POLS 5377 Scope & Method of Political Science

Week 15 Measure of Association - 1

Correlation

Healey. (2016) *Statistics: A Tool for Social Research,* Chapter 12 & 13 or 14

2

Key Questions:

- What is the logic of measures of association?
- How to compute a correlation for ordinal variables?
- How to compute a correlation for interval variables?
- What are the limitations of correlation?

3

Outline

- Measures of Association
- Ordinal Variables – Gamma (*G*)
- Ordinal Variables – Spearman rho (r_s)
- Interval Variables – Pearson's *r* (*r*)
- Correlation vs Causation

4

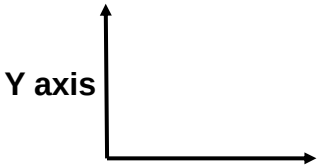
Measures of Association

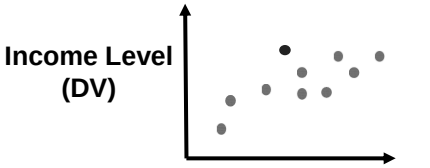
- A measure of association
  - We say two variables are associated, when one variable changes as the other changes.
  - If variables are associated, the score on one variable can be predicted from the score of the other variable.
- When we identify the association between two variables, there are three questions to ask:
  - **Does the association exist?**
  - **How strong is the association?**
  - **What is the direction of the association?**

5

Measures of Association

- We can present the relationship between two variables with a scatter plot
  - X-axis: Independent variable (IV)
  - Y-axis: Dependent variable (DV)
- For example: the relationship between education level and income level
  - Education level: independent variable (X axis)
  - Income level: dependent variable (Y axis)

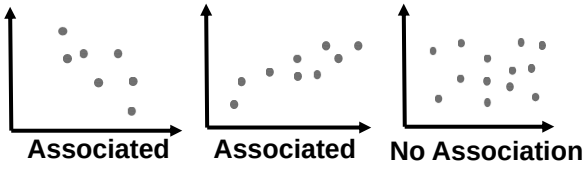




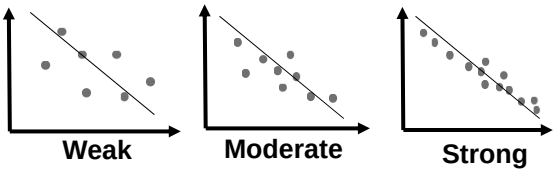
6

Measures of Association

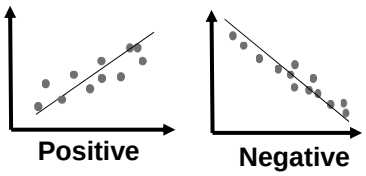
- Does the association exist?



- How strong is the association?



- What is the direction of the association?



7

Measures of Association

- Does the association exist?
  - The change of one variable accompanies with the change of another.
- How strong is the association?
  - Measured by the distance of each case from the trend line.
  - When the total distance is larger, the relationship is weaker.
  - When the distance is smaller, the relationship is stronger.
- What is the direction of the association?
  - Positive: when a variable increase, another variable increases.
  - Negative: when a variable increase, another variable decreases.

8

Measures of Association

- Although scatter plots can effectively present the relationships between two variables, they are not precise.
- Correlation is a measure of association
  - The calculation of Correlation Coefficient can help us measure the relationship between variables with precise statistical values.
- Three measures of association:
  - Ordinal variables with a few categories: **Gamma (*G*)**
  - Ordinal variables with a broad range of scores: **Spearman's rho (r_s)**
  - Interval/ratio variables: **Pearson's *r***

9

Ordinal Variables – Gamma (*G*)

- When measure the association for ORDINAL variables that have a few categories: Gamma (*G*)
  - The statistic of Gamma (*G*) can tell us the **strength** of the association and the **direction** of association between two ordinal variables.
- Remember, in an ordinal variable, the categories can be ranked in sequence.
- To measure the association of ranked variables, we ask: if a case ranks higher than another case on one variable, does it also rank higher on the other variable?
  - When a person A is ranked higher in education level than another person B, does A also rank higher on the income level than B?

