Data Analysis Project

James King

12Dec2020

Task 1: Simulation Study

Part A: Using repeated samples of size n = 10, 100 and 1000, describe the sampling distribution of the sample mean BMI in 2017. Include at least one plot. Report the means and standard deviations of the sampling distributions, and describe how they change with increasing sample size.

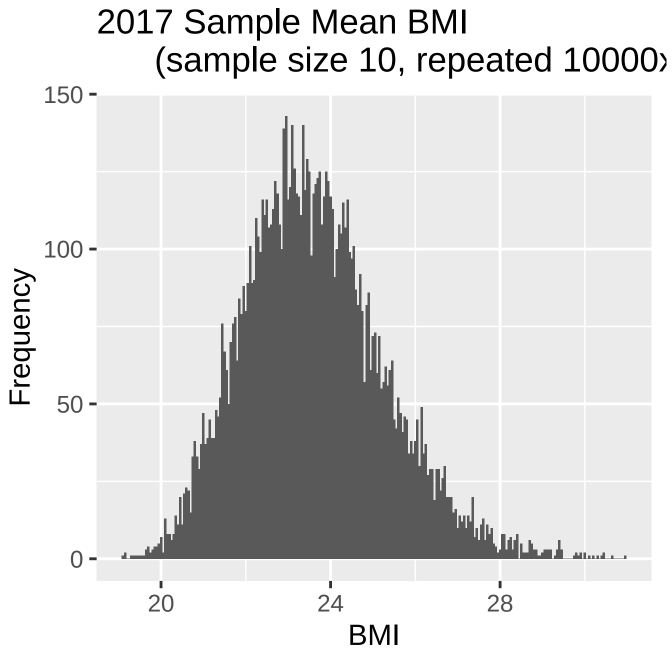
The sampling distributions of the sample mean of BMI from the 2017 data set are all normally distributed. As the sample size increases, the spread of the data decreases, and the frequency increases around the true population mean BMI. I have included each plot below. Note that the x and y axes auto scale with the data. I excluded any 'NA' values before running any of the analyses.

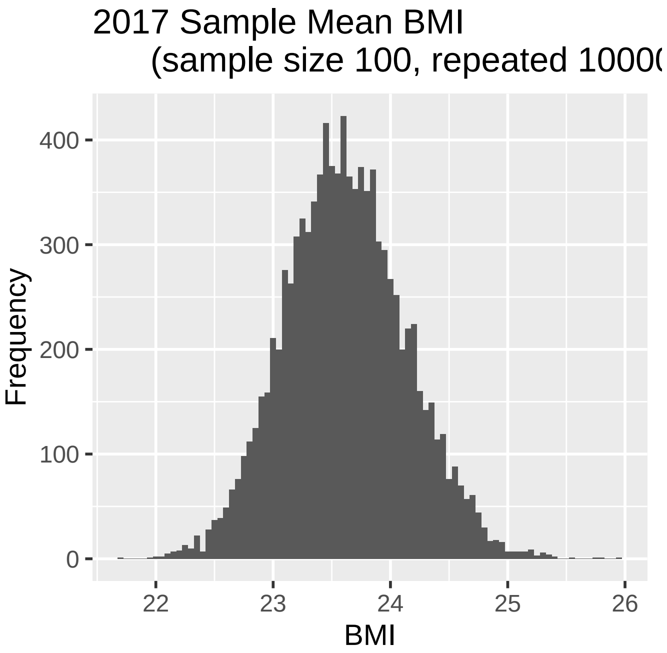
The sample means and sample standard deviations are as follows for each of the sample sizes:

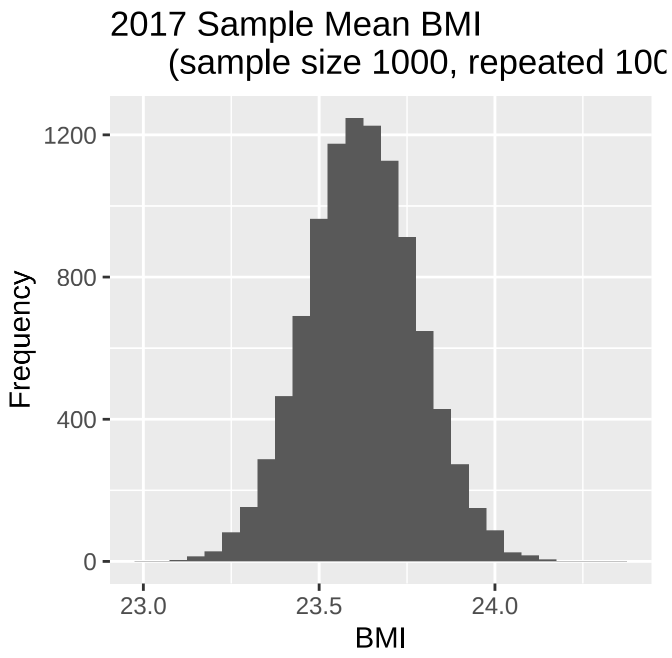
sample size = 10, sample mean = 23.656, sample SD = 1.656
sample size = 100, sample mean = 23.620, sample SD = 0.519
sample size = 1000, sample mean = 23.622, sample SD = 0.157

The sample means do not change very much with increasing sample size because I am replicating each simulation 10000 times. However, the sample standard deviation decreases as the sample size increases, as expected.

```
##        [,1]   [,2]   [,3]
## [1,] 23.617 23.617 23.622
## [2,]  1.663  0.518  0.157
```







Part B: Using repeated samples of size n = 10, 100 and 1000, describe the sampling distribution of the sample 25th percentile BMI in 2017. Include at least one plot. Report the 25th percentile and standard deviations of the sampling distributions, and describe how they change with increasing sample size.

