ST_516: Foundations of Data Analytics

Project Report

Submitted By:

Lalit Gupta

11/30/2018

Table of Contents

Introduction:

Welcome to the final project of the “Foundations of Data Analytics” course (Course Code: ST-516) facilitated by Prof. James Molyneux. In this project we will perform various statistical analysis on observational data from the Youth Risk Behavior Surveillance System (YRBSS). YRBSS is a large survey of high school students in the United States of America. There are two data sets one each from year 2003 and 2013. For more details on the YRBSS data and the variables please refer appendix A.

This project is divided into two different analyses:

1. Simulation Study
2. Data Analysis

As part of Simulation study we will perform various simulations to draw conclusions on various summary parameters and analyze what affect sample size has on those parameters.

In the second part, Data analysis, we will perform data analysis on multiple variables of interest and try to answer questions related to them.

Simulation Study:

Aims & Objectives

In the simulation study, we will investigate the properties of four sample statistics:

1. Mean
2. 25th Percentile/Quantile
3. Minimum
4. Difference in Medians

We will perform large simulations on the Body Mass Index (BMI) from the YRBSS study for year 2013 and 2003. The YRBSS dataset will be considered the population for the purposes of this study and using that we will draw random samples to conduct the simulation study.

Before moving to the simulation study of the sample statistics, let’s have a look at the population data.

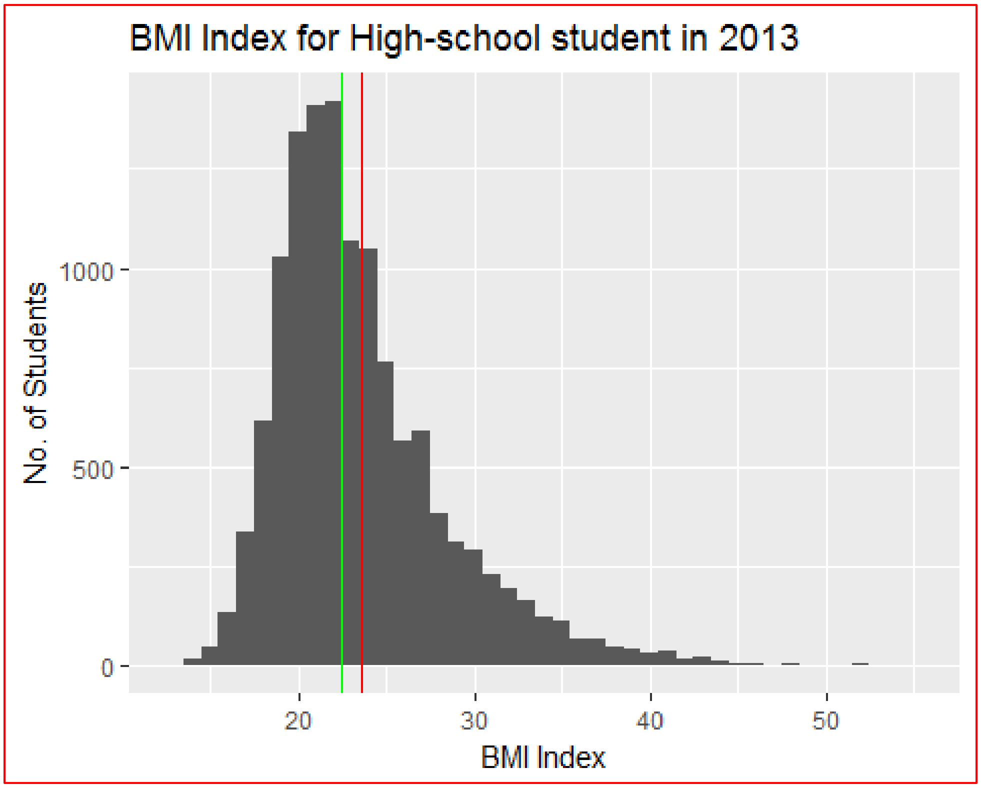

Figure 1 (BMI for high-school students – 2013)

As we see from Figure 1, the population has some outliers but is nearly normal. The mean (represented by red line on the graph) is towards the right of median (represented by green line on the graph) suggesting that the data is slightly right skewed.

Below table shows the population parameter:

	Mean	Median	Quantile	Minimum	SD
BMI - 2013	23.643	22.494	20.299	13.107	5.014

Simulation Study: Mean

First we would like to see how the sampling distribution changes for different sample sizes and then describe how the means and standard deviations of sampling distributions change with increasing sample size.

In order to do this we draw 10,000 samples of means of random samples of size 10, 100 and 1000 from the YRBSS dataset for year 2013 and get the respective sampling distributions.

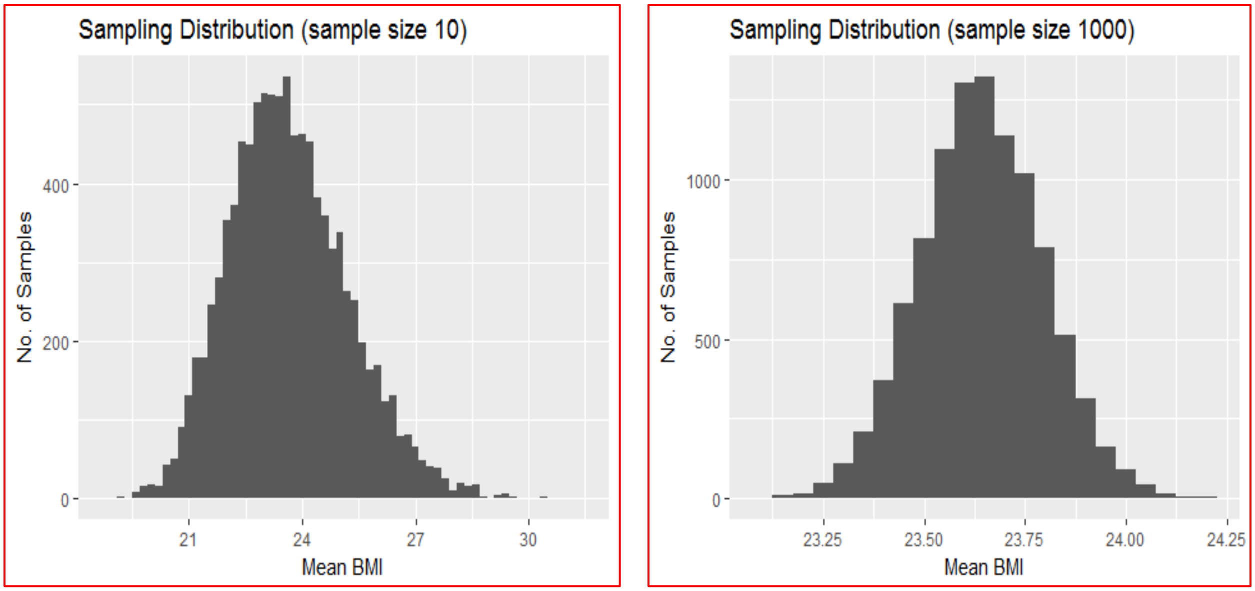

Figure 2 (Sampling Distributions: Mean BMI)

Looking at the sampling distributions of sample size 10 and 1000, we can observe that the sampling distribution of sample size 10 has outliers (not completely normal distribution) and is slightly skewed towards right. On the other hand, the sampling distribution of mean BMI for sample size 1000 is normal (even though the population is not, as we saw earlier), has no visible outliers, and the plot itself is not too spread out. Based on these graphs we can say that the sampling distribution is normal for large sample size even if the population is not.

