Lecture 6: Analysis of variance and Chi square

Ernesto F. L. Amaral

September 29, 2022
Introduction to Sociological Data Analysis (SOCI 600)

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapters 10 (pp. 247–275) and 11 (pp. 276–306).

TEXAS A&M UNIVERSITY

Outline

- Analysis of variance
- Chi square

Analysis of variance

- 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*)

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

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.

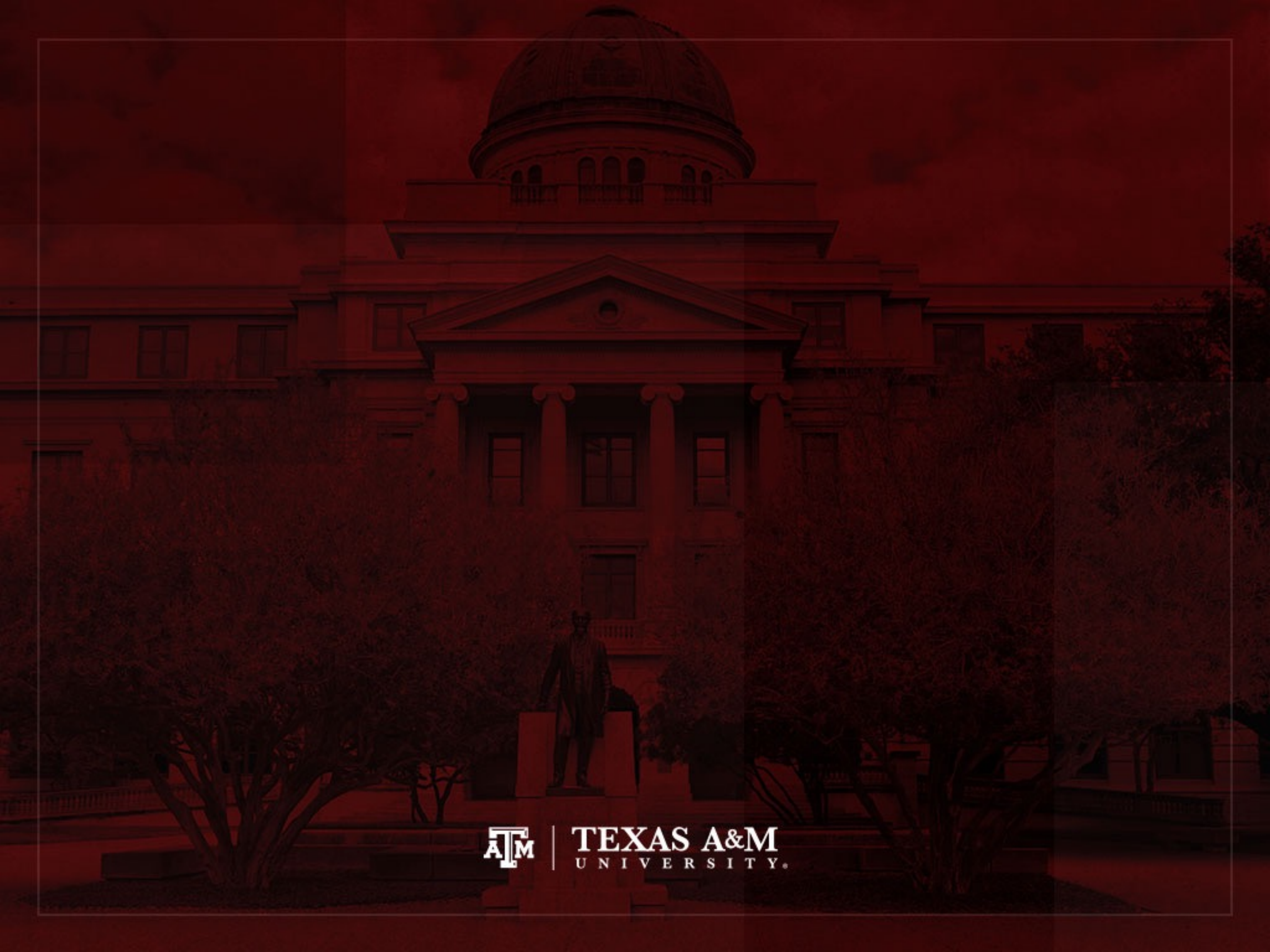
Edited table

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

Chi square

- Identify and cite examples of situations in which the chi square test is appropriate
- Explain the structure of a bivariate table and the concept of independence as applied to expected and observed frequencies in a bivariate table
- Explain the logic of hypothesis testing in terms of chi square
- Perform the chi square test using the five-step model and correctly interpret the results
- Explain the limitations of the chi square test and, especially, the difference between statistical significance and substantive significance (importance, magnitude)

The bivariate table

- Bivariate tables display the scores of cases on two different variables at the same time

Rates of Participation in Voluntary Associations by Marital Status for 100 Senior Citizens

	Marital Status		
Participation Rates	*Married*	*Unmarried*	TOTALS
High			50
Low			50
TOTALS	50	50	100

Source: Healey 2015, p.278.