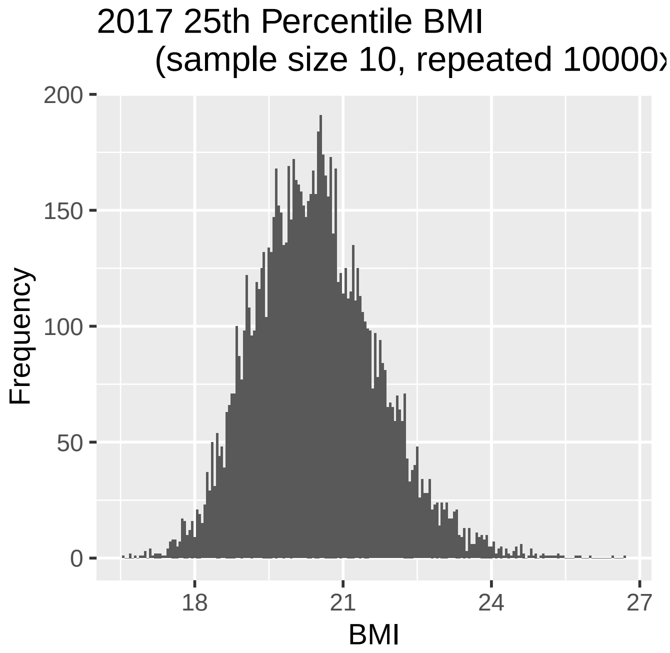
The sampling distributions of the 25th percentile BMI from the 2017 data set are not normally distributed, they have a slight skew to the right - the central limit theorem only applies to sample means. This makes sense because the majority of the values in the data set lie to the right of the 25th percentile. As the sample size increases, the spread of the data decreases, and the frequency around the true population 25th percentile BMI increases. I have included each plot below. Note that the x and y axes auto scale with the data and I excluded any ‘NA’ values before running the analyses.

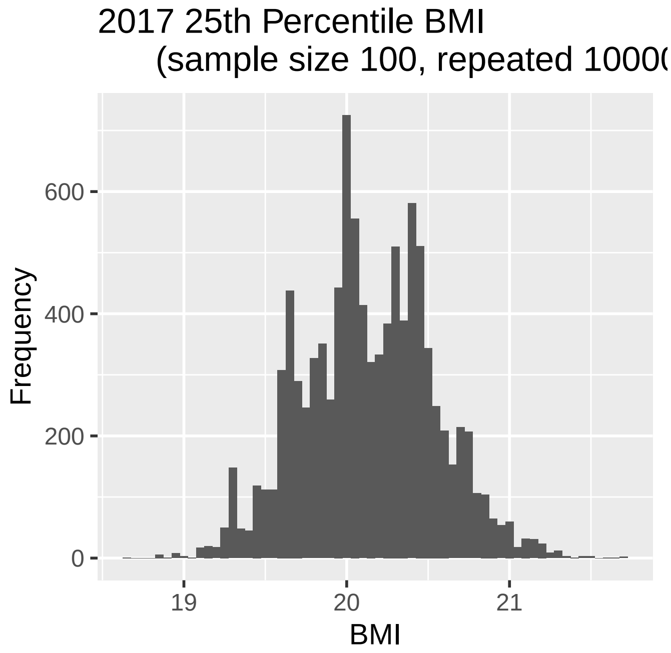
The sample 25th percentiles and sample standard deviations are as follows for each sample size:

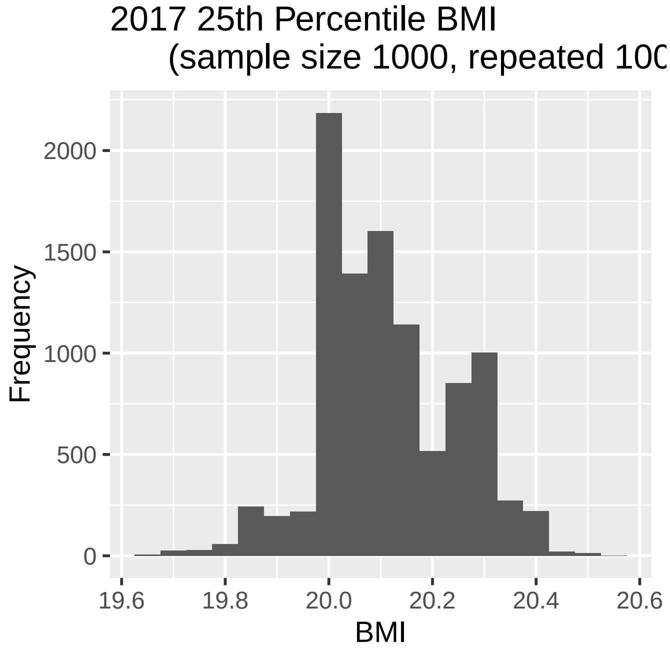
sample size = 10, sample 25th percentile = 20.461, sample SD = 1.244
sample size = 100, sample 25th percentile = 20.137, sample SD = 0.402
sample size = 1000, sample 25th percentile = 220.108, sample SD = 0.132

