Unit 2 – Section 2A

Inference for Two Population Proportions

Outline

- Variables and Data
- Hypothesis Test for $p_1 - p_2$
- Confidence Interval for $p_1 - p_2$
- Sample Size Calculations

Variables

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

Ex. Survey Study

- Obtaining a high response rate for surveys is very important to the overall validity of the results from a random sample of the population. A survey was administered to a random sample of around 10,000 doctors. Approximately half of the doctors were randomly selected to receive a letter before the survey. Researchers wanted to see if doctors who received the letter before the survey would respond to the survey at a different rate than doctors who did not receive the letter.

Ex. Variables

- Variable 2 = Response Variable
  - Did they return the survey?
  - Possible categories = Yes, No
- Variable 1 = Grouping Variable
  - Did they receive a letter?
  - Possible categories = Yes, No

Ex. Data

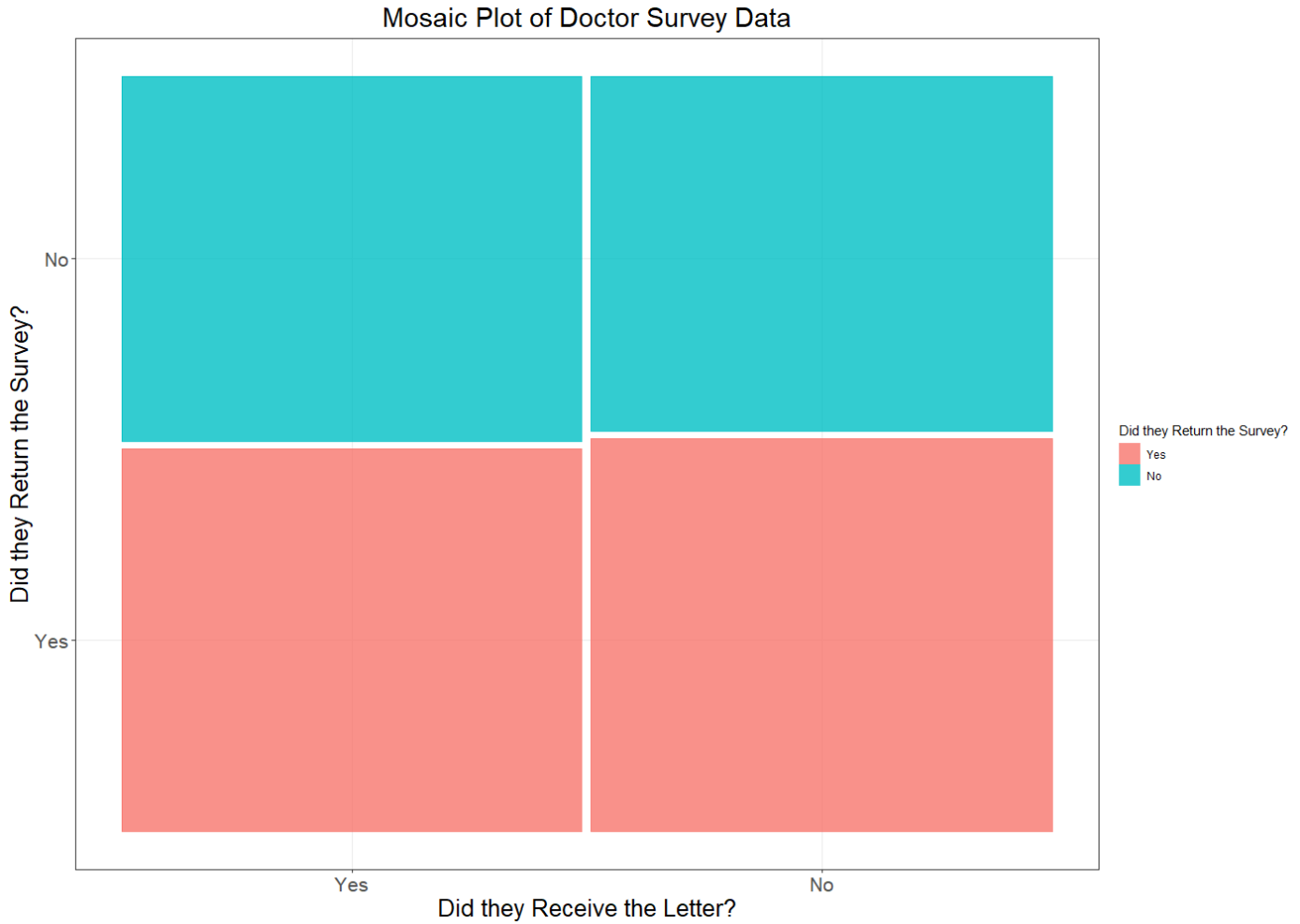
Receive Letter	Return Survey
Yes	Yes
Yes	Yes
Yes	Yes
⋮	⋮
⋮	⋮
No	No
No	No

