guptala_final_project

Lalit Gupta

November 29, 2018

Loading required libraries and setting seed to ensure consistency

```
library(ggplot2)
set.seed(2018)
```

Next, we will load the data to be analysed

```
yrbss_2003 <- readRDS("yrbss_2003.rds")
yrbss_2013 <- readRDS("yrbss_2013.rds")
```

Now, let's extract and save the variables of interest so they can be easily accessed

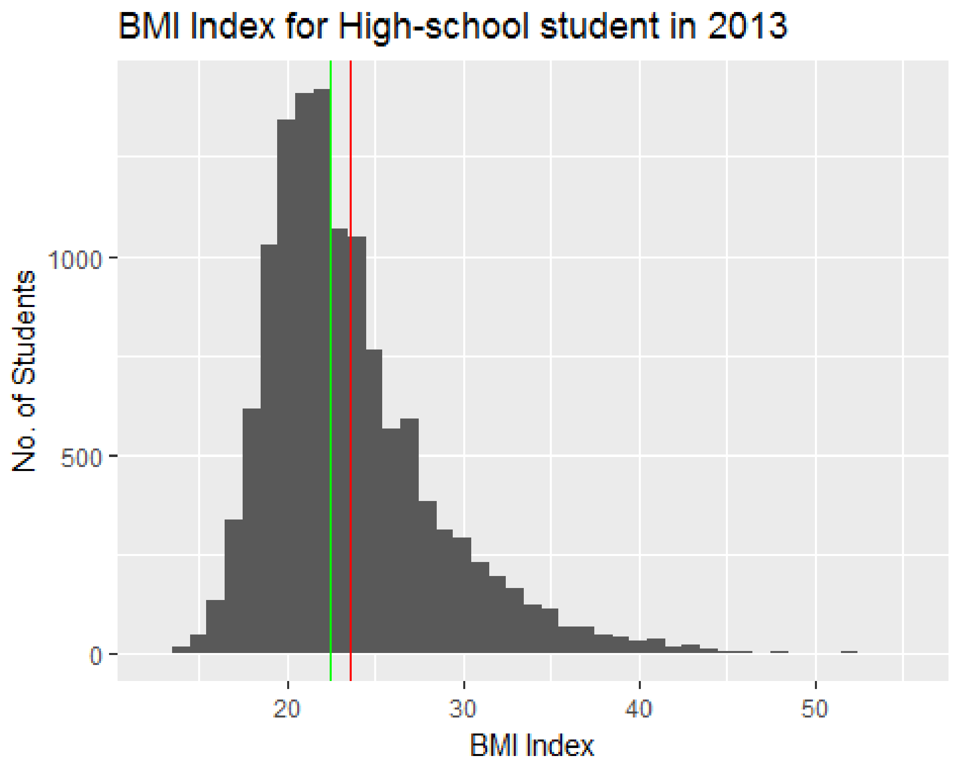
```
bmi_2003 <- yrbss_2003$bmi
bmi_2013 <- yrbss_2013$bmi
```

Simulation Study: Mean

We are interested in analysing the sampling distributions of means of random samples drawn from BMI index of high-school students from year 2013 and see how it changes with different sample sizes (10, 100, 1000) and also how they compare to the population mean and standard deviation.

Let's begin with getting a visual of 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")
```



The green line in the above chart shows the median and the red one shows the mean of the population. As we see, the population is slightly right skewed.

Let’s get the population parameters now.

```
# Population (bmi_2013) parameters for BMI Index
cbind(Pop_Mean = round(mean(bmi_2013),3),
      Pop_Median = round(median(bmi_2013), 3),
      Pop_quantile = round(quantile(bmi_2013, probs = 0.25), 3),
      Pop_min = round(min(bmi_2013), 3),
      Pop_SD = round(sd(bmi_2013),3)
      )

##     Pop_Mean Pop_Median Pop_quantile Pop_min Pop_SD
## 25%   23.643     22.494       20.299  13.107  5.014
```

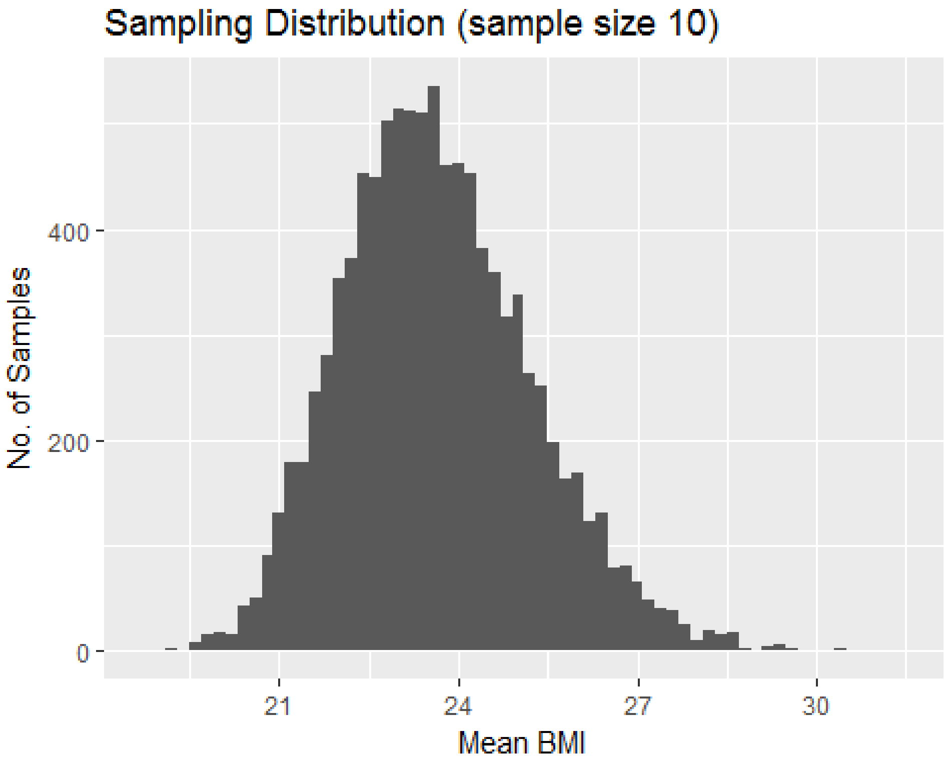
Okay, we are off to a good start. Now, let’s get means of 10000 random samples of size 10, 100 and 1000. This will give us our sampling distributions.