Aspects of the table

- Note the two dimensions: rows and columns
- What is the independent variable?
- What is the dependent variable?
- Where are the row and column marginals?
- Where is the total number of cases (*n*)?

Rates of Participation in Voluntary Associations by Marital Status for 100 Senior Citizens

	Marital Status		
Participation Rates	*Married*	*Unmarried*	TOTALS
High			50
Low			50
TOTALS	50	50	100

Source: Healey 2015, p.278.

Important information to report

- Must have a title
- Cells are intersections of columns and rows
- Subtotals are called marginals
- Sample size (*n*) or population size (*N*) is reported at the intersection of row and column marginals

Independent, dependent variables

- Columns are scores of the independent variable
  - There will be as many columns as there are scores on the independent variable
- Rows are scores on the dependent variable
  - There will be as many rows as there are scores on the dependent variable
- Each cell reports the number of times each combination of scores occurred
  - There will be as many cells as there are scores on the two variables combined

Test for independence

- Chi square as a test of statistical significance is a test for independence
  - Two variables are independent if the classification of a case into a particular category of one variable has no effect on the probability that the case will fall into any particular category of the second variable

Rates of Participation in Voluntary Associations by Marital Status for 100 Senior Citizens

	Marital Status		
Participation Rates	*Married*	*Unmarried*	TOTALS
High	25	25	50
Low	25	25	50
TOTALS	50	50	100

Source: Healey 2015, p.279.

Cross tabulations

- Chi square is a test of significance based on bivariate tables
  - Bivariate tables are also called cross tabulations, crosstabs, contingency tables
- We are looking for significant differences between
  - The actual cell frequencies observed in a table (f_o)
  - And frequencies that would be expected by random chance or if cell frequencies were independent (f_e)

Computation of chi square

$$f_e = \frac{Row\ marginal \times Column\ marginal}{n}$$

$$\chi^2(obtained) = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o = cell frequencies observed in the bivariate table

f_e = cell frequencies that would be expected if the variables were independent

Example

- Random sample of 100 social work majors
  - We know whether the Council on Social Work Education has accredited their undergraduate programs
  - And whether they were hired in social work positions within three months of graduation
- Is there a significant relationship between employment status and accreditation status?

Employment of 100 Social Work Majors by Accreditation Status of Undergraduate Program

	Accreditation Status		
Employment Status	*Accredited*	*Not Accredited*	TOTALS
Working as a social worker	30	10	40
Not working as a social worker	25	35	60
TOTALS	55	45	100

Source: Healey 2015, p.280.

Step 1: Assumptions,requirements

- Independent random samples
- Level of measurement is nominal
- Note the minimal assumptions
  - No assumption is made about the shape of the sampling distribution
  - The chi square test is nonparametric or distribution-free

Step 2: Null hypothesis

- Null hypothesis, H_0: $f_o = f_e$
  - The variables are independent
  - The observed frequencies are similar to the expected frequencies

- Alternative hypothesis, H_1: $f_o \neq f_e$
  - The variables are dependent of each other
  - The observed frequencies are different than the expected frequencies

Step 3: Distribution, critical region

- Sampling distribution
  - Chi square distribution (χ^2)
- Significance level (α) = 0.05
  - The decision to reject the null hypothesis has only a 0.05 probability of being incorrect
- Degrees of freedom (df) = $(r–1)(c–1)$
  - r = number of rows; c = number of columns
  - $df = (r–1)(c–1) = (2–1)(2–1) = 1$
- $\chi^2(critical) = 3.841$
  - If the probability (p-value) is less than 0.05
  - $\chi^2(obtained)$ will be beyond $\chi^2(critical)$

Step 4: Test statistic

Observed frequencies

	Accreditation Status		
Employment Status	*Accredited*	*Not Accredited*	TOTALS
Working as a social worker	30	10	40
Not working as a social worker	25	35	60
TOTALS	55	45	100

Expected frequencies

	Accreditation Status		
Employment Status	*Accredited*	*Not Accredited*	TOTALS
Working as a social worker	22	18	40
Not working as a social worker	33	27	60
TOTALS	55	45	100

Expected frequency (f_e) for the top-left cell

$$f_e = \frac{Row\ marginal \times Column\ marginal}{n} = \frac{40 \times 55}{100} = 22$$

Source: Healey 2015, p.281.

Computational table

(1)	(2)	(3)	(4)	(5)
f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
30	22	8	64	2.91
10	18	−8	64	3.56
25	33	−8	64	1.94
35	27	8	64	2.37
100	100	0		10.78

- χ^2(*obtained*) = 10.78

Source: Healey 2015, p.281.