Ex. Contingency Table

	Return Survey		
Receive Letter	Yes	No	Total
Yes	2570	2448	5018
No	2645	2384	5029
Total	5215	4832	10047

Ex. Mosaic Plot

- Proportions of each group that returned the survey are very similar; slightly more in the no letter group returned the survey.



Binomial Random Variables

- $Y_1 =$ number of successes in group 1 out of n_1 independent and identical trials
- $Y_2 =$ number of successes in group 2 out of n_2 independent and identical trials
- Independent & Identical as before
  - Both n_1 and n_2 are less than 10% of respective population sizes

Ex. Contingency Table

	Return Survey		
Receive Letter	Yes	No	Total
Yes	2570	2448	5018
No	2645	2384	5029
Total	5215	4832	10047

Population Proportions

- $p_1 =$ probability of success in group 1
- $p_2 =$ probability of success in group 2
- p_1 and p_2 are generally unknown parameters.
- Are these parameters the same? Is there an equal probability of success in each group?

Null and Alternative Hypotheses

- H_0: $p_1 - p_2 = 0$
- H_a: $p_1 - p_2 \neq 0$ or

  H_a: $p_1 - p_2 < 0$ or

  H_a: $p_1 - p_2 > 0$

Test Statistic

- Estimate $p_1 - p_2$ with $\hat{p}_1 - \hat{p}_2$
- What is the distribution of $\hat{p}_1 - \hat{p}_2$ when H_0 is true?

Test Statistic

- Call common value $p_1 = p_2 = p$
  - $E(\hat{p}_1) = E(\hat{p}_2) = p$
  - $V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2)$

$$= \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$$

  - $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}$

Test Statistic

- Value of common p is unknown.
- Use Pooled Estimate:

$$\hat{p}_{\text{pooled}} = \frac{y_1 + y_2}{n_1 + n_2}$$

	Response Variable		
Group	Success	Failure	Total
Yes	y_1	$n - y_1$	n_1
No	y_2	$n - y_2$	n_2
Total	$y_1 + y_2$	$n - (y_1 + y_2)$	n

Test Statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_2}}}$$

Test Statistic

- If H_0 is true and assuming
  - $n_1\hat{p}_{\text{pooled}} \geq 10$ and $n_1(1-\hat{p}_{\text{pooled}}) \geq 10$
  - $n_2\hat{p}_{\text{pooled}} \geq 10$ and $n_2(1-\hat{p}_{\text{pooled}}) \geq 10$

Distribution of z is approximately $N(0,1)$

P-value

- $H_a: p_1 - p_2 \neq 0$
  - p-value $= 2 * P(Z > |z|)$
- $H_a: p_1 - p_2 < 0$
  - p-value $= P(Z < z)$
- $H_a: p_1 - p_2 > 0$
  - p-value $= P(Z > z)$

