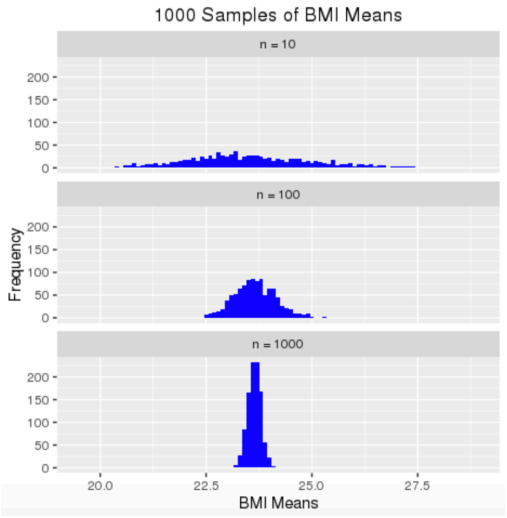
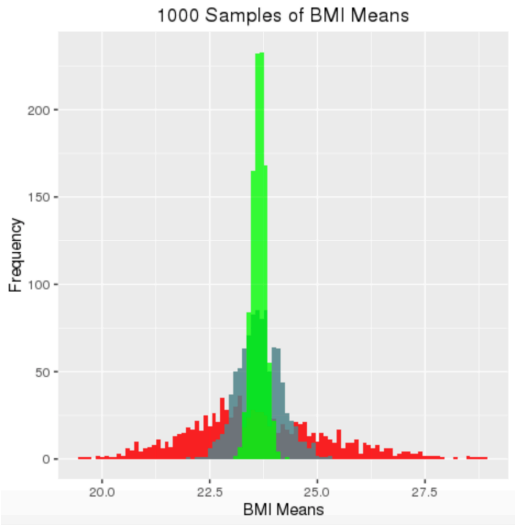
• Question 1 - Simulation Study

. (a) Using repeated samples of size n = 10, 100, 1000 from the BMI variable, describe the sampling distribution of the sample mean of BMI in 2013. 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.

Calculating the mean (n = 10, 100, 1000) from 1000 samples of the dataset, the standard deviation decreases from 1.63 to 0.496 to 0.159 as the sample size increases. This is shown in the merged graph with n = 10 as red, n = 100 as light blue, and n = 1000 as green. The resulting normal distribution is due to the Central Limit Theorem. The calculated mean is little changed showing the mean is unbiased to changes in sample size. As expected, standard deviation narrows in response to increases in sample size. By n = 1000 the calculated standard deviation and the true standard error agree at 0.159.



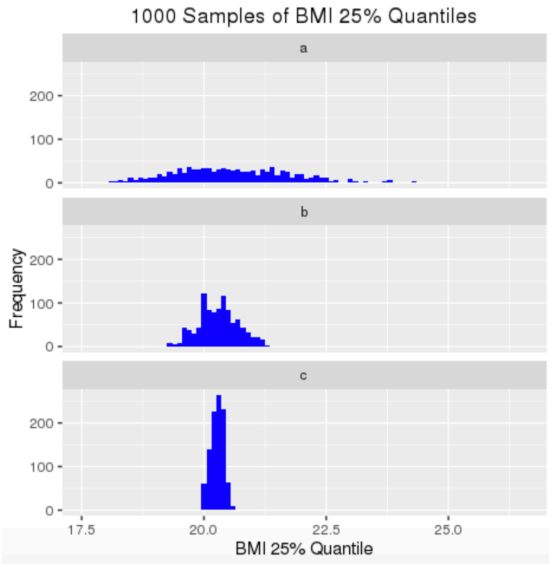


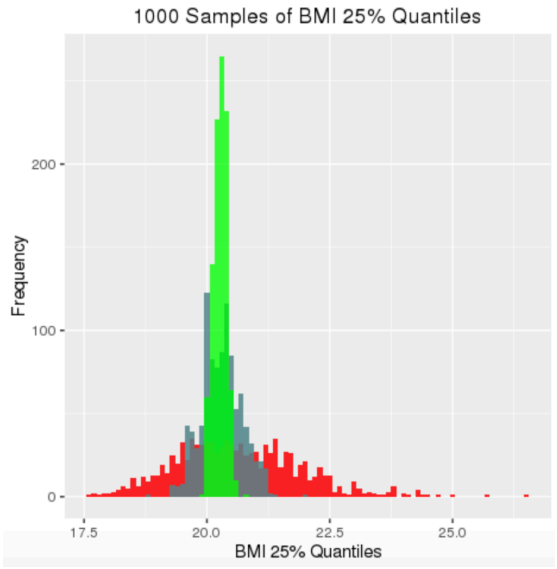
Summary findings for the BMI Means of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 samples	23.62	23.66	23.64
Calculated sd	1.625	0.496	0.159
True se	1.586	0.501	0.159

