Final Report
ST 516

Maryanne Evans

In this report we explore data from the **Youth Risk Behavior Surveillance System (YRBSS)** study using respondents from two years of observational data collected from high school students in the USA from years 2007 and 2017. These data include variables such as BMI, age, race, grade, sex, milk consumption, alcohol related car rides, hours of sleep per school night and time for physical activity per week.

For this report, the variable of primary interest is student BMI.

Task 1: Simulation Study

The properties of four sample statistics*; mean, 25th percentile, minimum, and a difference in medians* are investigated to explore the relationship these data exhibit after simulating samples to represent the high school student "population" of interest. By exploring these data distributions, assumptions we can guide data analysis techniques detailed in Task 2.

Sample Statistic: Mean

First, we explore the year 2017 BMI of high school students, simulating repeated samples for increasing sample sizes (n=10, 100 and 1000), and calculating the *standard deviation* (σ) and *mean ($\bar{x}$)* for each distribution for the sample statistic **mean.**

Sample Mean Visualization

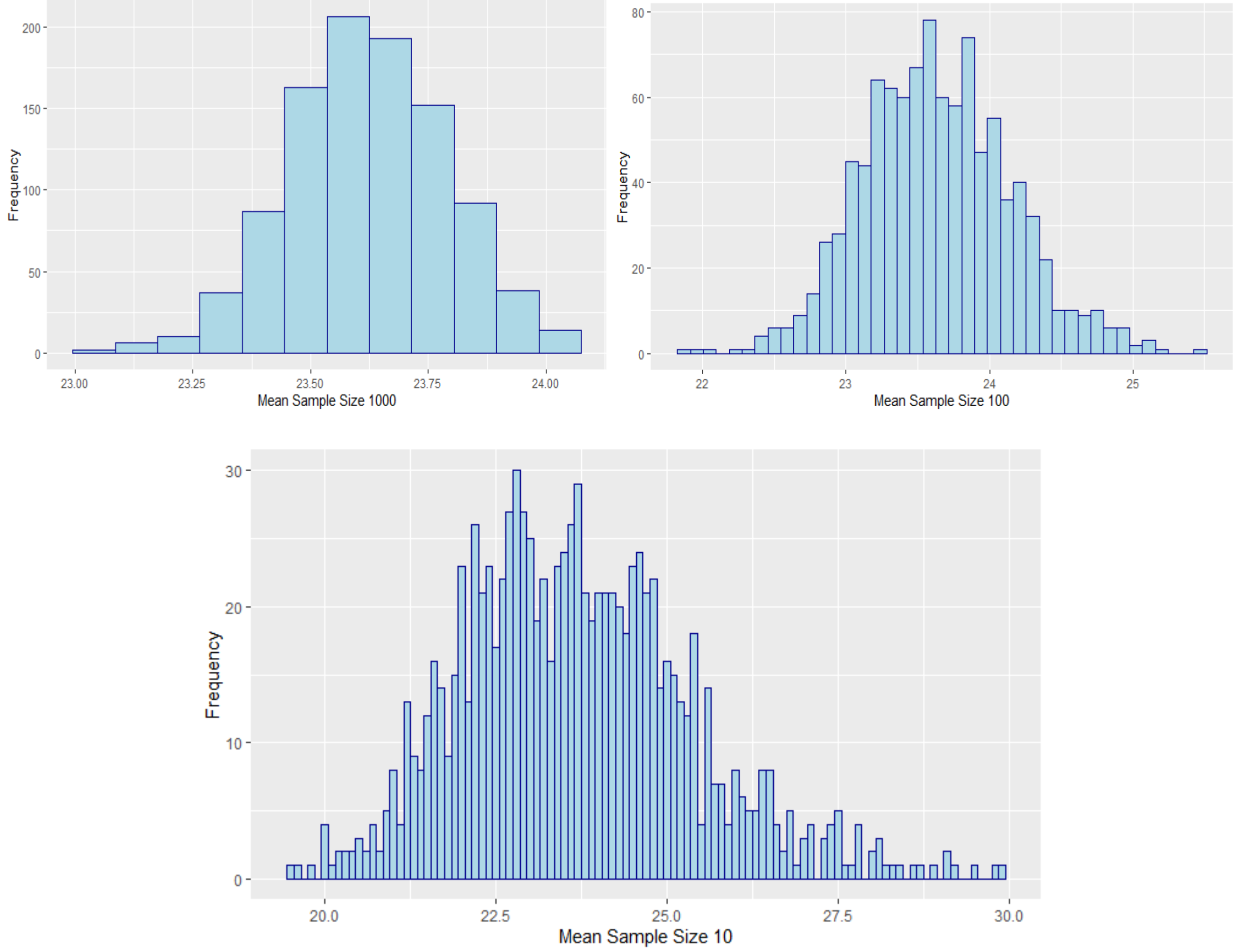

Figure 1: Above are three histograms displaying the distribution of the 2017 student BMI repeated samples for n = 1000, 100 and 10 of the ***sample mean****.*