10

Ordinal Variables – Gamma (*G*)

- Example of Education Level vs. Income Level

Income Level	Education Level		Total
	LOW	HIGH	
LOW	25 (52.1%)	20 (38.5%)	45
HIGH	23 (47.9%)	32 (61.5%)	55
Total	48 (100%)	52 (100%)	100

  - When compare the case from top-left cell (Low/Low) to the case from lower-right cell (High/High), we will find the two cases ranked the **same order** on both variables.
  - If we compare the person from top-right (High/Low) to the person from lower-left cell (Low/High), we will find the two cases ranked in **different order**.

11

Ordinal Variables – Gamma (*G*)

- The statistic of Gamma is a computation based on the number of pairs have the same order (N_s) and the numbers of pairs ranked in different order (N_d)

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

  - N_s= the total number of pairs of cases ranked in the same order on both variables
  - N_s= the total number of pairs of cases ranked in different order on both variables

12

Ordinal Variables – Gamma (*G*)

- In the example of education level and income level

Income Level	Education Level		Total
	LOW	HIGH	
LOW	25 (52.1%)	20 (38.5%)	45
HIGH	23 (47.9%)	32 (61.5%)	55
Total	48 (100%)	52 (100%)	100

N_s= (25)(32)=800 (# of paired cases ranked in the same order)

N_d= (20)(23)=460 (# of paired cases ranked in different order)

$$G = \frac{N_s - N_d}{N_s + N_d} = \frac{800 - 460}{800 + 460} = \frac{340}{1260} = +\mathbf{0.27}$$

13

Ordinal Variables – Gamma (*G*)

- Interpret Gamma
  - Direction of association:
    - If there is a more pairs of cases ranked in the same order than in different order ($N_s > N_d$), that means there is a positive relationship between the two variables.
    - "+" 0.27 indicates there is positive relationship between the two variable, education level and income level.
  - Strength of association:
    - if there is a big difference between N_s and N_d, that means the association is strong.
    - +"0.27" indicates the relationship between the two variable is weak

If the value is	Strength of the relationship
Between 0.00 and 0.30	Weak
Between 0.31 and 0.60	Moderate
Greater than 0.60	Strong

14

Ordinal Variables – Spearman's rho (r_s)

- When measure the association for ORDINAL variables that have a broad range of scores: Spearman's rho (r_s)
  - The statistic of Spearman's rho (r_s) can tell us the **strength** and the **direction** of association between two ordinal variables.
- Remember, in a ordinal variable, the categories can be ranked in sequence.
- To measure the association of ranked variables, we ask: if a case ranks high on one variable, does the case also rank high on the other variable?
  - When a person A ranks high in jogging involvement, does A also ranks high on the self-esteem?

15

Ordinal Variables – Spearman's rho (r_s)

- Example: Jogging involvement vs self-esteem (Do joggers have an enhanced sense of 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

16

Ordinal Variables – Spearman's rho (r_s)

- Computing Spearman's Rho
  - Step 1: rank cases from high to low on each variable
  - Step 2: use **RANK**, not the scores to calculate Rho (r_s)

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

$\sum D^2$= the sum of the differences in ranks, the quantity squared

N = number of cases

17

Ordinal Variables – Spearman's rho (r_s)

- Step 1: rank cases from high to low on each variable
  - The highest score ranked as 1
  - If there are identical scores (Christy and Patsy), average the two ranks (7 and 8), and apply the average rank to the both cases.

