# STAT540 Take-Home Midterm Exam Fall 2024

## Due: before Wednesday class on October 16

## Total Points: 140

This is a take-home midterm exam. You must work on it **independently**. **Do not discuss** any problems with your classmates. You can discuss with me if you have any confusion. Please hand in a hard copy of your midterm exam (compiled pdf file from R markdown) in class and email your R code to Yen-Yi Ho (hoyen@stat.sc.edu).

## Problem 1 Examine the effect of correlated data

### (a)

Simulate heights data for n=20 twin pairs (normal distribution with mean=160cm, sd=10cm). Assume the correlation coefficient within each twin pair ($\rho$) to be 0.5. Randomly assign the twin pairs into two groups so that one of a twin pair is in group 1 and the other is in group2. Present a scatter plot of the data. (10 points)

### (b)

Analyze this twin data using independent two-sample T test to determine whether the means of the two groups are equal. Write down the hypotheses, report p value, 95% Confidence interval for the hypothesis and interpret the results. (5 points) [Hint: use the function **mvrnorm** to simulated correlated data]

### (c)

Repeat (b) 1000 times and calculate type I error rate using the significance level 0.05 ($\alpha = 0.05$). (10 points)

$$\text{Type I error rate} = \frac{\#\text{ of times test results are significant}}{\#\text{ of simulation iterations}}.$$

### (d)

Repeat (b) (c) using a paired-test. Calculate type I error rate. (10 points)

### (e)

Repeat (b) (c) using permutation test and calculate type I error rate. (15 points)

### (f)

Varying $\rho$=0, 0.2, 0.5, 0.8, and n=10, 20, 100 and filled the type I error rate in the following table using two sample t-test for independent samples. Results must be presented in a table format to receive points. (15 points)

```
library(knitr)
mat<-matrix("", ncol=3, nrow=4)
```

```r
rownames(mat)<-paste("rho=", c(0, 0.2, 0.5, 0.8), sep="")
colnames(mat)<-c("n=10", "n=20", "n=100")
kable(mat, caption="Type I error rate using independent two sample t-test")
```

Table 1: Type I error rate using independent two sample t-test

| | n=10 | n=20 | n=100 |
|---|---|---|---|
| rho=0 | | | |
| rho=0.2 | | | |
| rho=0.5 | | | |
| rho=0.8 | | | |

### (g)

Perform the analysis in (f) using paired t-test, and permutation test. Results must be presented in a table format to receive points. (10 points)

### (h)

Comment on the tables obtained in (f) and (g). (10 points)

## Problem 2

Find the paper entitled "Analysis of the Electric Vehicles Adoption over the United States" posted on the course website. https://people.stat.sc.edu/hoyen/STAT540/Exam/ElectricVehicle.pdf. Use the data from Table 1 in the paper, and improve the plots presented in Figure 1 in the paper. (10 points)

## Problem 3

We will use the wine dataset from the paper by Cortez et al. (2009) "Modeling wine preferences by data mining from physicochemical properties." You can find the paper and the data (winequality-white.csv, winequality-red.csv, winequality-names.txt) used in the paper posted on the course website. To get started, a good source of information can be found at https://medium.com/@deepapandithu/a-data-science-approach-to-wine-tasting-exploring-the-wine-quality-dataset-part-1-4469c078cf5c.

