**STAT:2100/4200, Statistical Methods and Computing. Lab 5**

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 oc- cupational 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 population whose reason for leaving the field is **other than** taking time off to have and raise children.

* 1. What is the population?

occupational therapists

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

Female therapists

Other than to have

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

one-sample

* 1. Are the rules of thumb met so that we can use a normal approximation to carry out our test? To form a CI for p

At least 10 times the original

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

1-63%

* 1. What is the 95% confidence interval? What does the confidence interval mean?

± z\* sqt( (1-)/n ) z\* = 1.96

* 1. At the *α* = *.*01 significance level, carry out a hypothesis test of the hypotheses:

*H*0 : *p* = *.*25

*Ha* : *p ƒ*= *.*25

* 1. Can you reject *H*0? What does this mean?

Cal p – hypo p

( – p0) / sqt( (1-)/n )

p-value is relatively small, we reject the null hypothesis.

* 1. 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

Cumulative Cumulative 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

*Ha* : *p ƒ*= *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 ther- apists 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 ;

\*\*\*\*\*add the option to perform the test \*\*\*\*\*\*\*\*\*\*;

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 |

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 *− α* 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., your would specify alpha = .01. Note that this code also automatically produces a hypothesis test of *H*0 : *p* = 0*.*5.

proc freq data = leave ;

tables child / binomial alpha = .01 ; 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 |

Binomial Proportion for child = N

--------------------------------



|  |  |  |
| --- | --- | --- |
| Proportion |  | 0.3707 |
| ASE |  | 0.0183 |
| 99 Lower Conf | Bound | 0.3235 |
| 99 Upper Conf | Bound | 0.4178 |

Exact Conf Bounds



99 Lower Conf Bound 0.3238 99 Upper Conf Bound 0.4193



Test of H0: Proportion = 0.5 ASE under H0 0.0190

Z -6.8229

One-sided Pr < Z <.0001

Two-sided Pr > |Z| <.0001

1. 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 differ- ent 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 *α* = *.*05 significance level.

* 1. What is/are the populations of interest?

family w/ female depression family w/o female depression

* 1. What is/are the population parameters of interest?

proportion of alcoholism proportion of alcholism

* 1. Is this a one-sample, paired-sample, or two-independent-sample problem?
  2. Is the hypothesis one- or two-sided?

Do these data provide evidence that alcoholism occurs in a differ- ent proportion of families in

No certain directions

Η0: Pd = Pn

Ηa: Pd ≠ Pn

* 1. What are the null and alternative hypotheses for the test?
  2. Are the rules of thumb met so that we can use a normal approximation to carry out our test?

299-94 > 5 …

* 1. 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?

= 89+94 / 210+299 = 35.95

* 1. 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 ; weightassociation btw depress and alcohol count ;

run ;

SAS output

TABLE OF DEPRESS BY ALCOHOL DEPRESS ALCOHOL

Frequency| Percent | Row Pct |

Col Pct |N |Y | Total

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ---------+--------+--------+ | | | | | | |
| 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 (2-1)\*(2-1) | 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) (2-Tail)  Phi Coefficient |  | 0.112 | 0.996  7.46E-03  0.015 |
| Contingency Coefficient |  | 0.112 |  |
| Cramer’s V |  | 0.112 |  |
| Sample Size = 509 |  |  |  |

1. 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.

data milk ;

input age newburb termite above ; datalines;

|  |  |  |
| --- | --- | --- |
| 33 1 | 0 | 1 |
| 34 0 | 1 | 1 |
| ... |  |  |
| 23 0 | 1 | 0 |
| ; |  |  |

run ;

proc freq data = milk ;

tables above / binomial alpha = .01 ; run ;

Further, we want to test if the (population) proportion of women whose milk does not contain “above-average” levels of dieldrin are different for those whose house was pretreated for termites, versus those whose house was not pretreated for termites.

Two sample problem.

proc freq data = milk ;

tables termite \* above / chisq ; run ;