The sample 25th percentiles do not change very much with increasing sample size because I am replicating each simulation 10000 times. However, the sample standard deviation decreases as the sample size increases, as expected.

```
##         [,1]   [,2]   [,3]
## [1,] 20.459 20.134 20.109
## [2,]  1.254  0.406  0.134
```







Part C: Using repeated samples of size n = 10, 100 and 1000, describe the sampling distribution of the sample minimum BMI in 2017. Include at least one plot. Report the minimums and standard deviations of the sampling distributions, and describe how they change with increasing sample size.

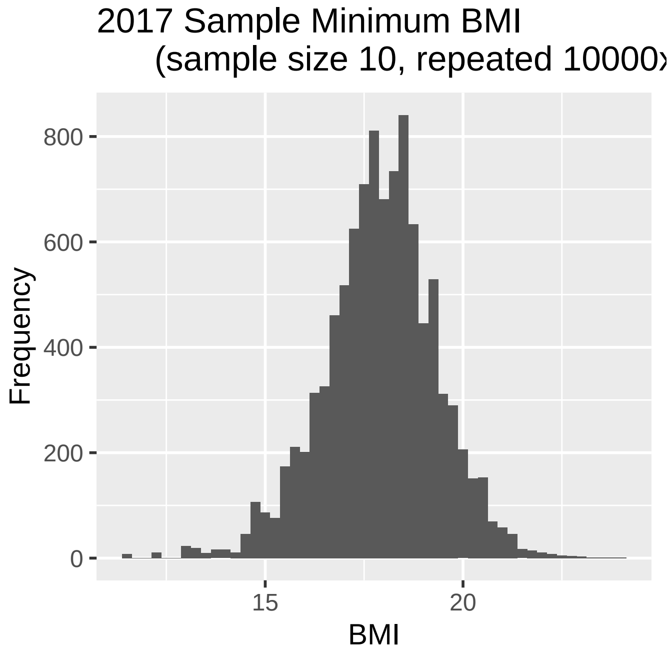
The sampling distributions of the minimum BMI from the 2017 data set are not normally distributed, they have some skew to the right - the central limit theorem only applies to sample means. This is to be expected as almost every value in the data set are to the right of the sample minimum. As the sample size increases, the spread of the data decreases, and the frequency around the true population minimum BMI increases. I have included each plot below. Note that the x and y axes auto scale with the data and I excluded any 'NA' values before running these analyses.

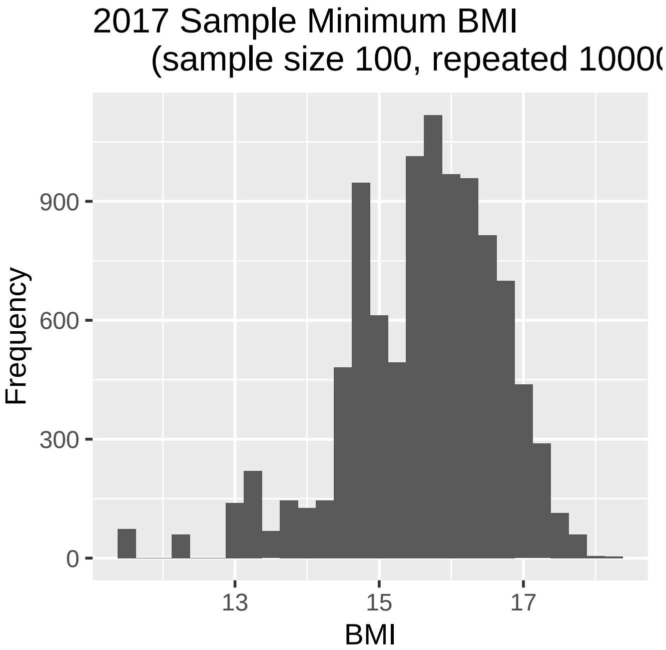
The sample minimums and sample standard deviations are as follows for each sample size:

