Lecture 10: Analysis of variance

Ernesto F. L. Amaral

March 28, 2023
Advanced Methods of Social Research (SOCI 420)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 10 (pp. 247–275).

TEXAS A&M UNIVERSITY

Outline

- Identify and cite examples of situations in which analysis of variance (ANOVA) is appropriate
- Explain the logic of hypothesis testing as applied to ANOVA
- Perform the ANOVA test, using the five-step model as a guide, and correctly interpret the results
- Define and explain the concepts of population variance, total sum of squares, sum of squares between, sum of squares within, mean square estimates
- Explain the difference between the statistical significance and the importance (magnitude) of relationships between variables

ANOVA application

- ANOVA can be used in situations where the researcher is interested in the differences in sample means across three or more categories
  - How do Protestants, Catholics, and Jews vary in terms of number of children?
  - How do Republicans, Democrats, and Independents vary in terms of income?
  - How do older, middle-aged, and younger people vary in terms of frequency of church attendance?

Extension of *t*-test

- We can think of ANOVA as an extension of *t*-test for more than two groups
  - Are the differences between the samples large enough to reject the null hypothesis and justify the conclusion that the populations represented by the samples are different?
- Null hypothesis, H_0
  - H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
  - All population means are similar to each other
- Alternative hypothesis, H_1
  - At least one of the populations means is different

Logic of ANOVA

- Could there be a relationship between age and support for capital punishment?
  - No difference between groups

Support for Capital Punishment by Age Group (fictitious data)

	18–29	30–45	46–64	65+
Mean	10.3	11.0	10.1	9.9
Standard deviation	2.4	1.9	2.2	1.7

  - Difference between groups

Support for Capital Punishment by Age Group (fictitious data)

	18–29	30–45	46–64	65+
Mean	10.0	13.0	16.0	22.0
Standard deviation	2.4	1.9	2.2	1.7

Source: Healey 2015, p.249.

Between and within differences

- If the H_0 is true, the sample means should be about the same value
  - If the H_0 is true, there will be little difference between sample means

- If the H_0 is false
  - There should be substantial differences **<u>between</u>** sample means (between categories)
  - There should be relatively little difference **<u>within</u>** categories
    - The sample standard deviations should be small within groups

Likelihood of rejecting H_0

- The greater the difference **between** categories (as measured by the means)
  - Relative to the differences **within** categories (as measured by the standard deviations)
  - The more likely the H_0 can be rejected

- When we reject H_0
  - We are saying there are differences between the **populations** represented by the sample

Computation of ANOVA

1. Find total sum of squares (*SST*)

$$SST = \sum(X_i^2) - n\bar{X}^2$$

2. Find sum of squares between (*SSB*)

$$SSB = \sum[n_k(\bar{X}_k - \bar{X})^2]$$

   - *SSB* = sum of squares between categories
   - n_k= number of cases in a category
   - $\bar{X}_k$= mean of a category

3. Find sum of squares within (*SSW*)

SSW = *SST* – *SSB*

4. Degrees of freedom

$$dfw = n - k$$

- dfw = degrees of freedom within
- n = total number of cases
- k = number of categories

$$dfb = k - 1$$

- dfb = degrees of freedom between
- k = number of categories

Final estimations

5. Find mean square estimates

$$Mean\ square\ within = \frac{SSW}{dfw}$$

$$Mean\ square\ between = \frac{SSB}{dfb}$$

6. Find the *F* ratio

$$F(obtained) = \frac{Mean\ square\ between}{Mean\ square\ within}$$

Example

- Support for capital punishment
- Sample of 16 people who are equally divided across four age groups

Support for Capital Punishment by Age Group (fictitious data)

18–29		30–45		46–64		65+	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
7	49	10	100	12	144	17	289
8	64	12	144	15	225	20	400
10	100	13	169	17	289	24	576
15	225	17	289	20	400	27	729
40	438	52	702	64	1058	88	1994
$\bar{X}_k = 10.0$		$\bar{X}_k = 13.0$		$\bar{X}_k = 16.0$		$\bar{X}_k = 22.0$	

$\bar{X} = 15.25$

Source: Healey 2015, p.252.

