Unit 2 – Section 2B

Inference for Multiple Population Proportions

Overview

- Hypothesis Test for Two Population Proportions Revisited
- Inference for More than Two Population Proportions

Variables

- Variable 2 = Response Variable
  - $J = 2$ categories
    - Success/Failure
    - Category of Interest/Not Category of Interest
- Variable 1 = Grouping Variable
  - I groups (categories)

$I = 2$ Groups Revisited

- Cross-classify data according to group and response variable category
- Y_{i1} = number of successes in group i
- Y_{i2} = number of failures in group i
- Enter data into contingency table

I = 2 Groups Contingency Table

	Response Variable		
Explanatory Variable	Success	Failure	Total
Group 1	Y_{11}	Y_{12}	n_1
Group 2	Y_{21}	Y_{22}	n_2
Total	$Y_{.1}$	$Y_{.2}$	n

Null and Alternative Hypotheses

- p_1 = probability of success in group 1
- p_2 = probability of success in group 2
  - H_0: $p_1 = p_2$
  - H_a: $p_1 \neq p_2$
- We can only
  - test for equality of two proportions
  - conduct a two-sided hypothesis test

Model

- If Null Hypothesis is true:
  - Set $p_1 = p_2 = p$
  - $E(Y_{i1}) = n_i p$
  - $E(Y_{i2}) = n_i(1 - p)$
- Value of p is unknown

Estimate of p

- $\hat{p}_{\text{pooled}} = \frac{Y_{11}+Y_{21}}{n_1+n_2} = \frac{Y_{.1}}{n}$

- $1 - \hat{p}_{\text{pooled}} = \frac{Y_{12}+Y_{22}}{n_1+n_2} = \frac{Y_{.2}}{n}$

	Response Variable		
Explanatory Variable	Success	Failure	Total
Group 1	Y_{11}	Y_{12}	n_1
Group 2	Y_{21}	Y_{22}	n_2
Total	$Y_{.1}$	$Y_{.2}$	n

Estimate of Expected Values

- $E(Y_{i1})$ estimated with $n_i\left(\frac{Y_{.1}}{n}\right)$
- $E(Y_{i2})$ estimated with $n_i\left(\frac{Y_{.2}}{n}\right)$

Estimate of Expected Values

- In general, estimate of $E(Y_{ij})$ is:

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

Test Statistic

- Compare Y_{ij} to $\widehat{E(Y_{ij})}$
  - If values are very different, evidence that p_1 and p_2 are different

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

P-value

- As long as $\widehat{E(Y_{ij})} \geq 5$ for all i and j, distribution of X^2 is well-approximated by χ_1^2.

$$p\text{-value} = P(\chi_1^2 > X^2)$$

Ex. Angina Treatment

- Angina pectoris is a chronic heart condition that inflicts periodic attacks of chest pain.
- Experiment
  - 160 patients randomly assigned to drug Timolol.
  - 147 patients randomly assigned to placebo.

Ex. Variables

- Variable 2 = Response Variable
  - Outcome
  - Categories: No.Angina, Angina
- Variable 1 = Grouping Variable
  - Drug
  - Categories = Timolol, Placebo

