## STAT-614 HW4 Yunting Chiu

STAT 614 - HW 4 Due: Thursday, October 22, 2020 in Blackboard by 11:59pm.
Instructions: Please type your solutions and upload the document as a pdf file in Blackboard. There is only one file to submit for this assignment. As part of this assignment, please take the completely anonymous Midterm Course Evaluation under the Survey tab in Blackboard.

Notes:

- For this HW you will need some concepts from chapter 3 on checking assumptions and transformations and chapter 4 on nonparametric methods.
- You will also be revisiting the "big ideas" around confidence intervals and hypothesis tests.

The food-frequency questionnaire (FFQ) is an instrument often used in dietary epidemiology to assess consumption of specific foods. A person is asked to write down the number of servings per day typically eaten in the past year of over 100 individual food items. A food-composition table is then used to compute nutrient intakes (protein, fat, etc.) based on aggregating responses for individual foods. The FFQ is inexpensive to administer but is considered less accurate than the diet record (DR) (the gold standard of dietary epidemiology). For the DR, a participant writes down the amount of each specific food eaten over the past week in a food diary and a nutritionist, using a special computer program, computes nutrient intakes from the food diaries. This is a much more expensive method of dietary recording. To validate the FFQ, 173 nurses participating in the Nurses' Health Study completed 4 weeks of diet recording about equally spaced over a 12-month period and an FFQ at the end of diet recording. Data are in Blackboard in the file valid.txt.

Consider the data on total alcohol consumption for both the DR and FFQ, **alco_dr** and **alco_ffq**, respectively. You are to assess whether the two methods, diet record and the food-frequency questionnaire, are comparable for total alcohol consumption. In particular, is there evidence that FFQ underestimates total alcohol consumption, in general? **Estimate by how much the FFQ generally underestimates total alcohol consumption.**

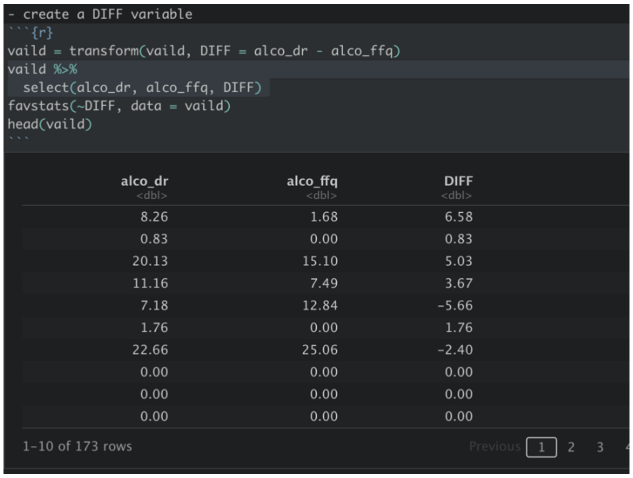
1. Explain why the initial model needed to address these research goals is a matched-pairs t-procedure.

   The matched-pairs t- procedure is test for difference in paired mean (In this case, there are 173 nurses have tested, they are from the same group). So, we need to define a new variable, which is based on the difference between paired values from alco_dr and alco_ffq.

Note: Two-sample t-test is used when the data of two samples are statistically independent, while the paired t-test is used when data is in the form of matched pairs. To use the two-sample t-test, we need to assume that the data from both samples are normally distributed and they have the same variances. The opposite of a matched sample is an independent sample, which deals with unrelated groups.

2. Use both the model notation we developed in class and a brief written description of the model (you may also use pictures) to illustrate the model. (Be careful! The matched-pairs procedure works on the difference in the two measures on each individual. Start with y = alco_dr – alco_ffq and describe the model for y!).