```
n_sim <- 10000 # number of simulations

# 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)
```

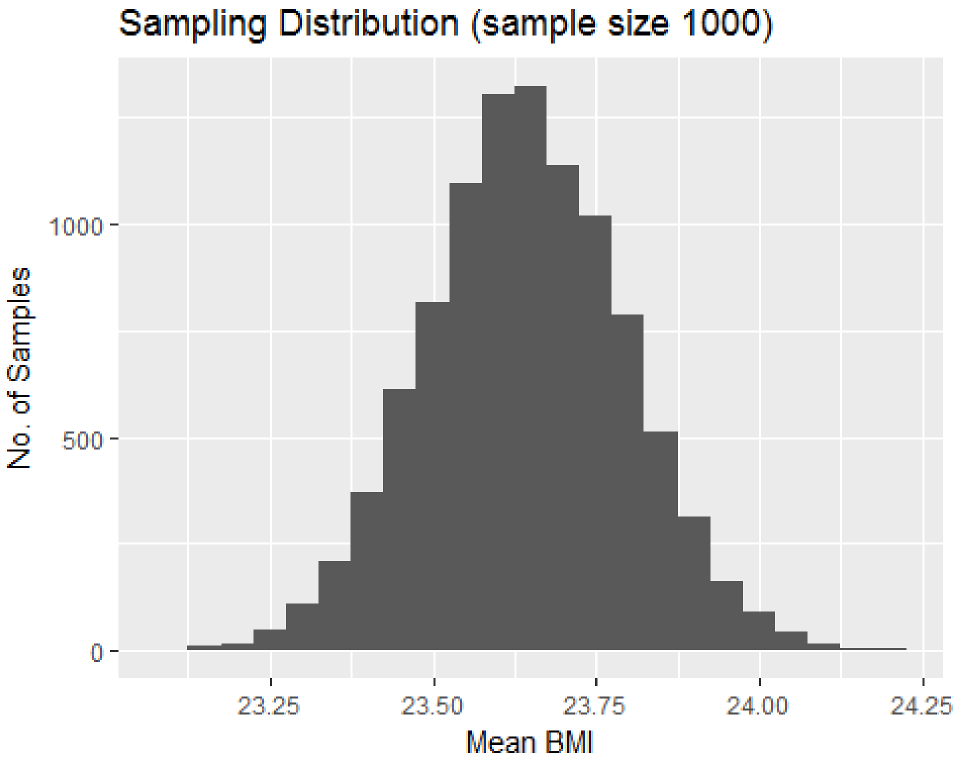
Now that we have got our sampling distributions, let's plot them (for sample size 10 and 1000) and see how they look.

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



Observe that the sampling distribution of sample size 10 is nearly normal with some outliers and also has a large spread overall. This implies that even though this will be a good approximation of the population, we will be not as confident because of the larger spread.

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



Comparatively, the sampling distribution of sample size 1000 is normal (even though the population is not, as we saw earlier). This should give us a better estimate of the population, also worth noting is that it is evenly spread with not too much deviation from the center.

Now, let's calculate the mean and standard deviation for our sampling distributions.

```
# 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))

##      SampleSize   Mean    SD
## [1,]         10 23.652 1.587
## [2,]        100 23.650 0.502
## [3,]       1000 23.642 0.153
```

Looking at the Mean and Standard Deviation for different sample size we see that the estimate of the mean gets closer to the polulation mean (23.643) as the sample size increases. Also, worth noting is the fact that the standard deviation decreases with increase in sample size. Meaning, with larger sample size we not only get closer to the population parameter but our confidence in our estimate also increases.

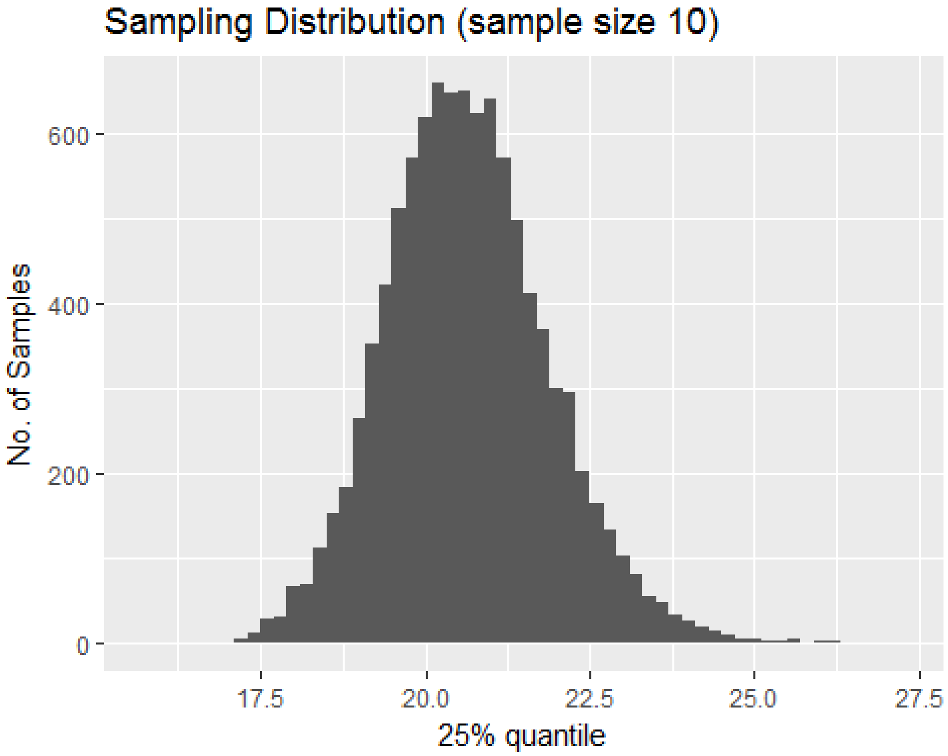
Simulation Study: 25 Percentile

So far we analysed the sampling distributions of means. Now, let's see how the sampling distributions and its parameters look like with 25% quantile.

