Tests About Proportions

In this exercise, we will learn how to perform various large sample Z-tests about proportions using software. We will use the dataset `head injuries.xls` posted on Blackboard. This dataset displays the results of a study investigating the effectiveness of bicycle safety helmets in preventing head injury. The data consist of a random sample of 793 individuals involved in bicycle accidents during a specified one-year period.

The SAS code in this exercise can be found in the file Lab15.sas, which is posted on Blackboard.

1 SAS

1.1 Read In The Data

We will read in the dataset `head injuries.xls` using the SAS Import Wizard. Follow the steps below to import the dataset into SAS.

1. From the File menu in SAS, select 'Import Data ...'.
2. Check 'Standard Data Source' and select 'Microsoft Excel 97/2000/2002/2003 Workbook' from the pulldown menu.
3. When the window pops up, click Browse and locate your file ('H:\bios600\Data\head injuries.xls'). Click Open and then OK.
4. Select the table 'head injuries$' (Sheet title) from the pulldown menu and hit Next.
5. Select 'WORK' to put the file in the Work library. Name your SAS dataset in the 'Member:' box by typing 'Injuries'.
6. Click Finish.

After completing these steps, click on the Log to view it. You should see the following statement in blue.

```
NOTE: WORK.INJURIES data set was successfully created.
```

1.2 Descriptive Statistics

In previous exercises, we begin with descriptive statistics on one or more continuous variables to initially look at how the data are distributed. In this setting, we have only categorical variables, so we can't compute the usual summary statistics. For this type of data, a frequency table summary is a useful first step in assessing the distribution of one or more categorical variables.

To compute frequency tables from our data, we will use the FREQ (or frequency) procedure. The first PROC FREQ asks for a frequency table describing the variables injury and helmet. This output is somewhat difficult to read, as it provides you with row, column and total percentages of each count. The second PROC FREQ suppresses this extra information.

```
*get 2x2 table of injury and helmet ;
PROC FREQ DATA=injuries;
  TABLES helmet*injury ;
RUN ;

*get 2x2 table of injury and helmet, without various types of percentages ;
PROC FREQ DATA=injuries;
  TABLES helmet*injury / NOROW NOCOL NOPERCENT ;
RUN ;
```

1.3 Large Sample Test of One Proportion

1.3.1 Null Value: 0.5

Suppose we define a Success to be "wearing a helmet". We would like to see if the people in this sample are equally likely to wear a helmet as not. That is, we want to perform a large sample (approximate) 2-sided Z test about the proportion of people who wear helmets while riding a bike. We write these hypotheses as follows.

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5$$

For this hypothesis test, we will use the Normal Approximation to the Binomial Z statistic discussed in Chapter 16. Recall that this test statistic is given by

$$Z = \frac{\hat{p} - p_0}{SE_p} = \frac{\hat{p} - 0.5}{SE_p}, \qquad \text{where } SE_p = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{0.5(0.5)}{n}}$$

To carry out a Z-test on a proportion in SAS, we use the FREQ procedure. We implement the above hypothesis test in SAS by submitting the following three lines of code.

```
PROC FREQ DATA=injuries ORDER=data;
  TABLES helmet / NOROW NOCOL NOPERCENT   BINOMIAL(P=0.5) ;
RUN ;
```

Assuming everything runs correctly, SAS will perform the desired one-sample 2-sided hypothesis test about a proportion and generate the results. The output is shown on the next page.

The FREQ Procedure

Helmet

Helmet	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	147	18.54	147	18.54
No	646	81.46	793	100.00

Binomial Proportion for Helmet = Yes

Proportion	0.1854
ASE	0.0138
95% Lower Conf Limit	0.1583
95% Upper Conf Limit	0.2124
Exact Conf Limits	
95% Lower Conf Limit	0.1589
95% Upper Conf Limit	0.2142

Test of H0: Proportion = 0.5

ASE under H0	0.0178
Z	-17.7200
One-sided Pr < Z	<.0001
Two-sided Pr > |Z|	<.0001

Sample Size = 793

SAS output often includes lots of things you don't need. The most important things that you'll need from this output are

1. First table - One-Way frequency table on the variable Helmet.
2. Second table - Estimates based on binomial distribution.
   - Estimated proportion of successes $\hat{p}$
   - Asymptotic and Exact standard errors.
   - Asymptotic and Exact 95% confidence intervals.