Below table shows the mean and standard deviation of the sampling distributions:

Sample Size	Mean	SD
10	23.652	1.587
100	23.650	0.502
1000	23.642	0.153

Comparing this to the population mean (23.643) we see that as the sample size increases the estimate of the mean gets closer to the population mean. Also, the standard deviation of the sampling distribution decreases meaning we will be more confident in our estimate.

Simulation Study: 25 Percentile

Now that we have analyzed the sampling distributions of mean, we will move on to the next parameter of interest, 25 percentile. In order to study how this parameter is affected by sample size we will draw 10,000 samples of 25 percentile values from random samples of size 10, 100 and 1000 from the YRBSS dataset for year 2013 and get the respective sampling distributions.

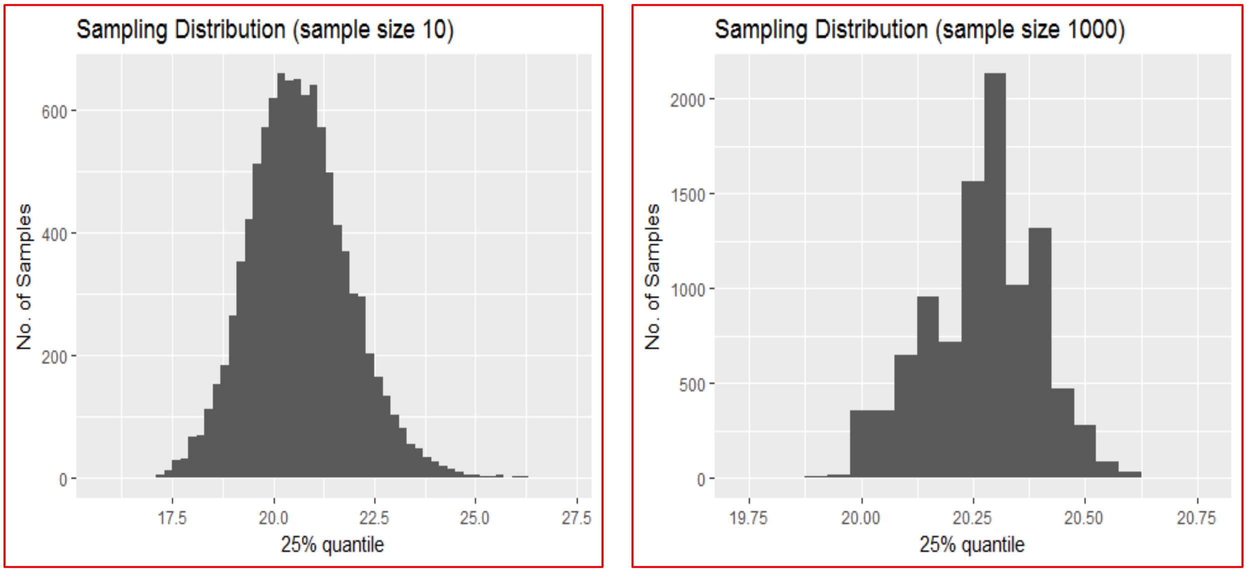

Figure 3 (Sampling Distributions: 25 percentile)

Looking at the sampling distributions above, we can see that the sampling distribution with sample size 10 is almost normal with a few outliers. Sampling distribution with sample size 1000 does not look normal but it has no visible outliers and the spread for the distribution is also small (less than 1). Based on these graphs we can say that with large sample size the sampling distribution tends to be tightly spread and it is centered around the population quantile.

Below table shows the mean, median and standard deviation for the sampling distributions:

Sample Size	Mean	Median	Standard Deviation
10	20.652	20.591	1.247
100	20.297	20.307	0.402
1000	20.272	20.294	0.129

The population 25th percentile is 20.299.

As we saw from the above graphs, the distribution does not appear to be normal especially for larger sample sizes, so median would be a better measure of center. Looking at the median we see that as the sample size increases the median tends to move closer to the population 25th percentile of 20.299. Another thing to observe is that the standard deviation decrease with the increase in sample size. That means that we can be more confident when estimating the population 25th percentile with larger sample size.

Simulation Study: Minimum

Now we are going to study the sample statistic “minimum”. To do this, we will draw 10,000 samples of minimum values from random samples of size 10, 100 and 1000 from the YRBSS dataset for year 2013 and get the respective sampling distributions.

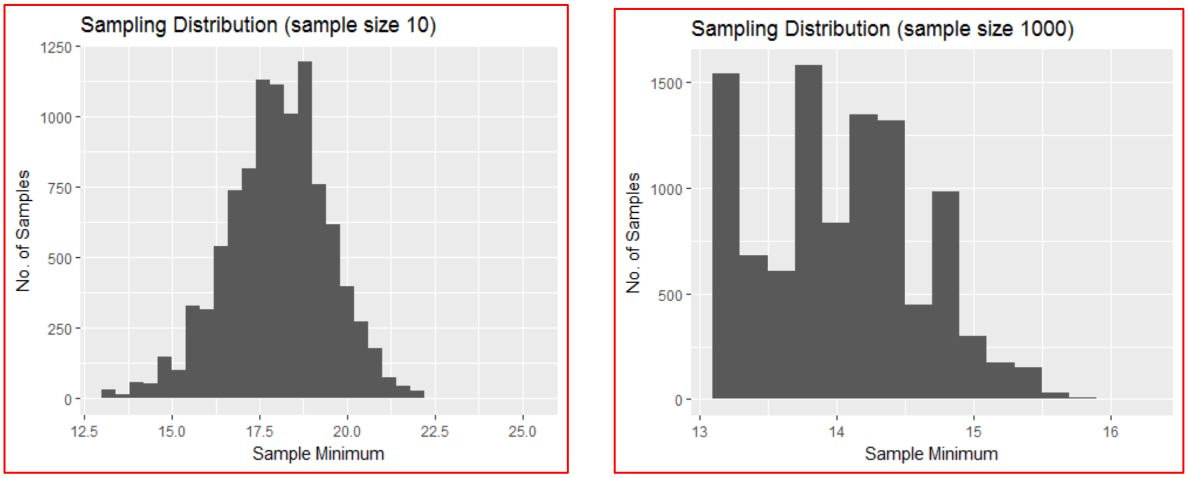

Figure 4 (Sampling Distributions: Minimum)

Looking at the sampling distributions of sample minimums for sample size 10 we see that it appears close to normal with a few outliers. The spread of the graph does not look too large. Now, looking at sampling distribution of sample size 1000, we see that the distribution is not normal but the spread is narrower than sampling distribution of sample size 10.

