Final Project

Nora Quick

2021-12-06

Task 1: Simulation Study

```
yrbss_2007 <- readRDS("yrbss_2007.rds")
yrbss_2017 <- readRDS("yrbss_2017.rds")
```

a) Using repeated samples of size n = 10, 100, and 1000 from the bmi variable, describe the sampling distribution of the *sample mean* of BMI in 2017. Include at least one plot to help describe your results. Report the means and standard deviations of the sampling distributions, and describe how they change with increasing sample size

```
#BMI for both years
bmi_2017 <- yrbss_2017$bmi
bmi_2007 <- yrbss_2007$bmi
```

```
#Repeated means found
m_10 <- replicate(1000, mean(sample(x = bmi_2017, size = 10)))
m_100 <- replicate(1000, mean(sample(x = bmi_2017, size = 100)))
m_1000 <- replicate(1000, mean(sample(x = bmi_2017, size = 1000)))
```

```
#Find average of replicated means
mean_10 <- mean(m_10)
mean_100 <- mean(m_100)
mean_1000 <- mean(m_1000)

#Print out
mean_10
```

```
## [1] 23.66819
```

```
mean_100
```

```
## [1] 23.62429
```

```
mean_1000
```

```
## [1] 23.61671
```

```r
#Repeated standard deviations found
s_10 <- replicate(1000, sd(sample(x = bmi_2017, size = 10)))
s_100 <- replicate(1000, sd(sample(x = bmi_2017, size = 100)))
s_1000 <- replicate(1000, sd(sample(x = bmi_2017, size = 1000)))
```

```r
#Find the average for replicated standard deviations
sd_10 <- mean(s_10)
sd_100 <- mean(s_100)
sd_1000 <- mean(s_1000)

#Print out
sd_10
```

```
## [1] 4.911743
```

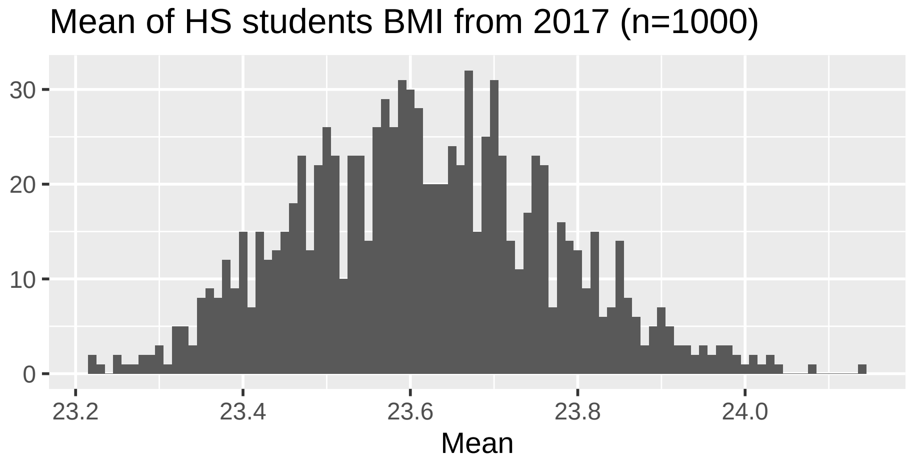
```r
sd_100
```

```
## [1] 5.157296
```

```r
sd_1000
```

```
## [1] 5.210257
```

```r
qplot(m_1000, binwidth = 0.01) +
  labs(
    title = "Mean of HS students BMI from 2017 (n=1000)",
    x = "Mean"
  )
```



- The means for the BMI of high school students in 2017 are as follows (n=10, 100, 1000) 23.53, 23.63, 23.61.
- The standard deviations for the BMI of high school students in 2017 are as follows (n=10, 100, 1000) 4.90, 5.17, 5.20.
- As the sample size increases the mean narrows down to reach closer to 23.61 and the standard deviation increases to 5.20. From the graph we can see that this causes the distribution to become more normal around 23.61.