With Figure 1, it is observed this sample statistic maintains a normal distribution as sample size increases, meeting the assumptions of the **Central Limit Theorem** of data becoming increasingly normal.

Along with the **Law of Large Numbers**, the assumption that with increasing sample size there is a distributive centralizing around the expected, true value. This becomes clear with the calculated $\bar{x}$ for each distribution, which remains relatively stable with increasing sample size; *$n = 10$, $\bar{x} = 20.471$, and $n = 100$, $\bar{x} = 20.129$ and $n = 1000$, $\bar{x} = 20.112$*, indicating a true estimate for the population mean (μ).

Meanwhile, the variance (σ^2), or spread, of the data decreases as represented with the standard deviations (σ) for each distribution as sample size increases towards infinity. This too can be explained with the Central Limit Theorem, in that for example, with *$n = 10$, $\sigma = 1.261$ and as sample size increases with $n = 100$, $\sigma = 0.405$, until $n = 1000$, $\sigma = 0.138$*. The distributions narrow over the true mean (μ).

Sample Statistic: 25th Percentile

Next, we explore the year 2017 BMI of high school students, simulating repeated samples for increasing sample sizes (n=10, 100 and 1000), and calculating the *standard deviation* (σ) and *mean ($\bar{x}$)* for each distribution for the sample statistic **25th percentile**.

