Analysis of the Behavior of Selected Sample Statistics on the Youth Risk Behavior Surveillance System Population

Sam Oliszewski

Oregon State University

Table of Contents

Introduction

This paper provides an analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS), as a population.

In the first section of the paper is the analysis of the YRBSS data. First, the sampling distribution of the sample mean of body mass indexes (BMI) of YRBSS students in 2013 was determined. Second, the sampling distribution of the sample 25^{th} percentile of BMIs of YRBSS students in 2013 was generated. Third, the sampling distribution of the sample minimum of BMIs of YRBSS students in 2013 was created. Finally, the sampling distribution of the difference in sample medians of BMIs of YRBSS students in 2003 and 2013 was produced.

In the second section of this paper is the inference on the population of high-school students based on the sample of YRBSS students. First, there is a determination as to whether high-school students are becoming more overweight. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day.

Ultimately, the information gathered from the YRBSS data can help make useful inferences about the population of high-school students. The research defined in this paper provide the context for such inferences.

Simulation Study

This section discusses the simulation study performed to investigate the properties of four sample statistics: the mean, 25^{th} percentile, minimum and difference in medians. The study is based on observational data on student BMIs from the YRBSS.

Mean

Methods. To generate a plot of the sampling distribution of the sample mean of the BMI of the Youth Risk Behavior Surveillance System in 2013, 10,000 samples of sizes 10 were taken. The sample mean was calculated for each sample and plotted on a graph. This process was then repeated for sample sizes of 100 and 1,000.

Results. The resulting distribution for each of the sample sizes became increasingly more normal as the sample size increased. The means of the sampling distribution for sizes 10, 100 and 1,000 were 23.636, 23.640, and 23.643, respectively. The standard deviations of the sampling distribution for sample sizes 10, 100 and 1,000 were 1.59, 0.50, and 0.15, respectively.

Figures. The following figures depict the normalization of the sampling distribution of the sample mean as the sample size increased:

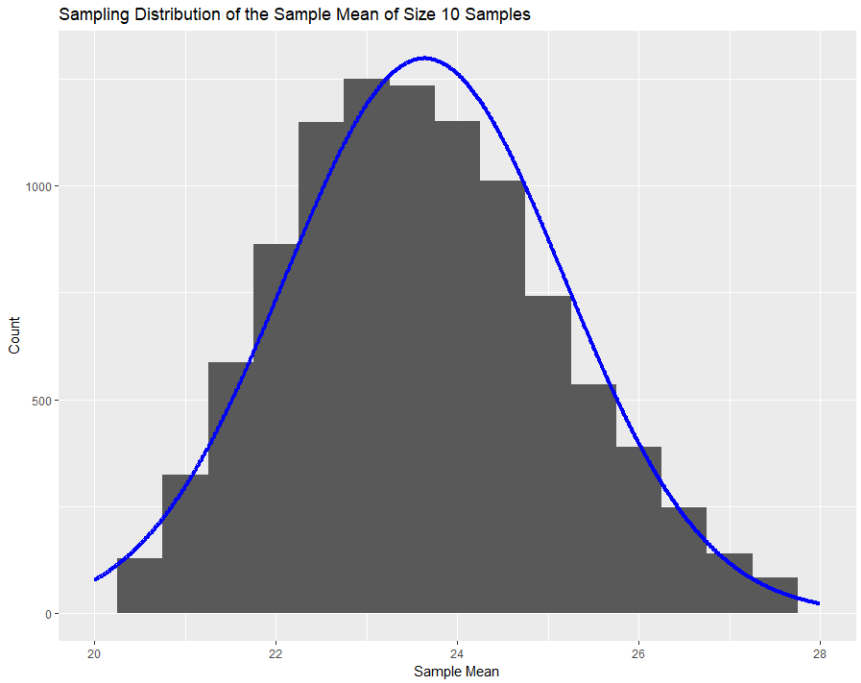

Figure 1. Sampling Distribution of the Sample Mean of Size 10 Samples with Normal Distribution Overlay.

