Q SCI 482 Story Assignment 5 (100 points)

Note for this assignment: for each test, at a minimum note whether these are one-tailed or two-tailed, the test statistic, and the resulting p-value.

Scenario (Q1, Q2, Q3): Which route is the best?

You have been arguing with room-mate (who also works in the same office as you) for quite some time about which route home from work is the fastest. They swear that traveling on the highway in rush hour traffic is still fastest, while you maintain that avoiding the highway and traveling on back routes is definitely faster. To settle the matter once and for all, you decide to both leave from work at exactly the same time, travel on your preferred route, and record the time of your journey. You repeat this experiment 30 times, with the data recorded in `"bestroute.csv"`. These are paired data since the traffic conditions are likely to be worse on both routes on bad traffic days, and better on both routes on good traffic days. In this exercise, you are going to compare the results from a non-parametric test with those from a parametric test, and then decide which one should be used to analyze the data.

Q1: Run the appropriate parametric paired-sample test on the data (14 points)

Read the data into R. Since the data are recorded in minutes and seconds, you will need to create new vectors `hwaytime` and `backtime` that contain the drive time converted to total seconds. Calculate the difference in time in seconds between these vectors (`hwaytime` minus `backtime`).

1a) Conduct a paired-sample parametric test the long way around to determine whether the highway time is significantly different from the back route time. (10 points)

1b) Conduct the test the short way around). Your answers should match. (4 points)

Q2: Run the appropriate non-parametric paired-sample test on the data (28 points)

2a) Conduct a Wilcoxon paired-sample test (this is non-parametric), the long way around, to determine whether the highway time is significantly difference from the back route time. Report the test statistic, p-value etc. (16 points)

2b) Conduct the same test but obtain the p-value using the normal approximation (use twice the area for a two-tailed test). (8 points)

2c) Conduct a Wilcoxon paired-sample test the short way around in R. If this differs from 2a) or 2b), explain why. (4 points)

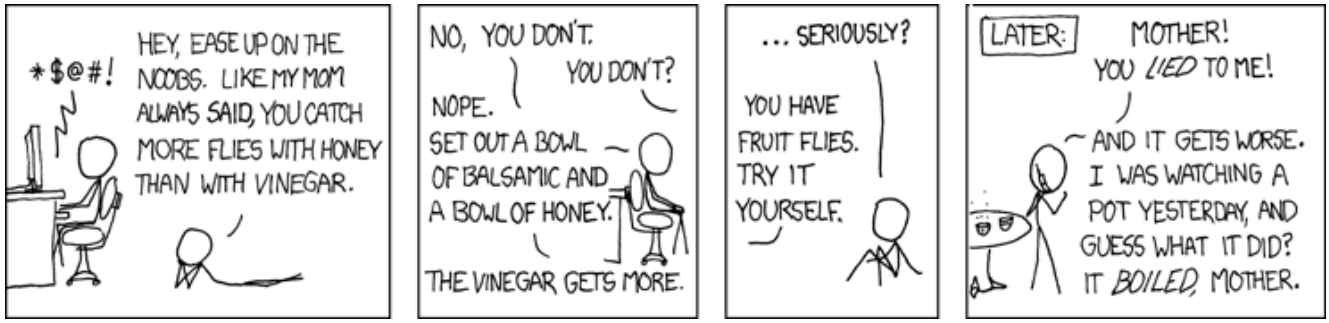
Q3: Are the differences normally distributed? (12 points)

3a) Test whether the differences in time are normally distributed using qqplot, histogram, and the Shapiro-Wilk test. What do you conclude about normality? (8 points)

3b) Given the answer in 3a), which of the tests used in Q1 or Q2 is appropriate, and what do you conclude about travel time on the two routes? Was your roommate correct or were you correct? (4 points)

Q4: Is it easier to trap flies with honey or vinegar? (16 points)

There is an old saying that "You catch more flies with honey than you do with vinegar", where the honey represents doing or saying pleasant things to get what you want, and vinegar represents doing unpleasant things to others to get what you want. In other words, you should try and do nice things rather than unpleasant things if you are trying to convince people to join your cause. As it turns out, however, there is some considerable debate about whether you actually do catch more (fruit) flies with honey than with vinegar, because balsamic vinegar contains ethanoic acid, which attracts fruit flies (Jouandet & Gallio 2015).



