Lab 4

Group Member Names - here

Lab Overview

All students attending class in the group can turn in a single document with each participants name. Students not attending class will need to complete their own lab.

You have been recruited and hired as a statistician by the Capital Bikeshare. The Capital Bikeshare company has created a bike share system throughout Washington D.C. allowing members to rent bikes from a station and return them to any station throughout the city. Capital Bikeshare has over 3,500 bikes in their fleet spread accross over 400 stations. One challenge they face is that popular stations run out of bikes or do not have any docking stations available for returns during peak times. Your first task is to create a SRS of stations to determine the proportion of stations having a bike checked out between midnight and 5AM. The dataset *bikes.csv* can be downloaded with the following code

```
bikes <- read.csv('http://math.montana.edu/ahoegh/teaching/stat446/bikes.csv', header = T)
```

and contains actual data from June 23, 2016. The variable *rental* denotes whether a bike was checked out during the specified time frame, where a values of 1 indicates that a bike was rented in that time frame.

1. (5 points)

Compute the necessary sample size (n) for estimating p, the proportion of stations with a bike checked out at 2AM, to a maximum allowable difference of 0.05. The dataset contains a total of 392 stations (N=392). Justify your procedure.

```
#### Sample Size Calculations for Proportions
d <- .05 # maximum allowable difference
alpha <- .05 # 95% confidence interval
N <- 392 # population size
p <- 0.5 # use p= 1/2 for conservative estimate
n.0 <- (qnorm(1-alpha/2)^2 * p *(1-p)) / d^2
n = ceiling(1 / (1/n.0 + 1/N))
```

To guarantee that the total number of samples is sufficient for a maximum allowable difference (with p = .5 to be conservative) a total of 195 samples should be selected.

2. (5 points)

Draw the specified number of samples, from question 1, and compute a confidence interval for p. (Include your R code or hand written computations).

```
set.seed(10042019)
sample <- sample(bikes$rental,n)
p_hat <- mean(sample)
alpha <- .05
p_interval <- p_hat + c(-1,1) * qnorm(1 - alpha / 2)* sqrt(((N-n)/N)*p*(1-p)/(n-1))
```

The confidence interval is 0.166, 0.265.

3. (5 points)

How does your confidence interval for p compare with the truth? How does the confidence interval change if you had drawn another sample?

The true value for p is 0.2091837 which is in the interval above. If we had a different samples, we get somewhat different results. The value $\hat{p}$ changes, which results in different point estimate and interval widths.

4. (5 points)

Suppose on the night you conducted your sample, there was a thunderstorm. How would your results generalize to a typical evening? Now assume you have the ability to collect data from another evening, explain (qualitatively) what you would hope to accomplish with the study from the second evening.

A basic principles with sampling is that we hope to have a representative sample. Ultimately the question goes back to what our target population is. If our goal is to make inferences about all nights in D.C., then ideally we'd hope to have data from some rainy evenings and some nicer evenings.