Public Policy 529
Anova

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 8, 2023

Outline

1. Comparing Several Means

2. Mathematical Details

3. Examples

Outline

Why Analysis of Variance (ANOVA)?

- We have learned how to perform comparison of means tests involving two means.
- When there were more than two categories, we could only compare two at a time.
- We instead may want to perform a comparison of means across many groups simultaneously.

  e.g. the mean final exam scores across five sections of a course that had different GSIs
- We can do this with ANOVA.

Example: Obama Thermometer and Party ID

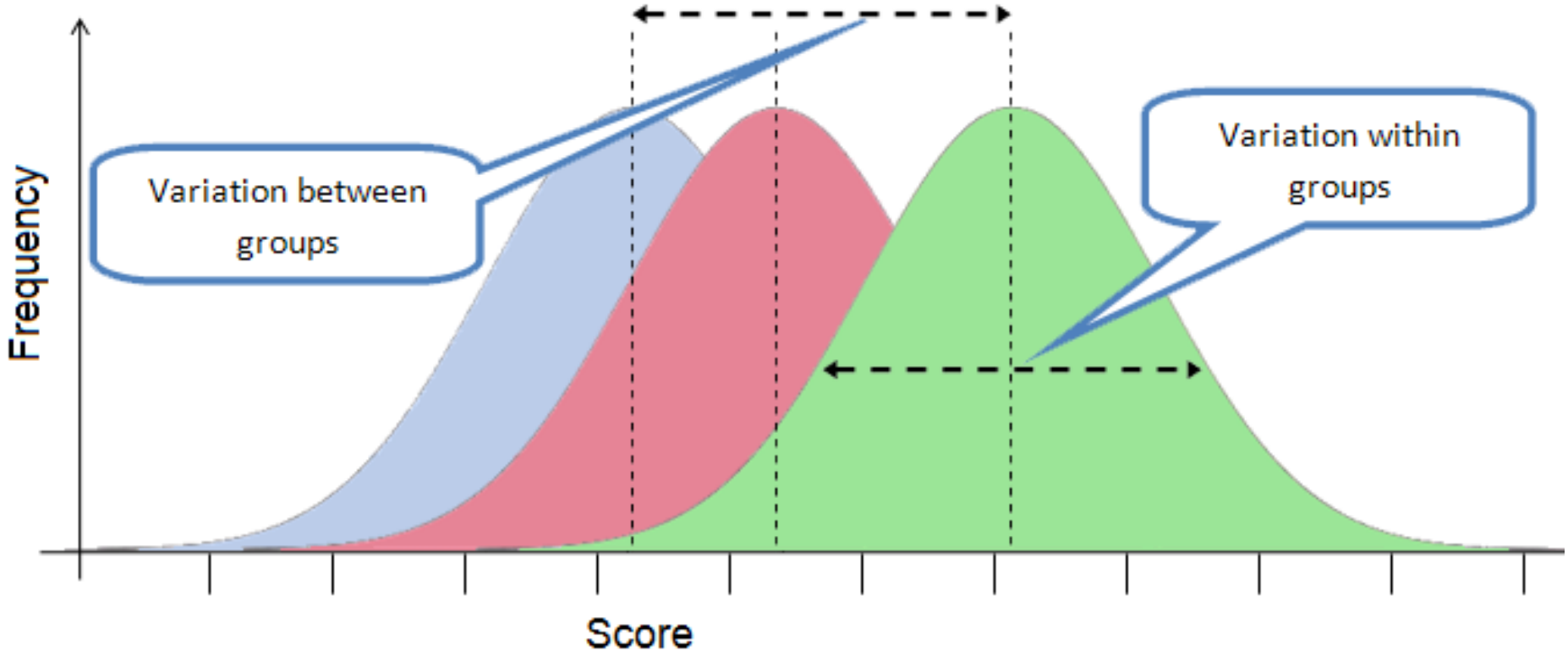
Party ID	Mean Thermometer	St. Dev.	Frequency
Democratic	85.31	19.1	40.1% (2,197)
Independent	55.73	31.5	36.2% (1,984)
Republican	26.62	26.7	23.6% (1,293)
Total	60.72	34.6	100% (5,474)

