Final Project

```
#install.packages("ggplot2")
library(ggplot2)
```

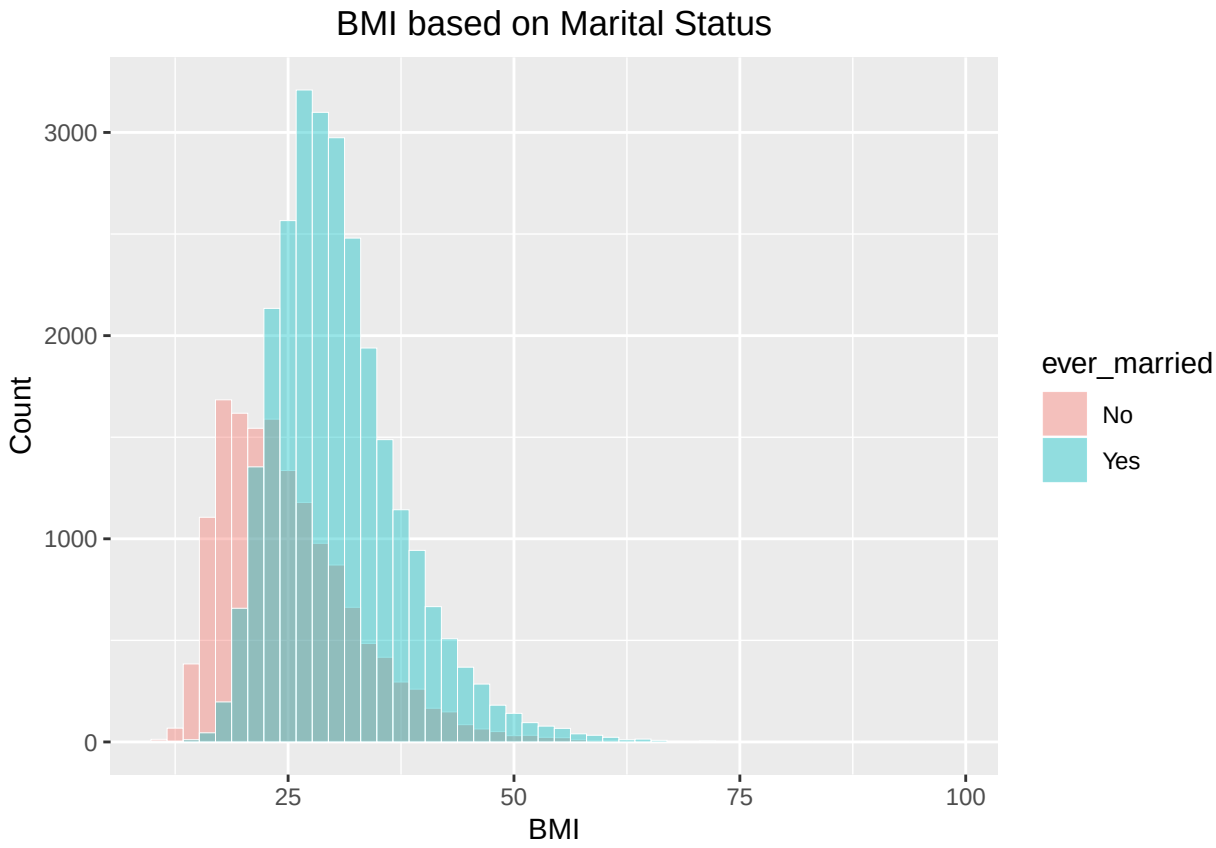
```
dataset <- read.csv("Dataset.csv")
summary(dataset)
```

```
##        id            gender           age         hypertension
##  Min.   :    1   Female:25665   Min.   : 0.08   Min.   :0.00000
##  1st Qu.:18038   Male  :17724   1st Qu.:24.00   1st Qu.:0.00000
##  Median :36352   Other :   11   Median :44.00   Median :0.00000
##  Mean   :36326                  Mean   :42.22   Mean   :0.09357
##  3rd Qu.:54514                  3rd Qu.:60.00   3rd Qu.:0.00000
##  Max.   :72943                  Max.   :82.00   Max.   :1.00000
##
##  heart_disease     ever_married         work_type     Residence_type
##  Min.   :0.00000   No :15462    children     : 6156   Rural:21644
##  1st Qu.:0.00000   Yes:27938    Govt_job     : 5440   Urban:21756
##  Median :0.00000                Never_worked :  177
##  Mean   :0.04751                Private      :24834
##  3rd Qu.:0.00000                Self-employed: 6793
##  Max.   :1.00000
##
##  avg_glucose_level      bmi                 smoking_status      stroke
##  Min.   : 55.00    Min.   :10.10                  :13292   Min.   :0.00000
##  1st Qu.: 77.54    1st Qu.:23.20   formerly smoked: 7493   1st Qu.:0.00000
##  Median : 91.58    Median :27.70   never smoked   :16053   Median :0.00000
##  Mean   :104.48    Mean   :28.61   smokes         : 6562   Mean   :0.01804
##  3rd Qu.:112.07    3rd Qu.:32.90                           3rd Qu.:0.00000
##  Max.   :291.05    Max.   :97.60                           Max.   :1.00000
##                    NA's   :1462
```

```
#Data cleaning: remove id, rows with gender=other, and rows with null bmi
final_data <- subset(dataset, gender != "Other" & !(is.na(bmi)), select = -(id))
```

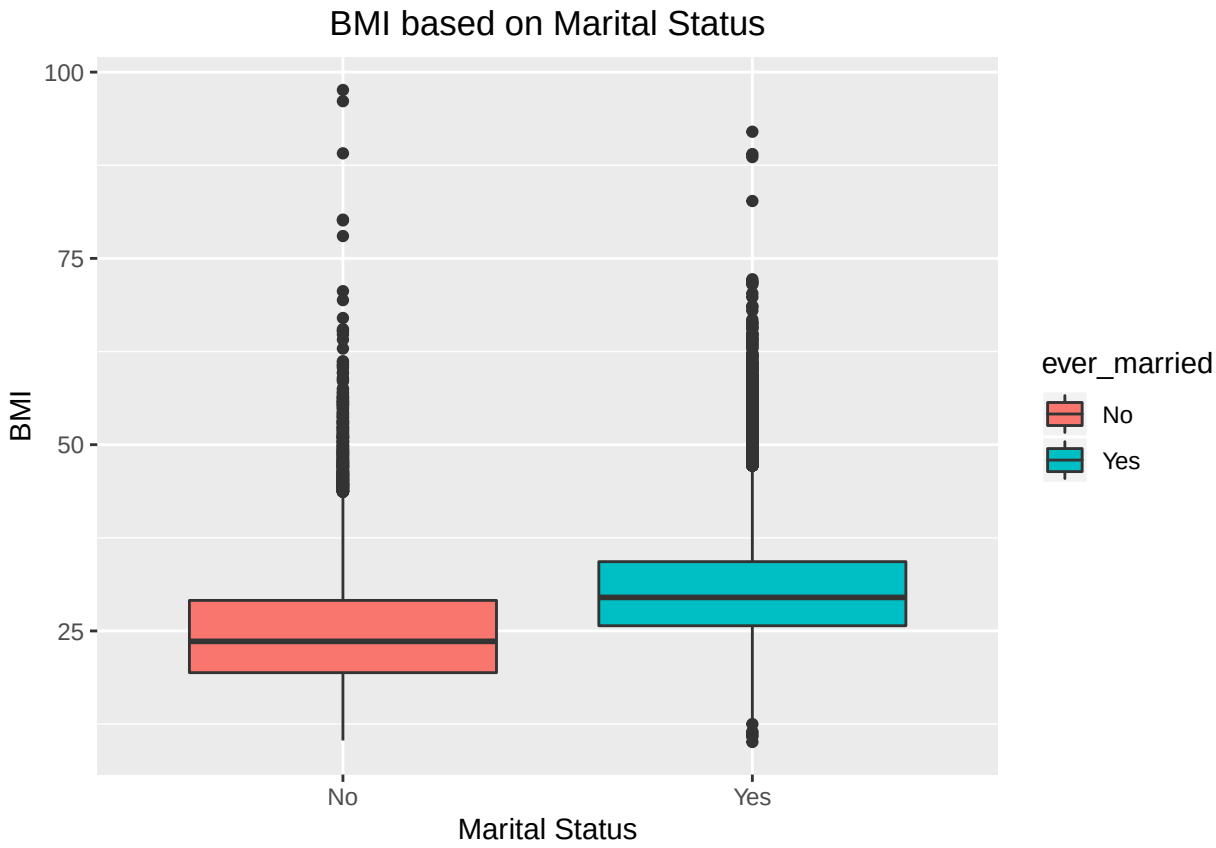
Exploratory Analysis

```
ggplot(final_data, aes(x=bmi, fill = ever_married)) + geom_histogram(bins = 50, alpha = 0.4, position =
                                 title = "BMI based on Marital Status",
                                 x = "BMI",
                                 y = "Count"
                               ) + theme(plot.title = element_text(hjust = .5))
```



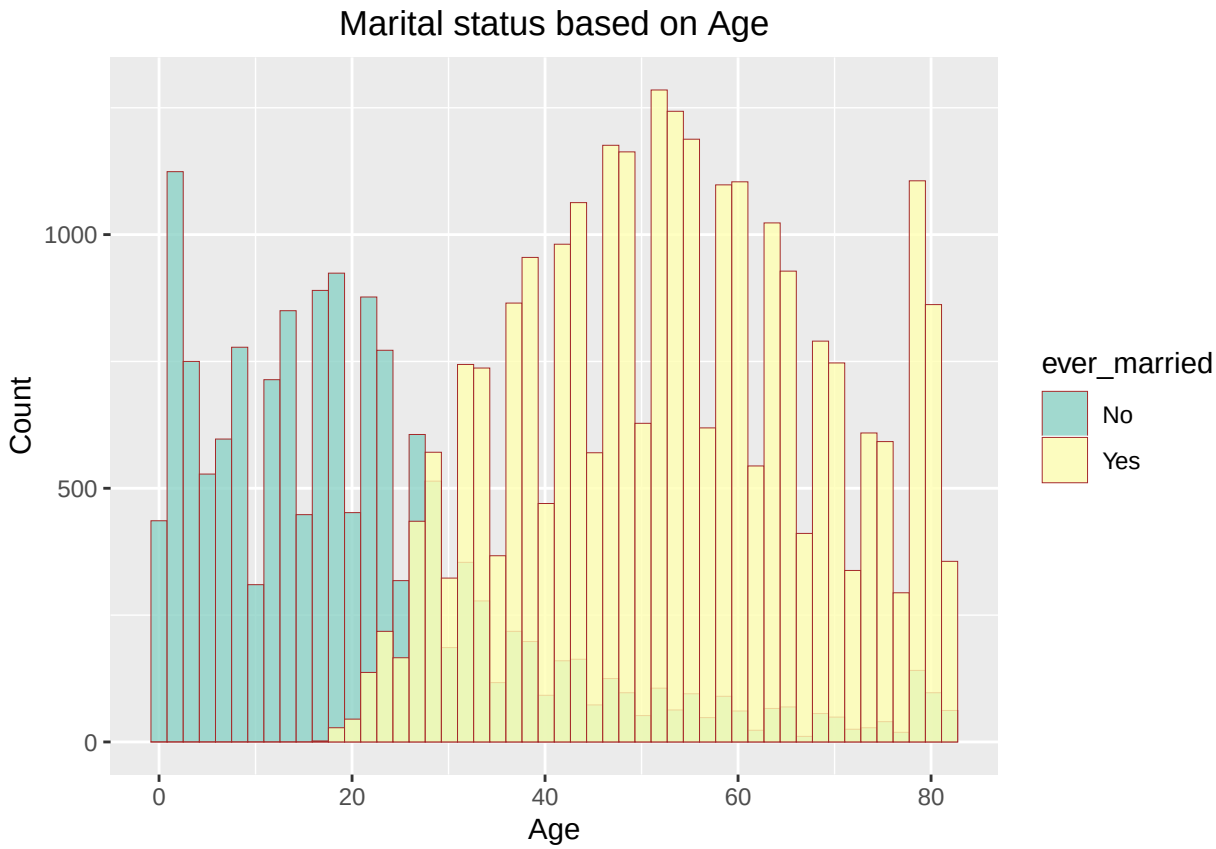

```
# seems like people who are married tend to have higher bmi (gain weight in marriage!!)
```

```
ggplot(final_data, aes(x=ever_married, y = bmi,fill = ever_married)) + geom_boxplot() + labs(
                                  title = "BMI based on Marital Status",
                                  x = "Marital Status",
                                  y = "BMI"
                                ) + theme(plot.title = element_text(hjust = .5))
```