Ex. Survey Study

- $p_1 =$ probability the doctors who receive letter will return survey
- $p_2 =$ probability the doctors who do not receive letter will return survey
- H_0: $p_1 - p_2 = 0$
- H_a: $p_1 - p_2 \neq 0$

Ex. Contingency Table

	Return Survey		
Receive Letter	Yes	No	Total
Yes	2570	2448	5018
No	2645	2384	5029
Total	5215	4832	10047

Ex. Data Summaries

- Receive Letter = Yes
  - $n_1 = 5018$
  - $y_1 = 2570$
  - $\hat{p}_1 = \frac{2570}{5018} = 0.5122$
- Receive Letter = No
  - $n_2 = 5029$
  - $y_2 = 2645$
  - $\hat{p}_2 = \frac{2645}{5029} = 0.5259$

Ex. Test Statistic

- Estimate p from samples

$$\hat{p}_{\text{pooled}} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{2570 + 2645}{5018 + 5029} = 0.5191$$

Ex. Test Statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}$$

$$= \frac{0.5122 - 0.5259}{\sqrt{\frac{(0.5191)(0.4809)}{5018} + \frac{(0.5191)(0.4809)}{5029}}}$$

$$= -1.38$$

Ex. P-value

- The assumptions are satisfied:
  - 5018(0.5191) and 5018(0.4809) ≥ 10
  - 5029(0.5191) and 5029(0.4809) ≥ 10
- p-value

$$2 * P(Z > 1.38) = 0.1665$$

Ex. Conclusion

- We do not have evidence to conclude the response rate for this survey was different between the group that received the letter and the group that did not receive the letter in this population of doctors.

Confidence Interval for $p_1 - p_2$

- Based on statistic $\hat{p}_1 - \hat{p}_2$
- Assumptions
  - $n_1\hat{p}_1 \geq 10$ and $n_1(1 - \hat{p}_1) \geq 10$
  - $n_2\hat{p}_2 \geq 10$ and $n_2(1 - \hat{p}_2) \geq 10$
- No pooling since no assumption on values of p_1 and p_2

Confidence Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Ex. Prostate Cancer Study

- There has been debate among doctors over whether surgery can prolong life among men suffering from prostate cancer. In 2003, in a study published in the New England Journal of Medicine, men diagnosed with prostate cancer were randomly assigned to either undergo surgery or not. The response variable is whether or not the men died from prostate cancer.

Ex. Variables

- Variable 2 = Response Variable
  - Did they die from Prostate Cancer?
  - Possible categories = Yes, No
- Variable 1 = Grouping Variable
  - Did they undergo surgery?
  - Possible categories = Yes, No