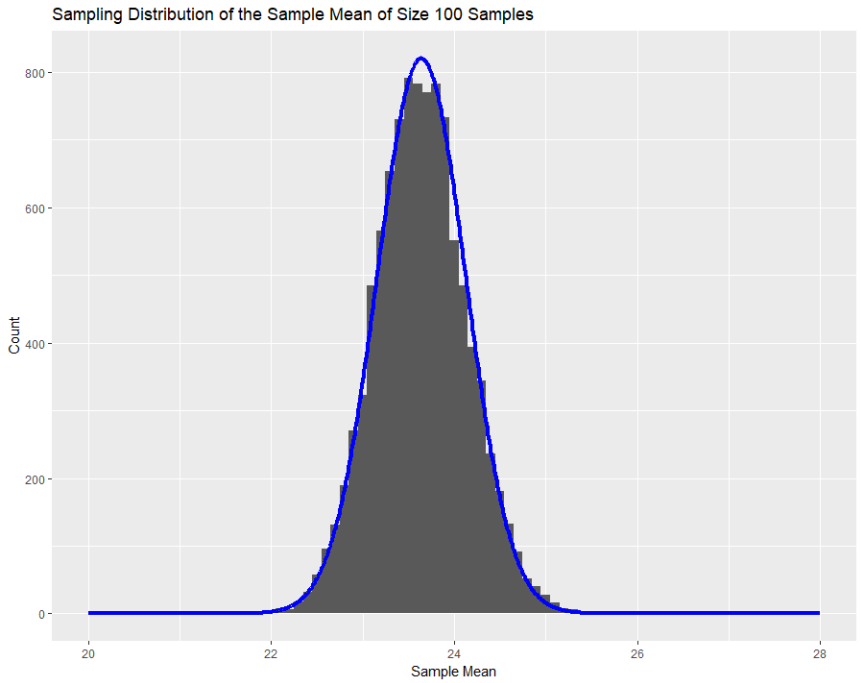

Figure 2. Sampling Distribution of the Sample Mean of Size 100 Samples with Normal Distribution Overlay.

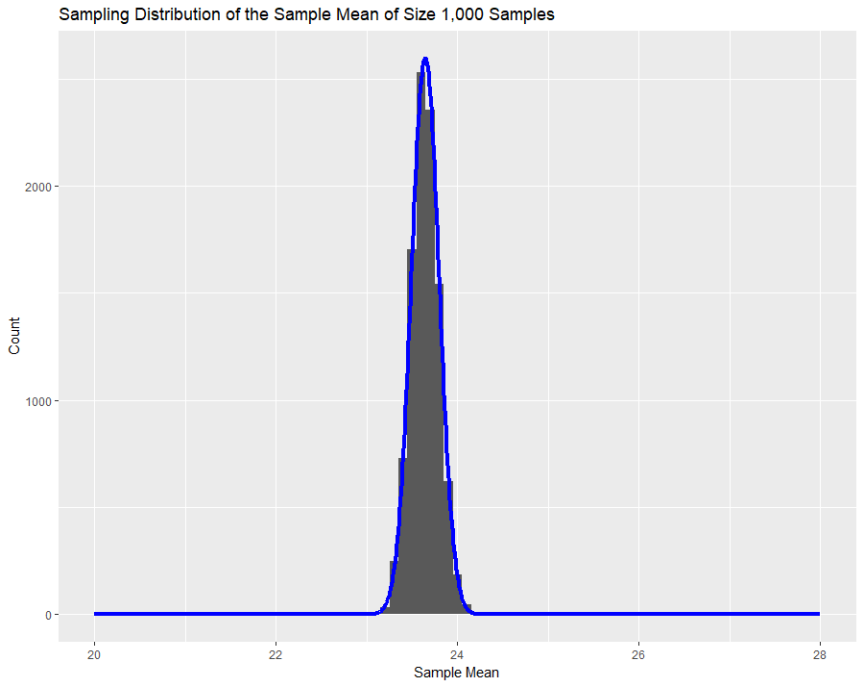

Figure 3. Sampling Distribution of the Sample Mean of Size 1,000 Samples with Normal Distribution Overlay.

Summary. As explained by the Law of Large Numbers, with increasing sample size, the mean of the sampling distribution converged to the approximate value of the true population mean (23.64). The standard deviation of the sampling distribution declined because as sample size increases, each sample represents a larger proportion of the population (until the entire population is studied) and thus repeated samples of a large size tend to reflect the true population parameter more accurately. With large samples, there is a greater chance of observing closer values which explains the smaller standard deviation. The converging to the true population mean and the decline in standard deviation as sample sizes increase reflects the Central Limit Theorem.

25th Percentile

Methods. To generate a plot of the sampling distribution of the sample 25th percentile of the BMI of the Youth Risk Behavior Surveillance System in 2013, 10,000 samples of sizes 10 were taken. The 25th percentile was calculated for each sample and plotted on a graph. This process was then repeated for sample sizes of 100 and 1,000.

Results. The resulting distribution for each of the sample sizes became increasingly more normal as the sample size increased. The means of the sampling distribution of the sample 25th percentile for sizes 10, 100 and 1,000 were 20.64, 20.30, and 20.28, respectively. The standard deviations of the sampling distribution for sample sizes 10, 100 and 1,000 were 1.25, 0.40, and 0.13, respectively.