```
# 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))
}

n_sim = 10000
# 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)
```

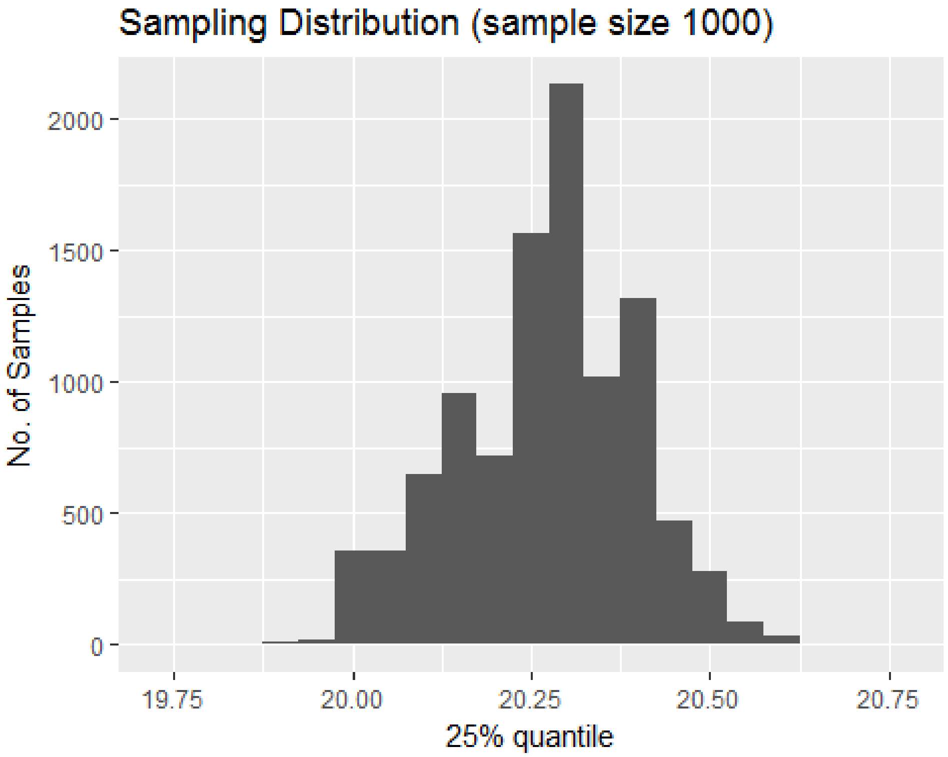
We have got our sampling distributions (with sample size 10, 100 and 1000). Let's see how they look like for sample size 10 and 1000.

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



We see that the sampling distribution with sample size 10 is almost normal. There are a few outliers but the spread is well contained. Now, let's compare this to the sampling distribution of sample size 1000.

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



Comparatively, the sampling distribution is well centered and there does not seem to be any outliers. The distribution although does not appear normal.

Next, let's get the mean, median and sd for our sampling distributions and compare them to our population 25% quantile.

```r
# 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)
      )
```

```
##      SampleSize   Mean Median    SD
## [1,]         10 20.652 20.591 1.247
## [2,]        100 20.297 20.307 0.402
## [3,]       1000 20.272 20.294 0.129

