Evans_Final_Project

Maryanne Evans

12/5/2021

#==============================Task 1: Simulation Study========================== For this part of the project you will need to perform a simulation study to investigate the properties of four sample statistics: the mean, the 25th percentile, the minimum, and a difference in medians. #Import Libraries

```
library(ggplot2)
```

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 Variable Sample Mean Analysis (Distribution) for 2017
#clean up the data, checking for and removing NA

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

sum(is.na(yrbss_2007[, "bmi"]))
```

```
## [1] 0
```

```
sum(is.na(yrbss_2017[, "bmi"]))
```

```
## [1] 0
```

```
#Focus on BMI Columns Only, Sub-setting
BMI07 <- yrbss_2017$bmi
BMI17 <- yrbss_2017$bmi

#View BMI Subsets
head(BMI07)
```

```
## [1] 24.2666 24.8047 24.5890 22.6338 21.7930 32.5062
```

```
head(BMI17)
```

```
## [1] 24.2666 24.8047 24.5890 22.6338 21.7930 32.5062
```

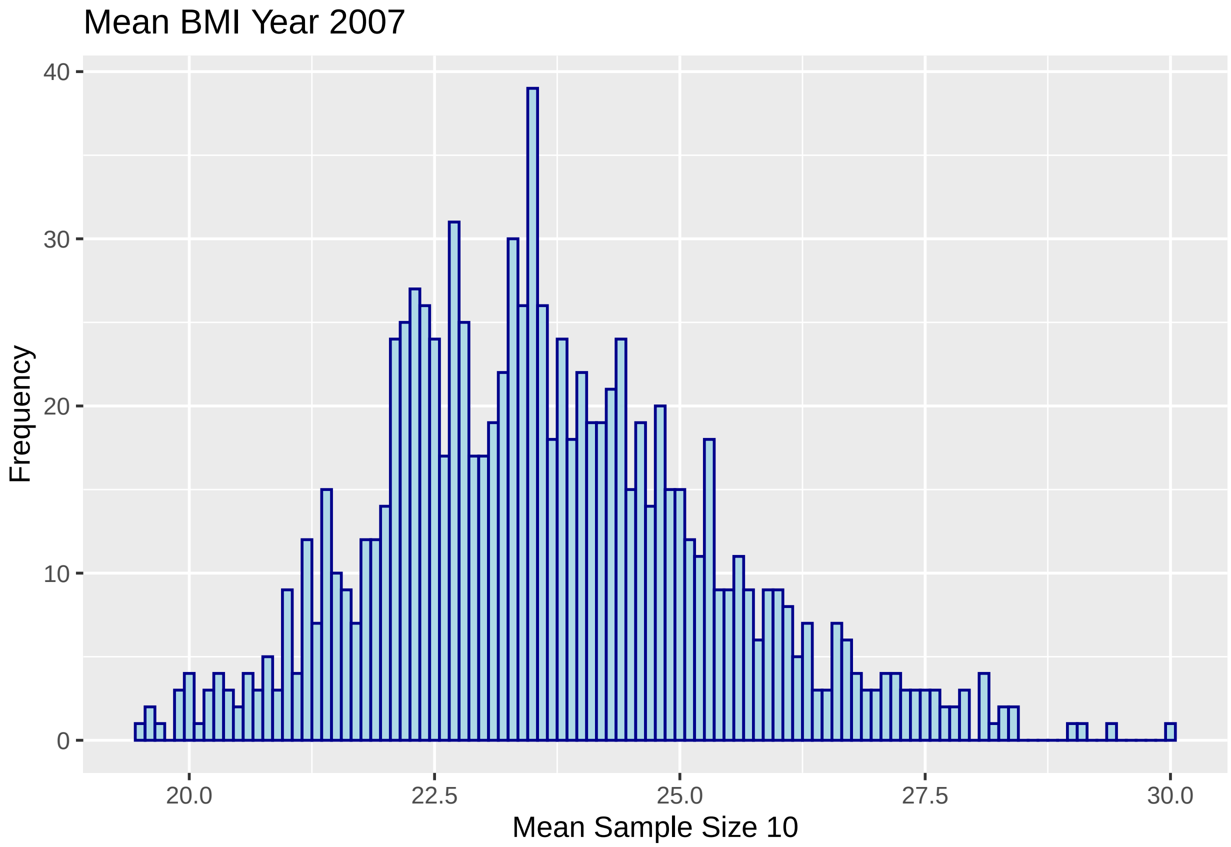
```
#=======Exploring the Data Distributions and Test Stats (mean and sd)==============

#create the SIMULATION
get_means <- function(n, n_sim=1000){
  replicate(n_sim, mean(sample(yrbss_2017$bmi, size = n, replace = TRUE)))
}

#Apply to each sample size, exploring the data
means <- lapply(c(10, 100, 1000), get_means)

#========SAMPLE n = 10=======
sample_sd_10 <- sd(means[[1]])
sample_mean_10 <- means[[1]]

#===TEST STATS====
#mean = 23.62922
sample_mean_10 <- means[[1]]
mean(sample_mean_10)
```

```
## [1] 23.65262
```

```
#standard deviation = 1.575878
sample_sd_10 <- sd(means[[1]])
sample_sd_10
```