Below table shows the mean, median and standard deviation for the sampling distributions:

Sample Size	Mean	Median	Standard Deviation
10	18.019	18.048	1.47
100	15.662	15.707	0.98
1000	14.028	14.001	0.584

The population minimum is 13.107.

We see that as the sampling size increases the mean and median becomes a better estimate of the population minimum (13.107), even though the sampling distribution of large sample size is not normal. Another key thing is that the standard deviation decreases with increase in sample size, meaning that we can be more confident about our estimate.

Simulation Study: Difference in Median

Now we will analyze the difference in sample median BMI between 2013 and 2003. We will perform this by creating sampling distributions by using 10,000 repeated samples of size (5, 5), (10, 10) and (100, 100). We will then calculate the mean and standard deviation of the sampling distributions and see how they change with different sample sizes.

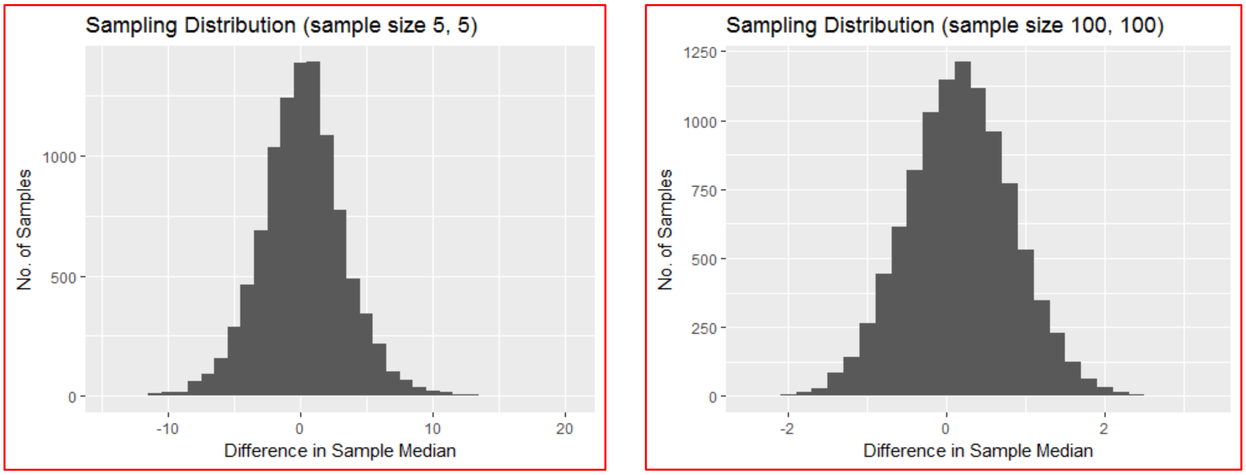

Figure 5 (Sampling Distributions: Difference in Sample Median)

The sampling distribution for sample size (5, 5) looks normal with no visible outliers. Same goes for sampling distribution for sample size (100, 100). The key difference between both is that the later one has smaller spread compared to the former one. Based on the given data and these graphs we can say that the sampling distribution of Difference in Sample Median is normal even for small sample size.

Below table shows the mean and standard deviations for the sampling distributions:

Sample Size	Mean	Standard Deviation
n1=5,n2=5	0.178	3.17
n1=10,n2=10	0.216	2.107
n1=100,n2=100	0.169	0.672

Population Difference of Median is 0.207

Looking at the above table we see that the mean of sampling distribution with sampling size (10, 10) is the closest to the population median difference. Surprisingly, the mean of sampling distribution with sample size (100, 100) is further away from population parameter even though standard deviation decreases as the sample size increases.

Summary

Based on our analysis of sampling distributions of various parameters with different sample sizes, we can say that atleast some sample statistic provide a better estimate of the population statistic with large sample size. As we saw, population estimates of mean, quantiles and minimum improve with large sample sizes as the center of the sampling distribution moves closer to the respective population parameter. Difference in Median became a better estimate when the sample size increased from (5, 5) to (10, 10), but it performed worse with sample size (100, 100). Based on the given data and current study we are not able to conclude the effect of sample size on Median Difference.

Data Analysis:

Aims & Objectives

In this section we will perform data analysis to study three variables of interest and try to answer some questions about them. We will be using the YRBSS data set and perform suitable data analysis tools and methodologies to draw inferences on broader population of high-school students in USA.

The questions of interest:

1. We want to study change in BMI for high-school students between 2013 and 2003. The question we want to answer is if high-schoolers in 2013 are getting overweight (BMI has increase on average) compared to 2003.
2. We want to study the difference in smoking habits of male and female high-school students and want to see if male students are more likely to smoke than female students in 2013.
3. Lastly, we want to study TV watching time of high-school students in 2013 and see if there are any differences based on gender. Another question of interest is does watching more TV lead to increased BMI.

Data Analysis: BMI

In this analysis we want to determine if BMI of high school students has increased on overage between 2013 and 2003. The data we will use to determine this is BMI data from YRBSS study from year 2013 and 2003.

Before we start analyzing, let’s plot the data and see if there are data points we need to focus on.

