Shamim Sherafati

R Practice 3 – ALY 6010

Module 3
Date: 2022/11/20

In this report, we will be conducting hypothesis testing on R data sets which is Student Food Survey. “Nutrition” variable as well as other variables will be used to conduct one-sample t-test. I will use boxplot and Shapiro test in further to be sure that if their distribution is normal or not.

First, I imported my dataset and its packages and start cleaning it. It starts from changing the columns name, dropping one column name “Timestamp” and deleting the ‘NA’ values in dataset. Then, after getting summary and string of data after cleaning which can be seen below, I start creating t-test.

```
> summary(Data_FoodSurvey) #get final summary after cleaning data
    Gender           Boarding           Grade            Athlete          Activities
Length:137         Length:137         Length:137         Length:137         Length:137
Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



breakfast.in.DH breakfast.NOTin.DH breakfast.in.Class BoxOfFood.in.DH BoxOfFood.NOTin.DH  Meal.in.Dorm
Min.   :0.000    Min.   :0.000      Length:137         Min.   :0.000    Min.   :0.0000     Min.   : 0.00
1st Qu.:0.000    1st Qu.:0.000      Class :character   1st Qu.:2.000    1st Qu.:0.0000     1st Qu.: 0.00
Median :2.000    Median :1.000      Mode  :character   Median :2.000    Median :0.0000     Median : 3.00
Mean   :3.095    Mean   :2.277                         Mean   :2.299    Mean   :0.5255     Mean   : 3.19
3rd Qu.:6.000    3rd Qu.:5.000                         3rd Qu.:3.000    3rd Qu.:1.0000     3rd Qu.: 5.00
Max.   :7.000    Max.   :7.000                         Max.   :6.000    Max.   :5.0000     Max.   :10.00
  Nutrition        Money
Min.   :0.000    Min.   :  0.0
1st Qu.:1.000    1st Qu.:  5.0
Median :3.000    Median : 15.0
Mean   :2.679    Mean   : 17.2
3rd Qu.:4.000    3rd Qu.: 25.0
Max.   :5.000    Max.   :100.0
>
```

```
> str(Data_FoodSurvey)
'data.frame':   137 obs. of  13 variables:
 $ Gender            : chr  "Female" "Male" "Female" "Male" ...
 $ Boarding          : chr  "Day" "Boarding" "Day" "Day" ...
 $ Grade             : chr  "12th" "10th" "9th" "11th" ...
 $ Athlete           : chr  "No" "Yes" "Yes" "No" ...
 $ Activities        : chr  "None of the above" "Soccer" "Tennis" "None of the above" ...
 $ breakfast.in.DH   : int  0 5 3 0 0 2 0 0 5 1 ...
 $ breakfast.NOTin.DH: int  2 2 7 4 5 0 6 7 2 2 ...
 $ breakfast.in.Class: chr  "No" "No" "Yes" "No" ...
 $ BoxOfFood.in.DH   : int  2 2 2 2 1 2 1 3 2 2 ...
 $ BoxOfFood.NOTin.DH: int  0 0 0 0 0 0 0 0 1 0 ...
 $ Meal.in.Dorm      : int  0 5 10 1 2 4 0 0 3 1 ...
 $ Nutrition         : int  0 4 4 2 3 2 4 3 4 2 ...
 $ Money             : num  40 15 50 4 20 5 25 12 10 12 ...
 - attr(*, "na.action")= 'omit' Named int [1:30] 6 7 8 10 12 14 19 21 33 42 ...
  ..- attr(*, "names")= chr [1:30] "6" "7" "8" "10" ...
>
```

A. One Sample t-test of Nutrition with two sided tailed

The One-Sample t-Test checks whether the mean of a population is statistically different from a known or hypothesized value. The One-Sample t-test is a parametric test.

Now, I start creating t-test for Nutrition: First I create a summary for Nutrition variable to get initial detail of this variable.

```
> summary(Data_FoodSurvey$Nutrition)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   3.000   2.679   4.000   5.000
>
```

Now, I want to create one Sample t-test of Nutrition with two sided tailed.

For this test, first I get the mean and standard deviation of nutrition variable which is like following.

Mean= `2.678832` SD= `1.603921`

Then, I create one sample t-test with two. Sided alternative and population mean = 0 for it.

```
> t.test(Data_FoodSurvey$Nutrition, mu = 0, alternative = "two.sided")

        One Sample t-test

data:  Data_FoodSurvey$Nutrition
t = 19.549, df = 136, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.407843 2.949822
sample estimates:
mean of x
 2.678832
```

Data_FoodSurvey$Nutrition: a numeric vector containing your data values

mu: the theoretical mean. Default is 0 and I keep it.

alternative: the alternative hypothesis which I put "two. Sided" as it default.