```
## [1] 1.702788
```

```
#====HISTO for N = 10 ====
sample_plot_10 <- qplot(sample_mean_10, binwidth = 0.1, main = "Mean BMI Year 2007", xlab = "Mean Sampl
sample_plot_10
```



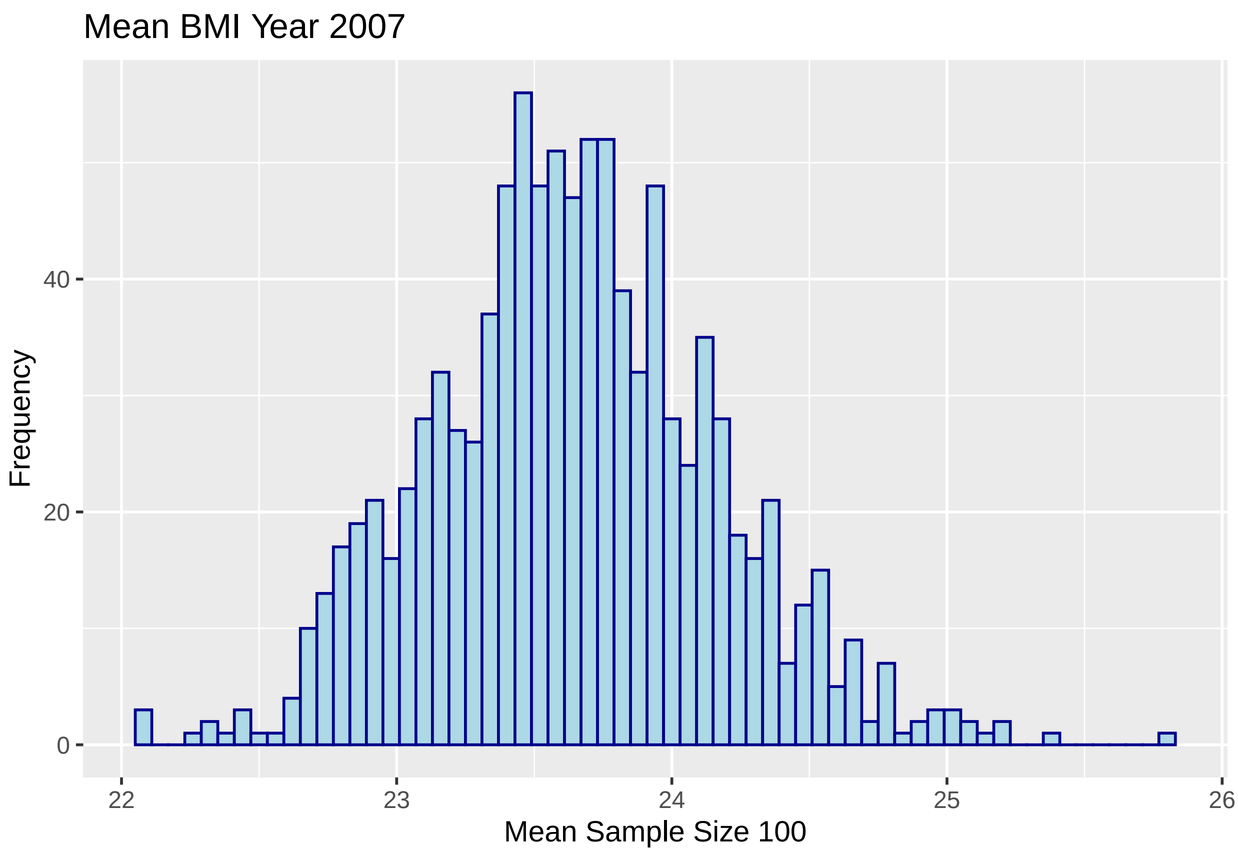
```
#========SAMPLE n = 100=======
sample_sd_100 <- sd(means[[2]])
sample_mean_100 <- means[[2]]

#===TEST STATS (SD and MEAN)====
#mean = 23.62378
sample_mean_100 <- means[[2]]
mean(sample_mean_100)
```

```
## [1] 23.64182
```

```
#standard deviation =  0.507905
sample_sd_100 <- sd(means[[2]])
sample_sd_100
```

```
## [1] 0.5182052
```

```
#====HISTO for N = 100 ====
sample_plot_100 <- qplot(sample_mean_100, binwidth = 0.06, main = "Mean BMI Year 2007", xlab = "Mean Sa
sample_plot_100
```



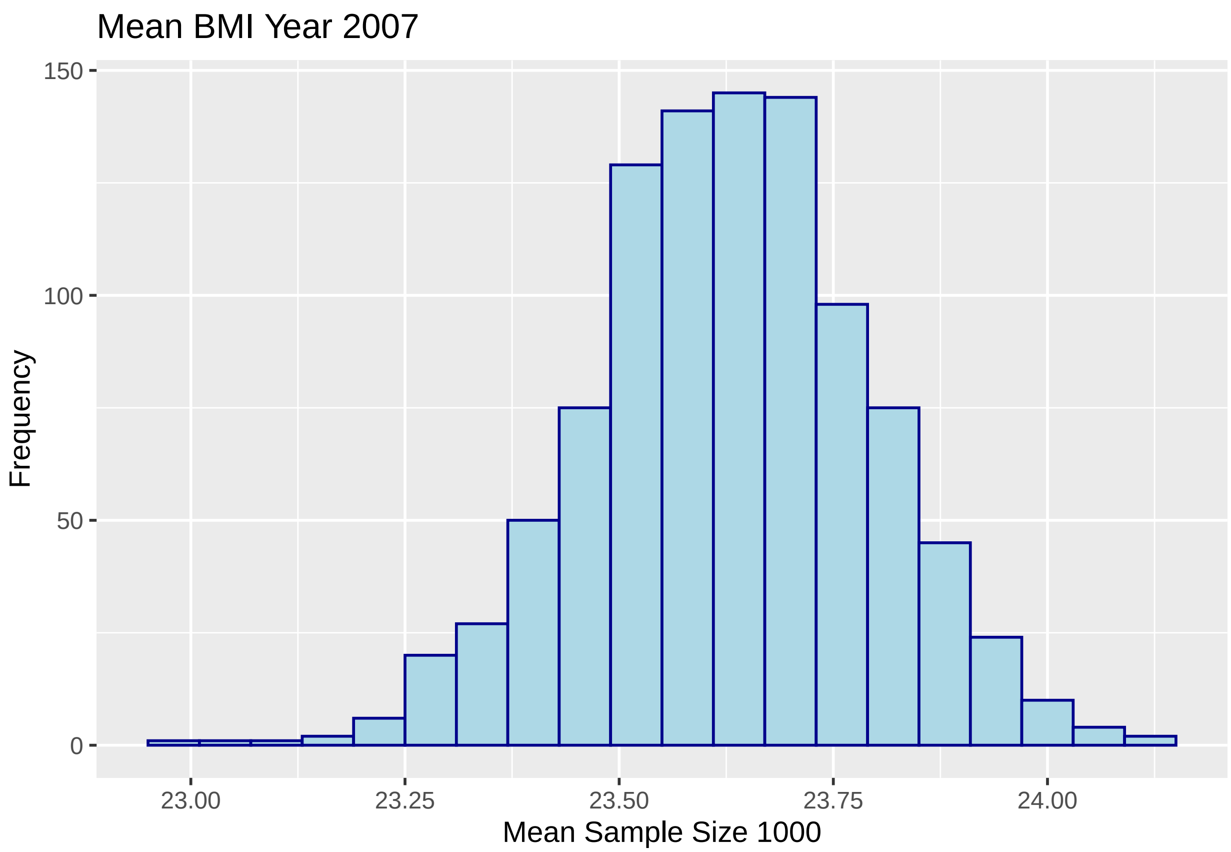
```
#========SAMPLE n = 1000=======
sample_sd_1000 <- sd(means[[3]])
sample_mean_1000 <- means[[3]]

#===TEST STATS (SD and MEAN)====
#mean = 23.62378
sample_mean_1000 <- means[[3]]
mean(sample_mean_1000)
```

```
## [1] 23.62776
```

```
#standard deviation = 0.1632048
sample_sd_1000 <- sd(means[[3]])
sample_sd_1000
```

```
## [1] 0.162872
```

```
#====HISTO for N = 100 ====
sample_plot_1000 <- qplot(sample_mean_1000, binwidth = 0.06, main = "Mean BMI Year 2007", xlab = "Mean

sample_plot_1000
```



```
#Tentative Explanation with increasing sample size: the data appears to become more normal with increas
```

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.