Ex. Data

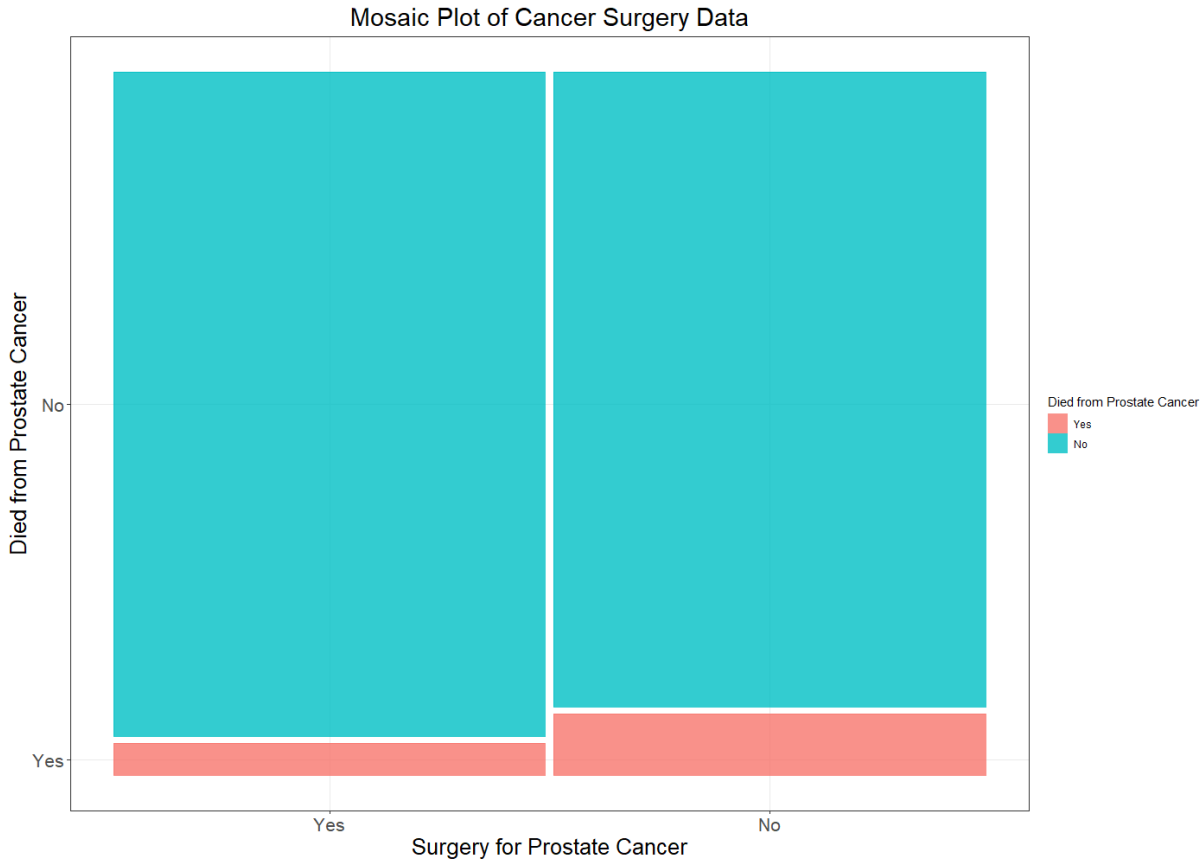
Surgery	Death from Prostate Cancer
Yes	Yes
Yes	Yes
Yes	Yes
⋮	⋮
⋮	⋮
No	No
No	No

Ex. Contingency Table

	Death from Prostate Cancer		
Surgery	Yes	No	Total
Yes	16	331	347
No	31	317	348
Total	47	648	695

Mosaic Plot

- Very large proportion of patients in each group did not die of prostate cancer.
- No Surgery group has higher proportion of deaths due to prostate cancer than Surgery group.



Ex. Data Summaries

- Surgery = Yes
  - $n_1 = 347$
  - $y_1 = 16$
  - $\hat{p}_1 = \frac{16}{347} = 0.0461$
- Surgery = No
  - $n_2 = 348$
  - $y_2 = 31$
  - $\hat{p}_2 = \frac{31}{348} = 0.0891$

Ex. Check Assumptions

- The assumptions are satisfied:
  - $n_1\hat{p}_1 = y_1 = 16$ and $n_1(1-\hat{p}_1) = n_1 - y_1 = 331$
  - $n_2\hat{p}_2 = y_2 = 31$ and $n_2(1-\hat{p}_2) = n_2 - y_2 = 317$

Ex. Confidence Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= (0.0461 - 0.0891) \pm 1.96 \sqrt{\frac{0.0461(0.9539)}{347} + \frac{0.0891(0.9109)}{348}}$$

$$= (-0.0802, -0.0058)$$

Ex. Interpretation

- We are 95% confident the proportion of men in this population with prostate cancer who died from prostate cancer is between 0.0058 and 0.0802 lower for men who have surgery than for men who do not.

Sample Size Calculation

- Confidence Interval
  - Assume $n_1 = n_2 = n$
  - Margin of Error for CI is:

$$ME = z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

  - Solve for n.

