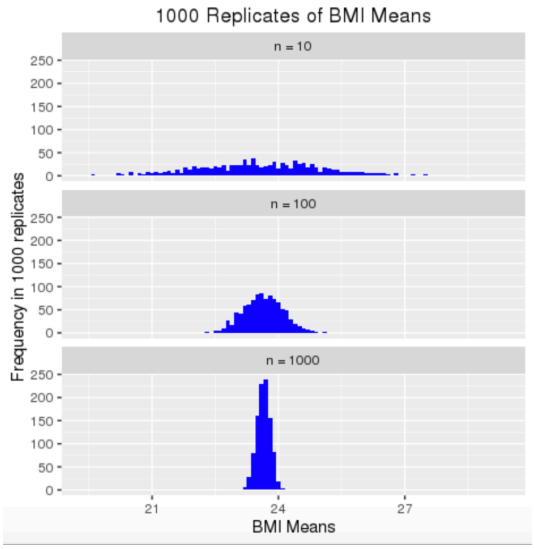
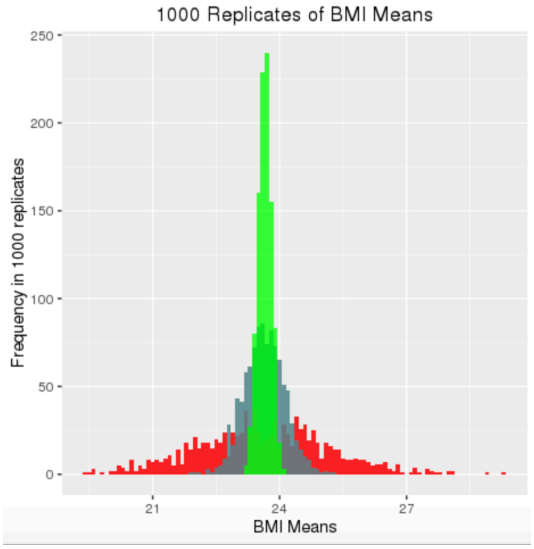
• Question 1 - Simulation Study

. (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.

The standard deviation calculated from 1000 replicates decreases significantly from 1.577 to 0.482 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. This is due to the Central Limit Theorem. The calculated mean is little changed showing the mean is robust to changes in sample size. Conversely, standard deviation narrows greatly in response to increases in sample size. By n = 1000 the calculated standard deviation and the standard error agree at 0.159.



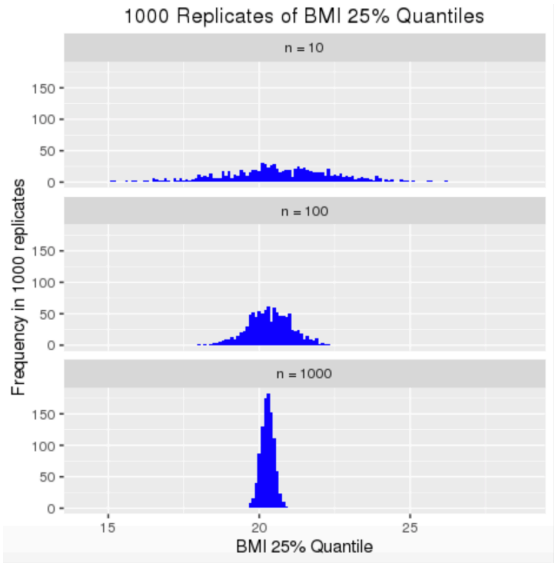


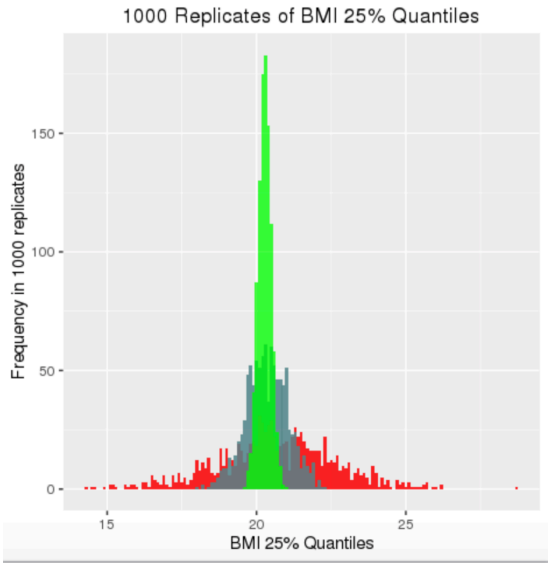
Summary findings for the BMI Means of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 replicates	23.61	23.64	23.65
Calculated sd	1.577	0.482	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.