25th Percentile Visualization

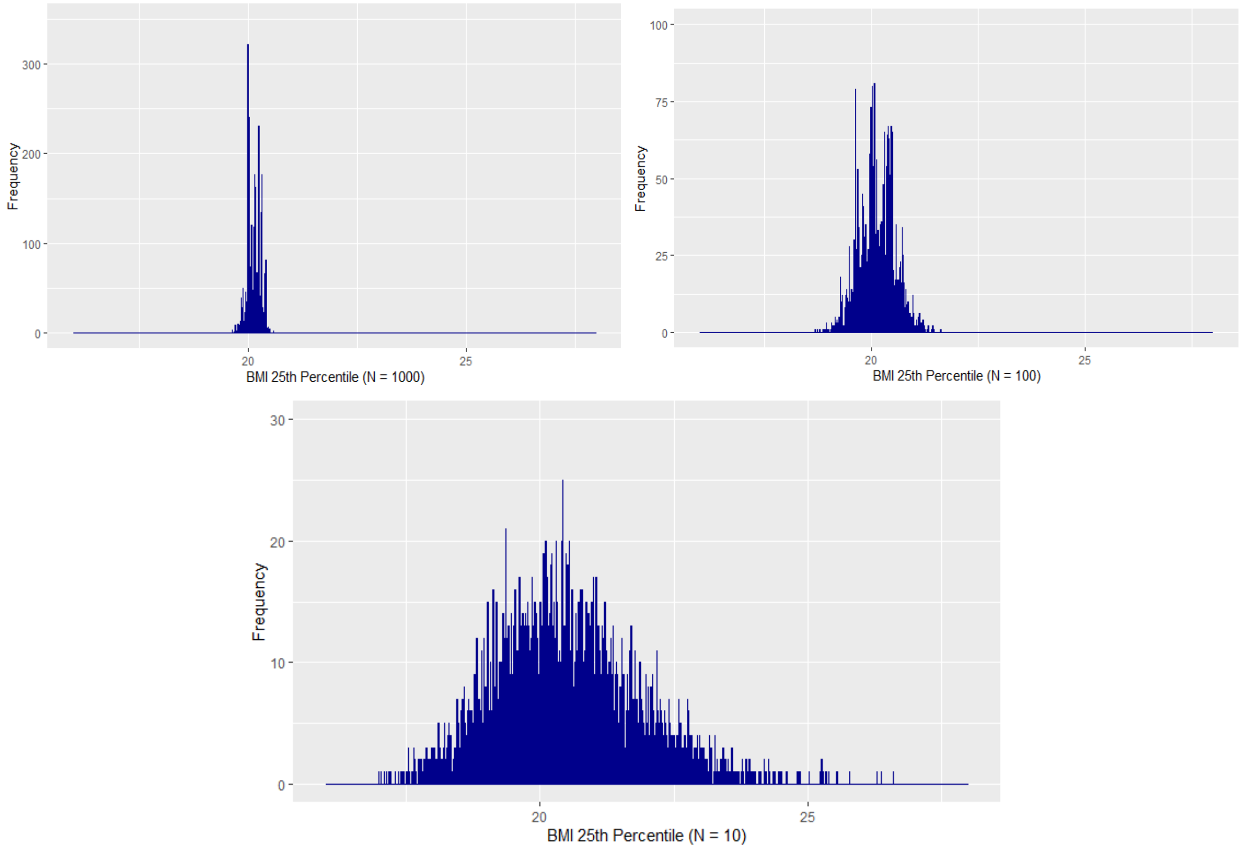

*Figure 2: Above are three histograms displaying the distribution of the 2017 student BMI repeated samples for n = 1000, 100 and 10 of the sample statistics **25th Percentile**.*

This sample statistic maintains a normal distribution as sample size increases, meeting the assumptions of the Central Limit Theorem (as applied in the context with the sample mean explanation above) with data centralizing around the true mean. Only, it is important to remember that in this case focused on only a portion of the dataset, narrowing it to the 25th percentile or the first quartile. The outcome for these distributions makes sense because the data were still independently collected and given these data are describing only the 25th Percentile, there is inherently less variability to begin with, represented with the initially narrow spread displayed on the Figure 2 histograms for n= 1000, and n= 100. The widespread data, n *=10, σ = 1.254*, still exhibits a large amount of variability. This makes sense, however, because the sample size is still small, notably less than n = 30, which is typically the standard for the application of the Central Limit Theorem.

The reported $\bar{x}$ for each distribution stay relatively stable; *n = 10, $\bar{x}$ = 20.456, and n = 100, $\bar{x}$ = 20.139 and n = 1000, $\bar{x}$ = 20.110*. Representative of a true estimate, particularly as sample size increases over n = 30.

As in Figure 1, with observing Figure 2 the distribution begins centralizing over the true estimate for each distribution. Meanwhile, the spread of the data decreases as represented with the standard deviations for each distribution. For example, with *n =10, σ = 1.254 and as sample size increases with n = 100, σ =0.406, until n=1000, σ = 0.139.* The distributions drastically narrow over the true mean (μ)., again.

Sample Statistic: Minimum

Additionally, we explore the year 2017 BMI of high school students, simulating repeated samples for increasing sample sizes (n=10, 100 and 1000), and calculating the *standard deviation* (σ) and *mean ($\bar{x}$)* for each distribution for the sample statistic **minimum**.