Ex. Data

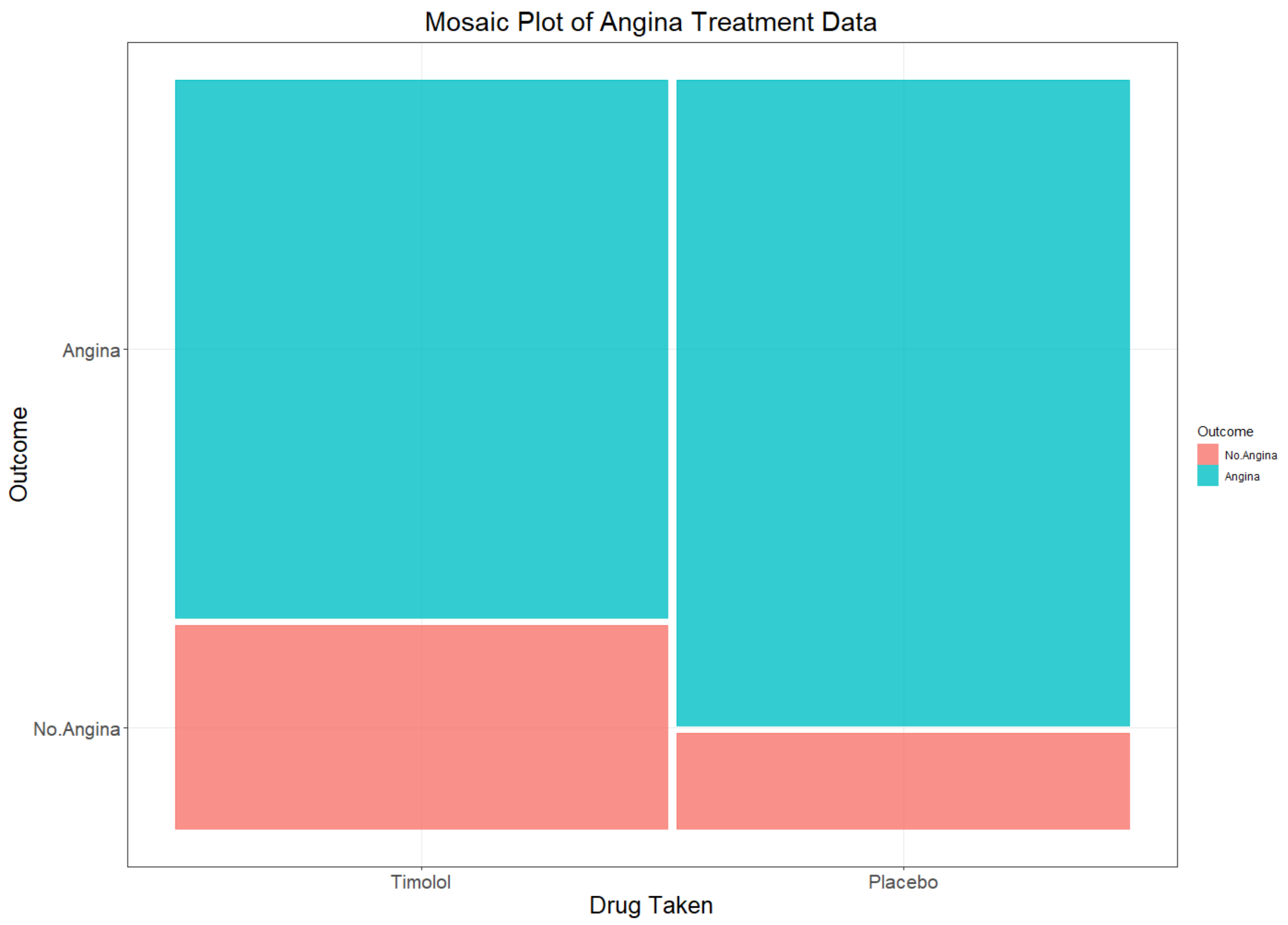
Outcome	Drug
No.Angina	Timolol
No.Angina	Timolol
No.Angina	Timolol
⋮	⋮
⋮	⋮
Angina	Placebo
Angina	Placebo

Ex. Contingency Table

	Outcome		
	No Angina	Angina	Total
Timolol	44	116	160
Placebo	19	128	147
Total	63	244	307

Ex. Mosaic Plot

- Large proportions of patients in each group still have angina.
- Proportion of patients with no angina is higher in the treatment group than in the placebo group.



Ex. Hypothesis Test

- p_1 is the proportion of patients receiving relief from angina in the Timolol group.
- p_2 is the proportion of patients receiving relief from angina in the placebo group.
  - H_0: $p_1 = p_2$
  - H_a: $p_1 \neq p_2$

Ex. Expected Counts

	No Angina	Angina	Total
Timolol	$\frac{160 * 63}{307} = 32.834$	$\frac{160 * 244}{307} = 127.166$	160
Placebo	$\frac{147 * 63}{307} = 30.166$	$\frac{147 * 244}{307} = 116.834$	147
Total	63	244	307

Ex. Test Statistic

$$X^2 = \frac{(44-32.834)^2}{32.834} + \frac{(116-127.166)^2}{127.166}$$

$$+ \frac{(19-30.166)^2}{30.166} + \frac{(128-116.834)^2}{116.834}$$

$$= 9.9782$$

Ex. P-value and Conclusion

- p-value
  - $P(\chi_1^2 > 9.9782) = 0.0016$
- Conclusion: There is strong evidence the proportion of patients receiving relief from angina in the Timolol group is different from the proportion of patients receiving relief from angina in the placebo group.

