Philip Ourso

ST516 Module 10: Simulation study and data analysis

1. Simulation study
    a. Investigation of the sampling distribution of the sample mean of BMI in 2013 and the effect of increasing sample size.
    *Please see R code in accompanying .R script.*

| | N=10 | N=100 | N=1000 |
|---|---|---|---|
| Mean | 23.642 | 23.642 | 23.643 |
| Standard deviation | 1.563 | 0.498 | 0.151 |

*Table 1 Sample mean averages and standard deviations*

As can be seen in the table above, the mean of the sampling distribution is fairly stable as sample size increases from 10 to 100 to 1000, but the standard deviation dramatically decreases with increasing sample size. This is illustrated in the histogram overlay below, wherein the means and mean ± standard deviation is indicated by dashed and dotted lines of the corresponding color.

In the below plot of overlaid histograms, blue represents the lowest sample size, 10, red 100 and green 1000. It is obvious that the distributions are centered about a similar value, while the standard deviation, and spread, decreases as sample size increases, consistent with the Central Limit Theorem. Some rightward skew is evident, particularly with a sample size of 10, and this is intuitive, as the human body can likely support a BMI further above the mean (heavier for a given height) than below.

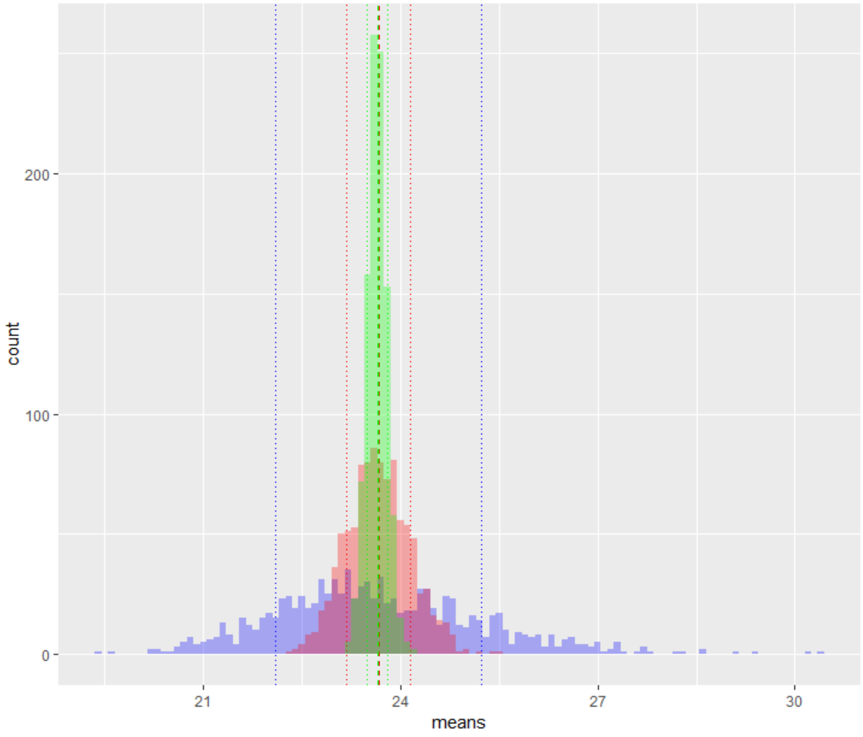

*Figure 1 Histograms of sampling distributions of sample mean*

b. Investigation of the sampling distribution of the sample 25th percentile of BMI in 2013 and the effect of increasing sample size.

| | N=10 | N=100 | N=1000 |
|---|---|---|---|
| Mean | 20.629 | 20.301 | 20.274 |
| Standard deviation | 1.308 | 0.407 | 0.128 |

*Table 2 Sample quantile averages and standard deviations*

As can be seen from the table values, the 25th percentile of the sampling distribution is fairly stable (although not as stable as the sample mean) as sample size increases from 10 to 100 to 1000, but the standard deviation again dramatically decreases with increasing sample size. This is illustrated in the histogram overlay, where the spread of the distribution is clearly decreasing as the sample size increases, but are still roughly centered about the same value. The same rightward skew is again evident, and as a whole, the distributions are similar to the sample mean distribution, although centered about a lower value.