Figures. The following figures depict the normalization of the sampling distribution of the sample 25th percentile as the sample size increased:

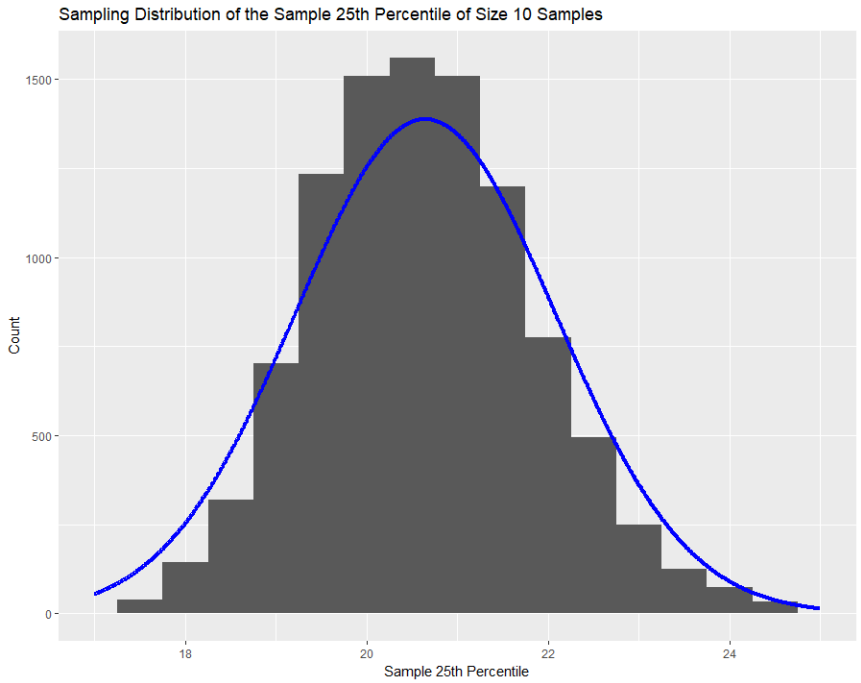

Figure 4. Sampling Distribution of the Sample 25th Percentile of Size 10 Samples with Normal Distribution Overlay.

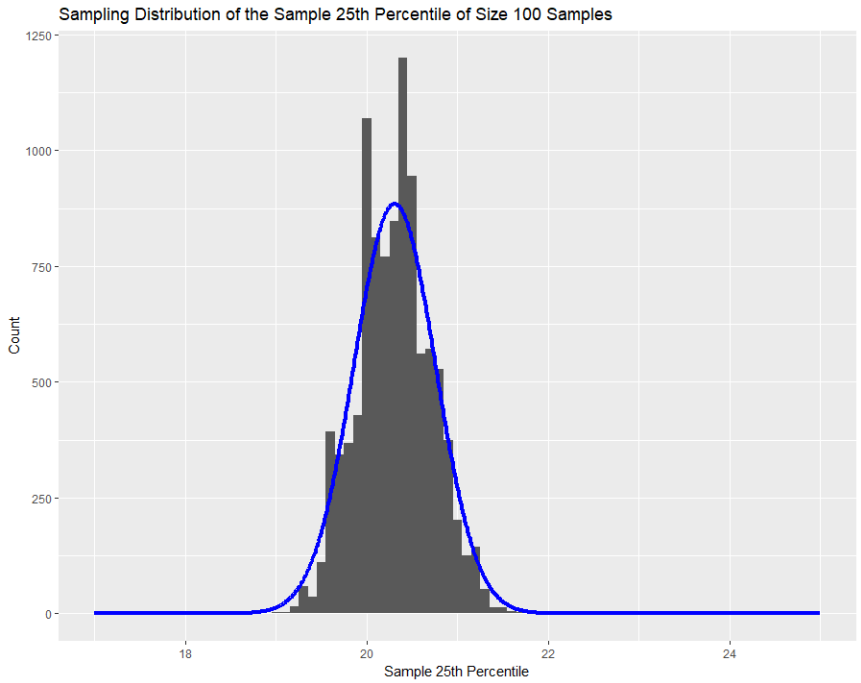

Figure 5. Sampling Distribution of the Sample 25th Percentile of Size 100 Samples with Normal Distribution Overlay.

