Chapter 8: Inference for Proportions

First Read: Chapter 8, then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Exercises:

Introduction

We often wish to make inferences concerning categorical variables. In this chapter we examine a binary variable (one with two categories); for example success or failure at a task, answer yes or no to a question, etc. The data in this chapter are counts **or** percentages or proportions obtained from counts. The parameter we wish to make inferences about is the proportion in the population of a category. We will examine the case of a single population as well as comparing two populations. The inference process for the one and two sample case discussed here is similar to the cases in Chap.7 where we compared one and two sample means.

View the Video – Inference for Proportions

Inference for a Single Proportion

We assume X_1, X_2,..., X_n is a sample from a distribution with the following properties:

P(X = 1) = p	**call this a success**
P(X = 0) = 1 – p	**call this a failure**

NOTE: "p" is unknown and there are only 2 possible values for X (i.e., pass or fail, live or die, etc.).

We want to make inferences about "p".

First estimate $\overline{X}$ = proportion of 1's in the sample.

We know that the sampling distribution of $\overline{X}$ is centered on "p", the proportion in the population:

$$\mu_{\overline{x}} = p$$

and the variability of the sampling distribution about "p" is measured by:

$$\sigma_{\bar{x}} = \sqrt{\frac{p\,(1-p)}{n}}$$

We estimate the accuracy of "X" by the **standard error** of the estimate, which is:

$$\sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}}$$

The textbook writes this as:

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

$$\text{since } \bar{x} = \hat{p}$$

Then the CLT says for large n

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

with (1-α) confidence interval for "p" given by:

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} Z_{\alpha/2}$$

Example: Application to Polling

Say a pollster randomly selects n people from the population π of eligible voters and asks "Will you vote Liberal in the next election?"

$$\text{let } X_i = \begin{cases} 1 \text{ for YES} \\ 0 \text{ for NO} \end{cases}$$

Then $\hat{p}$ is the estimate of p our population proportion

p = proportion of eligible voters who are going to vote Liberal in the next election

A 95% CI for p is given by:

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ 1.96$$

NOTE !!! $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ 1.96$

is called the **margin of error** for our estimate.

Example:

Consider the following problem of determining the proportion of coffee drinkers that prefer instant coffee. We have a preference test for coffee drinkers. In a sample of fifty coffee drinkers, n=50, we find 19 prefer instant coffee, the sample proportion is:

$$\hat{p} = \frac{19}{50} = .38$$

A coffee executive believes that less than 50% of coffee drinkers (p = 0.5) prefer instant coffee. Test this claim.

Test H_o: $p = 0.5$
H_a: $p < 0.5$

Substitute into Z and we have:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.38 - .5}{\sqrt{\frac{.5(.5)}{50}}} = \frac{-.12}{.07} = -1.7$$

A standardized score of -1.7 gives a p-value = .045.

The probability of .045 is small. Indicating the executive is correct in believing the proportion of coffee drinkers is less than .5. In other words we have evidence against H_o, that is, the proportion who prefer instant is less than .5 in the population.

A 95% CI is given by $\mathbf{.38 \pm \sqrt{.38(1 - .38)/ 50}\ (1.96)}$

Comparing Two Proportions

Previously we examined the problem of comparing one proportion. Now we see how to compare two proportions. We have two populations and take two independent samples. We want to test H_o: $p_1 = p_2$ and use the following Z statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{s_p(D)}$$

where D refers to the difference in the two proportions. The pooled standard deviation of the difference is

$$s_p(D) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$\hat{p} = \frac{X_1 + X_2}{n_1 \quad n_2}$$

refers to the overall proportion.

We substitute into the above formula and calculate the approximate p-values for the statistic.

Example:

A university financial aid office polled an SRS of undergraduate students to study their summer

employment. Not all students were employed the previous summer.

We have data on men and women plus their employment history given below. The researcher wishes to test if the proportion of men employed is equal to the proportion of women employed.

University Survey

	Men	Women
Employed	718	593
Not Employed	79	139
	797	732

The subscript M and W below refer to men and women respectively, whereas the subscript E refers to the employed group.

H_o: $p_{ME} = p_{WE}$ (proportion employed equal)

H_a: $p_{ME} \neq p_{WE}$ (proportion employed not equal)

$$\hat{p}_{ME} = \frac{718}{797} = .901 \qquad \hat{p}_{WE} = \frac{593}{732} = .810$$

overall proportion $\hat{p} = \frac{718 + 593}{797 + 732} = .857$

From the data above we calculate the following proportions. The standard error of the difference in proportions is:

$$s_p(D) = \sqrt{.857(1 - .857)\left(\frac{1}{797} + \frac{1}{732}\right)}$$

$$= .0179$$

the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{s_p(D)}$$

$$= \frac{(.901 - .810)}{.0179}$$

$$= 5.07$$

The Z statistic is about 5 standard deviations above the mean. The p-value will be very small given that

+ or - 3 standard deviations will give 99.7% of the data. Using table A, the p-value < .0002.

