Lecture (chapter 12): Bivariate association for nominal- and ordinal-level variables

Ernesto F. L. Amaral

November 5–7, 2018
Advanced Methods of Social Research (SOCI 420)

Source: Healey, Joseph F. 2015. ”Statistics: A Tool for Social Research.” Stamford: Cengage Learning. 10th edition. Chapter 12 (pp. 308–341).

TEXAS A&M UNIVERSITY

Chapter learning objectives

- Use measures of association to describe and analyze the importance (magnitude) vs. statistical significance of a bivariate correlation
- Define association in the context of bivariate tables
- List and explain the three characteristics of a bivariate correlation: (a) does it exist? (b) how strong is it? and (c) what is the pattern or direction of the association?
- Assess the association of variables in a bivariate table by: (a) calculating and interpreting column percentages and (b) computing and interpreting an appropriate measure of association
- Compute and interpret Spearman's rho, a measure of association for "continuous" ordinal-level variables

Basic concepts

- Two variables are said to be associated when they vary together, when one changes as the other changes
- Association can be important evidence for causal relationships, particularly if the association is strong
- If variables are associated, the score on one variable can be predicted from the score of the other variable
- The stronger the association, the more accurate the predictions
- Read the table from column to column, noting the differences across the “within-column” frequency distributions

Bivariate association

- Bivariate association can be investigated by finding answers to three questions
  1. Does an association exist?
  2. How strong is the association?
  3. What is the pattern and/or direction of the association?

Productivity by Job Satisfaction (frequencies)

	Job Satisfaction (X)			
Productivity (Y)	*Low*	*Moderate*	*High*	TOTALS
Low	30	21	7	58
Moderate	20	25	18	63
High	10	15	27	52
TOTALS	60	61	52	173

Source: Healey 2015, p.310.

Bivariate tables

- Most general rules
  - Calculate percentages within the categories of the independent variable
  - Compare percentages across the categories of the independent variable
- When independent variable is the column variable (as is generally the case, but not always)
  - Calculate percentages within the columns (vertically)
  - Compare percentages across the columns (horizontally)
- Briefest version
  - Percentage down
  - Compare across

Percentages

- To detect association within bivariate tables (assuming the column variable is the independent variable)
  - Compute percentages within the columns (vertically)
  - Compare percentages across the columns (horizontally)

Productivity by Job Satisfaction (percentages)

	Job Satisfaction (*X*)			
Productivity (*Y*)	*Low*	*Moderate*	*High*	TOTALS
Low	50.0%	34.4%	13.5%	33.5% (58)
Moderate	33.3%	41.0%	34.6%	36.4% (63)
High	16.7%	24.6%	51.9%	30.1% (52)
TOTALS	100.0% (60)	100.0% (61)	100.0% (52)	100.0% (173)

Source: Healey 2015, p.311.

1. Is there an association?

- An association exists if the conditional distributions of one variable change across the values of the other variable
- With bivariate tables, column percentages are the conditional distributions of Y for each value of X
- If the column percentages change, the variables are associated

2. How strong is the association?

- The stronger the correlation, the greater the change in column percentages (or conditional distributions)
- In weak correlations, there is little or no change in column percentages
- In strong correlations, there is marked change in column percentages

Maximum difference

- One way to measure strength is to find the "maximum difference"
  - The biggest difference in column percentages for any row of the table
  - This is a "quick and easy" method: easy to apply but of limited usefulness

The Relationship Between the Maximum Difference and the Strength of the Relationship

Maximum Difference	Strength
If the maximum difference is	*The strength of the relationship is*
Between 0 and 10 percentage points	Weak
Between 10 and 30 percentage points	Moderate
More than 30 percentage points	Strong

Source: Healey 2015, p.313.