From the 2012 American National Election Study

Basic Idea Behind ANOVA

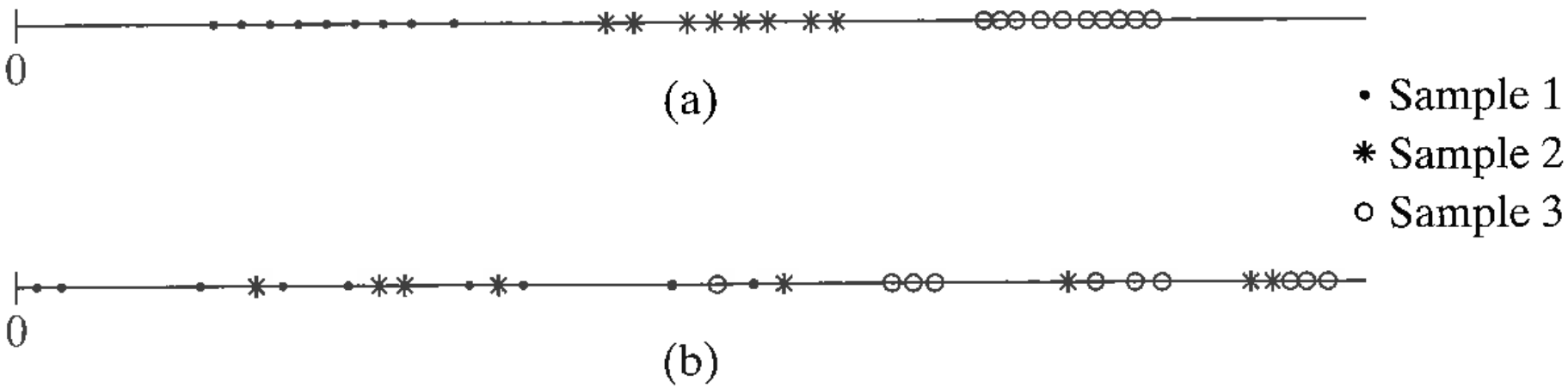
- We have an interval-level dependent variable and one or more more categorical independent variables.
- Does the population distribution of the dependent variable (y) differ for the groups formed by the independent variable(s)?
- ANOVA partitions variation in y into within-group variation in y and between-group variation in y.
- If between-group variation in y is high relative to within-group variation, we reject H_0 that the mean of y is the same for all groups.

Illustration



Note: variation between groups is about deviations between the individual group means and the overall sample mean of y.

Illustration



In (a), between-group variance is large relative to within-group variance. In (b), between-group variance is smaller on a relative basis. The sample means in (b) are less distinct from each other.

$\Rightarrow$ The ratio of between-group variance to within-group variance provides evidence about whether the categories of x are systematically associated with variation in y.

Forming a Test Statistic

The test involves the ratio of the between-group variance to within-group variance. This produces an F statistic:

