STAT:2100/4200, Statistical Methods and Computing. Lab 7

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

1. What is the population?
   female certified occupational therapists who leave the profession either permanently or temporarily.

2. What is/are the population parameter(s) of interest?
   The proportion in the population whose reason for leaving the field is **other than** taking time off to have and raise children. Let's denote it by p.

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

4. What population is actually sampled? Is it an SRS?

   It is not an SRS. People who left the field for controversial reasons would be less likely to respond than people who left for reasons they considered acceptable. For example, a woman who left the field after having been fired for stealing drugs from the hospital she worked for probably wouldn't answer the question. Some women would believe that leaving the field to have and raise children was a very good thing to do, and others would not. So this is a biased sample

5. Are the rules of thumb met so that we can use a normal approximation to carry out our test?
   Can the 696 subjects who responded to the survey be considered a simple random sample? No.
   $n\hat{p} = 696 - 438 > 10$ and $n(1-\hat{p}) = 438 > 10$.
   The population size is likely greater than 10*696.
   After all, the rules of thumb are basically met for us to be able to use the normal approximation method to make inference for p based on this data.

6. What is the point estimate for p, the proportion of occupational therapists who leave the field for reasons **other than** having and raising kids?
   $\hat{p} = (696 - 438)/696 \approx 37.07\%$

7. What is the 95% confidence interval? What does the confidence interval mean? We used to use the normal approximation method to get confidence interval for p, the formula is $\hat{p} +- z * \sqrt{\hat{p}(1-\hat{p}/n)}$. Now that we have SAS, this interval is provided in the output $(33.48\%, 40.66\%)$.
   Actually, it is better to report the EXACT confidence interval in SAS output $(33.47\%, 40.78\%)$.

8. At the $\alpha = .01$ significance level, carry out a hypothesis test of the hypotheses:

$$H_0 : p = 0.25$$

$$H_a : p \neq 0.25$$

9. Can you reject H_0? What does this mean substantively? According to the SAS output (last line of page 3), the two-sided p-value is $< .0001$.

10. Interpret the p-value. This means that if $H_0 : p = 0.25$ is indeed true, then the probability that we observe a z test statistic value as extreme as or more extreme than what the current data yields is smaller than .0001. Hence the data suggest strong evidence against H_0.

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

$$H_0 : p = p_0$$

$$H_a : p \neq p_0$$

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., your 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 = .01 ;
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
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

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), 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? There are two populations of interest here. One is the population of all families that have femails with clinical depression. The other is the population of all normal families, or more specifically, families that do not have females with clinical depression.

2. What is/are the population parameters of interest?
   p_D = population proportion of depressed families in which alcoholism is present.

   p_N = population proportion of normal families in which alcoholism is present.

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

4. Is the hypothesis one- or two-sided? two sided. Because the question asks "Do these data provide evidence that alcoholism occurs in a DIFFERENT proportion..."

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

$$H_0 : p_D = p_N$$

$$H_a : p_D \neq p_N$$

6. Are the rules of thumb met so that we can use a normal approximation to carry out our test? Yes
   0. Each sample can be considered as a Simple random sample (SRS) from its population.
   1. I believe both populations are at least 10 times as large as the samples.
   2. The counts of successes and failures are 5 or more in each sample.

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? $(89 + 94)/(210 + 299) = .3595$. This number is also available in the SAS output (find in in the last line on page 6.)

8. **What is your conclusion based on the statistical analysis?** The p-value of the chi-square test is .011, smaller than the prespecified significance level $\alpha = .05$. So we reject $H_0 : p_D = p_N$. In the context of this problem, this means that the data shows evidence that acoholism is related to depression. More specifically, we can see that $\hat{p}_D = 42.38\% > \hat{p}_N = 31.44\%$, which suggests that families with depression problems are also more likely to have alcoholism problems compared to normal families. And the chi-square test suggest this difference is very unlikely to have been caused by pure chance.

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		
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	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)			0.996
(Right)			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 > .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 pretreated for termites.

Two sample problem.

```
proc freq data = milk ;
tables termite * above / chisq ;
run ;
```