STAT 158 Take home final exam **Due Friday May 10th by 2pm**

Instructions: for each question, copy and paste all relevant R code and output (including plots) into a document. The code and output for each question should go together. Make sure you answer any direct questions that are asked.

Do not obtain assistance from anyone else for this exam. Don't work with classmates or talk to tutors. If you need clarification on a question, you can email me at ffn@colostate.edu

1. The following code simulates data on resting heart rate vs. age, then simulates running multiple regression analysis repeatedly and saving the coefficients (i.e. slopes) from the model. This model assumes a slightly quadratic relationship between heart rate and age. Copy and run this code, then answer the questions below.

```
n <- 30
coefs_age <- rep(0,1e4)
coefs_age2 <- rep(0,1e4)
for (i in 1:1e4) {
  Age=round(runif(n,min=18,max=70))
  Age2 <- Age^2
  HR <- 94-Age*0.5+Age2*0.0035+rnorm(n,sd=10)
  model <- lm(HR~Age+Age2)
  coefs_age[i] <- summary(model)$coefficients[2,1]
  coefs_age2[i] <- summary(model)$coefficients[3,1]
}
mean(coefs_age)
mean(coefs_age2)
```

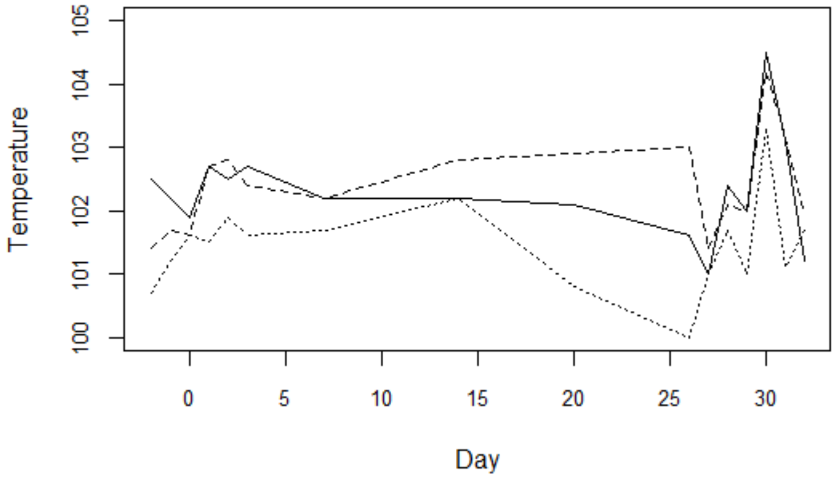
   a. What are the mean values of coefs_age and coefs_age2? Do these mean values seem "right" to you, based on the code in the simulation? Explain why or why not.
   b. Make a scatter plot of a single instance of the data simulation that is run 10,000 times in this for loop. Give this histogram the title "Heart rate vs. Age". Give the x axis the name "Age", and give the y axis the name "Heart rate". Make the data points a little bit smaller than their default.
   c. Add code that saves the p-values for the two slope coefficients in this model, and re-run the simulation. For each slope coefficient, report the estimated power.
   d. Modify the simulation so that "model" uses only Age as a predictor variable, not Age2. This means that the simulated regression will be incorrectly assuming a linear rather than quadratic relationship. Explain how this changes the results, both in terms of the values that the slope for Age takes on, and the values that the p-value for the slope takes on.

2. In the last homework, you ran a simulation to investigate how violating the assumption of equal variances can affect the properties of a t-test. For this problem, you will run a simulation to investigate how violating the assumption of normally distributed data can affect the properties of a t-test.

   a. The gamma distribution is skewed to the right. It contains a parameter called "shape". The R function for generating data from a gamma distribution is rgamma – you can read the details in R help.
      Make three historgrams, each of a sample of size n = 10,000 drawn from a gamma distribution, with shape = 1, shape = 0.5, and shape = 0.1. Use "breaks = 100" to force each histogram to have lots of bars. Describe what you see happening as the shape parameter gets smaller.
   b. Write a simulation that repeatedly draws two samples from a gamma distribution with shape = 1, then compares their means using a t-test. For this simulation, use n = 30 for the size of each sample. Write code that will save both the t-test statistic and p-value each time. Then make a histogram of the test statistics, and report the proportion of p-values less than 0.05. Note that, if the assumptions of the t-test are not violated, the p-value should be less than 0.05 5% of the time.
   c. Do the same thing in part b. two more times, using shape = 0.5 and shape = 0.1. Does this seem to have any effect on the distribution of the test statistics, or the proportion of p-values less than 0.05?
   d. Run the simulation three more times (once for each value of shape), using samples of size n = 10 rather than n = 30. Show the three histograms and three proportions of p-values less than 0.05. Did this have any noticeable effect on the results?

3. This question uses the "Ferret_Vaccine.csv" data set. This contains data (collected at CSU) on body temperatures and weights of ferrets, before and after being exposed to a strain of the influenza virus. Exposure was on day 28. Half the ferrets received a vaccination on day 0; half did not. The variable "Vaccine" specifies which group each ferret was in. Load the data set into R.

a. Report the mean and standard deviation for temperature and weight.
b. The plot below shows changes in temperature across days for the ferrets with ID#'s 574, 546, and 548:



Write code that re-creates this plot, and report the plot. You may want to consult the notes on plotting to make this. (Hint: the code "Temperature[Ferret.ID==574]" will get temperatures just for ferret #574)

c. Make another plot similar to the one in part b, but that just has two lines: one with the average temperature of ferrets who were vaccinated, and one with the average temperature of ferrets who were not vaccinated.
d. Make two new vectors that contain the maximum body temperature for each ferret across all days – one vector with maximum body temperatures for vaccinated ferrets, the other with maximum body temperatures for unvaccinated ferrets. Run a t-test comparing the two. Report the difference in these means, the test statistic, and the p-value.