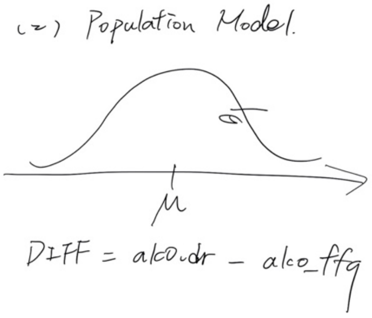
   We set DIFF = y = alco_dr – alco_ffq



~~~
- create a DIFF variable
```{r}
vaild = transform(vaild, DIFF = alco_dr - alco_ffq)
vaild %>%
  select(alco_dr, alco_ffq, DIFF)
favstats(~DIFF, data = vaild)
head(vaild)
```
~~~

```
       alco_dr        alco_ffq        DIFF
         <dbl>           <dbl>       <dbl>
          8.26            1.68        6.58
          0.83            0.00        0.83
         20.13           15.10        5.03
         11.16            7.49        3.67
          7.18           12.84       -5.66
          1.76            0.00        1.76
         22.66           25.06       -2.40
          0.00            0.00        0.00
          0.00            0.00        0.00
          0.00            0.00        0.00
1-10 of 173 rows                     Previous  1  2  3  4
```

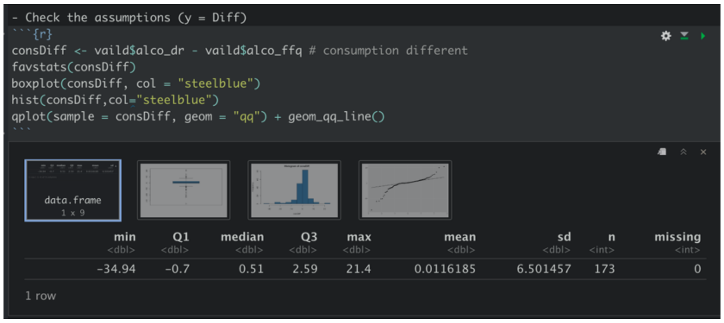
The Model notation is yi ~ N (mu, sigma^2) or yi = mu + ei
This is a matched pair sample. This means that yi is distributed normal with a random error term. Each observation has its own error.



Because the alco_dr and alco_ffq is paired, the initial analysis would be a matched pairs t-procedure. Each observation is taken one ID.

3. What are the model assumptions?
   Assumptions of the matched pairs t-procedure:
   - Sample of independent observations (study design)
   - From a normally distributed population (plot)
   - No influential outliers (plot)

   We focus on a DIFF variable first.

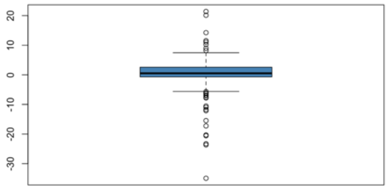


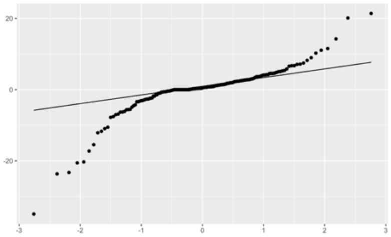
~~~
- Check the assumptions (y = Diff)
```{r}
consDiff <- vaild$alco_dr - vaild$alco_ffq # consumption different
favstats(consDiff)
boxplot(consDiff, col = "steelblue")
hist(consDiff,col="steelblue")
qplot(sample = consDiff, geom = "qq") + geom_qq_line()
```
~~~

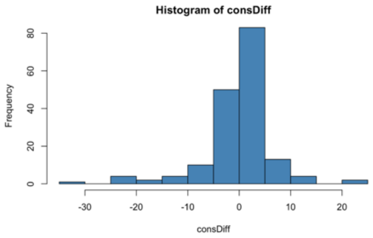
```
data.frame
  1 x 9

     min      Q1   median     Q3    max        mean        sd     n   missing
   <dbl>   <dbl>    <dbl>  <dbl>  <dbl>       <dbl>     <dbl> <int>     <int>
  -34.94    -0.7     0.51   2.59   21.4   0.0116185  6.501457   173         0

1 row
```

• Sample of independent observations (The observations are independent of one another) Because of the paired design, we focus on the difference in total alcohol consumption with two different record method. The first assumption of the matched pairs procedure is that we have a sample of independent observations (i.e. consumptions are independent of one another).