. (b) Repeat the simulation in part (a), but this time use the 25th percentile as the sample statistic. In R, the quantile() function will give you sample quantiles of a sample of data.

Calculating the 25% quantile (n = 10, 100, 1000) from 1000 samples of the dataset, the standard deviation decreases from 1.263 to 0.409 to 0.129 as the sample size increases from 10 to 100 to 1000 respectively. This is shown in the merged graph with n = 10 as red, n = 100 as light blue, and n = 1000 as green. The main difference when compared the to the means analysis above is the center of the data is shift ~23.6 to ~20.3. See results summary below.



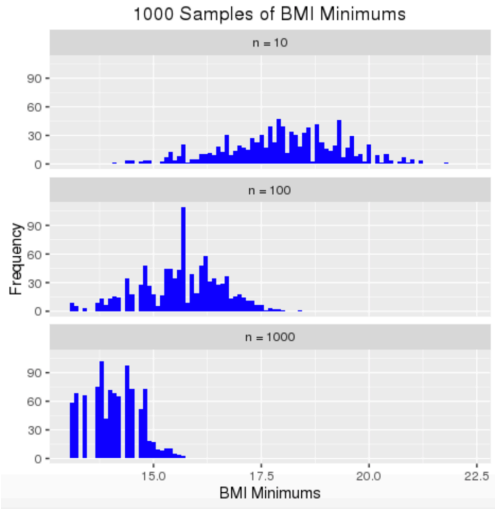


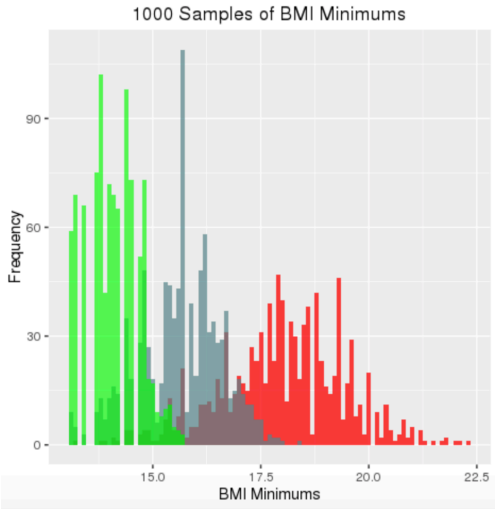
Summary findings for the BMI 25% Quantile of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 samples	20.63	20.29	20.28
Calculated sd	1.263	0.409	0.129

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

Determining the minimum (n = 10, 100, 1000) from 1000 samples of the dataset, the standard deviation decreases modestly from 1.426 to 0.944 to 0.593 as the sample size increases from 10 to 100 to 1000 respectively. This is shown in the merged graph with n = 10 as red, n = 100 as light blue, and n = 1000 as green. The main difference when compared to either the means or 25% quantile is that both the mean and standard deviation decrease as a function of samples size. Both the center and spread of the data are affected. The resulting distribution of minimums less normal or Gaussian when compared to the means or 25% quantile distribution. This is also due to the Central Limit Theorem because the minimum is not weighted by the rest of the dataset but rather is the most extreme data point for the one side. See results summary below.



