HW7

Arjun Ganesh

04/08/2020

1. [10] Our goal is to estimate the average departure delay (dep_delay), in minutes, for all the flights leaving the three airports in NYC area.

(a) [2] Identify and describe the population and parameter of interest.

The population is all the flights leaving the three airports in the NYC area. Our parameter of interest is the average departure delay for all the flights leaving the three aiports in the NYC area.

(b) [2] Write a code to construct a 95% confidence interval for the population mean. [Note:be sure to include the output of your code.]

```
library(nycflights13)
t.test(flights$dep_delay,conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  flights$dep_delay
## t = 180.16, df = 328520, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  12.50157 12.77657
## sample estimates:
## mean of x
##  12.63907
```

(c) [3] Write a sentence interpreting your confidence interval in part (b).

We are 95% confident that the population mean is between an interval of 12.50157 min to 12.77657 min.

(d) [3] If an air traffic controller claim that the average departure delay for all the flights leaving the three airports in NYC area is 10 minutes. Use your confidence interval to make a decision to either reject or not to reject their claim? Explain briefly how you made this decision.

We would reject the air traffic controller's claim that the average departure delay for all the flights leaving the three airports in NYC area is 10 mins. The reason is because the potential average falls outside our confidence interval.

2. [20] Our next goal is to estimate the proportion of all flights with American Airline (AA)carrier leaving the three airports in NYC area.

(a) [2] Identify and describe the population and parameter of interest.

The population for our next goal is all flights with American Airline carrier leaving the three airports in NYC area. The parameter of interest is the proportion of all flights with American Airlines carrier out of all the flights with all carriers.

(b) [3] Write a code to isolate and count the number of AA flights that left the three airports in the NYC area. [Note: be sure to include the output of your code.]

```r
total =length(flights$carrier)
aa = length(which(flights$carrier=="AA"))
print(total)
```

```
## [1] 336776
```

```r
print(aa)
```

```
## [1] 32729
```

(c) [2] Write a code to construct a 85% confidence interval for the population proportion.[Note: be sure to include the output of your code.]

```r
prop.test(32729, 336776, conf.level = .85)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  32729 out of 336776, null probability 0.5
## X-squared = 218581, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 85 percent confidence interval:
##  0.09644953 0.09792202
## sample estimates:
##          p
## 0.09718329
```

(d) [3] Write a sentence interpreting your confidence interval in part (c).

We are 85% confident that population proportion of flights with carrier AA to all flights leaving the NYC area is between 0.09644953 to 0.09792202.

(e) [5] Conduct a hypothesis test to investigate your conjecture that more than 12% of all flights leaving the three airports in NYC area were (AA) carrier. Use a 5% level of significance. [Note: clearly state the null and alternative hypotheses in words, write a code to output the p-value, and write your conclusion in context.]

```r
prop.test(32729,336776,p=0.12,alternative = "less")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  32729 out of 336776, null probability 0.12
## X-squared = 1660.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.12
## 95 percent confidence interval:
##  0.00000000 0.09802758
## sample estimates:
##          p
## 0.09718329
```

NH: more than 12% of all flights leaving the three airpots in NYC area were (AA) carrier AH: less than 12% of all flights leaving the three airpots in NYC area were (AA) carrier

After running our hypothesis test, since our p value of 2.2.e-16 is less than .05, we can say that we reject the null hypothesis and the true proportion of all flights leaving the three airports in NYC were AA carrier is less than 12%.

(f) [5] Refer to part (e). Write the Type I and Type II errors in this context. Which error would you consider as the worst?

Type I would be when we reject the null hypothesis but actually it is true. So we say we reject that more than 12% of all NYC flights are AA carrier and it's less than but really the true proportion is indeed more than 12%. Type II would be a false negative meaning we accept a false null hypothesis. So we say that indeed more than 12% of all NYC flights are AA carrier but the true proporiton is indeed less than 12%.

If I was CEO of AA and had budget cuts, I would see where my airlines flies in comparison to other airlines. And if I see that I don't take up a good portion of flights in NYC then I would remove flights there. But if it turns out that wasn't the case, Type 1 error, then that would alot of money I lost for a miscalculation.