b) Repeat the simulation in part (a), but this time use the *25th percentile* as the sample statistic. In R, `quantile(x, prob = 0.25)` will give you the 25th percentile of the values in `x`.

```
#Repeated percentiles found
p_10 <- replicate(1000, quantile(sample(x = bmi_2017, size = 10), prob = 0.25))
p_100 <- replicate(1000, quantile(sample(x = bmi_2017, size = 100), prob = 0.25))
p_1000 <- replicate(1000, quantile(sample(x = bmi_2017, size = 1000), prob = 0.25))
```

```
#Find the average perfentiles
percentile_10 <- mean(p_10)
percentile_100 <- mean(p_100)
percentile_1000 <- mean(p_1000)

#Print out
percentile_10
```

```
## [1] 20.47977
```

```
percentile_100
```

```
## [1] 20.13781
```

```
percentile_1000
```

```
## [1] 20.11304
```

c) Repeat the simulation in part (a), but this time use the *sample minimum* as the sample statistic.

```
#Repeated minimums found
min_10 <- replicate(1000, min(sample(x = bmi_2017, size = 10)))
min_100 <- replicate(1000, min(sample(x = bmi_2017, size = 100)))
min_1000 <- replicate(1000, min(sample(x = bmi_2017, size = 1000)))
```

```
#Average of the replicated minimums
minimum_10 <- mean(min_10)
minimum_100 <- mean(min_100)
minimum_1000 <- mean(min_1000)

#Print out
minimum_10
```

```
## [1] 17.91556
```

```
minimum_100
```

```
## [1] 15.60273
```

```
minimum_1000
```

```
## [1] 13.66876
```

d) Describe the sampling distribution of the *difference in the sample median BMI between 2017 and 2007*, by using repeated samples of size n_1 = 5, n_2 = 5, n_1 = 10, n_2 = 10 and n_1 = 100, n_2 = 100. Report the means and standard deviations of the sampling distributions, and describe how they change with the different sample sizes.

```r
#2017 medians
med_n15 <- replicate(1000, median(sample(x = bmi_2017, size = 5)))
med_n110 <- replicate(1000, median(sample(x = bmi_2017, size = 10)))
med_n1100 <- replicate(1000, median(sample(x = bmi_2017, size = 100)))

#2007 medians
med_n25 <- replicate(1000, median(sample(x = bmi_2007, size = 5)))
med_n210 <- replicate(1000, median(sample(x = bmi_2007, size = 10)))
med_n2100 <- replicate(1000, median(sample(x = bmi_2007, size = 100)))
```

```r
#Average of medians (2017)
median_n15 <- mean(med_n15)
median_n110 <- mean(med_n110)
median_n1100 <- mean(med_n1100)
#2017 print out
median_n15
```

```
## [1] 22.67422
```

```r
median_n110
```

```
## [1] 22.539
```

```r
median_n1100
```

```
## [1] 22.43902
```

```r
#Average of medians (2007)
median_n25 <- mean(med_n25)
median_n210 <- mean(med_n210)
median_n2100 <- mean(med_n2100)
#2007 print out
median_n25
```

```
## [1] 22.97134
```

```r
median_n210
```

```
## [1] 22.67843
```

```r
median_n2100
```

```
## [1] 22.61507
```

```
diff_2017 <- (median_n15 + median_n110 + median_n1100) / 3
diff_2007 <- (median_n25 + median_n210 + median_n2100) / 3

#Difference in medians
diff <- diff_2017 - diff_2007

#Print out
diff_2017
```

```
## [1] 22.55075
```

```
diff_2007
```

```
## [1] 22.75495
```

```
#Print out difference
diff
```

```
## [1] -0.2041979
```