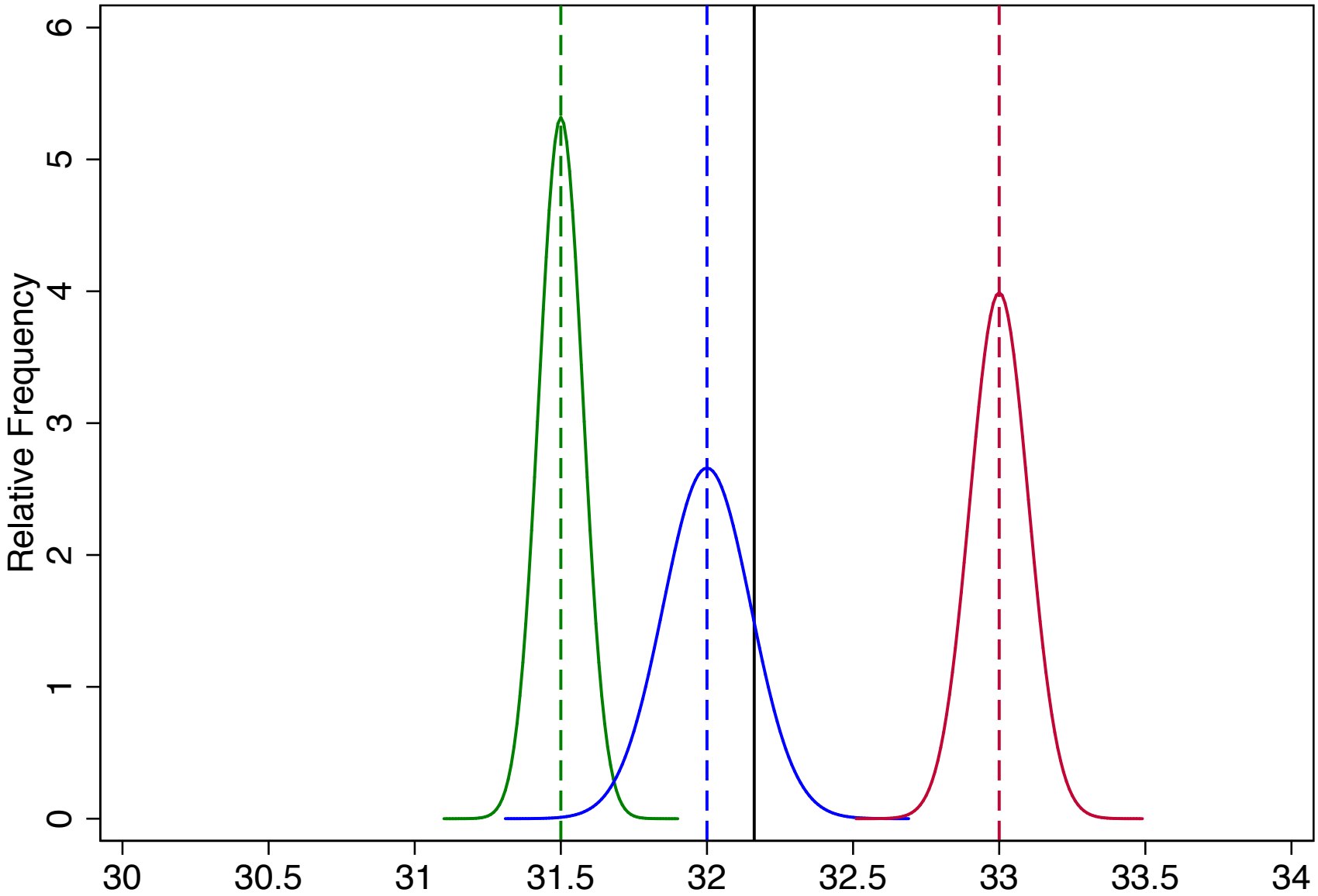
$$F = \frac{\text{Variance in } y \text{ between groups}}{\text{Variance in } y \text{ within groups}}$$

If there are G groups, this statistic has $G - 1$ degrees of freedom in the numerator, and $n - G$ degrees of freedom in the denominator.