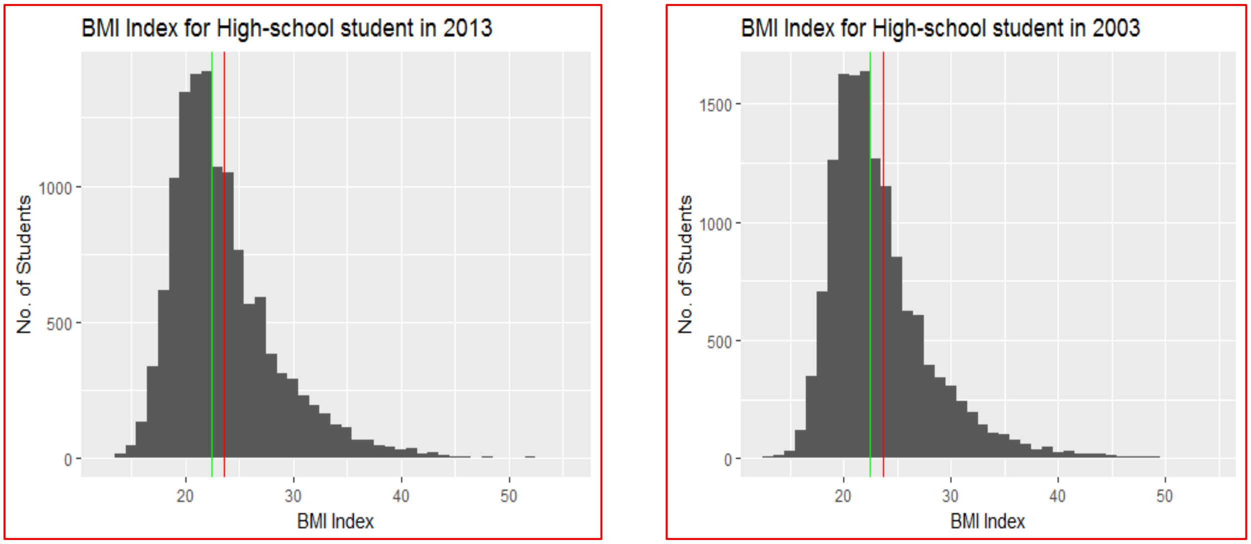

Figure 6 (BMI for High-school Students from 2013 and 2003)

Looking at the BMI data we see that there is a long right tail in data from both years and they do not appear to be normal but since we have large sample size we should be able to perform inferential study.

Assumptions

There are two samples of data – BMI from 2003 and BMI from 2013. The study does not states how the respondents were chosen, we are going to assume that they were chosen randomly and that the data is free from volunteer/convenience bias.

The data is less than 10% of the total population (around 15 million – Source: https://nces.ed.gov/fastfacts/display.asp?id=372) so we are going to assume that the samples are representative of data and are independent.

Since the studies are 10 years apart the chances of having the same student being part of the both studies is very unlikely, so we rule out the assumption of paired data. Worth noting is that there could some students in 2013 which are closely related to students in 2003 (coming from same household, community) and could be considered paired. Since the study does not provide any such data we are going to assume that there is no paired data.

We have large sample sizes (> 12000) so, based on Central Limit Theorem, we can assume normal approximation.

Procedure

We want to determine if the mean BMI of high-school students has increased from 2003 to 2013. So, we will use a Hypothesis test with null hypothesis that the mean BMI is same for high-school students between 2013 and 2003 and alternative hypothesis that the mean BMI is not same for high-school students between 2013 and 2003.

Since we have two independent samples and large sample size we will use Welch's Two Sample T-test with a 0.05 significance level.

Summary

There is convincing evidence (p = 0.0002, Welch's two sample t-test, df = 25988) that the mean BMI is not same for high-school students between 2013 and 2003. A 95% confidence interval for the mean difference is (0.108, 0.344) – high-school students from 2013 on average have greater BMI than high-school students from 2003. Although the results are statistically significant (we can say statistically that the mean BMI has increased) but looking at the 95% confidence interval the increase is estimated to have increased with very small quantity and that quantity may not be practically significant.

Data Analysis: Smoking Habits

In this analysis we want to study the smoking habits of male and female high-school students and determine if male students are more likely to smoke than female students in 2013.

Before diving into the analysis let's have a look at the data from YRBSS 2013 study.

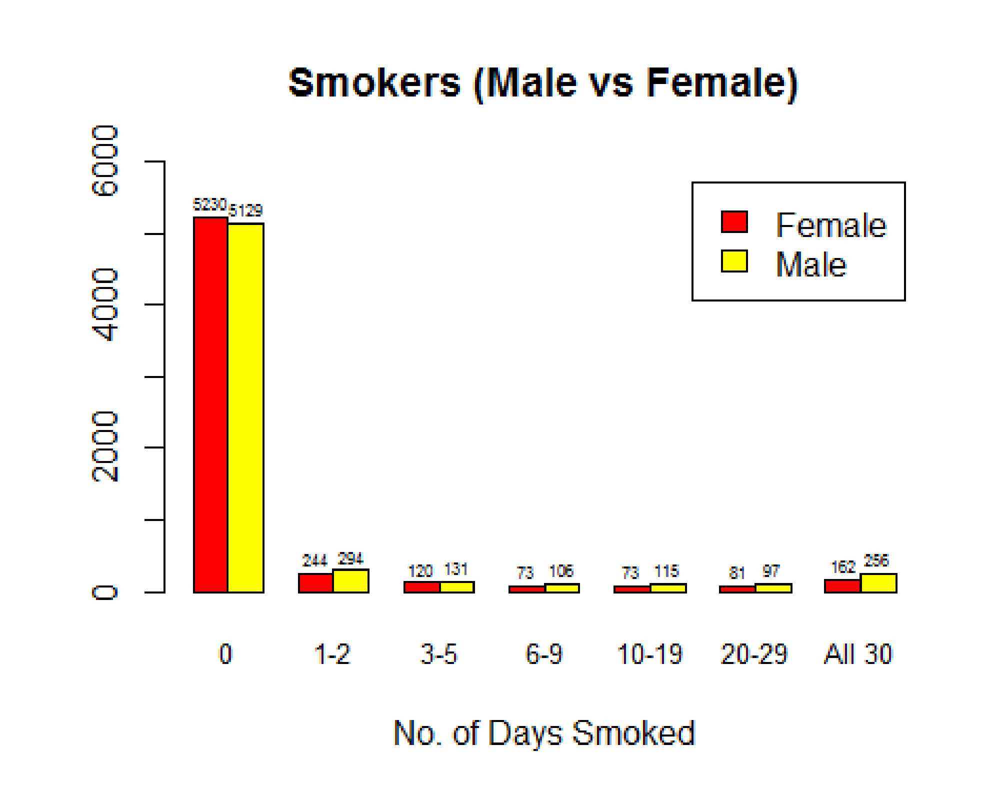

Figure 7 (Smoking habits: high-school students 2013 – Male vs Female)

Looking at the above plot the number of male high-schoolers who smoke is higher than the number of female high-schoolers that smoke in 2013. But we cannot infer that male high-schoolers smoke more than female high-schoolers solely based on this. This could simply be because there are more male high-schoolers than female high-schoolers. We will have to perform further analysis to establish statistical significance.

Assumptions

The data is less than 10% of the total population (around 15 million – Source: https://nces.ed.gov/fastfacts/display.asp?id=372) so we are going to assume that the samples are representative of data and are independent.

The study does not states how the respondents were chosen, we are going to assume that they were chosen randomly and that the data is free from volunteer/convenience bias.

There could be some paired data (for example male and female siblings, students from same households/community). Since the study does not provide any data on it, we are going to assume that all the data points are independent.

