Week 6 In Class Exercise

Instructions: Your instructor will provide a dataset that contains n=1338 observations on seven variables. The data originate with author Brett Lantz and were developed for a book called *Machine Learning with R.* These data have many interesting possibilities, but for today we will just be comparing medical procedure charges (variable name: charges) by region of the country (variable name: region). Your goal is to conduct a thorough ANOVA analysis using both conventional and Bayesian techniques and to write a paragraph that integrates the statistical results.

1. Read in the data. Against my usual advice, try the read.csv() function for this dataset: it will automatically turn the region names (which are character data in the csv file) into a factor. You can use read.csv() as part of the "Import Data" dialog under the Environment tab in the upper right corner of R-studio, or you can write a line of code like this:

   insurance <- read.csv("/MyComputer/MyFiles/dWeek6insurance.csv")

   You will need to substitute a real path name in place of /MyComputer/MyFiles/ or use the setwd() command to indicate the folder where R should look for the file.

2. Visualize the data with boxplot(). You will need to use the formula notation in your call to boxplot(). Remember to use the tilde character like this: depVar ~ indepVar

   Note the results in a comment for later reference.

3. Report the means and SDs of each group in a comment. Here's a trick to run group means and standard deviations: the tapply() command (prounounced "tee-apply):

   tapply(insurance$charges, insurance$region, mean)
   tapply(insurance$charges, insurance$region, sd)

4. Run a conventional ANOVA using the aov() command and report an overview of the results in a comment. Store the results in an object called aovOut.

5. Now conduct a Bayesian analysis. Refer to pages 104-109 of the text as needed. Don't forget that after running anovaBF, you will need to run the posterior() command to obtain the MCMC posterior distributions of the parameters from the ANOVA. Report an overview of the results in a comment.

6. Finally, write a paragraph-length comment that integrates the various pieces of evidence. It would be great to document the group means (from your answer to #3 above) before doing a read-out on the ANOVA results. You should also state a null hypothesis (and, optionally, an alternative hypothesis) as well as the alpha level.

7. The default stats package in R contains a procedure to run Tukey's Honestly Significant difference and make every pairwise comparison. Try this:
   TukeyHSD(aovOut)

   Examine the final column of adjusted p-values, labeled "p adj," and take note of any significant pairwise comparisons. Are there any significant pairwise differences? Document these in a comment. Also make note of the general pattern.

8. Install and library the emmeans package. This package computes estimated marginal means. If you are curious about the calculation and usage of these, try: https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html .

9. Run emmeans() on your aovOut object and plot the results. The plots shows each group mean as a dot, surrounded by the confidence interval around that mean. You can also view those confidence intervals with the summary command:
   emOut <- emmeans(aovOut, "region")
   plot(emOut)
   summary(emOut)

10. Examine a pairwise p-value plot (using the Tukey adjustment) with *pwwp(emOut)*. The p-values are on the X-axis. The Y-axis lists the groups, and the vertical lines connect the groups to show what is being compared.

11. Finally, write a paragraph-length comment that integrates the various pieces of evidence. Which groups are credibly different from one another? What additional insights do you have?