Step 5: Decision, interpret

- χ^2*(obtained)* = 10.78
  - This is beyond χ^2*(critical)* = 3.841
  - The obtained χ^2 score falls in the critical region, so we ***reject*** the H_0
  - Therefore, the H_0 is false and must be rejected

- There is a significant relationship between employment status and accreditation status in the population from which the sample was drawn

Interpreting chi square

- The chi square test tells us only if the variables are independent or not
- It does not tell us the pattern or nature of the relationship
- To investigate the pattern, compute percentages within each column and compare across the columns

Limitations of chi square

- Difficult to interpret
  - When variables have many categories
  - Best when variables have four or fewer categories
- With small sample size (*n*)
  - We cannot assume that chi square sampling distribution will be accurate
  - Small samples: High percentage of cells have expected frequencies of 5 or less
- Like all tests of hypotheses
  - Chi square is sensitive to sample size
  - As *n* increases, obtained chi square increases
  - Large samples: Trivial relationships may be significant
- Statistical significance is not the same as substantive significance (importance, magnitude)

GSS example

- Is opinion about immigration different by sex?
- The probability of not rejecting H_0 is big ($p>0.05$)
  - Opinion about immigration does not depend on respondent's sex

```
. tab letin1 sex if year==2016, chi col
```

Key
frequency
column percentage

number of immigrants to america nowadays should be	respondents sex male	female	Total
increased a lot	49 5.98	59 5.75	108 5.85
increased a little	104 12.70	114 11.11	218 11.82
remain the same as it	329 40.17	413 40.25	742 40.22
reduced a little	181 22.10	238 23.20	419 22.71
reduced a lot	156 19.05	202 19.69	358 19.40
Total	819 100.00	1,026 100.00	1,845 100.00

Pearson chi2(4) = 1.3515 Pr = 0.853

Source: 2016 General Social Survey.

Edited table

Table 1. Opinion of the U.S. adult population about how should the number of immigrants to the country be nowadays by sex, 2004, 2010, and 2016

Opinion About Number of Immigrants	Male (%)	Female (%)	Total (%)	Chi Square (df = 4)	p-value
2004				2.3397	0.6740
Increase a lot	3.17	4.30	3.78		
Increase a little	6.89	6.27	6.56		
Remain the same	35.01	34.05	34.49		
Reduce a little	27.68	28.72	28.24		
Reduce a lot	27.24	26.66	26.93		
Total	**100.00**	**100.00**	**100.00**		
(sample size)	**(914)**	**(1,069)**	**(1,983)**		
2010				7.0998	0.1310
Increase a lot	5.21	3.88	4.45		
Increase a little	7.90	11.40	9.91		
Remain the same	35.29	34.96	35.10		
Reduce a little	24.03	25.31	24.77		
Reduce a lot	27.56	24.44	25.77		
Total	**100.00**	**100.00**	**100.00**		
(sample size)	**(595)**	**(798)**	**(1,393)**		
2016				1.3515	0.8530
Increase a lot	5.98	5.75	5.85		
Increase a little	12.70	11.11	11.82		
Remain the same	40.17	40.25	40.22		
Reduce a little	22.10	23.20	22.71		
Reduce a lot	19.05	19.69	19.40		
Total	**100.00**	**100.00**	**100.00**		
(sample size)	**(819)**	**(1,026)**	**(1,845)**		

Source: 2004, 2010, 2016 General Social Surveys.

ACS example

- Does education attainment vary by race/ethnicity?
  - The probability of not rejecting H_0 is small ($p<0.01$)
  - Education attainment is dependent on race/ethnicity

```
. tab educgr raceth [fweight=perwt], col nofreq
```

educgr	raceth White	African A	Hispanic	Asian	Native Am	Ohter rac	Total
Less than high school	23.19	30.14	49.76	27.23	20.66	47.04	35.24
High school	26.55	29.72	26.11	16.23	34.00	17.85	26.09
Some college	20.38	22.79	14.40	12.29	25.15	16.42	17.82
College	19.92	11.04	7.12	23.26	15.36	12.51	13.78
Graduate school	9.95	6.31	2.62	20.99	4.83	6.17	7.07
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

```
. svy: tab educgr raceth, col
(running tabulate on estimation sample)

Number of strata =     212                 Number of obs   =    272,776
Number of PSUs   = 114,016                 Population size = 28,995,881
                                           Design df       =    113,804
Pearson:
  Uncorrected   chi2(20)          =  3.03e+04
  Design-based  F(19.11,  2.2e+06)=  676.9183    P = 0.0000
```