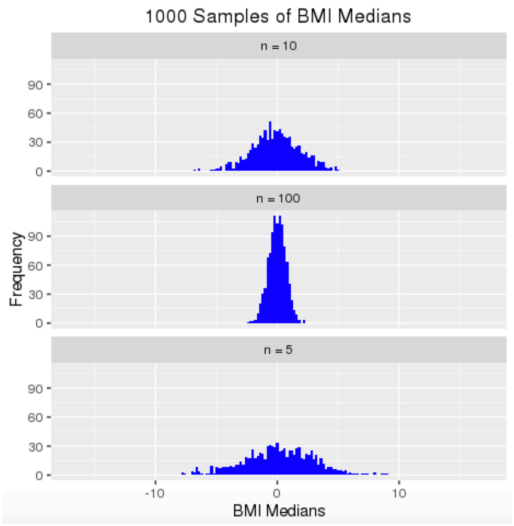


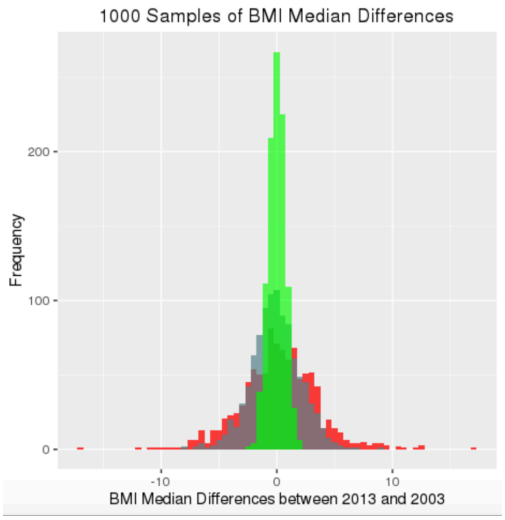
Summary findings for the BMI Minimums of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 samples	17.98	15.64	14.09
Calculated sd	1.426	0.944	0.593

. (d) Describe the sampling distribution of the difference in the sample median BMI between 2013 and 2003, 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.

Calculating the median (n = 5, 10, 100) from 1000 samples of the dataset, the standard deviation decreases from 3.367 to 2.127 to 0.709 as the sample size increases from 5 to 10 to 100 respectively. Additionally, the means of these sample sizes also decrease with increases in sample size. That is, mean is unbiased to sample size while standard deviation is not. Notice in the first plot n = 100 is the center panel which shows the narrowest distribution. Also note the magnitude of these differences is quite small (near zero). The shape of the median distributions are normal or Gaussian.



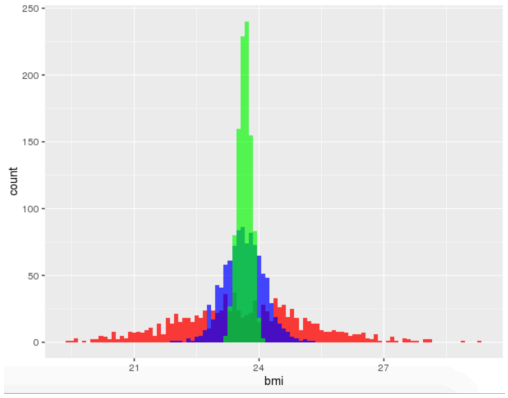


Summary findings for BMI Median Differences of Students from 2003 to 2013

	n = 5	n = 10	n = 100
Mean of 1000 samples	0.173	0.099	0.013
Calculated sd	3.367	2.127	0.709

. (e) Summarize your results.

It's clear from the analysis above that measures of center of data, mean and median, are relatively unbiased to changes in sample size. The measures of center, mean and median, change little in response to increases in sample size. As expected, a key measure of sample spread, standard deviation, decrease with increases in sample size. As shown above, standard deviation narrows or decreases significantly as sample size increases. Data spread and sample size have an inverse relationship. An example plot is shown again here to illustrate how the center of the distribution does not change while the spread does. The shape of the distribution is normal and goes from broad and flat to narrowly distributed.