Measures of association

- It is always useful to compute column percentages for bivariate tables
- It is also useful to have a summary measure (a single number) to indicate the strength of the association
- For nominal level variables, there are two commonly used measures of association
  - Chi Square based measures
    - Phi (ϕ) or Cramer's V
  - Proportional Reduction in Error (PRE) measure
    - Lambda (λ)

Phi (ϕ)

- Phi (ϕ) is the square root of chi square divided by *N*
- For 2 x 2 tables
- Ranges from 0.0 to 1.0

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Cramer's *V*

- Cramer's *V*
- For tables larger than 2 x 2
- Ranges from 0.0 to 1.0

$$V = \sqrt{\frac{\chi^2}{N(\min r - 1, c - 1)}}$$

1. Find the number of rows (*r*) and the number of columns (c) in the table. Subtract 1 from the lesser of these two numbers to find (min *r* – 1, *c* – 1)
2. Multiply the value you found in step 1 by *N*
3. Divide the value of chi square by the value you found in step 2
4. Take the square root of the quantity you found in step 3

Limitations

- Limitations of Chi Square based measures
- Phi and Cramer's *V* measure only the strength of the association
  - They do not identify the pattern/direction
- To assess pattern/direction, interpret the column percentages in the bivariate table
- Phi and *V* do not provide a true statistical interpretation
  - All we can say is whether the association is weak, moderate, or strong based on the value

Interpretation of strength

- To interpret the strength of an association using Phi or Cramer's *V* (Chi Square based measures), follow these guidelines

Guidelines for Interpreting the Strength of the Relationship for Nominal-Level Measures of Association

Measure of Association	Strength
If the value is	*The strength of the relationship is*
Between 0.00 and 0.10	Weak
Between 0.11 and 0.30	Moderate
Greater than 0.30	Strong

Source: Healey 2015, p.318.

PRE measures

- The logic of Proportional Reduction in Error (PRE) measures is based on two predictions
  - First prediction, E_1: How many errors in predicting the value of the dependent variable (Y) do we make if we **<u>ignore</u>** information about the independent variable (X)
  - Second prediction, E_2: How many errors in predicting the value of the dependent variable (Y) do we make if we take the independent variable (X) into account
- If the variables are associated, we should make fewer errors of the second kind (E_2) than we make of the first kind (E_1)

Lambda (λ)

- Like Phi and Cramer's *V*
  - Lambda (λ) is used to measure the **strength** of the association between nominal variables in bivariate tables
- Unlike Phi and Cramer's *V*
  - Lambda is a PRE measure and its value has a more **direct interpretation**
  - Phi and Cramer's *V* are only indexes of strength
  - Lambda tells us the improvement in predicting Y while taking X into account

Calculate Lambda (λ)

- To compute Lambda, find E_1 and E_2
- $E_1 = N$ – (largest row total)
- E_2 = for each column, subtract the largest cell frequency from the column total, then sum

$$\lambda = \frac{(E_1 - E_2)}{E_1}$$

Characteristics of Lambda (λ)

- Lambda is asymmetric
  - The value will vary depending on which variable is independent
- When row totals are very unequal, Lambda can be zero even when there is an association between the variables
  - For very unequal row marginals, it's better to use a Chi Square based measure of association

Limitations of Lambda (λ)

- Lambda gives an indication of the strength of the association only
- It does not give information about pattern
- To analyze the pattern of the association, use column percentages in the bivariate table

3. Pattern of the association

- Which scores of the variables go together?
- To detect, find the cell in each column which has the highest column percentage
- If both variables are ordinal, we can discuss the "direction" as well
  - In positive associations, the variables vary in the same direction
    - As one variable increases, the other variable increases
  - In negative associations, the variables vary in opposite directions
    - As one variable increases, the other variable decreases

Example of Phi, *V*, *λ*

- Various supervisors in the city government of Shinbone, Kansas, have been rated on the extent to which they practice authoritarian styles of leadership and decision making
- Efficiency of each department has also been rated

Efficiency	Authoritarianism		Total
	Low	High	
Low	10	12	22
High	17	5	22
Total	27	17	44

Source: Healey 2015, p.336, problem 12.7.

1. Is there an association?

- Calculate the column percentages taking each cell frequency, dividing by the column total, and multiplying by 100
- The column percentages show the efficiency of workers (Y) by the authoritarianism of supervisor (X)
- The column percentages change (differ across columns), so these variables are associated

Efficiency	**Authoritarianism**		**Total**
	Low	**High**	
Low	10 (*37.04%*)	12 (*70.59%*)	22
High	17 (*62.96%*)	5 (*29.41%*)	22
Total	27 (*100.00%*)	17 (*100.00%*)	44

Source: Healey 2015, p.336, problem 12.7.