Therefore, we have very strong evidence against H_o **or** the proportion of male students employed during the summer differs from the proportion of female students employed.

Example: Test of Uniformity

Suppose a researcher has "c" categories and wants to see if the probabilities/proportions are equal to one another in each category.

H_0: $p_1 = p_2 = ...=p_c = 1/c$
(the model under null hypothesis, is all proportions equal)

H_A: some p_i is different
(the model under the null hypothesis is not correct)

How do we test H_o? We will see how this is done using the following example:

Why do students select courses?

Example: Research Course Selection

A researcher wants to see why students select a course. Subsequently, n=50 students were asked to select one of the following four categories (c=4). When asked why they enrolled in the course, they were given the following choices:

1) Interest in topic
2) Ease in passing
3) Instructor
4) Time of day

The observed frequencies (f_o) are given below for our sample of 50 students:

	Interest	Ease	Inst	Time	Total
f_o	18	17	7	8	50

The expected frequencies (f_e), for the 4 categories under the null hypothesis is:

The proportion in each category equal to H_o: 1/c or ¼ = .25 (remember there are 4 categories).

If our model is correct then the proportion in each category is equal to p=.25. Since we have 50 observations, 50 (.25) = 12.5 is the expected number in each category under the null hypothesis, or model.

	Interest	Ease	Inst	Time	Total
f_e	12.5	12.5	12.5	12.5	50

The test statistic Chi-square is given below. This is the same Chi-square distribution introduced in Chapter 7 but the test statistic is different. Here we are testing if the model is reasonable for the data.

$$\chi^2 = \sum \frac{\textbf{(Observed – Expected)}^2}{\textbf{Expected}}$$

If the model under the null hypothesis is reasonable then the expected frequencies will be 'close' to the observed frequencies and the difference between the two small.

$$\sum \frac{(f_o - f_e)^2}{f_e}$$

If the differences between the observed and expected are small then the Chi-square value will be small and will yield a reasonable value from the Chi-square distribution (p-value >.05).

If they are large, the Chi-square statistic will be large and give a small p-value. In our example, we have (substituting into the above formula):

$$= \frac{(18 - 12.5)^2}{12.5} + \frac{(17 - 12.5)^2}{12.5} + \frac{(7 - 12.5)^2}{12.5} + \frac{(8 - 12.5)^2}{12.5}$$

$$= \frac{30.25}{12.5} + \frac{20.25}{12.5} + \frac{30.25}{12.5} + \frac{20.25}{12.5}$$

$$= 2.42 + 1.62 + 2.42 + 1.62 = \boxed{8.08}$$

Under H_o this is a Chi-square (c-1) distribution. Since we have 4 categories the distribution is a Chi-square with 3 degrees of freedom, written below as:

$$\chi^2(4 - 1) \quad \text{or} \quad \chi^2_3$$

The p-value using the chisquare Table in the text is approximately equal to p-value = .03 (between .025 and .05 in the tables). We have evidence against H_o and conclude that the model is not adequate, **all proportions are not equal**, students mention some factors more often than others. Note the general form of the test:

$$\frac{\textbf{Observed - Expected}}{\textbf{Expected}}$$

The Observed - Expected numerator gives a measure of how reasonable the model is for the data assuming the null hypothesis is true. Large deviations indicate a poor fit whereas small deviations indicate an adequate model. Each term in the sum is a called a **residual**. A zero residual indicates the model fits the data perfectly. We can examine the **standardized residual (Z)** which is equal to:

$$\frac{(O - E)}{\sqrt{E}}$$

to see where deviations from H_o occur.

A standardized residual squared is equal to a Chi-square variable with one degree of freedom.

$$Z^2 = \chi^2_1$$

As you can see, the Chi-square test is a test based on the residuals of the model. If the residuals are small, the model fits the data well. The resulting Chi-square statistic will then be a reasonable ($p>.05$) value from the Chi-square distribution. This fact is apparent since the square of the standardized residual is a Chi-square distribution.

Summary

Inference for a **population proportion (p)** is based on the sample proportion. The sampling distribution of the proportion is normal and the test of hypothesis is based on a **Z-statistic**.

The comparison of **two population proportions** is based on the difference D of proportions $p_1 - p_2$. The test of significance is based on a Z statistic.

Remember: Tips for Success

1) Read the text.
2) Read the notes.
3) Try the assignment.
4) If needed, try the exercise questions.
5) Try the simulations and view the videos if you need more help with a concept.
6) Try the self tests for practice on each chapter of the text at www.whfreeman.com/ips
7) Steady Work = Success