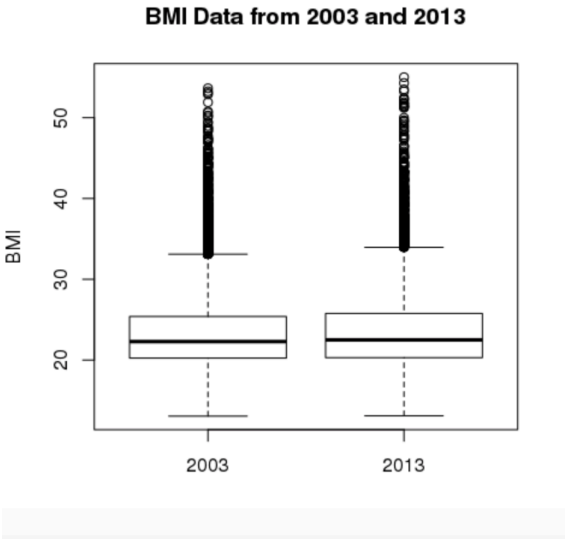
• Question 2 - Data Analysis

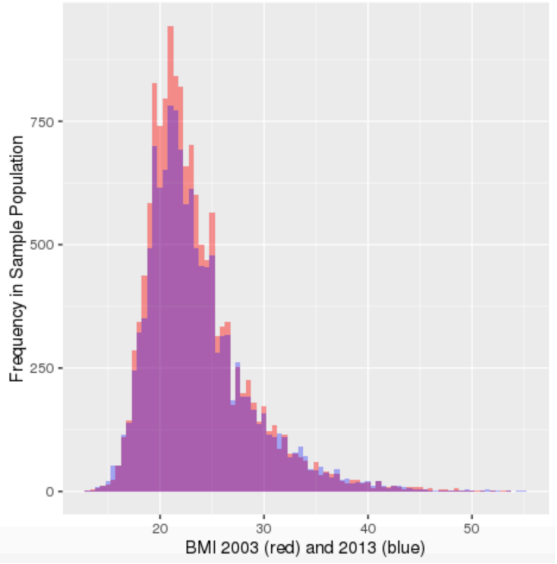
For this part of your assignment your task is to analyze the data to answer the questions of interest. Your solution must include a non-technical summary of your findings. Using the same data as the Simulation Study, but now treating the YRBSS as a sample from the population of all USA high-school students:

The Youth Risk Behavior Surveillance System (YRBSS) was developed in 1990 to monitor health risk behaviors of students. Behaviors included are: those that contribute to injury/violence, sexual behaviors, alcohol and tobacco use, and physical activity. For 2003 and 2013 that data sets are quite large (14057 observations for 2003 and 12580 observations for 2013). The schools taking the survey self-selected. Also, students at these schools were not presumably not taken at random but rather directed by the school's administration to take the survey. Further, this also means that the data cannot be assumed to be completely independent since there are likely some geographical proximity effects. A randomized study sample could be taken and compared to the results here to estimate the size of the geographical proximity effects.

- How has the BMI of high-school students changed between 2003 and 2013? Are high-schoolers getting more overweight?

There is statistical evidence that the average BMI of students from 2003 to 2013 has increased slightly (Welch Two Sample t-test, 95% Confidence Interval, p-value 0.000175, df = 25990). With 95% confidence the mean BMI in 2013 is between 0.108 and 0.344 units larger than in 2003. This difference, while statistically significant, may be of little practical importance because the average increase (0.23 on the BMI scale) is small. Indeed, visually the BMI scores from 2003 and 2013 are quite similar. Students from 2003 to 2013 are getting more overweight though the increase, as measured by BMI, is small. Further the differences from 2003 to 2013 may be exaggerated due to dependence. Since schools self selected there may be some school-to-school or region-to-region differences in nutrition and/or economic status that may exaggerate the above conclusion.





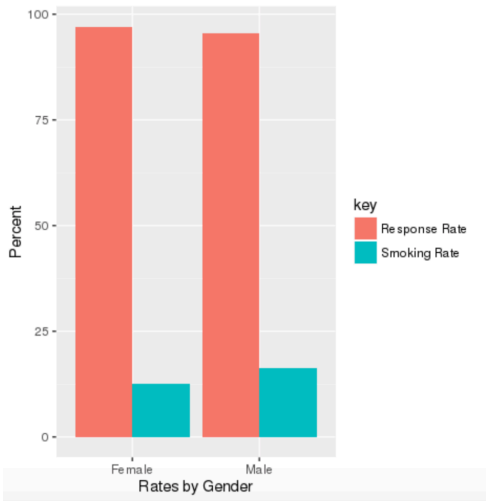
Summary findings for the BMI of students from 2003 and 2013

Year	Mean	Median	Standard Deviation
2003	23.42	22.29	4.784
2013	23.64	22.49	5.014

- Are male high-schoolers more likely to smoke than female high-schoolers in 2013?