Step 1: Assumptions,requirements

- Independent random samples
- Interval-ratio level of measurement
- Normally distributed populations
- Equal population variances

Step 2: Null hypothesis

- Null hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
  - The null hypothesis asserts there is no difference between the populations

- Alternative hypothesis, H_1
  - At least one of the populations means is different

Step 3: Distribution, critical region

- Sampling distribution
  - *F* distribution
- Significance level
  - Alpha (α) = 0.05
- Degrees of freedom
  - $dfw = n - k = 16 - 4 = 12$
  - $dfb = k - 1 = 4 - 1 = 3$
- Critical *F*
  - F(critical) = 3.49

Step 4: Test statistic

1. Total sum of squares (*SST*)

$$SST = \sum(X_i^2) - n\bar{X}^2$$

$$SST = (438 + 702 + 1058 + 1994) - (16)(15.25)^2$$
$$SST = 471.04$$

2. Sum of squares between (*SSB*)

$$SSB = \sum[n_k(\bar{X}_k - \bar{X})^2]$$

$$SSB = 4(10 - 15.25)^2 + 4(13 - 15.25)^2$$
$$+ 4(16 - 15.25)^2 + 4(22 - 15.25)^2 = 314.96$$

3. Sum of squares within (*SSW*)

$SSW = SST - SSB = 471.04 - 314.96 = 156.08$

4. Degrees of freedom

$dfw = n - k = 16 - 4 = 12$

$dfb = k - 1 = 4 - 1 = 3$

5. Mean square estimates

$$Mean\ square\ within = \frac{SSW}{dfw} = \frac{156.08}{12} = 13.00$$

$$Mean\ square\ between = \frac{SSB}{dfb} = \frac{314.96}{3} = 104.99$$

6. *F* ratio

$$F(obtained) = \frac{Mean\ square\ between}{Mean\ square\ within} = \frac{104.99}{13.00}$$
$$= 8.08$$

Step 5: Decision, interpret

- *F*(obtained) = 8.08
- This is beyond *F*(critical) = 3.49
- The obtained test statistic falls in the critical region, so we ***reject*** the H_0
- Support for capital punishment does differ across age groups

Limitations of ANOVA

- Requires interval-ratio level measurement of the dependent variable
- Requires roughly equal numbers of cases in the categories of the independent variable
- Statistically significant differences are not necessarily important (small magnitude)
- The alternative (research) hypothesis is not specific
  - It only asserts that at least one of the population means differs from the others

Example from 2016 GSS

- We know the average income by race/ethnicity

```
. tabstat conrinc [aweight=wtssall], by(raceeth) stat(mean sd n)

Summary for variables: conrinc
Group variable: raceeth (Race/Ethnicity)
```

raceeth	Mean	SD	N
White	38845.62	39157.17	1072
Black	23243.04	19671.53	273
Hispanic	23128.92	21406.31	215
Other	50156.35	59219.9	72
Total	34649.3	36722.06	1632

- Does at least one category of the race/ethnicity variable have average income different than the others?
  - This is not a perfect example for ANOVA, because the race/ethnicity variable does not have equal numbers of cases across its categories

Example from GSS: Result

- The probability of not rejecting H_0 is small (p<0.01)
  - At least one category of the race/ethnicity variable has average income different than the others with a 99% confidence level
  - However, ANOVA does not inform which category has an average income significantly different than the others in 2016

```
. oneway conrinc raceeth [aweight=wtssall]

                        Analysis of variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1.0142e+11      3   3.3806e+10     26.23     0.0000
 Within groups      2.0980e+12   1628   1.2887e+09
------------------------------------------------------------------------
    Total           2.1994e+12   1631   1.3485e+09

Bartlett's equal-variances test: chi2(3) = 292.7013    Prob>chi2 = 0.000
```

Source: 2016 General Social Survey.

Edited table

Table 1. One-way analysis of variance for individual average income of the U.S. adult population by race/ethnicity, 2004, 2010, and 2016

Source	Sum of Squares	Degrees of Freedom	Mean of Squares	F-test	Prob > F
2004					
Between groups	5.92e+10	3	1.97e+10	16.36	0.0000
Within groups	2.03e+12	1,682	1.21e+09		
Total	2.09e+12	1,685	1.24e+09		
2010					
Between groups	6.02e+10	3	2.01e+10	24.50	0.0000
Within groups	9.79e+11	1,195	819,590,864		
Total	1.04e+12	1,198	867,818,893		
2016					
Between groups	1.01e+11	3	3.38e+10	26.23	0.0000
Within groups	2.10e+12	1,628	1.29e+09		
Total	2.20e+12	1,631	1.35e+09		

Source: 2004, 2010, 2016 General Social Surveys.