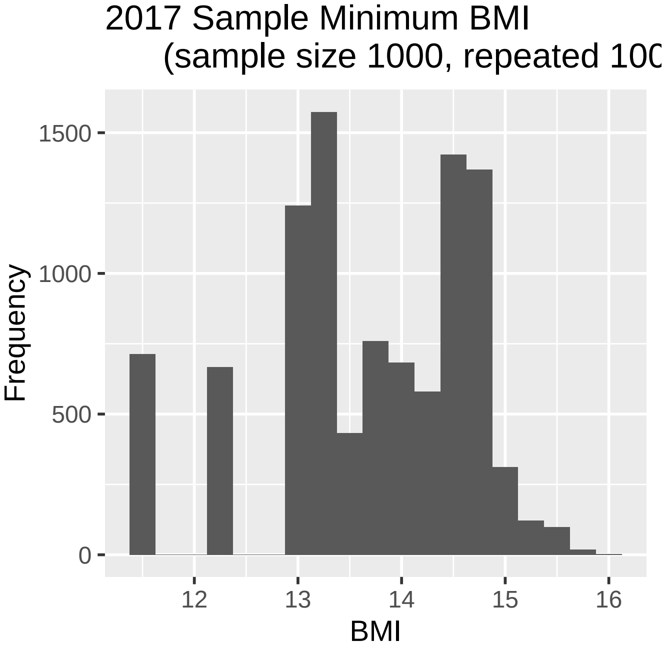
sample size = 10, sample minimum = 17.930, sample SD = 1.451
sample size = 100, sample minimum = 15.580, sample SD = 1.091
sample size = 1000, sample minimum = 13.645, sample SD = 0.989

The sample minimums do change fairly significantly with increasing sample size. I believe the reason for this is that there's a smaller chance a randomly selected data point from the sample set will lie close to the minimum value rather than the mean value. For this same reason, the sample standard deviation decreases as the sample size increases, but not nearly as much as when computing the mean or 25th percentile values.

```
##        [,1]   [,2]   [,3]
## [1,] 17.909 15.586 13.669
## [2,]  1.443  1.076  0.973
```







Part 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.

The sampling distributions of the median BMI from the 2007 and 2017 data sets are not normally distributed, they have a slight skew to the right - the central limit theorem only applies to sample means. That the data appear nearly normally distributed does not come as a surprise because the sample mean BMI is about 1 unit higher than the sample median for both 2007 and 2017. As the sample size increases, the spread of the data decreases, and the frequency around the true population median BMI increases for both years. There is not a large difference in measured sample medians between 2007 and 2017. The measured median BMI difference in 2017 is 0.19 BMI units less than in 2007. The sample standard deviations match almost exactly for each sample size between 2007 and 2017. I have included each plot below. Note that the x and y axes auto scale with the data and I excluded any 'NA' values before running these analyses.

The sample means of the population median and sample standard deviations are as follows for each sample size:

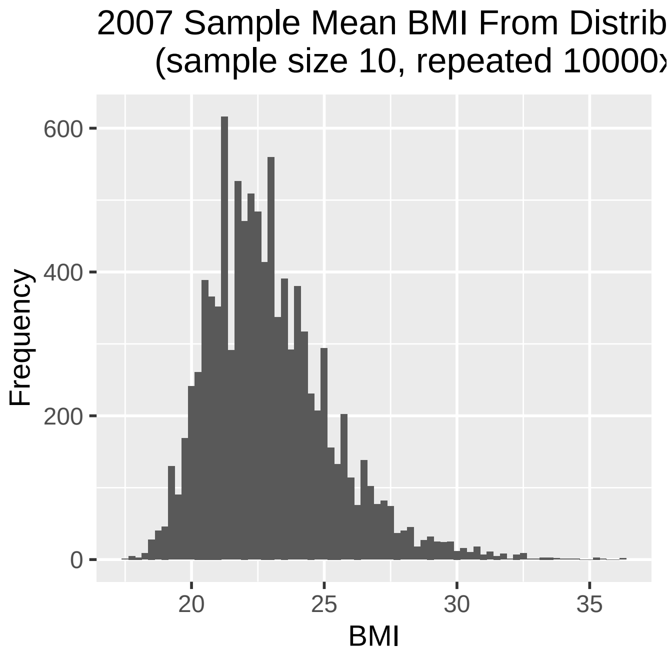
2007 data set: sample size = 10, mean sample median = 22.934, sample SD = 2.326
sample size = 100, mean sample median = 22.784, sample SD = 1.554
sample size = 1000, mean sample median = 22.609, sample SD = 0.486

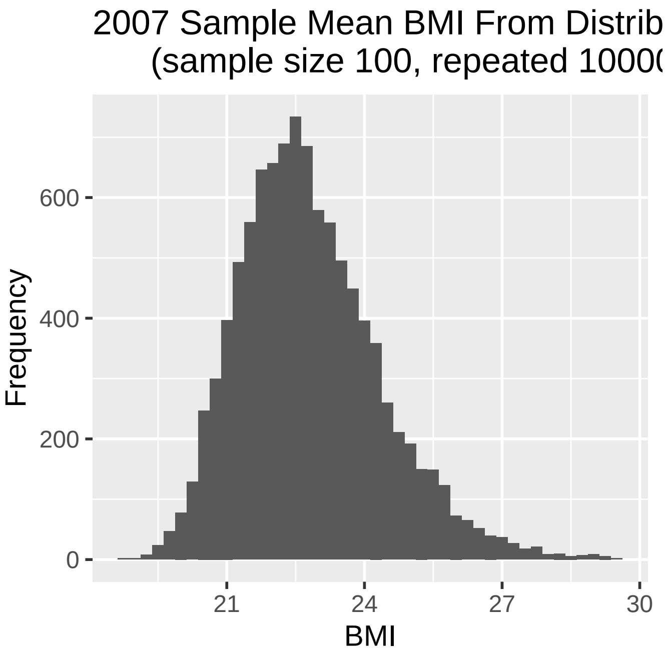
2017 data set: sample size = 10, mean sample median = 22.751, sample SD = 2.387
sample size = 100, mean sample median = 22.597, sample SD = 1.575
sample size = 1000, mean sample median = 22.423, sample SD = 0.512

