Week 6 Day 3

Stat140-04

Using built in functions: prop.test, t.test

RStudio has built in functions to do tests and intervals using the normal and t-distributions. The name of the function depends on the type of variables (categorical or quantitative). In short, if all variables are categorical, you should use `prop.test` to test for a single proportion or a difference in proportions; if all variabbles are quantitative, you should use `t.test` to test for a single mean or a difference in means.

1. Inference for Propotion

```
prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level =
0.95, data = NULL, success = NULL, ...)
```

- Default value of the null hypothesis is p=0.5.
- Two-sided test is the default.
- Just ignore the result of hypothesis test if you are looking for CI.
- Use the default `correct = FALSE` to match the formula you learned in class.
- In the output, X-squared is the square of the z-score, which is stored as `statistic` in the output list.

1.1 One-Sample Inference

(1) Suppose your data has 14 successes out of 50. The following performs a two-sided hypothesis testing (HT) with null value 0.3. ($H_0 : p = 0.3$ vs $H_a : p \neq 0.3$)

```
prop.test(x=14, n=50, p=0.3, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  14 out of 50, null probability 0.3
## X-squared = 0.095238, df = 1, p-value = 0.7576
## alternative hypothesis: true p is not equal to 0.3
## 95 percent confidence interval:
##  0.1747417 0.4166512
## sample estimates:
##    p
## 0.28
```

(2) The following performs a right-tail HT. ($H_0 : p = 0.3$ vs $H_a : p > 0.3$)

```r
prop.test(x=14, n=50, p =0.3, alternative="greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  14 out of 50, null probability 0.3
## X-squared = 0.095238, df = 1, p-value = 0.6212
## alternative hypothesis: true p is greater than 0.3
## 95 percent confidence interval:
##  0.1889395 1.0000000
## sample estimates:
##    p
## 0.28
```

1.2 Two-Sample Inference

(3) The following tests the alternative hypothesis that $p_1 - p_2$ is less than zero, with $\hat{p}_1 = 14/50$ and $\hat{p}_2 = 30/100$. ($H_0 : p_1 - p_2 = 0$ vs $H_a : p_1 - p_2 < 0$)

```r
prop.test(x=c(14, 30), n=c(50, 100), alternative = "less", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(14, 30) out of c(50, 100)
## X-squared = 0.064322, df = 1, p-value = 0.3999
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1088037
## sample estimates:
## prop 1 prop 2
##   0.28   0.30
```

2. Inference for Mean

```
t.test(x, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal
= FALSE, conf.level = 0.95, ...)
```

- Default value of the null hypothesis is mu=0.5.
- Two-sided test is the default.
- Just ignore the result of hypothesis test if you are looking for CI.
- `paired = FALSE` is the default for non-paired data.

2.1 One-Sample Inference I have data on the number of hours that 25 students slept. The dataset is loaded as `sleep` and the variable is named `hours`. Does this data set provide strong evidence that college students sleep 7 hours on avergae. ($H_0 : \mu = 7$ vs $H_a : \mu \neq 7$)

```r
t.test(sleep$hours, mu = 7)
```

2.2 Two-Sample Inference Since 2005, the American Community Survey polls approximately 3.5 million households yearly. We are interested in the distribution of salaries of males and females. I have loaded the dataset called `acs`. Is the average salaries of males higher than the average salaries of females in the U.S?

```r
# Save the data in two different vector

# extract income from all the females
female <- acs %>%
  filter(gender == "female") %>%
  pull(income)

# extract income from all the males
male <- acs %>%
  filter(gender == "male") %>%
  pull(income)

# Compute t-test
t.test(male, female, mu = 0, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = 8.1362, df = 1142.1, p-value = 5.281e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  14590.46      Inf
## sample estimates:
## mean of x mean of y
##  32627.30  14335.99
```

On your own

1. A large city's DMV claimed that 80% of candidates pass driving tests, but a survey of 90 randomly selected local teens who had taken the test found only 61 who passed. Does this finding suggest that the passing rate for teenagers is lower than the DMV reported? Set up the hypotheses, and use the `prop.test` function to compute the p-value.

2. Students were given words to memorize, then randomly assigned to take either a 90 min nap, or a caffeine pill. 2 and 1/2 hours later, they were tested on their recall ability. Is sleep or caffeine better for memory? Load the following data set in R, set up the hypotheses, and use the `t.test` function to compute the p-value.

```r
sleep.coffee <- read_csv(https://sshanshans.github.io/stat140/labs/data/sleepcoffee.csv)
```