2. How strong is the association?

- The “maximum difference” is 33.55% (70.59%–37.04%)
- This indicates a “strong” association

Efficiency	Authoritarianism		Total
	Low	High	
Low	10 (*37.04%*)	12 (*70.59%*)	22
High	17 (*62.96%*)	5 (*29.41%*)	22
Total	27 (*100.00%*)	17 (*100.00%*)	44

Source: Healey 2015, p.336, problem 12.7.

Phi

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{4.70}{44}} = 0.33$$

- Phi = 0.33
- This indicates a “strong” association

Source: Healey 2015, p.336, problem 12.7.

Cramer's *V*

$$V = \sqrt{\frac{\chi^2}{N(\min r - 1, c - 1)}} = \sqrt{\frac{4.70}{44(2-1)}} = 0.33$$

- Cramer's *V* = 0.33
- This indicates a "strong" association

Source: Healey 2015, p.336, problem 12.7.

Lambda

- E_1 = N – largest row total = 44 – 22 = 22
- E_2 = for each column, subtract largest cell frequency from the column total = (27 – 17) + (17 – 12) = 15

$$\lambda = \frac{(E_1 - E_2)}{E_1} = \frac{22 - 15}{22} = 0.32$$

- Lambda = 0.32
- We reduce our error in predicting the dependent variable by 32% when we take the independent variable into account

Source: Healey 2015, p.336, problem 12.7.

3. Pattern of the association

- Low on authoritarianism goes with high on efficiency
- High on authoritarianism goes with low on efficiency
- Therefore, the association is negative: as authoritarianism increases, efficiency decreases

Efficiency	**Authoritarianism**		**Total**
	Low	**High**	
Low	10 (*37.04%*)	12 (*70.59%*)	22
High	17 (*62.96%*)	5 (*29.41%*)	22
Total	27 (*100.00%*)	17 (*100.00%*)	44

Source: Healey 2015, p.336, problem 12.7.

Two types of ordinal variables

- Collapsed ordinal variables
  - Have just a few values or scores
  - Use Gamma (G)
  - e.g., social class measured as lower, middle, upper
- Continuous ordinal variables
  - Have many possible scores
  - Resemble interval-ratio level variables
  - Use Spearman's Rho (r_s)
  - e.g., scale measuring attitudes toward handgun control with scores ranging from 0 to 20

Gamma

- Gamma is used to measure the **strength** and **direction** of the association
  - Between two ordinal level variables that have been arrayed in a bivariate table
  - Gamma is based on pairs of cases
- Gamma (like Lambda)
  - Tells us the extent to which knowledge of one variable improves our ability to predict the other variable
  - Gamma predicts the order of pairs of cases
  - If two variables are related, the order of pairs on the dependent variable (Y) is predictable from their order on the independent variable (X)
- Before computing and interpreting Gamma, it will always be useful to find and interpret the column percentages

Calculate Gamma

- To compute Gamma, two quantities must be found
  - N_s is the number of pairs of cases ranked in the **same order** on both variables
  - N_d is the number of pairs of cases ranked in **different order** on the variables
  - Always make sure the “low-low” cell is the “top-left” cell in your table before calculation

$$G = \frac{N_s - N_d}{N_s + N_d}$$

Interpretation of Gamma

- The PRE interpretation refers
  - To the percentage of fewer errors made in predicting the order of pairs on the dependent variable (Y) from the order of pairs on the independent variable (X)
  - Compared to the number of errors made in predicting the order of pairs on the dependent variable (Y) while **ignoring** the independent variable (X)

Guidelines for Interpreting the Strength of the Relationship for Ordinal-Level Measures of Association

Measure of Association	Strength
If the value is	*The strength of the relationship is*
Between 0.00 and 0.30	Weak
Between 0.31 and 0.60	Moderate
Greater than 0.60	Strong

Source: Healey 2015, p.326.

Gamma: Strength and direction

- In addition to strength, gamma also identifies the direction of the association
- In a negative association, the variables change in different directions
  - e.g., as age increases, income decreases (or, as age decreases, income increases)
- In a positive association, the variables change in the same direction
  - e.g., as education increases, income increases (or, as education decreases, income decreases)

Example of Gamma: N_s

Efficiency	Authoritarianism: Low	Authoritarianism: High
Low	**10**	12
High	17	**5**

- To compute N_s, multiply each cell frequency by all cell frequencies ***below and to the right***
- $N_s = 10 \times 5 = 50$
- Regardless of how many cells a table has, this procedure is the same

Source: Healey 2015, p.336, problem 12.7.