# Population (bmi_2013) parameters for BMI Index
cbind(Pop_Quantile = round(quantile(bmi_2013, probs = 0.25),3))

##     Pop_Quantile
## 25%       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.

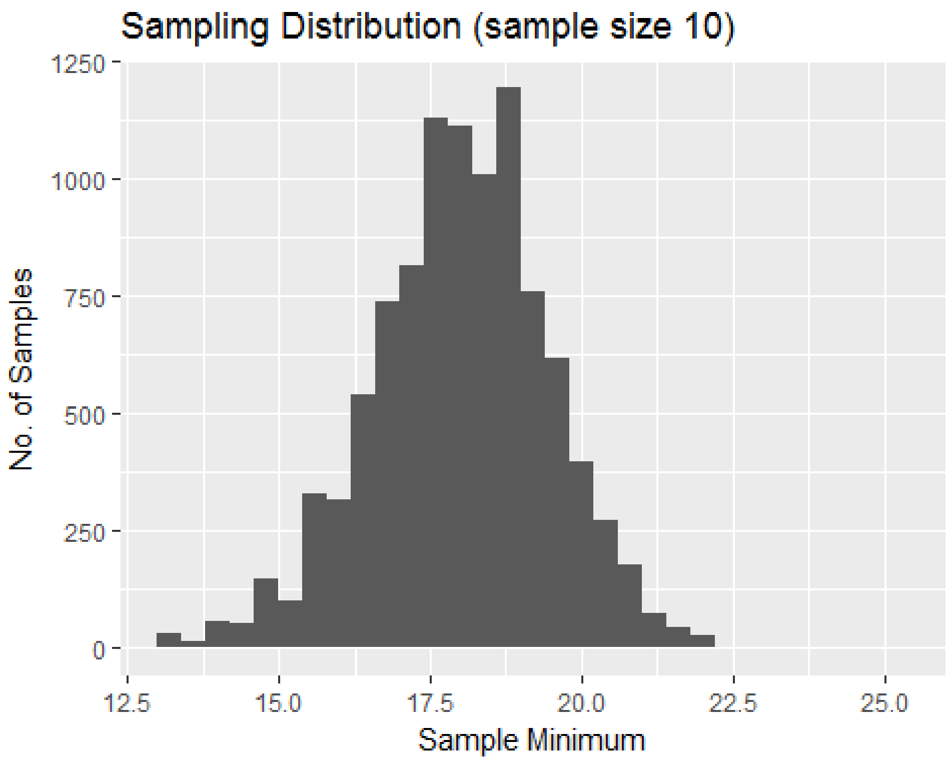
Simulation Study: Minimum

So far we have analysed sampling distributions of mean and quantile, next let’s analyse the sampling distributions of minimum for random samples of size 10, 100 and 1000.

```
# 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))))
}

n_sim = 10000
# 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)
```

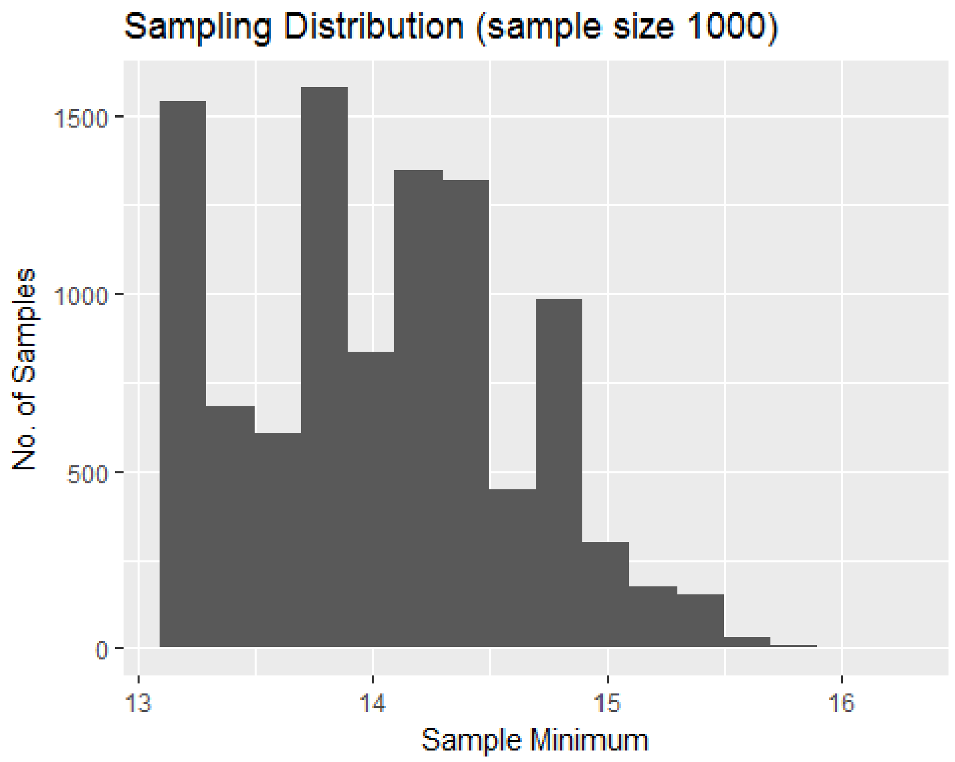
We have got our sampling distributions (with sample size 10, 100 and 1000). Let’s see how they look like for sample size 10 and 1000.

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



We see that the sampling distribution of minimums for sample size 10 has some outliers and the graph has some right skew. Let’s compare this to the sampling distribution with sample size 1000.

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



The sampling distribution with sample size 1000 does not look normal but it has no visible outliers and it does not have a very large spread.

Next, let's get the mean, median and sd for our sampling distributions and compare them to our population minimum.

```
# 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))

##      SampleSize   Mean Median    SD
## [1,]         10 18.019 18.048 1.470
## [2,]        100 15.662 15.707 0.980
## [3,]       1000 14.028 14.001 0.584

# Population (bmi_2013) parameters for BMI Index
cbind(Pop_Minimun = round(min(bmi_2013), 3))

##      Pop_Minimun
## [1,]      13.107
```