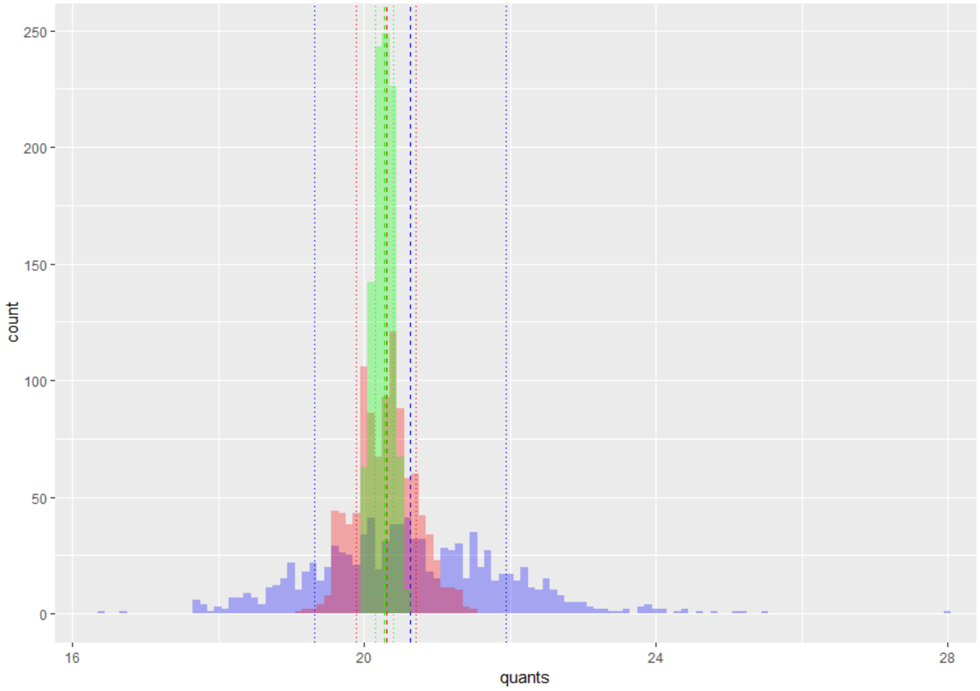

*Figure 2 Histograms of sampling distributions of sample quantile*

c. Investigation of the sampling distribution of the sample minimum of BMI in 2013 and the effect of increasing sample size.

| | N=10 | N=100 | N=1000 |
|---|---|---|---|
| Mean | 18.091 | 15.63 | 14.040 |
| Standard deviation | 1.480 | 0.96 | 0.581 |

*Table 3 Sample minimum averages and standard deviations*

In the case of the sample minimum, the standard deviation continues the expected trend: the values decrease with increasing sample size, and hence spread decreases. However, in the case of the summary statistic, the mean value decreases significantly. This is visualized in the histogram overlay as the distributions are no longer centered about the same approximate value. Furthermore, the distribution corresponding to the largest sample size (green, n = 1000) is truncated. This should be expected, as we'd expect the summary statistic to approach the population minimum as sample size increases.

