Module 2 – Section 5

Measures of Association

Overview

- φ Coefficient and Cramer's V
- Goodman-Kruskal γ

φ Coefficient

- Variable 1
  - $I = 2$ categories
  - Categories = (Yes, No) or (Success, Failure)
- Variable 2
  - $J = 2$ categories
  - Categories = (Yes, No) or (Success, Failure)

Population Proportions

	Variable 2		
Variable 1	Success (Yes)	Failure (No)	Total
Success (Yes)	p_{11}	p_{12}	$p_{1.}$
Failure (No)	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

φ Coefficient

- Population Correlation Coefficient

$$\varphi = \frac{p_{11} - p_{1.}p_{.1}}{\sqrt{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})}}$$

$$= \frac{p_{11} - p_{1.}p_{.1}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}}$$

Properties of φ Coefficient

- If two variables are independent:
  - $\varphi = 0$
- $p_{12} = 0$ and $p_{21} = 0$
  - $\varphi = 1$
- $p_{11} = 0$ and $p_{22} = 0$
  - $\varphi = -1$

Properties of φ Coefficient

- Minimum possible value for φ is

$$\max\left(-\sqrt{\frac{p_{1.}p_{.1}}{(1-p_{1.})(1-p_{.1})}}, -\sqrt{\frac{(1-p_{1.})(1-p_{.1})}{p_{1.}p_{.1}}}\right)$$

Properties of φ Coefficient

- Maximum possible value for φ is

$$\min\left(\sqrt{\frac{p_{1.}(1-p_{.1})}{p_{.1}(1-p_{1.})}}, \sqrt{\frac{p_{.1}(1-p_{1.})}{p_{1.}(1-p_{.1})}}\right)$$

Contingency Table

	Variable 2		
Variable 1	Success (Yes)	Failure (No)	Total
Success (Yes)	Y_{11}	Y_{12}	$Y_{1.}$
Failure (No)	Y_{21}	Y_{22}	$Y_{2.}$
Total	$Y_{.1}$	$Y_{.2}$	n

Sample Correlation Coefficient

$$r_{\varphi} = \frac{Y_{11}Y_{22} - Y_{12}Y_{21}}{\sqrt{Y_{1.}Y_{2.}Y_{.1}Y_{.2}}} = \mathrm{sign}(Y_{11}Y_{22} - Y_{12}Y_{21})\sqrt{\frac{X^2}{n}}$$

where X^2 is test statistic from test of independence of 2 x 2 table.

Ex. Smoking Study

	Student Smoking Status		
Parent Smoking Status	Non-Smoker	Smoker	Total
Neither	1168	188	1356
One or Both	3203	816	4019
Total	4371	1004	5375

Ex. Smoking Study

- $X^2 = 27.67658$
- p-value < 0.0001
- We have extremely strong evidence the smoking status of students and their parents is not independent.

Ex. Smoking Study

$$
\begin{aligned}
r_\varphi &= \frac{Y_{11}Y_{22} - Y_{12}Y_{21}}{\sqrt{Y_{1.}Y_{2.}Y_{.1}Y_{.2}}} \\
&= \frac{(1168)(816) - (188)(3203)}{\sqrt{(4371)(1004)(1356)(4019)}} \\
&= 0.0718
\end{aligned}
$$

Ex. Smoking Study

$$
\begin{aligned}
r_\varphi &= \operatorname{sign}(Y_{11}Y_{22} - Y_{12}Y_{21})\sqrt{\frac{X^2}{n}} \\
&= +\sqrt{\frac{27.67658}{5375}} \\
&= 0.0718
\end{aligned}
$$

Cramer V

- Variable 1
  - I categories
- Variable 2
  - J categories
- Compare association between different size contingency tables.