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

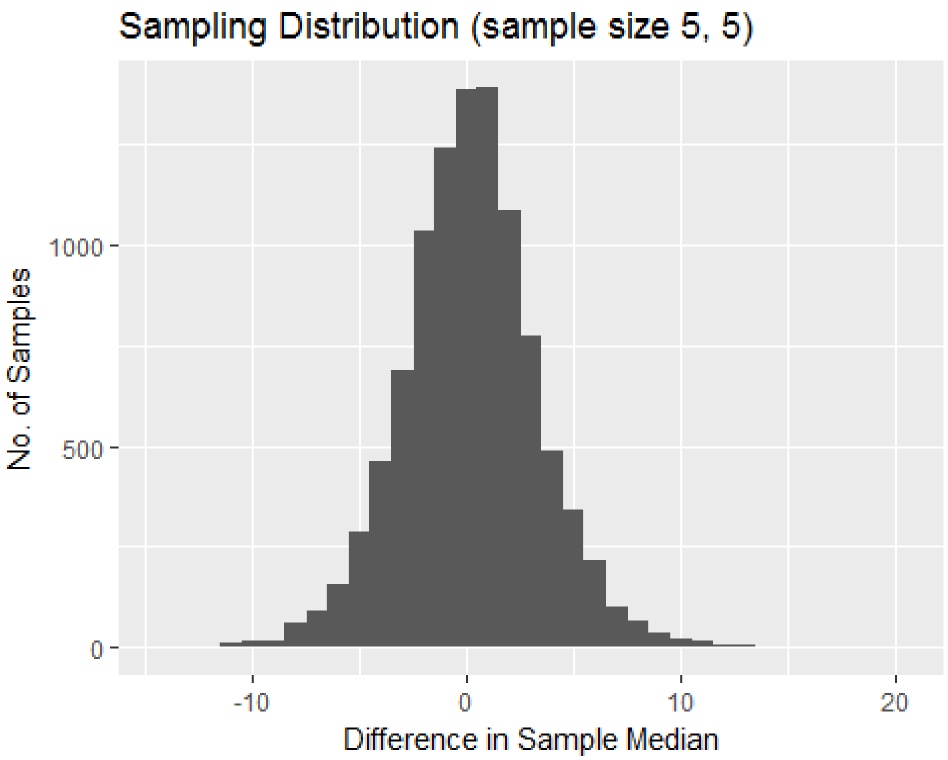
Simulation Study: Difference in Median

The next parameter of interest is the difference in the sample median BMI between 2013 and 2003. We will draw a sampling distribution by using repeated random samples of size (5, 5), (10, 10) and (100, 100) respectively from 2013 and 2003.

```
# 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
}

n_sim = 10000
# 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)
```

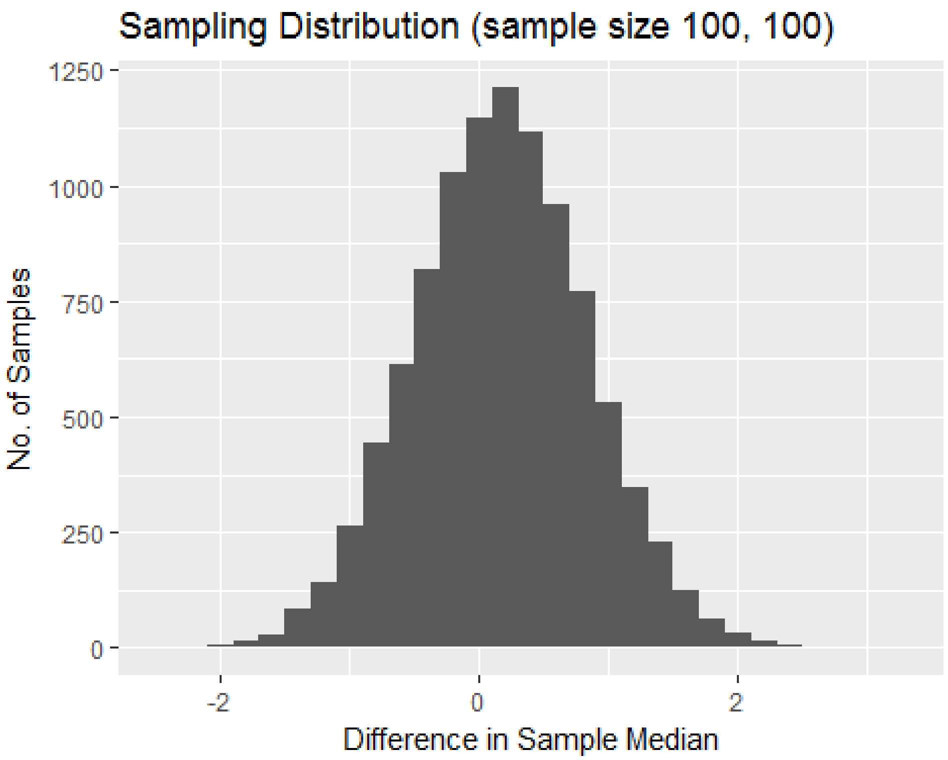
Now that we have got our sampling distributions lets see how they look.

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



This sampling distribution is nearly normal with a few outliers. The graph is also having a large spread.

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



Comparatively, the graph of sample median of sample size (100, 100) is centered around 0 and the spread is smaller with no visible outliers.

Now, let's get the mean and standard deviation of the sampling distributions and see how they change with the sample size. Also, let's get the population median difference of BMI between 2013 and 2003.

```
# 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))

##      SampleSize      Mean_Diff SD_Diff
## [1,] "n1=5,n2=5"     "0.178"   "3.17"
## [2,] "n1=10,n2=10"   "0.216"   "2.107"
## [3,] "n1=100,n2=100" "0.169"   "0.672"

# Population (bmi_2013 and bmi_2003 ) parameters for BMI Index
cbind(Pop_Med_Diff = round(median(bmi_2013) - median(bmi_2003), 3))

##      Pop_Med_Diff
## [1,]        0.207
```

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. The standard deviation decreases as the sample size increases.