There is convincing evidence that male students smoke at a higher proportion than do female students (Two-sample Proportion test with continuity correction, $p = 7.14e^{-5}$, $n_1 = 6128$, $n_2 = 5983$). In 2013, 16.3% of male students and 12.6% of female students reported smoking in the past 30 days. With 95% confidence male smokers is between 15.4% to 17.2% and female smokers is between 11.8% to 13.5%. With 95% confidence the difference in smoking rates between males and females is 2.5% to 5.0%.

One assumption in the higher rate of smoking among male students is that both genders respond honestly to this question at the same rate. There may be social pressures that cause male or female students to respond honestly at different rates. Another assumption is that the data are independent which may not be case with smoking. Peer pressure and other behavior influences may skew smoking rates at one school versus another.



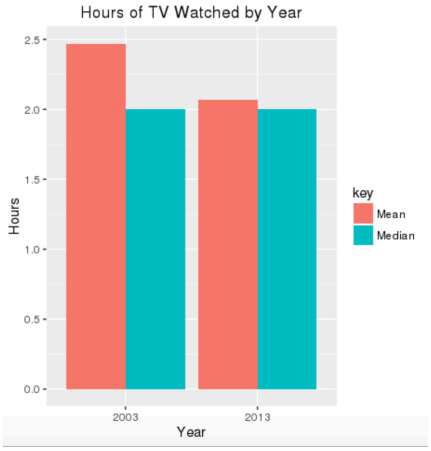
Male students had a response rate of 95.5% while female students responded at a rate of 97.0% to this question regarding smoking. This disparity of response rate may affect the above conclusion. Further, schools self-selected to participate and likely directed students to participate rather than having randomly selected students from all schools. The difference in male versus female smoking rates may be affected due to lack of independence.

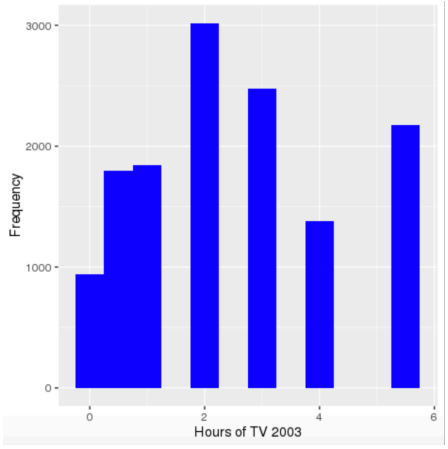
Summary of Smoking Rates by Gender for 2013

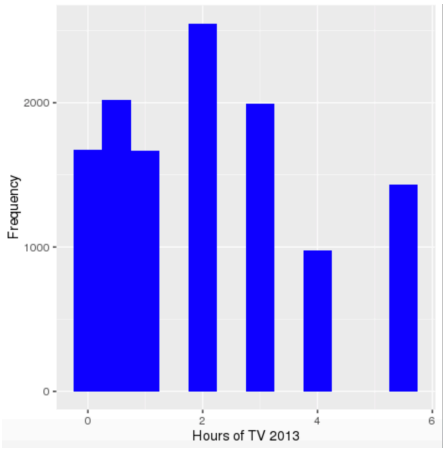
Gender	Responded	Did Not Respond	Response Rate	Reported Smoking in Past 30 Days
Male	6128	286	95.5%	16.3%
Female	5983	183	97.0%	12.6%

- How much TV do high-schoolers watch?

  The median hours of TV watched in both 2003 and 2013 is 2 (constructed 95% confidence interval 2, 2). Median is a much better estimate than mean because there is no good way to quantify a value for the categories 'Less than 1' and '5 or more.' One estimate of mean is to score 'Less than 1' as 0.5 hours and '5 or more' as 5.5 hours. The resulting mean and median hours of TV watched in 2003 and 2013 are summarized in the table below. Both 2003 and 2013 have similar response rates for this question (97.0 and 97.8 % respectively). As noted above, the data are not from a simple random sample but the sample sizes are large. Also, there may be some dependence in the data collected since schools self selected to participate. Schools are likely to come from different economic regions which may skew the amount of TV watched per day.







Summary of Hours of TV Watched by Students in 2003 and 2013

Year	Mean	Median	Confidence Interval	Response Rate
2003	2.47 (2.44, 2.50)	2 (2, 2)	95%	97.0%
2013	2.07 (2.04, 2.10)	2 (2, 2)	95%	97.8%