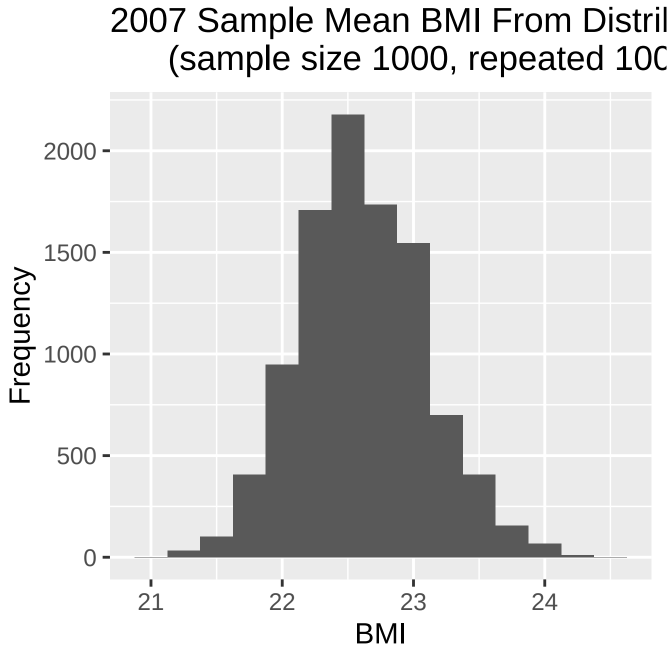
For the data sets from each year, the computed medians do not change very much with increasing sample size because I am replicating each simulation 10000 times. However, the sample standard deviation decreases as the sample size increases, as expected.

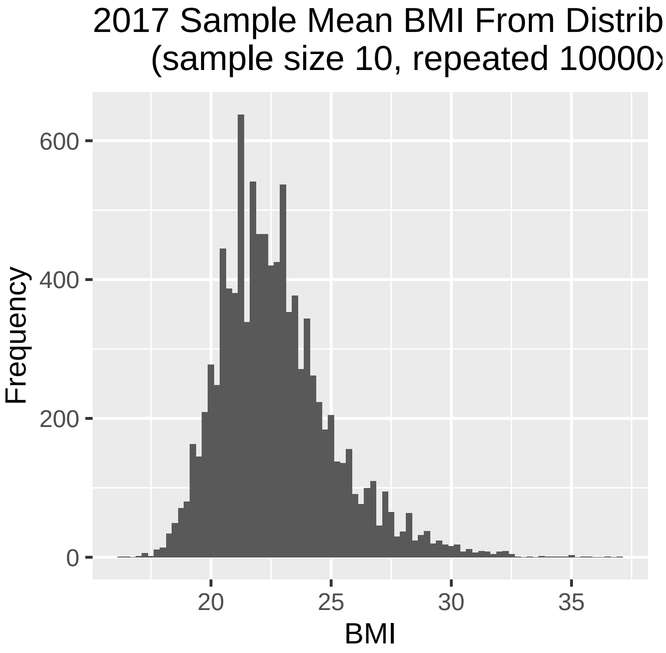
```
##        [,1]   [,2]   [,3]
## [1,] 22.957 22.808 22.609
## [2,]  2.346  1.547  0.482
```

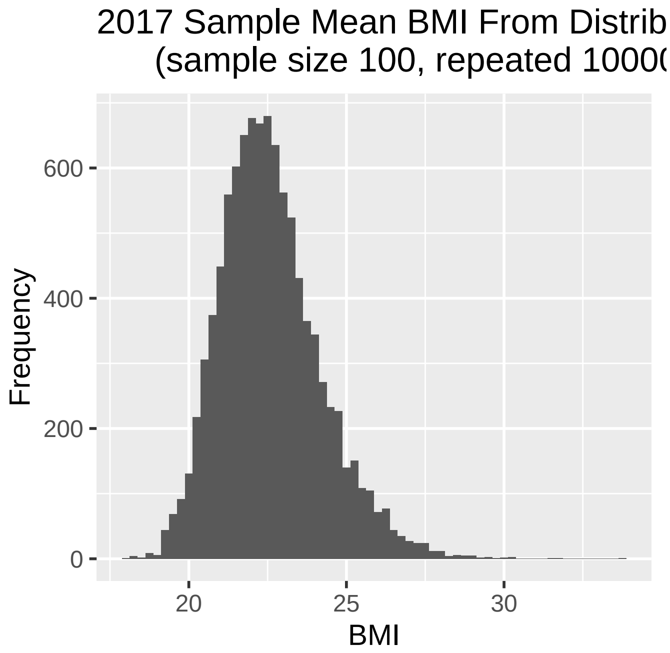
```
##        [,1]   [,2]   [,3]
## [1,] 22.738 22.597 22.423
## [2,]  2.407  1.610  0.506
```

