Lab 13: Basic inferences for proportions

STAT218

The learning objective for this lab is to perform tests and construct confidence intervals using a normal model for population proportions.

library(oibiostat)  
data("nhanes.samp.adult.500")  
nhanes <- nhanes.samp.adult.500  
load('data/prop-cases.RData')

## Inference for a single proportion

To illustrate inference for a single proportion, we’ll use data on diabetes status for 500 NHANES respondents to perform inference on the prevalence of diabetes among the U.S. population.

The variable of interest is a categorical variable indicating whether or not the respondent has diagnosed diabetes (Yes/No). It is convenient to format the vector of Yes/No responses as a factor. Notice how this encodes the values as integers (2 and 1).

# format as factor  
diabetes <- factor(nhanes$Diabetes, levels = c('Yes', 'No'))  
  
# inspect  
str(diabetes)

Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...

To estimate the population proportion, one can simply make a table of the frequency distribution and convert to proportions, then perform calculations manually.

# sample proportions  
diabetes.props <- table(diabetes) |> prop.table()  
  
# point estimate and se  
p.hat <- diabetes.props[1]  
p.hat.se <- sqrt(p.hat\*(1 - p.hat)/length(diabetes))

While manual calculations can be used to produce an interval and test statistic, the prop.test() function streamlines the analysis. Relevant arguments are:

* null specifies null value for the test
* alternative specifies direction of the test
* conf.level specifies confidence level for the interval

# test whether prevalence is 10% or not  
table(diabetes) |> prop.test(p = 0.1, alternative = 'two.sided', conf.level = 0.99)

1-sample proportions test with continuity correction  
  
data: table(diabetes), null probability 0.1  
X-squared = 0.93889, df = 1, p-value = 0.3326  
alternative hypothesis: true p is not equal to 0.1  
99 percent confidence interval:  
 0.0814852 0.1568967  
sample estimates:  
 p   
0.114

|  |
| --- |
| Your turn |
| Modify the command below to test whether diabetes prevalence among U.S. adults exceeds 9%. Identify and interpret both the hypothesis test (*i.e.*, -value) and confidence interval in the R output.  # does prevalence exceed 9%? modify below table(diabetes) |> prop.test(p = 0.1, alternative = 'two.sided', conf.level = 0.99)  1-sample proportions test with continuity correction  data: table(diabetes), null probability 0.1 X-squared = 0.93889, df = 1, p-value = 0.3326 alternative hypothesis: true p is not equal to 0.1 99 percent confidence interval:  0.0814852 0.1568967 sample estimates:  p  0.114 |

## Inference for two proportions

This part of the lab will illustrate how to perform hypothesis tests and construct intervals for a difference in proportions. This type of analysis arises when binomial data are collected for two groups.

We’ll use data from an experiment investigating the effect of vitamin C on preventing common colds. The data are already formatted in a table. Notice the layout:

* rows are groups
* columns are categories
* first column is category of interest

This formatting should be consistent when you apply this example to the analysis of other datasets.

Sample statistics can be obtained using prop.table() with the appropriate margin argument. Margin 1 corresponds to normalizing by row totals; margin 2, by column totals.

# summary statistics  
cold |> prop.table(margin = 1)

Cold NoCold  
Placebo 0.8150852 0.1849148  
VitC 0.7420147 0.2579853

To test whether the proportions differ and compute a 95% confidence interval for the difference:

# test whether the proportions are the same  
prop.test(cold, alternative = 'two.sided', conf.level = 0.95)

2-sample test for equality of proportions with continuity correction  
  
data: cold  
X-squared = 5.9196, df = 1, p-value = 0.01497  
alternative hypothesis: two.sided  
95 percent confidence interval:  
 0.01391972 0.13222111  
sample estimates:  
 prop 1 prop 2   
0.8150852 0.7420147

Take a moment to identify the relevant outputs.

|  |
| --- |
| Your turn |
| Modify the example above to test whether the vitamin C treatment lowers the probability of contracting the common cold.  # does vitamin C reduce the probability of a cold? modify the command below prop.test(cold, alternative = 'two.sided', conf.level = 0.95)  2-sample test for equality of proportions with continuity correction  data: cold X-squared = 5.9196, df = 1, p-value = 0.01497 alternative hypothesis: two.sided 95 percent confidence interval:  0.01391972 0.13222111 sample estimates:  prop 1 prop 2  0.8150852 0.7420147 |

## Practice problems

1. Researchers categorized 3,112 individuals in American Samoa according to whether they were obese and recorded whether subjects had cardiovascular disease (CVD).Test for a difference in disease rates between obese and non-obese populations, and produce an interval estimate for the difference in proportions.

# data  
obesity

Deaths NonDeaths  
Obese 16 2045  
NotObese 7 1044

# compute summary statistics  
  
# perform test

1. Researchers identified 86 lung cancer patients and 86 controls (without lung cancer), and categorized them according to whether they were smokers or non-smokers. Test for a difference in the proportion of smokers among cancer patients compared with controls, and produce an interval estimate for the difference.

# data  
smoking

Smokers NonSmokers  
Cancer 83 3  
Control 72 14

# compute summary statistics  
  
# perform test