Simulation Study: Summary

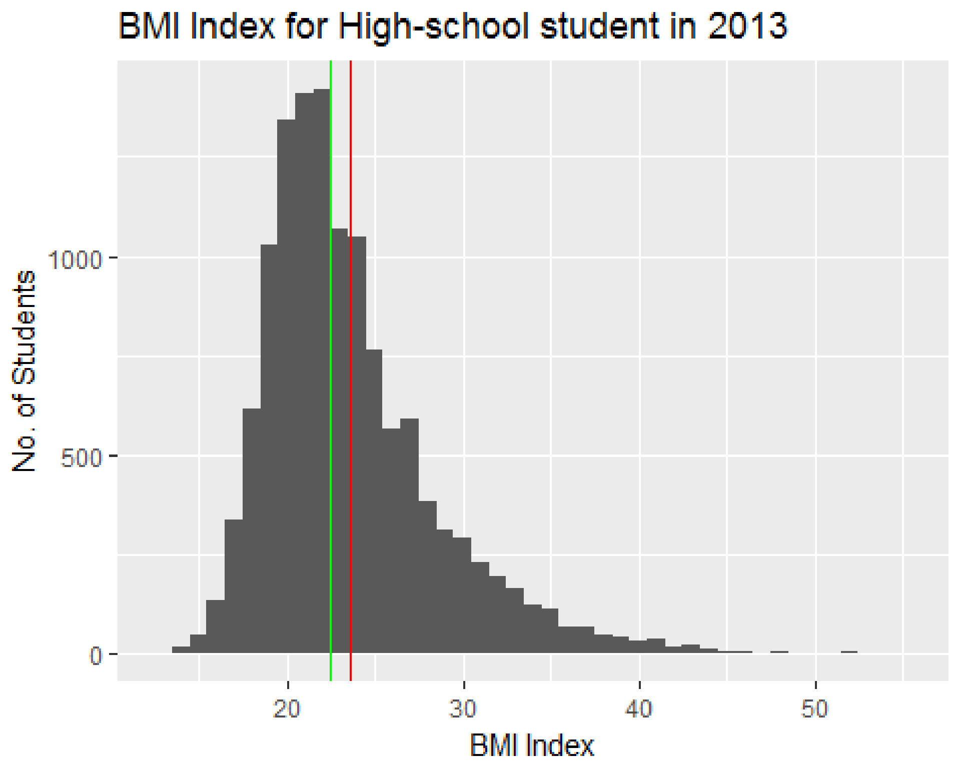
Now that we have seen sampling distributions of various parameters with different sample sizes, we can say that atleast some sample statistic provide a better estimate of the population statistic with large sample size. As we saw, estimates based on sample mean, quantiles and minimum improve with large sample sizes. 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: BMI

We want to study how the BMI of high-school students has changed between 2013 and 2003. We are especially interested in finding if the high-schoolers are getting more overweight.

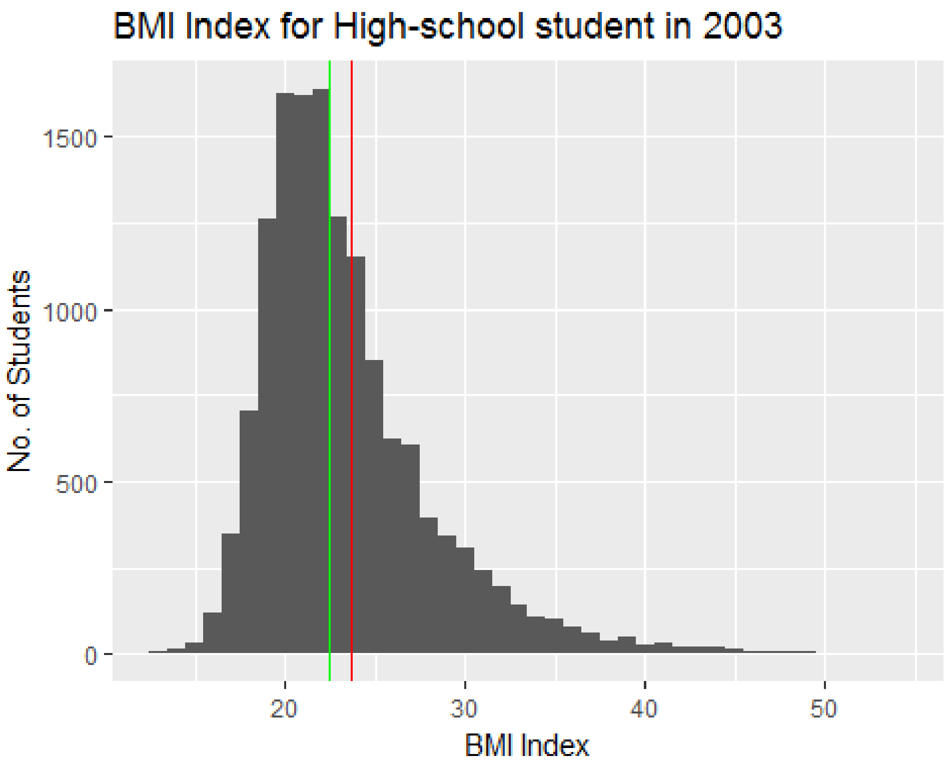
Before we start diving into statistical analysis, let's take a look at the variables of interest: BMI index for high school students in 2013 and 2003.

```r
# let's look the variables of interest

# BMI for High-school students in year 2013
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")
```



We see that the BMI for high school students in 2013 is right skewed.

```
# let's look the variables of interest

# BMI for High-school students in year 2003
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")
```



BMI for high-school students in 2003 is also right skewed.

In order to see if differnce between bmi of high-schoolers from 2013 and 2003 is statistically significant, let's perform Welch's two sample t-test.

```
# 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)

##
##  Welch Two Sample t-test
##
## data:  bmi_2013 and bmi_2003
## t = 3.7529, df = 25988, p-value = 0.0001752
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1079737 0.3440574
## sample estimates:
## mean of x mean of y
##  23.64326  23.41725
```