Difference between Rank1 and Rank 2

	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
					$\Sigma D = 0$	$\Sigma D^2 = 22.5$

18

Ordinal Variables – Spearman's rho (r_s)

- Computing Spearman's Rho
  - According to the result from the last slide, $\sum D^2$=22.5

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

$$r_s = 1 - \frac{6(22.5)}{10(10^2-1)} = 1 - \frac{135}{10(100-1)} = 1 - 0.14 = 0.86$$

$$\boldsymbol{r_s = +0.86}$$

19

Ordinal Variables – Spearman's rho (r_s)

- Interpret Spearman's Rho
  - Direction of association:
    - When Rho (r_s) is positive, there is a positive relationship between the two variables; When Rho (r_s) is negative, there is negative relationship between the two variables.
    - "**+**" 0.86 indicates there is positive relationship between jogging involvement and self-esteem. As jogging involvement rank increases, self-esteem rank also increases.
  - Strength of association:
    - Rho (r_s) ranges from 0 (no association) to ± 1 (perfect association)
    - + "**0.86**" indicates the relationship between the two variable is **strong**

If the value is	Strength of the relationship
Between 0.00 and 0.30	Weak
Between 0.31 and 0.60	Moderate
Greater than 0.60	Strong

20

Interval Variables – Pearson's *r* (*r*)

- When measure the association for INTERVAL variables: Pearson's *r* (*r*)
  - Pearson's *r* can indicate the **strength** and the **direction** of association between two ordinal variables.
- The logic of conducting Pearson's r is to identify the distance between case's scores and the means of the variables.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum(X - \bar{X})^2][\sum(Y - \bar{Y})^2]}}$$

X = independent variable; Y = dependent variable

21

Interval Variables – Pearson's *r* (*r*)

- Example: Number of children vs. Husband housework contribution (When number of children increase, does husband's contribution to housework increase?)

Average # of children ($\bar{X}$)=32/12=2.67

Husband's contribution to housework (hours/week)

Average hours of contribution ($\bar{Y}$)=40/12=3.33

# of children

X	$X-\bar{X}$	Y	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
1	−1.67	1	−2.33	3.89	2.79	5.43
1	−1.67	2	−1.33	2.22	2.79	1.77
1	−1.67	3	−0.33	0.55	2.79	0.11
1	−1.67	5	1.67	−2.79	2.79	2.79
2	−0.67	3	−0.33	0.22	0.45	0.11
2	−0.67	1	−2.33	1.56	0.45	5.43
3	0.33	5	1.67	0.55	0.11	2.79
3	0.33	0	−3.33	−1.10	0.11	11.09
4	1.33	6	2.67	3.55	1.77	7.13
4	1.33	3	−0.33	−0.44	1.77	0.11
5	2.33	7	3.67	8.55	5.43	13.47
5	2.33	4	0.67	1.56	5.43	0.45
32	−0.04	40	0.04	18.32	26.68	50.68

22

Interval Variables – Pearson's *r* (*r*)

- Computing Pearson's *r* (*r*)

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{[\sum(X-\bar{X})^2][\sum(Y-\bar{Y})^2]}}$$

$$r = \frac{18.32}{\sqrt{(26.68)(50.68)}} = \frac{18.32}{\sqrt{1352.14}} = \frac{18.32}{36.77} = 0.50$$

$$\boldsymbol{r = +0.50}$$

23

Interval Variables – Pearson's *r* (*r*)

- Interpret Pearson's *r* (*r*)
  - Direction of association:
    - When Pearson's *r* (*r*) is positive, there is a positive relationship between the two variables; When Pearson's *r* (*r*) is negative, there is negative relationship between the two variables.
    - "**+**" 0.50 indicates there is positive relationship between number of children and husband's housework contribution. As number of children increases, husband's contribution to housework also increases.
  - Strength of association:
    - Pearson's *r* (*r*) ranges from 0 (no association) to ± 1 (perfect association)
    - + "**0.50**" indicates the relationship between the two variable is **moderate**

If the value is	Strength of the relationship
Between 0.00 and 0.30	Weak
Between 0.31 and 0.60	Moderate
Greater than 0.60	Strong

24

Conducting Correlation within SPSS

- Correlation and causation are not the same things
  - Mathematics test score and shoe size are associated, but it didn't mean there is causality between the two variables.
- Strong associations may be used as evidence of causal relationships but they do not prove variables are causally related

25

After this lecture:

You should learn the following key concepts:

- The logic of measures of association
- How to compute a correlation for ordinal variables that with a few categories and with a broad range of scores.
- How to compute a correlation for interval variables.
- The limitation of correlations