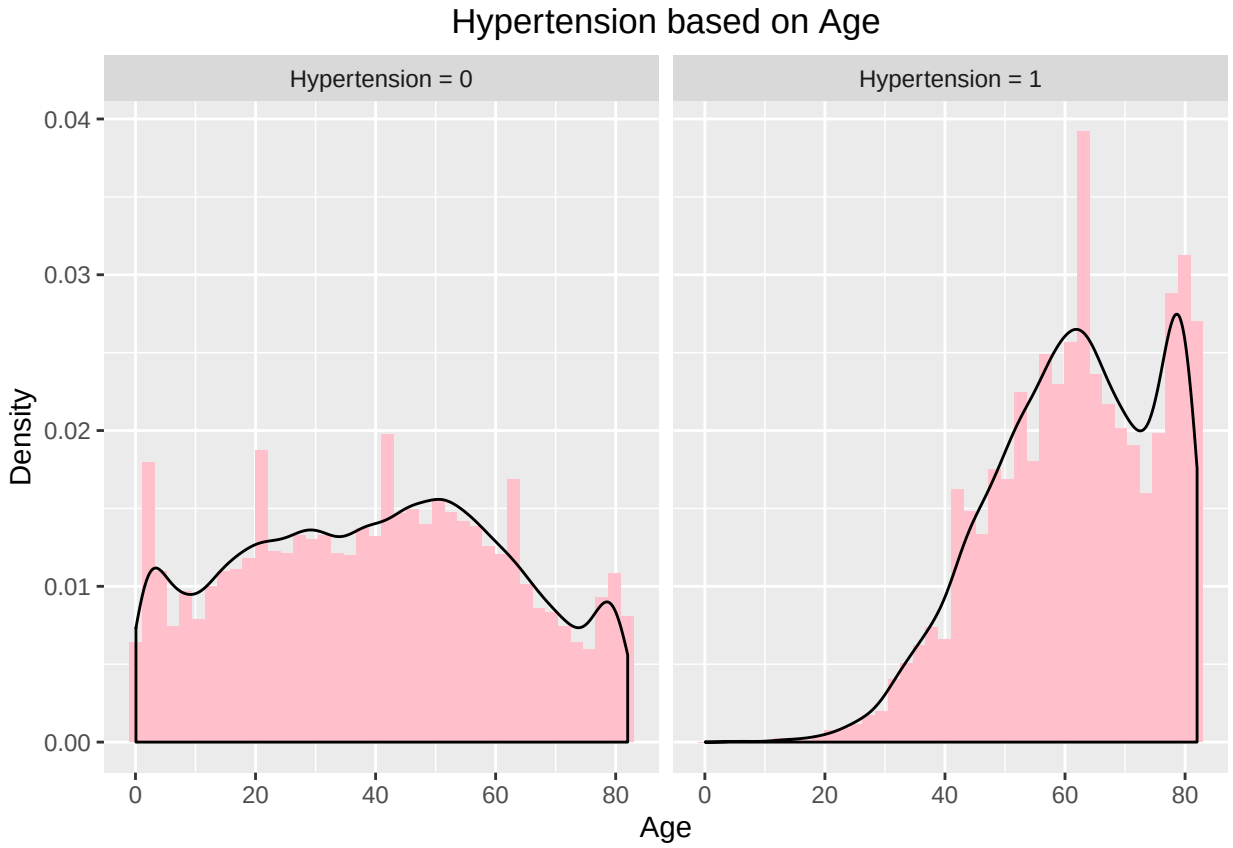



```
# This gives the average BMI of married and not married people, which clearly supports the above plot,
```

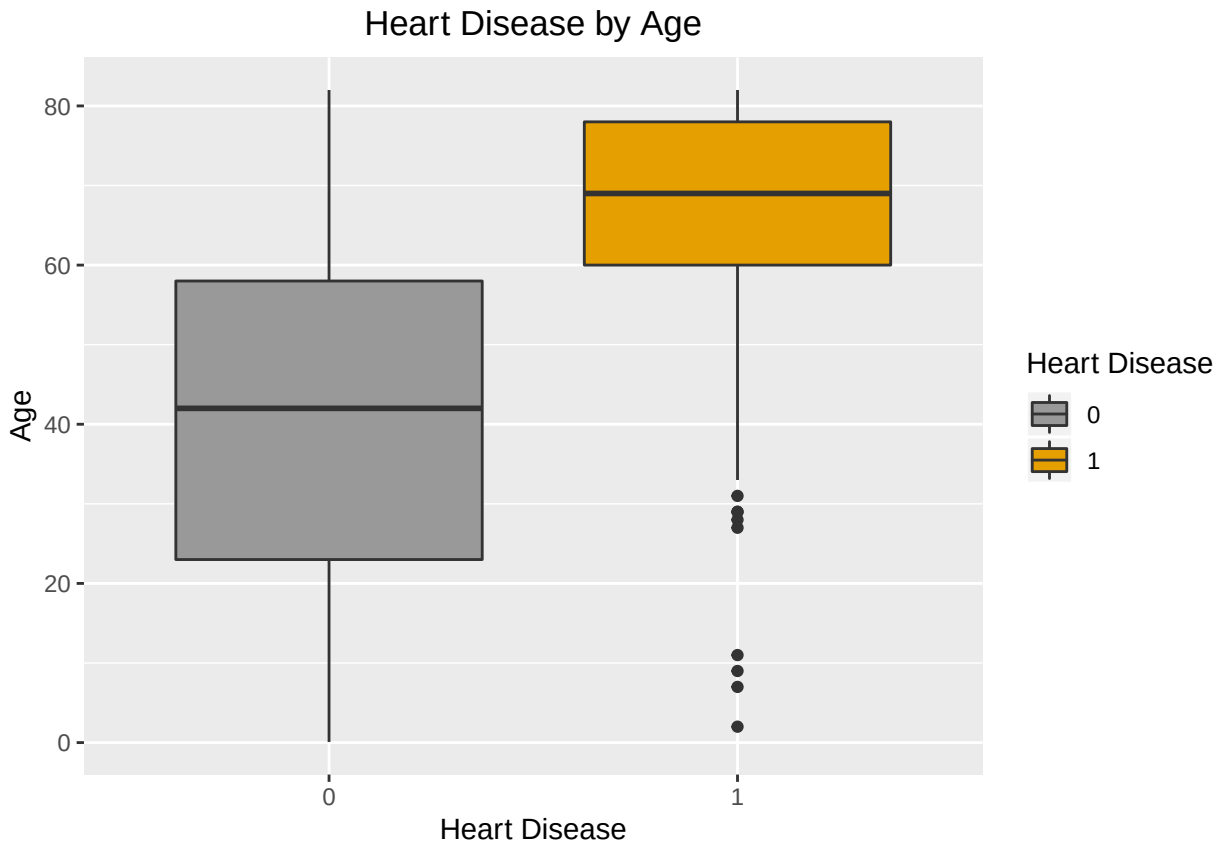
```
# Pretty obvious graph, that most of the people get married after the age of 18!
ggplot(final_data, aes(x=age, fill = ever_married)) + geom_histogram(bins = 50, alpha = 0.8, position =
                                  title = "Marital status based on Age",
                                  x = "Age",
                                  y = "Count"
                                ) + theme(plot.title = element_text(hjust = .5))
```



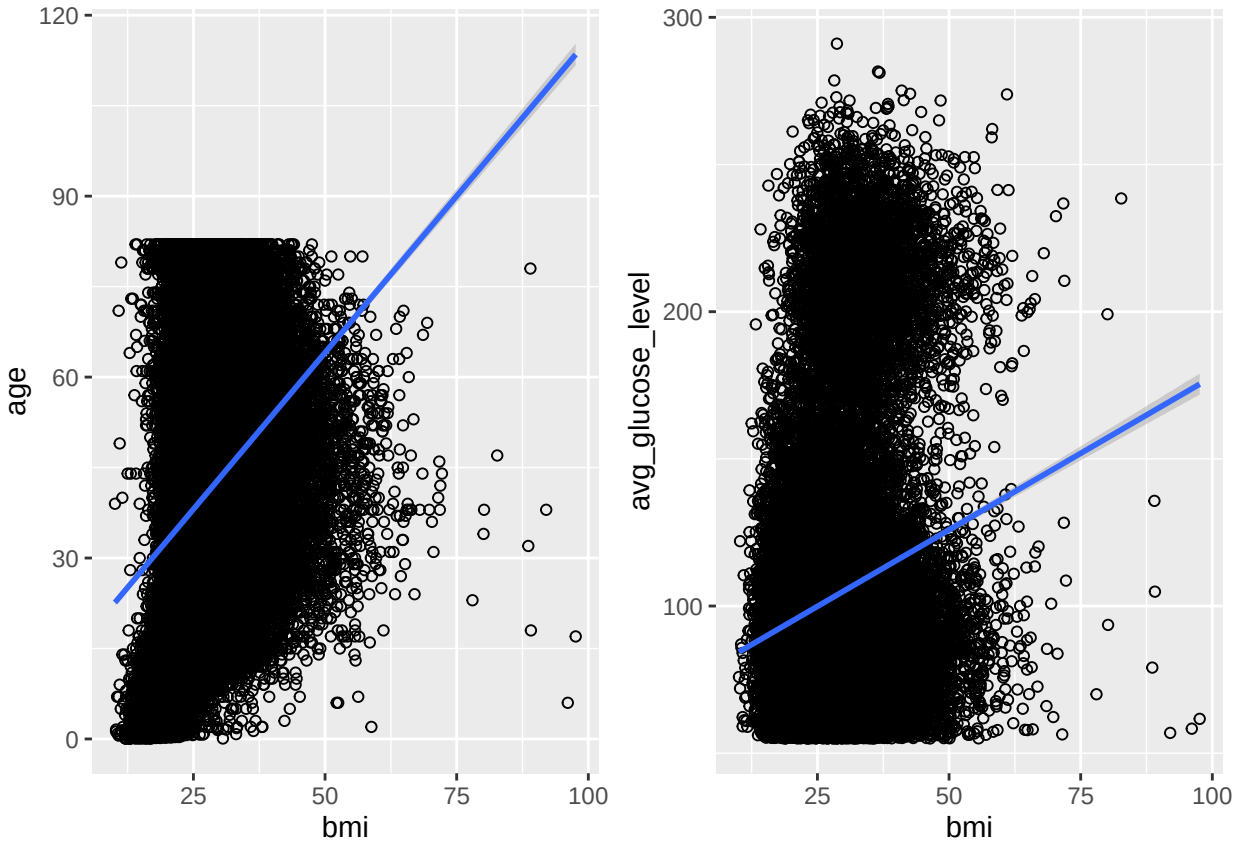

```
labels <- c('0' = "Hypertension = 0", '1' = "Hypertension = 1")
ggplot(final_data, aes(age))+ geom_histogram(bins=40,aes(y=..density..), fill = "pink")+ geom_density(c
                                  title = "Hypertension based on Age",
                                  x = "Age",
                                  y = "Density"
                                ) + theme(plot.title = element_text(hjust = .5))
```



