Homework Assignment 3
(Due Friday, September 23, 2022 at 5PM)
Total points: 123

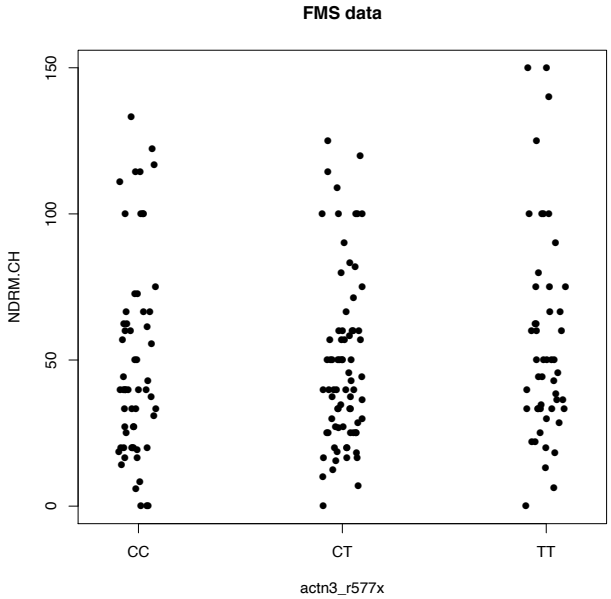
Please email your answer (compiled pdf file from R markdown) and R code to Yen-Yi Ho (hoyen@stat.sc.edu). Note: John Verzani's simpleR notes is available on https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf. To access the Simple dataset, install the **UsingR** package with the following code:

```
>install.packages( "UsingR" )
>library(UsingR)
>data(smokyph)  ### load smokyph dataset
```

1. Problem 10.2 From Verzani' s notes. This data set measures pH levels for water samples in the Great Smoky Mountains. Use the waterph column to test the null hypothesis that $\mu = 7$.
   (a) What is a reasonable alternative? (3 points)
   (b) Does a t-test seem appropriate? Why or why not? (5 points)
   [Hint: To access the waterph column in the dataset, the R command smokyph$waterph is more usually used.]
2. Problem 11.2 From Verzani' s notes. (7 points)
3. Use the FAMuss data to visualize if there are differences in nondominant arm muscle strength (NDRM.CH) between actn3_r577x genotype groups.

```
>fmsURL<-
"http://people.stat.sc.edu/hoyen/STAT704/Data/FMS_data.txt"
> fms<-read.delim(file=fmsURL, header=TRUE, sep="\t")
```

   (a) Plot the data shown in the figure below. Make sure the labels on the axes are correct as well as the main title. In addition, make sure the dots are solid dots (not hollow). [Hint: use stripchart] (5 points)
   (b) Remove two biggest observations in each genotype group, repeat (a). (10 points)



4. This exercise is for practicing central limit theorem.
   (a) Draw n=5 samples from uniform distribution and calculate sample means. Repeat this experiment 200 times, plot the distribution of sample means. (7 points) [Hint: To simulate **n** samples from uniform distribution, use **runif(n)**. Use **plot(density(x))**, where **x** is the vector contains the sample means from these 200 experiments.]
   (b) Repeat (a) but use n=100 (2 points).
   (c) Compare the sampling distributions obtained in (a) and (b), what do you observe? (3 points)
5. Perform the following steps in R:
   (a) Simulate 30 samples from Normal(mean=0, sd=1) (2 points)
   (b) Randomly assign 15 samples into control and 15 into treatment group (3 points) [Hint: Use **sample**]
   (c) Perform two sample T-test and report the p value. (2 points)
   (d) Randomly generate 1000 samples from uniform distribution, and plot the histogram of the 1000 samples. [Hint: Use **hist(x)** to plot a histogram of **x**.] (2 points)
   (e) Repeat **(a) (b) (c)** 1000 times, and stored the corresponding 1000 p values in a vector, plot a histogram using these 1000 p values. What is the distribution of p values? (10 points)

6. 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 (ρ) 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. Analyze this twin data using independent two-sample T test to determine whether the means of the two groups are equal. Write down the

ρ	n=10	n=20	n=100
0			
0.2			
0.5			
0.8			

hypothesis, report p value, 95% Confidence interval for $(\mu_1 - \mu_2)$ and interpret the results. (5 points)

(b) Repeat (a) 1000 times and calculate type I error rate. (7 points)

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

(c) Repeat (a) (b) using paired-test. Calculate type I error rate. (5 points)

(d) Repeat (a) (b) using permutation test and calculate type I error rate. (10 points)

(e) Varying ρ =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. (10 points)

(f) Perform the analysis in (e) using paired t-test, and permutation test (10 points).

(g) Comment on the table obtained in (e) and (f). (15 points)