STAT 614 - HW 4

By Sihyuan Han

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.

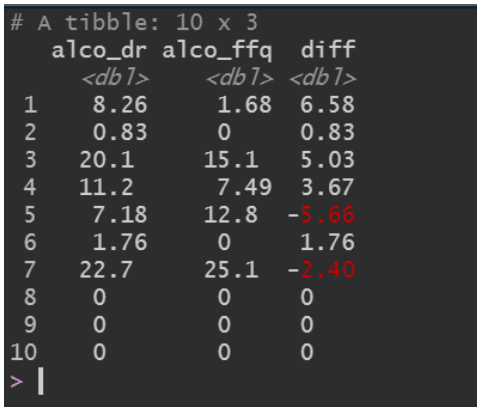
Consider the data on total **alcohol consumption** for both the **DR and FFQ**, **alco_dr and alco_ffq**, respectively. You are to assess whether the two methods, diet record and the food-frequency questionnaire, are comparable for total alcohol consumption. In particular, is there evidence that FFQ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

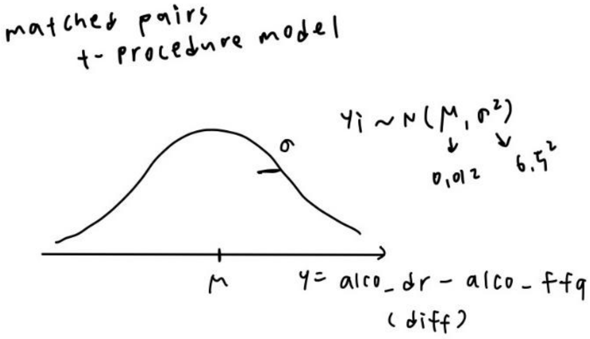
   Ans: Because this data is a paired data, this research is to test these 173 nurses who did both DR and FFQ, there is no other independent sample in this research.

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!)

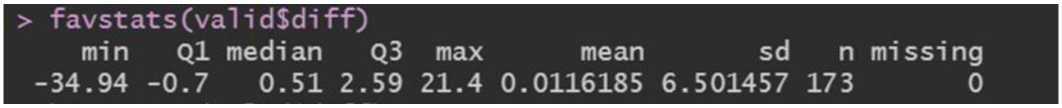
Ans:



```
# A tibble: 10 x 3
   alco_dr alco_ffq  diff
     <dbl>    <dbl> <dbl>
 1    8.26     1.68  6.58
 2    0.83     0     0.83
 3   20.1     15.1   5.03
 4   11.2      7.49  3.67
 5    7.18    12.8  -5.66
 6    1.76     0     1.76
 7   22.7     25.1  -2.40
 8    0        0     0
 9    0        0     0
10    0        0     0
> |
```



diff = alco_dr - alco_ffq



```
> favstats(valid$diff)
    min    Q1 median   Q3  max      mean       sd   n missing
 -34.94 -0.7   0.51 2.59 21.4 0.0116185 6.501457 173       0
```

3. What are the model assumptions?

Ans:

(1) Sample of independent observations

(2) From a normally distributed population

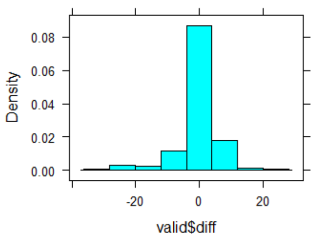
(3) Check (no influential) outliers!

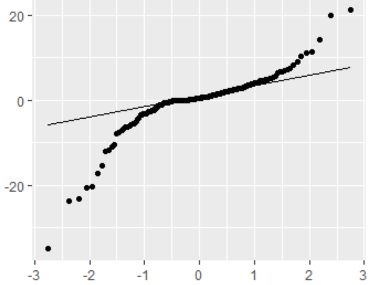
4. Which of the model assumptions are not met? Give and refer to specific output.

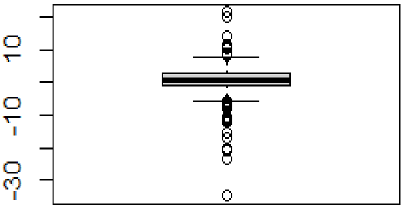
Ans:

This is not a normally distributed population. if normally distributed the qqplot should be mostly linear, but we can see the following plot has lots of points being apart from the line.

