Lab 11: Inference for binomial proportions

STAT218

The goal of this lab is to learn how to implement:

1. tests and intervals for binomial proportions from one sample
2. tests and intervals comparing two proportions from independent samples

The activity represents our first foray into categorical data analysis. You’ll reproduce examples from lecture using the NHANES data to estimate diabetets prevalence and the Vitamin C experiment; you’ll practice on a few additional datasets.

library(tidyverse)  
load('data/vitamin.RData')  
load('data/nhanes500.RData')  
load('data/obesity.RData')

### Inference for one proportion

#### Refresher: categorical frequency distributions

Inference for proportions – and for that matter, future material on inference for categorical data – will leverage frequency distributions to perform calculations.

You learned how to make these descriptive summaries at the beginning of the quarter, but a refresher may be helpful. Recall that a frequency distribution, for a categorical variable, is simply a set of counts of observations of each unique value of the variable. We made these using table(...).

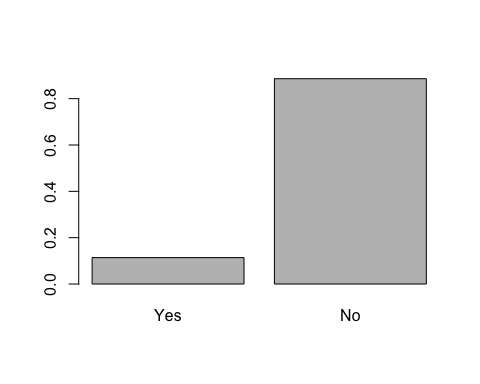
# extract variable of interest  
dia <- nhanes$diabetes  
  
# construct table of counts (frequency distribution)  
table(dia)

dia  
Yes No   
 57 443

# render as proportions (still frequency distribution, but normalized)  
table(dia) |> prop.table()

dia  
 Yes No   
0.114 0.886

# barplot  
table(dia) |> prop.table() |> barplot()



|  |
| --- |
| Your turn 1 |
| Compute the frequency distribution, in both counts and proportions, for the sleeptrouble variable in the NHANES dataset, which records whether the participant experiences sleep trouble. Also construct a barplot to visualize the frequency distribution |

#### Point estimation

The point estimate for a population proportion is simply the corresponding sample proportion:

This can be ascertained directly from the frequency distribution. In the diabetes example, .

The standard error for is:

This measures the sample-to-sample variability of . To compute the standard error for a binomial proportion, take the product of the two proportions and divide by the sample size, as shown below.

# point estimates \*are\* sample proportions in the frequency distribution table  
dia.p <- table(dia) |> prop.table()  
  
# point estimate (sample proportion of interest)  
dia.p.hat <- dia.p[1]  
dia.p.hat

Yes   
0.114

# standard error  
dia.n <- length(dia)  
dia.p.hat.se <- sqrt(prod(dia.p)/dia.n)  
dia.p.hat.se

[1] 0.01421295

This is interpreted as follows:

The proportion of U.S. adults with diagnosed diabetes is estimated to be 0.114 (SE = 0.0142).

|  |
| --- |
| Your turn 2 |
| Using the NHANES data, compute a point estimate and standard error for the proportion of U.S. adults who have sleep trouble. Interpret the estimate in context as shown above. |

#### Confidence intervals (manually)

The confidence interval is straightforward to compute by hand:

Once estimates are in hand, the calculation requires only finding an appropriate critical value . For a confidence interval, this is the quantile of the normal model, *i.e.*, the value satisfying:

Here are several examples:

# 90% interval  
cval <- qnorm(1 - 0.10/2)  
dia.p.hat + c(-1, 1)\*cval\*dia.p.hat.se

[1] 0.09062177 0.13737823

# 95% interval  
cval <- qnorm(1 - 0.05/2)  
dia.p.hat + c(-1, 1)\*cval\*dia.p.hat.se

[1] 0.08614313 0.14185687

# 99% interval  
cval <- qnorm(1 - 0.01/2)  
dia.p.hat + c(-1, 1)\*cval\*dia.p.hat.se

[1] 0.07738986 0.15061014

# 99.9% interval  
cval <- qnorm(1 - 0.001/2)  
dia.p.hat + c(-1, 1)\*cval\*dia.p.hat.se

[1] 0.0672319 0.1607681

To interpret the last interval:

With 99.9% confidence, the proportion of U.S. adults with diagnosed diabetes is estimated to be between 0.067 and 0.161, with a point estimate of 0.114 (SE 0.0142).

|  |
| --- |
| Your turn 3 |
| Using the NHANES data, construct and interpret a 98% confidence interval for the proportion of U.S. adults who have sleep trouble. |

#### Hypothesis tests

To test the hypothesis that a population proportion is some null value against a non-directional (two-sided) alternative, *i.e.*,

we use a test statistic that is *almost* the estimation error divided by the standard error, but we substitute the hypothetical value for the estimate in the standard error:

The -value is simply the proportion of the normal model for the sampling distribution of that exceeds the observed statistic in either direction: . This can be computed by hand as shown below.