```
#=======Exploring the Data Distributions and Test Stats (mean and SD of 25th percentile)==============

#subset the data to the 25th percentile, to use as sample stat
#create the SIMULATION for the 25th percentile
get_25 <- function(n, n_sim=5000){
  replicate(n_sim, quantile(sample(yrbss_2017$bmi, size = n, replace = TRUE), prob = 0.25))
}
#Apply to each sample size, exploring the data
get_25_all <- lapply(c(10, 100, 1000), get_25)

#N = 10
sample_10 <- get_25_all[[1]]
#TEST STATS
mean(sample_10)
```

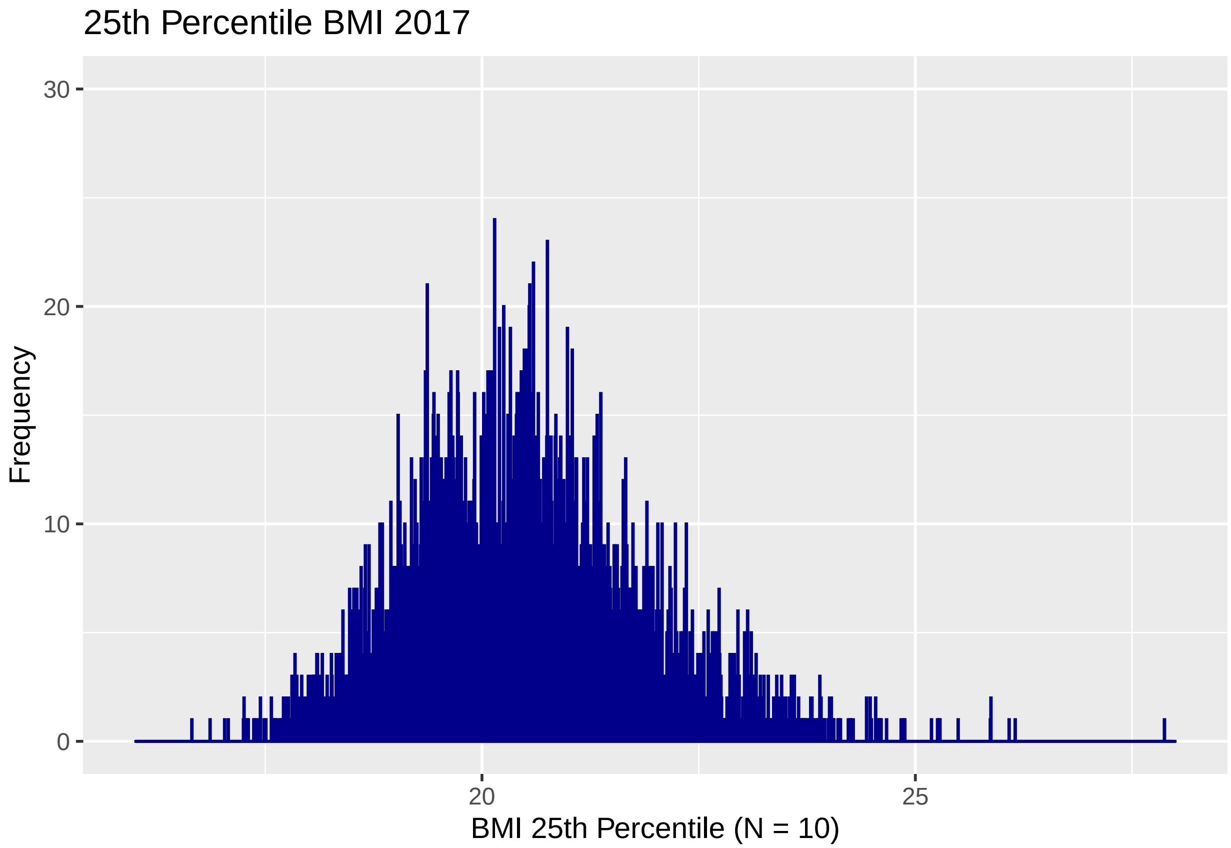
```
## [1] 20.47983
```

```
sd(sample_10)
```

```
## [1] 1.28584
```

```
#VIZ
sample_plot_10 <-  qplot(sample_10, binwidth = 0.007, main = "25th Percentile BMI 2017", xlab = "BMI 25
sample_plot_10
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
#N = 100
sample_100 <- get_25_all[[2]]
#TEST STATS
mean(sample_100)
```

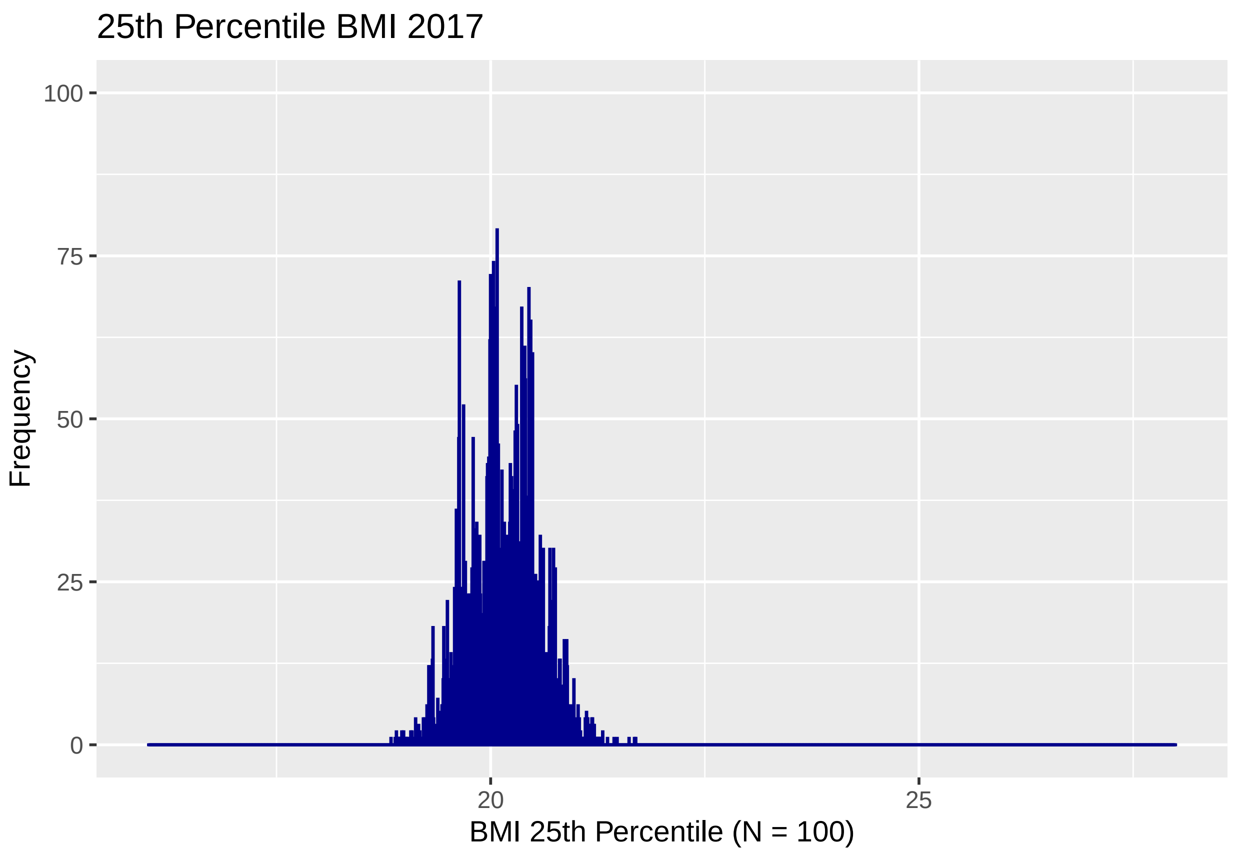
```
## [1] 20.14148
```