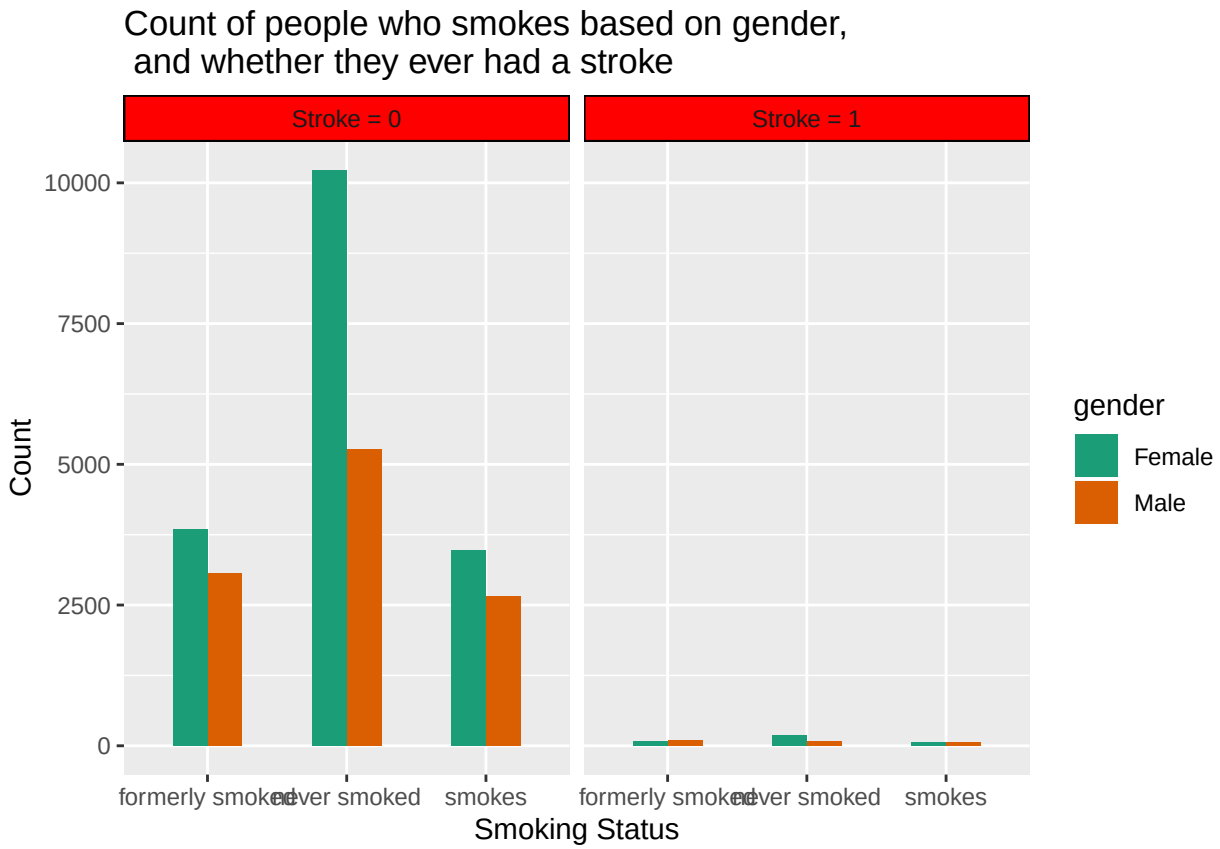

```
ggplot(final_data, aes(as.factor(heart_disease), age, fill = as.factor(heart_disease)))+
                           geom_boxplot()+ scale_fill_manual(name="Heart Disease",values=c("#999999",
                              labs(
                                 title = "Heart Disease by Age",
                                 x = "Heart Disease",
                                 y = "Age"
                               ) +
                               theme(plot.title = element_text(hjust = .5))
```



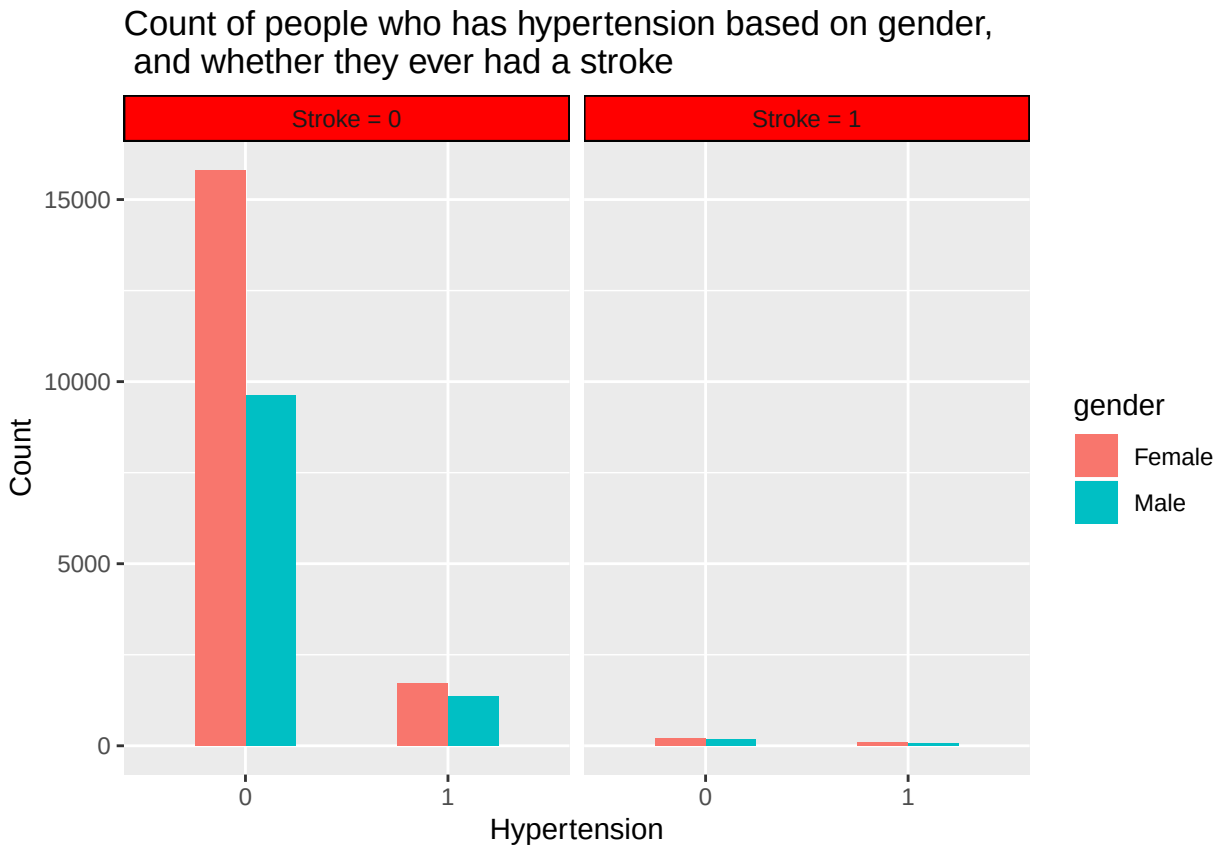

```
#install.packages("gridExtra")
library(gridExtra)
plot_1 <- ggplot(final_data, aes(bmi, age))+  geom_point(shape = 1)+  geom_smooth(method = "lm")
plot_2 <- ggplot(final_data, aes(bmi, avg_glucose_level)) + geom_point(shape = 1) + geom_smooth(method =
grid.arrange(plot_1, plot_2, ncol=2)
```



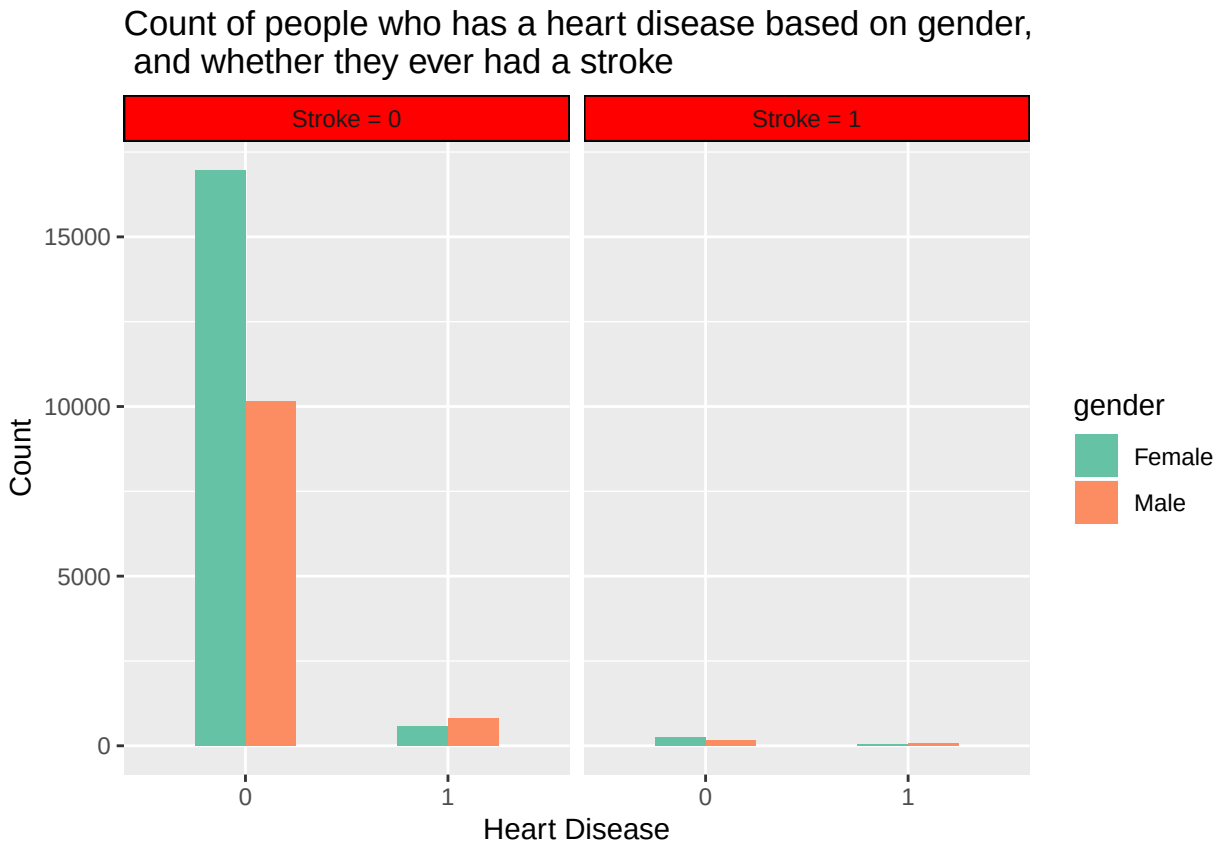

```
data_with_smoke = subset(final_data, smoking_status == 'never smoked' | smoking_status == 'smokes'| smol
labels <- c('0' = "Stroke = 0", '1' = "Stroke = 1")
ggplot(data_with_smoke, aes(fill = gender, x = smoking_status)) + geom_bar(width =0.5,position = "dodge
                              scale_fill_brewer(palette = "Dark2")  + xlab("Smoking Status") + ylab("Cou
```



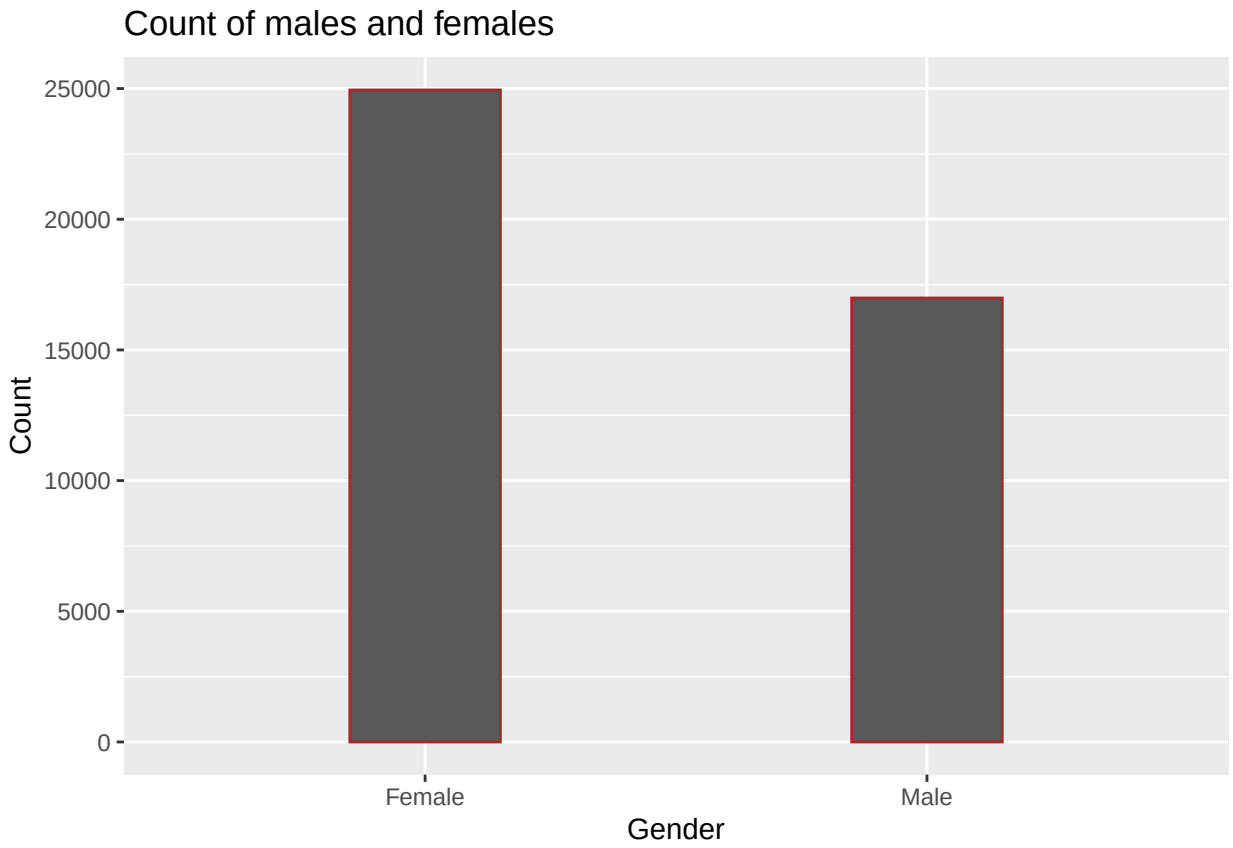

```
ggplot(data_with_smoke, aes(fill = gender, x = as.factor(hypertension))) + geom_bar(width =0.5,position
```




```
ggplot(data_with_smoke, aes(fill = gender, x = as.factor(heart_disease))) + geom_bar(width =0.5,positio
```

