PHP 2510 Homework 3

Kuan-Min Lee

Question 1:

a. This question requires the information of the Z distribution of the generated dataset, Bin(20,.25). Therefore, I first calculate the mean and standard deviation of the given data distribution using the following code:

```
# create and interpret a 90% confidence interval using z distribution
# conduct the mean of the data
mean_data_bin <- 20*.25
# conduct the variance of the data
var_data_bin = 20*(.25)*(1-.25)
# conduct the standard deviation of the data
std_data_bin <- sqrt(var_data_bin)
```

After the above calculation, I then generate the critical value for z distribution for 90% confidence interval:

```
# conduct the critical value for 90% confidence interval
lower_bound <- qnorm(0.05)
upper_bound <- qnorm(0.95)
```

In the final step, I calculate the lower and upper bound of the 90% confidence interval:

```
# conduct the confidence interval value
lower_confidence_interval <- mean_data_bin + lower_bound*std_data_bin/sqrt(50)
upper_confidence_interval <- mean_data_bin + upper_bound*std_data_bin/sqrt(50)
```

For the final results, I received the interval range as follow:

```
> lower_confidence_interval
[1] 4.549538
> upper_confidence_interval
[1] 5.450462
```

b. This question requires the information of the t distribution of the generated dataset. Firstly, I need to calculate the sample mean and sample deviation of the generated data as below:

```r
# create and interpret a 90% confidence interval using t distribution
mean_data_bin_n <- mean(data_bin)
std_data_bin_n <- sd(data_bin)
n <- length(data_bin)
```

After the above calculation, I calculate the critical calculation for the t-critical value for 90% confidence interval:

```r
# conduct the critical value for 90% confidence interval
t_lower_bound <- qt(0.05, n-1)
t_upper_bound <- qt(0.95, n-1)
```

In the end, I calculate the upper and lower bound of the t-distribution:

```r
# conduct the confidence interval value
lower_t_confidence_interval <- mean_data_bin_n + t_lower_bound*std_data_bin_n/sqrt(50)
upper_t_confidence_interval <- mean_data_bin_n + t_upper_bound*std_data_bin_n/sqrt(50)
```

For the final results, I received the interval range as follow:

```
> lower_t_confidence_interval
[1] 4.573925
> upper_t_confidence_interval
[1] 5.346075
```

c. Compared to the confidence interval in z distribution, confidence interval in t distribution is smaller. This can be expected since that t distribution assume the standard variance is unknown and takes only the data of the sample into account. As compared to the entire dataset, the standard deviation of the sample is expected to be smaller as Ill.

Question 2:

The question can be seen as a hypothesis test for the below conditions:

$$H_0\ (null\ hypothesis): \mu = 2$$

$$H_A\ (alternative\ hypothesis): \mu \neq 2$$

To conduct the above testing, I first grab out the variable veggies from the original data set:

```
# load data into work space
load("organic.rda")
# data is stored numerically in the variable "veggies".
data_process_veggie <- organic$veggies
```

Then I conduct the hypothesis test with the t.test() function in R:

```
t.test(x=data_process_veggie,mu=2)
```

The below is the printed-out outcome:

```
        One Sample t-test

data:  data_process_veggie
t = 31.99, df = 2670, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 2.883264 2.998616
sample estimates:
mean of x
  2.94094
```

From the above t-test results, I can see that the p-value is much lower to the significance level of 0.05, which leads us to reject the null hypothesis I established in the beginning.

Therefore, I can make the conclusion that the mean number of cups of vegetables per day doesn't equal to 2.

Question 3:

This portion can be view as the two-sample t test between two group of people. The first group of people is from the good health condition, and the other is the people from the less than good health condition. The hypothesis test can be viewed as below:

$$H_0\ (null\ hypothesis)\colon \mu_1 = \mu_2$$

$$H_A\ (alternative\ hypothesis): \mu_1 \neq \mu_2$$

I first grab out the good-health condition people and bad-health condition people using filter() function:

```
# test the hypothesis that there is a significant difference in the mean veggies
# response between those who report being of at least good health vs those with
# less than good health
# divide the entire dataset based on health condition
good_health_data <- organic %>% filter(health=="Excellent" | health=="Very Good" | health=="Good")
bad_health_data <- organic %>% filter(health=="Fair" | health=="Poor")
```