```
sd(sample_100)
```

```
## [1] 0.4067081
```

```
#VIZ
sample_plot_100 <- qplot(sample_100, binwidth = 0.007, main = "25th Percentile BMI 2017", xlab = "BMI 2
sample_plot_100
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
#N = 1000
sample_1000 <- get_25_all[[3]]
#TEST STAT
mean(sample_1000)
```

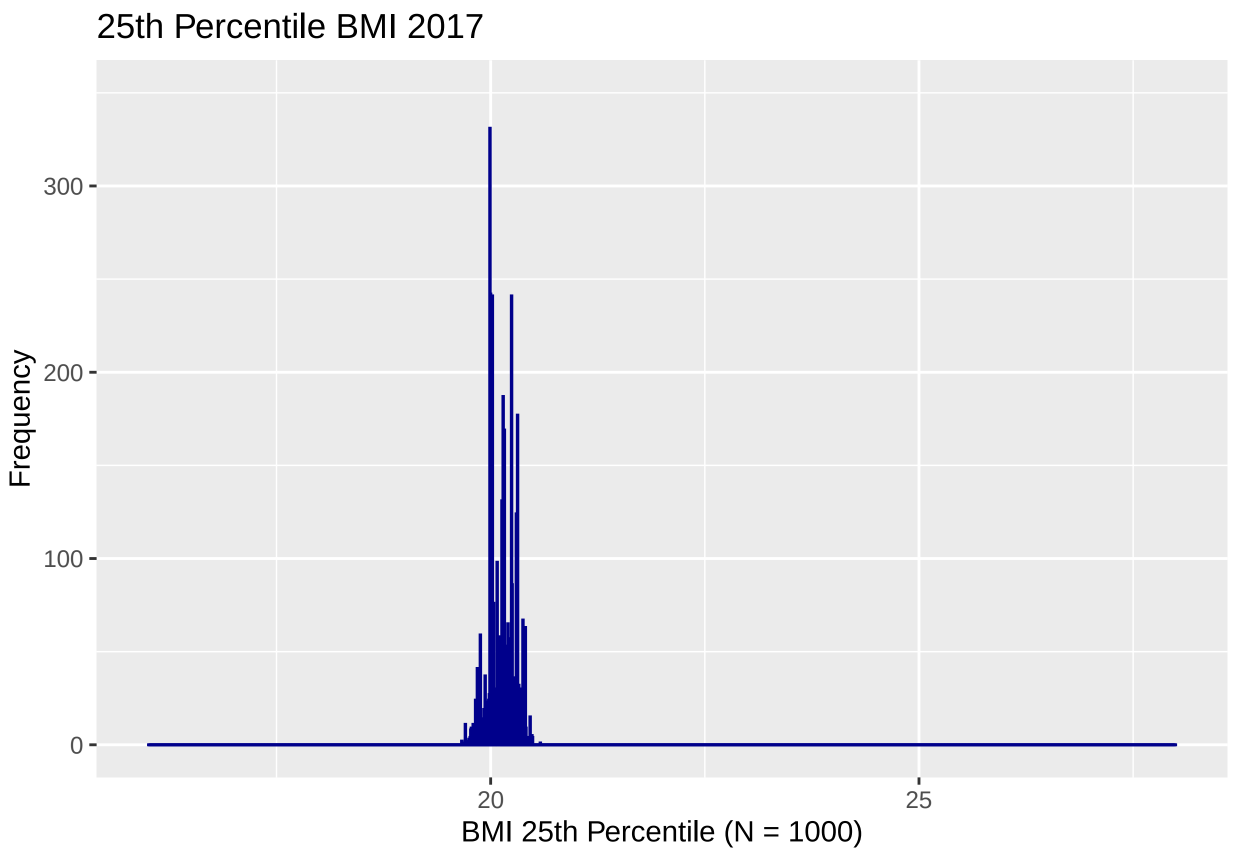
```
## [1] 20.10834
```

```
sd(sample_1000)
```

```
## [1] 0.1380338
```

```
#VIZ
sample_plot_1000 <- qplot(sample_1000, binwidth = 0.007, main = "25th Percentile BMI 2017", xlab = "BMI
sample_plot_1000
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



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

```
#BMI Variable Sample Minimum Analysis (Distribution) for 2017
get_min <- function(n, n_sim=1000){
  replicate(n_sim, min(sample(yrbss_2017$bmi, size = n, replace = TRUE)))
}

mins <- lapply(c(10, 100, 1000), get_min)

#=======Defined for Sample N = 10==========
sample_min_10 <- mins[[1]]
head(sample_min_10)
```

```
## [1] 14.8828 16.6728 17.8945 18.7813 18.4388 21.4414
```