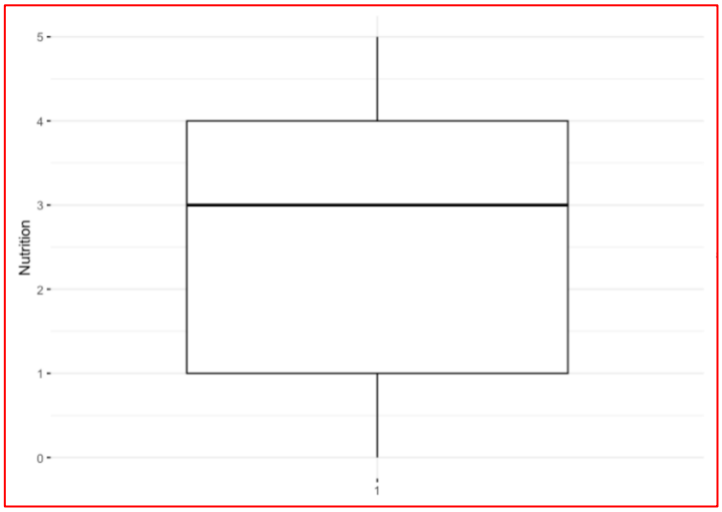
From the test above, we observed that difference is 136 and p value is less than 2.2e-16 which is not close to mean and close to zero. Also, value of t test is t = 19.549.

- Null Hypothesis **H0**: Average of Nutrition is 2.678832.
- Alternative Hypothesis **H1**: Average of Nutrition is not 2.678832.

Here, the alternative hypothesis is show that true mean is not equal to 0 which means there is a statistical difference between the two means.
Also, as the p-value is smaller than significant level (As confidence level is 95% so significant level is 0.05) then the alternative hypothesis is accepted.

Visualize Nutrition using box plots for test A

Now, I create a box plot for Nutrition value to examine the t-test in other way which make it more understandable for us; The Nutrition value, which is between 0 to 5 counted, is in the third quartile and is larger than 75%.



Preliminary test to check one-sample t-test assumptions for test-A

Now one question appears that;
Is this a large sample? - No, because n < 10.
Since the sample size is not large enough (less than 10, central limit theorem), we need to check whether the data follow a normal distribution.

Briefly, it's possible to use the Shapiro-Wilk normality test and to look at the normality plot.
Shapiro-Wilk test:

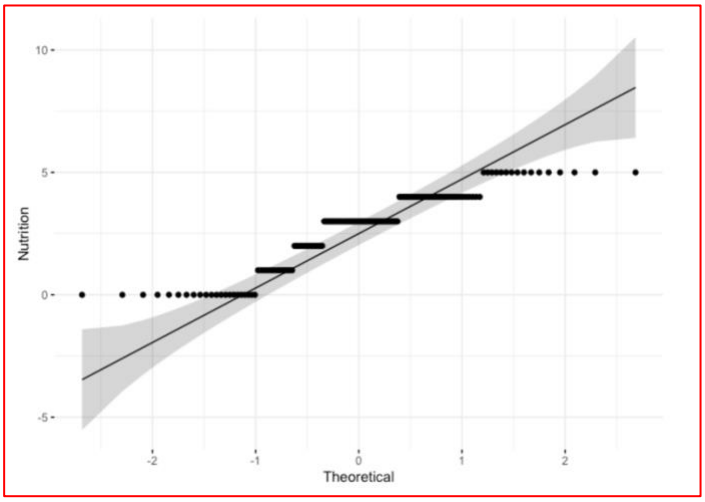
- Null hypothesis: the data are normally distributed
- Alternative hypothesis: the data are not normally distributed

```
> shapiro.test(Data_FoodSurvey$Nutrition) # => p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  Data_FoodSurvey$Nutrition
W = 0.89561, p-value = 2.392e-08

>
```

From the output, we can see that our p-value of 2.2e–16 is much smaller than 0.05, so we can reject the null hypothesis of no difference and say with a high degree of confidence that the true difference in means is not equal to zero.

Visual inspection for test A

Visual inspection of the data normality using Q-Q plots (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution.



From the normality plots, we conclude that the data may come from normal distributions.