You can read the data in R using the programming codes below as an example. Perform analyses to answer the following questions.

```r
urlwhite<-"https://people.stat.sc.edu/hoyen/STAT540/Exam/winequality-white.csv"
white<-read.csv(file=urlwhite, header=T, sep=";")
white[1:3,]
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
```

```
## 2       6
## 3       6
```

```
urlred<-"https://people.stat.sc.edu/hoyen/STAT540/Exam/winequality-red.csv"
red<-read.csv(file=urlred, header=T, sep=";")
red[1:3,]
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
##   quality
## 1       5
## 2       5
## 3       5
```

## (a)

What are the difference between red and white wines? Specifically create the figures (similar to the figures below) for all the variables in the data. Plot the data in log scale if the distributions are very skewed (as shown in the figures below). Comment on the differences you observed between red and white wines. (15 points)

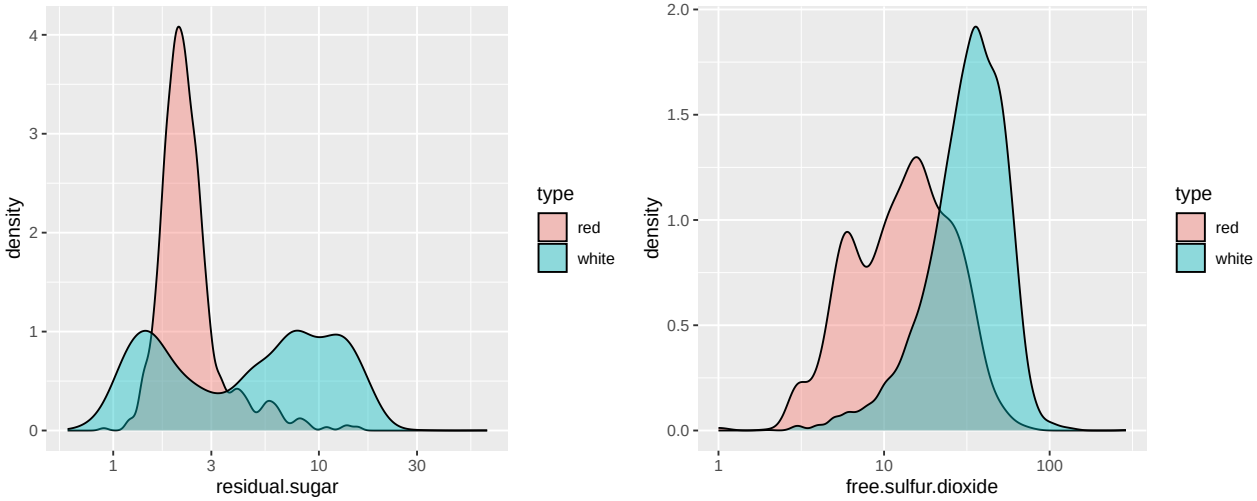


Figure 1: Density distributions of physicochemical properties in red and white wines.

## (b)

Perform correlation between the physicochemical properties and quality in red and white wines separately. Specifically create two heatmaps of correlation matrices (similar to Figure 2) for red and white wines, respectively. Comments on the correlations you observed in red and white wines, respectively. In addition, comment on the differences in the correlations between red versus white wines. (15 points)

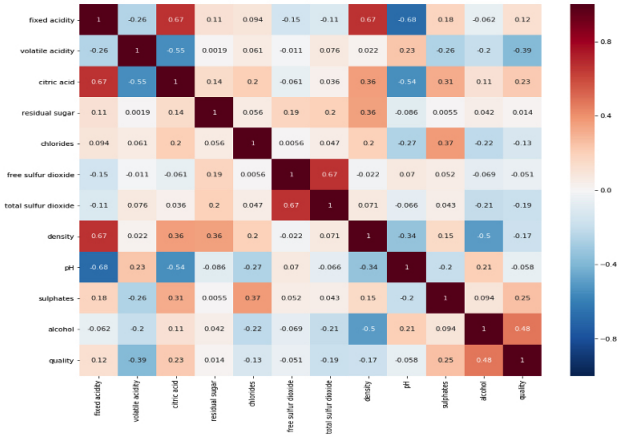


Figure 2: Heatmap of a correlation matrix.

## (c)

In red wines, what are the important factors contributing to wine quality? How about white wines? Are there any differences in the factors that determine wine quality between red and white wine? Perform your own analyses using statistical analysis and data visualization tools. (15 points)