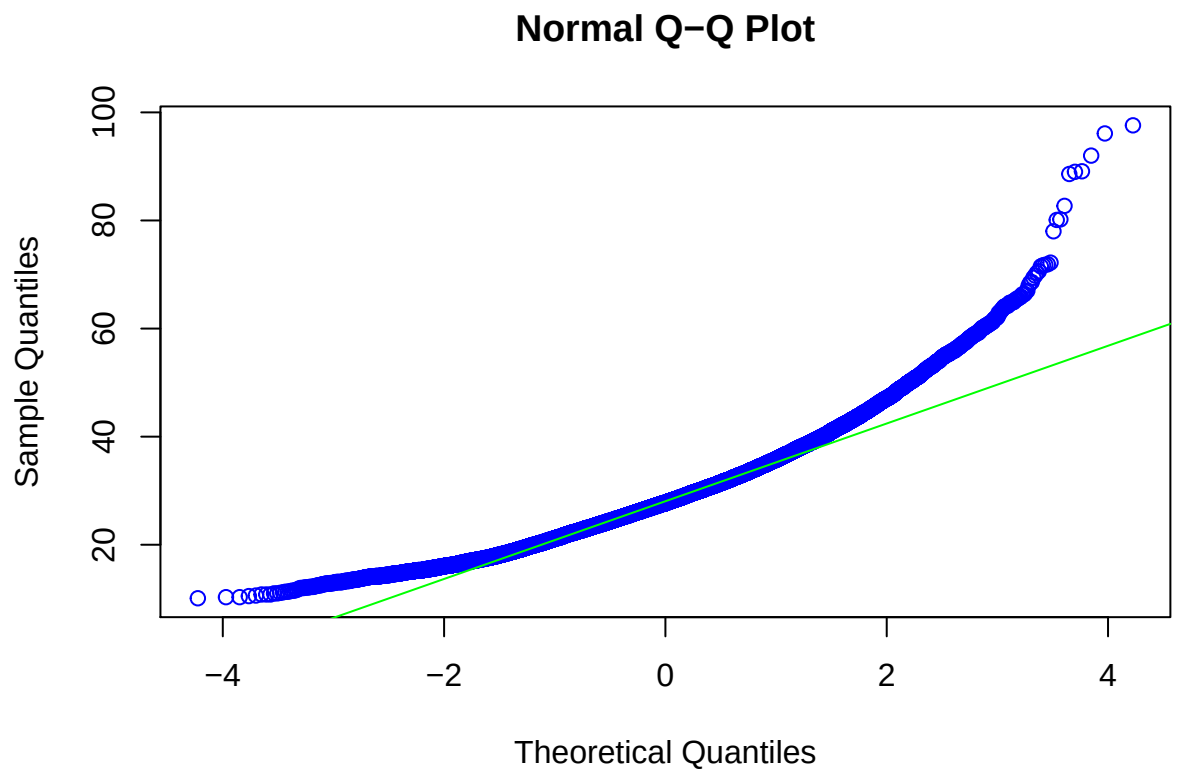



```
ggplot(final_data, aes(x = gender)) + geom_bar(width = 0.3, color = "brown") + scale_fill_brewer(palett
```

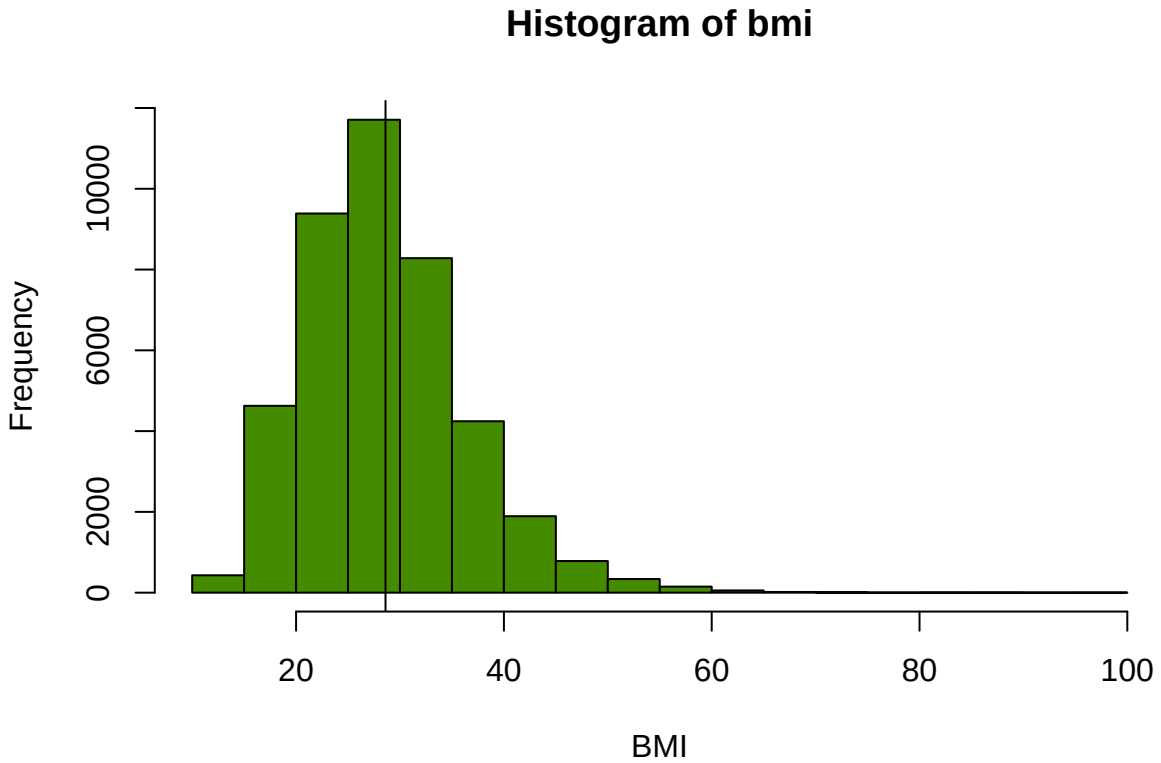


One sample t- Test ## Question: What is the average BMI for all the individuals? Let's say that the average BMI is 21 (normal or healthy weight!)

```
qqnorm(final_data$bmi, col= 'blue')
qqline(final_data$bmi,col='green')
```




```
mx <- mean(final_data$bmi)
hist(final_data$bmi, main = "Histogram of bmi", xlab = "BMI", col = "chartreuse4")
abline(v = mx, col = "black")
```



We want to see that the data fall along the line with most points concentrated in the center of the line and fewer points towards the ends. This is a skewed graph(right-skewed), the normality assumption is not met and so

we should be concerned about the reliability of our p-value and confidence intervals.

Population Parameter

The population parameter we want to make inference to is μ.

Population Variance is unknown- So we will use One sample t-test!

Sample Statistic

The sample statistic is the sample mean bmi for individuals $\overline{x}$

Test Statistic

Test statistic :

$$t_{n-1} = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

Distribution of the test Statistic

$$t_{n-1} = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

##Hypothesis

Two sided:

$$H_o : \mu = 21$$

The Average BMI of our population is 21

$$H_o : \mu \neq 21$$

#Traditional approach!

```
# the parts of the test statistic
# sample mean
x_bar <- mean(final_data$bmi)
# null-hypthesized population mean
mu_0 <- 21
# sample standard deviation
s <- sd(final_data$bmi)
# sample size
n <- length(final_data$bmi)
# t-test test statistic
t <- (x_bar - mu_0)/(s/sqrt(n))
# two-sided p-value- so multiplying by 2
two_sided_t_pval <- pt(q=t, df = n-1, lower.tail = FALSE)*2
two_sided_t_pval
```

```
## [1] 0
```

```
# Since, the p-value is very low, R shows it as 0
```

```
# lower bound
x_bar + (qt(0.025, n-1) * (s/sqrt(n)))
```

```
## [1] 28.53078
```

```
# upper bound
x_bar + (qt(0.975, n-1) * (s/sqrt(n)))
```

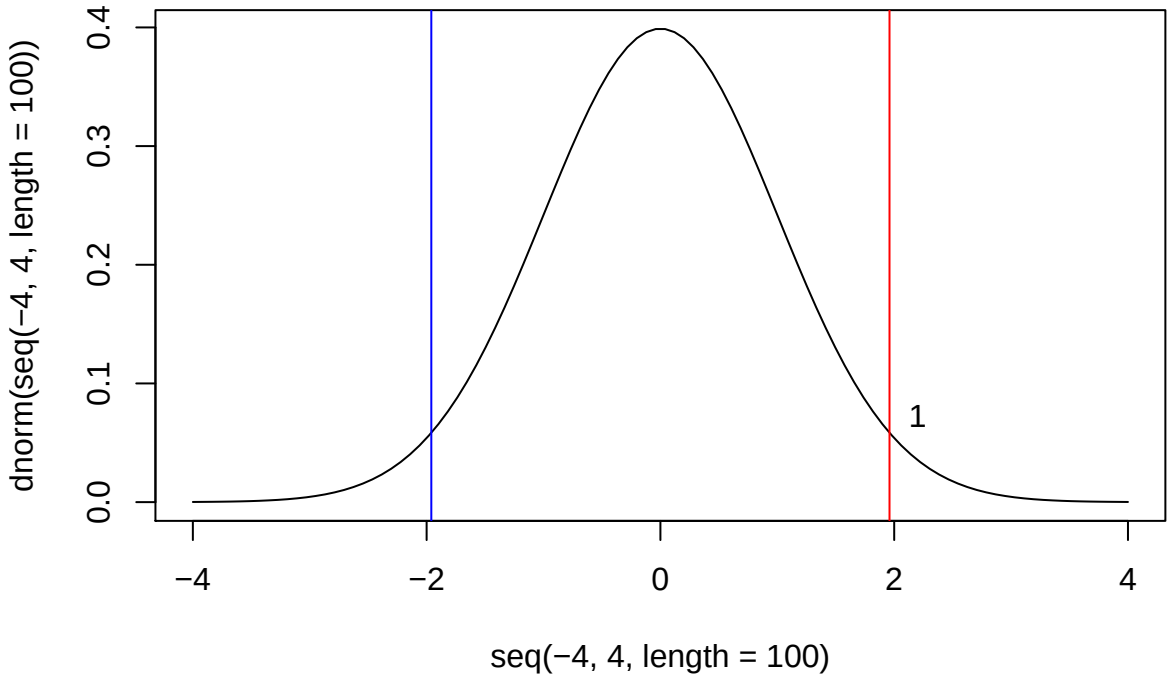
```
## [1] 28.67953
```

```
# Sanity check!
t.test(final_data$bmi, alternative = "two.sided", mu =21)
```

```
##
##  One Sample t-test
##
## data:  final_data$bmi
## t = 200.42, df = 41930, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 21
## 95 percent confidence interval:
##  28.53078 28.67953
## sample estimates:
## mean of x
##  28.60516
```

Plotting Confidence interval

```
plot(x = seq(-4, 4, length = 100), dnorm(seq(-4, 4, length = 100)),type = 'l')
abline(v = qt(0.975, n-1),col = "Red")
abline(v = qt(0.025, n-1), col = "Blue")
text(qt(0.025, n-1), 0.07 , "", srt=0.2,pos=2)
text(qt(0.975, n-1), 0.07,srt=0.2,pos=4)
text(t,.025,"t=-17.389962",srt=0.2,pos=4)
text(-t,.025,"t=17.389962",srt=0.2,pos=2)
```