Example from 2019 ACS, Texas

- We know the average income by race/ethnicity

```
. tabstat income if income!=0 & income!=. [fweight=perwt], by(raceth) stat(mean sd n)

Summary for variables: income
Group variable: raceth
```

raceth	Mean	SD	N
White	63199.24	74601.04	6081513
African American	40079.03	40410.99	1766063
Hispanic	36595.08	38076.88	5250789
Asian	66528.88	73827.69	776722
Native American	44246.01	57666.53	44743
Other races	46151.98	58649.93	235029
Total	50285.44	60567.56	1.42e+07

```
. svy, subpop(if income!=0 & income!=.): mean income, over(raceth)
(running mean on estimation sample)

. estat sd
```

(correct standard deviation)

Over	Mean	Std. dev.
c.income@ raceth		
White	63199.24	81952.97
African A..	40079.03	33729.03
Hispanic	36595.08	34417.96
Asian	66528.88	71633.26
Native Am..	44246.01	57876.89
Other races	46151.98	56501.55

```
. svy, subpop(if income!=0 & income!=.): mean income
(running mean on estimation sample)

. estat sd
```

	Mean	Std. dev.
income	50285.44	59920.72

- Does at least one category of race/ethnicity have average income different than the others?
  - This is not a perfect example for ANOVA, because race/ethnicity does not have equal numbers of cases across its categories

Source: 2019 American Community Survey, Texas.

Example from ACS: Result

- The probability of not rejecting H_0 is small (p<0.01)
  - At least one category of the race/ethnicity variable has average income different than the others with a 99% confidence level
  - However, ANOVA does not inform which category has an average income significantly different than the others

```
. oneway income raceth if income!=0 & income!=. [aweight=perwt]

                        Analysis of variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2.2032e+13      5   4.4065e+12   1259.17     0.0000
 Within groups      4.5608e+14 130325   3.4995e+09
                                                   (statistical significance)
------------------------------------------------------------------------
    Total           4.7811e+14 130330   3.6685e+09

Bartlett's equal-variances test: chi2(5) =  1.2e+04    Prob>chi2 = 0.000
```

Source: 2019 American Community Survey, Texas.

Example from 2019 ACS: n, N

```
. ***Sample size of each category of race/ethnicity and missing cases
. tab raceth if income!=0 & income!=., m
```

raceth	Freq.	Percent	Cum.
White	69,043	52.98	52.98
African American	11,574	8.88	61.86
Hispanic	40,359	30.97	92.82
Asian	6,879	5.28	98.10
Native American	424	0.33	98.43
Other races	2,052	1.57	100.00
Total	130,331	100.00	

```
. ***Population size of each category of race/ethnicity
. tab raceth if income!=0 & income!=. [fweight=perwt]
```

(correct percentage distribution)

raceth	Freq.	Percent	Cum.
White	6,081,513	42.96	42.96
African American	1,766,063	12.48	55.44
Hispanic	5,250,789	37.10	92.54
Asian	776,722	5.49	98.02
Native American	44,743	0.32	98.34
Other races	235,029	1.66	100.00
Total	14,154,859	100.00	

