MSE719_Assignment1

Kurt Anand

2025-01-30

Q1. Throughout this assignment, we will work with a significance level of 0.02.

We will first import the Patients_Dataset.csv into R before we start investigating relationships.

```
patients_dataset <- read.csv("C:\\Users\\Local Admin\\Downloads\\Patients_Dataset.csv")
# Only the relevant columns of patients_dataset will be considered.
patients_dataset <- patients_dataset[c("Patient_ID", "Parental_History", "Gender",
                                       "Age_Category", "Status")]
```

This dataset has 2000 observations but the assignment states that there are 1000 observations. We will solve the questions as though there are 2000 observations. We will use head to see the first six rows.

```
print(head(patients_dataset, 6))
```

```
##   Patient_ID Parental_History Gender      Age_Category Status
## 1          1              Yes   Male Between 25 and 60    Yes
## 2          2              Yes Female          Above 60    Yes
## 3          3              Yes Female          Below 25     No
## 4          4              Yes   Male          Below 25     No
## 5          5               No Female Between 25 and 60     No
## 6          6               No   Male Between 25 and 60    Yes
```

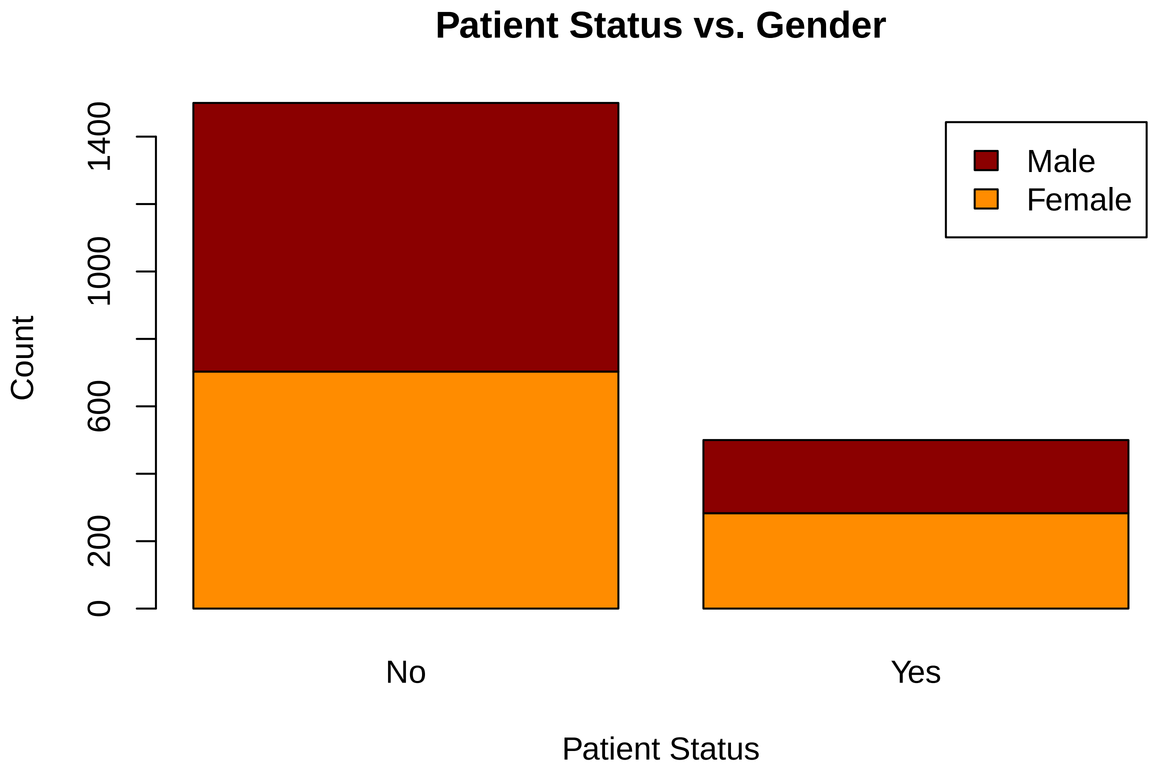
(1) We need to examine the association between patient status and gender. We will first create a contingency table by leveraging the table() function and using the two variables of interest as the arguments.

```
print(table(patients_dataset$Gender, patients_dataset$Status))
```

```
##
##           No Yes
##   Female 703 283
##   Male   797 217
```

We can visualize this contingency table by using the barplot() function.

```
barplot(table(patients_dataset$Gender, patients_dataset$Status),
        xlab = "Patient Status", ylab = "Count",
        main = "Patient Status vs. Gender",
        col = c("darkorange", "darkred"),
        legend.text = c("Female", "Male"))
```



To determine the association between patient status and gender, we ask ourselves the following question: do the proportion of answering “Yes” (or answering “No”) differ significantly between males and females? Based on the contingency table and the barplot, it appears as though there is a significant relationship between patient status and gender because females appear to have a higher proportion of “Yes” compared to males (in other words, males appear to have a higher proportion of “No” compared to females). To determine if this association is significant, we can conduct a Chi-squared distribution test.

