Lab2: Introduction to the data

We will focus on a random sample of 20,000 people from the BRFSS survey (http://www.cdc.gov/brfss) conducted in 2000. While there are over 200 variables in this data set, we will work with a small subset.

We begin by loading the data set of 20,000 observations into the R workspace: **source**("http://www.openintro.org/stat/data/cdc.R")

To view the names of the variables, type the command: **names**(cdc)

This returns the names genhlth, exerany, hlthplan, smoke100, height, weight, wtdesire, age, and gender. Each one of these variables corresponds to a question that was asked in the survey. For example, for **genhlth**, respondents were asked to evaluate their general health, responding either excellent, very good, good, fair or poor. The **exerany** variable indicates whether the respondent exercised in the past month (1) or did not (0). Likewise, **hlthplan** indicates whether the respondent had some form of health coverage (1) or did not (0). The **smoke100** variable indicates whether the respondent had smoked at least 100 cigarettes in her lifetime. The other variables record the respondent's **height** in inches, **weight** in pounds as well as their desired weight, **wtdesire**, **age** in years, and **gender**.

1. Check the size of the data frame: **dim**(cdc)
2. Display the first and last few entries (rows): **head**(cdc) and **tail**(cdc)
3. A good first step in any analysis is to get a few summary statistics and graphics. Display a numerical summary (minimum, first quartile, median, mean, second quartile, and maximum) for weight: **summary**(cdc$weight)
4. Compute summary statistics one by one: **mean**(cdc$weight), **var**(cdc$weight) and **median**(cdc$weight)
5. For categorical data, consider the sample frequency or relative frequency distribution. The function **table** counts the number of times each kind of response was given. For example, to see the number of people who have smoked 100 cigarettes in their lifetime, type **table**(cdc$smoke100) or instead look at the relative frequency distribution by typing: **table**(cdc$smoke100)/20000
6. Next, make a bar plot of the entries in the table : **barplot**(**table**(cdc$smoke100))
7. You could also break this into two steps by typing the following: smoke <- **table**(cdc$smoke100) and then **barplot**(smoke)
8. The **table** command can be used to tabulate any number of variables that you provide. Examine which participants have smoked across each gender: **table**(cdc$gender,cdc$smoke100)
9. Create a mosaic plot of this table: **mosaicplot**(**table**(cdc$gender,cdc$smoke100))
10. Access a subset of the full data frame using row-and-column notation. What do the following commands display: cdc[567,6], cdc[1:10,6] and cdc[1:10,]
11. It's often useful to extract all individuals (cases) in a data set that have specific characteristics. We accomplish this through *conditioning* commands. Consider expressions like cdc$gender == "m" or cdc$age > 30
12. Extract just the data for the men in the sample: mdata <- **subset**(cdc, cdc$gender == "m")
13. Use several of these conditions together: m_and_over30 <- **subset**(cdc, gender == "m" & age > 30)
14. Two common ways to visualize quantitative data are with box plots and histograms. Construct a box plot for the height: **boxplot**(cdc$height)

15. Compare the locations of the components of the box by examining the summary statistics: **summary**(cdc$height) and confirm that the median and upper and lower quartiles reported in the numerical summary match those in the graph.
16. Compare the heights of men and women with **boxplot**(cdc$height ~ cdc$gender)

(The ~ character can be read *versus* or *as a function of).*

17. Display a histogram for the age of respondents: **hist**(cdc$age)
18. Control the number of bins by adding an argument to the command: **hist**(cdc$age, breaks = 50)
19. Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables. Try the following commands and check the documentation of the different functions.

```
library(tidyverse)
ggplot(data = cdc) + geom_point(mapping = aes(x = weight, y = wtdesire),color="blue")
```

or **ggplot**(data = cdc, mapping = **aes**(x = weight, y = wtdesire)) + **geom_point**() + **geom_smooth**()

On your own

20. Consider a new variable: the difference between desired weight (wtdesire) and current weight (weight). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called wdiff.
21. What type of data is wdiff? If an observation wdiff is 0, what does this mean about the person's weight and desired weight? What if wdiff is positive or negative?
22. Describe the distribution of wdiff in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?
23. Using numerical summaries and a side-by-side box plot, determine if men tend to view their weight differently than women.
24. Find the mean and standard deviation of weight and determine what proportion of the weights is within one standard deviation of the mean.
25. Display all plot types studied in class using **tidyverse** library (Use any variables of your choice in the cdc dataset. Hint: Check the documentation of **geom_boxplot**(), **geom._histogram**(),**geom_bar**(), **geom._col**(), **coord_polar**(), **coord_flip**(), and **labs**().