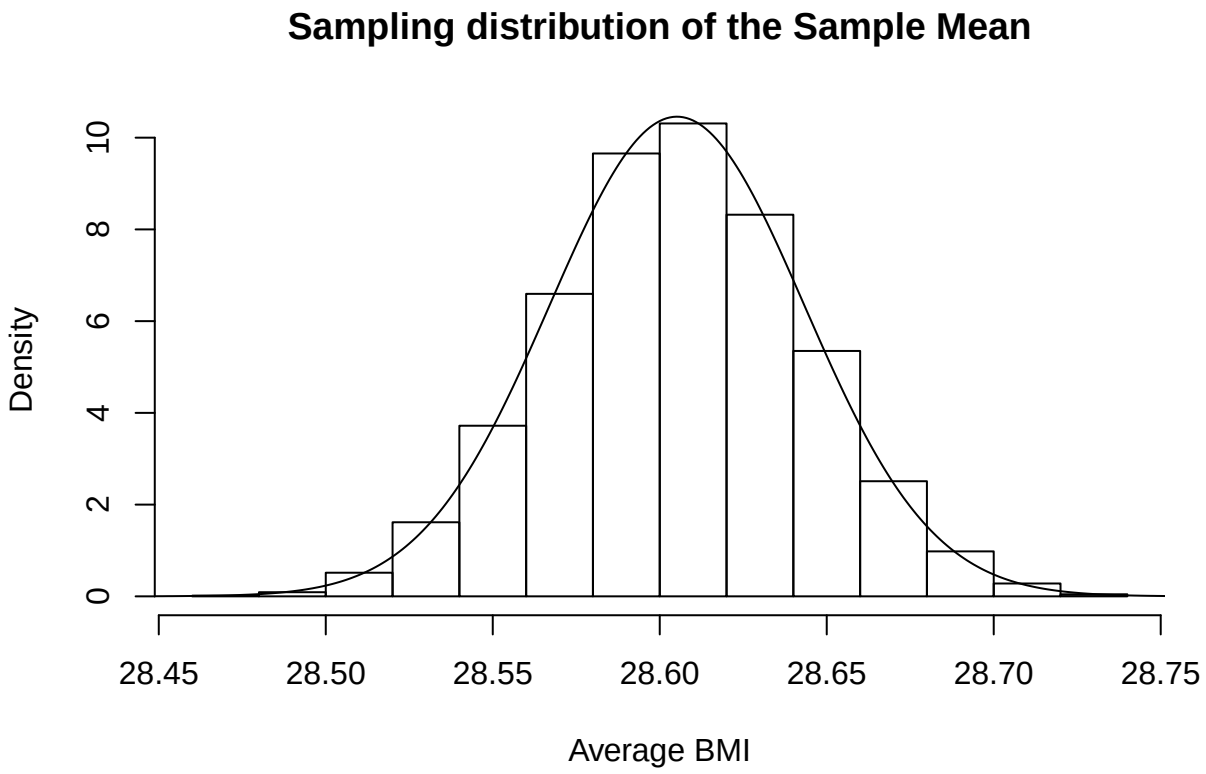
Interpretation

There is a strong evidence (p-value of nearly 0) to suggest that the true mean bmi is different than 21. We reject the null hypothesis that the true mean bmi is 21. With 95% confidence, the true mean bmi is between 28.53078 and 28.67953, which suggests that the true mean bmi is greater than 21.

Bootstrap Approach

```
set.seed(0)
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  results[i]= mean(sample(x= final_data$bmi, size = n, replace = TRUE))
}
```

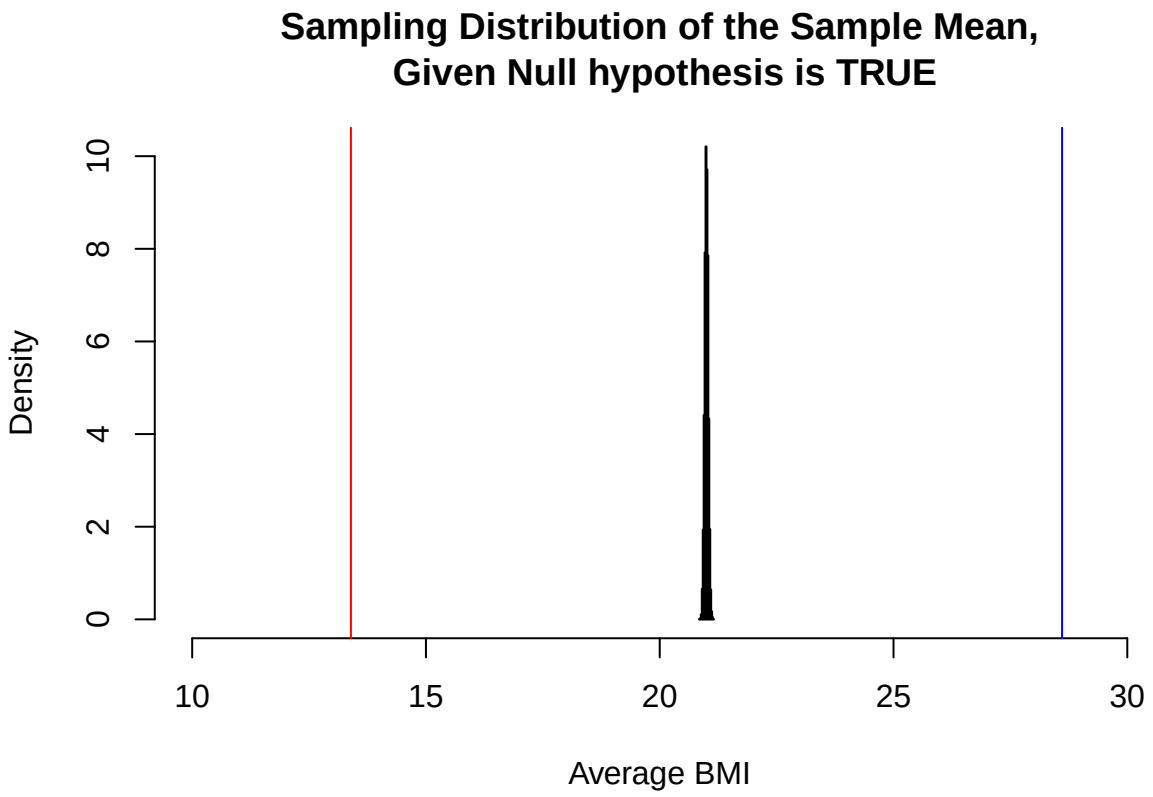
```
# Plot the results
hist(results, freq = FALSE, main = 'Sampling distribution of the Sample Mean', xlab = 'Average BMI', yla

# estimating a normal curve over this
lines(x = seq(28.3, 28.9, 0.001), dnorm(seq(28.3, 28.9, 0.001), mean =x_bar, sd = sd(results)))
```




```
# Shift the mean value so tha the null hypothesis is true
bmi_given_H0_true <- final_data$bmi - mean(final_data$bmi) + mu_0
# This data is really skewed so even though my data is large, I'm going to do a lot of simulations
num_sims <- 10000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
```

```r
# A loop for completing the simulations
for(i in 1:num_sims){
  results_given_H0_true[i] <- mean(sample(x = bmi_given_H0_true, size = n, replace = TRUE))
}
```

```r
hist(results_given_H0_true, freq = FALSE, main = "Sampling Distribution of the Sample Mean,\n Given Null
# Add lines to show values more extreme on upper end
abline(v=x_bar, col = "blue")
# Add lines to show values more extreme on lower end
low_end_extreme <- mean(results_given_H0_true) + (mean(results_given_H0_true)-x_bar)
abline(v=low_end_extreme, col = "red")
```




```r
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= x_bar)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
cat("bootstrap p-value: ",bootstrap_pvalue, "\n")
```

```
## bootstrap p-value:  0
```

```r
cat("two-sided p-value: ",two_sided_t_pval, "\n")
```

```
## two-sided p-value:  0
```

```r
# Bootstrap Confidence interval
# need the standard error which is the standard deviation of the results
bootstrap_SE_X_bar <- sd(results)
# an estimate is to use the formula statistic +/- 2*SE
cat("Bootstrap confidence interval: ", c(x_bar - 2*bootstrap_SE_X_bar, x_bar + 2*bootstrap_SE_X_bar))
```

```
## Bootstrap confidence interval:  28.52886 28.68146
```

```
# you can also use the 5th and 95th quantiles to determine the bounds:
cat("Interval using quantiles: ",c(quantile(results, c(.025, .975))))
```

```
## Interval using quantiles:  28.53044 28.68073
```

```
# compare to our t-methods
cat("Confidence interval using t-method: ",c(x_bar+(qt(0.025, n-1)*(s/sqrt(n))), x_bar+(qt(0.975, n-1)*
```

```
## Confidence interval using t-method:  28.53078 28.67953
```

One-Sample Test of Proportion

What proportion of population has a residence_type as "Urban"? Is it equal to 0.5?

```
x <- sum(final_data$Residence_type == "Urban")
n <- length(final_data$Residence_type)
p_cap <- x/n
p_0 = 0.5
z <- (p_cap - p_0)/sqrt((.5*(1-.5))/n)
p_val <- pnorm(z, lower.tail = FALSE)
p_val
```

```
## [1] 0.3643973
```

```
# confidence interval for one sample proportions
p_cap - qnorm(0.05,lower.tail =  FALSE) * sqrt((.05 *(1-.05))/n)
```

```
## [1] 0.499096
```

```
# one sided upper exact
binom.test(x, n, p=(.5), alternative="greater")
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 21001, number of trials = 41931, p-value = 0.3662
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4968184 1.0000000
## sample estimates:
## probability of success
##              0.5008466
```

```
cat("Confidence Interval using exact binomial test: ", "\n")
```

```
## Confidence Interval using exact binomial test:
```

```
binom.test(x, n , p_0, alternative="greater")$conf.int
```

```
## [1] 0.4968184 1.0000000
## attr(,"conf.level")
## [1] 0.95
```

```
cat("Confidence interval using Normal approximation: ", c(p_cap - z*sqrt(((p_cap)*(1-p_cap))/100), 1),
```

```
## Confidence interval using Normal approximation:  0.4835102 1
```

Bootstrap Approach

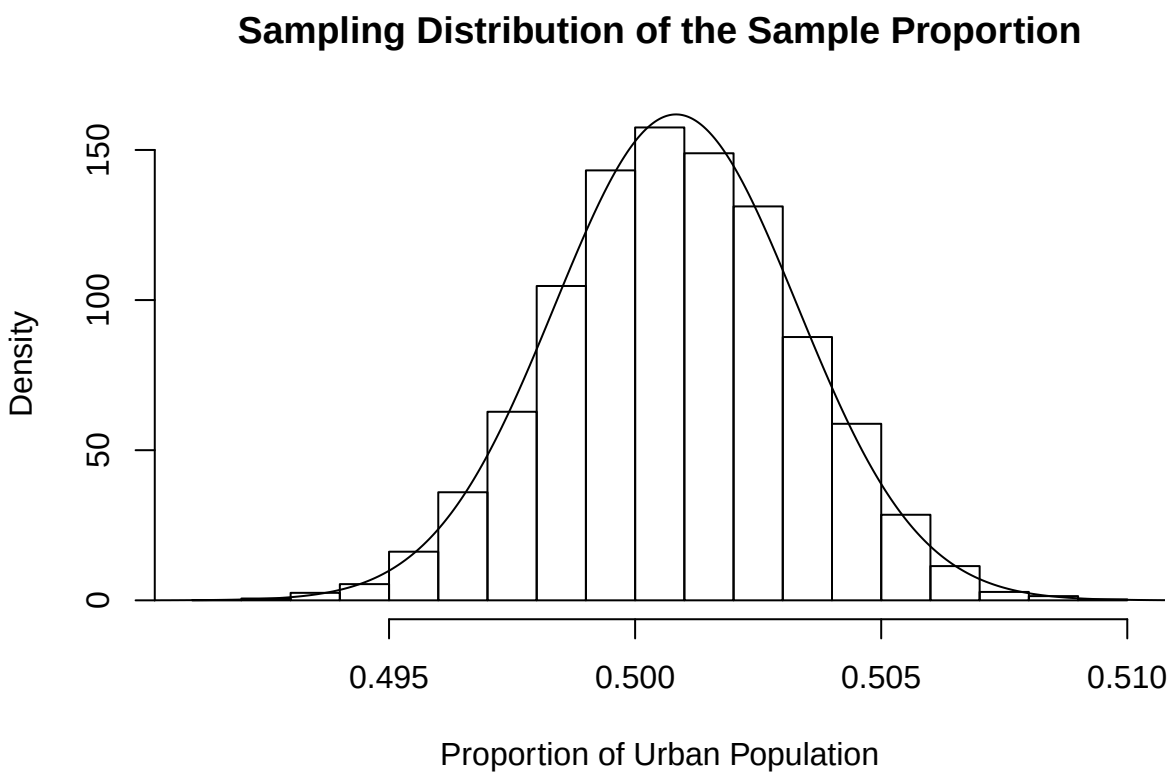
```
urban_population <- factor(rep(c("Urban", "Rural"), c(x, n-x)))
table(urban_population)
```

```
## urban_population
## Rural Urban
## 20930 21001
```

```
urban_population <- rep(c(1, 0), c(x, n-x))
```

```
set.seed(500)
num_sims <- 10000
results <- rep(NA, num_sims)
for(i in 1:num_sims){
          results[i] <- mean(sample(x = urban_population,
                                    size = n,
                                    replace = TRUE))
}
```

```
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion', xlab = 'Proportion o
# estimate a normal curve over it
lines(x = seq(.4, .55, .00001), dnorm(seq(.4, .55, .00001), mean = mean(results), sd = sd
(results)))
```




```
#Using this sampling distribution to find the 5th and 95th percentiles and compare to the other methods
cat("Bootstrap Confidence Interval", "\n")
```

```
## Bootstrap Confidence Interval
```

```r
c(quantile(results, c(.05, 1)))
```

```
##        5%      100%
## 0.4967685 0.5097422
```

```r
# exact binomial test
cat("Exact Binomail Test\n")
```

```
## Exact Binomail Test
```

```r
binom.test(x, n , p_0, alternative="greater")$conf.int
```

```
## [1] 0.4968184 1.0000000
## attr(,"conf.level")
## [1] 0.95
```

```r
cat("Normal Approximation\n")
```