# test stat  
dia.z <- (dia.p.hat - 0.1)/sqrt(0.1\*0.9/dia.n)  
  
# two sided p value  
dia.pval <- 2\*pnorm(abs(dia.z), lower.tail = F)  
dia.pval

Yes   
0.2967175

The result is interpreted as follows in standard narrative style:

The data provide no evidence that the proportion of U.S. adults with diagnosed diabetes differs from 0.1 (Z = 1.04, *p* = 0.2967).

|  |
| --- |
| Your turn 4 |
| Test the hypothesis that the proportion of U.S. adults who have sleep trouble is 0.3 against a two-sided alternative at the 5% significance level. Interpret the result in context.  # test stat  # two sided p value |

A more straightforward way to perform the inference – both the test and interval – utilizes prop.test(...). The command below matches our prior results:

# pass table (counts) to prop.test  
table(dia) |>   
 prop.test(p = 0.1, alternative = 'two.sided',   
 conf.level = 0.95, correct = F)

1-sample proportions test without continuity correction  
  
data: table(dia), null probability 0.1  
X-squared = 1.0889, df = 1, p-value = 0.2967  
alternative hypothesis: true p is not equal to 0.1  
95 percent confidence interval:  
 0.0890369 0.1448491  
sample estimates:  
 p   
0.114

The resulting test and interval would be reported jointly:

The data provide no evidence that the proportion of U.S. adults with diagnosed diabetes differs from 0.1 (Z = 1.04, *p* = 0.2967). > With 95% confidence, the proportion of U.S. adults with diagnosed diabetes is estimated to be between 0.089 and 0.145, with a point estimate of 0.114 (SE 0.0142).

The prop.test(...) function also makes doing directional tests easier. Below are two examples:

# upper sided test/interval: does p exceed 0.09?  
table(dia) |>   
 prop.test(p = 0.09, alternative = 'greater',   
 conf.level = 0.95, correct = F)

1-sample proportions test without continuity correction  
  
data: table(dia), null probability 0.09  
X-squared = 3.5165, df = 1, p-value = 0.03038  
alternative hypothesis: true p is greater than 0.09  
95 percent confidence interval:  
 0.09266984 1.00000000  
sample estimates:  
 p   
0.114

# upper sided test/interval: is p under 0.14?  
table(dia) |>   
 prop.test(p = 0.14, alternative = 'less',   
 conf.level = 0.95, correct = F)

1-sample proportions test without continuity correction  
  
data: table(dia), null probability 0.14  
X-squared = 2.8073, df = 1, p-value = 0.04692  
alternative hypothesis: true p is less than 0.14  
95 percent confidence interval:  
 0.000000 0.139485  
sample estimates:  
 p   
0.114

|  |
| --- |
| Your turn 5 |
| Test whether the proportion of U.S. adults who experience sleep trouble is under 0.3 at the 1% significance level. Provide an upper confidence bound along with your test.  # test whether the proportion with sleep trouble is under 0.3 |

The default approach is to apply the continuity correction (set correct = T or omit this argument). We omitted it so that results would match the manual calculations above. You should apply the correction in practice.

### Inference for two proportions

Inference comparing two proportions proceeds from a two-way table or “contingency” table. You may recall that this is a bivariate frequency distribution of two categorical variables. Take a moment to refresh your memory on how to construct these tables:

# variables of interest  
trt <- vitamin$vitamin  
out <- vitamin$autism  
  
# construct contingency table  
vitamin.tbl <- table(trt, out)  
vitamin.tbl

out  
trt autism no autism  
 no vitamin 111 70  
 vitamin 143 159

For inference to work appropriately with prop.test(...), it is important that the outcome be shown in the column dimension and the groups be shown in the row dimension. Further, the outcome of interest should be the first column (not the second).

|  |
| --- |
| Your turn 6 |
| Construct a contingency table for the obesity data.  # variables of interest  # contingency table |

The contingency table can be used directly to perform inference on a difference in proportions:

# test for difference in proportions  
prop.test(vitamin.tbl,   
 alternative = 'two.sided',   
 conf.level = 0.95)

2-sample test for equality of proportions with continuity correction  
  
data: vitamin.tbl  
X-squared = 8.3131, df = 1, p-value = 0.003936  
alternative hypothesis: two.sided  
95 percent confidence interval:  
 0.04475176 0.23474771  
sample estimates:  
 prop 1 prop 2   
0.6132597 0.4735099

|  |
| --- |
| Your turn 7 |
| Test whether the rate of CHD deaths differs among obese and non-obese populations. |