Module 2 – Section 4

Chi-square Test of Independence

Variables

- Variable 2
  - J categories
- Variable 1
  - I categories
- Neither variable is considered the response variable.

Data

- Random sample of size n from population
- Gather information on two categorical variables

Data Summary

- Cross-classify data according to categories of two variables
- Form into contingency table

Ex. 3 x 4 Contingency Table

	Variable 2				
Variable 1	Cat 1	Cat 2	Cat 3	Cat 4	Total
Cat 1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
Cat 2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$Y_{2.}$
Cat 3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	n

Example

- A study involving more than 5000 students looked at the relationship between smoking habits of students and the smoking habits of their parents.

Ex. Variables

- Variable 2
  - Student Smoking Status
  - Categories: Non-smoker, Smoker
- Variable 1
  - Parent Smoking Status
  - Categories: Neither Smokes, One Smokes, Both Smoke

Ex. Data

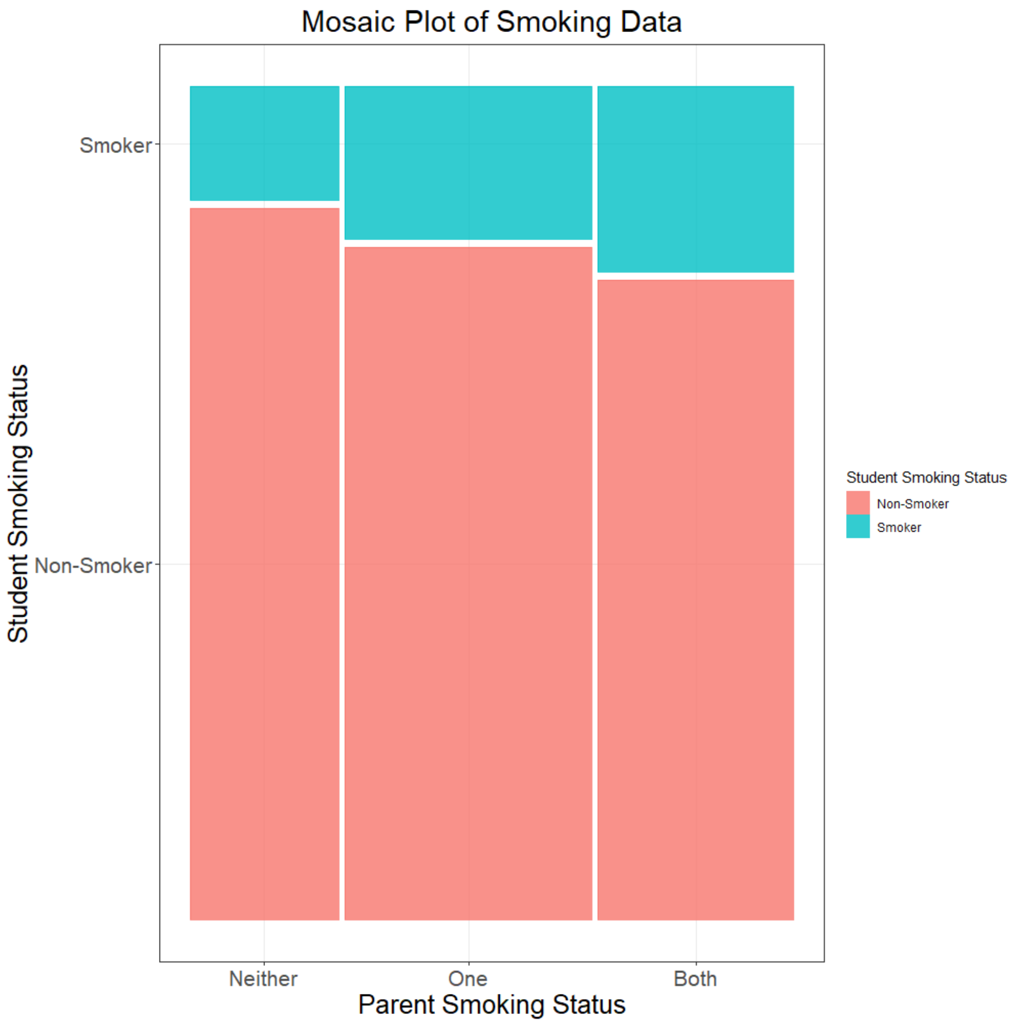
Parent Smoking Status	Student Smoking Status
Neither Smokes	Non-smoker
Neither Smokes	Non-smoker
Neither Smokes	Non-smoker
⋮	⋮
⋮	⋮
Both Smoke	Smoker
Both Smoke	Smoker

Ex. Contingency Table

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

Ex. Mosaic Plot

- A small proportion of students in the study are Smokers.
- The proportion of students who are Smokers is the lowest when Neither Parent Smokes and highest when Both Parents Smoke



Population Proportions

- p_{ij} = population proportion in category i of Variable 1 and category j of Variable 2.
- $p_{i.}$ = population proportion in category i of Variable 1.
- $p_{.j}$ = population proportion in category j of Variable 2.