- The difference between the medians of 2017 and 2007 is about 0.13. It appears that the median BMI for high schoolers between 2017 and 2007 has lowered about 0.13.

```
#2017 means
m17_5 <- replicate(1000, mean(sample(x = bmi_2017, size = 5)))
m17_10 <- replicate(1000, mean(sample(x = bmi_2017, size = 10)))
m17_100 <- replicate(1000, mean(sample(x = bmi_2017, size = 100)))

#2007 means
m07_5 <- replicate(1000, mean(sample(x = bmi_2007, size = 5)))
m07_10 <- replicate(1000, mean(sample(x = bmi_2007, size = 10)))
m07_100 <- replicate(1000, mean(sample(x = bmi_2007, size = 100)))
```

```
#2017 average of means
mean17_5 <- mean(m17_5)
mean17_10 <- mean(m17_10)
mean17_100 <- mean(m17_100)

#Print out
mean17_5
```

```
## [1] 23.54801
```

```
mean17_10
```

```
## [1] 23.71391
```

```
mean17_100
```

```
## [1] 23.61867
```

```r
#2007 average of means
mean07_5 <- mean(m07_5)
mean07_10 <- mean(m07_10)
mean07_100 <- mean(m07_100)

#Print out
mean07_5
```

```
## [1] 23.75101
```

```r
mean07_10
```

```
## [1] 23.77004
```

```r
mean07_100
```

```
## [1] 23.78493
```

```r
#2017 standard deviations
s17_5 <- replicate(1000, sd(sample(x = bmi_2017, size = 5)))
s17_10 <- replicate(1000, sd(sample(x = bmi_2017, size = 10)))
s17_100 <- replicate(1000, sd(sample(x = bmi_2017, size = 100)))

#2007 standard deviations
s07_5 <- replicate(1000, sd(sample(x = bmi_2007, size = 5)))
s07_10 <- replicate(1000, sd(sample(x = bmi_2007, size = 10)))
s07_100 <- replicate(1000, sd(sample(x = bmi_2007, size = 100)))
```

```r
#2017 average standard deviations
sd17_5 <- mean(s17_5)
sd17_10 <- mean(s17_10)
sd17_100 <- mean(s17_100)

#Print out (2017)
sd17_5
```

```
## [1] 4.693042
```

```r
sd17_10
```

```
## [1] 4.839601
```

```r
sd17_100
```

```
## [1] 5.157175
```

```r
#2007 average standard deviations
sd07_5 <- mean(s07_5)
sd07_10 <- mean(s07_10)
sd07_100 <- mean(s07_100)

#Print out (2007)
sd07_5
```

```
## [1] 4.464196
```

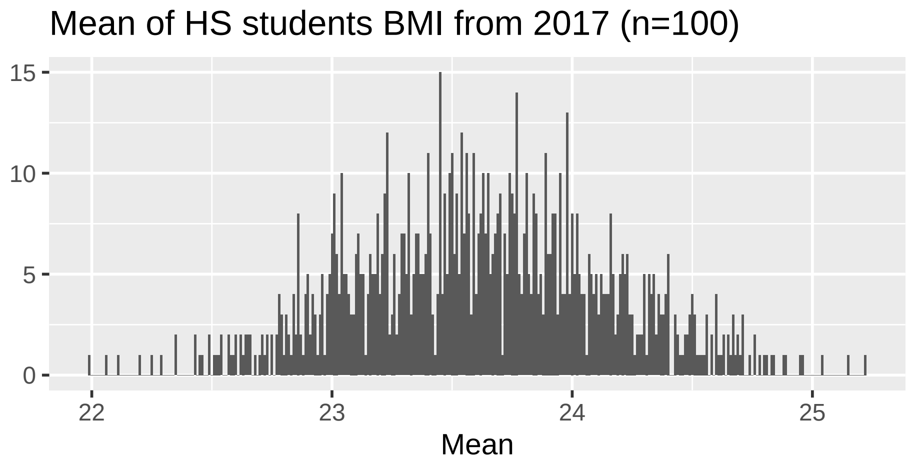
```
sd07_10
```

```
## [1] 4.665349
```

