ST 516 - Homework 9 Paul ReFalo 11/26/17

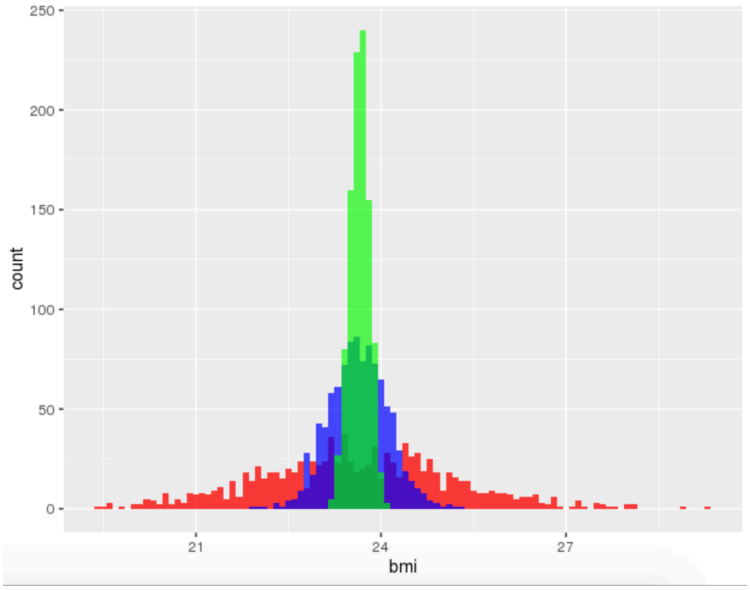
- **Question 1 - Simulation Study**

1. Submit R code to complete part 1(a). This will help you verify you know how to complete the coding part of this section, and allow us check you are on the right path.

```
> yrbss_2003 <- readRDS("yrbss_2003.rds")
> yrbss_2013 <- readRDS("yrbss_2013.rds")

> set.seed(8226)
> n_sim <- 1000
> pop_sd <- sd(yrbss_2013$bmi)
> pop_mean <- mean(yrbss_2013$bmi)

> get_means <- function(n, n_sim, pop_sd, pop_mean) {
+   replicate(n_sim, mean(rnorm(n, sd = pop_sd, mean = pop_mean)))
+ }
> ns <- c(10, 100, 1000)
> means <- lapply(ns, get_means, n_sim = n_sim, pop_sd = pop_sd, pop_mean = pop_mean)
```




```
> # === Graph Distributions ===
> bmi10 <- unlist(means[1])
> bmi100 <- unlist(means[2])
> bmi1000 <- unlist(means[3])

> df <- data.frame(bmi =
c(bmi10, bmi100, bmi1000), yy =
rep(letters[1:3], each = 1000))
```

Distribution of:

- n = 10 (red)
- n = 100 (blue)
- n = 1000 (green)

```
> ggplot(dat,aes(x = bmi)) +
+   geom_histogram(data = subset(df, yy == 'a'), fill = "red", alpha = 0.75, binwidth
= 0.1) +
+   geom_histogram(data = subset(df, yy == 'b'), fill = "blue", alpha = 0.7, binwidth
= 0.1) +
+   geom_histogram(data = subset(df, yy == 'c'), fill = "green", alpha = 0.65,
binwidth = 0.1)
```

```
> spread_sampdist <- sapply(means, sd)  # get SD of means vector containing means for
n = 10, 100, 1000
> spread_sampdist
[1] 1.576819 0.481948 0.158867
>
> true_se <- pop_sd/sqrt(ns)
> rbind(round(spread_sampdist, 3), round(true_se, 3))
      [,1]  [,2]  [,3]
[1,] 1.577 0.482 0.159
[2,] 1.586 0.501 0.159
>
> meansTally <- sapply(means, mean) # get mean of these data sets
> meansTally
[1] 23.6124 23.6370 23.6470
```

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 replicates	23.61	23.64	23.65
Calculated sd	1.577	0.482	0.159
True se	1.586	0.501	0.159

2. A tentative written answer for part 1(a).

The standard deviation calculated from 1000 replicates decreases significantly from 1.577 to 0.482 to 0.159 as the sample size increases from 10 to 100 to 1000 respectively. This is shown in the above graph with n = 10 as red, n = 100 as blue, and n = 1000 as green. This is due to the Central Limit Theorem. The calculated mean is little changed across the respective sample sizes as is the difference in the calculated standard deviation and the true standard error. By n = 1000 the calculated standard deviation and the standard error agree at 0.159.

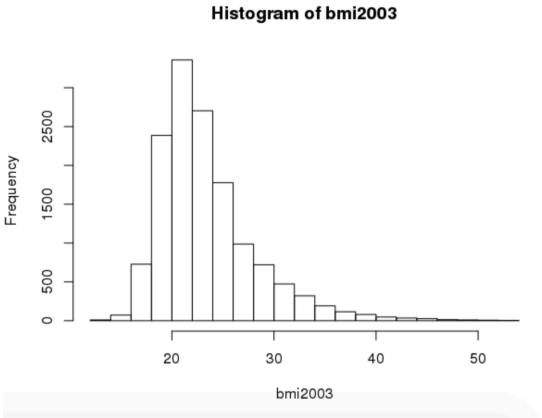
• **Question 2 - Data Analysis**

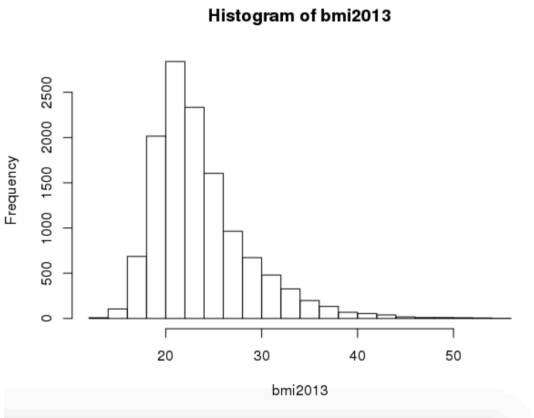
For each part, a translation of the questions of interest into inferential questions about parameters in a statistical model and a description of the method(s) you will use to answer the questions of interest. You do not need to have completed any of the analysis.

Because the sample sizes are large (x for yrbss of 2003 and y of yrbss of 2013) the Central Limit Theorem applies. The schools taking the survey self-selected. Also, students at these schools were not presumably not taken at random but rather directed by the school's administration to take the survey.

- How has the BMI of high-school students changed between 2003 and 2013? Are high-schoolers getting more overweight?

For this question, since the sample size is large and there is not evidence of skewness, I plan to perform a two-sample t-test on yrbss_2003$bmi and yrbss_2013$bmi in order perform hypothesis testing and report confidence intervals to see if the difference in bmi from 2003 is different from 2013.





- Are male high-schoolers more likely to smoke than female high-schoolers in 2013?

  Again the sample sizes are large but not from a randomized collection. For this I plan to also use a proportion test to compare the males in yrbss_2013$q33 that have non-zero reported days of smoking to that of females in 2013 (excluding NA results).

- How much TV do high-schoolers watch?

  Again the sample sizes are large but not from a randomized collection. For this question there is no notion of a treatment. This is an observational study. I plan to analyze the data from 2003, then from 2013, and lastly the combined data from 2003 and 2013 results for the q81 column and then report the distributions and report the mean, median and standard deviation. The median is probably more useful here as some students watch no, very little, or quite a lot of TV. Those outliers may skew the interpretation of the results.