Test of Independence

- Two categorical variables are independent if

$$p_{ij} = p_{i.}p_{.j} \text{ for all } i \text{ and } j$$

Test of Independence

- H_0: the two variables are independent
  - $p_{ij} = p_{i.}p_{.j}$ for all i and j
- H_a: the two variables are not independent
  - At least one $p_{ij} \neq p_{i.}p_{.j}$ for some i and j

Test of Independence

- If H_0 is true,

$$E(Y_{ij}) = np_{ij} = np_{i.}p_{.j}$$

- Population proportions $p_{i.}$ and $p_{.j}$ are unknown.

Test of Independence

- Estimate with sample proportions from table

$$\widehat{E(Y_{ij})} = n\hat{p}_{i.}\hat{p}_{.j}$$

$$= n\left(\frac{Y_{i.}}{n}\right)\left(\frac{Y_{.j}}{n}\right)$$

$$= \frac{Y_{i.}Y_{.j}}{n}$$

	Var. 2				
Var. 1	Cat 1	Cat 2	Cat 3	Cat 4	Total
Cat 1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
Cat 2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$Y_{2.}$
Cat 3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$Y_{3.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	n

Test of Independence

- If H_0 is true:

$$\widehat{E(Y_{ij})} = \frac{Y_{i.}Y_{.j}}{n} = \frac{(\text{row } i \text{ total})(\text{column } j \text{ total})}{\text{table total}}$$

Test Statistic

- Compare observed cell value Y_{ij} to estimated expected cell value $\widehat{E(Y_{ij})}$:

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left(Y_{ij} - \widehat{E(Y_{ij})}\right)^2}{\widehat{E(Y_{ij})}}$$

- Large values of X^2 indicate evidence the two variables are not independent.

P-value

- If $\widehat{E(Y_{ij})} > 5$ for each cell, the distribution of X^2 is well approximated by a $\chi^2_{(I-1)(J-1)}$ distribution.

$$p\text{-value} = P\left(\chi^2_{(I-1)(J-1)} > X^2\right)$$

Ex. Null and Alternative Hypotheses

- H_0: The smoking status of students and their parents are independent.
- H_a: The smoking status of students and their parents are not independent.

Ex. Expected Values

	Student Smoking Status		
Parent Smoking Status	Non-smoker	Smoker	Total
Neither Smokes	1102.712	253.288	1356
One Smokes	1820.776	418.224	2239
Both Smoke	1447.512	332.488	1780
Total	4371	1004	5375

Ex. Test Statistic and P-value

- Test Statistic

$$X^2 = \sum_{j=1}^{2}\sum_{i=1}^{3}\frac{\left(Y_{ij} - \widehat{E(Y_{ij})}\right)^2}{\widehat{E(Y_{ij})}} = 37.5663$$

- P-value

$$P(\chi_2^2 > 37.5663) < 0.0001$$

Ex. Conclusion

- We have extremely strong evidence that the smoking status of students is not independent from the smoking status of their parents.

Study of Relationship

- Cell Expected Values
- Cell Residuals
- Contribution of Cell to X^2 statistic

Ex. Smoking

- We found extremely strong evidence that the smoking status of students is not independent from the smoking status of the parents.
- Where is the relationship?

Ex. Contingency Table with Expected Values

	Student Smoking Status		
Parent Smoking Status	Non-smoker	Smoker	Total
Neither Smokes	1168 (1102.712)	188 (253.288)	1356
One Smokes	1823 (1820.776)	416 (418.224)	2239
Both Smokes	1380 (1447.512)	400 (332.488)	1780
Total	4371	1004	5375

Ex. Smoking

- Under the assumption of independence:
  - When neither parent smoked, we expect more students to smoke than did.
  - When both parents smoke, we expect less students to smoke than did.
  - When one parent smoked, the expected number of students who smoked is very close to the observed number.

Connections and Similarities

- Analyses for differences in proportions and multinomial response probabilities are similar to analysis for independence.
  - Same Expected Values, Test Statistic, degrees of freedom, p-value.
  - Hypotheses and conclusions are different.

Which one to use?

- Proportions and Multinomial response probabilities
  - Always when group sizes are fixed prior to data collection.
    - Experiment
    - Stratified Sampling
  - Usually when Variable 1 is a grouping variable.

Which one to use?

- Test of Independence
  - Always when Variable 1 is not a grouping variable.
  - Sometimes when group sizes are not fixed prior to data collection.