Discussion 2

Student's Name Here

Warmup

Work on the Week 2 Quiz question 6 with your neighbor. Choose the answers you think make sense and explain how you know. What improvement could be made to the graphs to make this question easier to answer?

Comparison of Two Sets of Data

Possible relationships between mothers' characteristics and birth outcomes have been studied extensively. The low birth weight data set `lbw.csv` was collected as part of a larger study at Bayside Medical Center, Springfield, Massachusetts in 1986. It contains information on 189 births to women that were seen in the obstetrics clinic.

Exercise 1. Load the data into your environment by reading in the .csv file (`lbw.csv`) to the variable `lbw`.

a. Download the .csv file into the same folder as this Rmd file (drag and drop this file directly to the folder as opening the .csv file in other programs such as Numbers can cause issues)

b. Set the folder that holds both the Discussion 2 files as your working directory by navigating to that folder in the Files pane of RStudio, and select "Session > Set Working Directory > To Files Pane Location" from the top RStudio menu.

c. Run the code below to define the variable lbw. The lbw variable is a data frame that has 189 observations of 11 variables. Confirm that it shows up in the Environment tab of RStudio. Click on the blue arrow next to the lbw name to see some information about the 11 variables it contains.

```
lbw <- read.csv("lbw.csv", header = TRUE)
```

I have included vectors of the relevant data at the bottom of this page if you have having issues importing and would rather just define vectors of data.

d. View the data frame to see what the data looks like, by running `View(lbw)` in the console. Or, click on the table icon in the Environment tab. (This will run `View(lbw)` in the console for you.)

Exercise 2. We will be focusing on the variables `bwt`, `low`, and `smoke`.

- `bwt` is the birth weight (in grams) for each of the 189 babies observed.
- `low` is a variable that is 0 if the birth weight is at least 2500 g and 1 if the birth weight is less than 2500 g.
- `smoke` is a variable that is 0 if mother reported not smoking during pregnancy and 1 if mother reported smoking during pregnancy.

a. Identify the variable type for each of the three variables, then run the following code to see how R has identified the variables in lbw. Identify whether any of the three are saved as the incorrect type.

```
str(lbw)
```

```
## 'data.frame':    189 obs. of  11 variables:
##  $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

We want `low` and `smoke` to be categorical variables, but R is currently recognizing them as integers (numeric).

b. Run the following code to resave low and smoke as categorical vectors in the lbw data frame. Notice the `$` after the data frame name `lbw` pulls up a list of all of the columns that are defined in `lbw`. And the `as.factor()` function changes the variable type to categorical.

c. Reference the Environment tab or rerun `str(lbw)` to confirm low and smoke have been updated correctly.

```
lbw$low <- as.factor(lbw$low)
lbw$smoke <- as.factor(lbw$smoke)

str(lbw)
```

```
## 'data.frame':    189 obs. of  11 variables:
##  $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ low  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ smoke: Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 1 2 2 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

Exercise 3. Consider comparing the variable for birthweight (`bwt`) across the groups with `smoke = 0` (non smoking) and `smoke = 1` (smoking).

a. Create comparative boxplots and histograms of the variable `bwt` between babies with mothers who had smoking status of 0 = No and 1 = Yes.

(i) Save off two data frames, SmokeY and SmokeN, to hold the data for those babies whose mothers smoked and those whose did not.

```
# First make a data frame `SmokeY` for just those babies whose mother smoked:
SmokeY <- subset(lbw, smoke == "1")
```

```
# Then make a data frame `SmokeN` for just those babies whose mother did not smoke:

# After the above two steps look at your environment tab to see
# what the variables have stored in them. (Click the arrow next to each name)
```

(ii) Then save off the two vectors of bwt for those two dataframes into bwtSmokeY and bwtSmokeN.

```
# Define `bwtSmokeY` to be the btw values for babies whose mothers smoked
bwtSmokeY <- SmokeY$bwt

# Define `bwtSmokeN` to be the btw values for babies whose mothers did not smoke

# After the above two steps look at your environment tab to see
# what the variables have stored in them.
```

(iii) Update the following boxplot code to include labels that show which data is which. If you want to knit the .Rmd, set `eval = TRUE` at the top of the chunk.

```
boxplot(bwtSmokeN, bwtSmokeY,
        main = "Birthweights", xlab = "Birthweight (g)")
```

(iv) Update the following histogram code so that both histgrams have x axis classes from 0 to 5000 with a width of 250. Also, choose a more useful y axis for both graphs. Why is it important that the x and y axis are consistent across the two histograms? If you want to knit the .Rmd, set `eval = TRUE` at the top of the chunk.

```
# this makes two rows and 1 column for graphs
par(mfrow = c(2,1))

hist(bwtSmokeY,
     main = "Smoking Mothers", xlab = "Birthweight")

hist(bwtSmokeN,
     main = "Non Smoking Mothers", xlab = "Birthweight")

# this makes one row and one column for graphing (restoring to default)
par(mfrow = c(1,1))
```

b. Compare the center, variability and shape of the two graups' data using the graphs and numeric summaries. Write your own code to find the numeric summaries.

If you had issues loading the data from the .csv file in problem 1:

Birthweight (Smoking): 2557, 2594, 2600, 2663, 2665, 2769, 2769, 2782, 2821, 2906, 2920, 2948, 2948, 2977, 2977, 2922, 3005, 3033, 3042, 3076, 3076, 3090, 3132, 3147, 3203, 3260, 3303, 3317, 3321, 3331, 3374, 3430, 3444, 3572, 3629, 3637, 3643, 3651, 3651, 3756, 3856, 3884, 3940, 4238, 709, 1135, 1790, 1818, 1885, 1928, 1928, 1936, 2084, 2084, 2125, 2126, 2187, 2211, 2225, 2296, 2296, 2353, 2367, 2381, 2381, 2410, 2410, 2414, 2424, 2466, 2466, 2466, 2495, 2495

Birthweight (Non-Smoking): 2523, 2551, 2622, 2637, 2637, 2722, 2733, 2750, 2750, 2778, 2807, 2835, 2835, 2836, 2863, 2877, 2877, 2920, 2920, 2920, 2977, 2977, 3062, 3062, 3062, 3080, 3090, 3090, 3100, 3104, 3175, 3175, 3203, 3203, 3225, 3225, 3232, 3232, 3234, 3274, 3274, 3317, 3317, 3374, 3402, 3416, 3459, 3460, 3473, 3475, 3487, 3544, 3572, 3586, 3600, 3614, 3614, 3629, 3651, 3651, 3699, 3728, 3770, 3770, 3770, 3790, 3799, 3827, 3860, 3860, 3884, 3912, 3941, 3941, 3969, 3983, 3997, 3997, 4054, 4054, 4111, 4153, 4167, 4174, 4593, 4990, 1021, 1330, 1474, 1588, 1588, 1701, 1729, 1893, 1899, 1928, 1970, 2055, 2055, 2082, 2100, 2187, 2240, 2240, 2282, 2301, 2325, 2353, 2381, 2395, 2438, 2442, 2450, 2495, 2495