R Assignment 5

> Question 1: a. What are the cases in the dataset? b. What is the sample proportion of US residents that have health insurance?

a. The cases in the *ACS* data set are 3.5 million randomly selected US residents.

```
# read ACS data set
acs = read.csv('ACS.csv')

countHI = table(acs$HealthInsurance)
prop.table(countHI)
```

```
##
##     0     1
## 0.139 0.861
```

b. US residents that have health insurance are coded with a 1 in the `HealthInsurance` variable from the data set. As we can see from the output, 86.1% of US residents in this sample have health insurance.

> Question 2: a. What type of estimate is the one you found in question 1: a point estimate or an interval estimate? b.Which do you think is a better estimate to report, a point estimate or an interval estimate? Explain your reasoning!

a. The estimate found in question 1 is a point estimate. It is one calculated value based on the selected sample.
b. The better estimate to report would be an interval estimate. Sample statistics as the proportion calculated in question 1 vary from the population parameter and with just reporting the sample statistic, one would not know, how much the population parameter is likely to vary or how big the margin of error is for that statistic.

> Question 3: Suppose we want to construct a confidence interval. Are the conditions met to assume the sampling distribution of sample proportions is approximately normal (i.e., the CLT is valid)? Explain.

The conditions to assume that a sampling distribution of sample proportions is approximately normal are that the sample observations are independent, the sample was taken randomly, and n is less than 10% of the population. Another conditions that needs to be met is that there are at least 10 expected successes and 10 expected failures. With the ACS, we can only assume that the sample observations are independent. The ACS data is a sample of 1% of US resident, so n = 1%, which is smaller than 10% and the sample was taken randomly. Additionally, there are more than 10 cases for both success and failure. Assuming the observations are independent, the CLT is valid and the conditions are met to assume that the sampling distribution of sample proportions is approximately normal.

Question 4: What is the value of the estimated standard error? Use the formula from the Week 5 slides and estimate the standard error using the normal distribution.

The formula to estimate the standard error is $\sqrt{\frac{p(1-p)}{n}}$, substituting $\hat{p}$ for p, we get $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ as the formula to estimate the standard error. Using the in question 1 calculated value 0.861 for $\hat{p}$ and defining *having health insurance* as success and n = 1000, we get:

$$\sqrt{\frac{0.861(1-0.861)}{1000}}$$

```
## [1] 0.01240564
```

The estimated standard error from the sample mean is 0.0124.

Question 5: a. Find a confidence interval for the true proportion of US residents who have health insurance based on a confidence level that you choose. b. Explain why you chose the confidence interval that you did. Use qnorm() to find the z needed. c. Interpret this confidence interval.

```
xbar = 0.861
seHI = 0.0124
zHI = qnorm(0.99, xbar, seHI)

lowerBd = xbar - zHI * seHI
upperBd = xbar + zHI * seHI
```

The 99% CI of US residents who have health insurance is [0.85, 0.872]. The confidence interval of99% is chosen because it gives a high confidence of 99% while the margin of error increases only slightly compared to a smaller confidence interval.

Question 6: What is the value of the estimated standard error? Use bootstrap simulations like in HW 4 to find the standard error.

```
n = length(acs$HealthInsurance)
##Generating our bootstrap distribution
boot.phats <- c() #Initializing the vector
for(i in 1:10000){ #i is a sample and we are taking 10000 samples
  boot.samp <- sample(acs$HealthInsurance, n, replace = TRUE) #Take a random sample
  #Now we need to calculate our bootstrap statistic (this is analogous to the sample  statistic
 we compute from a sample)
  boot.k <- length(which(boot.samp == 1)) #how many events or "successes" do we have in our samp
le
  boot.phat <- boot.k/n #a bootstrap statistic
  boot.phats <- c(boot.phats, boot.phat) #I am added the newly computed bootstrap statistic
#to the vector of bootstrap statistics
}

SE <- sd(boot.phats) # estimate of the SE for the sampling distribution of the proportion. We es
timate the SE by computing the standard deviation of our bootstrap distribution.
```

The value of the estimated standard error using bootstrap simulations is 0.0109.

> Question 7: Find a confidence interval for the true proportion of US residents who have health insurance based on a confidence level that you choose and the standard error you calculated in question 6.

```
zHI = qnorm(0.99, xbar, SE)

lowerBdBoot = xbar - zHI * SE
upperBdBoot = xbar + zHI * SE
```

The 99% confidence interval for the true proportion of US residents who have health insurance using the standard error calculated with the bootstrap simulations is [0.851, 0.871].

> Question 8: Suppose we'd like to test if the true proportion of US residents who have health insurance is 80% vs. the true proportion of US residents who have health insurance is NOT 80%. What would be the hypotheses for this test? Please write your hypotheses in nontechnical language AND using notation. Specify which hypothesis is which (null or alternative).

The null hypothesis in this case is that the true proportion of US residents who have health insurance equals 80%. The alternative hypothesis is that the true proportion of US residents who have health insurance does not equal 80%.

H_0: $p = 0.8$

H_a: $p \neq 0.8$

Question 9: Conduct a hypothesis test for the hypotheses specified in Question 8 using the confidence interval calculated in Question 5. State your conclusions in layman's terms and in the context of this question. Hint: look at the Week 5 part 2 slides.

The 99% confidence interval for p is [0.85, 0.872]. The null hypothesis is H_0: p = 0.8. The data provides evidence that the true proportion of Us residents that have health insurance is not equal to 0.8. The data suggest between 85% and 87.2% of US residents have health insurance. Therefor I reject the null hypothesis.