Compute one-sample t-test for test-A

We want to know if the average Nutrition of the students differs from 5 (two-tailed test)?

The result shows:

t is the t-test statistic value (t = -16.939),

df is the degrees of freedom (df= 136),

p-value is the significance level of the t-test (p-value < 2 .2e-16

conf.int is the confidence interval of the mean at 95% (c onf.int = [2.407843 2.949822]).

sample estimate is the mean value of the sample (mean = 2.678832).

```
> # One-sample t-test
> test1 <- t.test(Data_FoodSurvey$Nutrition, mu = 5)
> # Printing the results
> test1

        One Sample t-test

data:  Data_FoodSurvey$Nutrition
t = -16.939, df = 136, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 2.407843 2.949822
sample estimates:
mean of x
 2.678832

>
```

Interpretation of the result

The p-value of the test is 2.2e-16 which is less than the significance level alpha = 0.05. We can conclude that the average Nutrition of the students differs from 5 is with a p-value = 2.2e-16

B. One Sample t-test of Having Meal.in.Dorm with right tailed

I start creating t-test for Having Meal in Dorm: First I get the mean and SD for this variable

```
> mean(Data_FoodSurvey$Meal.in.Dorm) #get the mean of Meal.in.Dorm
[1] 3.189781
> sd (Data_FoodSurvey$Meal.in.Dorm) #get the Standard Deviation of Meal.in.Dorm
[1] 2.959364
>
```

Now, I want to create one Sample t-test for Having Meal in Dorm with right-tailed population mean = 7 for it;

- We want to know, if the average of Having Meal in Dorm of the students greater than 10 (greater-tailed test), So:

```
> test2 <- t.test(Data_FoodSurvey$Meal.in.Dorm, mu = 10, alternative = 'greater') #One Sample t-test of Meal.in.Dorm
> test2

        One Sample t-test

data:  Data_FoodSurvey$Meal.in.Dorm
t = -26.935, df = 136, p-value = 1
alternative hypothesis: true mean is greater than 10
95 percent confidence interval:
 2.771051      Inf
sample estimates:
mean of x
 3.189781

>
```

Data_FoodSurvey$Meal.in.Dorm: a numeric vector containing your data values

mu: the theoretical mean which is 10 here.

alternative: the alternative hypothesis which I put “right. Sided”.

From the test above, we observed that difference is 136 and p value is 1 which is not close to mean. Also, **value of t test is t = -26.935**.

- Null Hypothesis **H0**: Average of Nutrition is 3.189781.
- Alternative Hypothesis **H1**: Average of Nutrition is not 3.189781.

Here, the alternative hypothesis show that true mean is greater than 10 which means the null hypothesis is rejected.
Also, as the p-value is greater than significant level (As confidence level is 95% so significant level is 0.05) then the null hypothesis is accepted.

Visualize of Having Meal.in.Dorm using box plots for test B

Now, I create a box plot for Having Meal.in.Dorm value to examine the t-test in other way which make it more understandable for us; This variable's value which is between 0 to 10 counted.

Preliminary test to check one-sample t-test assumptions for test-B

Now one question appears that,

- Is this a large sample? – Yes, its large enough which is n = 10.

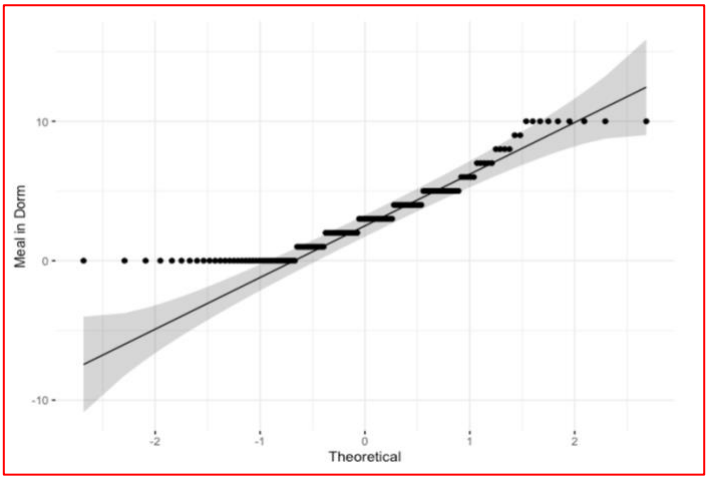
Since the sample size is large enough (equal to 10, central limit theorem), we don't need to check whether the data follow a normal distribution, but I also provide this test for it, too.

```
> shapiro.test(Data_FoodSurvey$Meal.in.Dorm) # => p-value = 1

        Shapiro-Wilk normality test

data:  Data_FoodSurvey$Meal.in.Dorm
W = 0.88627, p-value = 7.902e-09

>
```

From the output, we can see that our p-value of 1 is greater than 0.05, so we can accept the null hypothesis say with a high degree of confidence that the true difference in means is greater than population mean.

