Math 207
Introduction to Statistics

NHANESIII Data Set: Report II

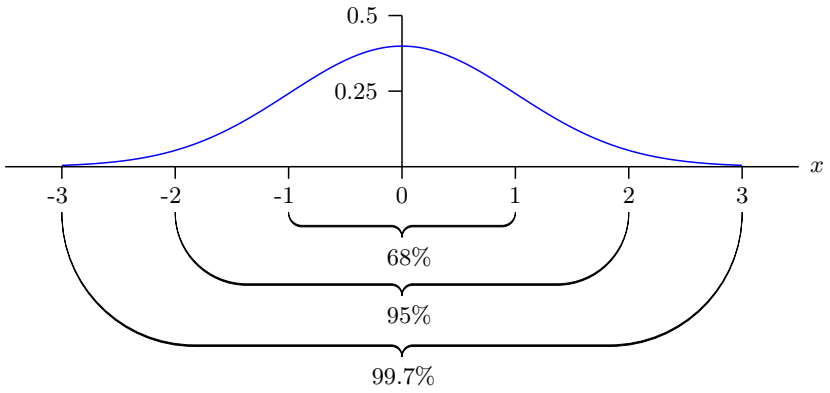
In a prior assignment you were asked to review the NHANES code books then decide on four questions about the data that interest you and that you would like to explore. You were asked to include variables of each of the four types (continuous, discrete, ordered and unordered).

For this assignment, please prepare an RMarkdown file that answers the questions below. Please include your R code. Include text to explain what you are calculating and what the results mean. Recall that the following commands read the NHANES data files into `R`.

```
adult = read.csv("adult.csv", header=TRUE)
exam  = read.csv("exam.csv",  header=TRUE)
lab   = read.csv("lab.csv",   header=TRUE)
youth = read.csv("youth.csv", header=TRUE)
```

Your operating system may require you to include full or partial path names rather than just the file names.

1. **Normal Curve**. The standard normal (Gaussian) distribution is an ideal histogram which we can often use as a model data.



a) Does the continuous (quantitative) variable that you chose to study fit a normal curve?

a.i) As a first test, compute the fraction of the data that is within 1, 2 and 3 standard deviations of the mean. For example,

```
e1 = exam[exam$BMPBMI != 8888, ]
e1bmi = e1$BMPBMI
m  = mean(e1bmi, na.rm=T)
s  = sd(e1bmi, na.rm=T)
length(e1bmi[(m - s < e1bmi) & (e1bmi < m + s)]) / length(e1bmi)
# Repeat for 2*s and 3*s
```

Do the percents roughly fit 68%, 95% and 99.7%? If so, what does this tell you about how well the data fits a normal curve?

a.ii) As a second test, make a QQ plot comparing the quantiles of your data to the quantiles of a normal curve.

```
qqnorm(e1bmi)
```

If your data fits a normal curve, the QQ plot should be a straight line. Comment on your plot and the conclusion you reached in a.i).

a.iii) The Kolmogorov-Smirnov test can be used to test the null hypothesis that a given data set was generated from a specified distribution. A larger p-value for the test (e.g., at least 5%) indicate that the data probably differs from the expected distribution simply due to chance. A small p-value suggests that the difference is not due to chance and that the data probably came from some other distribution. Run the test on your data:

```
ks.test(e1bmi, "pnorm", m, s)
```

Report the p-value and comment on its significance. How does your result relate to your conclusions in a.i) and a.ii)?

b) Repeat part a) using the weight distribution variable `exam$BMPWT`.

2. **Correlation**. In many situations, we may seek to use measurements of one or more variables to predict the value of another variable. For example, we could seek to predict the sale price of a house based on the sale prices of other houses in the neighborhood, size of the house, number of bedrooms, and interest rates. We could predict the stopping distance of a car using speed, weight, and road conditions. Standardized test scores, high school grade point average, and enrollment in advanced classes are used in college admissions because these variables can be used to estimate college success.

The correlation coefficient r is a measure of linear association between two lists $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ of variables. It is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \text{mean}_x}{\text{sd}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{sd}_y} \right) = \text{mean of the } x \text{ and } y \text{ values measured in standard units}$$

Observations about r:

- The correlation r is a number between -1 and 1.
- It measures linear association between the lists.
- It measures the clustering of the (x_i, y_i) points around a line.
- If r is close to 1 or -1, the points fit rather tightly along a line.
- The correlation between X and Y is equal to the correlation between Y and X.

a) Make a plot of your two quantitative variables. Be sure to label the axes.

```
plot( x variable, y variable, xlab="X axis label", ylab="Y axis label")
```

Does there appear to be a linear association between the variables? If so, is it strong (close to ± 1) or weak (closer to 0)?

b) Compute the correlation between the two variables and comment on your findings. What, if anything, do the sign and magnitude of the correlation indicate?

```
cor(x variable, y variable)
cor(y variable, x variable)
```

c) A strong correlation can indicate that one variable can cause the other (e.g., between hours spent reading and student GPA). It can mean that there is some underlying common cause for both (e.g., between drinking and lung cancer where the confounding factor is an association between smoking and drinking). It can also be simply a statistical artifact (e.g., between sunspot activity and stock returns over some conveniently selected time periods). Does the correlation between your variables suggest a cause and effect relationship, an underlying confounding factor or a statistical artifact?