The standard deviation calculated from 1000 replicates decreases significantly from 2.01 to 0.711 to 0.214 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.



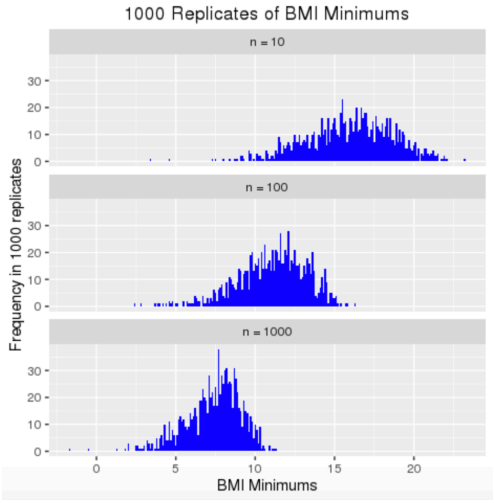


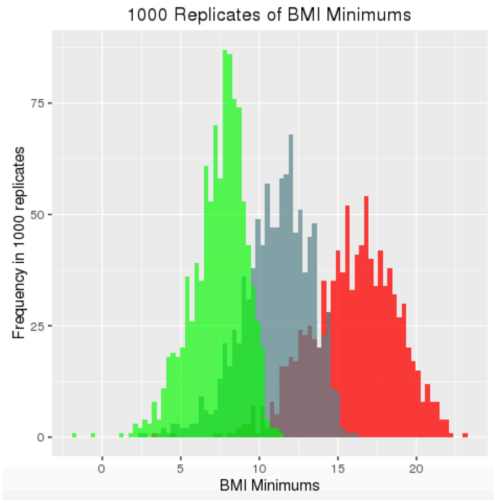
Summary findings for the BMI 25% Quantile of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 replicates	20.7	20.3	20.3
Calculated sd	2.01	0.711	0.214

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

The standard deviation calculated from 1000 replicates decreases modestly from 2.768 to 2.153 to 1.713 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 the mean as a function of sample size decreases while standard deviation shows little change. See results summary below.



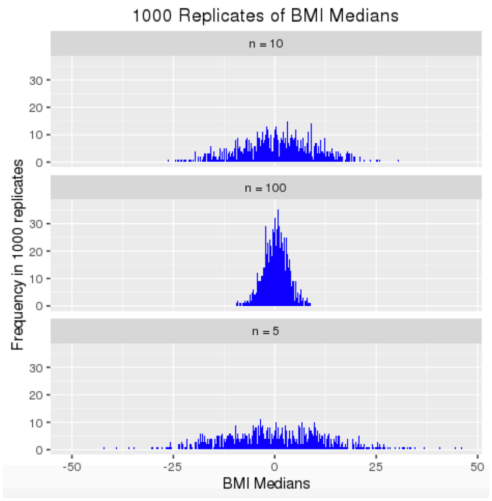


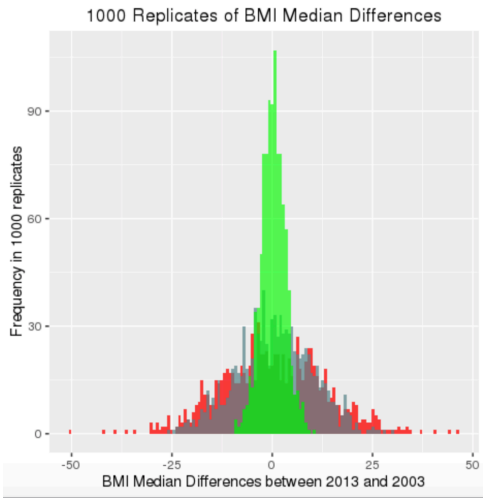
Summary findings for the BMI Minimums of Students from 2013

	bmi10 (n = 10)	bmi100 (n = 100)	bmi1000 (n = 1000)
Mean of 1000 replicates	15.9	11.1	7.39
Calculated sd	2.758	2.153	1.713

. (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.