Visual inspection for test B

From the normality plots, we conclude that the data may come from normal distributions.



Interpretation of the result

The p-value of the test is 1 which is more than the significance level alpha = 0.05. We can conclude that the average Nutrition of the students greater from 10 is with a p-value =1

C. One Sample t-test of breakfast.in.Dining Hall with left tailed

I start creating t-test for breakfast.in.Dining Hall: First I get the mean and SD for this variable

```
> mean(Data_FoodSurvey$breakfast.in.DH) #get the mean of breakfast.in.Dinng Hall
[1] 3.094891
> sd(Data_FoodSurvey$breakfast.in.DH) #get the standard deviation of breakfast.in.Dinng Hall
[1] 2.645427
>
```

Now, I want to create one Sample t-test for breakfast.in.Dining Hall with left-tailed population mean = 7 for it.

- We want to know, if the average for breakfast.in.Dining Hall of the students less than 7 (greater-tailed test), So:

```
> test3 <- t.test(Data_FoodSurvey$breakfast.in.DH, mu = 7, alternative = 'less') #One Sample t-test
 of breakfast.in.DH with
> test3

        One Sample t-test

data:  Data_FoodSurvey$breakfast.in.DH
t = -17.278, df = 136, p-value < 2.2e-16
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
   -Inf 3.4692
sample estimates:
mean of x
 3.094891

>
```

Data_FoodSurvey$breakfast.in.DH: a numeric vector containing your data values

mu: the theoretical mean which is 7 here.

alternative: the alternative hypothesis which I put “left. Sided”.

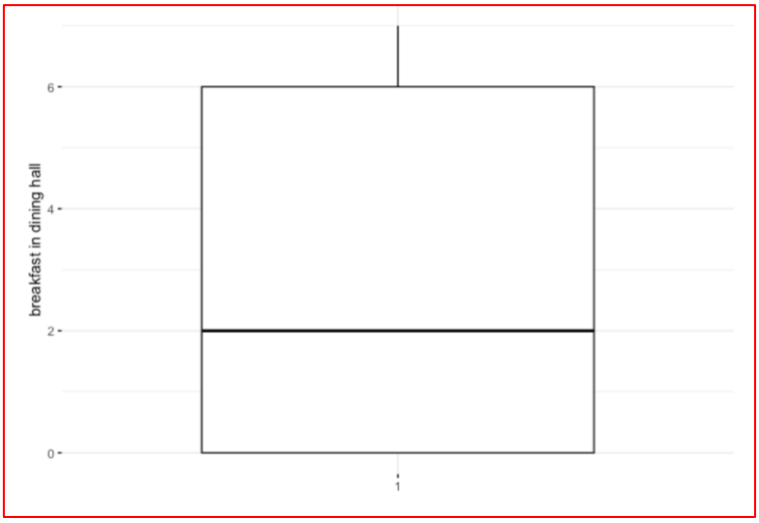
From the test above, we observed that difference is 136 and p value is 1 which is not close to mean. Also, value of t test is t = -17.278.

- Null Hypothesis H0: Average of Nutrition is 3.094891.
- Alternative Hypothesis H1: Average of Nutrition is not 3.094891.

Here, the alternative hypothesis show that true mean is less than 7 which means the null hypothesis is rejected.
Also, as the p-value is less than significant level (As confidence level is 95% so significant level is 0.05) then the alternative hypothesis is accepted.

Visualize of having breakfast in dining hall using box plots for test C

Now, I create a box plot for having breakfast in dining hall value to examine the t-test in other way which make it more understandable for us; This variable's value which is between 0 to 6 counted.



Preliminary test to check one-sample t-test assumptions for test-C

Now one question appears that;

- Is this a large sample? - No, because n < 10.

Since the sample size is not large enough (less than 10, central limit theorem), we need to check whether the data follow a normal distribution.

```
> shapiro.test(Data_FoodSurvey$breakfast.in.DH) # => p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  Data_FoodSurvey$breakfast.in.DH
W = 0.86192, p-value = 5.637e-10

>
```

From the output, we can see that our p-value of 2.2e–16 is much smaller than 0.05, so we can reject the null hypothesis of no difference and say with a high degree of confidence that the true difference in means is not equal to zero.