Check the outliers, as it shows in the boxplot below, it has lots of outliers!







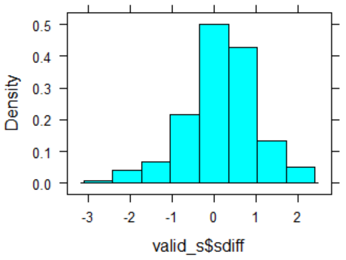
5. Consider a square root transformation of the alcohol data: salcoDR = $\sqrt{\text{alco\_dr}}$ and salcoFFQ = $\sqrt{\text{alco\_ffq}}$. Are the model assumptions met for the transformed data? Give and refer to specific output.

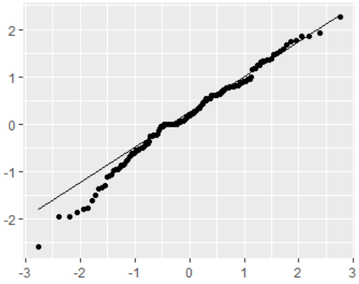
   Ans:

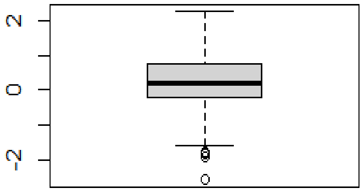
   This is a sample of independent observations.

   As we can see in the qqplot, it is mostly linear, so we can conclude that the population is normally distributed.

   There are some outliers as we can check from the boxplot, but no influential outliers!







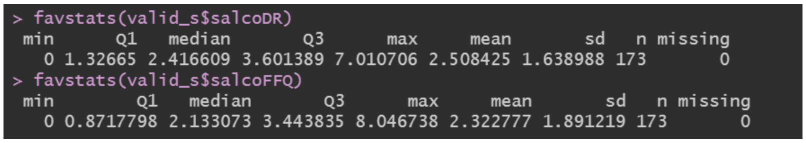
6. Conduct the appropriate test on the **square root transformed** data and interpret the results. Be sure to address the research questions stated above.

Ans:

RQ: Is there evidence that FFQ *underestimates* total alcohol consumption, in general?

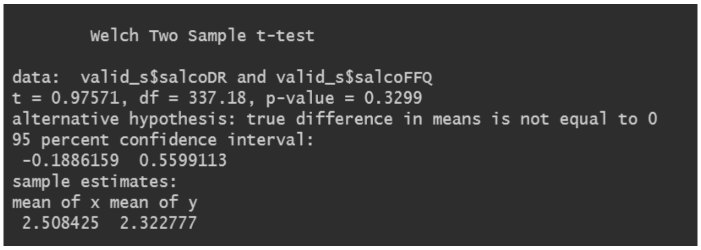
H0(null hypothesis): there is no difference in total alcohol consumption between DR and FFQ (alco_dr = alco_ffq)

Ha(alternative hypothesis): there is difference in total alcohol consumption between DR and FFQ (alco_dr > alco_ffq)



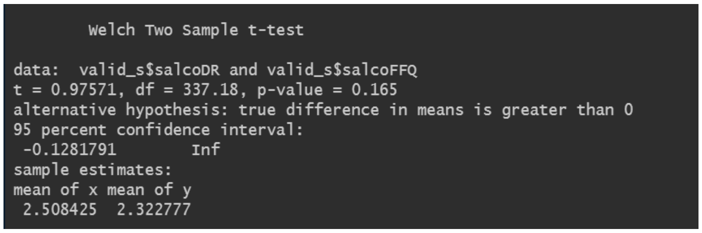
```
> favstats(valid_s$salcoDR)
 min      Q1   median       Q3      max     mean       sd   n missing
   0 1.32665 2.416609 3.601389 7.010706 2.508425 1.638988 173       0
> favstats(valid_s$salcoFFQ)
 min        Q1   median       Q3      max     mean       sd   n missing
   0 0.8717798 2.133073 3.443835 8.046738 2.322777 1.891219 173       0
```