4. Which of the model assumptions are not met? Give and refer to specific output.







• From a normally distributed population (should be approximately normally distributed) There are many points that deviate from others in the qq plot. The sample size (n = 173) in this study is enough. Therefore, we cannot consider the difference of total alcohol consumption is follower a normal distribution.

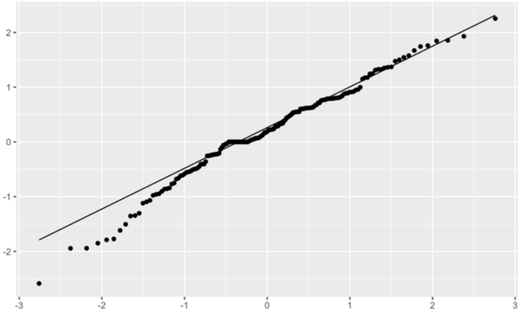
- No influential outliers

When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot. Based on the above boxplot we can see many outliers, that is, we need to check if outliers are impacting our conclusions.

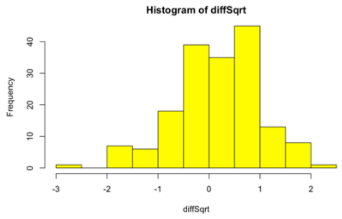
5. Consider a square root transformation of the alcohol data: salcoDR = √alco_dr and salcoFFQ = √alco_ffq. Are the model assumptions met for the transformed data? Give and refer to specific output.

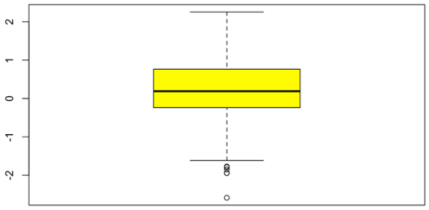
- Assumption1: Sample of independent observations (study design)
  Still met, because the study design has not changed.

- Assumption 2: From a normally distributed population (plot)
  we cloud see the points forming a line that's roughly straight. As the normal distribution is symmetric, so the below qqplot has no skew (the mean is approximately equal to the median).



- Assumption 3: No influential outliers (plot)
  After we transformed the data, the below boxplot and histogram show there are only a few outliers in the histogram or boxplot, which follows the assumption 3.





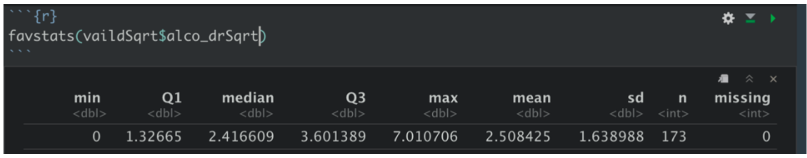
6. Conduct the appropriate test on the square root transformed data and interpret the results. Be sure to address the research questions stated above.

According to the above qqplot and boxplot, the transformed variables alco_drSqrt and alco_ffqSqrt meet the three assumptions. Therefore, the appropriate test is one-sided t-test for this study.

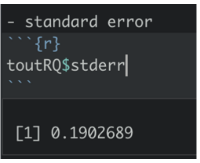
RQ: Is there evidence that FFQ underestimates total alcohol consumption?

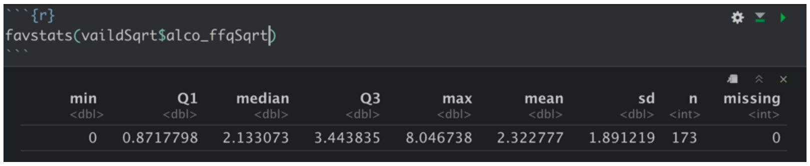
H0: The distribution of total alcohol consumption is **the same** between the DR and FFQ method. (DR = FQ).