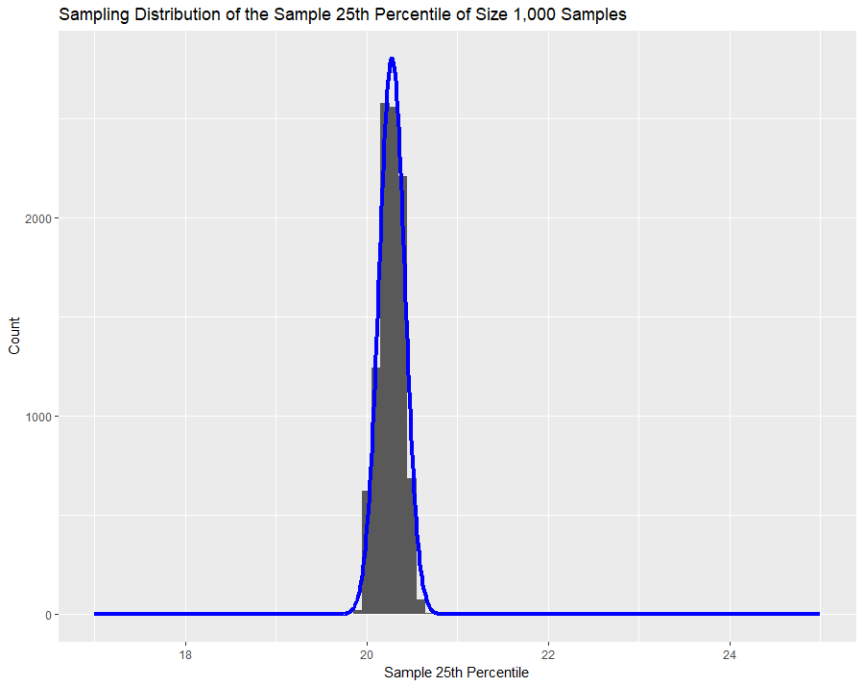

Figure 6. Sampling Distribution of the Sample 25th Percentile of Size 1,000 Samples with Normal Distribution Overlay.

Summary. With increasing sample size, the sampling distribution of the sample 25th percentile normalized, and the standard deviation of the sampling distribution declined. This behavior of the standard deviation is because as sample size increases, each sample represents a larger proportion of the population (until the entire population is studied) and thus repeated samples of a large size tend to reflect the true population parameter more accurately. Therefore, with large samples, there is a greater chance of observing closer values which reflects the smaller standard deviation.

Minimum

Methods. To generate a plot of the sampling distribution of the sample minimum of the BMI of the Youth Risk Behavior Surveillance System in 2013, 10,000 samples of sizes 10 were taken. The sample minimum was calculated for each sample and plotted on a graph. This process was then repeated for sample sizes of 100 and 1,000.

Results. The resulting distribution for each of the sample sizes became increasingly more uniform as the sample size increased. The means of the sampling distribution of the sample minimum for sizes 10, 100 and 1,000 were 18.04, 15.64, and 14.03, respectively. The standard deviations of the sampling distribution for sample sizes 10, 100 and 1,000 were 1.46, 0.99, and 0.58, respectively.

Figures. The following figures depict the sampling distribution of the sample minimum becoming more uniform as the sample size increased:

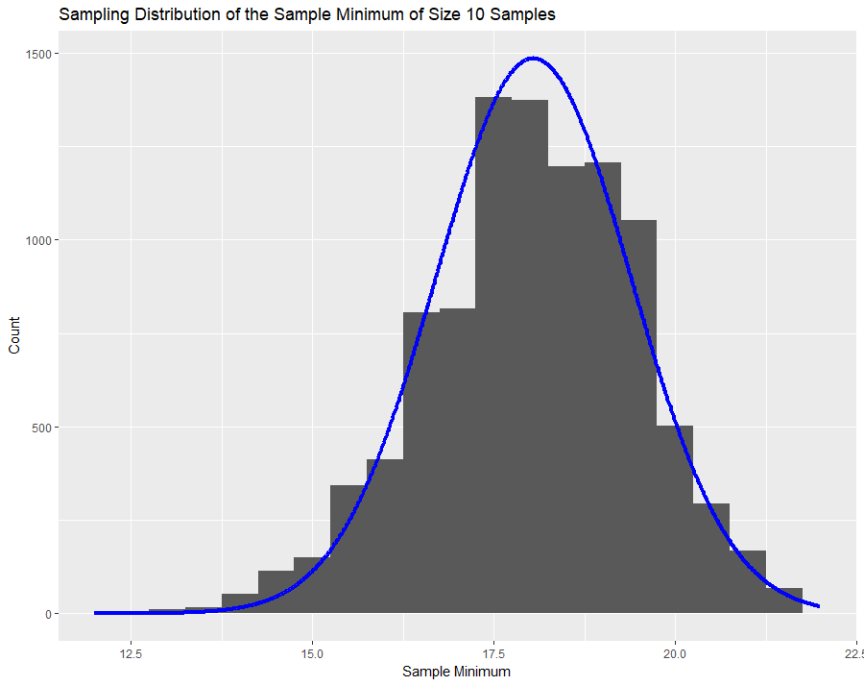

Figure 7. Sampling Distribution of the Sample Minimum of Size 10 Samples with Normal Distribution Overlay.