Minimum Visualization

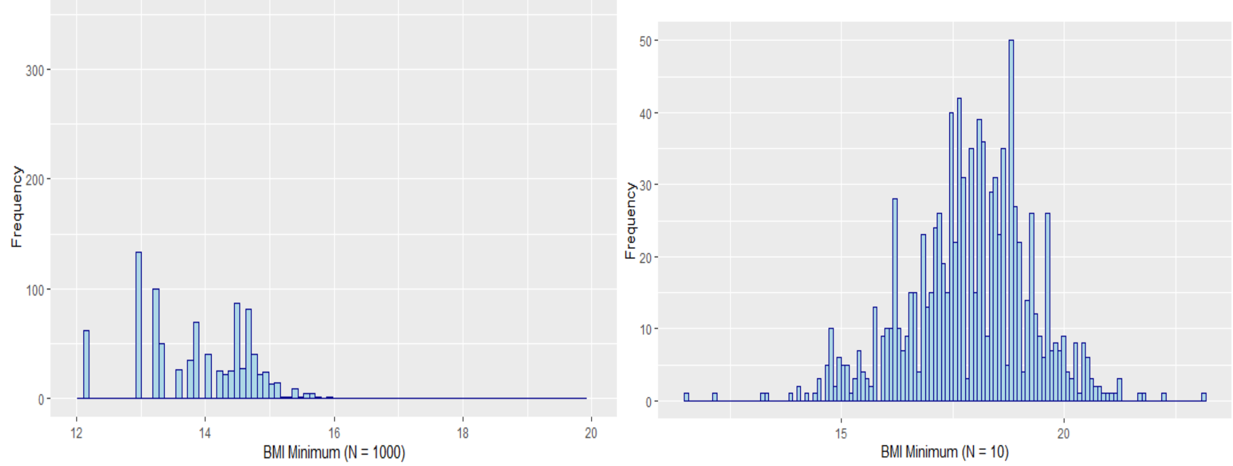

*Figure 3: Above are two histograms displaying the distribution of the 2017 student BMI repeated samples for n = 1000, and 10 of the sample statistics **Minimum.***

Figure 3 it can be observed that this sample statistic does not become more normal with increasing sample size. In fact, it appears the data exhibits a lack of centralization around a parameter at all, possibly because a minimum is typically a singular value for any given dataset. Applying a simulation to calculate this value **helps determine a limiting, lower range for these data or can be an indicator for significant outliers**. The Central Limit Theorem does not apply, but the values can help inform quartiles (such as the 25th percentile sample statistic).

The reported $\bar{x}$ for each distribution; *n = 10, $\bar{x}$ = 17.910, and n = 100, $\bar{x}$ = 15.624 and n = 1000, $\bar{x}$ = 13.742.*

Meanwhile, the variance (σ^2) is relatively unaffected. The spread of the data does not drastically respond, as reported with standard deviations for each distribution. For example, with *n =10, σ = 1.423 and as sample size increases with n = 100, σ =1.093, until n=1000, σ = 0.935*.

Sample Statistic: Difference in the Sample Median Years 2017 and 2007 Student BMI

Next, we explore the year 2017 BMI of high school students, simulating repeated samples for increasing sample sizes (n_1 = 5, n_2 = 5, and n_1 = 10, n_2 = 10, and n_1= 100, n_2= 100), then calculating the *standard deviation* (σ) and *mean* ($\bar{x}$) for each distribution for the sample statistic **difference in sample median for BMI years 2017 and 2007**.