3. Third table - Estimates based on Normal approximation.

   - Asymptotic standard error under H_0. $\left(SE_p = \sqrt{\frac{p_0 q_0}{n}}\right)$
   - Large sample test statistic $Z = \frac{\hat{p} - 0.5}{SE_p}$
   - 1- and 2-sided p-values.

1.3.2 Null Value: Not 0.5

In §1.3.1, we performed a typical one-sample Z-test about a proportion, where we simply wanted to know where the probability of success was different from the probability of failure. Occasionally however, we want to compare a proportion to some other null value that is thought to describe the general population. That is, we are interested in the more general hypothesis test

$$H_0 : p = p_0$$
$$H_A : p \neq p_0$$

where p_0 can now be any number. For instance, if we have reason to believe the probability of wearing a helmet is 20% in the general population, then we may want to compare this sample with the general population.

To do this in SAS, then we modify our previous code very slightly. In the `TABLES` statement, we change the value for the option `P=` to whatever we have chosen as our value of p_0. In this example, we choose $p_0 = 0.2$, so we would submit the following code.

```
PROC FREQ DATA=injuries ORDER=data;
  TABLES helmet / NOROW NOCOL NOPERCENT   BINOMIAL(P=0.2) ;
RUN ;
```

1.3.3 Varying Significance Level

Both of the Z-tests above were conducted at the significance level $\alpha = 0.05$, which is the default level. We can change this in the `TABLES` statement by adding the SAS Keyword option `ALPHA=`, along with whatever value we prefer. For instance, to conduct the previous test at the $\alpha = 0.01$ significance level, we would use the following code.

```
PROC FREQ DATA=injuries ORDER=data;
  TABLES helmet / NOROW NOCOL NOPERCENT   BINOMIAL(P=0.2) ALPHA=0.01;
RUN ;
```

1.4 Large Sample Comparison of Two Proportions

1.4.1 Calculations By Hand

We would now like to compare head injury rates between the population of people who wear helmets (group 1) and those who do not (group 2). That is, we want to perform the 2-sample test of proportions using the Normal approximation to the binomial from Chapter 17.

First we define Success as "received head injury". To compare proportion of head injuries between the two groups, we will conduct the following hypothesis test.

$$H_0 : p_1 = p_2$$

$$H_A : p_1 \neq p_2$$

Our test statistic is given by

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\hat{p}_1 - \hat{p}_2}}, \qquad \text{where } SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ and } \bar{p} = \frac{a_1 + a_2}{n_1 + n_2}$$

Remember that $\bar{p}$ is the proportion of successes (head injuries) in the study (both groups combined). If we perform the calculations by hand, we get

$$\hat{p}_1 = \frac{17}{147} = 0.1156, \quad \hat{p}_2 = \frac{218}{646} = 0.3375, \quad \bar{p} = \frac{235}{793} = 0.2963$$

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{235}{793}\left(\frac{558}{793}\right)\left(\frac{1}{147} + \frac{1}{646}\right)} = 0.04173$$

$$Z = \frac{0.1156 - 0.3375}{0.04173} = \frac{-0.2219}{0.04173} = -5.3156$$

.

1.4.2 Equivalence to χ^2 Test

SAS does not actually perform this test. However, we can obtain the results of this test using the equivalent chi-square test of general association for frequency tables discussed in Chapter 18. In this setting, we have a row variable with r levels (or rows), and a column variable with c levels (or columns). The hypotheses of this test are

$$H_0 : \text{ row and column variables are independent (not associated)}$$

$$H_A : \text{ row and column variables are not independent}$$

For our data, our row variable is `Helmet` with $r = 2$ levels and the column variable is `Injury` with $c = 2$ levels. We can calculate a χ^2 statistic where the degrees of freedom are given by $df = (r-1)(c-1)$.

To perform the χ^2 test in SAS, we use the following code.

```
PROC FREQ DATA=injuries ORDER=data ;
  TABLES helmet*injury / NOROW NOCOL NOPERCENT   EXPECTED CHISQ ;
RUN ;
```

The output is displayed here and described on the next page.

