Chi-Square

Test for Independence

Course: Statistics 1

Lecturer: Dr. Courtney Pindling

Review: Goodness of Fit

- Uses sample data to test hypothesis about the shape or proportion of a population distribution
- Test how well the sample distribution fits the population distribution specified by H_0
- Null Hypothesis, H_0:
  - ***No Preference****: The proportion is equally divided among the categories* ***or***
  - ***No Difference from Know Population****: The proportion of one population is no different from the proportion of another*

Test for Independence

- Two variables are independent when:
  - there is no consistent, predictable relationship between them
  - The frequency distribution for one sample is not related to (independent to) the categories of the second sample
- *When two variable are independent: for each individual, the value obtained from one variable is not related to (or influenced by) the value of the second variable*
- Null Hypothesis, H_0:
  - ***Version 1:*** *There is no relationship between variables* ***or***
  - ***Version 2:*** *The distributions have equal proportions (same shape)*

Chi-Square Distribution, χ^2

Ch-Square Distribution:

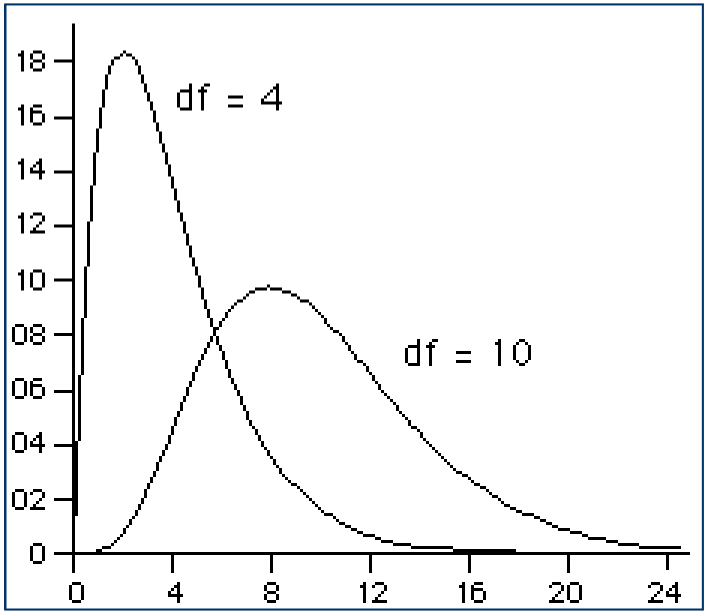
Independent Samples,

df = (***R – 1***)(***C – 1)***,

C is number of columns (variables)

R is number of rows (categories)

1. Shape of Chi-Square depends on ***df***
2. Family of chi-square distributions (*df*)



Frequencies

- **Observed Frequency**, f_o:

  The number of individuals from the sample who are classified in a particular category. Each individual is counted as one-and-only one category

- **Expected Frequency**, f_e:

  For each category, is the frequency value that is predicted from the marginal row and column totals and the sample size (n).

  $f_e = (C \times R)/n$, where C is column total and R is row total (by cell)

Contingency Table

	Variable 1	**Variable 2**	**Variable 3**	***Row Total***
Category A	$f_{A1} = (A)(1)/n$	$f_{A2} = (A)(2)/n$	$f_{A3} = (A)(3)/n$	Category A **Row Total** **(A)**
Category B	$f_{B1} = (B)(1).n$	$f_{B1} = (B)(1)/n$	$f_{B1} = (B)(1)/n$	Category B **Row Total** **(B)**
Column Total	Variable 1 **Column Total** **(1)**	Variable 2 **Column Total** **(2)**	Variable 3 **Column Total** **(3)**	***Grand Total*** $= n$

Sample Test for Independence

- A researcher is investigating the relationship between academic performance (AP: High, Low) and self-esteem (SE: Low, Medium, High). A sample of n = 150 ten-year-old children is obtained and each child is classified by levels of academic performance and self-esteem. The **observed frequency** distribution along with column and row totals are shown below (3 x 2 contingency table).

	High	**Medium**	**Low**	***AP Row Total***
High	*17*	*32*	*11*	60
Low	*13*	*43*	*34*	90
SE Col Total	30	75	45	***n = 150***

AP = Academic Performance and ***SE*** = Self-Esteem

Chi-Square Statistics

Steps to calculate χ^2

1. Find f_e for each variable and category
2. Compute f_0 - f_e and Square the difference
3. Divide Step 1 by f_e
4. Add values from all rows or columns, this is the χ^2 *statistics*

$$\text{chi-square} = \chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Expected Frequency Distribution Table

	High	Medium	Low	*AP Row Total*
High	*(30 x 60) / 150 = 12*	*30*	*18*	60
Low	*18*	*45*	*(45 x 90) / 150 = 27*	90
SE Col Total	30	75	45	***n = 150***

No more than 20% of cell should have f_e less than 5

Chi-Square Statistics

Cell value = $(f_o - f_e)^2 / f_e$

	High	Medium	Low	*Row Total*
High	$(17 - 12)^2 / 12 =$ 2.08	0.13	2.72	4.93
Low	1.39	0.09	$(34 - 27)^2 / 27 =$ 1.81	3.29
Column Total	3.47	0.22	4.53	

$\chi^2 = \mathbf{8.22}$

df = (R – 1)(C – 1) = (2 – 1)(3 – 1) = ***2***

Chi-Square: Critical Value

df	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
1	0.000039	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.01	0.0201	0.0506	0.1026	0.2107	4.61	**5.99**	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.8

- The critical region of the chi-square test is the region above *1- a*; so for a = 0.05 and *df* =(1)(2) = 2, χ^2 = 5.99 ($\chi^2_{0.95}$)

Decision and Conclusion

- Chi-Square statistics of 8.22 > Chi-Square Critical or 8.22 > *5.99* at a = 0.05 level
- Reject H_0 and so

- Conclude that there is a significant relationship between academic performance and self-esteem or there is a significant difference between the distribution of self-esteem for high academic performance versus low academic performance.

Effect Size for 2 x 2 Table

- Cramer Phi coefficient, F_C
- Interpret like Pearson *r*

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

Effect Size for Larger Table

- Cramer's *V*
- *df** is smaller of (R – 1) or (C – 1)
- For example: *df** = 1 and Cramer's *V* = *0.23*
- ***Small effect size***

$$V = \sqrt{\frac{\chi^2}{n(df*)}} = \sqrt{\frac{8.22}{150(1)}} = 0.23$$

Interpreting Cramer's *V*

For *df** = 1 *e.g. V = 0.23 is small*	$0.10 < V < 0.30$ $0.30 < V < 0.50$ $V > 0.50$	Small effect Medium effect Large effect
For *df** = 2	$0.07 < V < 0.21$ $0.21 < V < 0.35$ $V > 0.35$	Small effect Medium effect Large effect
For *df** = 3	$0.06 < V < 0.17$ $0.17 < V < 0.29$ $V > 0.29$	Small effect Medium effect Large effect

Assumptions for Chi-Square

- Data must be in frequency form
- Each observation must be independent of each other
- Sample size must be adequate
  - *For 2 x 2 table, Chi-Square, $n >= 20$*
  - *No more than 20% of cell should have $f_e < 5$*
- Distribution assumptions must be decided before data collection
- Sum of f_o must equal sum of f_e