Difference in Sample Median Visualization

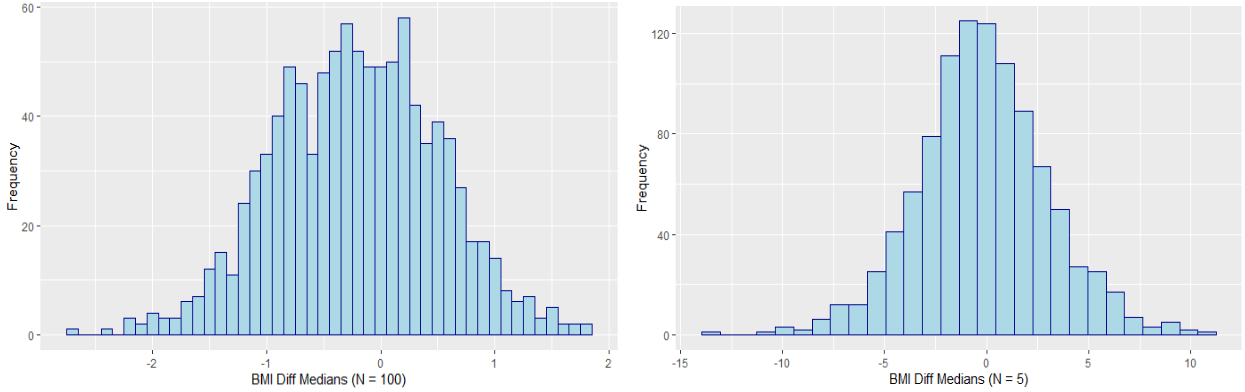

Figure 4: Above are two histograms displaying the distribution of the 2007 and 2017 student BMI repeated samples for n_1 *&* n_2 *= 100, compared to* n_1 *&* n_2 *= 5 of the sample statistics* ***difference in sample median.***

The reported *means* for each distribution; n_1& n_2 *= 5, $\bar{x}$ = -0.159, and* n_1 *&* n_2*= 10, $\bar{x}$ = -0.265 and* n_1 *&* n_2 *= 100, $\bar{x}$ = -0.196*. The negative values are acceptable for the means because the difference in medians can produce negative values. The *median* is the sample statistic describing the middle value of the higher and lower ranges for these data.

When observing Figure 4, as sample size increases, little change occurs in the spread of the data, or the variation (σ^2). This can be seen by considering the standard deviations for each distribution. For example, with n_1 *&* n_2 *=5, σ = 3.311 and as sample size increases with* n_1 *&* n_2 *= 10, σ = 2.257, until* n_1 *&* n_2*=100, σ = 2.257.* There is relatively little change in the variability (σ^2), although there is a general centralization over the true parameter as seen with the gradually decreasing spread between n_1 *&* n_2 *=5* and n_1 *&* n_2*=100.*

Summary of Results

In summary, we determined the following for the four sample statistics results of interest:

The **sample mean** responds accordingly to the Central Limit Theorem and the Law of Large Numbers. As sample size increases, the distribution centralizes around the true value parameter for the population of interest, in this case USA high school students, variation decreases and the spread narrows. The histograms for these data display normalized distributions. The sample mean value is meaningful for addressing questions regarding the larger population, as it has a true parameter, and assumptions have been met for test statistics that require independence, and normalized datasets.

The **25th Percentile** resembles what is seen with the sample mean. This in part is because it is essentially a subset of the same data. However, it is worth noting that the standard deviations are less drastically impacted with the increasing sample size, in part because quantiles, or "bucketing data" inherently reduces variation by selecting portions of the dataset above or below a set value. These data meet assumptions for test statistics requiring independence and normalized datasets.

The **minimum** does not appear normal, which makes sense because this is a unique value for any given dataset, and can be used to understand outliers, quantiles, and the range of data. The CLT and Law of Large Numbers does not apply to this sample statistic. The distribution cannot center around the minimum because there is not a possibility for getting a value lower than the true minimum. Therefore, the histogram is stratified (particularly for n=1000) and does not centralize around a true parameter value. This does not mean that the minimum is not of value, it simply indicates a separate use for this sample statistic.

The **difference in medians** has a normal distribution and relatively little change in variation and spread as the sample size increases. This is a product of the median's role in a relative dataset. It simply separates lower values from higher values and finding the difference can only explain a relative relationship between the years in terms of which year's distribution falls on a positive or negative side of a distribution curve. Arguably, the mode would be a more useful comparative statistic because it would be comparing the most common values between the datasets. There is no true difference in medians, and this test statistic does not meet any necessary assumptions for the data analysis steps of this report.

Task 2: Data Analysis

Now that we have evaluated the sample statistics of primary interest (*mean, 25th percentile, minimum and difference in medians*), we can explore the importance of these sample statistics in extrapolated applications, such as what these data are describing for the entire population of all USA high school students.

Answering the question, "**How has the BMI of high school students changed between 2007 and 2017? Are high schoolers getting more overweight?"**, these data can be compared using *two independent populations*: 1) high school students in 2007 and 2) high school students in 2017. We create a null hypothesis for these data and run a test statistic for rejection or acceptance of the null.

H_o: High schoolers are getting more overweight as time goes on between the years 2007 and 2017.