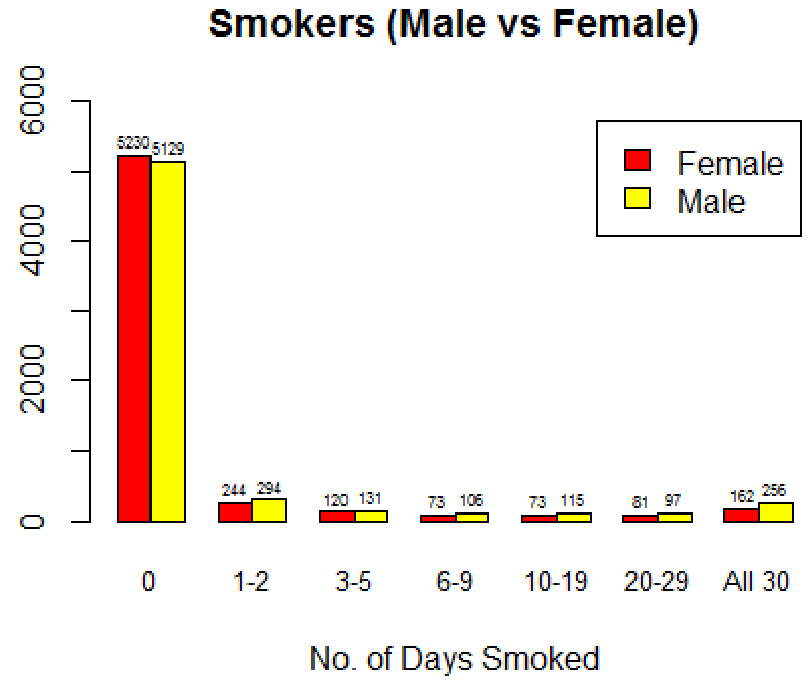
Based on the welch's two sample test we see that the difference in means is not equal to 0 and it seems to have increased in 2013 compared to 2003.

Data Analysis: Smoking Habits

Next we want to identify if male high-scholers are more likely to smoke than female high-schoolers in 2013.

Before we start applying statistical tests to determine the answer, let's take a look at the variable of interest

```r
# 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"),
        cex.names = 0.8,
        xlab = "No. of Days Smoked")
text(bp, smoke, smoke, cex = 0.5, pos = 3, offset = 0.2)
```



As we see from the above plot the number of males high-schoolers who smoke is higher than the number of female high-schoolers that smoke. But we cannot infer only based on this that male high-schoolers smoke more than female high-schoolers. This could simply be because there are more male high-schoolers than female high-schoolers.

So, let's perform a proportions test to determine if there is statistical significance to the difference in proportions of males and females.

In order to perform a proportions test though, we will have to code the number of days smoked to either a smoker or non-smoker. Let's say that anyone smoking for less than 20 days (smokes once every two days) is a non-smoker otherwise a smoker.

```
# 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)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  sex_smoke_table
## X-squared = 18.675, df = 1, p-value = 7.75e-06
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.05677747 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.4984802 0.4077181

# performing two sided proportions test to get confidence intervals
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "t",
          correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  sex_smoke_table
## X-squared = 18.675, df = 1, p-value = 1.55e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05026691 0.13125733
## sample estimates:
##    prop 1    prop 2
## 0.4984802 0.4077181
```

Based on the proportions test we see that the null hypothesis that both the population proportions are same is rejected. We accept the alternative hypothesis that the male high-schoolers are more likely to smoke than female high-schoolers in 2013.

A key point to note with this test was that we have coded and thus grouped students who smoke 20 or more days as smokers. Let's see if a change in coding gives us a different result. This time let's code any one who smokes more than 0 days as smoker.

```
# 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)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  sex_smoke_table
## X-squared = 33.795, df = 1, p-value = 3.062e-09
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.05401528 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5048750 0.4297945

# performing two sided proportions test to get confidence intervals
prop.test(sex_smoke_table, conf.level = 0.95, alternative = "t",
          correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  sex_smoke_table
## X-squared = 33.795, df = 1, p-value = 6.125e-09
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04997975 0.10018118
## sample estimates:
##    prop 1    prop 2 
## 0.5048750 0.4297945
```

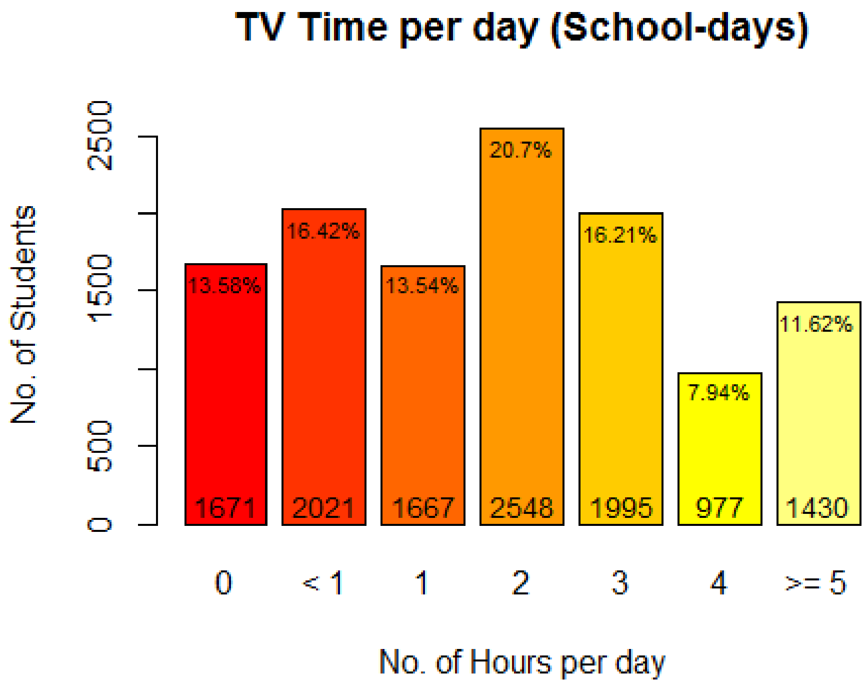
We see that even with this coding the result is same, i.e. male high-schoolers are more likely to smoke than female high-schoolers in 2013.

Data Analysis: TV time

Next, we want to estimate how much TV do high-schoolers watch in 2013.

