Correlation Between Continuous & Categorical Variables

CIVL 7012/8012

Association between variables

- Continuous and continuous variable
  - Pearson's correlation coefficient
- Categorical and categorical variable
  - Chi-square test
  - Cramer's V
  - Bonferroni correction
- Categorical and Continuous variable
  - Point biserial correlation

Association between Continuous Variables

- Compute Pearson’s correlation coefficient (r)
  - r<0.3, weak correlation
  - 0.3<r<0.7, moderate correlation
  - r>0.7, high correlation

Association between categorical variables

- Pearson’s correlation coefficient can not be applied.
- What are some of the methods
- How to compute them
- What will be the conclusion

Set up hypothesis

- **Null hypothesis:** Assumes that there is no association between the two variables.
- **Alternative hypothesis:** Assumes that there is an association between the two variables.

Categorical variable

Example: Two categorical variables: marital status and gender

Question: How do we measure degree of association?

Since these are categorical variables Pearson's correlation coefficient will not work

Observed	*Male*	*Female*
Married	456	516
Widowed	58	123
Divorced	142	172
Separated	29	50
Never married	188	207

Reference: https://peterstatistics.com

Pearson Chi-square test for independence

- Calculate estimated values

Expected	Male	Female
Married	437.1747	534.8253
Widowed	81.40804	99.59196
Divorced	141.2272	172.7728
Separated	35.53168	43.46832
Never married	177.6584	217.3416

Observed	*Male*	*Female*
Married	456	516
Widowed	58	123
Divorced	142	172
Separated	29	50
Never married	188	207

$$E_{i,j} = \frac{R_i \times C_j}{N}$$

Calculate chi-sq for each pair

(O-E)2/E	Male	Female
Married	0.810646	0.662634
Widowed	6.730738	5.501811
Divorced	0.004229	0.003457
Separated	1.2007	0.981471
Never married	0.601988	0.492074

$$\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Pearson Chi-square value (sum of all cells): 16.98975

Degrees of freedom and significance

- Degrees of freedom = (r-1) *(c-1)
  - In this example: (5-1)*(2-1) = 4
- Significance: Chi-square (16.98975, 4) = 0.00194

- Reject null hypothesis
- Conclusion: there is an association between the two variables.

Cramer's V (1)

$$\text{Cramer's V} = sqrt\ (\chi^2 / [n(q-1)])$$

- q= min (# of rows, # of columns)
- Cramer's V interpretation
  - 0: The variables are not associated
  - 1: The variables are perfectly associated
  - 0.25: The variables are weakly associated
  - .75: The variables are moderately associated

Cramer's V (2)

- In this case
  - Not associated

Observed>	*Male*	*Female*	Total
Married	456	516	972
Widowed	58	123	181
Divorced	142	172	314
Separated	29	50	79
Never married	188	207	395
Total	873	1068	1941

Pearson Chi-square value:		16.98975
	# of rows (r)	5
	# of cols (c)	2
	q	2
	Cramer's V	0.093558

Bonferroni correction

Observed	*Male*	*Female*	Total
Married	456	516	972
Widowed	58	123	181
Divorced	142	172	314
Separated	29	50	79
Never married	188	207	395
Total	873	1068	1941

Expected	Male	Female
Married	437.1747	534.8253
Widowed	81.40804	99.59196
Divorced	141.2272	172.7728
Separated	35.53168	43.46832
Never married	177.6584	217.3416

Adjusted Residuals

(O-E)2/E	Male	Female
Married	1.717883	-1.71788
Widowed	-3.67295	3.672949
Divorced	0.095753	-0.09575
Separated	-1.50823	1.508229
Never married	1.172004	-1.172

$$\chi^2 Adjusted\ Residual\_i,j = \frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j} * \left(1 - \frac{R_i}{n}\right)\left(1 - \frac{C_j}{n}\right)}}$$

Significance level	0.05
# of tests	10
Adjusted sig level	0.005

Only widowed male and female has significance association

Correlation between continuous and categorial variables

- Point Biserial correlation
  - product-moment correlation in which one variable is continuous and the other variable is binary (dichotomous)
  - Categorical variable does not need to have ordering
  - Assumption: continuous data within each group created by the binary variable are normally distributed with equal variances and possibly different means

Point Biserial correlation

- Suppose you want to find the correlation between
  - a continuous random variable Y and
  - a binary random variable X which takes the values zero and one.
- Assume that n paired observations (Yk, Xk), k = 1, 2, …, n are available.
  - If the common product-moment correlation r is calculated from these data, the resulting correlation is called the point-biserial correlation

Point Biserial correlation

- Point biserial correlation is defined by

$$r_{pb} = \left(\frac{\bar{Y}_1 - \bar{Y}_0}{s_Y}\right)\sqrt{\frac{np_0(1-p_0)}{n-1}}$$

where

$$s_Y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{n-1}}$$

$$\bar{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}$$

$$p_1 = \frac{\sum_{k=1}^{n} X_k}{n}$$

$$p_0 = 1 - p_1$$

Hypothesis test

The hypothesis that ρ = 0 can be tested using the following test which is equivalent to the two-sample t-test.

$$t_{pb} = \frac{r_{pb\sqrt{n-2}}}{\sqrt{1 - r_{pb}^2}}$$

This test statistic follows Student's t distribution with n – 2 degrees of freedom.