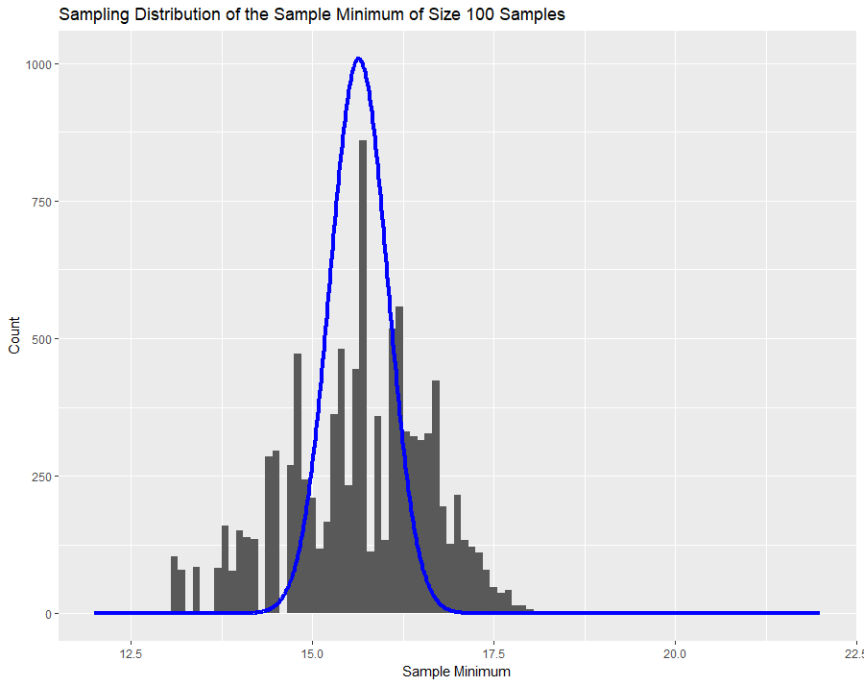

Figure 8. Sampling Distribution of the Sample Minimum of Size 100 Samples with Normal Distribution Overlay.

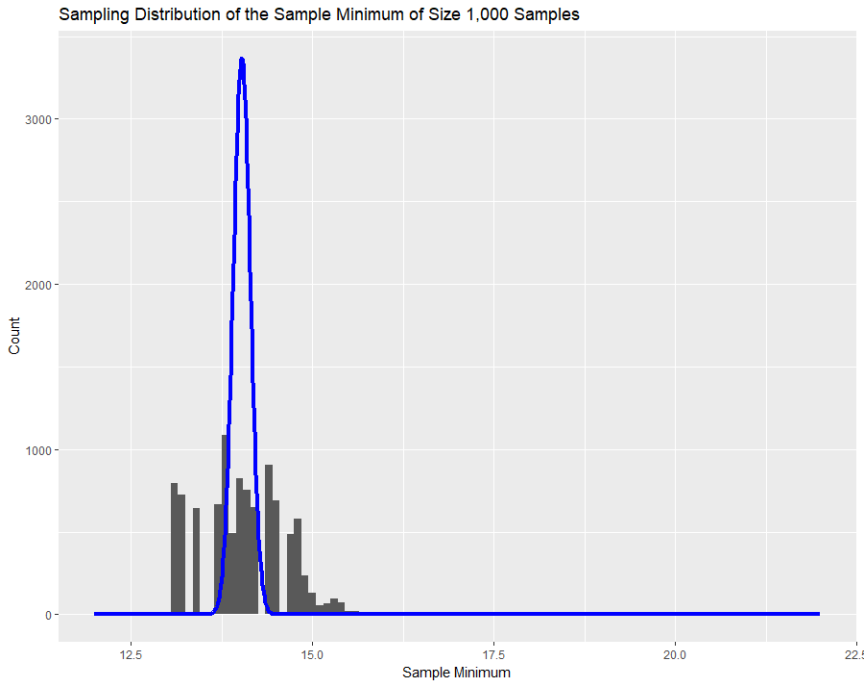

Figure 9. Sampling Distribution of the Sample Minimum of Size 1,000 Samples with Normal Distribution Overlay.

Summary. With increasing sample size, the sampling distribution of the sample minimum became more uniform and the standard deviation of the sampling distribution of the sample minimum declined. This behavior of the standard deviation is because as sample size increases, each sample represents a larger proportion of the population (until the entire population is studied) and thus repeated samples of a large size tend to reflect the true population parameter more accurately. Therefore, with large samples, there is a greater chance of observing closer values which reflects the smaller standard deviation.

Difference in Medians

Methods. To generate a plot of the sampling distribution of the difference in sample median BMIs between 2013 and 2003 of the Youth Risk Behavior Surveillance System, 10,000 samples of size 5 were taken from the population in 2013 and the population in 2003. Then the sample medians were taken from each sample and the difference between the sample medians of 2013 and 2003 was calculated and plotted on a graph. This process was then repeated for sample sizes of 10 and 100.

