Stress-Free Stats

7) Assessing Relationships

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Overview

- Categorical on Interval Variables
- Categorical on Categorical Variables
- χ^2 test
- Interval on Interval Variables
- Third Variables

Introduction

		Dependent Variable	
		Categorical	Interval
Indep. Variable	Categorical	Crosstabs χ^2	Compare means Diff of means test
	Interval	Logits	Correlation / Scatter Regression

Categorical on Interval Variable

Interval DV, Categorical IV

- Does the predominant religion of a country affect its income?
- GDP $<-$ religion
- Compare means

	N	Mean	Std. Dev.	Min	Max
Protestant	30	30321.38	5199.057	22386.6	41245.8
Mixed	18	28380.27	6094.755	22295.1	38826.8
Catholic	48	19612.95	5732.95	10942.8	30669.4

- What do you want to know?

Means difference test Protestant v. Catholic

- $T = \frac{H_a - H_0}{se_{diff}}$; $se_{diff} = \sqrt{se_1^2 + se_2^2}$; $se = \frac{s}{\sqrt{N}}$
- Here:
  $se_{diff} = \sqrt{(5199.057/\sqrt{30})^2 + (5732.95/\sqrt{48})^2} = 1259.2576$
- $T = \frac{10708.43 - 0}{1259.2576} = 8.5037$
- Where is that on the T-distribution?
- Far out! Reject H_0, and conclude that there is a significant difference in income between Protestant and Catholic countries.

Categorical on Categorical Variable

Categorical DV, Categorical IV

- You claim that women are more likely to watch the Academy Awards than men.
- Your friend tells you that he has a male friend who always watches the Oscars, and that you cannot 'generalize'.
- Can you generalize?

Testing Categorical DV on Categorical IV

- Collect data

Obs.	Gender	Watch
1	F	Y
2	F	N
3	M	Y
4	F	N
5	F	Y
6	M	N
7	F	Y
8	M	Y
...	...	...
1004	F	Y

- A bit overwhelming...

Categorical DV, Categorical IV

- Crosstabulation

	Female	Male
Watch	331	170
Don't Watch	210	293

- Would be easy if it were something like this:

	Female	Male
Watch	502	50
Don't Watch	50	402

Categorical DV, Categorical IV

- Need to compare the values of the DV across the IV
- Calculate proportions of columns (IV), and compare across rows (DV)
- Watch out, sometimes DV is in columns, so need to inverse the process

	Female	Male	Total
Watch	331	170	501
	61%	37%	
Don't Watch	210	293	503
	39%	63%	
Total	541	463	1004
	100%	100%	

Categorical DV, Categorical IV

- Are viewers more likely to be female than male?
- Calculate proportions of rows (IV), and compare across columns (DV)

	Female	Male	Total
Watch	331	170	501
	66%	34%	100%
Don't Watch	210	293	503
	42%	48%	100%
Total	541	463	1004

χ^2 test

Testing relationships between categorical variables

- We want to test how cases are dispersed across the dependent variable
- $H_0 =$ every category of the IV should have the same distribution as the total, i.e. the IV does not matter.

Party ID and career crosstabulation

		Law	Politics	Business	Education	Total
Republican	N	6	2	5	1	14
	%	42.9	14.3	35.7	7.1	100
Democrat	N	10	10	2	2	24
	%	41.7	41.7	8.3	8.3	100
Other	N	6	5	7	3	21
	%	28.6	23.8	33.3	14.3	100
Total	N	22	17	14	6	59
	%	37.3	28.8	23.7	10.2	100

χ^2 Test

- To test H_0, we use the χ^2 (read chi-squared) test
- This test compares each observed frequency (fo) with the expected (total) frequency (fe)
  - E.g. if H_0 is correct, 37.3% of the 14 republicans (=5.22) should want to go to into law; and 28.8% of the 14 Republicans (=4.03) should want to go into politics
  - Test: sum the squared differences and divide by the expected frequency for all cells: $\chi^2 = \sum_{i=1}^{N} \frac{(fo_i - fe_i)^2}{fe_i}$; where N=number of cells (12)

Party ID and career crosstabulation

		Law	Politics	Business	Education	Total
Republican	N	6	2	5	1	14
	%	42.9	14.3	35.7	7.1	100
Democrat	N	10	10	2	2	24
	%	41.7	41.7	8.3	8.3	100
Other	N	6	5	7	3	21
	%	28.6	23.8	33.3	14.3	100
Total	N	22	17	14	6	59
	%	37.3	28.8	23.7	10.2	100

χ^2 Test

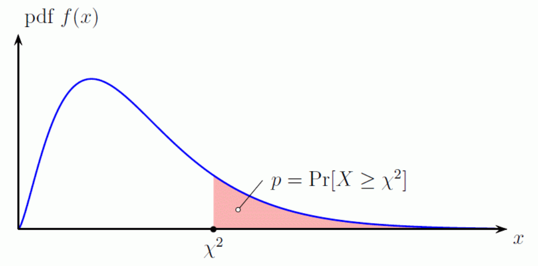
- The χ^2 test: $\chi^2 = \sum_{i=1}^{N} \frac{(fo_i - fe_i)^2}{fe_i} = (6-5.2)^2/5.2 + (2-4.0)^2/4.0 + \ldots = 7.87$
- Apply this value to the χ^2 distribution with appropriate degrees of freedom
- Df=(number of rows - 1)*(number of columns - 1) = (3-1)*(4-1)=6