Procedure

We want to determine if the proportion of male students who smoke is greater than female students in 2013. In order to do this we will have to convert our YRBSS data from 2013 into two groups male and female and then use them as two independent populations. Then we will perform a hypothesis test with null hypothesis that the population proportion for male students who smoke is same as female students who smoke in 2013. Since we want to determine if male students are more likely to smoke, our alternative hypothesis will be the proportion of male students who smoke is greater than female students who smoke.

Since we want to compare the population proportions for two different populations (male and female students), we will use two sample proportions test.

Although, in order to perform the proportions test we will have to code the students as either smokers or non-smokers. Since, we do not have a guideline for this coding we are going to repeat the test with two different coding and see if that changes the outcome:

1. Anyone smoking more than 19 days is classified as "smoker"
2. Anyone smoking more than 0 days is classified as "smoker"

Summary

Based on coding 1 (more than 19 days -> smoker) we found convincing evidence ($p = 0$, two-sample proportions test, $df = 1$) that the proportion of male students who smoke is greater than the female students who smoke. A 95% confidence interval for difference in proportion is (0.05, 0.13).

Based on coding 2 (more than 0 days -> smoker) we found convincing evidence (p = 0, two-sample proportions test, df = 1) that the proportion of male students who smoke is greater than the female students who smoke. A 95% confidence interval for difference in proportion is (0.05, 0.10).

We see that even with two different coding we get the same result. So, we can say with confidence that the proportion of male students who smoke is greater than female students who smoke. In other words, male high-school students are more likely to smoke than female high-school students in 2013.

Data Analysis: TV Time

In this analysis we want to study the TV watching habits of high-school students in 2013. Then we will see if there are any differences in TV watching habits between males and females. Lastly we would like to determine if watching more TV has some correlation with BMI. Specifically we want to know if there is evidence of increase in BMI for students who watch TV for more than 2 hours per school-day.

Assumptions

The data is less than 10% of the total population (around 15 million – Source: https://nces.ed.gov/fastfacts/display.asp?id=372) so we are going to assume that the samples are representative of data and are independent.

The study does not states how the respondents were chosen, we are going to assume that they were chosen randomly and that the data is free from volunteer/convenience bias.

There could be some paired data (for example male and female siblings, students from same households/community). Since the study does not provide any data on it, we are going to assume that all the data points are independent.

Procedure

To study the TV watching habit of high-school students in 2013 we will perform exploratory data analysis. Then we will do the same based on gender and see if there are any major differences in the two gender groups. Lastly, to study the relationship between TV time and BMI we will do exploratory analysis on both the variables. If we see some evidence of correlation we will try to find if there is statistical significance to the observed correlation. In order to do this we will determine if watching TV for more than 2 hours show evidence of increased mean BMI.

In order to answer the above question we will perform a Hypothesis test with null hypothesis that there is no difference in the mean BMI between students who watch TV for 2 hours or less and students who watch TV for more than 2 hours. The alternative hypothesis being the mean BMI between students who watch TV for 2 hours or less and students who watch TV for more than 2 hours is different.

Since we have two independent samples and large sample size we will use Welch's Two Sample T-test with a 0.05 significance level.

To get the two samples we will split the BMI data into two groups; Students who watch TV for 2 hours or less and Students who watch TV for more than 2 hours. We will use these two groups as the two samples in t-test.

Summary

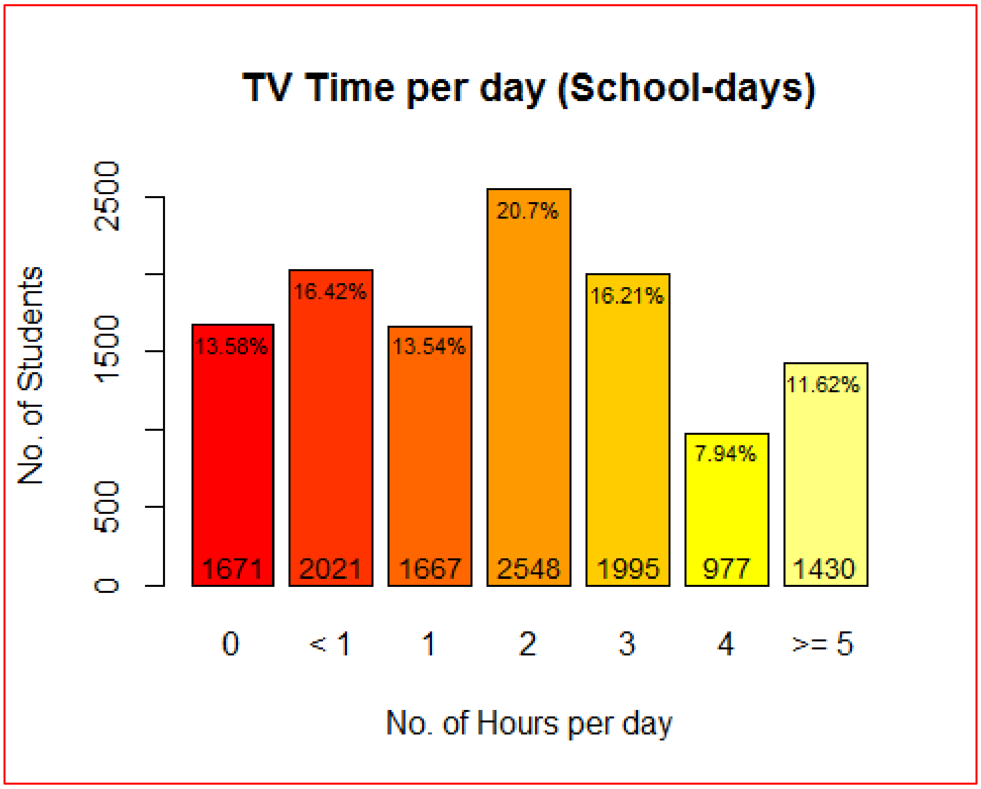

Figure 8 (TV time per School-day: high-school students 2013)

Looking at the above figure we see that most of the students (around 50%) watch TV for 1 to 3 hours. Also, 13.5% of the students do not watch any TV and 11.5% students watch TV for more than 5 hours per school-day.

Now, let's see if there is any difference in TV watching habits between male and female students.