Example of Gamma: N_d

Efficiency	Authoritarianism	
	Low	**High**
Low	10	**12**
High	**17**	5

- To compute N_d, multiply each cell frequency by all cell frequencies ***below and to the left***
- $N_d = 12 \times 17 = 204$
- This procedure is the same for any size table

Source: Healey 2015, p.336, problem 12.7.

Calculate Gamma

$$G = \frac{N_s - N_d}{N_s + N_d} = \frac{50 - 204}{50 + 204} = -0.61$$

Efficiency	Authoritarianism	
	Low	**High**
Low	10	12
High	17	5

Source: Healey 2015, p.336, problem 12.7.

Interpretation of direction

- Gamma = –0.61
- Gamma is negative, so the association between authoritarianism and efficiency is negative
- As one variable decreases the other variable increases

Source: Healey 2015, p.336, problem 12.7.

Interpretation of strength

- Gamma = –0.61
- The absolute value of Gamma is 0.61
  - According to the guideline table this indicates a strong association
- PRE interpretation
  - We would make 61% fewer errors if we predicted the order of pairs on efficiency (Y) from the order of pairs on authoritarianism (X)
  - Compared to predicting the order of pairs on efficiency (Y) while **ignoring** authoritarianism (X)

Source: Healey 2015, p.336, problem 12.7.

Spearman's Rho (r_s)

- Measure of association for ordinal-level variables with a broad range of different scores and few ties between cases on either variable
- Computing Spearman's Rho
  1. Rank cases from high to low on each variable
  2. Use ranks, not the scores, to calculate Rho

$$r_s = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

where $\sum D^2$ is the sum of the squared differences in ranks

Example of Spearman's Rho (r_s)

Scores on Involvement in Jogging and Self-Esteem

Jogger	Involvement in Jogging (X)	Self-Esteem (Y)
Wendy	18	15
Debbie	17	18
Phyllis	15	12
Stacey	12	16
Evelyn	10	6
Tricia	9	10
Christy	8	8
Patsy	8	7
Marsha	5	5
Lynn	1	2

Source: Healey 2015, p.329.

Computing Spearman's Rho (r_s)

Computing Spearman's Rho

	Involvement (X)	Rank	Self-Image (Y)	Rank	D	D^2
Wendy	18	1	15	3	−2	4
Debbie	17	2	18	1	1	1
Phyllis	15	3	12	4	−1	1
Stacey	12	4	16	2	2	4
Evelyn	10	5	6	8	−3	9
Tricia	9	6	10	5	1	1
Christy	8	7.5	8	6	1.5	2.25
Patsy	8	7.5	7	7	0.5	0.25
Marsha	5	9	5	9	0	0
Lynn	1	10	2	10	0	0
					$\Sigma D = 0$	$\Sigma D^2 = 22.5$

Source: Healey 2015, p.330.

Result of Spearman's Rho (r_s)

- In the column headed D^2, each difference is squared to eliminate negative signs
- The sum of this column is $\sum D^2$, and this quantity is entered directly into the formula

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(22.5)}{10(100-1)} = 0.86$$

Interpreting Spearman's Rho (r_s)

- Rho is positive, therefore jogging and self-image share a positive association
  - As jogging rank increases, self-image rank also increases
- On its own, Rho does not have a good strength interpretation
  - But Rho^2 is a PRE measure
  - For this example, $Rho^2 = (0.86)^2 = 0.74$
  - We would make 74% fewer errors if we used the rank of jogging (X) to predict the rank on self-image (Y) compared to if we ignored the rank on jogging

