Advanced Statistics HW4

Due date: October 30, 2018

Exercises 1

Load the `WHEATUSA2004` data frame from the `PASWR2` package.

(a) Find the quantiles, deciles, mean, maximum, minimum, interquartile range, variance, and standard deviation for the variable acres. Comment on what the most appropriate measures of center and spread would be for this variable. What is the USA's 2004 total harvested wheat surface area?

(b) Which states are below the 20th percentile? Which states are above the 80th percentile? In which quantile is WI (Wisconsin)?

(c) Create a frequency and a density histogram in the same graphics device using square plotting regions of the values in ACRES.

(d) Add vertical lines to the density histogram from (c) to indicate the location of the mean and the median.

(e) Create a boxplot of the acres and locate the outliers' communities and their values.

(f) Determine the state with the largest harvested wheat surface in acres. Remove this state from the data frame and compute the mean, median, and standard deviation of acres. How do these values compare to the values for these statistics computed in (a)?

Exercises 2

Access the data from url http://www.stat.berkeley.edu/users/statlabs/data/babies.data and store the information in an object named `BABIES`. A description of the variables can be found at http://www.stat.berkeley.edu/users/statlabs/labs.html.

These data are a subset from a much larger study dealing with child health and development.

(a) Create a "clean" data set that removes subjects if any observations on the subject are "unknown." Note that `bwt`, `gestation`, `parity`, `age`, `height`, `weight`, and `smoke` use values of 999, 999, 9, 99, 99, 999, and 9, respectively, to denote "unknown." Store the modified data set in an object named `CLEAN`.

(b) Use the information in `CLEAN` to create a density histogram of the birth weights of babies whose mothers have never smoked (`smoke=0`) and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke (`smoke=1`). Make the range of the x-axis 30 to 180 (ounces) for both histograms. Superimpose a density curve over each histogram.

(c) Based on the histograms in (b), characterize the distribution of baby birth weight for both non-smoking and smoking mothers.

(d) What is the mean weight difference between babies of smokers and non-smokers? Can you think of any reasons not to use the mean as a measure of center to compare birth weights in this problem?

(e) Create side-by-side boxplots to compare the birth weights of babies whose mothers never smoked and those who currently smoke. Use traditional graphics (`boxplot()`), lattice graphics (`bwplot()`), and `ggplot` graphics to create the boxplots.