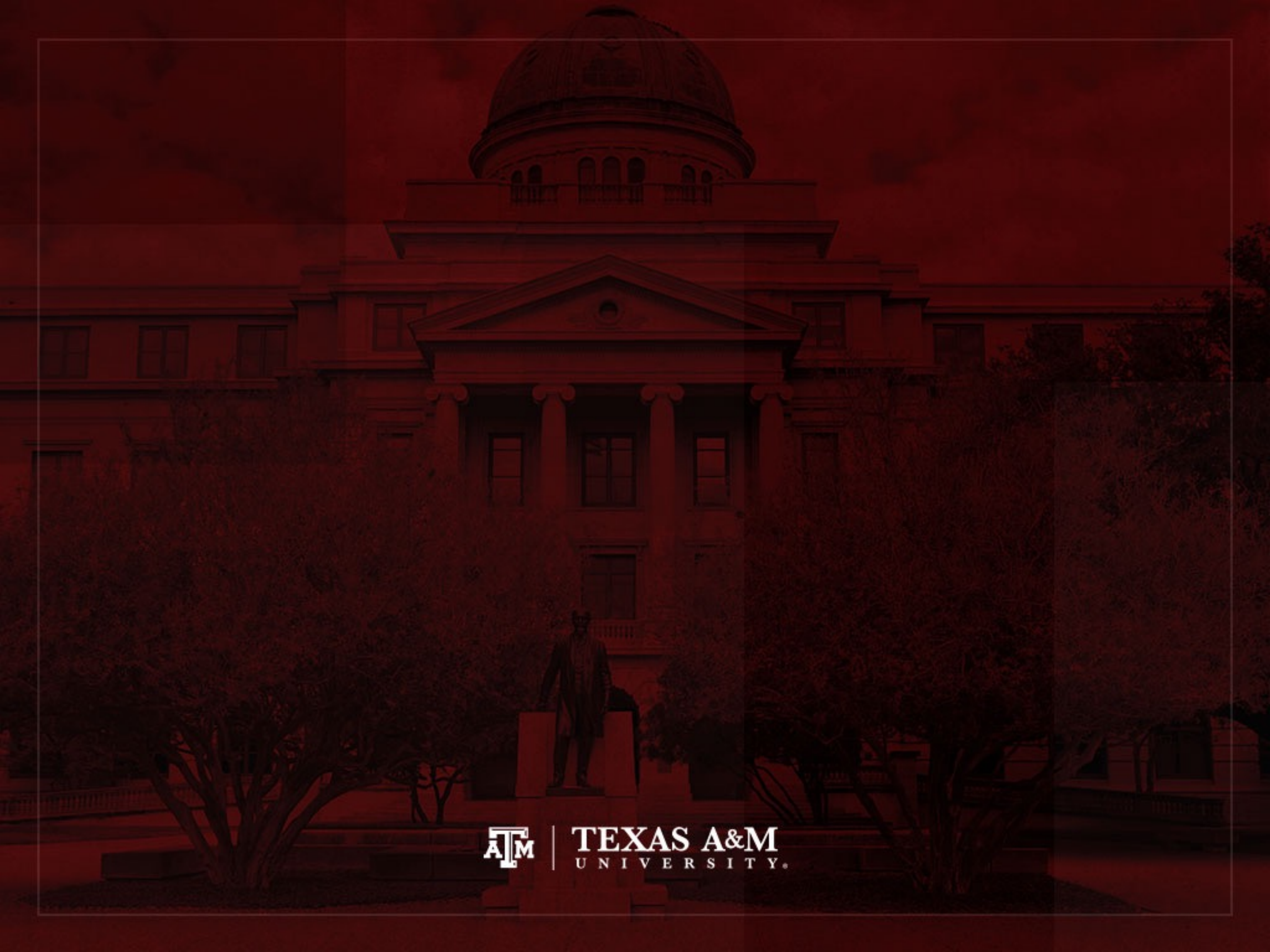
Source: 2019 American Community Survey, Texas.

Edited table

Table 1. Percentage distribution of population by educational attainment and race/ethnicity, Texas, 2019

Educational attainment	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Native American	Other races	Total
Less than high school	23.19	30.14	49.76	27.23	20.66	47.04	35.24
High school	26.55	29.72	26.11	16.23	34.00	17.85	26.09
Some college	20.38	22.79	14.40	12.29	25.15	16.42	17.82
College	19.92	11.04	7.12	23.26	15.36	12.51	13.78
Graduate school	9.95	6.31	2.62	20.99	4.83	6.17	7.07
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Population size (*N*)	11,929,840	3,445,104	11,527,412	1,444,220	79,394	569,911	28,995,881
Chi square (*df* = 20)	3.03e+04						
Design-based *F* (19.11, 2.2e+06)	676.92						
p-value	0.0000						

Source: 2019 American Community Survey.


TEXAS A&M
UNIVERSITY