Lecture 11: Chi square

Ernesto F. L. Amaral

March 25–April 03, 2024
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 11 (pp. 276–306).

TEXAS A&M UNIVERSITY

Outline

- 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
  - $df = (r-1)(c-1) = (2-1)(2-1) = 1$
- χ^2(*critical*) = 3.841
  - If the probability (*p*-value) is less than 0.05
  - χ^2(*obtained*) will be beyond χ^2(*critical*)

Step 4: Test statistic

Observed frequencies

Employment Status	Accreditation Status: *Accredited*	Accreditation Status: *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

Employment Status	Accreditation Status: *Accredited*	Accreditation Status: *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
```

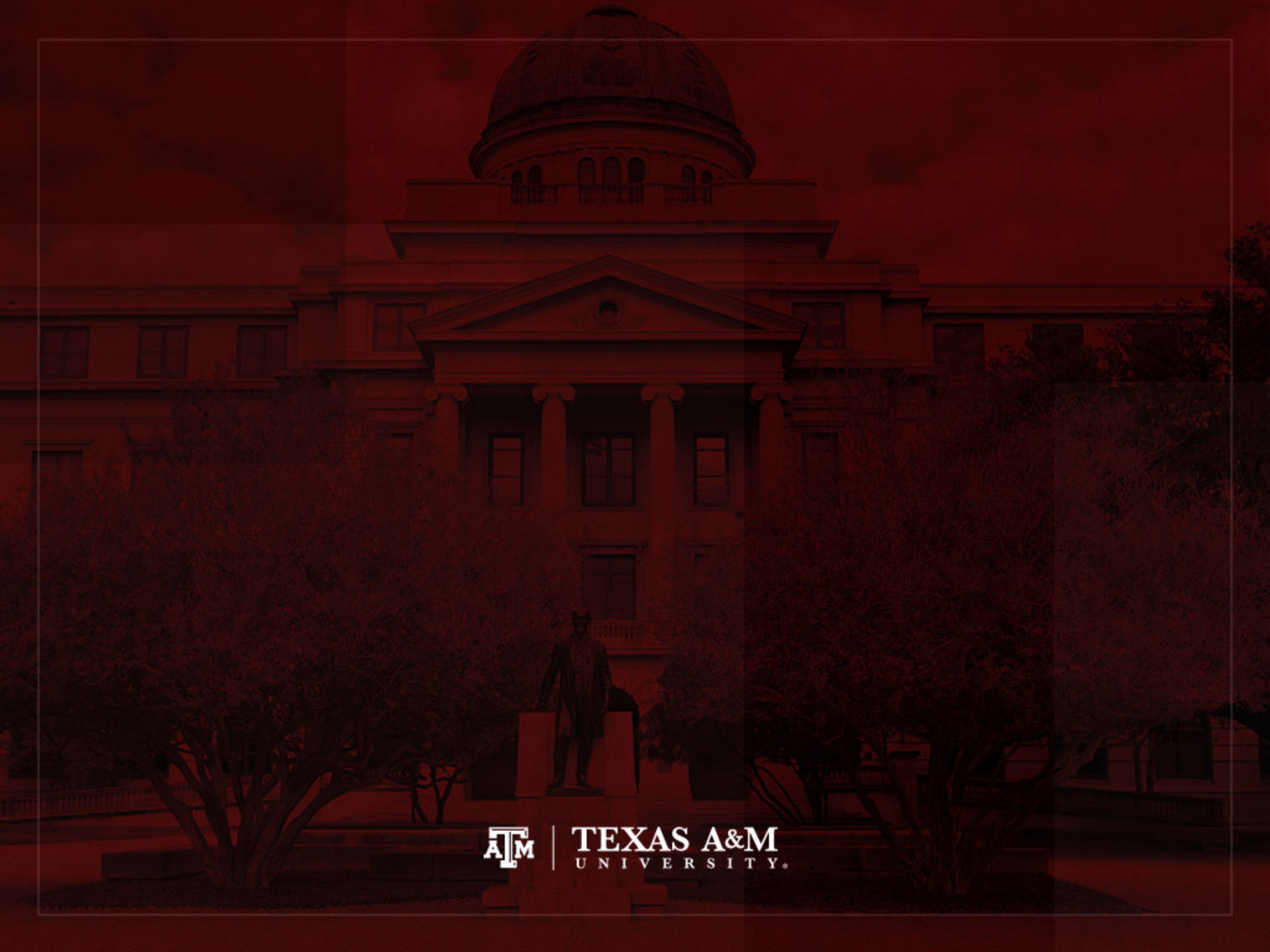
Source: 2019 American Community Survey, Texas.

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.


ĀTM
TEXAS A&M
UNIVERSITY