Results. The resulting distribution for each of the sample sizes were non-normal with long tails that saw a decrease in the standard deviation as the sample size increased. The means of the sampling distribution of the difference in sample medians for sizes 5, 10 and 100 were 0.22, 0.20, and 0.17, respectively. The standard deviations of the sampling distribution for sample sizes 5, 10 and 100 were 3.20, 2.12, and 0.67, respectively.

Figures. The following figures depict the non-normal, long-tailed sampling distribution of the difference in sample medians with declining standard deviation as sample size increased:

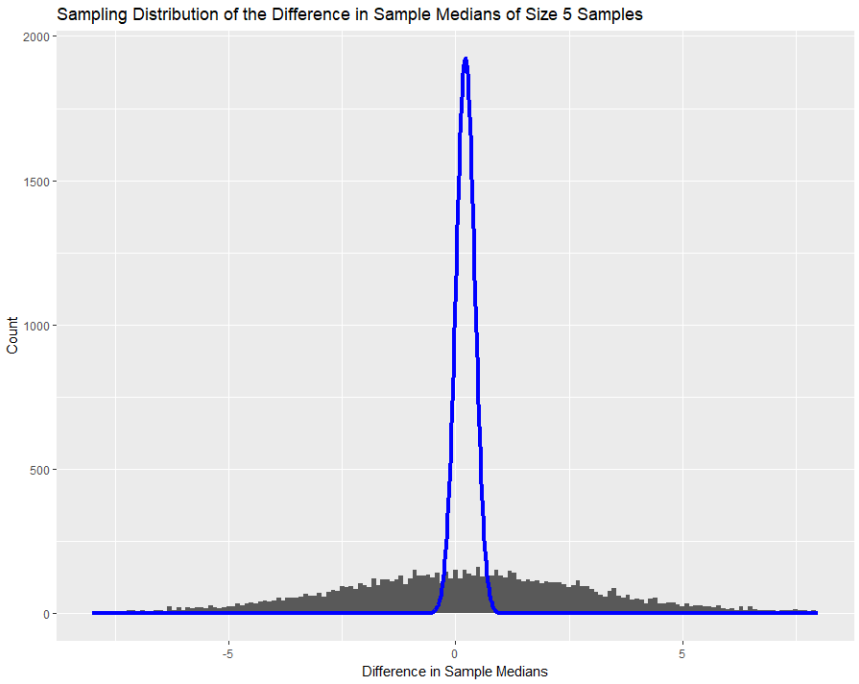

Figure 10. Sampling Distribution of the Difference in Sample Medians of Size 5 Samples with Normal Distribution Overlay.

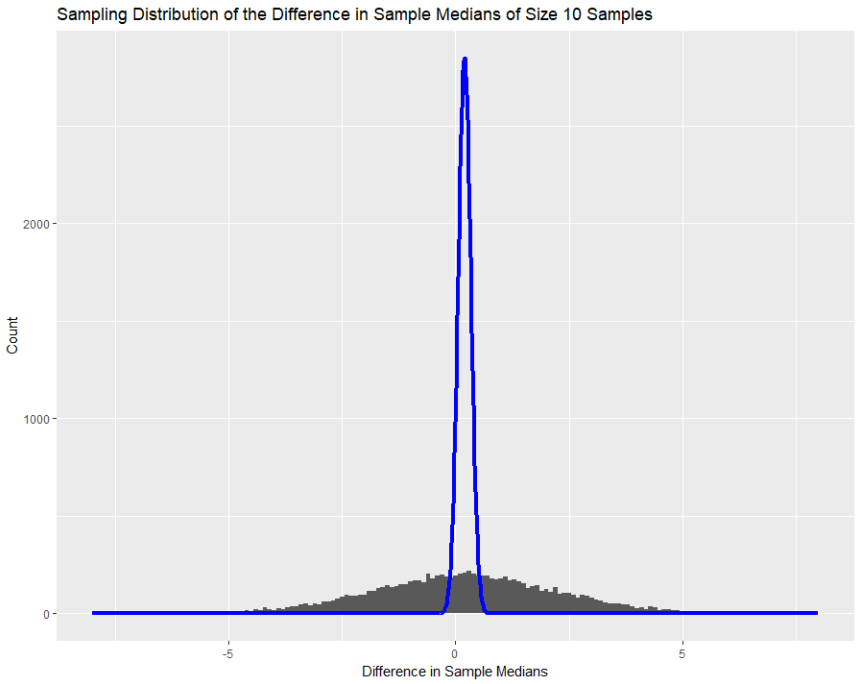

Figure 11. Sampling Distribution of the Difference in Sample Medians of Size 10 Samples with Normal Distribution Overlay.