A *t-test* is performed, in this case the unpaired **Welch's two sample t-test**, and *confidence intervals* are constructed to describe the difference in *population means* (μ) for the BMI variable. Then, the t-test results for a statistically significant test statistic are applied to evaluate the null hypothesis for the question. There is statistically supported evidence (p-value=0.013) that there is no difference between mean values for either high school population (*mean_'17 = 23.776, and mean_'07 = 23.620*), with a confidence interval range of at the 5% level ranging from a high of 0.0325 to a low of 0.279. *Therefore, the null hypothesis is rejected.*

Simply, there is no statistical support that high school students are getting more overweight as time goes on between the years 2007 and 2017.

Answering the question, "**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?"** These data are approached as if there are *two independent populations*: 1) high school students in 2007 and 2) high school students in 2017. There is a null hypothesis for these data and a test statistic for rejection or acceptance of the null.

H_o: Year 2017, 12th graders are more likely than 9th graders to be "physically active" at least 60 minutes per day on 5 or more days.

A *two-sample proportions test* is preformed, as accompanied with binary observations of “true” and “false” values, and a difference in proportions across the populations is investigated. These data suggest that there is statistically significant evidence (p-value < $2.2e^{-16}$) that there is no difference in the level of physical activity at least 60 minutes per day on 5 or more days from year 2007 and 2017 12th graders when comparing the proportional values for *proportion 1* (0.382) and *proportion 2* (0.512). *Therefore, the null hypothesis is rejected.*

Simply, there is no statistically significant difference in 12th grader activity levels per day on 5 or more days for the USA high schooler population.

Answering the question: “**How much sleep do highschoolers get?”**, given these data are not numerical, and therefore no test statistic is needed to answer this question, instead, the reported values are a useful tool to visualize and evaluate differences across these data using histograms. One way of doing this would be by bucketing 25th, 50th and 75th percentiles, before drawing conclusions. The approach for this report, was simply plotting the data for 2017 versus 2007, and visually comparing. It is noted that the distributions are relatively similar in counts and quantity of reported hours of sleep per night. There is a slight up-tick in the number of students sleeping 6 hours per night for year 2017. One way to investigate discrepancies in the histogram heights would be to run a Wilcoxon test on numerically reported data, after first evaluating the visual distributions.

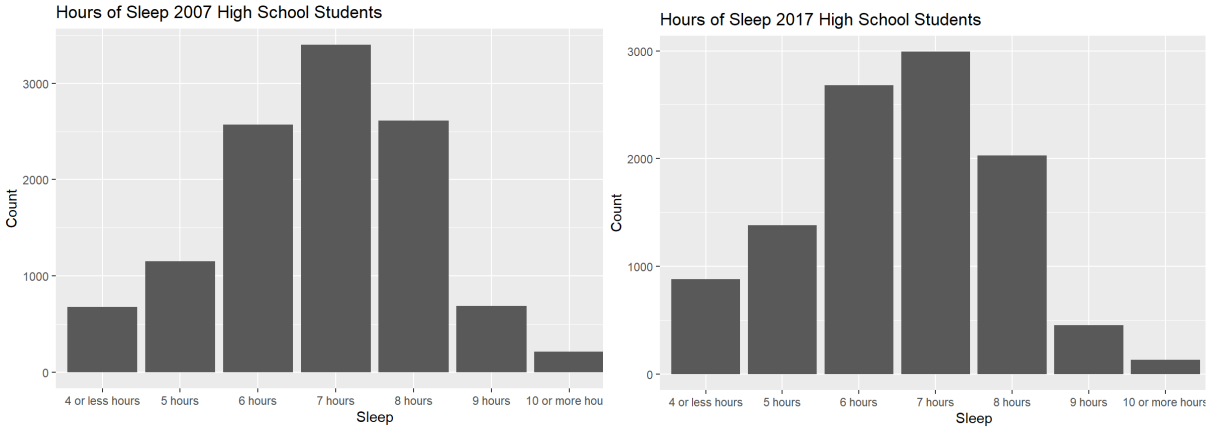

Figure 5: Two histograms displaying number of reported hours of sleep for high school students in years 2017 and 2007 from the YRBSS dataset.