Party ID and career crosstabulation

		Law	Politics	Business	Education	Total
Republican	N	6	2	5	1	14
	exp N	5.2	4.0	3.3	1.4	14
	%	42.9	14.3	35.7	7.1	100
Democrat	N	10	10	2	2	24
	exp N	8.9	6.9	5.7	2.4	24
	%	41.7	41.7	8.3	8.3	100
Other	N	6	5	7	3	21
	exp N	7.8	6.1	5.0	2.1	21
	%	28.6	23.8	33.3	14.3	100
Total	N	22	17	14	6	59
	%	37.3	28.8	23.7	10.2	100

χ^2 Test

- Our value of χ^2 is 7.78
- What is the critical value of χ^2 at the 0.05 confidence level with 6 df? ▸ Chi2-table
- The answer is 12.592. Our χ^2 is smaller than the critical value, so it is possible that 7.87 could occur more than 5 times out of 100 by random chance.
- We fail to reject H_0; there is no statistically significant difference between party ID and career choice.



Interval on Interval Variable

Measures of Association

- Is a level of one variable associated with the level of another?
- **Sample Covariance**: $Cov(XY) = S_{(XY)} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{N-1}$
- **Sample Correlation**: $Corr(XY) = r_{(XY)} = \frac{\sum(\frac{x_i - \bar{X}}{s_X})(\frac{y_i - \bar{Y}}{s_Y})}{N-1}$
  - Correlation standardizes Covariance by dividing covariance by the standard deviations of X and Y.
  - Hence correlation is bounded between -1 and 1.

Scatterplot

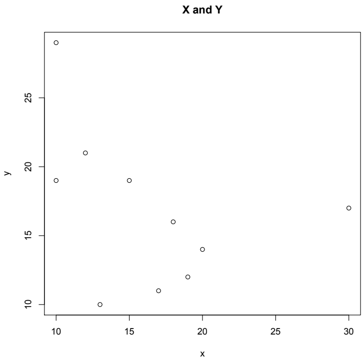


Figure: Little association: $r_{XY} = -0.38$

Scatterplot

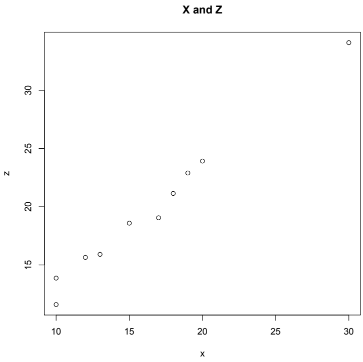


Figure: Strong association: $r_{XZ} = 0.99$

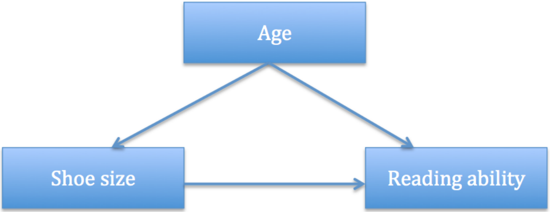
Third Variables

Third Variables

- In reality, we are not just interested in the relationship between two variables
- We want to be sure that the relationship between X and Y takes into account other, potentially intervening, factors.
- How can third variables matter?
  1. Spurious relationships = hidden variable
  2. Multivariate relationships = omitted variable
  3. Conditioned relationships = interaction or moderation

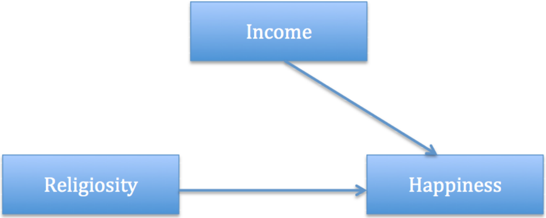
Spurious relationships

- The relationship between X and Y is caused by a hidden third variable Z that causes both X and Y.
- When Z is controlled for, the relationship between X and Y is not significant (not there).
  - Shoe size $\rightarrow$ reading ability
  - Spurious on age
  - If we consider the relationship (shoe size $\rightarrow$ reading ability) *within each age category* (year), relationship disappears.



Multivariate relationships

- The relationship between X and Y stands, but an omitted third variable also causes Y.
- When Z is controlled for, the relationship between X and Y is altered (weakened or strengthened).
  - Religiosity $\rightarrow$ happiness
  - Happiness is also caused by income, and income is correlated with religiosity.
  - If we control for income, the relationship between religiosity and happiness is altered.



Conditioned relationships

- The relationship between X and Y is moderated by a third variable Z.
- The relationship between X and Y changes as the values of Z change.
  - Economic left-right ideology $\rightarrow$ support for EU integration
  - Moderated by country
  - In Britain, the left is supportive of EU integration, while the right is opposed.
  - In Sweden, the left is opposed to EU integration while the right is more supportive...