Compute one-sample t-test for test-C

We want to know, if the average Nutrition of the students less than 7 (left-tailed test)?

```
> # One-sample t-test
> test2 <- t.test(Data_FoodSurvey$breakfast.in.DH, mu = 7)
> # Printing the results
> test2

        One Sample t-test

data:  Data_FoodSurvey$breakfast.in.DH
t = -17.278, df = 136, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 2.647934 3.541847
sample estimates:
mean of x
 3.094891

>
```

In the result above:

t is the t-test statistic value (t = -17.278),

df is the degrees of freedom (df= 136),

p-value is the significance level of the t-test (p-value < 2.2e-16

conf.int is the confidence interval of the mean at 95% (conf.int = [2.647934 3.541847]

sample estimate is the mean value of the sample (mean = 3.094891).

Interpretation of the result

The p-value of the test is less than 2.2e-16 which is less than the significance level alpha = 0.05. We can conclude that the average Nutrition of the students less than 7 is with a p-value < 2.2e-16.

Hypothesis testing for p-value

Now, I created Wilcoxon rank for my two variables to compare them that these data are distributed symmetrically around the median or not. The paired test, we set the "paired" argument as TRUE. As the p-value turns out to be 1.441e-09, and is more than the .05 significance level, we accept the null hypothesis.

```
> A <- Data_FoodSurvey$BoxOfFood.NOTin.DH  #create data A for one variable n dataset
>
> B <- Data_FoodSurvey$breakfast.NOTin.DH # #create data B for another variable n dataset
>
> wilcox.test(A,B, paired = TRUE, correct = TRUE)  #Wilcoxon rank sum exact test for A and B

        Wilcoxon signed rank test with continuity correction

data:  A and B
V = 776.5, p-value = 1.441e-09
alternative hypothesis: true location shift is not equal to 0

>
>
```

I also get the p value of Nutrition and other variables which I made t-test for them, based on that which I put both TRUE and FALSE for lower tail to get the probability cumulative density of these variables which can be seen below.

```
> t = (mean(Data_FoodSurvey$Nutrition)-2.5)/(sd(Data_FoodSurvey$Nutrition)/sqrt(length
(Data_FoodSurvey$Nutrition)))
> 2*pt(-abs(t),df=length(Data_FoodSurvey$Nutrition)-1)
[1] 0.1940841
> pt(t, df=136, lower.tail=T)
[1] 0.902958
> pt(t, df=136, lower.tail=F)
[1] 0.09704203
>
```

```
> t = (mean(Data_FoodSurvey$breakfast.in.DH)-2.5)/(sd(Data_FoodSurvey$breakfast.in.DH)/sqrt(length
(Data_FoodSurvey$breakfast.in.DH)))
> 2*pt(-abs(t),df=length(Data_FoodSurvey$breakfast.in.DH)-1)
[1] 0.009466932
> pt(t, df=136, lower.tail=T)
[1] 0.9952665
> pt(t, df=136, lower.tail=F)
[1] 0.004733466
>
```

```
> t = (mean(Data_FoodSurvey$Meal.in.Dorm)-2.5)/(sd(Data_FoodSurvey$Meal.in.Dorm)/sqrt
(length(Data_FoodSurvey$Meal.in.Dorm)))
> 2*pt(-abs(t),df=length(Data_FoodSurvey$Meal.in.Dorm)-1)
[1] 0.007209708
> pt(t, df=136, lower.tail=T)
[1] 0.9963951
> pt(t, df=136, lower.tail=F)
[1] 0.003604854
>
```

References

I. D. Iacobucci, On p-values, Journal of Consumer Research 32 (June 2005), no. 1, 6–12. http://www.journals.uchicago.edu/cgi-bin/resolve?JCR320199PDF

II. R. Mason, D. Lind, and W. Marchal, Statistics, An Introduction; 5th edition; Duxbury Press; Brooks/Cole Publishing Company; 1998.

III. R. Savage and K. W. Deutsch, A Statistical Model of the Gross Analysis of Transaction Flows Econometrica 28 (1960), no. 3 551–572.

IV. Y. M. M. Bishop and S. E. Fienberg, Incomplete Two-Dimensional Contingency Tables, Biometrics 25 (1969), no. 1, 119–128.