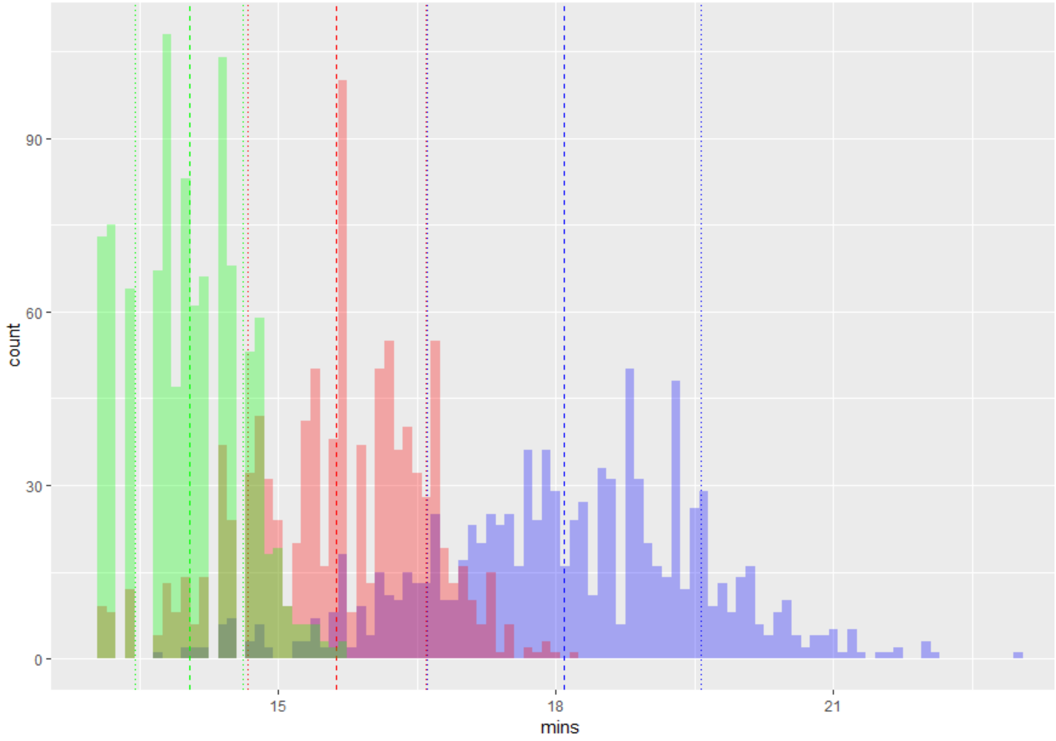

*Figure 3 Histograms of sampling distributions of sample minimum*

d. Investigation of the sampling distribution of the sample median of BMI in 2013 and the effect of increasing sample size.

| | $N_1$, $N_2$ = 5 | $N_1$, $N_2$ = 10 | $N_1$, $N_2$ = 100 |
|---|---|---|---|
| Mean | 0.237 | 0.131 | 0.181 |
| Standard deviation | 3.115 | 2.122 | 0.686 |

*Table 4 Sample median difference averages and standard deviations*

The difference in median BMI continues to exhibit the expected trend in standard deviation, decreasing as sample size increases. The summary statistic is approximately normal and roughly centered around the same value, in this case close to zero. This is to be expected, as the samples are drawn from the same population, and hence should reflect a similar central value. Furthermore, given that the median is simply the 50th percentile, a single median would be expected to be approximately normal. The histogram overlay suggests that the difference in normal distributions is also approximately normal. There continues to be some outliers in the upper extremes of the distributions, consistent with above observations of high BMI values.

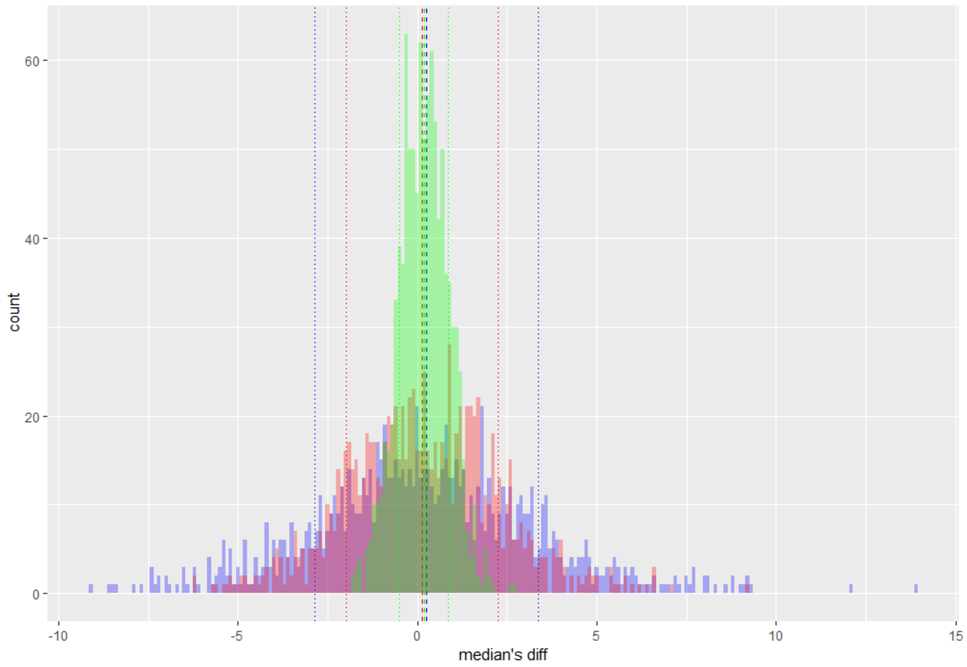

*Figure 4 Histograms of sampling distributions of sample median difference*

e. Comparison and summary of simulation studies.