Just like smoking days, this is also a categorical variable. So let's explore the variable with a plot

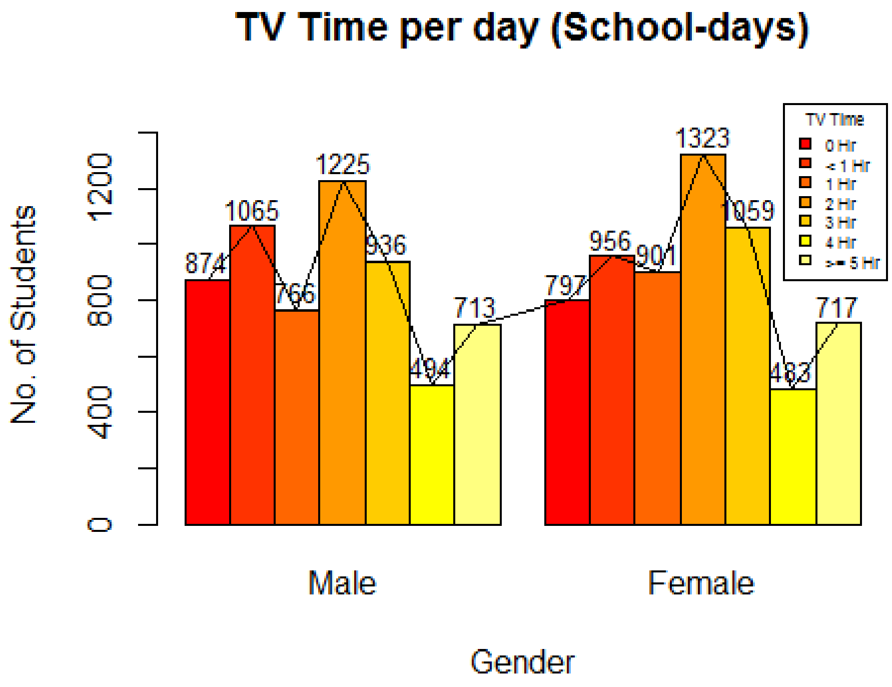
```
# Let's explore the variable of interest
tv_time <- table(yrbss_2013$q81)

# 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=0.7, pos=1, offset = 0.3)
```



Looking at the plot 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 look at this data based on gender of the student and see if there is a difference in TV watching habits of male vs females.

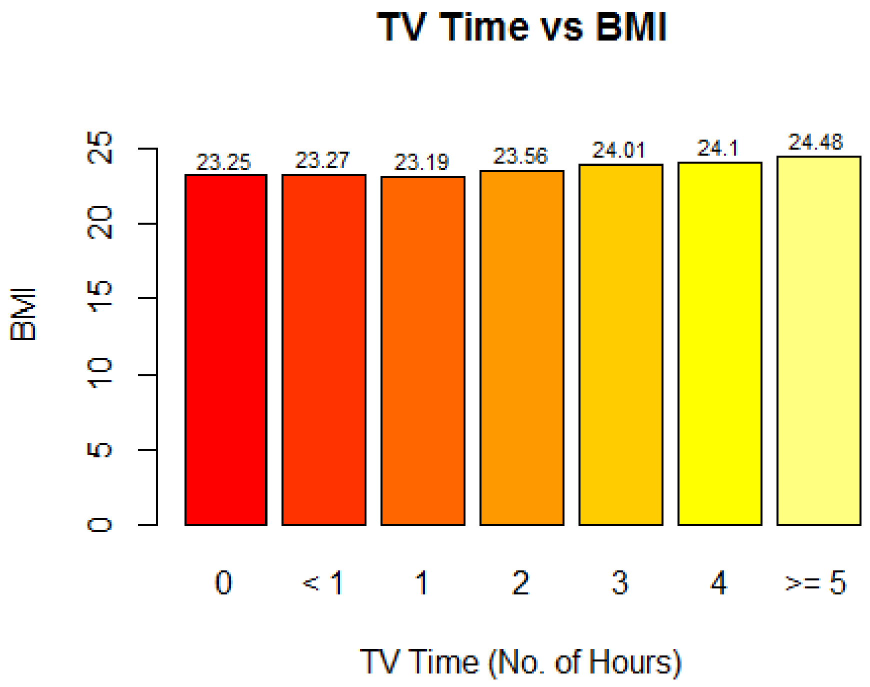
```
# Let's see how TV time differs between male and female students
tv_time_sex <- table(yrbss_2013$q81, yrbss_2013$sex)

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 = 0.5))
text(bp, y = tv_time_sex, tv_time_sex, cex = 0.8, pos=3,
     offset = 0.2)
lines(x = bp, y = tv_time_sex)
```



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

While we are on the subject of TV watching, it will be interesting to see if there is any correlation between TV watching habits and BMI. Is it possible that students who watch more TV have higher mean BMI?

```
# 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$bmi, ylim = c(0,28),
              xlab = "TV Time (No. of Hours)", ylab = "BMI",
              col = heat.colors(7), width = 1.2,
              names.arg = c("0", "< 1", "1",
                            "2", "3", "4",
                            ">= 5"),
              main = "TV Time vs BMI")
text(bp, y = tv_bmi$bmi, round(tv_bmi$bmi,2), cex = 0.7,
     pos=3, offset = 0.2)
```



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 will answer this question in a slightly different way. We will determine if watching TV for more than 2 hours per schoolday results in an increase in BMI.

In order to do this we will have to translate number of hours per day of TV to two categories: 1. Watch TV for 2 or less number of hours
2. Watch TV for more than 2 hours

Then we will perform Welch's two sample t-test to determine statistical significance.

```
# 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)
```

```
##
##  Welch Two Sample t-test
##
## data:  tv_2hr_more and tv_2hr_less
## t = 8.5451, df = 8053.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6449253 1.0289045
## sample estimates:
## mean of x mean of y
##  24.17958  23.34266
```

Based on the two sample t-test there is significant difference between the two samples, in other words there is evidence that the BMI increases for students who watch TV for more than 2 hours per school day in 2013.