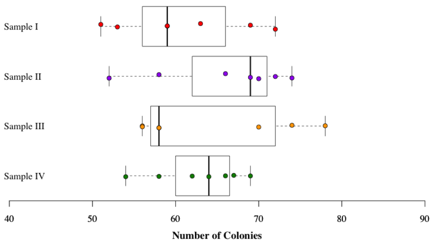
Ronald Fisher introduced the ANOVA procedure to the world in 1925 in is book *Statistical Methods for Research Workers*. In that book, he described the following experiment:

> *Collect a sample of pond water. Divide that water amongst four different beakers. Separate the beakers to ensure that there is no cross-contamination. For each beaker, take four samples and count and record the number of amœba present.*

The results of this experiment, he provided in Table 3.8 in his book.

They are also available as the fisher38 datafile.

Example 2: Fisher 38



Example 2: Fisher 38

Here is the entire script I used to determine if the average number of amœba differed among the four beakers.

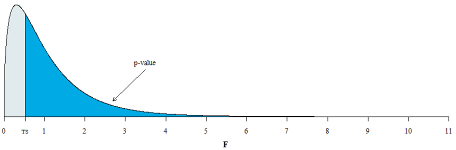
```
dt=read.csv("http://rfs.kvasaheim.com/data/fisher38.csv")
attach(dt)

ickymod = aov(colonies ~ sample)
summary(ickymod)
```

Here is the output:

```
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
sample        3      95    31.65    0.525   0.669
Residuals    24    1446    60.25
```


Example 2: Fisher 38

Here is the distribution of the test statistic and what we observed. Note that the p-value — the area to the right of the observed test statistic value — is not that small. Thus, we would *not* expect to reject the null hypothesis.



Example 2: Fisher 38

Brief Conclusion:

Because the p-value of 0.6690 is greater than our usual $\alpha = 0.05$, we cannot reject the null hypothesis. We did not detect a difference in the average number of amœba in the samples across the four beakers.

- Does this result make sense?

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. if the mean of several groups are the same
2. whether a categorical variable helps explain a numeric
3. if a numeric and a categorical variable are independent

Today's R Functions

Here is what we used the following R functions:

- `mod = aov(x ~ g)`
  performs the ANOVA procedure
- `summary(mod)`
  provides the results from the above ANOVA

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning the Analysis of Variance (ANOVA) procedure.

- SCA 42a and 42b

Please note that there are a couple of examples in these SCAs that use the Kruskal-Wallis test, a non-parametric version of ANOVA. We will cover the Kruskal-Wallis test in the next lecture.

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding the theory of the analysis of variance (ANOVA) procedure:

- Hawkes Learning: Section 11.6
- Intro to Modern Statistics: Chapter 22
- R for Starters: Chapter 7

- Wikipedia: ANOVA

Please do not forget to be familiar with the `allProcedures` document that provides all of the statistical procedures we will use in R.