In all cases of the simulation study, the standard deviations of the summary statistic decreases with increasing sample size. For the sample mean, this is understood from the Central Limit Theorem as the variance of the sample distribution is equal to the population variance divided by the sample size. Hence, as sample size increases, the variance decreases.

The sampling distribution itself is apparently normal for the sample mean, again consistent with the CLT. This appears to be true of the 0.25 quantile as well, although for seemingly more advanced reasons not laid out thus far in the coursework. The difference in medians also exhibits an approximately normal distribution. This is essentially a difference in 50th percentiles, which are normal, and hence is a difference in normal distributions.

The sample minimum, however, doesn't exhibit the same general behavior as the other summary statistics: as sample size is increased, a greater range of values is included in the sample, and hence the probability of a lower minimum increases. This is evident by the leftward shift of the mean sample minimum (towards zero).

2. Data analysis

a. How has BMI changed of high-schoolers changed between 2003 and 2013? Are high-schoolers becoming more overweight?

To answer this question, the data is already in acceptable form: a continuous numeric of type double. An appropriate analysis here would be to compare two population means, those of the 2003 and 2013 data. A one-sided, two-sample t-test could be performed on our summary statistic, mean BMI, to determine if it had appreciably increased from 2003 to 2013, or a two-sided analysis could be performed to detect a potential difference.

As a first step, a simple inspection of the data can help guide the subsequent analysis. In R, calls to head() give an idea of data frame contents and example values. Using the qplot() function from the ggplot2 library reveals the sample distributions for the 2003 and 2013 BMI data. Looking at the below plots it is evident that the distributions are skewed to the right, given the prevalence of observations with high BMI. This is intuitive, as there is presumably a more limited range of possible values on the lower end of the distribution, where there is some hard limit for the minimum BMI.

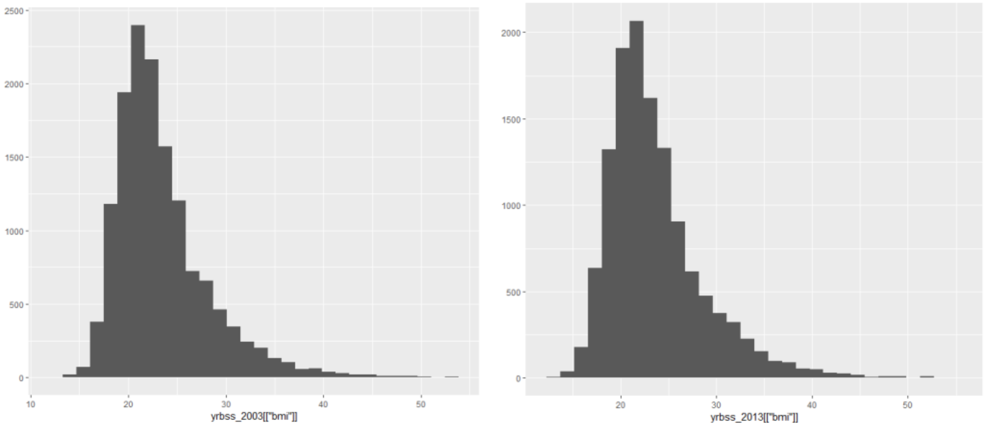

*Figure 5 2003 and 2013 High-school students' BMI*

In order to compare the distributions and infer to the population, it is necessary to choose a summary statistic. For this question, comparing mean BMI can tell us whether the average BMI is increasing, decreasing, or staying the same between the two distributions. The point estimate for the summary statistic is the sample mean.

Given the summary statistic of the population mean, an obvious statistical procedure is the two-sample t-test. Since the observations do not map 1:1 between samples, a paired test is not appropriate. The data is presumably random; if this assumption is violated the results would potentially be invalidated. The data is considered independent, as it's easily less than 10% of the general population of high-schoolers. Even though the data appear to be skewed, the t-test is robust to reasonable skew given the large sample size.