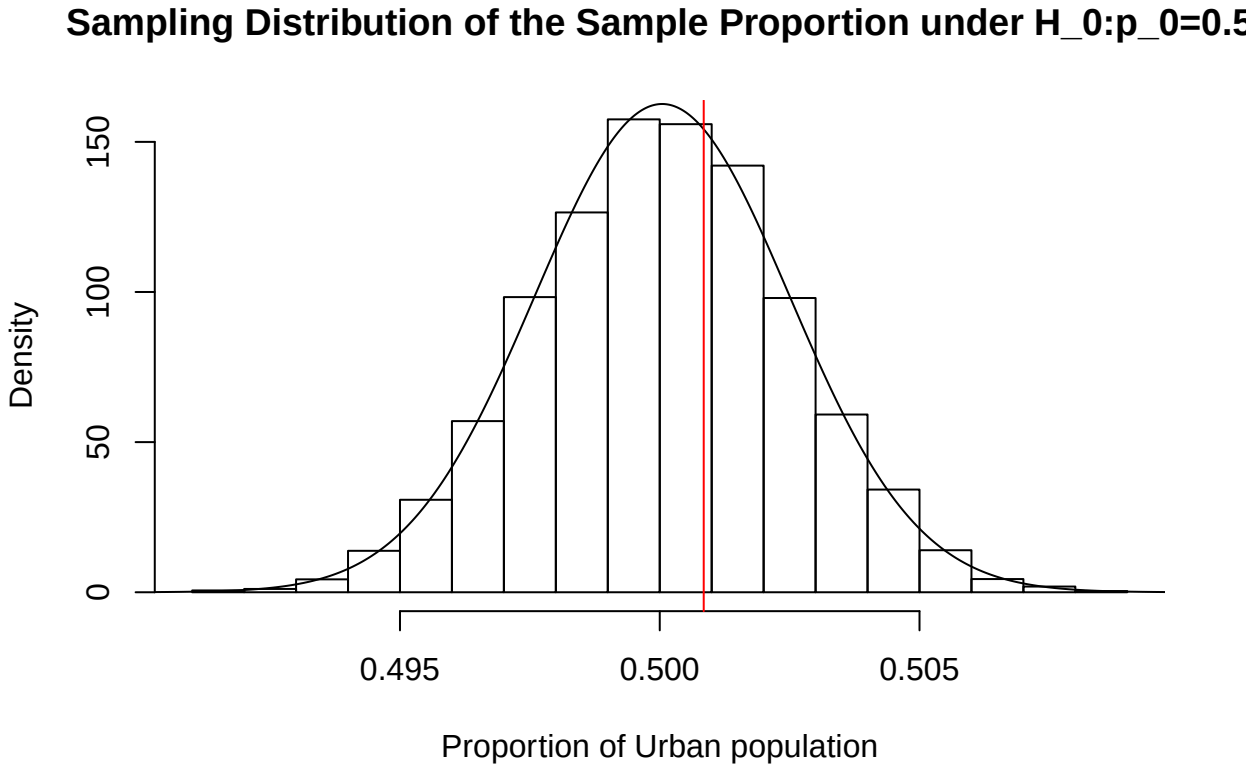
```
## Normal Approximation
```

```r
c(p_cap - z*sqrt(((p_cap)*(1-p_cap))/100), 1)
```

```
## [1] 0.4835102 1.0000000
```

```r
# Under the assumption that the null hypothesis is true, we have 50% Urban population
urban_population <- rep(c(1, 0), c(50, 100-50))
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
                    results[i] <- mean(sample(x = urban_population,
                    size = n,
                    replace = TRUE))
}
```

```r
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion under H_0:p_0=0.5', xla
# estimate a normal curve over it!
lines(x = seq(.4, .55, .00001), dnorm(seq(.4, .55, .00001), mean = mean(results), sd = sd(results)))
abline(v=p_cap, col="red")
```




```
count_of_more_extreme_upper_tail <- sum(results >= p_cap)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
cat("Bootstrap p-value\n")
```

```
## Bootstrap p-value
```

```
bootstrap_pvalue
```

```
## [1] 0.3771
```

```
## Exact Binomial p-value
cat("Exact Binomail p-value\n")
```

```
## Exact Binomail p-value
```

```
binom.test(x, n, p_0, alternative="greater")$p.value
```

```
## [1] 0.3662336
```

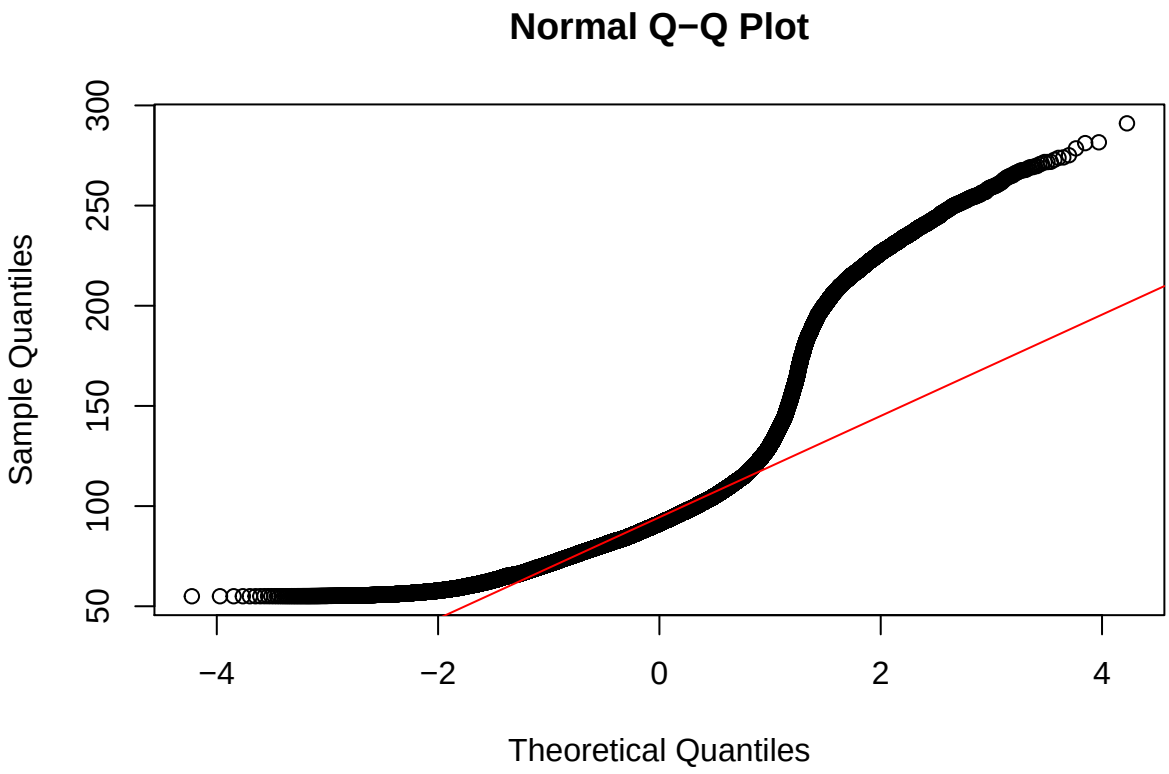
```
c(quantile(results, c(.05, 1)))
```

```
##        5%      100%
## 0.4959815 0.5083351
```

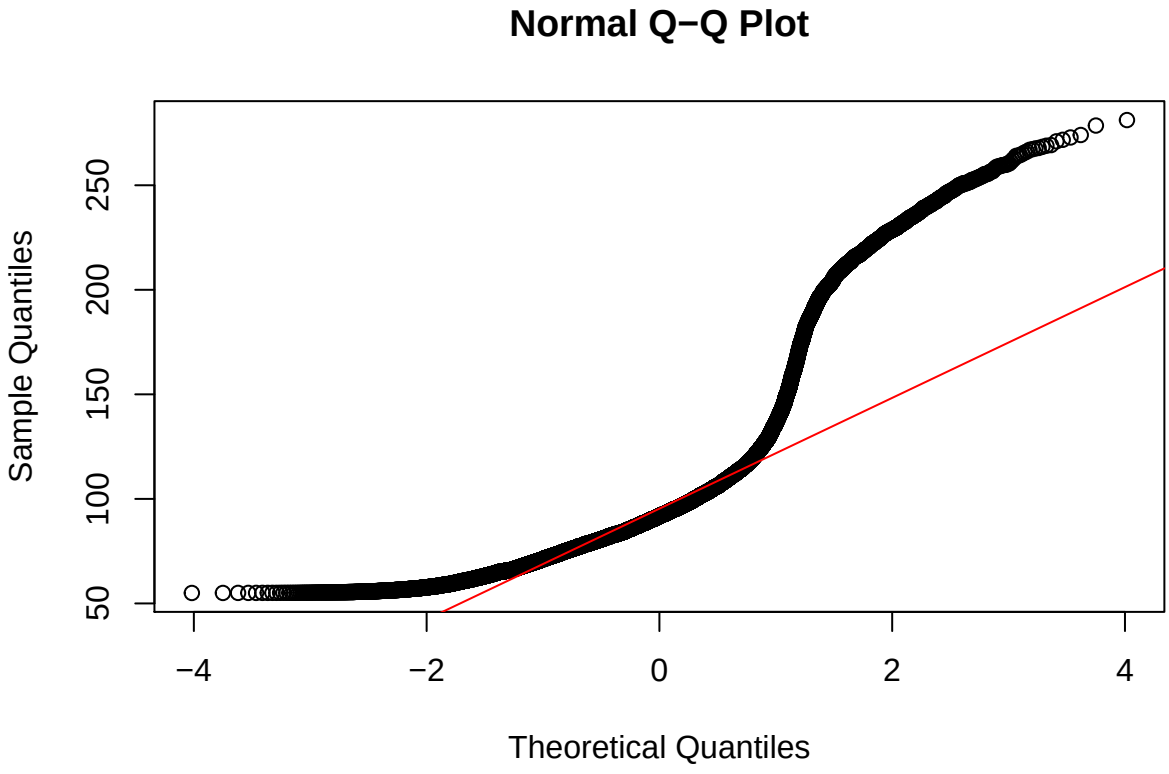
t-Test for Difference in Means

Question: Is there a difference in average glucose level between females and males?

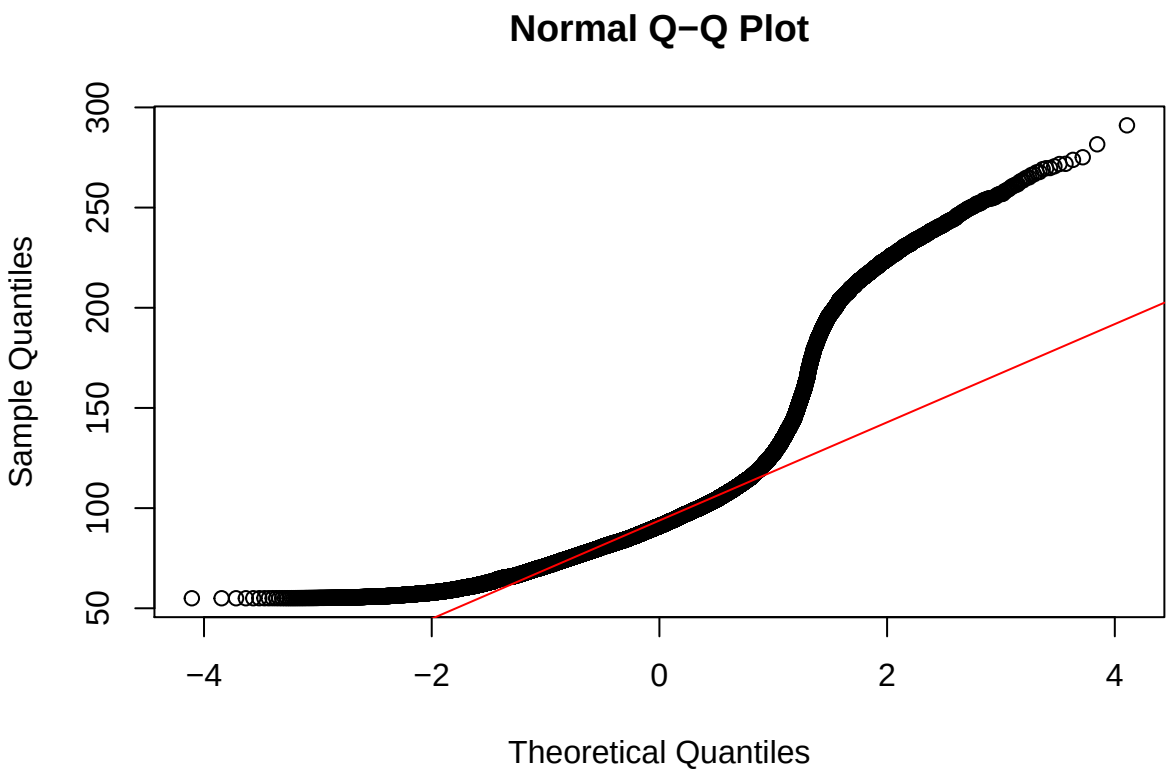
```
qqnorm(final_data$avg_glucose_level)
qqline(final_data$avg_glucose_level, col = "red")
```




```
qqnorm(final_data$avg_glucose_level[final_data$gender == 'Male'])
qqline(final_data$avg_glucose_level[final_data$gender == 'Male'], col="red")
```




```
qqnorm(final_data$avg_glucose_level[final_data$gender == 'Female'])
qqline(final_data$avg_glucose_level[final_data$gender == 'Female'], col="red")
```