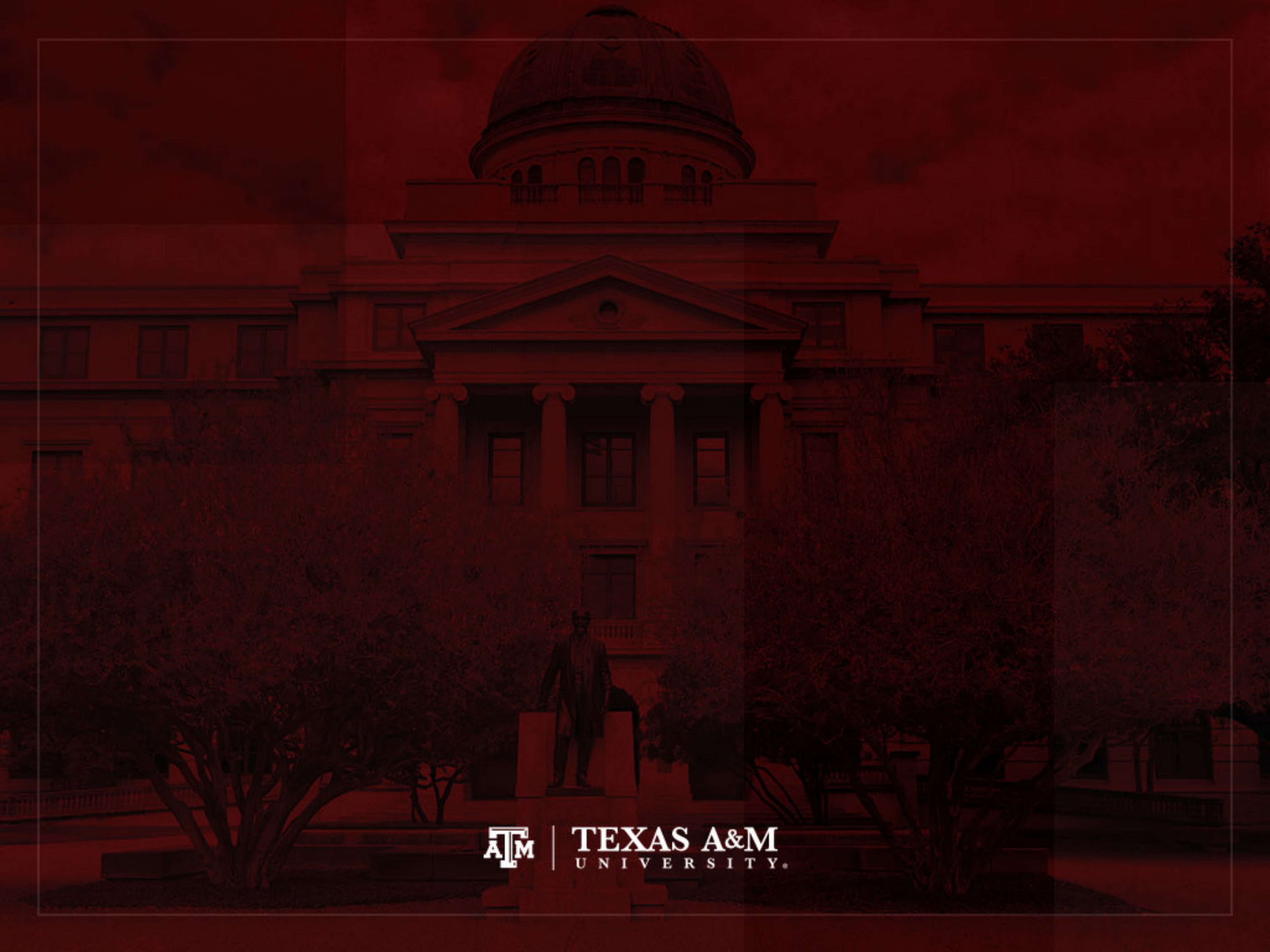
Source: 2019 American Community Survey, Texas.

Edited table

Table 1. One-way analysis of variance for wage and salary income by race/ethnicity, Texas, 2019

Race/ethnicity	Income		Population percentage		
	Mean	Standard deviation			
White	63,199.24	81,952.97	42.96		
African American	40,079.03	33,729.03	12.48		
Hispanic	36,595.08	34,417.96	37.10		
Asian	66,528.88	71,633.26	5.49		
Native American	44,246.01	57,876.89	0.32		
Other races	46,151.98	56,501.55	1.66		
Total	50,285.44	59,920.72	100.00		
Population size	—	—	14,154,859		
Sample size	—	—	130,331		
ANOVA	**Sum of squares**	**Degrees of freedom**	**Mean of squares**	**F-test**	**Prob > F**
Between groups	2.20e+13	5	4.41e+12	1,259.17	0.0000
Within groups	4.56e+14	130,325	3.50e+09		
Total	4.78e+14	130,330	3.67e+09		

Source: 2019 American Community Survey.


TEXAS A&M
UNIVERSITY