In the case of the one-sided, two-sample t-test, the null and alternate hypotheses are as follows:

$H_0$ : there is no difference in the population means of high-schoolers' BMI
$H_a$ : the difference in population means (2013 vs 2003) is greater than 0

Using t.test in R produces the following output:

```
Welch Two Sample t-test

data:  yrbss_2013[["bmi"]] and yrbss_2003[["bmi"]]
```

```
t = 3.7529, df = 25988, p-value = 8.758e-05
alternative hypothesis: true difference in means is
greater than 0
95 percent confidence interval:
 0.1269528       Inf
sample estimates:
mean of x mean of y
 23.64326  23.41725
```

The p-value is extraordinarily small, giving us confidence that there is a statistical difference in the two populations' means. Furthermore, the 0.95 confidence interval is positive, giving us further confidence in our below conclusion.

*There is convincing evidence that the difference in mean BMI of 2003 and 2013 high-schoolers is not zero (Welch's two-sample t-test, p-val = 8.76e-5). With 95% confidence, we estimate that the mean BMI has increased more than 0.127 from 2003 to 2013.*

It should be noted that the lower bound of the confidence interval, 0.127, and the difference in point estimates, 0.22, are both less than 1% of the 2013 BMI. So, despite statistical significance, the increase in BMI might not have practical significance. As an example, for a 5'10" person weighing 160 lb, an increase in their BMI consistent with the observed trend would be equivalent to gaining roughly 2 lb. The practical significance of this would depend on the desired goal of the analysis.

b. Are male high-schoolers more likely to smoke than female high-schoolers, given 2013 data?

There are a few ways to interpret this question, which have implications for the choice of a test statistic and an analysis procedure. Two ways to interpret "likely to smoke" are as follows:

1. Classification of high-schoolers as smokers and non-smokers, and interpretation of "likely to smoke" based upon this binomial distribution.
2. Quantification of frequency of smoking, and comparison of counts of these "bins". Higher frequency of past smoking would then indicate a greater likelihood of smoking.

Approach #1 requires the transformation of the data into two classes: smokers and non-smokers, with the below definitions:

- non-smokers: didn't smoke in the last 30 days
- smokers: smoked at least once in the last 30 days

This results in a binomial distribution that can be compared on the basis of proportions. The below bar charts indicate these distributions for both genders.

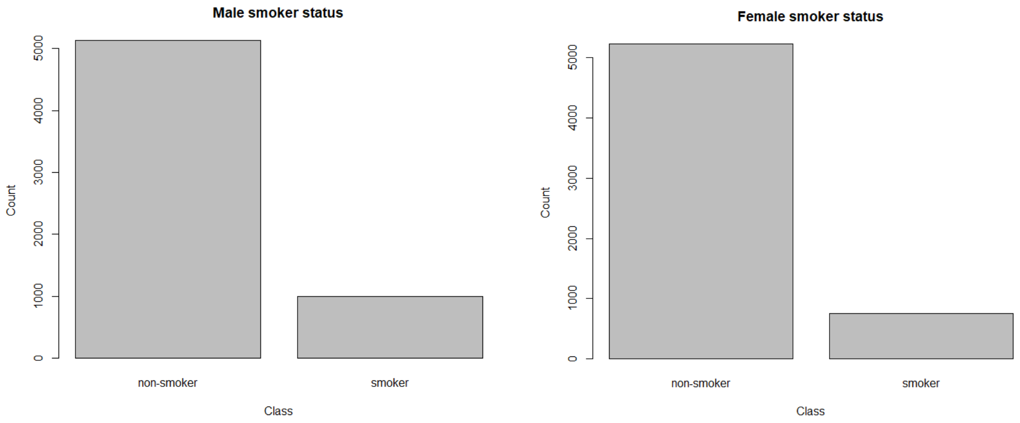

*Figure 6 Bar charts of male and female smoker status*

In this case, tabular data of percentages is clearer, and more efficiently communicates smoker status:

| | Non-smoker | Smoker |
|---|---|---|
| Male | 0.837 | 0.163 |
| Female | 0.874 | 0.126 |

*Table 5 Percentage of male and female non-smokers and smokers*

As can be seen above, the percentages are similar, but with fewer female smokers.

Given proportional data, a two-sample test for equality of proportions is conducted, the output of which follows the statement of null and alternate hypotheses.