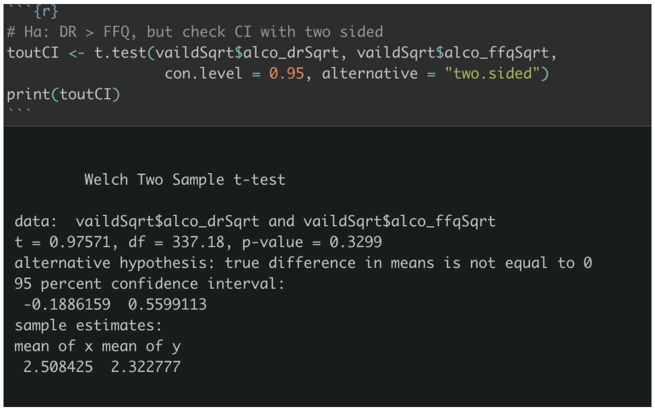
Ha: The distribution of total alcohol consumption in the diet record (DR: the gold standard of dietary epidemiology) is **greater than** food-frequency questionnaire (DR >FFQ).



~~~
```{r}
favstats(vaildSqrt$alco_drSqrt)
~~~

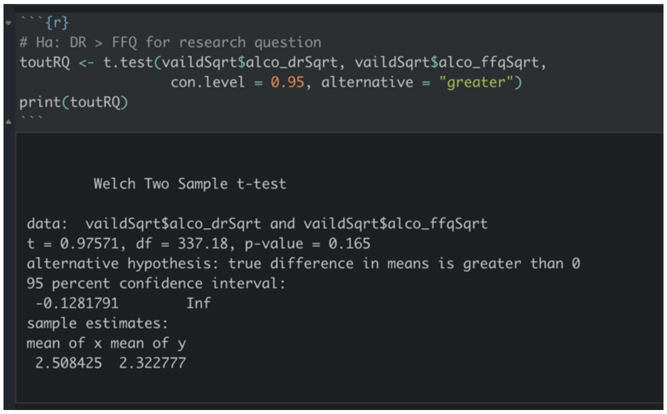
    min       Q1   median       Q3      max     mean       sd     n  missing
  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <int>    <int>
      0  1.32665 2.416609 3.601389 7.010706 2.508425 1.638988   173        0
```



```
- standard error
```{r}
toutRQ$stderr
```

[1] 0.1902689
```



```
```{r}
favstats(vaildSqrt$alco_ffqSqrt)
```

    min        Q1   median       Q3      max     mean       sd     n  missing
  <dbl>     <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <int>    <int>
      0 0.8717798 2.133073 3.443835 8.046738 2.322777 1.891219   173        0
```

In the t-procedure, the DR method is greater than FFQ method -0.189 to 0.556 units of total alcohol consumption with 95 % CI. The DR sample estimated mean is 2.51, the FFQ sample estimated mean is 2.32, and the difference in sample mean is 0.19 with 95% CI, respectively.



```
```{r}
# Ha: DR > FFQ, but check CI with two sided
toutCI <- t.test(vaildSqrt$alco_drSqrt, vaildSqrt$alco_ffqSqrt,
                 con.level = 0.95, alternative = "two.sided")
print(toutCI)
```

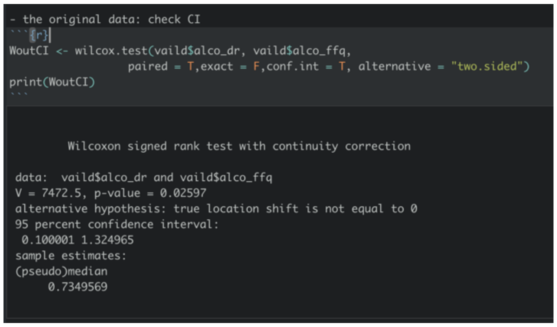
	Welch Two Sample t-test

data:  vaildSqrt$alco_drSqrt and vaildSqrt$alco_ffqSqrt
t = 0.97571, df = 337.18, p-value = 0.3299
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1886159  0.5599113
sample estimates:
mean of x mean of y 
 2.508425  2.322777 
```

Then we use one-sided test to conduct the hypothesis. Based on the result, the p-value is 0.165, so we fail to reject the null hypothesis. The t-test tell us we cannot say that FFQ underestimates total alcohol consumption.
```