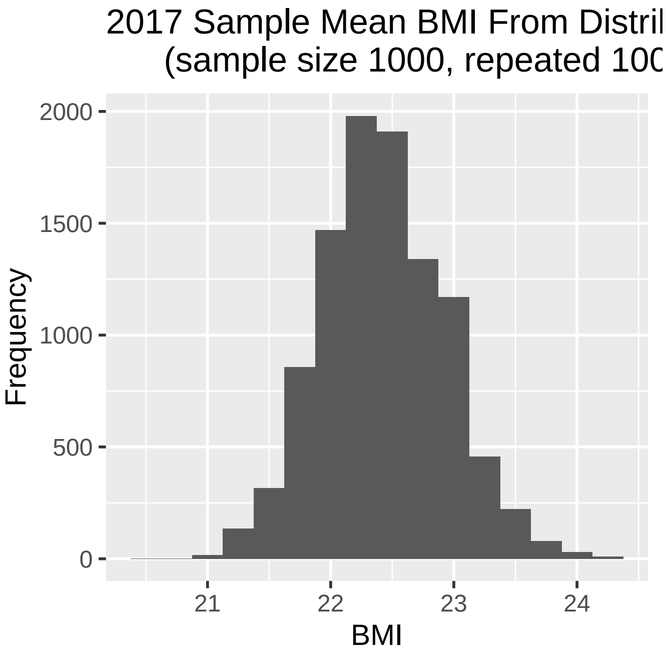












Part E: Comment on the center, spread and shape for the sampling distribution of each statistic as the sample size increases. You should also contrast the behavior of the sample statistics.

To consider the validity of the above analyses, I will consider the two assumptions that need to be met for the data sets. The first is that the sample set needs to be sufficiently large to be able to rely upon the results. At around 13000 per data set, it's safe to assume the sample sets are sufficiently large. The second assumption that needs to be met is independence of individual samples. On the CDC's website regarding these YRBS data sets, it explains that the design included a two-stage cluster sample collection design from public schools around the country to produce a representative sample of students in grades 9-12, so it's safe to say this assumption is also met.

The behavior of the sample statistics vary based on the specific statistic in question. Mean and median appear to have the least amount of skew while 25th percentile and minimum appear to have significantly more skew in their distributions. There was a noticeably greater spread in the distributions of sample minimum and sample median. For the sample minimum, this makes sense for reasons I explained above. For the sample median, the standard error is about 1.25 times greater than the sample mean, and this shows in the computed values as well. The sample mean and 25th percentiles had very small spreads at the largest sample sizes.

In all of the above analyses, it was observed that the sampling distribution was centered around the "true" parameter value. The precision of the estimate also appeared to increase with the square root of the sample size. With the exception of the sample minimum, it appeared that the sampling distributions always moved toward a normal distribution as the sample size increased, though the 25th percentile distribution did maintain some skew to the right at the largest sample size.

Task 2: Data Analysis

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

In order to answer this question I ran the same simulations on the 2007 data set as in Task 1 Part A above and included the plots, sample means and sample standard deviations below. Note that the x and y axes auto scale with the data. I also excluded any 'NA' values before running any of the analyses.

The 2007 sample means and sample standard deviations are as follows for each of the sample sizes:

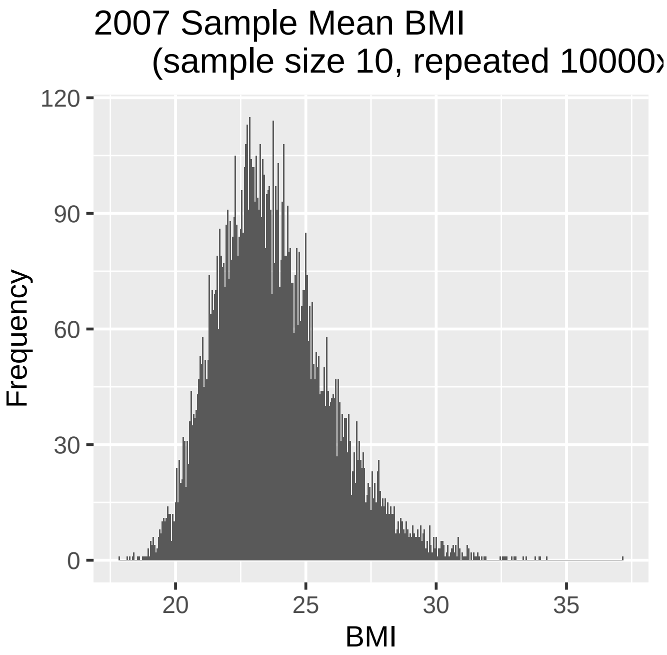
sample size = 10, sample mean = 23.774, sample SD = 2.225
sample size = 100, sample mean = 23.767, sample SD = 0.564
sample size = 1000, sample mean = 23.870, sample SD = 0.500