$H_0$ : there is no difference in the population proportions of male and female high-school smokers

$H_a$ : is a non-zero difference in the population proportions of male and female high-school smokers

```
2-sample test for equality of proportions without
continuity correction

data:  sm out of total
X-squared = 33.795, df = 1, p-value = 6.125e-09
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02466874 0.04966246
sample estimates:
   prop 1    prop 2
0.1630222 0.1258566
```

Again, the p-value is quite small, and the 0.95 confidence interval indicates a range of 2.5 to 5%, leading to a reasonably strong conclusion below.

*There is convincing evidence that the difference in proportions of male and female high school smokers in 2013 is not zero (two-sample test for equal proportions, p-val = 6.13e-9), and hence infer that male high-schoolers are more likely to smoke than* *female high-schoolers. With 95% confidence, we estimate that the difference in proportions is (0.025, 0.05), a practical difference of between 2.5 and 5%.*

In this case, the results appear to be both statistically and practically significant: given the notable health risks associated with smoking, 2.5 to 5% difference might justify a change in the approach of an awareness campaign regarding the ill effects of smoking, for example. Furthermore, it might prompt a follow-on study to determine if this gap appears to be increasing or decreasing with respect to 2003 data.

A second, albeit slightly more involved approach would be to convert the categorical data to ordinal data, with the new data indicating increasing frequency of smoking.

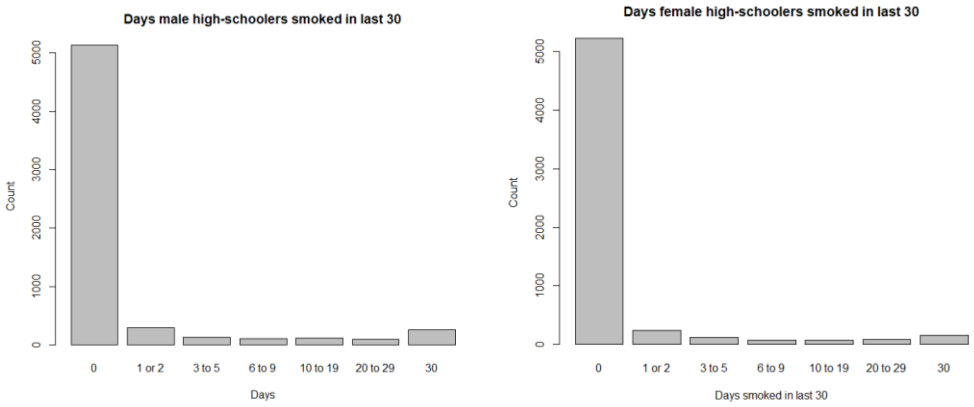

*Figure 7 Bar charts of male and female frequency of smoking*

Once again, tabular display of proportions allows a more detailed comparison, although the number of columns makes the comparison less straightforward.

| | 0 | 1 or 2 | 3 to 5 | 6 to 9 | 10 to 19 | 20 to 29 | 30 |
|---|---|---|---|---|---|---|---|
| Male | 0.837 | 0.048 | 0.021 | 0.017 | 0.019 | 0.016 | 0.042 |
| Female | 0.874 | 0.041 | 0.020 | 0.012 | 0.012 | 0.014 | 0.027 |

*Table 6 Percentages of male and female frequency of smoking*

A Chi-squared test for given probabilities can be performed to compare the male proportions to female proportions (termed "expected probabilities") to determine the likelihood of equivalence between the two populations, the results of which follows, however this does not answer the original question but rather simply indicates a difference in populations.

```
   Chi-squared test for given probabilities

data:  table(x = yrbss_2013[yrbss_2013$sex == "Male",
"smoke"])
X-squared = 103.96, df = 6, p-value < 2.2e-16
```

While the result above indicates a difference, the first approach more accurately answers the question posed, subject to the stated definitions of "smoker" and "non-smoker".

c. How much TV do high-schoolers watch?

Accurately answering the question in this case would require reporting some number that indicated amount of television watched by 2013 high-schoolers on an average school night, likely in hours. The data, however, are categorical in nature, and indicate ranges of television watched.

The data could be transformed into a form more easily analyzed, the most obvious transformation being the replacement of categories with hours, in integer form. There are a few issue with this approach, however.

The responses given represent ranges, and replacing them with integers can hide this fact, and lead to an approach analyzing the mean, which would have a loss of accuracy, given that watching 1 hour and 50 minutes of television might result in a response of 1 or 2 hours per day. Furthermore, the extreme values would need to be handled:

- "Less than 1 hour per day" doesn't have an obvious integer value, and would need to be included in the "No TV" or "1 hour" bins.
- "5 or more hours per day" would likely be converted to an integer value of 5, which could lead to underestimating the mean if there were significant outliers of 6 or more hours.