After doing this, I grab out variable veggies from these two group respectively, and concatenate them together.

```
# grab out veggies from good health people and bad health people
good_health_data_veggies <- good_health_data$veggies
bad_health_data_veggies <- bad_health_data$veggies
health_data_veggies <- c(good_health_data_veggies,bad_health_data_veggies)
```

After conducting that, I create an indicator variable, which shows up as 1 for good health condition data and 0 for bad health condition data. This variable is created to form the dual grouping for the later t test:

```
# create a good health indicator
good_health_data_ind <- rep(1,length(good_health_data_veggies))
bad_health_data_ind <- rep(0,length(bad_health_data_veggies))
health_ind <- c(good_health_data_ind,bad_health_data_ind)
```

In the very last step, the health_data_veggies and health_ind data are concatenated together to form the final data frame for t test.

```
# create new tested data
test_data <- data.frame(health_data_veggies,health_ind)

t.test(health_data_veggies~health_ind,data=test_data,var.equal=F)
```

And the following are the printed-out results:

```
        Welch Two Sample t-test

data:  health_data_veggies by health_ind
t = -2.898, df = 786.58, p-value = 0.00386
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.34557716 -0.06646993
sample estimates:
mean in group 0 mean in group 1
       2.772866        2.978889
```

From the above test, we can see that the generated p-value is less than the significance level of 0.05. As a result, we reject the null hypothesis I established in the beginning. Therefore, the conclusion that the mean number of veggies for good and bad health condition people are not the same.

Appendix: Source Code for Homework

Question 1:

```
install.packages("dplyr")

library(dplyr)

# generate 50 random values from a Bin(20,.25) distribution

data_bin <- rbinom(50,20,.25)


# create and interpret a 90% confidence interval using z distribution

# conduct the mean of the data

mean_data_bin <- 20*.25

# conduct the variance of the data

var_data_bin = 20*(.25)*(1-.25)

# conduct the standard deviation of the data

std_data_bin <- sqrt(var_data_bin)
```

```
# conduct the criticial value for 90% confidence interval
lower_bound <- qnorm(0.05)
upper_bound <- qnorm(0.95)
# conduct the confidence interval value
lower_confidence_interval <- mean_data_bin + lower_bound*std_data_bin/sqrt(50)
upper_confidence_interval <- mean_data_bin + upper_bound*std_data_bin/sqrt(50)

# create and interpret a 90% confidence interval using t distribution
mean_data_bin_n <- mean(data_bin)
std_data_bin_n <- sd(data_bin)
n <- length(data_bin)
# conduct the critical value for 90% confidence interval
t_lower_bound <- qt(0.05, n-1)
t_upper_bound <- qt(0.95, n-1)
# conduct the confidence interval value
lower_t_confidence_interval <- mean_data_bin_n + t_lower_bound*std_data_bin_n/sqrt(50)
upper_t_confidence_interval <- mean_data_bin_n + t_upper_bound*std_data_bin_n/sqrt(50)
```

Question 2 and 3:

```
# load data into work space
load("organic.rda")
# data is stored numerically in the variable "veggies".
data_process_veggie <- organic$veggies
t.test(x=data_process_veggie,mu=2)

# test the hypothesis that there is a significant difference in the mean veggies
# response between those who report being of at least good health vs those with
```

```
# less than good health

# divide the entire dataset based on health condition

good_health_data <- organic %>% filter(health=="Excellent" | health=="Very Good" | health=="Good")

bad_health_data <- organic %>% filter(health=="Fair" | health=="Poor")

# grab out veggies from good health people and bad health people

good_health_data_veggies <- good_health_data$veggies

bad_health_data_veggies <- bad_health_data$veggies

health_data_veggies <- c(good_health_data_veggies,bad_health_data_veggies)

# create a good health indicator

good_health_data_ind <- rep(1,length(good_health_data_veggies))

bad_health_data_ind <- rep(0,length(bad_health_data_veggies))

health_ind <- c(good_health_data_ind,bad_health_data_ind)

# create new tested data

test_data <- data.frame(health_data_veggies,health_ind)


t.test(health_data_veggies~health_ind,data=test_data,var.equal=F)
```