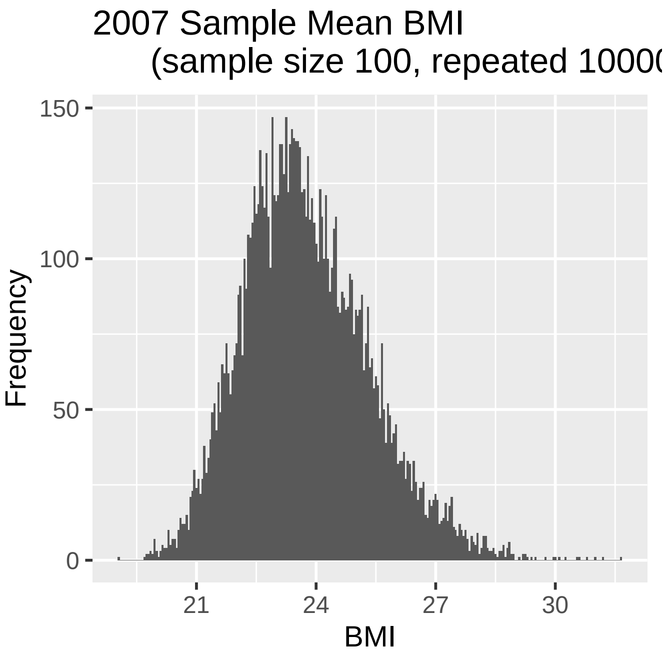
I then ran a two sample t-test on the difference in mean BMIs between 2007 and 2017. The results from the t-test are as follows: With a p-value of 0.01, we can reject the null hypothesis in favor of there being a difference in sample mean BMI between students 2007 and 2017. With 95% confidence, the difference in sample means is between 0.03 and 0.3 BMI units lower in 2017 than 2006. This measure alone would be an example of statistical but not practical significance since the difference is so small.

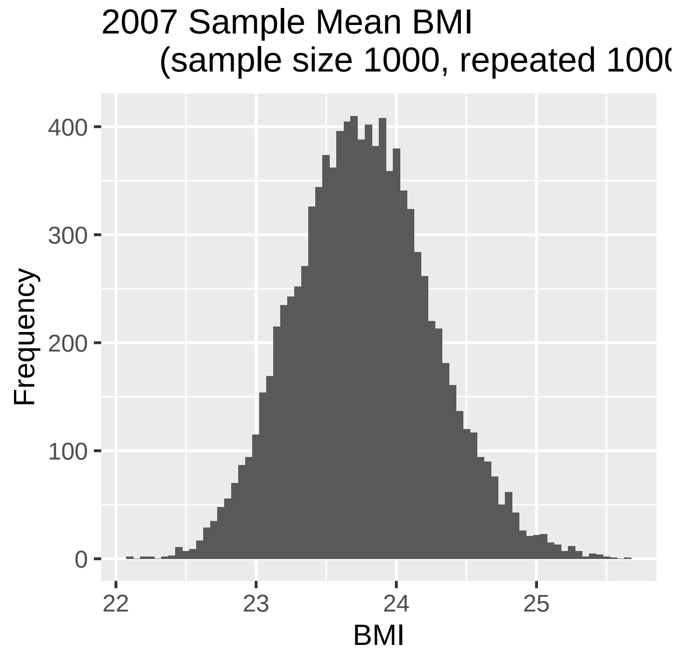
Finally, I computed the 0th, 25th, 50th, 75th and 100th percentile BMIs from each year, which are plotted below. What we can see here is that despite the sample mean and 25th through 75th percentile values being essentially the same, the 0th and 100th percentile BMIs are significantly more extreme in 2017.

To conclude, based on these sample sets, it appears that in the US high-schoolers overall are not necessarily becoming more overweight, but the outlier values have become more extreme from 2007 2017. It has been established that the data evaluated are representative of high-school students in the US, so this conclusion applies to all US high-school students.

```
##         [,1]   [,2]   [,3]
## [1,] 23.753 23.755 23.770
## [2,]  2.195  1.577  0.493
```







```
##
##  Welch Two Sample t-test
##
## data:  data_2007$bmi and data_2017$bmi
## t = 2.4775, df = 26125, p-value = 0.01324
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03249091 0.27864773
## sample estimates:
## mean of x mean of y
##  23.77572  23.62015
```

```
## [1] 23.77572
```

```
## [1] 4.953279
```

```
##      0%     25%     50%     75%    100%
## 13.2994 20.4048 22.5762 25.8255 54.6554
```

```
## [1] 23.62015
```

```
## [1] 5.202513
```

```
##      0%     25%     50%     75%    100%
## 11.5461 20.0893 22.4429 25.7701 62.4793
```

Part 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”?

This is a true/false question, so I ran a two-sample proportion test to determine if 9th graders or 12th graders were more likely to be physically active at least 60 minutes per day on 5 or more days per week. I excluded any ‘NA’ values, so my sample sets that I ran the proportion test on decreased to 3284 and 3009 for 9th and 12th graders versus the original sample set of sizes of 3479 and 3119.

