STAT 451

Hand-in Homework 2

Problems 1-2 due Friday, Sep. 11 by midnight; problem 3 (grad) due Mon, Sep. 14 by midnight

Problems 1 and 2 are both be done in a single R Markdown document. Problem 3 is for graduate students only and is to be done in a separate R Markdown document and submitted separately.

1. Problem 54, p. 156, Chapter 5 problems. Answer the questions with complete sentences using in-line R code, e.g., "The proportion of tires that will last 40,000 or more miles is ...".

2. A study of hearing impairment in Dalmatian dogs examined the relationship between eye color and degree of impairment: deaf, unilateral hearing (one ear), bilateral hearing (both ears) (G.M. Strain, Deafness prevalence and pigmentation and gender associations in dog breeds at risk, *The Veterinary Journal*, 167(2004), pp.23-32). The data are in the file `dalmatian.csv`. Note that `dalmatian.csv` does not have a row for each individual dog, but summarizes the data using a count column. Write a short report which includes the following:

   - A two-way table of the counts with marginal totals.
   - A two-way table showing the conditional distribution of the hearing variable given each eye color.
   - A segmented bar chart comparing the conditional distributions in the previous table.
   - Two or three sentences (complete, grammatically correct, including context) describing the association between these two variables.

3. Graduate students only. To be done in a separate R Markdown file. The file `EPAtccb.csv` contains measurements of 1,2,3,4-Tetrachlorobenzene (TcCB) concentrations in parts per billion (ppb) from soil samples at a Reference site and a Cleanup area. In this exercise, you will assess whether a normal model is appropriate for the logarithm of the TcCB levels for the Cleanup site. If the logarithm of a variable follows a normal distribution, then the variable itself is said to follow a lognormal distribution. To pull out the Cleanup data from the full data set, use the `filter` command like

   ```
   cleanup <- filter(d,Area=='Cleanup')
   ```

   where `d` is what you have named the full data set. Questions (a)-(c) are to be answered for only the Cleanup area.

   (a) First, examine the raw TcCB levels in the Cleanup area by creating a histogram and normal probability plot and put them side-by-side. What's the shape of the distribution?

   (b) Assess whether a lognormal model looks appropriate by constructing a histogram and a normal probability plot of the log (base 10) of TcCB (put plots side-by-side). Discuss. Note: the R command is `log10(x)` where `x` is a variable.

(c) Estimate the parameters of the normal model for the log TcCB levels from the mean and standard deviation of the log TcCB levels. According to this model, what proportion of samples would have TcCB levels exceeding each of the following levels: 0.2 ppb, 0.5 ppb, 1 ppb, 20 ppb, 50 ppb? Then calculate the corresponding proportions for the actual data. How well do the observed data values match the model predictions?

(d) The Cleanup site is where a cleanup has taken place and the Reference site is an uncontaminated site. Compare the two distributions with either side-by-side boxplots or density plots (on the same plot) of the log TcCB levels.. Does the Cleanup plot appear to have been remediated to be like the Reference site?

Note for part (c): `pnorm` can calculate proportions for a vector of values:, e.g. `pnorm(c(1,2,4,8),10,4)` (try it) and the result will be a vector of the proportions less than 1,2,4 and 8 for a N(10,4) distribution. I can then create a table of the results with commands like this:

```
x <- c(1,2,4,8)
p <- pnorm(x,10,4)
data.frame(Value=x,Probability=p)
```

I can add a third column with observed proportions. Try this code to see what it gives you. You might want to round some of the variables in the table.