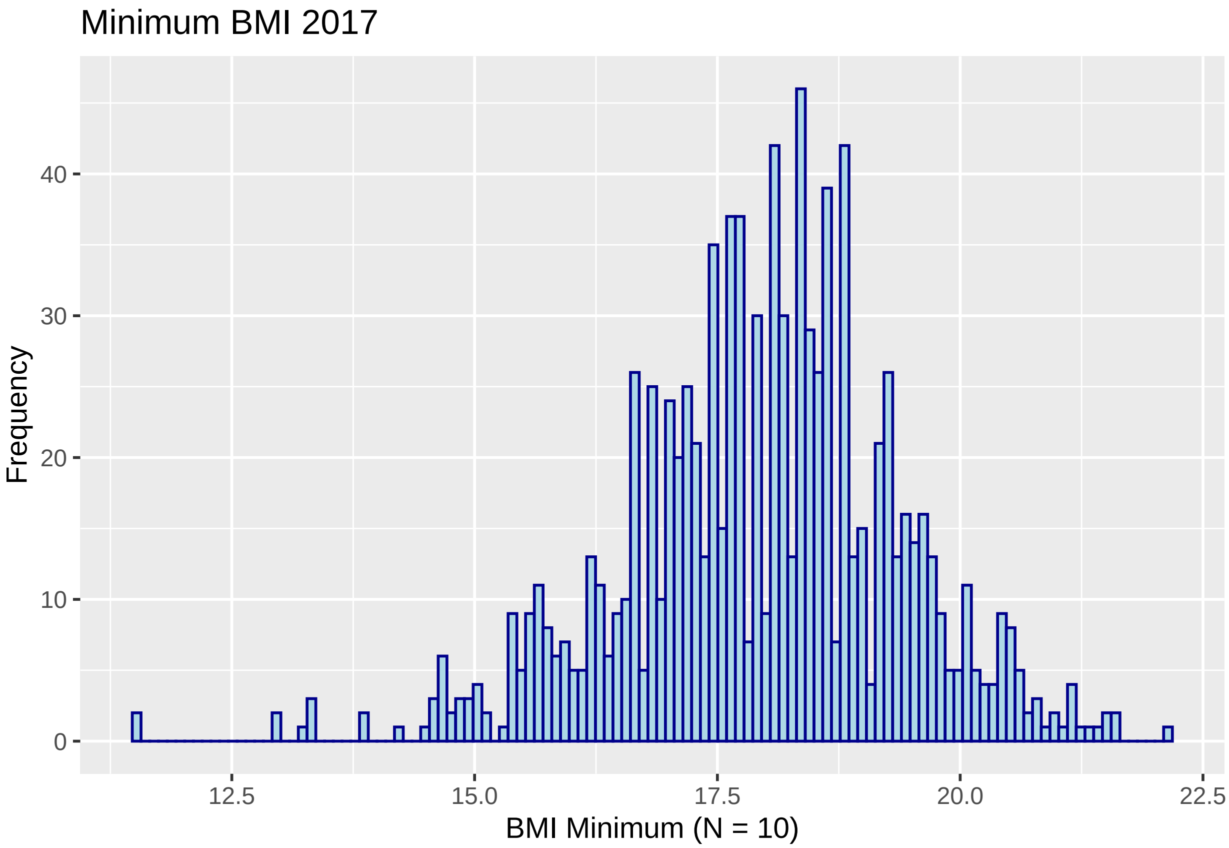
```
#===TEST STATS====
#mean =  17.90709
sample_min_10 <- mins[[1]]
mean(sample_min_10)
```

```
## [1] 17.94503
```

```
#standard deviation = 1.425519
sample_sd_10 <- sd(mins[[1]])
sample_sd_10
```

```
## [1] 1.434259
```

```
#====PLOT for N = 10 ====
sample_plot_10 <- qplot(sample_min_10, binwidth = 0.09, main = "Minimum BMI 2017", xlab = "BMI Minimum
sample_plot_10
```



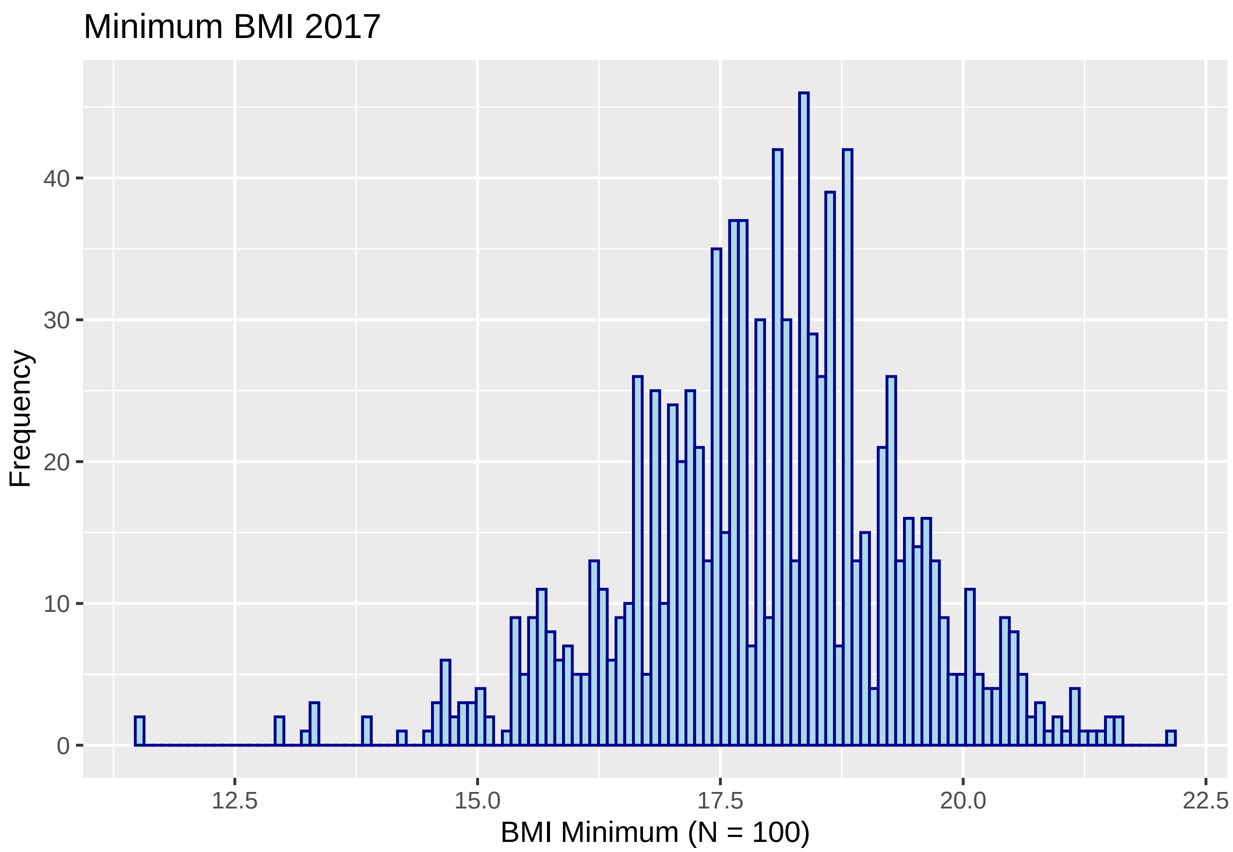
```
#=======Defined for Sample N = 100==========
sample_min_100 <- mins[[2]]
#===TEST STATS====
sample_min_100 <- mins[[2]]
mean(sample_min_100)
```

```
## [1] 15.60552
```

```
sample_sd_100 <- sd(mins[[2]])
sample_sd_100
```

```
## [1] 1.077644
```

```
#N = 100
sample_plot_100 <- qplot(sample_min_10, binwidth = 0.09, main = "Minimum BMI 2017", xlab = "BMI Minimum
sample_plot_100
```



```
#=======Defined for Sample N = 1000==========
sample_min_1000 <- mins[[3]]
head(sample_min_1000)
```

```
## [1] 15.0328 12.9954 14.4629 11.5461 14.4394 14.3516
```