Table of Helmet by Injury

Helmet	Injury		
Frequency| Expected |	Yes	No	Total
Yes	17	130	147
	43.562	103.44	
No	218	428	646
	191.44	454.56	
Total	235	558	793

Statistics for Table of Helmet by Injury

Statistic	DF	Value	Prob
Chi-Square	1	28.2555	<.0001
Likelihood Ratio Chi-Square	1	32.5432	<.0001
Continuity Adj. Chi-Square	1	27.2018	<.0001
Mantel-Haenszel Chi-Square	1	28.2199	<.0001
Phi Coefficient		-0.1888	
Contingency Coefficient		0.1855	
Cramer's V		-0.1888	

Fisher's Exact Test

Cell (1,1) Frequency (F)	17
Left-sided Pr <= F	1.249E-08
Right-sided Pr >= F	1.0000
Table Probability (P)	9.395E-09
Two-sided Pr <= P	2.273E-08

Sample Size = 793

1. First Table - Display observed and expected frequencies for calculation of test statistic.
2. Second Table - Various χ^2 test statistics, degrees of freedom and p-values. The χ^2 test statistic for general association (what we want) is the first entry.
3. Third Table - Test statistic and multiple p-values for Fisher's Exact Test of 2×2 tables.

From this output we get $\chi^2 = 28.2555$ with $df = (2-1)(2-1) = 1$ and $p < 0.0001$, so we reject H_0 and conclude that the two variables are not independent. However, we notice that

$$\chi^2 = 28.2555 = (-5.3156)^2 = Z^2$$

Therefore, the χ^2 test of general association is equivalent to the large sample Normal approximation test of 2 proportions. So, by this conclusion, we can equivalently reject the previous H_0 from §1.4.1. That is, we have strong evidence to suggest that the proportion of head injuries differs significantly between the two populations. The risk of a head injury while riding a bicycle is significantly smaller for people who wear helmets than people who don't wear helmets.

2 Excel

Unfortunately, the Analysis Toolpak does not have any functions available to tests of proportions. There are various macros available online, that may be downloaded and used easily. However, these are written by other users of Excel and it is important to check that their macros are actually doing what you want them to do (or think they are doing).

Rather than looking around for one of these, we can just use the Excel built-in functions to manually calculate the frequency tables from our data and the statistics necessary to perform our hypothesis test.

2.1 Large Sample Test of One Proportion

As in the previous section, we will start with a one-sample test about the proportion of helmet-wearers in our sample.

2.1.1 Summarize Data in Frequency Table

The first step is to create a frequency table of the Helmet data, so we can get an idea of what the data look like and how the variable is distributed.

1. Click Insert to select the Insert tab.
2. Click PivotTable (left-most entry in tab).
3. In the Table/Range entry box, press the diagonal red arrow button and select the Helmet Data in cells B1:B794.

4. Click the vertical red arrow button to return to the PivotTable dialog box.
5. At the bottom, select "Existing Worksheet" and press the diagonal red arrow button.
6. Select cell H3 to place the table, then press the the vertical red arrow button.
7. Click OK.

Now that we have a place for our frequency table, we need to modify the layout of the table. To do this we will use the Pivot Table Field List at right.

1. Select Helmet in 'Choose Fields' box at top.
2. Drag Helmet from 'Choose Fields' box to 'Values' box at bottom right.
3. Close the Pivot Table Field List.

2.1.2 Null Value: 0.5

Since the Analysis Toolpak doesn't have any functions for testing proportions, we must calculate the test statistic and p-value manually using built-in Excel functions. As before, we will test $H_0 : p = 0.5$ vs. $H_A : p \neq 0.5$, to see if the probability of success (i.e., wearing a Helmet) is the same as the probability of failure (i.e., not wearing a Helmet).

1. **Calculate $\hat{p}$**: Select cell E4. Type "=147/793" and press Enter.
2. **Calculate $\hat{q}$**: Select cell E5. Type "=1-E4" and press Enter.
3. **Calculate q_0**: Select cell E8. Type "=1-E7" and press Enter.
4. **Calculate $SE_{\hat{p}}$**: Select cell E9. Type "=SQRT(E7*E8/793)" and press Enter.
5. **Calculate Z_{stat}**: Select cell E11. Type "=(E4-E7)/E9" and press Enter.
6. **Calculate 1-sided p-value**: Select cell E12. Type "=NORMDIST(E11,0,1,1)" and press Enter.
7. **Calculate 2-sided p-value**: Select cell E13. Type "=2*E12" and press Enter.