Cramer V

- Denoted as φ_C

$$\varphi_C = \sqrt{\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\frac{\left(p_{ij} - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}}{\min(I-1, J-1)}}$$

Properties of Cramer V

- $0 \leq \varphi_C \leq 1$
- $\varphi_C = 0$
  - No association between the two variables
- $\varphi_C = 1$
  - Complete association between the two variables

Estimate of Cramer V

$$\hat{\varphi}_C = \sqrt{\frac{X^2/n}{\min(I-1, J-1)}}$$

Ex. Smoking Study

	Student Smoking Status		
Parent Smoking Status	Non-smoker	Smoker	Total
Neither	1168	188	1356
One	1823	416	2239
Both	1380	400	1780
Total	4371	1004	5375

Ex. Smoking Study

- $X^2 = 37.5663$
- p-value < 0.0001
- We have extremely strong evidence the smoking status of students and their parents is not independent.

Ex. Smoking Study

$$
\begin{aligned}
\hat{\varphi}_C &= \sqrt{\frac{X^2/n}{\min(I-1, J-1)}} \\
&= \sqrt{\frac{37.5663/5375}{1}} \\
&= 0.0836
\end{aligned}
$$

Goodman-Kruskal γ

- Variable 1
  - I ordinal categories
- Variable 2
  - J ordinal categories

Goodman-Kruskal γ

- Is there a “directional” relationship between the ordinal variables?
  - Is a higher (lower) category for one variable associated with a higher (lower) category for the other variable?
  - Is a higher (lower) category for one variable associated with a lower (higher) category for the other variable?

Contingency Table

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Concordant Pairs

- Take a pair of observations (i_1, j_1) and (i_2, j_2)
- Pair of observations are concordant if either:

  $i_1 < i_2$ and $j_1 < j_2$

  or

  $i_1 > i_2$ and $j_1 > j_2$

Concordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Concordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Concordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Concordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Number of Concordant Pairs

$$
\begin{aligned}
P &= Y_{11}(Y_{22} + Y_{23} + Y_{32} + Y_{33}) \\
&\quad + Y_{12}(Y_{23} + Y_{33}) \\
&\quad + Y_{21}(Y_{32} + Y_{33}) \\
&\quad + Y_{22}(Y_{33})
\end{aligned}
$$

Discordant Pairs

- Take a pair of observations (i_1, j_1) and (i_2, j_2)
- Pair of observations are discordant if either:

  $i_1 < i_2$ and $j_1 > j_2$

  or

  $i_1 > i_2$ and $j_1 < j_2$

Discordant Pairs

Variable 1	Variable 2 Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Discordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Discordant Pairs

	Variable 2			
Variable 1	Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Discordant Pairs

Variable 1	Variable 2 Cat 1 (Low)	Cat 2 (Medium)	Cat 3 (High)	Total
Cat 1 (Low)	Y_{11}	Y_{12}	Y_{13}	$Y_{1.}$
Cat 2 (Medium)	Y_{21}	Y_{22}	Y_{23}	$Y_{2.}$
Cat 3 (High)	Y_{31}	Y_{32}	Y_{33}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Number of Discordant Pairs

$$
\begin{aligned}
Q \;= & \; Y_{13}(Y_{21} + Y_{22} + Y_{31} + Y_{32}) \\
& + Y_{12}(Y_{21} + Y_{31}) \\
& + Y_{23}(Y_{31} + Y_{32}) \\
& + Y_{22}(Y_{31})
\end{aligned}
$$

Goodman-Kruskal γ

$$\hat{\gamma} = \frac{P - Q}{P + Q}$$

- Possible values of $\hat{\gamma}$: $-1 < \hat{\gamma} < 1$
- If the two variables are independent: $\hat{\gamma} \approx 0$

Properties of γ

- $\hat{\gamma} > 0$
  - Positive relationship between two variables
- $\hat{\gamma} < 0$
  - Negative Relationship between two variables
- Closer to -1 and 1 indicates "stronger" directional relationship

Ex. Employment Survey

- In 1974, the Danish National Institute for Social Science Research interviewed a random sample of Danes between 20 and 69 years old in order to investigate the general welfare in Denmark. The survey respondents were asked to categorize the physical and psychological demands of their employment. Here are the results for female respondents.

Ex. Employment Survey

Physically Demanding	Psychologically Demanding Seldom	Sometimes	Usually	Total
Seldom	542	179	100	821
Sometimes	179	89	33	301
Usually	202	109	100	411
Total	923	377	233	1533