###

Mean glucose level of female - mean glucose level of male # Traditional approach

```
x1_bar <- mean(final_data$avg_glucose_level[final_data$gender == 'Female'])
x2_bar <- mean(final_data$avg_glucose_level[final_data$gender == 'Male'])
se <- sd(final_data$avg_glucose_level[final_data$gender == 'Female'])**2/length(final_data$avg_glucose_l

t <- (x1_bar - x2_bar) / sqrt(se)

n <- min(length(final_data$avg_glucose_level[final_data$gender == 'Female']),length(final_data$avg_gluco

two_sided_t_pval <- pt(t, df=n-1, lower.tail = TRUE)*2
two_sided_t_pval
```

```
## [1] 1.031689e-10
```

```
# confidence Interval
c((x1_bar - x2_bar)+qt(0.025,df=n-1)*sqrt(se),(x1_bar - x2_bar)+qt(0.975,df=n-1)*sqrt(se))
```

```
## [1] -3.590117 -1.919891
```

```
# p-value and confidence interval
# two-sided based on alternate hypothesis
t.test(final_data$avg_glucose_level[final_data$gender == 'Female'], final_data$avg_glucose_level[final_d
```

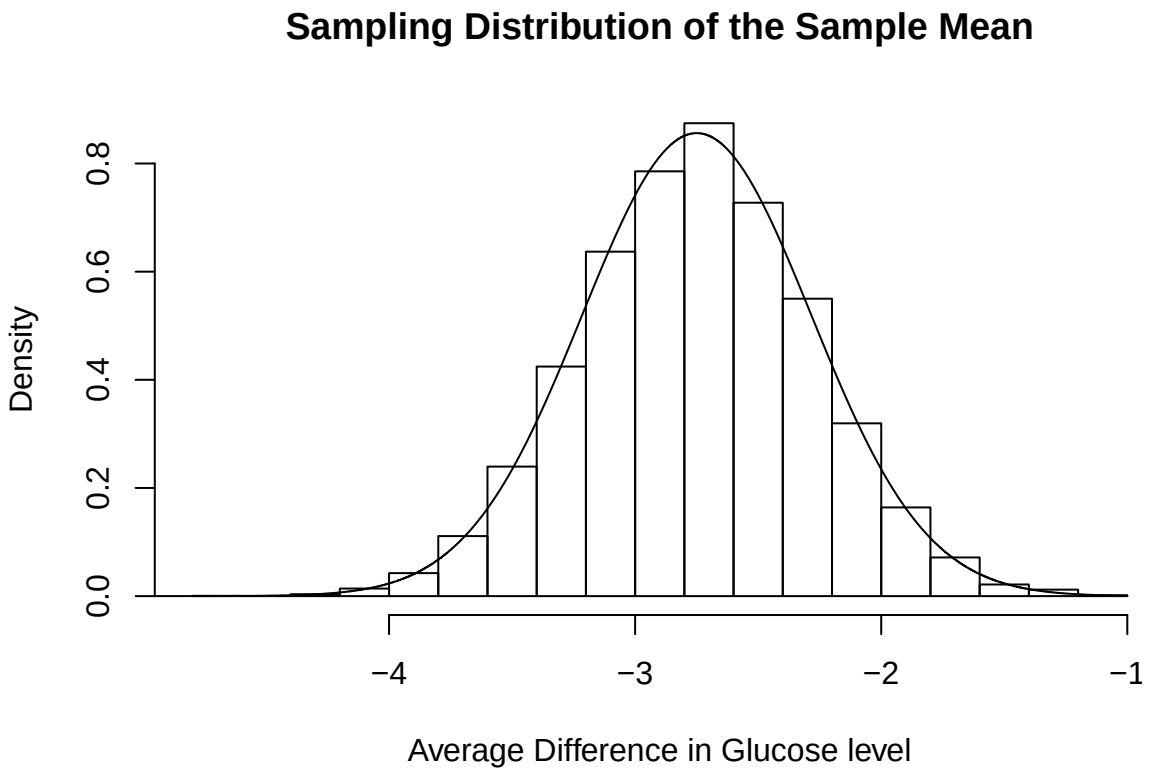
```
##
##  Welch Two Sample t-test
##
## data:  final_data$avg_glucose_level[final_data$gender == "Female"] and final_data$avg_glucose_level[f
## t = -6.4663, df = 34587, p-value = 1.018e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.590087 -1.919921
```

```
## sample estimates:
## mean of x mean of y
##  102.5178  105.2728
```

Bootstrap approach

```
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
                    mean_Female <- mean(sample(x = final_data$avg_glucose_level[final_data$gender == 'F
                    size = n,
                    replace = TRUE))
                    mean_Male <- mean(sample(x = final_data$avg_glucose_level[final_data$gender == 'Mal
                            size = n,
                            replace = TRUE))
                results[i] <- mean_Female - mean_Male
}
```

```
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Difference
lines(x = seq(-5, -1, .0001), dnorm(seq(-5, -1, .0001), mean = mean(results), sd = sd(results)))
```




```
# Bootstrap CI
c(quantile(results, c(.025, .975)))
```

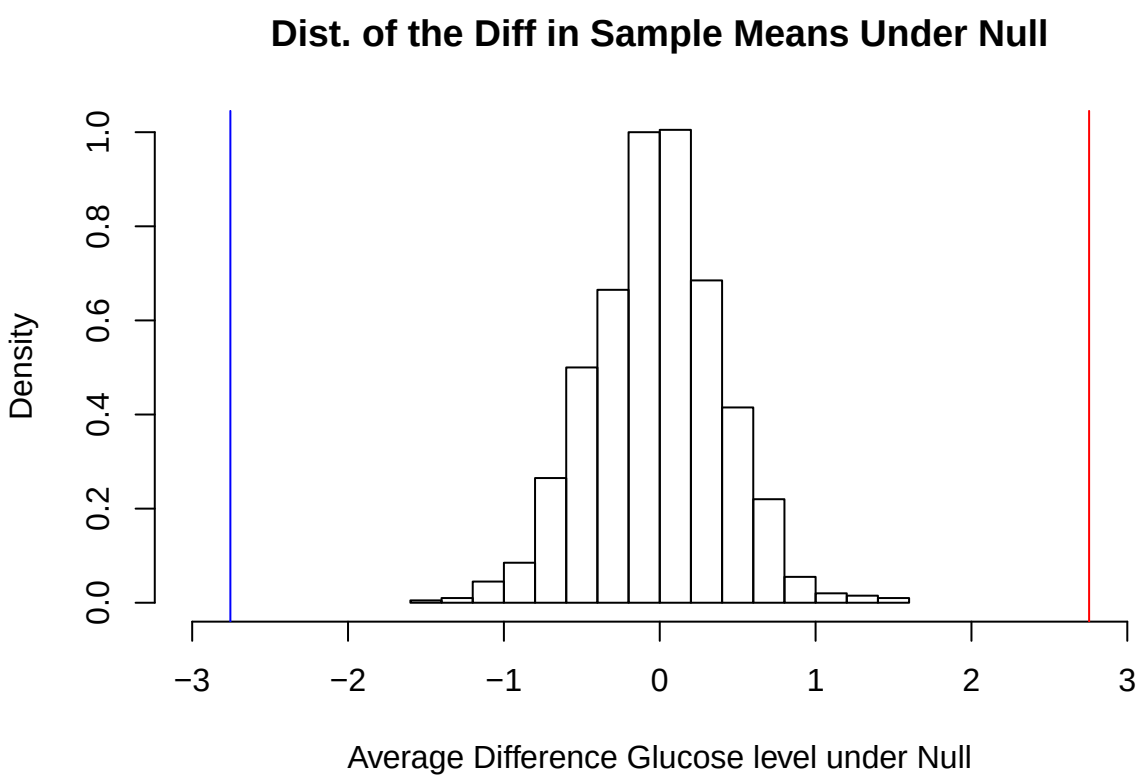
```
##      2.5%     97.5%
## -3.661286 -1.842207
```

```r
# compare to our t-methods
t.test(final_data$avg_glucose_level[final_data$gender == 'Female'], final_data$avg_glucose_level[final_
```

```
## [1] -3.590087 -1.919921
## attr(,"conf.level")
## [1] 0.95
```

```r
num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
# idea here is if there is no relationshipm we should be able to shuffle the groups
                    shuffled_groups <- transform(final_data, gender=sample(final_data$gender))
                    mean_Female <- mean(shuffled_groups$avg_glucose_level[shuffled_groups$gender=="Fema
                    mean_Male <- mean(shuffled_groups$avg_glucose_level[shuffled_groups$gender=="Male"]
                    results_given_H0_true[i] <- mean_Female - mean_Male
}
```

```r
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,main='Dist. of the Diff in Sample Means Under Null',xlab = 'Av
diff_in_sample_means <- mean(final_data$avg_glucose_level[final_data$gender=="Female"]) - mean(final_da
abline(v=diff_in_sample_means, col = "blue")
abline(v=abs(diff_in_sample_means), col = "red")
```




```r
# counts of values more extreme than the test statistic in our original sample, given H0  is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= abs(diff_in_sample_means))
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
```

```r
cat("Bootstrap p-value: ", bootstrap_pvalue, "\n")
```

```
## Bootstrap p-value:  0
```

```r
cat("Bootstrap confidence interval: ",c(quantile(results, c(0.25, 0.975))),"\n")
```

```
## Bootstrap confidence interval:  -3.062212 -1.842207
```

Difference in Proportions!

Question: Is the proportion of males who experienced stroking same as the proportion of females who experienced stroke?

```r
# the parts of the test statistic
# sample props
p_hat_men <- sum(final_data$gender == 'Male' & final_data$stroke == 1)/sum(final_data$gender == 'Male')
p_hat_women <- sum(final_data$gender == 'Female' & final_data$stroke == 1)/sum(final_data$gender == 'Fe
# null hypothesized population prop difference between the two groups
p_0 <- 0
# sample size
n_men <- sum(final_data$gender == 'Male')
n_women <- sum(final_data$gender == 'Female')
# sample variances
den_p_men <- (p_hat_men*(1-p_hat_men))/n_men
den_p_women <- (p_hat_women*(1-p_hat_women))/n_women
# z-test test statistic
z <- (den_p_men - den_p_women - p_0)/sqrt(den_p_men + den_p_women)
# two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval
```

```
## [1] 0.999737
```

```r
# lower bound
(p_hat_men - p_hat_women)+(qnorm(0.025)*sqrt(den_p_men + den_p_women))
```