The standard deviation calculated from 1000 replicates decreases modestly from 12.9 to 9.42 to 3.00 as the sample size increases from 5 to 10 to 100 respectively. Conversely, the means of these sample sizes does not change much. That is, mean is robust 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)



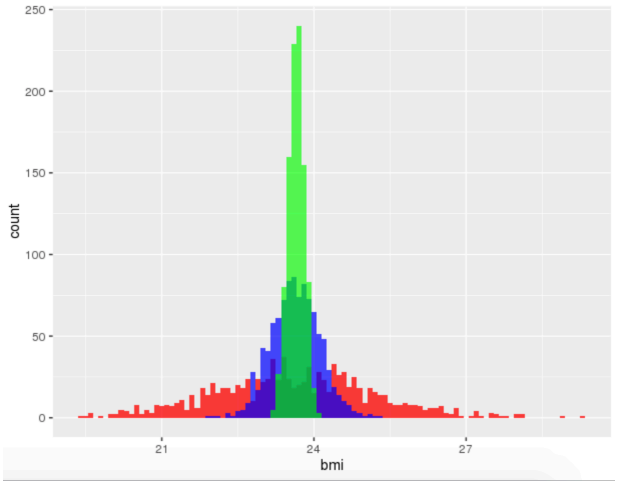


Summary findings for the BMI Medians Differences of Students from 2003 to 2013

	n = 5	n = 10	n = 100
Mean of 1000 replicates	0.278	0.548	0.216
Calculated sd	12.9	9.42	3.00

. (e) Summarize your results.

It's clear from the analysis above that measures of center of data, mean and median, are quite robust to changes in sample size. The measures of center change little in response to increases in sample size. Conversely, a key measure of sample spread, standard deviation, are quite sensitive to increases in sample size. As shown above, standard deviation narrows or decreases significantly as sample size increases. Data spread and sample size have an inversely proportional relationship. An example plot is shown again here to illustrate how the center of the distribution does not change while the spread does.



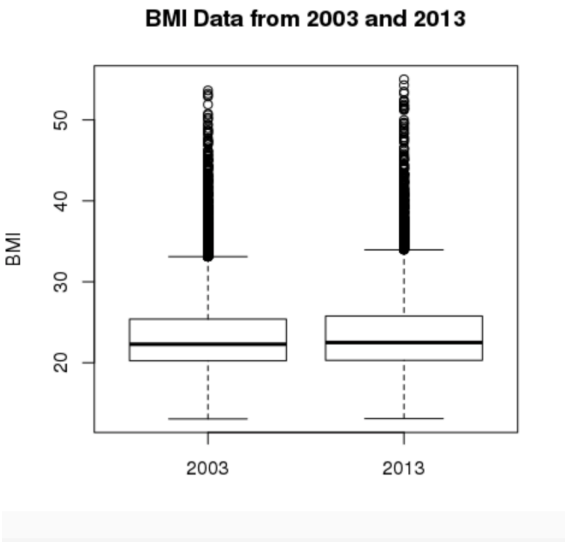
• Question 2 - Data Analysis

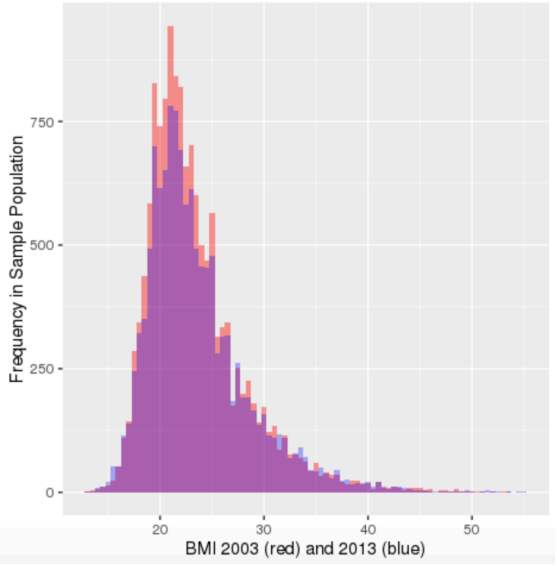
For each part, a translation of the questions of interest into inferential questions about parameters in a statistical model and a description of the method(s) you will use to answer the questions of interest. You do not need to have completed any of the analysis.

Because the sample sizes are large (14057 observations for yrbss of 2003 and 12580 observations of yrbss of 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 and possible temporal proximity effects.

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

There is some evidence that the average BMI of students from 2003 to 2013 has increased slightly (Welch Two Sample t-test, 95% Confidence Interval -3.44 to -1.03, p-value 0.000175, df = 25990). This difference, while statistically significant, may be of little practical significance because the average increase (0.23 on the BMI scale) is small. Indeed, visually the BMI scores from 2003 and 2013 are quite similar. Yes, from 2003 to 2013 students 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 or economic status the may exaggerate the above conclusion.



