Making Inferences with Binary Data

Blake Pappas

10/19/2021

Inference on One Proportion Using R

Example: Defective Widgets

An engineer wants to learn the percentage of defective widgets that come out of the production line. The plant manager believe that 3% are defective, but she believes it might be higher. A simple random sample of 320 widgets is selected one day and the number of defects is counted. The code below produces some binary data that could arise in this scenario. (In this simulation, the engineer is correct and the true proportion is 0.04, not 0.03). The variable `defective` will equal 0 if the widget is not defective and 1 if it is.

```
# Generate a Data Set of Simulated Widgets

n <- 320  # Number of widgets to sample from the line
true.pi <- 0.04  # True proportion of defective widgets
defects <- rbinom(n, 1, prob = true.pi) # Indicator of whether it is defective

# Put It All Into a Data Frame
widget_data <- data.frame(widget_no = 1:n, defective = defects)
head(widget_data)
```

```
##   widget_no defective
## 1         1         1
## 2         2         1
## 3         3         0
## 4         4         0
## 5         5         0
## 6         6         0
```

Descriptive Analysis

First, we'll summarize the `defect` variable using the `table` and `prop.table` functions.

```
table(widget_data$defective)
```

```
##
##   0   1
## 308  12
```

```r
prop.table(table(widget_data$defective))
```

```
##
##      0      1
## 0.9625 0.0375
```

Confidence Interval for the Proportion

To make a 90% confidence interval for pi, the true proportion of widgets that are defective, we can use the `prop.test` function as shown below.

```r
successes <- sum(widget_data == 1)
n <- nrow(widget_data)
prop.test(successes, n, conf.level = 0.9)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  successes out of n, null probability 0.5
## X-squared = 268.28, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.02477101 0.06482758
## sample estimates:
##       p
## 0.040625
```

By default, this function prints a lot of information.

If you are only interested in the confidence interval, you can extract it by appending `$conf.int` to the call of `prop.test`.

```r
ci_pi <- prop.test(successes, n, conf.level = 0.9)$conf.int
ci_pi
```

```
## [1] 0.02477101 0.06482758
## attr(,"conf.level")
## [1] 0.9
```

Remember that the default settings in R use a continuity correction and a small-sample correction that will not give the exact same results as the large-sample intervals discussed in lecture.

Testing for pi

The same `prop.test` function performs a hypothesis test for π. By default, it tests the hypotheses

$$H_0 : \pi = 0.5; \quad H_A : \pi \neq 0.5.$$

You can change the null value (π_0) using the `p` option. You can change the alternative using the `alternative` option (the options are "two.sided", "less", and "greater"). The option `correct = FALSE` will result in the large-sample test from lecture being performed.

```r
test_result <- prop.test(successes, n, p = 0.03, alternative = 'greater', correct = FALSE)
print(test_result)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  successes out of n, null probability 0.03
## X-squared = 1.2414, df = 1, p-value = 0.1326
## alternative hypothesis: true p is greater than 0.03
## 95 percent confidence interval:
##  0.02599412 1.00000000
## sample estimates:
##        p
## 0.040625
```

You can extract the p-value for the test as follows:

```r
test_result$p.value
```

```
## [1] 0.1325998
```

You can also find the absolute value of the z_0 statistic using the code below. It's good practice to report this value as well as the p-value.

```r
sqrt(test_result$statistic)
```

```
## X-squared
##  1.114185
```

Example: Poisonous Mushrooms

The file `mushrooms.csv` contains records that represent a simple random sample of mushroom species from the Agaricus and Lepiota family. Today we'll look at the variable `edible`, which is equal to "e" if the species is edible and "p" if it is toxic.

a. How many species in the sample are edible and how many are poisonous? What proportion are edible?

```r
mushrooms <- read.csv("mushrooms.csv")

table(mushrooms$edible)
```

```
##
##   e   p
## 219 206
```

```r
prop.table(table(mushrooms$edible))
```

```
##
##         e         p
## 0.5152941 0.4847059
```

Answer: 219 species in the sample are edible, while 206 species are poisonous. Approximately 51.53% of mushrooms in the sample are edible.

b. Calculate a 99% large-sample confidence interval for the proportion of mushrooms that are edible. Use the formula for a large-sample interval.

```
point_estimate <- 219 / 425
point_estimate
```

```
## [1] 0.5152941
```

```
multiplier <- 1 - (0.01 / 2)
multiplier1 <- qnorm(multiplier)

standard_error <- sqrt(point_estimate * (1 - point_estimate) / 425)
standard_error
```

```
## [1] 0.02424221
```

```
upper_bound <- point_estimate - multiplier1 * standard_error
upper_bound
```

```
## [1] 0.4528503
```

```
lower_bound <- point_estimate + multiplier1 * standard_error
lower_bound
```

```
## [1] 0.5777379
```

