Recitation 11/27

Review of Inference for Categorical Data

Chapter 9: Inference for Proportions & Counts

- One sample proportions:
  - Large Sample: Normal Approximation
  - Small Sample: Binomial Test
- Two sample proportions:
  - Large Sample: Normal Approximation
  - Small sample: Fisher’s Exact Test
  - Matched Pairs: MacNemar’s Test
- Chi Squared:
  - Goodness of Fit
  - Independence/Homogeneity

One Sample Inference

Large Sample Test: Normal Approximation

Conditions:

- Single Random Sample
- Binary Data
- *$n*p_0 > 10$, $n*(1-p_0)>10$*

Then our test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{y - n p_0}{\sqrt{n p_0 q_0}}.$$

And we can use the normal distribution to calculate the p-value.

Small samples: Binomial Test

Conditions:

- Single Random Sample
- Binary Data
- Does NOT meet large counts condition

Then the p-value is calculated using the binomial distribution:

$$P\text{-value} = P(Y \geq y \mid p = p_0) = \sum_{i=y}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \quad \text{(Upper One-Sided).}$$

$$P\text{-value} = P(Y \leq y \mid p = p_0) = \sum_{i=0}^{y} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \quad \text{(Lower One-Sided).}$$

Two Sample Inference

Large Sample Test: Normal Approximation

Conditions:

- Two Independent Random Samples
- Binary Data
- $n*\hat{p} > 10,\ n*(1-\hat{p})>10$ for BOTH samples

Then our test statistic is:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}.$$

or

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \quad \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x + y}{n_1 + n_2}$$

And we can use the normal distribution to calculate the p-value.

Small samples: Fisher's Exact Test

Conditions:

- Two Independent Random Samples
- Binary Data
- Does NOT meet large counts condition

Then the p-value is calculated using the hypergeometric distribution:

$$\text{Upper One-Sided } P\text{-value} = P_U = P(X \geq x \mid X + Y = m) = \sum_{i \geq x} \frac{\binom{n_1}{i}\binom{n_2}{m-i}}{\binom{n}{m}};$$

$$\text{Lower One-Sided } P\text{-value} = P_L = P(X \leq x \mid X + Y = m) = \sum_{i \leq x} \frac{\binom{n_1}{i}\binom{n_2}{m-i}}{\binom{n}{m}};$$

Matched Pairs: McNemar's Test

Conditions:

- Random Samples
- Binary Data
- Matched Pairs

Then the p-value (for an "upper" test) is calculated using a conditional binomial distribution:

		Condition 2 Response	
		Yes	No
Condition 1 Response	Yes	a	b
	No	c	d

$$P\text{-value} = P(B \geq b \mid B + C = m) = \left(\frac{1}{2}\right)^m \sum_{i=b}^{m} \binom{m}{i}.$$

Chi Squared Tests

Goodness of Fit

Conditions:

- Single Categorical Variable
- Multinomial
- All “expected counts” > 1, and at least 80% of “expected counts”>5

$$e_i = np_{i0} \qquad (i = 1, 2, \ldots, c).$$

Then our test statistic is:

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - e_i)^2}{e_i}.$$

And we can use the Chi Squared distribution with c-1 df to calculate the p-value.

Independence OR Homogeneity

Conditions:

- Two Categorical Variables OR Two Independent Samples and one Categorical variable
- Multinomial
- All “expected counts” > 1, and at least 80% of “expected counts”>5

$$\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}.$$

Then our test statistic is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}.$$

And we can use the Chi Squared distribution with (r-1)*(c-1) df to calculate the p-value.

Summary

Chapter 9: Inference for Proportions & Counts

- One sample proportions:
  - Large Sample: Normal Approximation
  - Small Sample: Binomial Test
- Two sample proportions:
  - Large Sample: Normal Approximation
  - Small sample: Fisher's Exact Test
  - Matched Pairs: MacNemar's Test
- Chi Squared:
  - Single Categorical Variable: Goodness of Fit
  - Two-Way Categorical Variables: Independence/Homogeneity

Examples

To gauge a change in opinion regarding the public view on bilingual education, a telephone poll was taken in September 1993 and again in September 1995.[9] The results based on the survey of 1000 American adults contacted in each poll were that 40% from the 1993 poll and 48% from the 1995 poll favored teaching all children in English over bilingual alternatives. Has there been a significant change in opinion? Answer by doing a two-sided test for the significance of the difference in two proportions at $\alpha = .05$. Why is a two-sided alternative appropriate here?

A study evaluated the urinary-thromboglobulin excretion in 12 normal and 12 diabetic patients.[11] Summary results are obtained by coding values of 20 or less as "low" and values above 20 as "high," as shown in the following table.

	Excretion	
	Low	High
Normal	10	2
Diabetic	4	8

(a) Set up the hypotheses to determine whether there is a difference in the urinary-thromboglobulin excretion between normal and diabetic patients. Which statistical test is appropriate to test the hypotheses?

(b) Calculate the P-value of the test. What is your conclusion using $\alpha = .05$?

People at high risk of sudden cardiac death can be identified using the change in a signal averaged electrocardiogram before and after prescribed activities. The current method is about 85% accurate. The method was modified, hoping to improve its accuracy. The new method is tested on 50 people and gave correct results on 46 patients. Is this convincing evidence that the new method is more accurate?

(a) Set up the hypotheses to test that the accuracy of the new method is better than that of the current method.

(b) Perform a test of the hypotheses at $\alpha = .05$. What do you conclude about the accuracy of the new method?

In a speech class two persuasive speeches, one pro and the other con, were given by two students on requiring guest lists for fraternity/sorority parties. The opinions of the other 52 students in the class were obtained on this issue before and after the speeches with the following responses.

		After	
		Pro	Con
Before	Pro	2	8
	Con	26	16

(a) Set up the hypotheses to determine whether or not there is a change in opinion of the students. Which statistical test is appropriate to test the hypotheses?

(b) Calculate the P-value of the test. What is your conclusion using $\alpha = .05$?

Use the following data to test the hypothesis that a horse's chances of winning are unaffected by its position on the starting lineup. The data give the starting position of each of 144 winners, where position 1 is closest to the inside rail of the race track.[13]

Starting Position	1	2	3	4	5	6	7	8
Number of Wins	29	19	18	25	17	10	15	11

State the hypotheses and perform a test at $\alpha = .05$.