GSS example

- Is opinion about immigration different by sex?

```
. svy: tab letin1 sex if year==2016, col
(running tabulate on estimation sample)

Number of strata   =        65                  Number of obs     =      1,845
Number of PSUs     =       130                  Population size   = 1,841.4241
                                                Design df         =         65

──────────────────────────────────────
number of │
immigrant │
s to      │
america   │
nowadays  │    respondents sex
should be │   male  female   Total
──────────┼───────────────────────────
 increase │   .056   .0607   .0586
 increase │   .122   .1115   .1163
 remain t │  .4108   .3961   .4028
  reduced │  .2241    .236   .2305
  reduced │  .1871   .1957   .1918
          │
    Total │      1       1       1
──────────────────────────────────────
  Key:  column proportion
```

```
***Commands for measures of association:

***Lambda
lambda letin1 sex if year==2016

***Chi square, Cramer's V, Gamma
tab letin1 sex if year==2016, chi V gamma

***Test statistic for Gamma: Z = gamma / ASE
di 0.0321/0.035 // test statistic
di 1-normal(0.91714286) // p-value

***Spearman's rank correlation coefficient
spearman letin1 sex if year==2016
```

Source: 2016 General Social Survey.

Lambda

```
. ***Lambda
. lambda letin1 sex
```

number of immigrants to america nowadays should be	respondents sex male	female	Total
increased a lot	49	59	108
increased a little	104	114	218
remain the same as it	329	413	742
reduced a little	181	238	419
reduced a lot	156	202	358
Total	819	1,026	1,845

```
lambda_a    0.0000
lambda_b    0.0000
lambda      0.0000
```

Source: 2016 General Social Survey.

Chi square, Cramer's V, Gamma

```
. ***Chi square, Cramer's V, Gamma
. tab letin1 sex, chi V gamma

number of immigrants |
 to america nowadays |    respondents sex
           should be |      male     female |     Total
---------------------+----------------------+----------
     increased a lot |        49         59 |       108
  increased a little |       104        114 |       218
remain the same as it |      329        413 |       742
    reduced a little |       181        238 |       419
       reduced a lot |       156        202 |       358
---------------------+----------------------+----------
               Total |       819      1,026 |     1,845

          Pearson chi2(4) =   1.3515   Pr = 0.853
               Cramér's V =   0.0271
                    gamma =   0.0321  ASE = 0.035

.
. ***Test statistic for Gamma: Z = gamma / ASE
. di 0.0321/0.035 // test statistic
.91714286

. di 1-normal(0.91714286) // p-value
.17953389
```

Source: 2016 General Social Survey.

Spearman’s rho

```
. ***Spearman's rho (rank correlation coefficient)
. spearman letin1 sex

 Number of obs =     1845
Spearman's rho =        0.0212

Test of Ho: letin1 and sex are independent
    Prob > |t| =        0.3637
```