Sample Size Calculation

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{ME}\right)^2 \left(\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)\right)$$

- Same problem as one sample case
  - Values of $\hat{p}_1$ and $\hat{p}_2$ come from samples
  - Unknown until after data collected
- Same two solutions as one sample case

Sample Size Calculation

- Solution 1:

$$n \geq \left(\frac{z_{1-\alpha/2}}{ME}\right)^2 (0.5)$$

- Solution 2:

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{ME}\right)^2 \left(\hat{\hat{p}}_1\left(1-\hat{\hat{p}}_1\right) + \hat{\hat{p}}_2\left(1-\hat{\hat{p}}_2\right)\right)$$

Ex. Sample Size Calculation

- Prostate Cancer – Surgery-No Surgery Example
  - ME = 0.03
  - $\hat{\hat{p}}_1 = 0.05$ and $\hat{\hat{p}}_2 = 0.09$
  - 95% Confidence Level

Ex. Sample Size Calculation

- Solution 2 is more appropriate:

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{ME}\right)^2 \left(\hat{\hat{p}}_1\left(1-\hat{\hat{p}}_1\right)+\hat{\hat{p}}_2\left(1-\hat{\hat{p}}_2\right)\right)$$

$$= \left(\frac{1.96}{0.03}\right)^2 \left((0.05)(0.95)+(0.09)(0.91)\right)$$

$$= 552.3367$$

Ex. Sample Size Calculation

- To get desired margin of error of 0.03 with 95% confidence, we need approximately 553 patients in each treatment group.

Sample Size Calculation

- Hypothesis Test
  - Assume $n_1 = n_2 = n$
  - Power $1 - \beta$
  - Size of difference $p_1 - p_2$
  - Assumption for values for p_1 and p_2

Sample Size Calculation

- Intermediate calculations

$$p_0 = \frac{p_1 + p_2}{2}$$

$$R = \sqrt{\frac{2p_0(1-p_0)}{p_1(1-p_1) + p_2(1-p_2)}}$$

Sample Size Calculation

- H_a: $p_1 - p_2 \neq 0$

$$n \geq \frac{\left(z_{1-\beta} + Rz_{1-\frac{\alpha}{2}}\right)^2 (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2}$$

Sample Size Calculation

- H_a: $p_1 - p_2 < 0$ or H_a: $p_1 - p_2 > 0$

$$n \geq \frac{\left(z_{1-\beta} + Rz_{1-\alpha}\right)^2 (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2}$$

Ex. Sample Size Calculation

- New Survey to Doctors – Testing Incentive
  - Assume $p_1 = 0.55$ and $p_2 = 0.5$
  - Power $= 0.9$ for $p_1 - p_2 = 0.05$
  - $\alpha = 0.05$

Ex. Sample Size Calculation

$$p_0 = \frac{p_1 + p_2}{2}$$

$$= \frac{0.55 + 0.5}{2}$$

$$= 0.525$$

$$R = \sqrt{\frac{2p_0(1-p_0)}{p_1(1-p_1) + p_2(1-p_2)}}$$

$$= \sqrt{\frac{2(0.525)(0.475)}{0.55(0.45) + 0.5(0.5)}}$$

$$= 1.001255$$

Ex. Sample Size Calculation

- H_a: $p_1 - p_2 \neq 0$

$$n \geq \frac{\left(z_{1-\beta} + Rz_{1-\frac{\alpha}{2}}\right)^2 \left(p_1(1-p_1) + p_2(1-p_2)\right)}{(p_1 - p_2)^2}$$

$$= \frac{\left(1.282 + 1.001255(1.96)\right)^2 \left(0.55(0.45) + 0.5(0.5)\right)}{(0.05)^2}$$

$$= 2094.153$$

Ex. Sample Size Calculation

- In order to detect a difference in response rates of 5% between the two treatment groups with probability 0.9, the sample size from both groups must be approximately 2,095 doctors.