Source: https://xkcd.com/357/

Simple fruit fly traps can be set up by adding a substance to a glass, wrapping cellophane over the top, fastened with a rubber band, and poking a few holes in the top with a fork. An experiment was designed in a house containing a large number of fruit flies, that compared the number of fruit flies attracted to the following substances: ripe bananas, balsamic vinegar, honey, and sugar water. The results after one week of trapping were as follows:

Banana	Vinegar	Honey	Sugar water
23	18	5	10

4a) Run a chi-square goodness-of-fit test the long way around to determine if the distribution of fruit flies among the four categories could be due to random luck, or are due to a preference of fruit flies for some substances over others. (12 points)

4b) Run a direct R test. (4 points)

Q5: Do anadromous parents produce anadromous offspring? (30 points)

Catherine Austin at SAFS used otolith microchemistry (chemistry in fish ear stones) to determine if bull trout travel to the ocean and then come back to freshwater to spawn (this is called anadromy), or whether they remain in freshwater their whole lives. In particular, she is interested in whether anadromous mothers also have children that are anadromous. She has collected the following data:

	Offspring anadromy	
Maternal anadromy	No	Yes
No	16	13
Yes	5	1

We will be comparing the p-values from three forms of the chi-square test, asking whether the data support the idea that there is a link between anadromous mothers and anadromous offspring, or not. In each of 5a, 5b, and 5c, run the test the long way around.

5a) Run a chi-square test with no corrections (9 points)

5b) Run a chi-square test with the Yates continuity correction (9 points)

5c) Now compare the results of 5a-5b with the built-in chi-square test in R. Which form is used by default in R for a 2x2 contingency table? (4 points)

5d) What conclusions can be drawn from these data and the test? (2 points)

5e) The data in Q5 can only have a certain number of possible outcomes, given the number of anadromous mothers, and the number of non-anadromous mothers. For example, there were six anadromous mothers, and thus the data for offspring could only be 0-6, 1-5, 2-4, 3-3, 4-2, 5-1, or 6-0 for the offspring. In this question, you will need to explore what it would take to change the result of the test (this is related to the power of the test). If the test was significant, what is the smallest change in the data that would lead to a non-significant result? If the test was not significant, what is the smallest change in the data that would lead to a significant result? What does this tell you about the available sample size compared to the required sample size to reject the null hypothesis? (6 points)

R functions and constants that might be useful for this assignment

```
%o%        #outer product of two vectors (returns matrix)
abs()      #absolute value of the parameter
c()        #concatenates values into a vector
chisq.test #runs a chi-square test for goodness of fit or contingency tables
colSums()  #sum of each column in a matrix
det()      #determinant of a matrix. For 2x2 matrix, this is f11*f22 - f12*f21
hist()     #plots a histogram from a vector of data. Use "breaks" to control.
length()   #number of elements in a vector
log()      #natural logarithm (not log10)
matrix()   #creates a matrix from a vector of values
mean()     #mean of a vector of data
min()      #smallest value of a series of numbers
ncol()     #number of columns in a matrix
nrow()     #number of rows in a matrix
pchisq()   #probabilities of a chi-square distribution
pnorm()    #probabilities from a normal distribution
prod()     #product of all values in parameters (including vectors)
psignrank() #probabilities of Wilcoxon T distribution
qchisq()   #quantiles of the chi-square distribution
qqnorm()   #produces a QQ plot, used for examining normality
rank()     #returns the ranks (order number) of a vector of numbers
shapiro.test  #tests if vector of values is significantly different from normal
read.csv() #read in a CSV file, e.g. xdata <- read.csv(file="values.csv")
round()    #rounds number to specified number of decimal places
rowSums()  #sum of the rows in a matrix
```

```
sum()      #returns the sum of all values in a vector or matrix
wilcox.test() #conducts a Wilcoxon paired sample test.
t.test()   #runs a t-test, either one-sample or two-sample on a set of data
var()      #given the tvalue and d.f., returns the area to the left (CDF)
```

References

Jouandet GC & Gallio M (2015) Olfaction: catching more flies with vinegar. eLife 4:e10535.