Inference about a single population proportion

Diana M. Bailey (The American Journal of Occupational Therapy, 1990) conducted a study to examine the reasons why occupational therapists have left the field of occupational therapy. Her sample consisted of female certified occupational therapists who had left the profession either permanently or temporarily. Out of 696 subjects who responded to the data-gathering survey, 438 (or 63%) had planned to take time off from their jobs to have and raise children. On the basis of these data, we wish to compute a confidence interval for the unknown proportion in the sampled population whose reason for leaving the field is other than taking time off to have and raise children.

1. What is the sampled population?

2. What is/are the population parameter(s) of interest?

3. Is this a one-sample, paired-sample, or two-independent-sample problem?

4. Are the rules of thumb met so that we can use a normal approximation to carry out our test?

5. What is the point estimate for \( p \), the proportion of occupational therapists who leave the field for reasons other than having and raising kids?

6. What is the 95% confidence interval? What does the confidence interval mean?

7. At the \( \alpha = .01 \) significance level, carry out a hypothesis test of the hypotheses:

\[
H_0 : p = 0.25 \quad \quad \quad H_a : p \neq 0.25
\]

8. Can you reject \( H_0 \)? What does this mean substantively?

9. Interpret the p-value.

SAS code

Creating the dataset:

```sas
data leave ;
input child $ count ;
datalines ;
Y 438
N 258 ;
```

Proc freq makes a table of counts and percents.

```sas
proc freq data = leave ;
tables child ;
weight count ;
run ;
```

SAS output

<table>
<thead>
<tr>
<th>child</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>258</td>
<td>37.07</td>
<td>258</td>
<td>37.07</td>
</tr>
<tr>
<td>Y</td>
<td>438</td>
<td>62.93</td>
<td>696</td>
<td>100.00</td>
</tr>
</tbody>
</table>
To carry out a one-sample \( z \) test of the hypothesis

\[ H_0 : p = p_0 \]

\[ H_a : p \neq p_0 \]

add the \( \text{binomial (} p = p_0 \) \) option on the end of the \( \text{tables} \) statement. The following code tests the null hypothesis that the population proportion of occupational therapists leaving the field for reasons other than to have and raise kids is 0.25. Note that it also automatically produces a 95% c.i. for \( p \).

```bash
proc freq data = leave ;
tables child / binomial (p = 0.25) ;
weight count ;
run ;
```

SAS output

<table>
<thead>
<tr>
<th>child</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>258</td>
<td>37.07</td>
<td>258</td>
<td>37.07</td>
</tr>
<tr>
<td>Y</td>
<td>438</td>
<td>62.93</td>
<td>696</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Binomial Proportion for child = N

- Proportion: 0.3707
- ASE: 0.0183
- 95% Lower Conf Bound: 0.3348
- 95% Upper Conf Bound: 0.4066
- Exact Conf Bounds
  - 95% Lower Conf Bound: 0.3347
  - 95% Upper Conf Bound: 0.4078

Test of \( H_0 \): Proportion = 0.25

- A\(SE\) under \( H_0 \): 0.0164
- Z: 7.3532
- One-sided Pr > Z: <.0001
- Two-sided Pr > |Z|: <.0001

To get a level \( 1 - \alpha \) confidence interval for the true population proportion \( p \), add the \( \text{binomial alpha = alpha0} \) option to the end of the \( \text{tables} \) statement. This code requests a 95% c.i. To get a 99% c.i., you would specify alpha = .01. Note that this code also automatically produces a hypothesis test of \( H_0 : p = 0.5 \).

```bash
proc freq data = leave ;
tables child / binomial alpha = .05 ;
weight count ;
run ;
```

SAS output

<table>
<thead>
<tr>
<th>child</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>258</td>
<td>37.07</td>
<td>258</td>
<td>37.07</td>
</tr>
<tr>
<td>Y</td>
<td>438</td>
<td>62.93</td>
<td>696</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Binomial Proportion for child = N

- Proportion: 0.3707
- ASE: 0.0183
- 95% Lower Conf Bound: 0.3348
- 95% Upper Conf Bound: 0.4066
- Exact Conf Bounds
  - 95% Lower Conf Bound: 0.3347
  - 95% Upper Conf Bound: 0.4078

