STAT 210
Applied Statistics and Data Analysis:
Homework 3

Due on Sep. 25/2022

Question 1

Access the data from url http://www.stat.berkeley.edu/users/statlabs/data/babies.data and store the information in an object named `BABIES` using the function `read.table()`. Use the option that reads the first line as header.

A description of the variables can be found at http://www.stat.berkeley.edu/users/statlabs/labs.html. Look for the data set Birth Weight II. These data are a subset from a much larger study dealing with child health and development.

(i) 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`. The function `subset` may be useful here.

(ii) 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`). Use a common range of the x-axis for both histograms. Superimpose a density curve over each histogram. Use informative titles and labels for your graphs. Comment on what you observe.

(iii) The body weight index or body mass index (bmi) is defined as the weight of a person divided by the height squared and is measured in units of kg/m^2. Compute the bmi for each mother in `CLEAN`. Observe that you have to convert the measurements in the data frame to metric (0.0254 m= 1 in., and 0.45359 kg= 1 lb.). Modify the variables `weight` and `height` so that they now appear in metric units (kg and m), and add `bmi` to `CLEAN` and store the result in `CLEANP`. Count how many subjects have `bmi` above 30.

Question 2

The file `data_q4.csv` has four simulated samples of size 20 coming from the following distributions

- Standard Cauchy, (`rcauchy(20)`)
- Chi-square with 2 degrees of freedom, χ^2_2, (`rchisq(20,2)`)
- Lognormal with standard parameters, (`rlnorm(20)`)
- Weibull with shape parameter 2 (`rweibull(20,2)`)

You have to identify which is which using quantile plots. Since you will need to draw quantile plots with respect to distributions other than the normal, it will be convenient to use a new function named `qqPlot` in the package `car`. You will need to install this package. If you are using RStudio, select the `Packages` tab on the panel on the right and then select the `Install` tab. Type `car` on the pop-up window and click install. After installing, you need to load the package using `library(car)`.

The function `qqPlot` has syntax

```
qqPlot(x, dist = 'weibull', shape = 2)
```

for plotting a quantile graph of vector x with respect to the Weibull distribution with shape parameter 2. The default distribution for qqPlot is the normal distribution. You can find more details in the help for qqPlot. By default, this function draws confidence bands which I find in many cases of little use, and in some cases misleading. If you don't want them in your graph, add envelope = FALSE in your call.

Explain clearly the reasons for your choices.