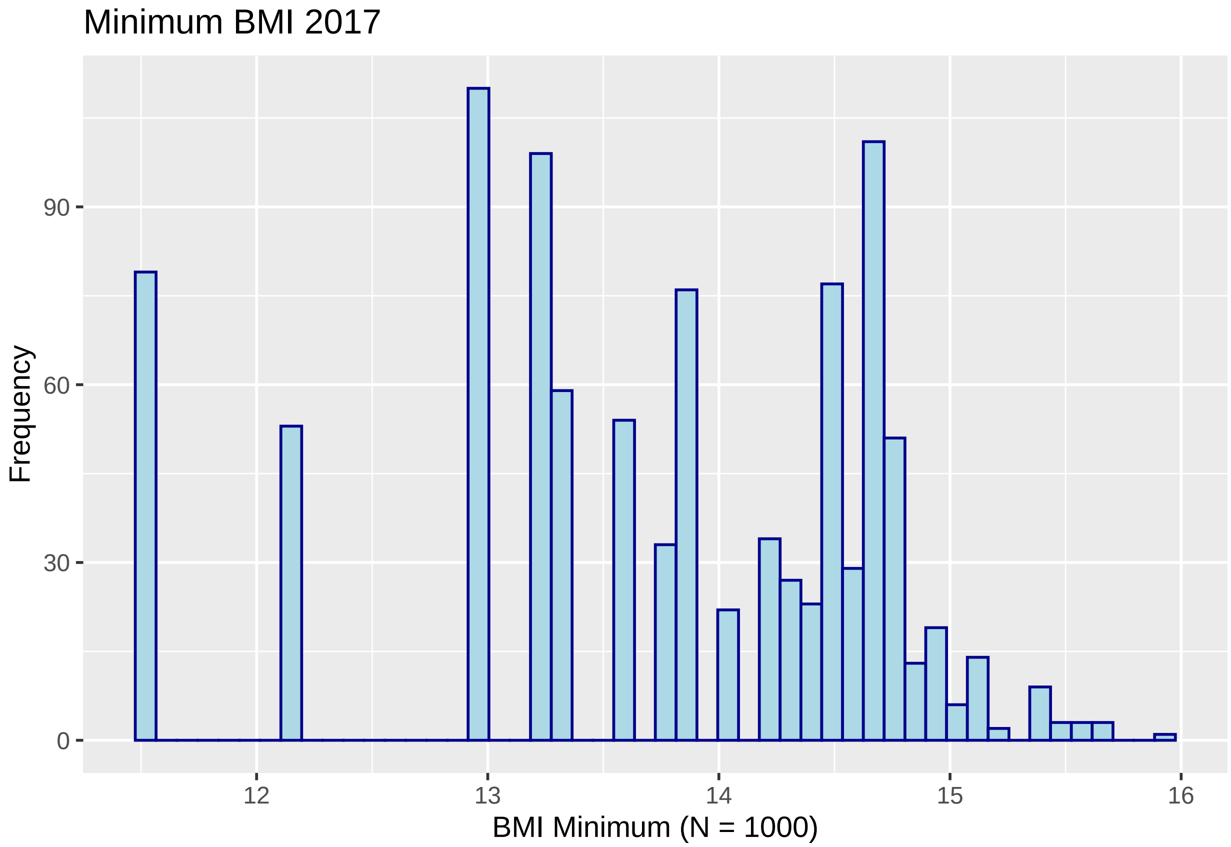
```
#===TEST STATS====
#mean = 13.7178
sample_min_1000 <- mins[[3]]
mean(sample_min_1000)
```

```
## [1] 13.7092
```

```
#standard deviation = 0.9702329
sample_sd_1000 <- sd(mins[[3]])
sample_sd_1000
```

```
## [1] 0.9932978
```

```
#====PLOT for N = 100 ====
sample_plot_1000 <- qplot(sample_min_1000, binwidth = 0.09, main = "Minimum BMI 2017", xlab = "BMI Mini
sample_plot_1000
```



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.

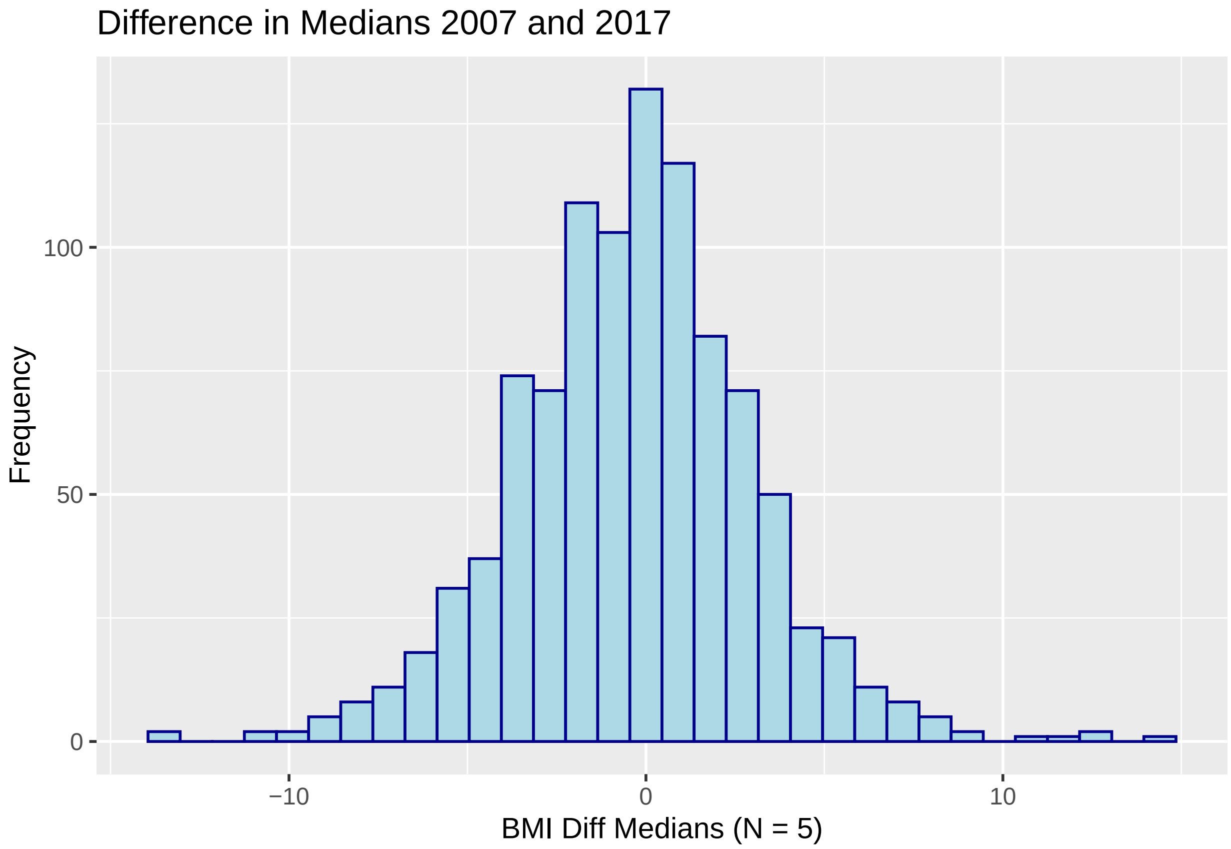
```
#SIMULATION for 2017 and 2007
#sample size = 5
median_bmi_5 <- replicate(1000, median(sample(yrbss_2017$bmi, size = 5) - median(sample(yrbss_2007$bmi,
mean(median_bmi_5)
```

```
## [1] -0.3710798
```

```
sd(median_bmi_5)
```

```
## [1] 3.336651
```

```
qplot(median_bmi_5, binwidth = 0.9, main = "Difference in Medians 2007 and 2017", xlab = "BMI Diff Medi
```



