Unit 2 – Section 2C

Inference for Multiple Multinomial Variables

Variables

- Variable 2 = Response Variable
  - $J > 2$ categories
- Variable 1 = Grouping Variable
  - I groups (categories)

Contingency Table (3 x 3 example)

	Response Variable			
Explanatory Variable	Cat 1	Cat 2	Cat 3	Total
Group 1	Y_{11}	Y_{12}	Y_{13}	n_1
Group 2	Y_{21}	Y_{22}	Y_{23}	n_2
Group 3	Y_{31}	Y_{32}	Y_{33}	n_3
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Null and Alternative Hypothesis

- p_{ij} = Population proportion in category j in group i, $i = 1, \dots, I$

- H_0: $p_{i1}, p_{i2}, \cdots, p_{iJ}$ is same for all $i = 1, \dots, I$
  - Distribution of response variable is same for each group
- H_a: at least one $p_{i1}, p_{i2}, \cdots, p_{iJ}$ is different, $i = 1, \dots, I$
  - Distribution of response variable varies between groups

Model

- If Null Hypothesis is true:
  - Set $p_{11} = p_{21} = p_{31} = \cdots = p_{I1} = p_{.1}$
  - Set $p_{12} = p_{22} = p_{32} = \cdots = p_{I2} = p_{.2}$
  - $\vdots$
  - Set $p_{1J} = p_{2J} = p_{3J} = \cdots = p_{IJ} = p_{.J}$

Model

- Expected Values
  - $E(Y_{ij}) = n_i p_{.j}$
- Each $p_{.j}$ is unknown.

Estimate of p

- Example (3 x 3)
  - $\hat{p}_{.j} = \frac{Y_{1j}+Y_{2j}+Y_{3j}}{n_1+n_2+n_3} = \frac{Y_{.j}}{n}$

	Response Variable			
Explanatory Variable	Cat 1	Cat 2	Cat 3	Total
Group 1	Y_{11}	Y_{12}	Y_{13}	n_1
Group 2	Y_{21}	Y_{22}	Y_{23}	n_2
Group 3	Y_{31}	Y_{32}	Y_{33}	n_3
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	n

Estimate of Expected Values

- $E\left(Y_{ij}\right)$ estimated with $n_i\left(\frac{Y_{.j}}{n}\right)$

$$\widehat{E\left(Y_{ij}\right)} = \frac{n_i Y_{.j}}{n}$$

$$= \frac{\text{row } i \text{ total} * \text{column } j \text{ total}}{\text{table total}}$$

Test Statistic

- Compare Y_{ij} to $\widehat{E(Y_{ij})}$
  - If values are very different, evidence that some of the $p_{i1}, p_{i2}, \cdots, p_{iJ}$ are different for some groups i.

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

P-value

- As long as $\widehat{E(Y_{ij})} \geq 5$ for all i and j, distribution of X^2 is well-approximated by $\chi^2_{(I-1)(J-1)}$.

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

Ex. Smoking and Sex

- As a part of a survey of a given population, over 400 people indicated their smoking status (Non-Smoker, Past Smoker, Current Smoker) and their Sex (Female, Male).

Ex. Variables

- Variable 2 = Response Variable
  - Smoking Status
  - Categories = Non-Smoker, Past Smoker, Current Smoker
- Variable 1 = Grouping Variable
  - Sex
  - Categories = Female, Male