For General I

- p_i = Probability of success in group i, $i = 1, \ldots, I$
- H_0: $p_1 = p_2 = \cdots = p_I$
- H_a: at least one p_i is different, $i = 1, \ldots, I$

Example ($I = 3$)

	Response Variable		
Explanatory Variable	Success	Failure	Total
Group 1	Y_{11}	Y_{12}	n_1
Group 2	Y_{21}	Y_{22}	n_2
Group 3	Y_{31}	Y_{32}	n_3
Total	$Y_{.1}$	$Y_{.2}$	n

Model

- If Null Hypothesis is true:
  - Set $p_1 = p_2 = p_3 = \cdots = p_I = p$
  - $E(Y_{i1}) = n_i p$
  - $E(Y_{i2}) = n_i(1 - p)$
- Value of p is unknown

Estimate of p

- Example ($I = 3$)
  - $\hat{p}_{\text{pooled}} = \dfrac{Y_{11}+Y_{21}+Y_{31}}{n_1+n_2+n_3} = \dfrac{Y_{.1}}{n}$
  - $1 - \hat{p}_{\text{pooled}} = \dfrac{Y_{12}+Y_{22}+Y_{32}}{n_1+n_2+n_3} = \dfrac{Y_{.2}}{n}$

	Response Variable		
Explanatory Variable	Success	Failure	Total
Group 1	Y_{11}	Y_{12}	n_1
Group 2	Y_{21}	Y_{22}	n_2
Group 3	Y_{31}	Y_{32}	n_3
Total	$Y_{.1}$	$Y_{.2}$	n

Estimate of Expected Values

- $E(Y_{i1})$ estimated with $n_i\left(\frac{Y_{.1}}{n}\right)$
- $E(Y_{i2})$ estimated with $n_i\left(\frac{Y_{.2}}{n}\right)$

Estimate of Expected Values

- In general, estimate of $E(Y_{ij})$ is:

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

Test Statistic

- Compare Y_{ij} to $\widehat{E(Y_{ij})}$
  - If values are very different, evidence that some of the p_i are different

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

P-value

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

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

Ex. Diodes

- Diodes used on a printed circuit board are produced in lots of size 4000. To study the homogeneity of lots with respect to a demanding specification, random samples of size 300 from 5 consecutive lots were taken and the diodes tested.

Ex. Variable

- Variable 2 = Response Variable
  - Status
  - Categories: Non-Conforming, Conforming
- Variable 1 = Grouping Variable
  - Lot
  - Categories = 1, 2, 3, 4, 5

Ex. Data

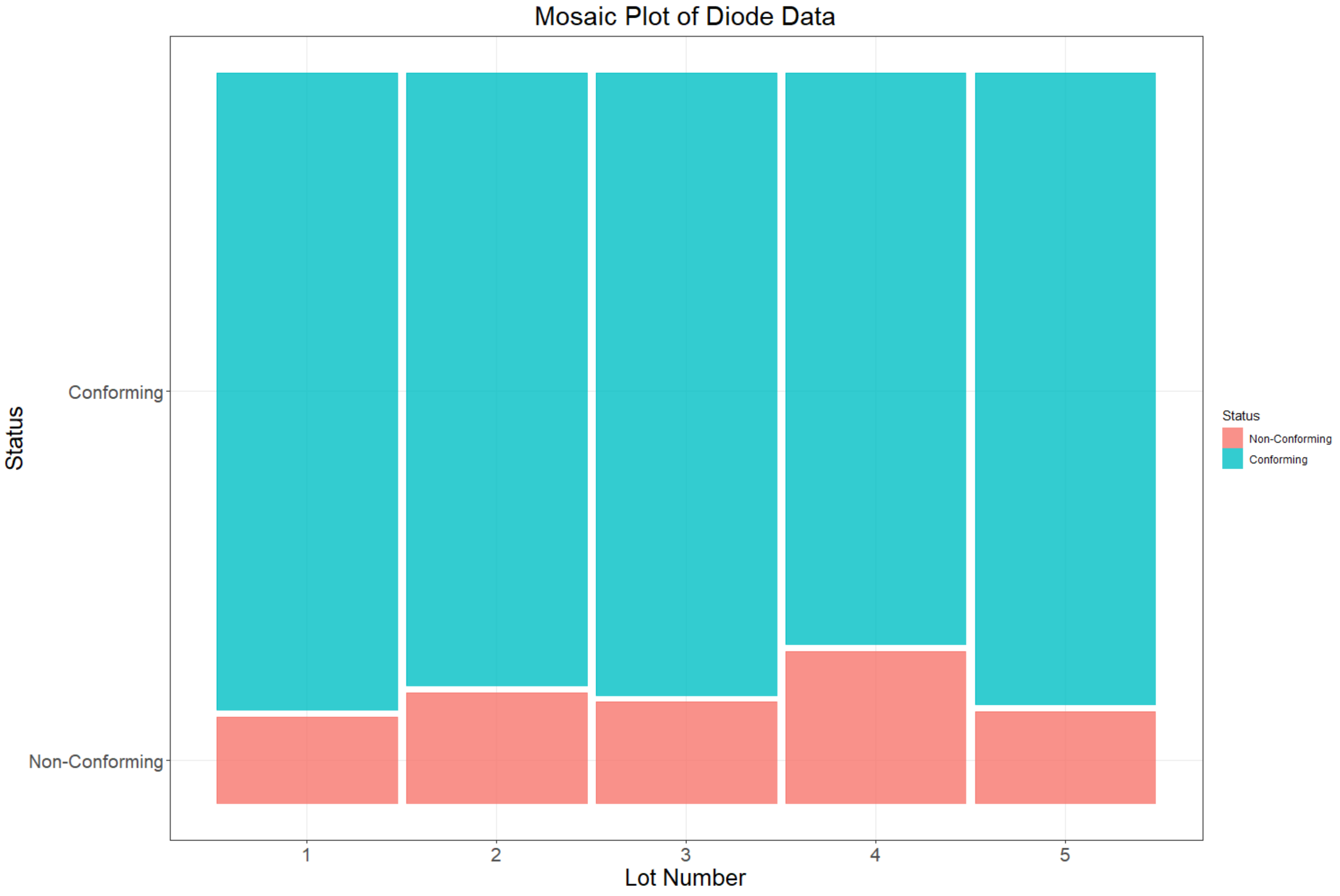
Status	Lot
Non-Conforming	1
Non-Conforming	1
Non-Conforming	1
⋮	⋮
⋮	⋮
Conforming	5
Conforming	5