My findings were as follows: Based on a p-value of 2.2e-16, there is strong evidence to support that 9th graders in the US are more likely than 12th graders to be physically active at least 60 minutes per day on 5 or more days per week. There’s a probability of ~0.51 that 9th graders meet this exercise standard and a probability of ~0.38 that 12th graders meet this exercise standard. With 95% confidence, 9th graders have a 0.11 to 0.15 higher probability of meeting this exercise standard than 12th graders in the US.

It has been established that the data evaluated are representative of high-school students in the US, so this conclusion applies to all US high-school students.

```
## 
##  2-sample test for equality of proportions with continuity correction
## 
## data:  x out of n
## X-squared = 106.25, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1050340 0.1543997
## sample estimates:
##    prop 1    prop 2 
## 0.5115713 0.3818544
```

```
length(ninth_raw)
```

```
## [1] 3479
```

```
length(ninth)
```

```
## [1] 3284
```

```
length(twelfth_raw)
```

```
## [1] 3119
```

```
length(twelfth)
```

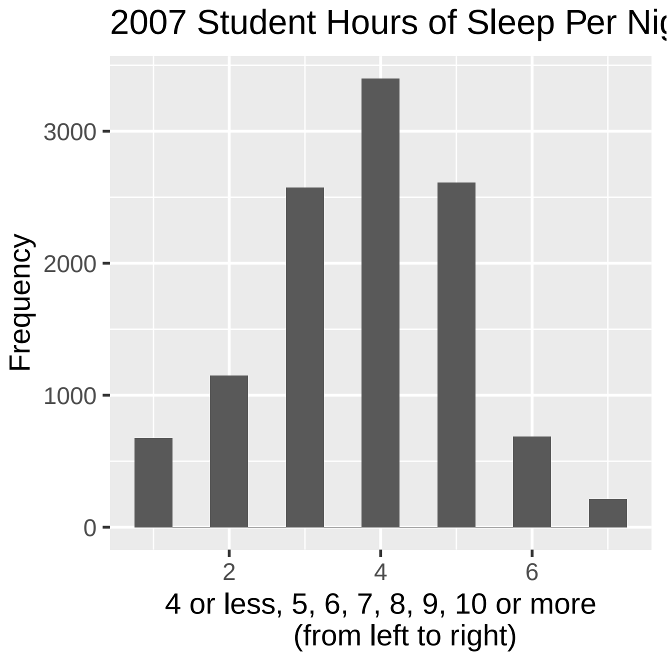
```
## [1] 3009
```

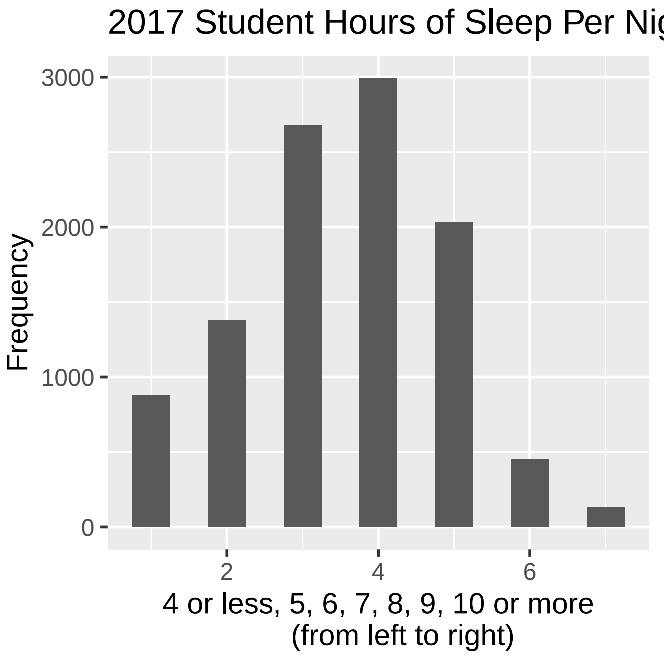
Part 3: How much sleep do high-schoolers get?

In order to answer this question I converted the factor type data to integer based data on the amount of sleep that high-schoolers were reporting to get in 2007 versus 2017. I then plotted the data to look for trends between the two years. I excluded any ‘NA’ values before plotting the data. In 2007 and 2017 there were 11309 and 10552 respondents, respectively.

My findings are as follows: The data appear to be more or less normally distributed for each year. Roughly the same proportion of students report getting 7 hours of sleep in 2007 and 20017. In 2007 there were a greater proportion of students reporting 8, 9, or 10+ hours of sleep per night than in 2017. In 2017, there were a greater proportion of students reporting 4 or less, 5, or 6 hours of sleep than in 2007. So the overall trend appears to be that high-school students were getting less sleep in 2017 as compared to 2007, even if the greatest proportion of students (~0.3) were reporting to get 7 hours of sleep per night on average in both 2007 and 2017.

It has been established that the data evaluated are representative of high-school students in the US, so this conclusion applies to all US high-school students.





```
length(sleep_2007_num)
```

```
## [1] 11309
```

```
length(sleep_2017_num)
```

```
## [1] 10552
```