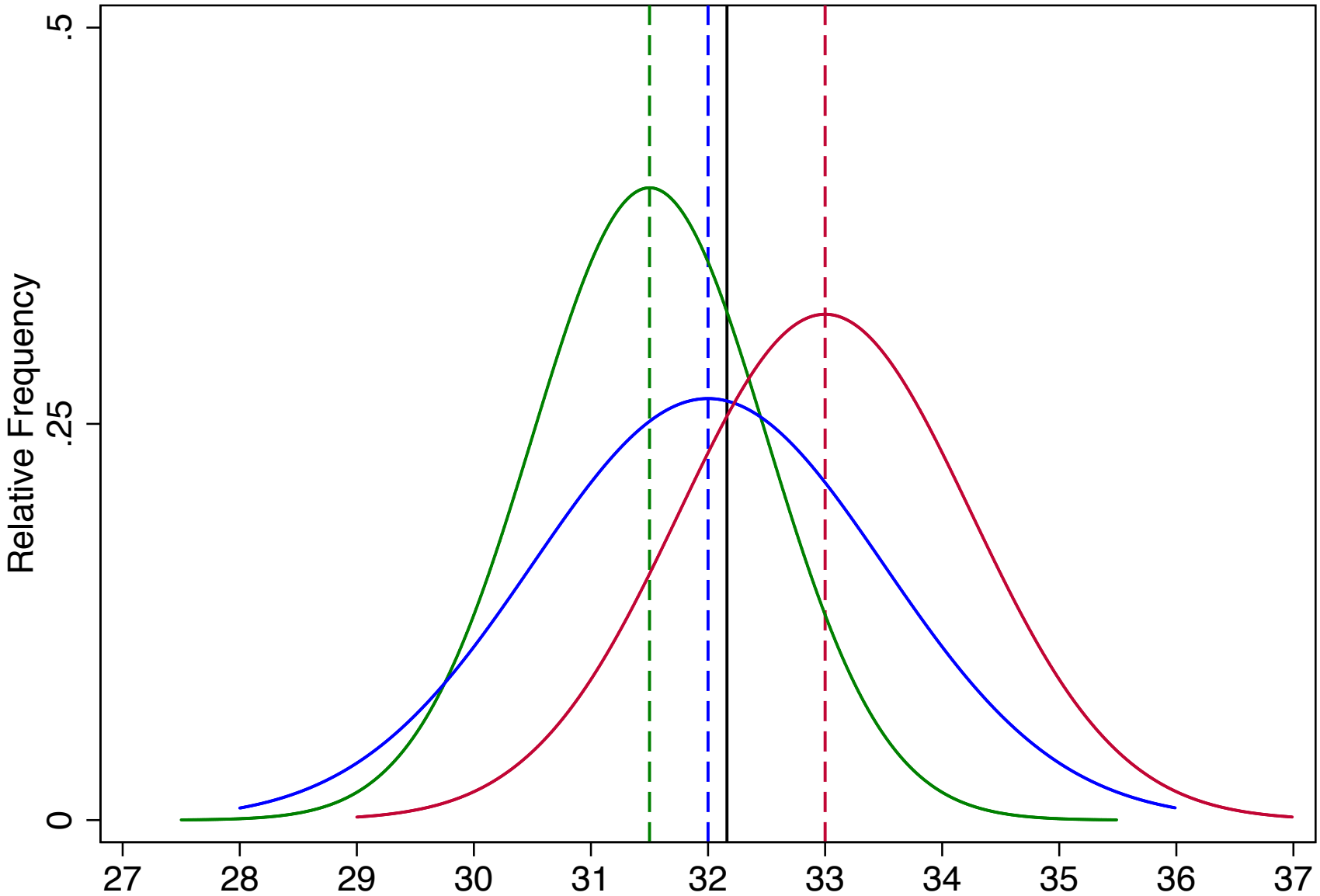
Illustration

- Suppose the overall mean of y is 32.2. The means of y for three categories of x are: 31.5, 32, and 33.
- Let's take two scenarios. First, the within-group variance is small. Second, the within-group variance is large.
- The between-group variance is the same in both scenarios.
- We will see why the F-statistic will be much larger in the first scenario.

Scenario #1: Within-Group Variance is Small



Scenario #2: Within-Group Variance is Large



Outline

Stating the Hypotheses

Suppose the number of groups is G. The null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G$$

The alternative hypothesis is:

H_A : at least two population means are unequal.

In other words, H_0 can be rejected in a variety of circumstances.

Within- vs. Between-Group Variance

Suppose we have a sample of size n. Observation y_i is the ith observation of y, where i goes from 1 to n.

Now suppose this sample is divided according to the categories of another variable (x) so that there are G different groups. The sample size of group g is n_g.

- Within-group variance is about deviations of y_{gi} from $\bar{y}_g$, which is the mean of y in group g.
- Between-group variance is about deviations of $\bar{y}_g$ from $\bar{y}$, the overall mean of y.