(t.test: two-sided & "greater")



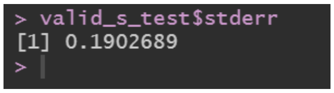
```
        Welch Two Sample t-test

data:  valid_s$salcoDR and valid_s$salcoFFQ
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 
```



```
        Welch Two Sample t-test

data:  valid_s$salcoDR and valid_s$salcoFFQ
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 
```



```
> valid_s_test$stderr
[1] 0.1902689
> 
```

we can see total alcohol consumption mean of DR is approximately 2.51 and mean of FFQ is nearly 2.32, with 95% confidence interval (-0.19, 0.56), SE ≈ 0.19

The p-value of this test is ≈ 0.165 > 0.05, meaning we can't reject the null hypothesis, so there is insufficient evidence to say FFQ underestimates total alcohol consumption

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.

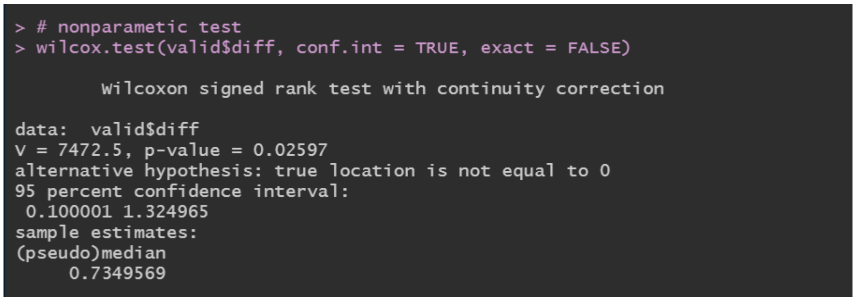
Ans:

Because of using nonparametric method, the hypothesis will be as following,

H0(null hypothesis): population distribution of alcohol consumption (diff in alcohol consumption) is centered at 0

Ha(alternative hypothesis): population distribution of alcohol consumption is shifted to right (alcohol consumption DR > FFQ)

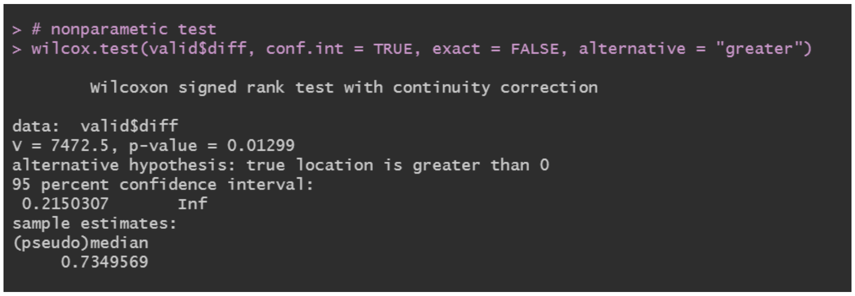
Rank test:



```
> # nonparametric test
> wilcox.test(valid$diff, conf.int = TRUE, exact = FALSE)

        Wilcoxon signed rank test with continuity correction

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



```
> # nonparametric test
> wilcox.test(valid$diff, conf.int = TRUE, exact = FALSE, alternative = "greater")

        Wilcoxon signed rank test with continuity correction

data:  valid$diff
V = 7472.5, p-value = 0.01299
alternative hypothesis: true location is greater than 0
95 percent confidence interval:
 0.2150307       Inf
sample estimates:
(pseudo)median
     0.7349569
```

With a p-value ≈ 0.013 < 0.05 there is evidence to reject the null hypothesis

There is difference in total alcohol consumption between DR and FFQ, with 95% confidence interval (0.1, 1.32) total alcohol consumption DR is more than FFQ, so we can conclude that there is evidence that FFQ underestimates total alcohol consumption.

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

Ans: I think in this case is better to use (6) square root transformed parametric method, because first, we have large sample (n=173) in this study. Second, the mean is more accurately represents the center of the distribution in this data set instead median. Third, between parametric and nonparametric, parametric is more powerful because it uses the actual number of the data, for nonparametric it usually use for smaller sample and is less accurate.