Ex. Data

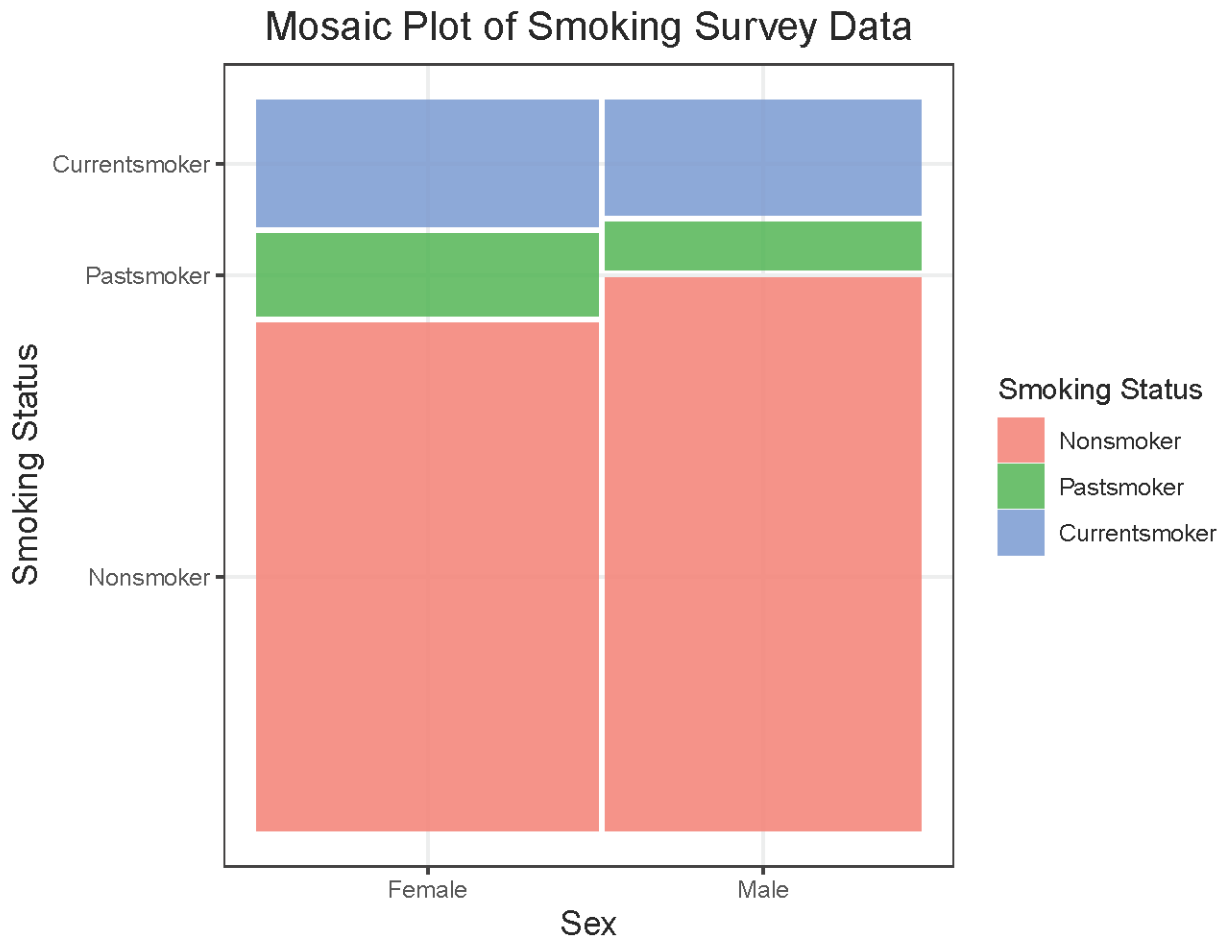
Smoking Status	Sex
Nonsmoker	Male
Nonsmoker	Male
Nonsmoker	Male
⋮	⋮
⋮	⋮
Currentsmoker	Female
Currentsmoker	Female

Ex. Contingency Table

	Smoking Status			
Sex	Non-smoker	Past Smoker	Current Smoker	Total
Female	148	24	37	209
Male	149	13	31	193
Total	297	37	68	402

Ex. Mosaic Plot

- Females have slightly lower proportion of Non-Smokers, and slightly higher proportion of Past-smokers and Current-Smokers than Males.



Ex. Null and Alternative Hypotheses

- Distribution of smoking status is same for each Sex.
  - H_0: p_{i1}, p_{i2}, p_{i3} is the same for all $i = 1, 2$.
- Distribution of smoking status is different between Sexes.
  - H_a: at least one p_{i1}, p_{i2}, p_{i3} is different, $i = 1, 2$.

Male

Ex. Expected Values

	Smoking Status			
Sex	Non-smoker	Past Smoker	Current Smoker	Total
Female	154.4104	19.2363	35.3532	209
Male	142.5896	17.7637	32.6468	193
Total	297	37	68	402

Ex. Test Statistic

$$X^2 = \frac{(148 - 154.4104)^2}{154.4104} + \frac{(24 - 19.2363)^2}{19.2363} +$$

$$\cdots + \frac{(31 - 32.6468)^2}{32.6468}$$

$$= 3.1713$$

Ex. P-value and Conclusion

- p-value $= P(\chi_2^2 > 3.1713) = 0.2048$
- Conclusion: We do not have evidence the smoking status of members of this population is different between Sexes.