~~~
```{r}
# Ha: DR > FFQ for research question
toutRQ <- t.test(vaildSqrt$alco_drSqrt, vaildSqrt$alco_ffqSqrt,
                  con.level = 0.95, alternative = "greater")
print(toutRQ)
~~~

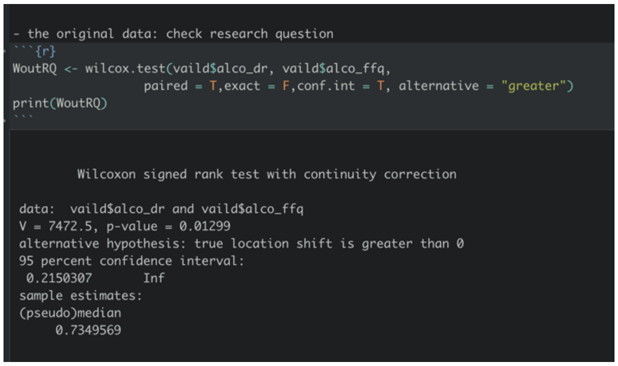
	Welch Two Sample t-test

data:  vaildSqrt$alco_drSqrt and vaildSqrt$alco_ffqSqrt
t = 0.97571, df = 337.18, p-value = 0.165
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1281791        Inf
sample estimates:
mean of x mean of y
 2.508425  2.322777
```

7. Consider a nonparametric method for addressing the research questions. What null and alternative hypotheses are addressed by the appropriate nonparametric method? Carry out and interpret the results of the nonparametric method. Include and interpret the confidence interval estimate.

   The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. Because the original data didn’t pass the assumption of one sample mean. We will be using a nonparametric method. Therefore, we need to formulate our hypothesis:

   RQ: Is there evidence that FFQ underestimates total alcohol consumption?
   H0: The distribution of total alcohol consumption is the same between the DR and FFQ method (DR = FQ).
   Ha: The distribution of total alcohol consumption in the diet record (DR: the gold standard of dietary epidemiology) is greater than food-frequency questionnaire (DR >FFQ).

   We find the regular two-sided confidence interval first. We estimate that the distribution of FFQ is shifted 0.735 below that of DR, a shift of 0.10 to 1.32 below with 95% confidence interval, on average.



```
- the original data: check CI
```{r}
WoutCI <- wilcox.test(vaild$alco_dr, vaild$alco_ffq,
                  paired = T,exact = F,conf.int = T, alternative = "two.sided")
print(WoutCI)
```

	Wilcoxon signed rank test with continuity correction

data:  vaild$alco_dr and vaild$alco_ffq
V = 7472.5, p-value = 0.02597
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.100001 1.324965
sample estimates:
(pseudo)median
     0.7349569
```
```

Then we use one-sided test to conduct the hypothesis. With the small p-value 0.01299, we conclude there is sufficient evidence against the null hypothesis in favor of the alternative hypothesis that the population distribution of total alcohol consumption in the DR method is greater than FFQ method.



~~~
- the original data: check research question
```{r}
WoutRQ <- wilcox.test(vaild$alco_dr, vaild$alco_ffq,
                      paired = T,exact = F,conf.int = T, alternative = "greater")
print(WoutRQ)
```


        Wilcoxon signed rank test with continuity correction

 data:  vaild$alco_dr and vaild$alco_ffq
 V = 7472.5, p-value = 0.01299
 alternative hypothesis: true location shift is greater than 0
 95 percent confidence interval:
  0.2150307       Inf
 sample estimates:
 (pseudo)median
      0.7349569
~~~

8. Which of the results in (6) or (7) do you prefer to use to draw conclusions for this study and why?

   I think using a parametric test to conclude this study is more suitable, because parametric tests (t-procedure) are more powerful than nonparametric tests. Secondly, our sample size is large enough (n = 173). By contrast, nonparametric tests are less powerful because they use less information in their calculation. In summary, with the 0.165 p-value of one tailed paired t-test, we cannot conclude that FFQ underestimates total alcohol consumption.