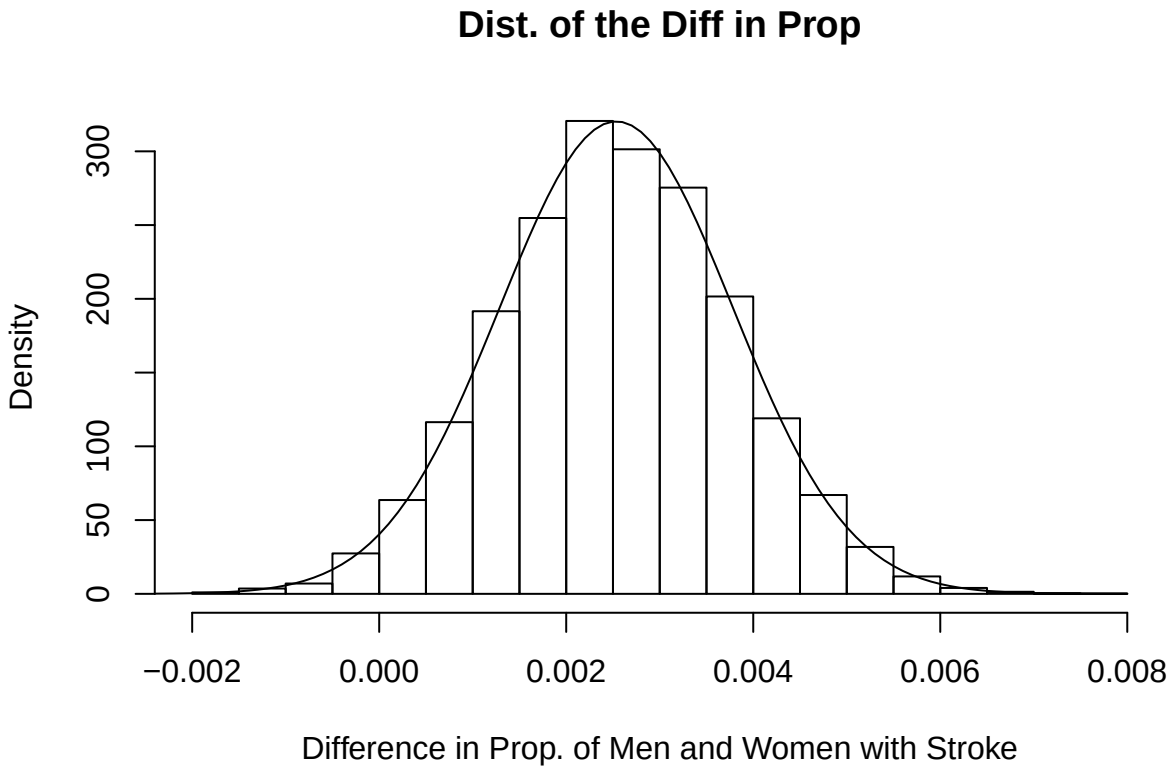
```
## [1] 9.360199e-05
```

```r
# upper bound
(p_hat_men - p_hat_women)+(qnorm(0.975)*sqrt(den_p_men + den_p_women))
```

```
## [1] 0.004958218
```

```r
# Bootstrap + Randomization Approach
# Make the data
men <- rep(c(1, 0), c(sum(final_data$gender == 'Male' & final_data$stroke == 1), n_men - sum(final_data
women <- rep(c(1,0), c(sum(final_data$gender == 'Female' & final_data$stroke == 1), n_women - sum(final_
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
prop_men <- mean(sample(men, size = n_men, replace = TRUE))
prop_women <- mean(sample(women, size = n_women, replace = TRUE))
results[i] <- prop_men - prop_women
}
```

```r
# Finally plot the results
hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Difference in Prop. of Men and Wo
lines(x = seq(-0.004, 0.008, .0001), dnorm(seq(-0.004, 0.008, .0001), mean = mean(results), sd = sd(resu
```




```r
cat("Bootstrap", "\n")
```

```
## Bootstrap
```

```r
## Bootstrap
c(quantile(results, c(.025, .975)))
```

```
##         2.5%        97.5%
## 0.0001320469 0.0049873471
```

```r
cat("Normal Approximation", "\n")
```

```
## Normal Approximation
```

```r
c((p_hat_men - p_hat_women)+(qnorm(0.025)*sqrt(den_p_men + den_p_women)), (p_hat_men - p_hat_women)+(qn
```

```
## [1] 9.360199e-05 4.958218e-03
```

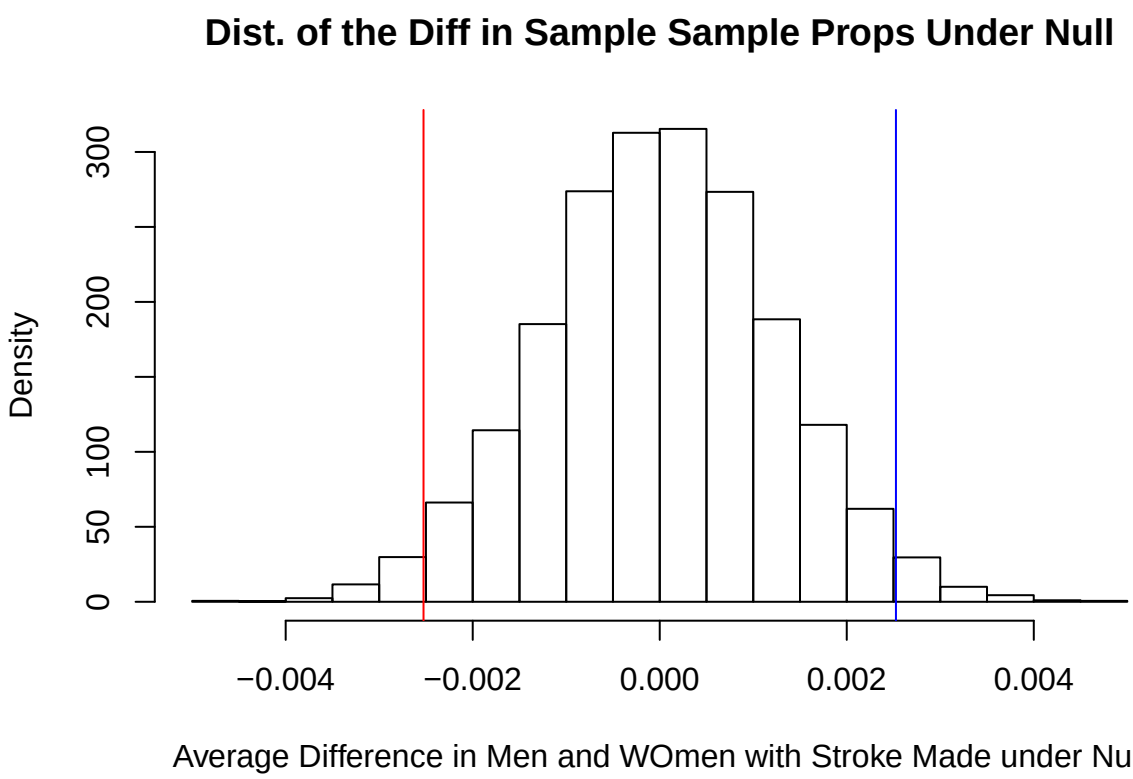
Randomization:

```r
# Make the data
df_combined <- data.frame("Stroke_data" = c(men, women),"gender" = rep(c("Male", "Female"), c(n_men, n_w
# Sanity checks
summary(df_combined$gender)
```

```
## Female   Male
##  24945  16986
```

```
# 16986
# 24945
```

```
num_sims <- 10000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
# idea here is if there is no relationshipm we should be able to shuffle the groups
shuffled_groups <- transform(df_combined, gender=sample(gender))
prop_men <- mean(shuffled_groups$Stroke_data[shuffled_groups$gender=="Male"])
prop_women <- mean(shuffled_groups$Stroke_data[shuffled_groups$gender=="Female"])
results_given_H0_true[i] <- prop_men - prop_women
}
# Finally plot the results
hist(results_given_H0_true, freq = FALSE, main='Dist. of the Diff in Sample Sample Props Under Null',
     xlab = 'Average Difference in Men and WOmen with Stroke Made under Null',
     ylab = 'Density')
diff_in_sample_props <- p_hat_men - p_hat_women
abline(v=diff_in_sample_props, col = "blue")
abline(v=-diff_in_sample_props, col = "red")
```




```
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
```

```
## Bootstrap p-value
cat("Bootstrap p-value: ",bootstrap_pvalue,"\n")
```

```
## Bootstrap p-value:  0.0413
```

```
## Normal Approx p-value
cat("Normal Approximation p-value: ", two_sided_diff_prop_pval,"\n")
```

```
## Normal Approximation p-value:  0.999737
```

```
cat("Bootstrap", "\n")
```

```
## Bootstrap
```

```
## Bootstrap
c(quantile(results, c(.025, .975)))
```

```
##         2.5%        97.5%
## 0.0001320469 0.0049873471
```

Chi-square Goodness of Fit test!

Question: Are the number of people who smokes different than those who formerly smoked or never smoked?
So, we have 3 sample statistic here: p_never_smoked, p_formerly_smoked, p_smokes

```
smoke_data <- subset(final_data, smoking_status == c('never smoked','smokes','formerly smoked'), select
```

```
n <- length(smoke_data)
table(smoke_data)
```

```
## smoke_data
##                 formerly smoked    never smoked          smokes
##               0            2430            5267            2126
```

```
prop.table(table(smoke_data))
```

```
## smoke_data
##                 formerly smoked    never smoked          smokes
##       0.0000000       0.2473786       0.5361906       0.2164308
```

```
n <- length(smoke_data$smoking_status)
# Expected count
E <- n * 0.3333
```

```
test_statistic <- sum(((table(smoke_data)-E)/E)^2)
test_statistic
```

```
## [1] 1.559961
```

```
pchisq(test_statistic, df = 3-1, lower.tail = FALSE)
```

```
## [1] 0.4584149
```

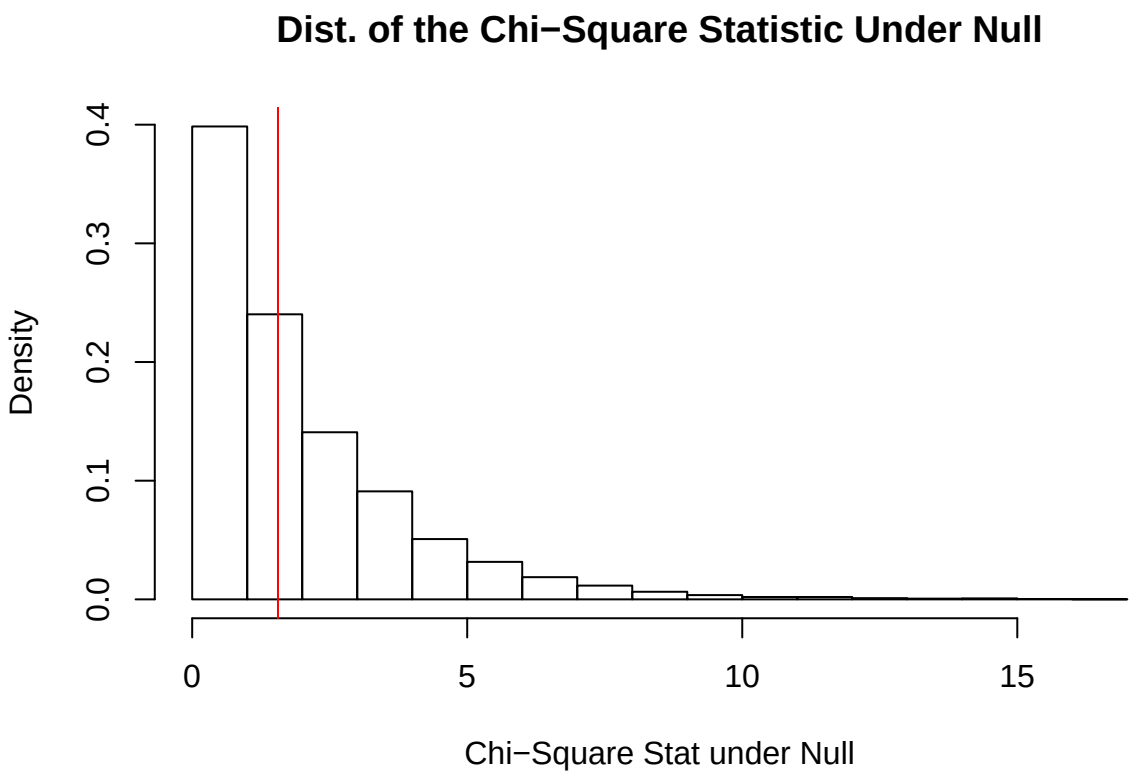
Randomization Approach

```
# Create our data under the assumption that H_0 is true
solutions_under_H_0 <- rep(c("formerly smoked", "never smoked", "smokes"), E)
# Sanity Check
table(solutions_under_H_0)
```

```
## solutions_under_H_0
## formerly smoked    never smoked          smokes
##            3274            3274            3274
```

```
num_sims <- 10000
# A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
new_samp <- sample(solutions_under_H_0, n, replace = T)
chisq_stats_under_H0[i] <- sum(((table(new_samp) - E)^2)/E)
}
```

```
hist(chisq_stats_under_H0, freq = FALSE, main='Dist. of the Chi-Square Statistic Under Null', xlab = 'Ch
abline(v=sum(((table(smoke_data)-E)/E)^2), col="red")
```



The randomization p-value

```
sum(chisq_stats_under_H0 >= sum(((table(smoke_data)-E)/E)^2))/num_sims
```

```
## [1] 0.4537
```