Between-Group Variance

1. Find the sum of the squared deviations between groups.

$$\begin{aligned} BSS &= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + \ldots + n_G(\bar{y}_G - \bar{y})^2 \\ &= \sum_{g=1}^{G} n_g(\bar{y}_g - \bar{y})^2 \end{aligned}$$

2. Divide by the degrees of freedom, which is $G - 1$, to get the estimate of the between-group variance:

$$\frac{BSS}{df_{BSS}} = \frac{\sum n_g(\bar{y}_g - \bar{y})^2}{G - 1}$$

This is the numerator of the F statistic. It has $G - 1$ degrees of freedom.

Within-Group Variance

1. Add up the sums of the squared deviations within each group.

$$WSS = SS_1 + SS_2 + \ldots + SS_g$$

Where, each SS_g is:

$$SS_g = \sum_{i=1}^{n_g} (y_i - \bar{y}_g)^2$$

2. Divide by the degrees of freedom, which is $n - G$, to get the estimate of the within-group variance:

$$\frac{WSS}{df_{WSS}} = \frac{SS_1 + SS_2 + \ldots + SS_g}{n - G}$$

This is the denominator of the F statistic. It has $n - G$ degrees of freedom.

Brief Aside: Calculating WSS

$$SS_g = \sum_{i=1}^{n_g} (y_i - \bar{y}_g)^2$$

The above looks familiar because it is the numerator of the formula for the sample variance:

$$s_g^2 = \frac{\sum (y_i - \bar{y}_g)^2}{n_g - 1}$$

Thus, an equivalent way to calculate the WSS is:

$$WSS = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_g - 1)s_g^2$$

Putting it All Together

We are now ready to create our F-statistic:

$$F_{G-1,n-G} = \frac{\sum n_g(\bar{y}_g - \bar{y})^2/(G-1)}{(SS_1 + SS_2 + \ldots + SS_g)/(n-G)}$$

Reminder: this is the ratio of the between-group variance to the within-group variance.

If the F statistic is sufficiently high, we can reject the null hypothesis that there is no difference in the means between the groups.

Outline

One-Way ANOVA

- One-Way ANOVA tests whether variation in the dependent variable is connected with the categories of one independent variable.

  e.g. are there significant differences in Obama thermometer scores across different categories of party identification?

  e.g. are there significant differences in student test scores across different classrooms or schools?

- If there are only two categories, we could do this with two-sample t-tests. This is not efficient for more than two categories.

Example: Obama Thermometer and Party ID

Party ID	Mean Thermometer	St. Dev.	Frequency
Democratic	85.31	19.1	40.1% (2,197)
Independent	55.73	31.5	36.2% (1,984)
Republican	26.62	26.7	23.6% (1,293)
Total	60.72	34.6	100% (5,474)

From the 2012 American National Election Study

Example: Between-Group Variance

1. Find the sum of the squared deviations between groups.

$$\begin{aligned} BSS &= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2 \\ &= 2,197(85.31 - 60.72)^2 + 1,984(55.73 - 60.72)^2 \\ &\quad +1,293(26.62 - 60.72)^2 \\ &= 2,881,370.9 \end{aligned}$$

2. Divide by the degrees of freedom, which is $G - 1$, to get the estimate of the between-group variance:

$$\frac{BSS}{df_{BSS}} = \frac{\sum n_g(\bar{y}_g - \bar{y})^2}{G - 1} = \frac{2,881,370.9}{2} = 1,440,685.5$$

This is the numerator of the F statistic.

Example: Within-Group Variance

1. Add up the sums of the squared deviations within each group.

$$\begin{aligned} WSS &= SS_1 + SS_2 + SS_3 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \\ &= 2,196 * 19.1^2 + 1,983 * 31.5^2 + 1,292 * 26.7^2 \\ &= 3,689,808.4 \end{aligned}$$

2. Divide by the degrees of freedom, which is $n - G$, to get the estimate of the within-group variance:

$$\frac{WSS}{df_{WSS}} = \frac{3,689,808.4}{5,474 - 3} = 674.4$$

This is the denominator of the F statistic.

