ELEMENTS OF STATISTICS AND PROBABILITY
R EXERCISE
Submission Deadline: Friday 2 October 2026, 23.59

The data frame "babies" contains data from a larger study dealing with child health and development. The variables included are the following:

bwt: Birth weight in ounces (999 unknown)

gestation: Length of pregnancy in days (999 unknown)

parity: 0= first born, 9=unknown

age: mother's age in years

height: mother's height in inches (99 unknown)

weight: mother's pre-pregnancy weight in pounds (999 unknown)

smoke: smoking status of mother (0=not now, 1=yes now, 9=unknown)

1. Create a "clean" data set that removes subjects if any observations on the subject are "unknown." Note that bwt, gestation, parity, height, weight, and smoke use values of 999, 999, 9, 99, 999, and 9, respectively, to denote "unknown." Store the modified data set in an object named CLEAN.
2. Use the information in CLEAN to create a 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.
3. Based on the histograms in (2), characterize the distribution of baby birth weight for both non-smoking and smoking mothers.
4. 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 centre to compare birth weights in this problem?
5. Create side-by-side boxplots to compare the birth weights of babies whose mother has never smoked and those who currently smoke.

6. What is the median weight difference between babies who are firstborn and those who are not?
7. What is the mean pre-pregnancy weight difference between mothers who do not smoke and those who do? Can you think of any reasons not to use the mean as a measure of centre to compare pre-pregnancy weights in this problem?
8. Compute the body weight index (BWI) for each mother in CLEAN. Recall that BWI is defined as kg/m2 (0.0254 m= 1 in., and 0.45359 kg= 1 lb.).
9. Add the variables weight in kg, height in m, and BWI to CLEAN and store the result in CLEANP.
10. Characterize the distribution of BWI.
11. Group pregnant mothers according to their BWI quartile.
12. Following the previous question find the mean and standard deviation for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Find the median and IQR for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke.
13. Based on your answers to the previous question, would you characterize birth weight in each group as relatively symmetric or skewed?
14. Create histograms of bwt conditioned on BWI quartiles and whether the mother smokes to verify your previous assertions about the shape.
15. Does it appear that BWI is related to the birth weight of a baby? Create scatterplot of birth weight (bwt) versus BWI while conditioning on BWI quartiles and whether the mother smokes to help answer the question.
16. Create a table of smoke by parity. Display the numerical results in a graph. What percent of mothers did not smoke during the pregnancy of their first child?

Deliverables

- A short report with comments on your findings
- Your script in an R file

Grading

Your score at the assignment will count for 20% of your final score.