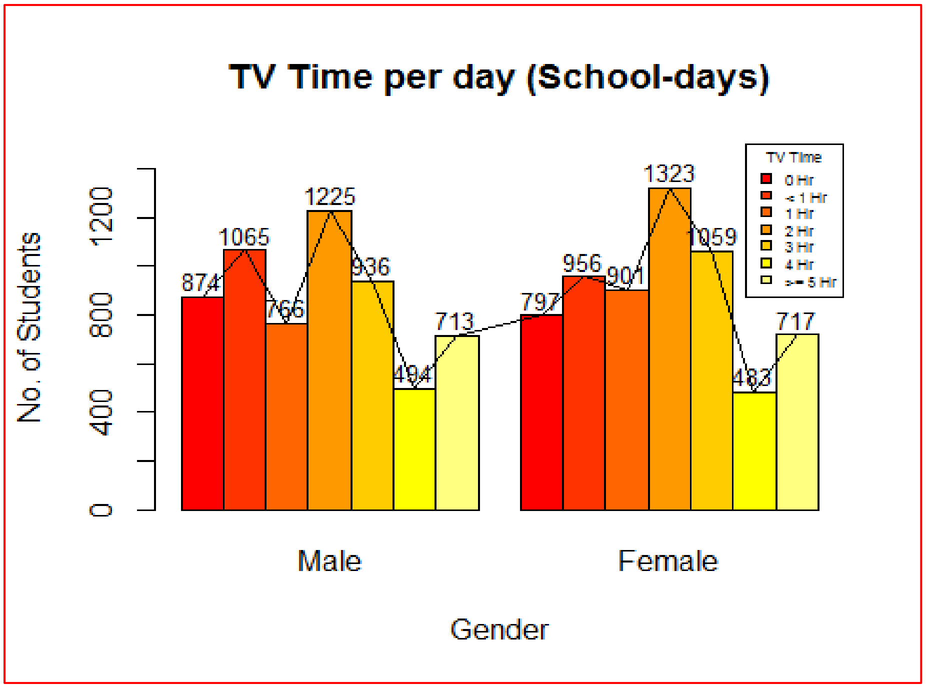

Figure 9 (TV time per day: Male vs Female)

Looking at the pattern of the graph, there seems to be no difference in the TV watching habits of Male vs Female students.

Now, let see the graph for mean BMI and TV time.

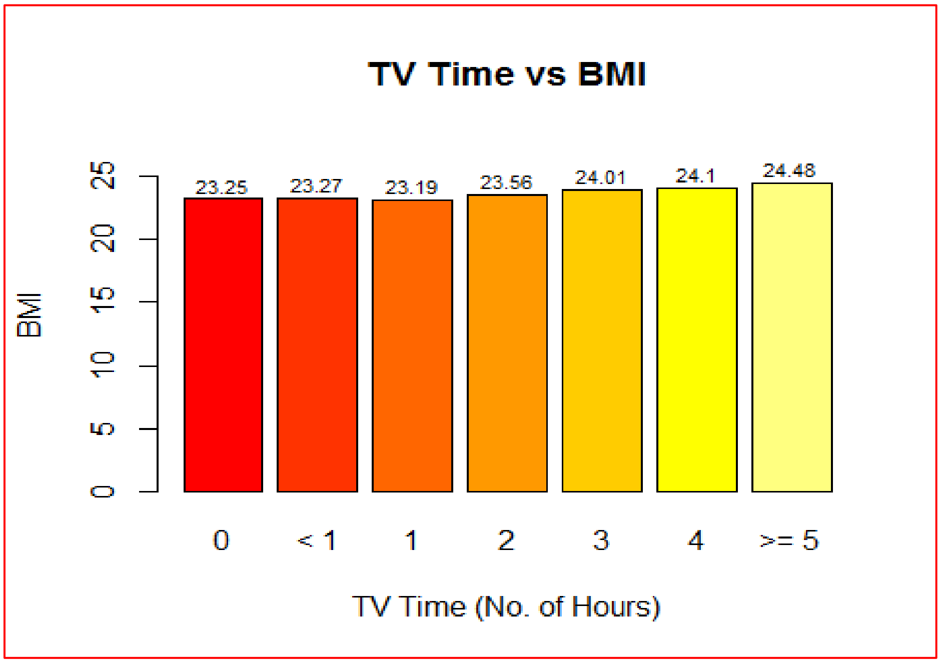

Figure 10 (TV time and BMI: high-school students 2013)

We see from the graph that the BMI on average increases with the TV watching hours. But is the increase in BMI statistically significant?

We answered this question in a different way. We performed a Welch's two-sample hypothesis test to determine is students who watch TV for more than 2 hours per day have a higher BMI on average.

We found convincing evidence (p = 0, Welch's two sample t-test, df = 8053) that the mean BMI of high-school students who watch TV for more than 2 hours per day is not same as the BMI of high-school students who watch TV for 2 or less hours per day. A 95% confidence interval for the mean difference is (0.645, 1.029) – suggesting that the mean BMI of students who watch TV for more than 2 hour per day is higher. There may be confounding factors that make it difficult to attribute this increase in mean BMI directly to TV watching habits. For Example, students may be excessively snacking while watching TV which may the causing the increase in BMI. We will have to conduct experimental studies to establish causation.

Appendix A (YRBSS Dataset and Variables)

YRBSS data set variables.

Variable	Description
year	4-digit year of survey – 1991, 1993, etc.
age	Data from: How old are you? A. 12 years old or younger B. 13 years old C. 14 years old D. 15 years old E. 16 years old F. 17 years old G. 18 years old or older
sex	Data from: What is your sex? A. Female B. Male
grade	Data from: In what grade are you? A. 9thgrade B. 10thgrade C. 11thgrade D. 12thgrade E. Ungraded or other grade
race4	4-level variable from race and ethnicity questions: 1 = "White" 2 = "Black or African American" 3 = "Hispanic/Latino" 4 = "All Other Races"
stheight	Data from: How tall are you without your shoes on?
stweight	Data from: How much do you weigh without your shoes on?
bmi	Body mass index (BMI)

q9	Q9. How often do you wear a seat belt when riding in a car driven by someone else? A. Never B. Rarely C. Sometimes D. Mostofthetime E. Always
q33	Q33. During the past 30 days, on how many days did you smoke cigarettes? A. 0 days B. 1or2days C. 3to5days D. 6to9days E. 10 to 19 days F. 20 to 29 days G. All 30 days
q77	Q77. During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not count diet soda or diet pop.) A. I did not drink soda or pop during the past 7 days B. 1 to 3 times during the past 7 days C. 4 to 6 times during the past 7 days D. 1 time per day E. 2 times per day F. 3 times per day G. 4 or more times per day
q80	Q80. During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physicalactivity that increased your heart rate and made you breathe hard some of the time.) A. 0 days B. 1 day C. 2 days D. 3 days E. 4 days F. 5 days G. 6 days H. 7 days

q81	A. I do not watch TV on an average schoolday B. Less than 1 hour per day C. 1 hour per day D. 2 hours per day E. 3 hours per day F. 4 hours per day G. 5 or more hours per day