2.1.3 Null Value: Not 0.5

Now we would like to test the more general hypotheses $H_0 : p = p_0$ vs $H_A : p \neq p_0$. As before, we are willing to assume that the probability of wearing a helmet in the general population is 20%, so we choose $p_0 = 0.2$.

In this setting, we must update our q_0, standard error $SE_{\hat{p}}$, test statistic Z_{stat} and p-value.

1. **Calculate q_0**: Select cell F8. Type "=1-F7" and press Enter.

2. **Calculate $SE_{\hat{p}}$**: Select cell F9. Type "=SQRT(F7*F8/793)" and press Enter.

3. **Calculate Z_{stat}**: Select cell F11. Type "=(F4-F7)/F9" and press Enter.

4. **Calculate 1-sided *p*-value**: Select cell F12. Type "=NORMDIST(F11,0,1,1)" and press Enter.

5. **Calculate 2-sided *p*-value**: Select cell F13. Type "=2*F12" and press Enter.

2.2 Large Sample Comparison of Two Proportions

As in the SAS section, we would now like to compare head injury rates between the two groups (helmet-wearers and non-helmet-wearers). That is, we want to test $H_0 : p_1 = p_2$ versus the two-sided alternative $H_A : p_1 \neq p_2$. We will continue to use the Normal approximation to the binomial distribution.

2.2.1 Summarize Data in Frequency Table

The first step is to create a frequency table of the Helmet and Injury data, so we can get an idea of what the data look like and how the variables are distributed.

1. Click Insert to select the Insert tab.

2. Click PivotTable (left-most entry in tab).

3. In the Table/Range entry box, press the diagonal red arrow button and select the Injury and Helmet Data in cells A1:B794.

4. Click the vertical red arrow button to return to the PivotTable dialog box.

5. At the bottom, select "Existing Worksheet" and press the diagonal red arrow button.

6. Select cell H18 to place the table, then press the the vertical red arrow button.

7. Click OK.

Now that we have a place for our frequency table, we need to modify the layout of the table. To do this we will use the Pivot Table Field List at right.

1. Drag Helmet from 'Choose Fields' box to 'Row Labels' box at bottom left.

2. Drag Injury from 'Choose Fields' box to 'Column Labels' box at right.

3. Drag Helmet from 'Choose Fields' box to 'Values' box at bottom right.

4. Close the Pivot Table Field List.

Since the Analysis Toolpak doesn't have any functions for testing proportions, we must calculate the test statistic and p-value manually using built-in Excel functions.

1. **Calculate n_1**: Select cell E21. Type "=147" and press Enter.
2. **Calculate n_2**: Select cell F21. Type "=646" and press Enter.
3. **Calculate $\hat{p}_1$**: Select cell E22. Type "=17/147" and press Enter.
4. **Calculate $\hat{p}_2$**: Select cell F22. Type "=218/646" and press Enter.
5. **Calculate $\hat{q}_1$**: Select cell E23. Type "=1-E22" and press Enter.
6. **Calculate $\hat{q}_2$**: Select cell F23. Type "=1-F22" and press Enter.
7. **Calculate $\bar{p}$**: Select cell E26. Type "=235/793" and press Enter.
8. **Calculate $\bar{q}$**: Select cell E27. Type "=1-E26" and press Enter.
9. **Calculate $SE_{\hat{p}_1-\hat{p}_2}$**: Select cell E28. Type "=SQRT(E26*E27*(1/147+1/646))" and press Enter.
10. **Calculate Z_{stat}**: Select cell E30. Type "=(E22-F22)/E28" and press Enter.
11. **Calculate 1-sided p-value**: Select cell E31. Type "=NORMDIST(E30,0,1,1)" and press Enter.
12. **Calculate 2-sided p-value**: Select cell E32. Type "=2*E31" and press Enter.

3 References

1. UCLA ATS- Statistical Tests in SAS
   http://www.ats.ucla.edu/stat/sas/whatstat/whatstat.htm