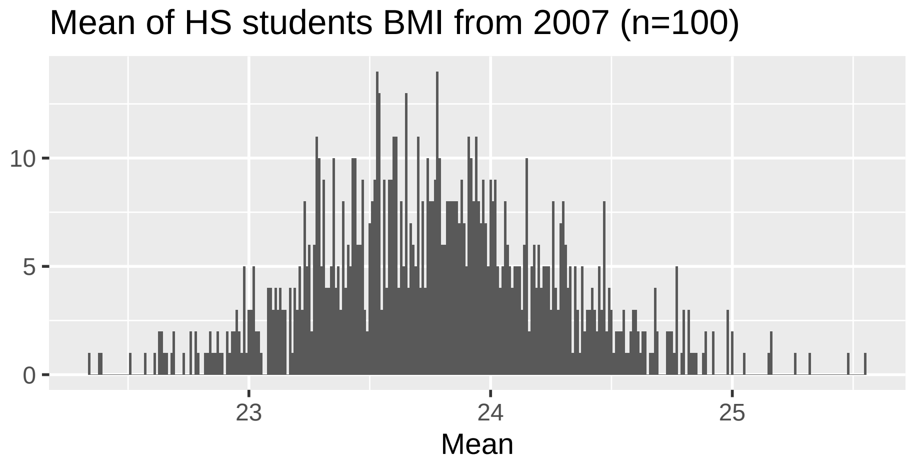
```
sd07_100
```

```
## [1] 4.9002
```

```
qplot(m17_100, binwidth = 0.01) +
  labs(
    title = "Mean of HS students BMI from 2017 (n=100)",
    x = "Mean"
  )
```



```
qplot(m07_100, binwidth = 0.01) +
  labs(
    title = "Mean of HS students BMI from 2007 (n=100)",
    x = "Mean"
  )
```



- The mean for the BMI of high school students in 2017 as the sample size grows larger gets closer to 23.6 and the standard deviation grows closer to 5.2.
- The mean for the BMI of high school students in 2007 as the sample size grows larger gets closer to 23.8 and the standard deviation grows closer to 4.9.
- Both hovered around the same values getting more accurate as the sample size got bigger. Additionally both had their sample sizes grow to a more accurate number as well. They both start to distribute out more normally as the sample size grows each having their median around 22.6/22.7. It's important to note that the difference in median that we found earlier (0.1) can also be seen in the averages of the means up above. This means that our data grows more accurate to the difference in median we found.

e) *Summarize your results.*

- The BMI of high schoolers in 2017 becomes a more normal shape as the sample size increases. It begins to center more closely around the mean and the spread becomes more narrow as the data collects near the center. We can see a visual of n=100 above showing a increasingly normal shape.
- The BMI of the high schoolers in 2007 also becomes more normal shape as the sample size increases. The speard becomes smaller as the data approaches the center. We can see a visual of n=100 above showing the increasing normal shape.

Task 2: Data Analysis

1) How has the BMI of high-school students changed between 2007 and 2017? Are high-schoolers getting more overweight?

- The BMI decreased between 2007 and 2017. We can see this with both the mean and median. The mean decreasing from 23.8 to 23.6 and the median going from 22.8 to 22.6. From this we can determine that high schoolers are not becoming more overweight. In fact, we can determine that they are slightly fitter in 2017 than they were in 2007.

2) In 2017, are 12th graders more or less likely than 9th graders to be "physically active at least 60 minutes per day on 5 or more days"?

```
qn79_true <- yrbss_2017[,"qn79"] == "TRUE"
sum(qn79_true, na.rm = TRUE)
```

```
## [1] 5677
```

```
twelve <- sum((!is.na(yrbss_2017[,"qn79"] == "TRUE")) && (!is.na(yrbss_2017[,"grade"] == "12th")))
twelve
```

```
## [1] 1
```

```
nine <- (!is.na(yrbss_2017[,"qn79"] == "TRUE")) && (!is.na(yrbss_2017[,"grade"] == "9th"))
sum(nine)
```

```
## [1] 1
```

- 12th graders do not excercise more or less than the 9th graders. It appears that they excercise the same amount.

3) How much sleep do highschoolers get?

```
#Sum of hours slept 2017
q17_88 <- yrbss_2017$q88
new17_q88 <- as.numeric(q17_88)
mean(new17_q88, na.rm = TRUE)
```

```
## [1] 3.549185
```

```
#Sum of hours slept 2007
q7_88 <- yrbss_2007$q88
new7_q88 <- as.numeric(q7_88)
mean(new7_q88, na.rm = TRUE)
```

```
## [1] 3.798568
```

```
#Average of both years
avg <- (mean(new17_q88, na.rm = TRUE) + mean(new7_q88, na.rm = TRUE)) / 2
avg
```

```
## [1] 3.673876
```

- In the year of 2017 high schoolers get about 3.55 hours of sleep. In the year of 2007 high schoolers get about 3.8 hours of sleep.
- The average amount of sleep for both year combined is about 3.67 hours of sleep.