Answer: (0.4528503, 0.5777379)

c. Use the `prop.test` function to find a 99% confidence interval for the proportion of mushrooms that are edible. Does it differ from the one you found in part b? If so, is the difference substantial enough to lead you to different conclusions?

```
prop.test(x = 219, n = 425, conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  219 out of 425, null probability 0.5
## X-squared = 0.33882, df = 1, p-value = 0.5605
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.4519339 0.5781753
## sample estimates:
##         p
## 0.5152941
```

Answer: The confidence interval based on the `prop.test` function is (0.4519339, 0.5781753). This indeed does differ from the one I found in part b. However, I do not think the difference is substantial enough to lead me to a different conclusion.

d. Use the `prop.test` function in R to test whether the data provide strong evidence that the proportion of edible species is not equal to 0.5. Report the test statistic, p-value, decision, and summarize your conclusion. Use $\alpha = 0.01$.

```
test_results <- prop.test(point_estimate, 425, conf.level = 0.99)
test_results
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  point_estimate out of 425, null probability 0.5
## X-squared = 420.95, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  8.257156e-08 1.979205e-02
## sample estimates:
##           p
## 0.001212457
```

```
sqrt(test_results$statistic)
```

```
## X-squared
##  20.51703
```

Answer: See above for the results of the hypothesis test. The test statistic is 20.51703 and the p-value is 0.001212457. Based on these findings, we can reject the null hypothesis. We can conclude that there is strong evidence in favor of the proportion of edible species not being equal to 0.5.

Example: Mushroom Gills

Estimate (make a confidence interval for) the proportion of mushrooms whose `gill.size` is broad. (For this variable, b = broad; n = narrow.) Report and interpret the interval. Use a reasonable confidence level.

```
table(mushrooms$gill.size)
```

```
##
##   b   n
## 288 137
```

```
prop.table(table(mushrooms$gill.size))
```

```
##
##         b         n
## 0.6776471 0.3223529
```

```r
b <- 288 / 425
b
```

```
## [1] 0.6776471
```

```r
prop.test(x = b, n = 425, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  b out of 425, null probability 0.5
## X-squared = 420.3, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.0000177256 0.0139026588
## sample estimates:
##           p
## 0.001594464
```

Answer: I used a 95% confidence level for the above confidence interval. The results I received were (0.0000177256, 0.0139026588). Therefore, we are 95% confident that between 0% and 1.4% of of mushrooms have broad gills.

Example: Colored Mushrooms

a. Look at the mushrooms for whose `cap.color` equal `w` (white). Test whether there is evidence that the proportion of these mushrooms that are edible is greater than 0.5. Use $\alpha = 0.05$.

Report the test statistic, p-value, and conclusion.

```r
sset <- subset(mushrooms, mushrooms$cap.color == 'w')
b <- table(sset$edible)
b
```

```
##
##  e  p
## 41 15
```

```r
sset1 <- prop.test(x = 41, n = 56, conf.level = 0.95, alternative = 'greater')
sset1
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  41 out of 56, null probability 0.5
## X-squared = 11.161, df = 1, p-value = 0.0004177
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.6163287 1.0000000
## sample estimates:
##         p
## 0.7321429
```

```r
sqrt(sset1$statistic)
```

```
## X-squared
##  3.340766
```

Answer. The test statistic for the above hypothesis test is 3.340766 and the p-value is 0.7321429. Based on these findings, we can fail to reject the null hypothesis. There is strong evidence in favor of the proportion of edible white species not being greater than 0.5.

b. Compare the findings to Exercise 1. Do the findings suggest any association between cap color and edibility?

Answer. These findings do not suggest any association between cap color and edibility.

Example: Writing a Function

a. Write an R function to calculate the large-sample confidence interval for π.

```r
my.pi <- function(x)
{
 standard_deviation <- sd(x)
 return(standard_deviation)
}
```

```r
my.pi <- function(x, n, conf)
{
large_sample_conf <- prop.test(x, n, conf.level = conf, correct = FALSE)
return(large_sample_conf)
}

# Example of Function in Use
my.pi(41, 56, 0.95)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 12.071, df = 1, p-value = 0.000512
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6040544 0.8304269
## sample estimates:
##         p
## 0.7321429
```