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

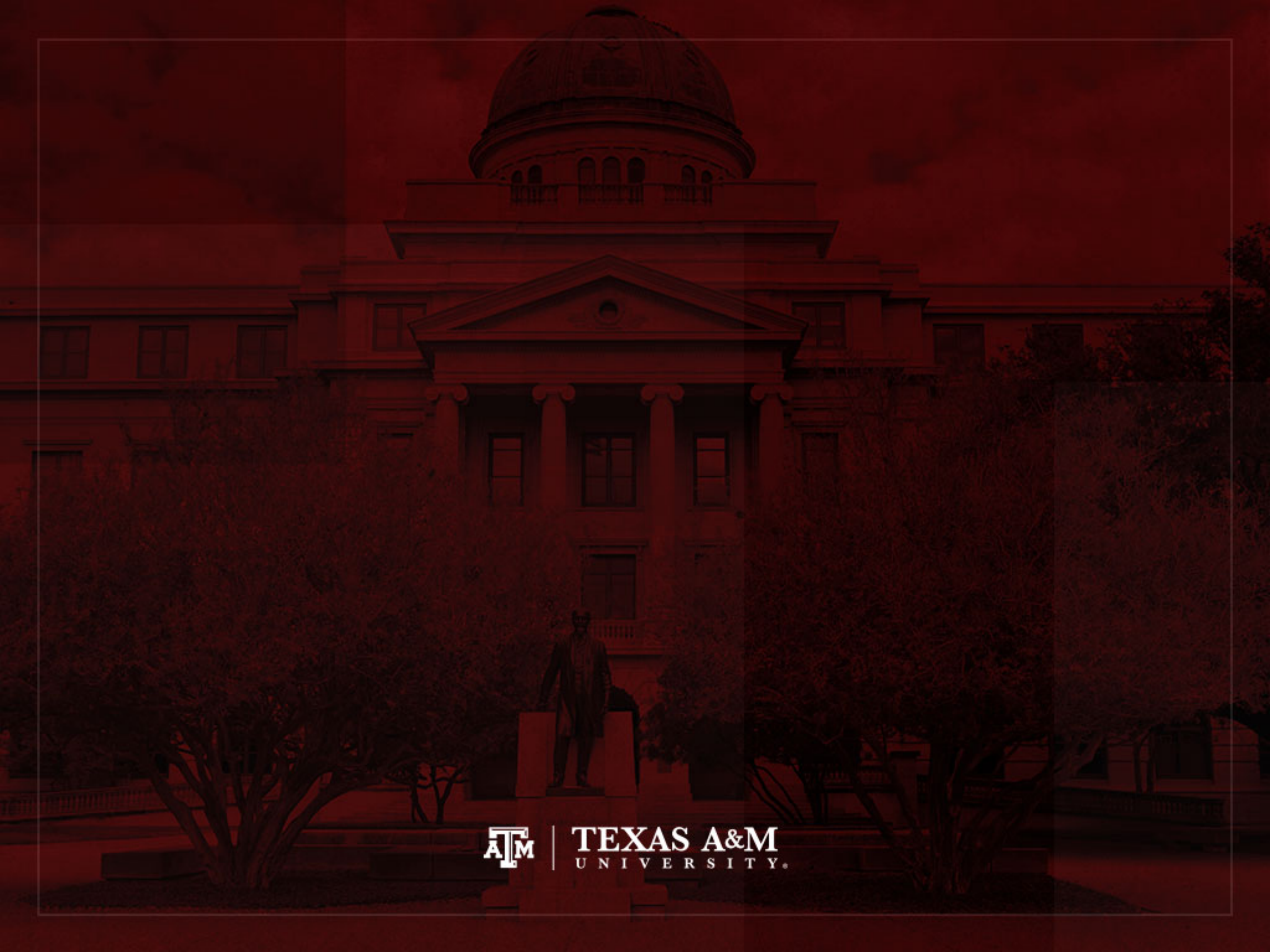
Opinion About Number of Immigrants	Male (%)	Female (%)	Total (%)	Measures of association		p-value
2004						
Increase a lot	3.19	3.74	3.48	Chi square:	2.3397	0.6740
Increase a little	6.55	6.53	6.54	Cramer's V:	0.0343	
Remain the same	36.25	34.22	35.17	Lambda:	0.0000	
Reduce a little	27.61	28.90	28.30	Gamma:	–0.0050	0.4415
Reduce a lot	26.40	26.61	26.51	Spearman's rho:	–0.0032	0.8852
Total (sample size)	**100.00 (914)**	**100.00 (1,069)**	**100.00 (1,983)**			
2010						
Increase a lot	4.84	3.80	4.26	Chi square:	7.0998	0.1310
Increase a little	7.33	11.10	9.44	Cramer's V:	0.0714	
Remain the same	36.44	35.46	35.89	Lambda:	0.0000	
Reduce a little	25.17	24.01	24.52	Gamma:	–0.0472	0.1248
Reduce a lot	26.22	25.62	25.88	Spearman's rho:	–0.0310	0.2476
Total (sample size)	**100.00 (595)**	**100.00 (798)**	**100.00 (1,393)**			
2016						
Increase a lot	5.60	6.07	5.86	Chi square:	1.3515	0.8530
Increase a little	12.20	11.15	11.63	Cramer's V:	0.0271	
Remain the same	41.08	39.61	40.28	Lambda:	0.0000	
Reduce a little	22.41	23.60	23.05	Gamma:	0.0321	0.1795
Reduce a lot	18.71	19.57	19.18	Spearman's rho:	0.0212	0.3637
Total (sample size)	**100.00 (819)**	**100.00 (1,026)**	**100.00 (1,845)**			

Note: Column percentages were estimated taking into account the complex survey design of the General Social Survey.
Source: 2004, 2010, 2016 General Social Surveys.

Correlation vs. causation

- Correlation and causation are different
  - Strong associations (correlation) may be used as evidence of causal relationships (causation)
  - Associations do not prove variables are causally related
- We might have problems of reverse causality
  - e.g., immigration increases competition in the labor market and affects earnings
  - Availability of jobs and income levels influence migration

Migration $\longleftrightarrow$ **Earnings**


TEXAS A&M
UNIVERSITY