Given transformed data, two possible procedures are:

1. One-sample t-test to infer the population mean
2. Wilcoxon signed-rank test for centrality

The problem with the first approach follows from the initial transformation, and is exacerbated by handling of the upper and lower bins.

The second approach requires that a median, or central, value be specified. Inspecting the distribution of responses could provide such a central value (e.g. "2 hours" or an ordinal value of 4), after which the signed-rank test could proceed. However, this test answers the notion of "centrality" and assumes a symmetric distribution about this central value. Failing this test would provide unclear results: does the test fail because the central value is not in fact 2 hours, or is the distribution not symmetric about this value?

Given the discrepancy between the question posed and the data at hand, in this case a statistical test to infer to the general population might be less meaningful, and a better approach is to simply visualize the data and draw conclusions based on this visual summary. Doing so, and including proportional data in the below table, the following observations can be made:

- Over half of the respondents watch more 2 or more hours of television on an average school day. This seems quite high.
  - Furthermore, the percentage of respondents that watch 4 or more hours is also quite high (~20%).
- The most common response is "2 hours per day", although this is not dramatically different from other bins (~20% vs ~10-15% in other bins)

- Including “less than 1 hour” with 0 hours would make this the highest bin, and close to 30%.

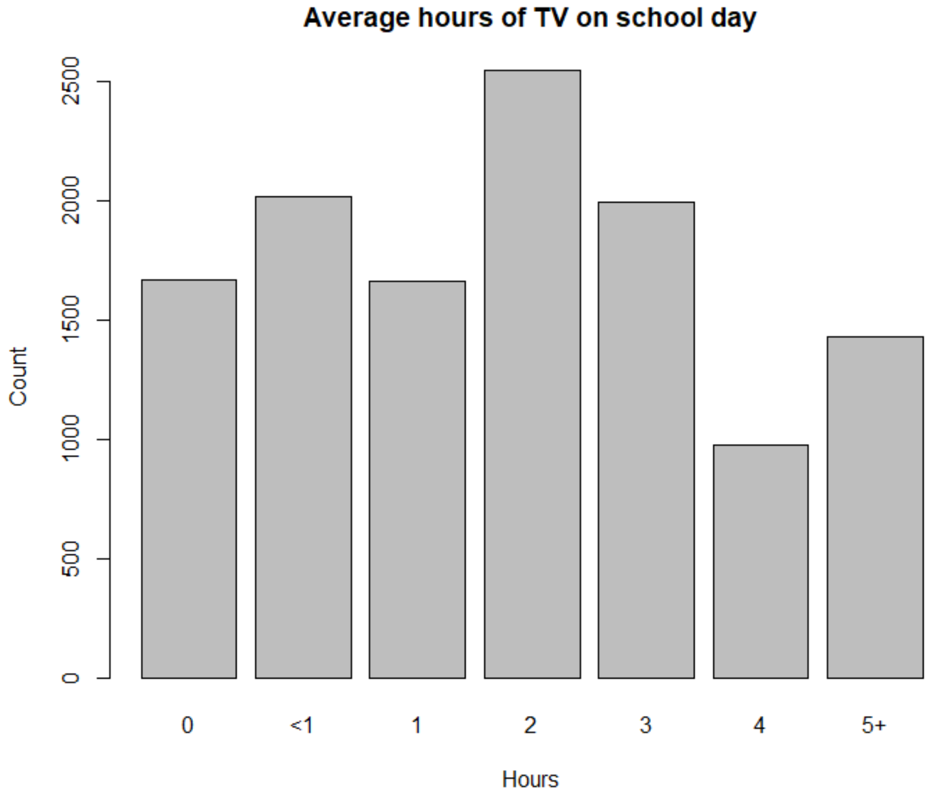

*Figure 8 Bar chart of hours of TV on a school day*

| 0 | <1 | 1 | 2 | 3 | 4 | 5+ |
|---|---|---|---|---|---|---|
| 0.136 | 0.164 | 0.135 | 0.207 | 0.162 | 0.079 | 0.116 |

*Table 7 Percentages of hours of television viewing*