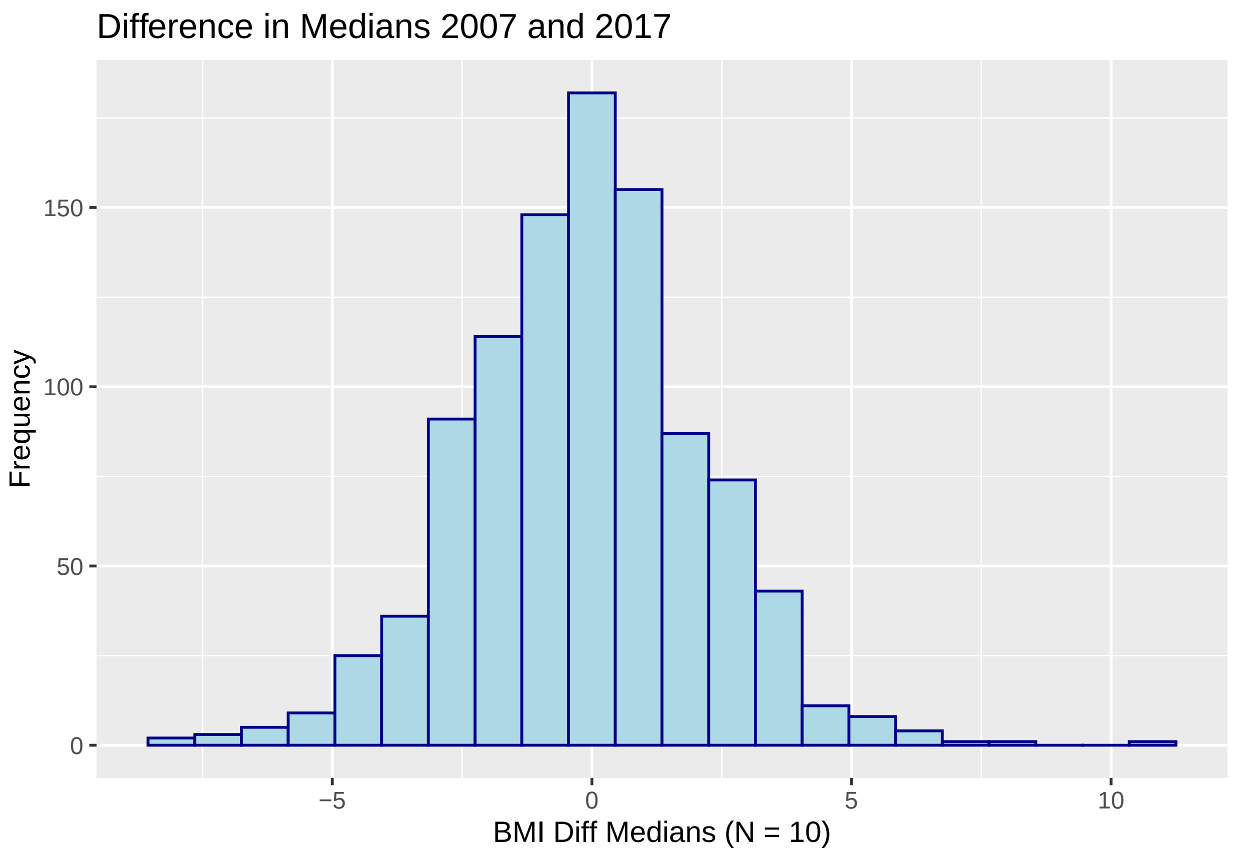
```
#sample size = 10
median_bmi_10 <- replicate(1000, median(sample(yrbss_2017$bmi, size = 10) - median(sample(yrbss_2007$bmi
mean(median_bmi_10)
```

```
## [1] -0.1375372
```

```
sd(median_bmi_10)
```

```
## [1] 2.257955
```

```
qplot(median_bmi_10, binwidth = 0.9, main = "Difference in Medians 2007 and 2017", xlab = "BMI Diff Medi
```