Appendix B (R code)

```
# load libraries

library(ggplot2)


# set seed so we get consistent results on every run

set.seed(2018)


# ==   Simulation Study      =====


# load the data

yrbss_2003 <- readRDS("yrbss_2003.rds")

yrbss_2013 <- readRDS("yrbss_2013.rds")


n_sim <- 10000 # number of simulations




# ==   Simulation Study: Mean      =====



# extract and save the variable of interest so they

# can be accessed easily

bmi_2003 <- yrbss_2003$bmi

bmi_2013 <- yrbss_2013$bmi


# Let's look at the Population (bmi_2013) for BMI Index

# plotting the population

qplot(bmi_2013, binwidth = 1) +
  ggtitle("BMI Index for High-school student in 2013") +
  xlab("BMI Index") +
  ylab("No. of Students") +
  geom_vline(xintercept = mean(bmi_2013), color = "red") +
  geom_vline(xintercept = median(bmi_2013), color = "green")

# Population (bmi_2013) parameters
cbind(Pop_Mean = round(mean(bmi_2013),3),
      Pop_SD = round(sd(bmi_2013),3),
      Pop_Median = round(median(bmi_2013), 3),
      Pop_quantile = round(quantile(bmi_2013, probs = 0.25), 3)
)

# Writing a function to get means of random samples of
# size n from population x
get_mean <- function(n, n_sim, x = bmi_2013){
  replicate(n_sim, mean(sample(x, n, replace=FALSE)))
}

# get means for n_sim repeated random samples of size 10,
# 100 and 1000
ns <- c(10, 100, 1000)
means <- lapply(ns, get_mean, n_sim = n_sim)

# Let's see how the sampling distribution looks like
# for sample size 10 and 1000
qplot(means[[1]], binwidth = 0.2) +
  ggtitle("Sampling Distribution (sample size 10)") +
  xlab("Mean BMI") +
  ylab("No. of Samples")
qplot(means[[3]], binwidth = 0.05) +
  ggtitle("Sampling Distribution (sample size 1000)") +
  xlab("Mean BMI") +
  ylab("No. of Samples")

# Let's get the mean and SD for our sampling Distributions
mean_sam_dist <- sapply(means, mean)
sd_sam_dist <- sapply(means, sd)
cbind(SampleSize = ns,
      Mean = round(mean_sam_dist, 3),
      SD = round(sd_sam_dist, 3))



# == Simulation Study: 25 percentile  =====


# Writing function to get 25% quantile of random samples of
# size n from population x
get_quantile <- function(n, n_sim, x = bmi_2013){
  replicate(n_sim, quantile((sample(x, n, replace=FALSE)),
                            probs = 0.25))
}

# get 25% quantiles for n_sim repeated random samples of size
# 10, 100 and 1000
ns <- c(10, 100, 1000)
quantiles <- lapply(ns, get_quantile, n_sim = n_sim)
```

```
# Let's see how the sampling distribution looks like
# for sample size 10 and 1000
qplot(quantiles[[1]], binwidth = 0.2) +
  ggtitle("Sampling Distribution (sample size 10)") +
  xlab("25% quantile") +
  ylab("No. of Samples")
qplot(quantiles[[3]], binwidth = 0.05) +
  ggtitle("Sampling Distribution (sample size 1000)") +
  xlab("25% quantile") +
  ylab("No. of Samples")

# Let's get the mean, median and SD for our sampling Distributions
mean_quantile <- sapply(quantiles, mean)
median_quantile <- sapply(quantiles, median)
sd_quantile <- sapply(quantiles, sd)
cbind(SampleSize = ns,
      Mean = round(mean_quantile, 3),
      Median = round(median_quantile, 3),
      SD = round(sd_quantile, 3)
)

# Population (bmi_2013) parameters for BMI Index
cbind(Pop_Quantile = quantile(bmi_2013, probs = 0.25))



# == Simulation Study: Minimum  =====
# function to get minimum of random samples of size n
# from population x
get_min <- function(n, n_sim, x = bmi_2013){
  replicate(n_sim, min((sample(x, n, replace=FALSE))))
}

# get minimum for n_sim repeated random samples of size 10,
# 100 and 1000
ns <- c(10, 100, 1000)
mins <- lapply(ns, get_min, n_sim = n_sim)

# Let's see how the sampling distribution looks like
# for sample size 10 and 1000
qplot(mins[[1]], binwidth = 0.4) +
  ggtitle("Sampling Distribution (sample size 10)") +
  xlab("Sample Minimum") +
  ylab("No. of Samples")
qplot(mins[[3]], binwidth = 0.2) +
  ggtitle("Sampling Distribution (sample size 1000)") +
  xlab("Sample Minimum") +
  ylab("No. of Samples")

# Let's get the mean, median and sd for our sampling Distributions
mean_mins <- sapply(mins, mean)
median_mins <- sapply(mins, median)
sd_mins <- sapply(mins, sd)
cbind(SampleSize = ns, Mean = round(mean_mins, 3),
      Median = round(median_mins, 3),
   SD = round(sd_mins, 3))

# Population (bmi_2013) parameters for BMI Index
cbind(Pop_Minimun = min(bmi_2013))



# == Simulation Study: Diffrence in Median  =====


# function to get minimum of difference of sample median BMI
# between 2013 and 2003 by using sample size n1 and n2 respectively
get_median_diff <- function(n1, n2, n_sim, x1 = bmi_2013,
                            x2 = bmi_2003){
  median_2013 <- replicate(n_sim, median((sample(x1, n1,
                                         replace=FALSE))))
  median_2003 <- replicate(n_sim, median((sample(x2, n2,
                                         replace=FALSE))))
  median_2013-median_2003
}

# get Difference in sample medians between 2013 and 2003 for
# n_sim repeated random samples of size 5,5, 10,10 and 100,100
med_diffs <- list()
med_diffs[[1]] <- get_median_diff(5, 5, n_sim)
med_diffs[[2]] <- get_median_diff(10,10, n_sim)
med_diffs[[3]] <- get_median_diff(100, 100, n_sim)

# Let's see how the sampling distribution looks like
# for sample size 5,5 and 100,100
qplot(med_diffs[[1]], binwidth = 1) +
  ggtitle("Sampling Distribution (sample size 5, 5)") +
  xlab("Difference in Sample Median") +
  ylab("No. of Samples")
qplot(med_diffs[[3]], binwidth = 0.2) +
  ggtitle("Sampling Distribution (sample size 100, 100)") +
  xlab("Difference in Sample Median") +
  ylab("No. of Samples")

# Let's get the mean and SD for our sampling Distributions
mean_med_diff <- sapply(med_diffs, mean)
sd_med_diff <- sapply(med_diffs, sd)
cbind(SampleSize = c('n1=5,n2=5', 'n1=10,n2=10',
                    'n1=100,n2=100'),
      Mean_Diff = round(mean_med_diff, 3),
      SD_Diff = round(sd_med_diff, 3))

# Population (bmi_2013 and bmi_2003 ) parameters for BMI Index
median(bmi_2013) - median(bmi_2003)