Ex. Contingency Table

	Status		
Lot	Non-Conforming	Conforming	Total
Lot 1	36	264	300
Lot 2	46	254	300
Lot 3	42	258	300
Lot 4	63	237	300
Lot 5	38	262	300
Total	225	1275	1500

Ex. Mosaic Plot

- Lot 4 has the highest proportion of non-conforming diodes.
- Small differences between the proportion of non-conforming diodes in the other lots.



Ex. Null and Alternative Hypotheses

- p_i = proportion of non-conforming diodes in lot i
  - H_0: $p_1 = p_2 = \cdots = p_5$
  - H_a: at least one p_i is different, $i = 1, \ldots, 5$

Ex. Expected Value

- If H_0 is true:

$$\widehat{E(Y_{i1})} = \frac{300 * 225}{1500} = 45$$

$$\widehat{E(Y_{i2})} = \frac{300 * 1275}{1500} = 255$$

Ex. Test Statistic

$$X^2 = \frac{(36-45)^2+(46-45)^2+(42-45)^2+(63-45)^2+(38-45)^2}{45}$$

$$+\frac{(264-255)^2+(254-255)^2+(258-255)^2+(237-255)^2+(262-255)^2}{255}$$

$$= 12.131$$

Ex. P-value and Conclusion

- p-value = $P(\chi_4^2 > 12.131) = 0.0164$
- Conclusion: We have moderately strong evidence to conclude at least one of the lots has a different proportion of non-conforming diodes than the others.

What's next?

- If we reject H_0, then which p_i values are different?
  - Pairwise Hypothesis Tests
- Test:

$$H_0: p_i = p_l$$

$$H_a: p_i \neq p_l$$

  for each pair of groups (i, l)
- Beware multiple comparisons

Ex. Diodes

- Previously, we concluded there was moderately strong evidence that at least one of the groups has a different proportion of non-conforming diodes.
- $10 = \binom{5}{2}$ pairs of groups

Ex. Pairwise Tests

Group i	Group l	Test Statistic (X^2)	P-value Adjusted
1	2	1.4126	0.570
1	3	0.5305	0.777
1	4	8.8187	0.042
1	5	0.0617	0.901
2	3	0.2131	0.810
2	4	3.2400	0.226
2	5	0.8859	0.684
3	4	5.0909	0.105
3	5	0.2308	0.810
4	5	7.4406	0.044

Ex. Pairwise Tests

- We have moderately strong evidence to conclude Lot 4 has a different proportion of non-conforming diodes than Lots 1 and 5.