CS 3130 / ECE 3530: Probability and Statistics for Engineers **Due Th 12/7**

Homework 8: Hypothesis Testing and Linear Regression

Instructions: Submit a single R Markdown file (.Rmd) of your work on Canvas by 11:59pm on the due date. You may also submit diagrams, drawings, etc. as image files (.png, .jpg, .gif)—they must be formatted into your .Rmd document (we won't look at them separately). **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, and do not show your answers to anyone.

1. Here we are going to test some hypotheses about cardiac measurements from the following data: `http://www.stat.ucla.edu/projects/datasets/cardiac.dat`

   Download this data set and load it into R with the following command:
   `cardiac = read.csv("cardiac.dat")`

   To understand what the variables mean, read the description of the data set here:
   `http://www.stat.ucla.edu/projects/datasets/cardiac-explanation.html`

   You want to test the hypothesis that men are more likely to have hypertension than women. Hypertension is the variable `hxofHT` (be careful, `hxofHT = 0` indicates they **do** have hypertension) and `gender` is male = 0, female = 1. Use the Fisher exact test to get a p value for this hypothesis test. Can you reject the null hypothesis at the $\alpha = 0.05$ level?

2. Now we will test the hypothesis that women have on average a lower resting blood pressure than men. Blood pressure is a continuous variable. You can get the values for males and females with this R code:

   ```
   maleBP = cardiac$basebp[cardiac$gender == 0]
   femaleBP = cardiac$basebp[cardiac$gender == 1]
   ```

   (a) What is the null hypothesis? What is the alternate hypothesis?

   (b) Plot two boxplots (in the same plot) for the male blood pressures and the female blood pressures.

   (c) Perform a two-sample t-test of the difference in the means of the two groups (assuming both are a Gaussian distribution). You may **not** use the function `t.test` in R, use the steps we went through in class and in the example R code. You **may** use `t.test` to check your answer! What is the t statistic? What is the p-value? At a significance level of $\alpha = 0.05$, can you reject the null hypothesis?

   (d) Now you want to test the hypothesis that people who had a cardiac event (heart attack, bypass, etc.) have higher blood pressure than those who did not. Replace the variable `gender` with `any.event` and repeat steps (a)-(c) above for this hypothesis.

(e) Lastly, you want to test the hypothesis that people who had a cardiac event have less efficient heart pumping, as measured by baseline cardiac ejection fraction (variable `baseEF` in this data). Again, repeat steps (a)-(c) to test this hypothesis.

3. In this exercise you will test hypotheses involving the correlation between two variables. We will use the `iris` data set, which is built in to R. Let's assume we have two random variables X and Y that are Gaussian distributed. The null hypothesis will be that these two random variables have zero correlation:

$$H_0 : \rho(X, Y) = 0$$

Under this null hypothesis, the sample correlation coefficient, r, can be transformed into the statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}},$$

which will have a Student's t distribution with $n-2$ degrees of freedom. As usual, n denotes the sample size, which in the case of the iris data is $n = 150$. Remember, you can compute the sample correlation of two vectors in R using the `cor` command. Answer the following: (**Hint:** You might try using the R function `cor.test` to double-check the answers you get, but you may not use it as the R commands that are asked for below.)

(a) Say you hypothesize that irises with long petals will also have long sepals, and short petals will coincide with short sepals (`x = iris$Petal.Length` and `y = iris$Sepal.Length`). What is the alternative hypothesis here for $\rho(X, Y)$?

(b) Plot a scatterplot of the x and y data.

(c) For a significance level of $\alpha = 0.05$ what is the critical value for the t statistic? Give the R command (one line) for computing this. What does it return?

(d) What is the value for the t statistic above?

(e) Give the R command (one line) for computing the p-value. What value does it return?

(f) Now say you hypothesize that longer sepals will be skinnier and shorter sepals will be fatter. What is the alternative hypothesis in terms of sepal width and length? (X is `iris$Sepal.Width`, and Y is `iris$Sepal.Length`)

(g) Repeat steps (b), (c), and (d) for this hypothesis.

4. Write an R function `my.regression(x, y)` that computes the regression of an independent variable `x` and a dependent variable `y`. It should return the estimated slope and intercept.

(a) Run your regression command on the `faithful` data, with `waiting` as the `x` and `eruptions` as the `y` variables. What is the slope and intercept? Draw the resulting regression line over top of a scatterplot of the data.

(b) Say you are watching the Old Faithful geyser, and you time the interval between two eruptions to be 82 minutes. Based on your regression analysis, how long should you expect this next eruption to last?

(c) Run the R command `lm` on the same data and test that your `my.regression` function gives the same results.

(d) Using the `summary` command on the results of your `lm` command, what is the p-value for the slope of your regression? What conclusion can you make from this?