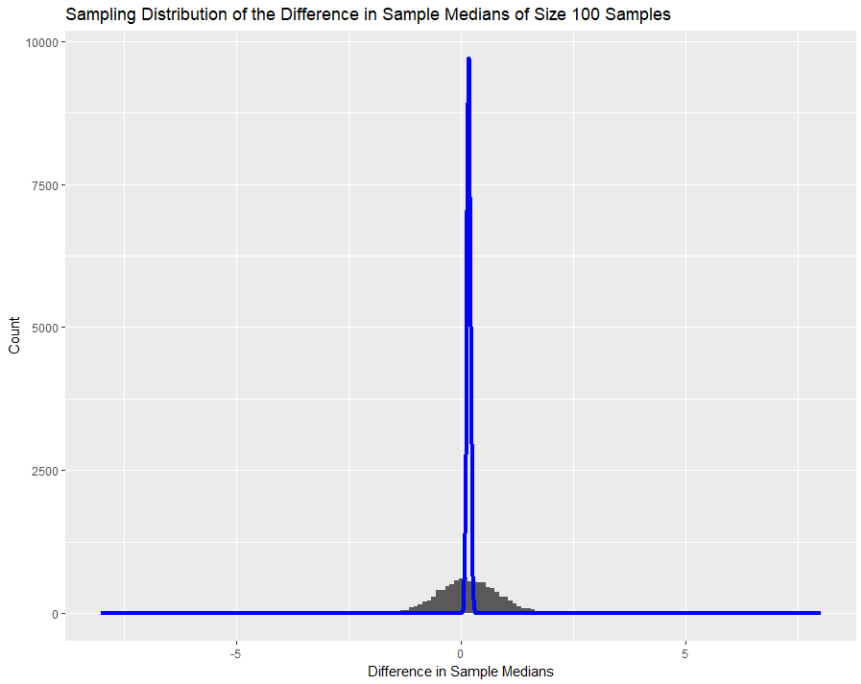

Figure 12. Sampling Distribution of the Difference in Sample Medians of Size 100 Samples with Normal Distribution Overlay.

Summary. The sampling distribution of the difference in sample medians was non-normal with long tails. With increasing sample size, the standard deviation of the sampling distribution of the difference in sample medians declined because as sample size increases, each sample represents a larger proportion of the population (until the whole population is studied) and thus repeated samples of a large size tend to reflect the true population parameter more accurately. Therefore, with large samples, there is a greater chance of observing closer values which reflects the smaller standard deviation.

Summary of Simulation Study

For the mean, 25th percentile, minimum, and difference in medians the true population values for each parameter are 23.64, 20.30, 13.11, and 0.21, respectively. These values are the captured by the sampling distribution for each sample statistic near the center of the distribution. It was observed that as sample size increased for the mean and 25th percentile, the sampling distribution for each sample statistic normalized. This differs from the behavior of the sample minimum, which became more uniform, and the difference in sample medians, which was long-tailed with a declining standard deviation as the sample size increased.

Data Analysis

Are High-Schoolers Getting More Overweight?

Methods. To study the BMI of high-school students between 2003 and 2013, a Welch's two-sample t-test was run with a null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013 and an alternative hypothesis that the mean BMI of students in 2013 is greater than the mean BMI of students in 2003. The variances are assumed to be unequal between the populations since there is no clear evidence that they should be assumed equal. The data from 2003 and 2013 are also assumed to be independent because they are not tracking the same students. A two-sample t-test is an appropriate choice for this study because the parameter in question is the mean. This test can be applied because the sample size is large allowing the Central Limit Theorem to apply and validating the use of the approximately normal test statistic. Further, a t-distribution is a reasonable comparison to use, since the sample is approximately normal, but not guaranteed to be normal.

Results. The two-sample t-test yielded a p-value of 8.76e-05, suggesting that there is significant evidence to reject the null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013. It is estimated that the mean BMI of students in 2013 is 23.64 and the mean BMI of students in 2003 is 23.41. With 95% confidence, the mean BMI of students in 2013 is at least 0.13 units larger than the mean BMI of students in 2003.

Figures. The following figure depicts the increase in high-value outliers in the 2013 BMI data that caused the increase in mean BMI from 2003 to 2013:

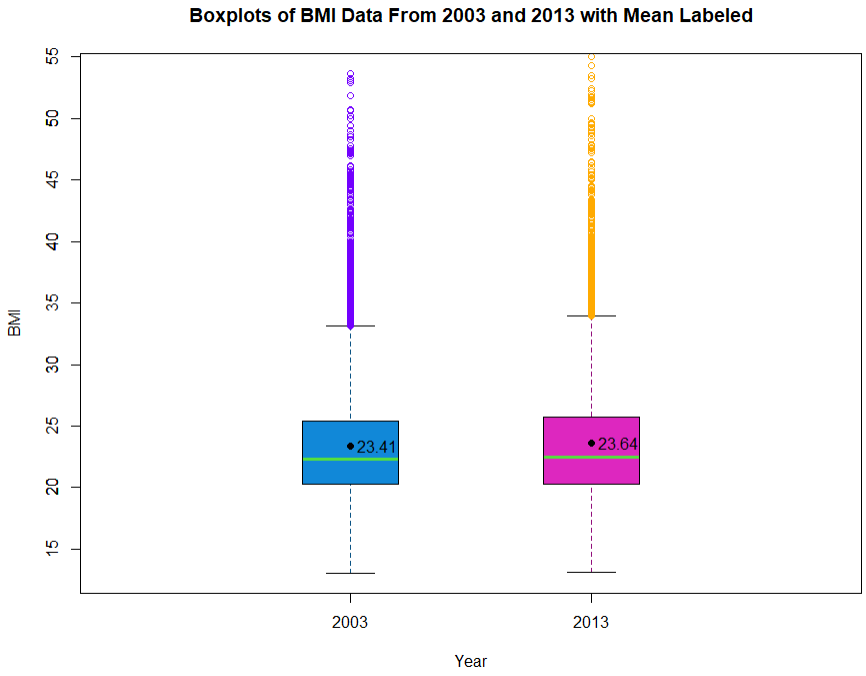

Figure 13. Boxplots of BMI Data From 2003 and 2013 with Mean Labeled.

Summary. The result of this test showed that high-school students in 2013 are more overweight than the students from 2003 (Welch's two sample t-test, t = 3.75, df = 25988, p-value = 8.76e-05).

Are Male High-Schoolers More Likely to Smoke than Female High-Schoolers?

Methods. To study the likelihood that male high-schoolers are more likely to smoke than female high-schoolers, a two-sample proportion test was run with a null hypothesis that the proportion of male high-school smokers was equal to the proportion of female high-school smokers and an alternative hypothesis that the proportion of male high-school smokers is greater than the proportion of female high-school smokers. A two-sample proportion test is appropriate for this study because proportions are an appropriate tool for comparing binary data. In the case of determining whether students are smokers, the answer would be "yes" or "no" for the purpose of this question. Therefore, a proportion of "yes" responses would indicate how many students are smokers. Then, the comparison of proportions between the two sexes would indicate whether one is more likely to smoke than the other. The two-sample proportion test can be used because the following conditions are met: 1. The samples of female and male students are sufficiently large, and 2. The samples of female and male students are independent.

Results. The two-sample proportion test yielded a p-value of 1.02e-07, suggesting there is significant evidence to reject the null hypothesis that the proportion of male high-school smokers is equal to the proportion of female high-school smokers. It is estimated that the proportion of male high-school smokers is 10.99% and the proportion of female high-school smokers is 8.25%. With 95% confidence, the proportion of male high-school smokers is between 2% and 100% larger than the proportion of female high-school smokers.

Data. The following table displays the number of smokers versus total students and the calculated proportion of smokers by sex:

High-School Student Smoker Proportions			
Sex	***Number of Smokers***	***Number of Students***	***Proportion of Student Smokers***
Male	705	6414	10.99%
Female	509	6166	8.25%

Table 1. High-School Student Smoker Proportions.

Summary. The result of this test showed that male high-schoolers are more likely to smoke than female high-schoolers (two sample proportion test, x-squared = 27.00, df =1, p-value = 1.02e-07).

How Much TV Do High-Schoolers Watch?

Methods. To study how much TV high-schoolers watch, the median TV time value reported by high-schoolers in 2013 was calculated. This value is assumed to reflect the population of high-schoolers in the present because it is the most recent data available. Since the median function in R requires numerical data, the student responses were converted from ordinal

categories to numerical values and then the median was determined. The numerical value was ultimately converted back into the corresponding categorical value for interpretation purposes. The median is an appropriate measure of the average amount of time high-schoolers watch TV because the data recorded was in ordinal categories and therefore the median is an appropriate measure of the center of the data because other statistics, such as the mean, wouldn't apply to this type of data.

Results. The median value reported by students in 2013 was "2 hours per day".

Data. This table displays the student responses to the amount of TV they watch per day and the frequency of each response:

Frequency of TV Watching Responses	
TV Watching Time	***Frequency***
No TV on average school day	1,671
Less than 1 hour per day	2,021
1 hour per day	1,667
2 hours per day	2,548
3 hours per day	1,995
4 hours per day	977
5 or more hours per day	1,430

Table 2. Frequency of TV Watching Responses.

Summary. The amount of TV that high-schoolers watch is estimated to be two hours per day.

Conclusion

Based on the analysis of the YRBSS data, inferences about the population of high-school students can be drawn. It was observed that students are becoming more overweight and that male students are more likely to smoke than females. It is also known that students on average are watching approximately two hours of TV per day. Understanding this data allows law-makers and educators to guide student behavior in a way that could improve student health.