Calculate the F-statistic

We are now ready to create our F-statistic:

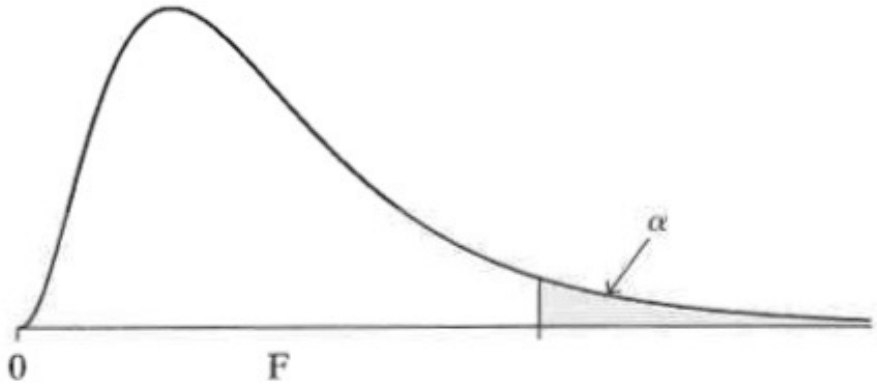
$$
\begin{aligned}
F_{G-1,n-G} &= \frac{\text{between-group variance}}{\text{within-group variance}} \\
&= \frac{1,440,685.5}{674.4} \\
&= 2,136.2
\end{aligned}
$$

We compare this statistic to the critical value that comes from the F table with 2 degrees of freedom in the numerator and 5,471 degrees of freedom in the denominator.

According to the table, the critical value is between 3.07 and 2.99. We can reject the null hypothesis.

Using the F Table

TABLE D: *F* Distribution



	$\alpha = .05$									
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Example: Estimating with Stata

The basic command is: anova dep_variable ind_variable

```
anova Obama_therm pid_3
```

```
                   Number of obs =      5,474    R-squared     =  0.4386
                   Root MSE      =    25.9629    Adj R-squared =  0.4384
```

Source	Partial SS	df	MS	F	Prob>F
Model	2881001.3	2	1440500.6	2137.01	0.0000
pid_3	2881001.3	2	1440500.6	2137.01	0.0000
Residual	3687850.8	5,471	674.07253		
Total	6568852.1	5,473	1200.2288		

There are slight differences from the previous slides due to rounding.

Example: Estimating with R

The basic command is: aov(dep_variable ~ ind_variable)

```
> anova <- aov(Obama_therm ~ as.factor(pid_3), data = nes2012)
> summary(anova)
                  Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(pid_3)    2 2881001 1440501    2137 <2e-16 ***
Residuals        5471 3687851     674
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
442 observations deleted due to missingness
```

The independent variable should be of the factor class. In the command, I use the `as.factor()` function to have R convert a numeric variable to a factor.

Example: Student Test Scores

- There is variation across students in terms of math test scores (y).
- Some of this variation might be due to the fact that students are located in different schools (x).
- We can test whether variation in test scores across schools helps explain overall variation in test scores.
- This is equivalent to testing whether school-mean math test scores differ across the schools.

Example: Student Test Scores

- Data: 8,064 student state math test scores from 25 different schools.

  H_0: the mean test score is the same across all the schools.

  H_a: the mean test scores differ from each other for at least some schools.

Example: Student Test Scores

```
anova Math_testscore SchoolID
```

Number of obs =	8,064	R-squared =	0.0871
Root MSE =	37.5021	Adj R-squared =	0.0843

Source	Partial SS	df	MS	F	Prob>F
Model	1078162.8	24	44923.45	31.94	0.0000
SchoolID	1078162.8	24	44923.45	31.94	0.0000
Residual	11306119	8,039	1406.4086		
Total	12384282	8,063	1535.9397		

The F-statistic is 31.94 and the p-value is .0000. We can reject the null hypothesis that there are no differences in school-level means.

