

22S:30/105, Statistical Methods and Computing

Instructor: Cowles
Lab 5, Apr. 8, 2009
Inference for Proportions

1 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 to take 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:

$$H_0 : p = 0.25$$

$$H_a : p \neq 0.25$$

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

9. Interpret the p-value.

SAS code

Creating the dataset:

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

Proc freq makes a table of counts and percents.

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

SAS output

The FREQ Procedure				
			Cumulative	Cumulative
child	Frequency	Percent	Frequency	Percent
N	258	37.07	258	37.07
Y	438	62.93	696	100.00

To carry out a one-sample z test of the hypothesis

$$H_0 : p = p_0$$

$$H_a : p \neq p_0$$

add the *binomial* ($p = p_0$) option on the end of the *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 .

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

SAS output

The FREQ Procedure				
child	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	258	37.07	258	37.07
Y	438	62.93	696	100.00

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 H0: Proportion = 0.25	
ASE under H0	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 *binomial alpha = alpha0* option to the end of the *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 also produces a hypothesis test of $H_0 : p = 0.5$.

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

SAS output

The FREQ Procedure				
child	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	258	37.07	258	37.07
Y	438	62.93	696	100.00

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 H0: Proportion = 0.5	
ASE under H0	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-

plored 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 OF DEPRESS BY ALCOHOL

DEPRESS	ALCOHOL		Total
	N	Y	
N	205	94	299
	40.28	18.47	58.74
	68.56	31.44	
	62.88	51.37	
Y	121	89	210
	23.77	17.49	41.26
	57.62	42.38	
	37.12	48.63	
Total	326	183	509
	64.05	35.95	100.00

STATISTICS FOR TABLE OF DEPRESS BY ALCOHOL

Statistic	DF	Value	Prob
-----------	----	-------	------

```

-----
Chi-Square          1      6.415      0.011
Likelihood Ratio Chi-Square  1      6.385      0.012
Continuity Adj. Chi-Square  1      5.949      0.015
Mantel-Haenszel Chi-Square  1      6.402      0.011
Fisher's Exact Test (Left)
                    (Right)          0.996
                    (2-Tail)        7.46E-03
                    (2-Tail)        0.015

Phi Coefficient          0.112
Contingency Coefficient  0.112
Cramer's V              0.112

```

Sample Size = 509

3 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 $\geq .009$ 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.

```

data milk ;
infile '/group/ftp/pub/kcowles/datasets/dieldrin.dat' ;
input age newburb termite above ;
run ;

proc freq data = milk ;
tables above / binomial alpha = .01 ;
run ;

```

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.

Number of employees	AIDS educ provided?	
	Yes	No
0-50	2	20
50-500	5	11
500+	11	5

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?