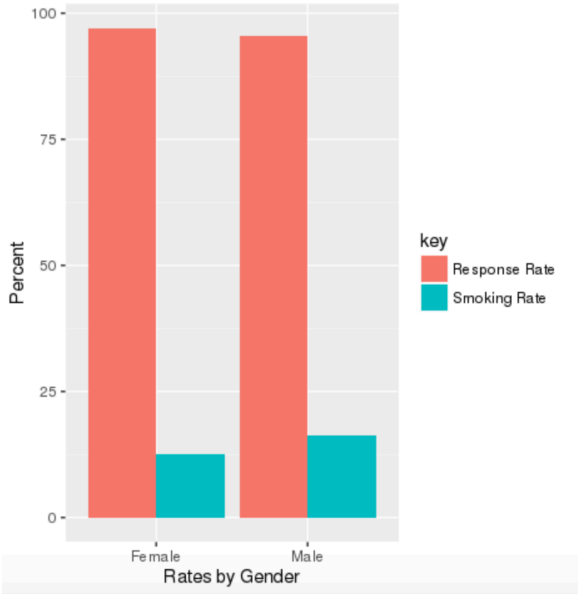


Summary findings for the BMI of students from 2003 and 2013

Year	Mean	Median	Variance
2003	23.4	22.3	22.9
2013	23.6	22.5	25.1

- 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 (p-value = 7.14e-5 for a two-sided Proportion test, $n_1 = 6128$, $n_2 = 5983$). A 95% Confidence Interval 16.3% (2.45, 5.00) of male respondents reported smoking in the past 30 days while 12.6% of female respondents did the same. 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. There is also the possibility that male students at a school influence (e.g. via peer pressure) female students or vice versa meaning the data here may not be independent. 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 exaggerated here due to dependence.

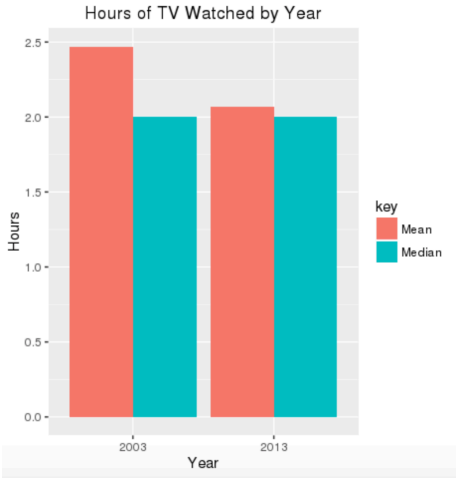


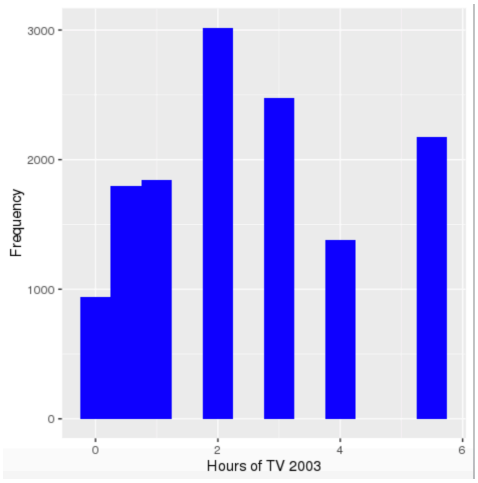
Summary of Smoking Rates by Gender for 2013

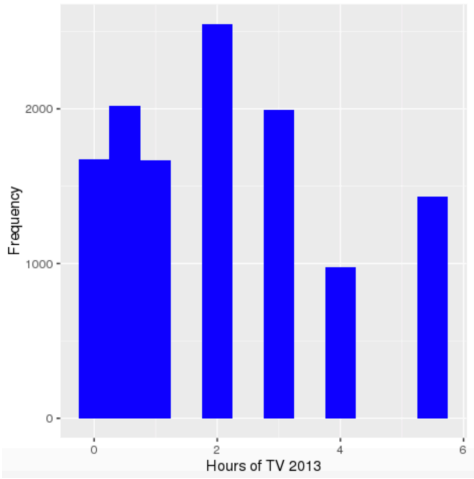
Gender	Responded	Did Not Respond	Response Rate	Reported Smoking in Past 30 Days
Male	6128	286	95.5%	16.3%
Female	5983	183	97.0%	12.6%

- How much TV do high-schoolers watch?

The median hours of TV watched in both 2003 and 2013 is 2 (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 randomized study 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%