Two-Way ANOVA

- Two-Way ANOVA tests whether variation in the dependent variable is connected with the categories created by different combinations of two independent variables.

- The test is whether the population means are identical across the categories of one variable when controlling for the other one.

  e.g. are there significant differences in Obama thermometer scores across different combinations of party identification and region (South, non-South)?

  e.g. are there significant differences in life expectancy across the categories of democracy/non-democracy and categories of GDP per capita (low, medium, high)?

Example: Obama Thermometer, Party ID, and Region

	Mean Thermometer Score		
Party ID	Non-South	South	Total
Democratic	81.9	86.1	83.2
Independent	54.2	51.5	53.4
Republican	27.1	24.7	26.3
Total	56.8	55.6	56.4

From the 2012 American National Election Study

Example: Obama Thermometer

The basic command is: anova dep_variable ind_var1 ind_var2

```
. anova Obama_therm pid_3 south
```

Number of obs =	5,474	R-squared =	0.4390	
Root MSE =	25.9553	Adj R-squared =	0.4387	

Source	Partial SS	df	MS	F	Prob>F
Model	2883822.5	3	961274.16	1426.90	0.0000
pid_3	2879442	2	1439721	2137.10	0.0000
south	2821.176	1	2821.176	4.19	0.0408
Residual	3685029.7	5,470	673.68001		
Total	6568852.1	5,473	1200.2288		

Party ID matters for the Obama thermometer even after controlling for region. Region matters after controlling for party ID. See the respective F-statistics and p-values.

Example: Using R

```
> anova2 <- aov(Obama_therm ~ as.factor(pid_3) + as.factor(south), data = nes2012)
> summary(anova2)
                   Df  Sum Sq Mean Sq  F value Pr(>F)
as.factor(pid_3)    2 2881001 1440501 2138.256 <2e-16 ***
as.factor(south)    1    2821    2821    4.188 0.0408 *
Residuals        5470 3685030     674
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
442 observations deleted due to missingness
```

Interpretation

- We can reject the null hypothesis that the mean Obama thermometer score is the same in different categories of the PartyID variable, even after we control for whether or not a respondent lives in the South.
- We reject the null hypothesis that the mean Obama thermometer scores are the same for Southerners and non-Southerners, even after we control for PartyID.

Example: Democracy, GDP, and Life Expectancy

Mean Level of Life Expectancy

	Level of GDP/capita			
Democracy?	Low	Medium	High	Total
No	58.8	67.4	72.6	64.5
	(27)	(23)	(11)	(61)
Yes	63.1	73.0	78.3	72.5
	(28)	(31)	(43)	(102)
Total	61.0	70.6	77.1	69.5
	(55)	(54)	(54)	(163)

Example: Democracy, GDP, and Life Expectancy

```
anova cialifex democ_regime08 gdp_cap3
```

Number of obs =	163	R-squared =	0.5404
Root MSE =	6.59158	Adj R-squared =	0.5317

Source	Partial SS	df	MS	F	Prob>F
Model	8122.6016	3	2707.5339	62.32	0.0000
democ_re~08	947.4889	1	947.4889	21.81	0.0000
gdp_cap3	5677.217	2	2838.6085	65.33	0.0000
Residual	6908.3823	159	43.448945		
Total	15030.984	162	92.783851		

Both both democracy and GDP per capita are strongly related to life expectancy even when controlling for the other variable. See the respective F-statistics and p-values.

Interpretation

- We reject the null hypothesis that mean life expectancy is the same in democracies and non-democracies, even after we control for GDP per capita category.
- We reject the null hypothesis that mean life expectancy is the same for countries in different categories of GDP per capita, even after we control for the country's regime type.

Summary

- When the independent variable has more than two categories, or when we have multiple categorical independent variables, pairwise t-tests are too cumbersome.
- ANOVA is better for these circumstances. We test whether differences between the categories explains significant variance in the dependent variable.
- There is a strong relationship between ANOVA and linear regression (properly specified). The two approaches will give the same results.