Test of \( H_0 \): Proportion = 0.5

- A\(SE\) under \( H_0 \): 0.0190
- Z: -6.8229
- One-sided Pr < Z: <.0001
- Two-sided Pr > |Z|: <.0001

2  Comparing two population proportions

Research has suggested that alcoholism may be related to clinical depression. An investigation by Winokur and Coryell (American Journal of Psychiatry, 1991), ex-
explored this possible relationship. In 210 families of females with clinical depression, they found that alcoholism was present in 89. In 299 control families, alcoholism was present in 94. Do these data provide evidence that alcoholism occurs in a different proportion of families in which unipolar major depression occurs than in which there is no diagnosis of depression? Carry out a hypothesis test at the $\alpha = .05$ significance level.

1. What is/are the populations of interest?

2. What is/are the population parameters of interest?

3. Is this a one-sample, paired-sample, or two-independent-sample problem?

4. Is the hypothesis one- or two-sided?

5. What are the null and alternative hypotheses for the test?

6. Are the rules of thumb met so that we can use a normal approximation to carry out our test?

7. If the null hypothesis is true, what is our best estimate based on this data of the common proportion of alcoholism in both populations of families?

8. What is your conclusion based on the statistical analysis?

First we must key in our data.

data depress ;
input depress $ alcohol $ count ;
datalines ;
Y Y 89
Y N 121
N Y 94
N N 205
run ;

Next we use the Chi square test option of proc freq to do the hypothesis test.

proc freq data = depress ;
tables depress * alcohol / chisq ;
weight count ;
run ;

SAS output

<table>
<thead>
<tr>
<th>TABLE OF DEPRESS BY ALCOHOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEPRESS</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Row Pct</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STATISTICS FOR TABLE OF DEPRESS BY ALCOHOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Proc freq for data read in from a dataset of individual observations

Do not use the weight statement in proc freq if each observation should be given weight = 1. Here is an example problem based on the datasets “dieldrin.dat” from the course web page.

Stacy, Perriman, and Whitney (1985) studied pesticide residues in human milk in Western Australia in 1979-80. Earlier research had discovered high pesticide levels. Stacey et al. hoped to show that levels had decreased due to stronger government controls over the use of pesticides on food crops. They did find decreases for several types of pesticides, but levels of dieldrin had increased substantially.

This dataset has information from 45 donors. The variables are:

- age in years
- whether they lived in a new suburb (0 = old, 1 = new)
- whether their house was treated for termites within the past 3 years (0 = no, 1 = yes, two missing values)
- whether their milk contained above-average levels of dieldrin (0 = no, 1 = yes; above average defined as > .009 parts per million)

Termites are a common problem in Western Australia, and dieldrin is often used to control them. By law, new houses must be pretreated for termites.

If this sample of 45 donors can be considered a simple random sample of Western Australian mothers who live in suburbs, find a point estimate and 99% confidence interval for the proportion of such women whose milk does not contain above-average levels of dieldrin.

4 The Chi-square test for differences among more than two population proportions

Patty J. Hale (Family and Community Health, 1990) mailed a questionnaire to survey businesses as to whether they provided AIDS education for employees. The following table shows her results, broken down by size of business.

<table>
<thead>
<tr>
<th>Number of employees</th>
<th>AIDS educ provided?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>2</td>
</tr>
<tr>
<td>50-500</td>
<td>5</td>
</tr>
<tr>
<td>500+</td>
<td>11</td>
</tr>
</tbody>
</table>

May we conclude on the basis of these data that the proportion of businesses that provide AIDS education to their employees is different for different sizes of companies?

1. How should we enter this data into a SAS data step?

2. What is/are the populations of interest?

3. What is/are the population parameters of interest?
4. Is this a one-sample, paired-sample, or two-independent-sample problem, or something else?

5. What are the null and alternative hypotheses for the test?

6. Are the rules of thumb met so that we can use a normal approximation to carry out our test?

7. What is your conclusion based on the statistical analysis?