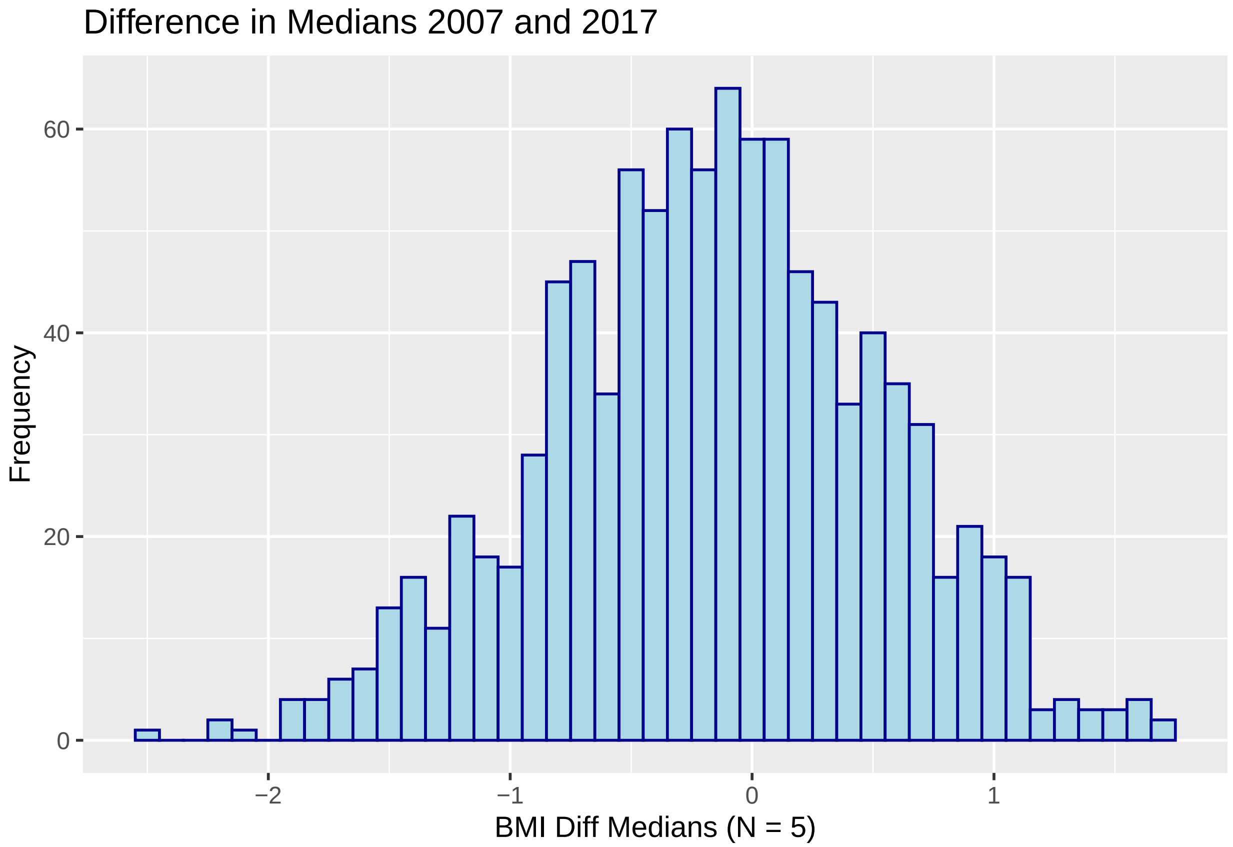
```
#sample size = 100
median_bmi_100 <- replicate(1000, median(sample(yrbss_2017$bmi, size = 100) - median(sample(yrbss_2007$b
#stats
mean(median_bmi_100)
```

```
## [1] -0.1636746
```

```
sd(median_bmi_100)
```

```
## [1] 0.6895322
```

```
#vis
qplot(median_bmi_100, binwidth = 0.1, main = "Difference in Medians 2007 and 2017", xlab = "BMI Diff Me
```



#============================Summarize your Results============================

Make sure you comment on the center, spread and shape for the sampling distribution of each statistic as the sample size increases.

Tentative Summary of Results:

Means: The means respond accordingly to the Central Limit Theorem, and as sample size increases, the parameter centralizes around the true value. The histograms for this data display normalized data, though that can be tested to clarify accuracy.

25th Quantile: The standard deviation did not vary greatly, and the mean was relatively stable. I am wondering if my coding is incorrect given the consistency in values for sd and mean? The distribution is skewed in all cases, despite sample size increasing.

Difference in Medians: My values were all over the place, and I know this portion of my analysis requires more work. Visually, it may be more explanatory to plot both years in different colors, and visually display overlapping values, and demonstrate the variation and differences in median values.

Mins: Minimum doesn't tell you much, and does not appear normal, and looks less and less normal as sample size increases. The distribution cannot center around the minimum because there is not a possibility for getting a value lower than the true minimum. Histogram is stratified, and does not centeralize around a clear value.

#=========================TASK 2: Data Analysis===============================

For this part of your assignment your task is to analyze the data to answer the questions of interest. Your solutions must include a non-technical summary of your findings. Using the same data as the Simulation Study, but now treating the survey as a sample from the population of all USA high-school students:

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

Example R Code: t.test(year ~ bmi, data =)

For this question, we are comparing two independent populations (students in 2007 to students in 2017). I would preform a t-test, such as the Welch's two sample t-test and construct a confidence intervals for the difference in means in weight, testing the hypothesis: "Highschoolers are getting more overweight as time goes on between the years 2007 and 2017".

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"?

Associated Code:

```
#Subset Data
Active_12 <- yrbss_2017[yrbss_2017$grade == "12th", "qn79"]
Active_12
```

```
## # A tibble: 3,119 x 1
##    qn79
##    <lgl>
##  1 NA
##  2 NA
##  3 FALSE
##  4 NA
##  5 NA
##  6 NA
##  7 NA
##  8 NA
##  9 NA
## 10 NA
## # ... with 3,109 more rows
```

```
#Count Response Active 12th
True_12 <- length(which(Active_12 == TRUE))
False_12 <- length(which(Active_12 == FALSE))

True_12
```

```
## [1] 1149
```

```
False_12
```

```
## [1] 1860
```

```
#9th Grader Response
Active_9 <- yrbss_2017[yrbss_2017$grade == "9th", "qn79"]
Active_9
```

```
## # A tibble: 3,479 x 1
##    qn79
##    <lgl>
##  1 NA
##  2 NA
##  3 FALSE
##  4 FALSE
##  5 NA
##  6 TRUE
##  7 TRUE
##  8 FALSE
##  9 TRUE
## 10 NA
## # ... with 3,469 more rows
```

```
#Count Response Active 9th
True_9 <- length(which(Active_9 == TRUE))
False_9 <- length(which(Active_9 == FALSE))

True_9
```

```
## [1] 1680
```

```
False_9
```

```
## [1] 1604
```

```
#prop test for two populations

X <- c(1149, 1680)
n <- c((1149+1860), (1680+1604))

prop.test(X,n,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X out of n
## X-squared = 106.77, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1540812 -0.1053524
## sample estimates:
##    prop 1    prop 2
## 0.3818544 0.5115713
```

Given I would use a two sample prop test to address this question, because there are two different populations with a difference in proportions to investigate, with binary observations.

3) How much sleep do highschoolers get?

The data is not numerical, and therefore no test stat is needed to answer this question. Instead, I would explore the sample quartiles and compare them across the highschoolers. I would visualize these data using histograms, bucketing 25th, 50th and 75th percentiles, before trying to draw conclusions.