# == Data Analysis: BMI =====


# let's look the variables of interest
qplot(yrbss_2013$bmi, binwidth = 1) +
  ggtitle("BMI Index for High-school student in 2013") +
  xlab("BMI Index") +
  ylab("No. of Students") +
  geom_vline(xintercept = mean(bmi_2013), color = "red") +
  geom_vline(xintercept = median(bmi_2013), color = "green")

qplot(yrbss_2003$bmi, binwidth = 1) +
  ggtitle("BMI Index for High-school student in 2003") +
  xlab("BMI Index") +
  ylab("No. of Students") +
  geom_vline(xintercept = mean(bmi_2013), color = "red") +
  geom_vline(xintercept = median(bmi_2013), color = "green")

# perform two sample t-test to see if the BMI changes are
# significant enough to draw statistical significance.
t.test(bmi_2013, bmi_2003, mu = 0, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)



# == Data Analysis: Smoking Habits   =====


# let's look at the variable of interest
smoke <- table(yrbss_2013$sex, yrbss_2013$q33)
bp <- barplot(smoke, main = "Smokers (Male vs Female)",
        legend = c("Female","Male"), beside = TRUE,
        col=heat.colors(2), ylim = c(0,6000),
        names.arg = c("0", "1-2", "3-5",
                 "6-9","10-19", "20-29",
              "All 30"),
      xlab = "No. of Days Smoked")
text(bp, smoke, smoke, cex = 0.8, pos = 3, offset = 0.2)

# let's perform a one sided proportions test with significance
# level 0.05 and alternative hypothesis that population proportion
# of male high-schoolers who smoke is higher than population
# proportion of female high-schoolers

# converting qualitative variables to quantitative variables:
# anyone smoking less than 20 days is considered non-smoker;
# considered smoker otherwise
yrbss_2013$smokers <- as.character(yrbss_2013$q33)
yrbss_2013$smokers[yrbss_2013$smokers %in% c("0 days",
                    "1 or 2 days",
                    "3 to 5 days",
                    "6 to 9 days",
                    "10 to 19 days")] <- 0
yrbss_2013$smokers[yrbss_2013$smokers %in% c("20 to 29 days",
                    "All 30 days")] <- 1

# performing one sided proportion test
sex_smoke_table <- table(yrbss_2013$smokers, yrbss_2013$sex)
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "g",
      correct = FALSE)
# performing two sided proportions test to get confidence intervals
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "t",
      correct = FALSE)
```

```
# Let's see if the result is different when we classify smokers
# more rigidly, i.e. anyone smoking for more than 0 days is
# classified as smoker
yrbss_2013$smokers <- as.character(yrbss_2013$q33)
yrbss_2013$smokers[yrbss_2013$smokers %in% c("0 days")] <- 0
yrbss_2013$smokers[yrbss_2013$smokers %in% c("1 or 2 days",
                         "3 to 5 days","6 to 9 days",
                         "10 to 19 days",
                         "20 to 29 days",
                         "All 30 days")] <- 1

# performing one sided proportion test
sex_smoke_table <- table(yrbss_2013$smokers, yrbss_2013$sex)
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "g",
      correct = FALSE)
# performing two sided proportions test to get confidence intervals
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "t",
      correct = FALSE)



# == Data Analysis: TV Time  =====



# Let's explore the variable of interest
tv_time <- table(yrbss_2013$q81)

# let's summarize the variable with a bar graph
bp <- barplot(tv_time, main = "TV Time per day (School-days)",
        col = heat.colors(7), ylab = "No. of Students",
        names.arg = c("0", "< 1", "1", "2",
                "3", "4", ">= 5"), width = 1.2,
        args.legend = list(title = "TV Time", x = "topright",
                   cex = .7),
        xlab = "No. of Hours per day",
        beside = TRUE, ylim = c(0, 2700))
tv_time_percent <- as.character(round((tv_time*100/sum(tv_time)),2))
tv_time_percent <- paste(tv_time_percent, "%", sep = "")
text(bp, 0, tv_time, cex=0.9, pos=3, offset = 0.2)
text(bp, tv_time, tv_time_percent, cex=0.9, pos=1, offset = 0.3)

# Let's see how TV time differs between male and female students
tv_time_sex <- table(yrbss_2013$q81, yrbss_2013$sex)

bp <- barplot(tv_time_sex, main = "TV Time per day (School-days)",
        col = heat.colors(7), xlab = "Gender",
        names = c("Male", "Female"), beside = TRUE,
        ylab = "No. of Students", ylim = c(0,1500),
        legend = c("0 Hr", "< 1 Hr", "1 Hr", "2 Hr", "3 Hr",
              "4 Hr", ">= 5 Hr"),
        args.legend = list(title = "TV Time", x = "topright",
                   cex = 0.48))
text(bp, y = tv_time_sex, tv_time_sex, cex = 0.9, pos=3,
   offset = 0.2)
lines(x = bp, y = tv_time_sex)

# Let's see if watching more TV has any correlation with BMI
# index
```

```
tv_bmi <- aggregate(yrbss_2013$bmi, list(yrbss_2013$q81), mean)
colnames(tv_bmi) <- c("TV-Time", "bmi")

bp <- barplot(tv_bmi$bmi, ylim = c(0,28),
        xlab = "TV Time", ylab = "BMI",
        col = heat.colors(7), width = 1.2,
        names.arg = c("0 Hr", "< 1 Hr", "1 Hr",
                "2 Hr", "3 Hr", "4 Hr",
                ">= 5 Hr"),
        main = "TV Time vs BMI")
text(bp, y = tv_bmi$bmi, round(tv_bmi$bmi,2), cex = 0.7,
   pos=3, offset = 0.2)

# Time spent watching TV and BMI index do seem to be positively
# correlated. Let's perform a two sample t-test to find out if
# watching tv for more than 2 hours per day results in increased
# BMI.
yrbss_2013$tv <- as.character(yrbss_2013$q81)
yrbss_2013$tv[yrbss_2013$tv %in% c("No TV on average school day",
                "Less than 1 hour per day",
                "1 hour per day",
                "2 hours per day")] <- 0
yrbss_2013$tv[yrbss_2013$tv %in% c("3 hours per day",
                "4 hours per day",
                "5 or more hours per day")] <- 1

tv_2hr_less <- yrbss_2013$bmi[yrbss_2013$tv == 0]
tv_2hr_more <- yrbss_2013$bmi[yrbss_2013$tv == 1]
t.test(tv_2hr_more, tv_2hr_less, mu = 0, alternative = "t",
paired = FALSE, var.equal = FALSE)
```