The null hypothesis is that there is no significant association between patient status and gender, and the alternate hypothesis is that there is indeed a significant association between patient status and gender. Like mentioned earlier, we are using a significance level of 0.02. Since we used a significance level of 0.02 in the 1998 Data Analysis and 2000 Data Analysis tutorials, we will use that significance level in this assignment as well.

We will need to calculate the expected cell count for each cell.

```r
# Calculate row and column totals in the contingency table
sum_Female_Total <- 703 + 283 # total of female row
sum_Male_Total <- 797 + 217 # total of male row
sum_No_Total <- 703 + 797 # total of no column
sum_Yes_Total <- 283 + 217 # total of yes column

# Calculate grand total.
grand_total <- 703 + 797 + 283 + 217

# Calculate expected cell count.
e_Female_No <- (sum_Female_Total * sum_No_Total) / grand_total # Female X No
e_Male_No <- (sum_Male_Total * sum_No_Total) / grand_total # Male X No
```

```
e_Female_Yes <- (sum_Female_Total * sum_Yes_Total) / grand_total # Female X Yes
e_Male_Yes <- (sum_Male_Total * sum_Yes_Total) / grand_total # Male X Yes
```

We calculate the test statistic of the Chi-Squared Test of Independence.

```
# Calculate test statistic by using the formula.
test_statistic_status_gender <- (703 - e_Female_No) ** 2 / e_Female_No +
  (797 - e_Male_No) ** 2 / e_Male_No + (283 - e_Female_Yes) ** 2 / e_Female_Yes +
  (217 - e_Male_Yes) ** 2 / e_Male_Yes

cat("The test statistic is: ", test_statistic_status_gender)
```

```
## The test statistic is:  14.21345
```

Now that we have the test statistic, we will now calculate the p-value by using the pschisq function. If the p-value is strictly less than the significance level (in this case, 0.02), then we reject the null hypothesis and there is indeed a significant association between these two variables.

Note that there is only one degree of freedom because there are two rows and two columns are there would be $(2-1)*(2-1)=1$ degree of freedom.

```
cat("The p-value is: ", pchisq(test_statistic_status_gender, df = 1,
                               lower.tail = FALSE))
```

```
## The p-value is:  0.0001631996
```

Note: If we check the p-value using the chisq.test() function, we will get a slightly different result. The test-statistic would also be different.

```
chisq.test(table(patients_dataset$Gender, patients_dataset$Status))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(patients_dataset$Gender, patients_dataset$Status)
## X-squared = 13.827, df = 1, p-value = 0.0002005
```

This is because they use Yates' continuity correction (subtracting 0.5 from the absolute difference between the observed and expected cell counts and then squaring rather than squaring the absolute difference between the observed and expected cell counts). Thus, we will calculate the test statistic by using the formula from class for the remainder of this assignment.

As the p-value is less than the significance level of 0.02, we reject the null hypothesis and claim that there is a a significant relationship between patient status and gender. The observed pattern that the proportion of "Yes" among the females is higher than that of males (and the proportion of "No" among the males is higher than that of females) is indeed significant according to our Chi-squared distribution test.

We can use scientific facts to understand why Status vs. Gender has a strong association. According to a research conducted at Upper East Side Cardiology's Vein Institute, one of the reasons why females are more likely to develop varicose veins is that their blood volumes increase by around 45% during their first trimester of pregnancy. As the extra blood uses the existing blood vessel network for blood circulation during pregnancy, the extra volume that is unused puts additional pressure on the superficial blood vessels in the legs which can prevent valves from closing completely. Another contributor to this strong association is the extreme fluctuations in some hormones, particularly progesterone and estrogen. More information can be found on: https://www.bhusriheart.com/blog/why-women-are-more-likely-to-develop-varicose-veins#:~:text=Another%20issue%20that%20causes%20more,of%20your%20blood%20vessels%20strong.

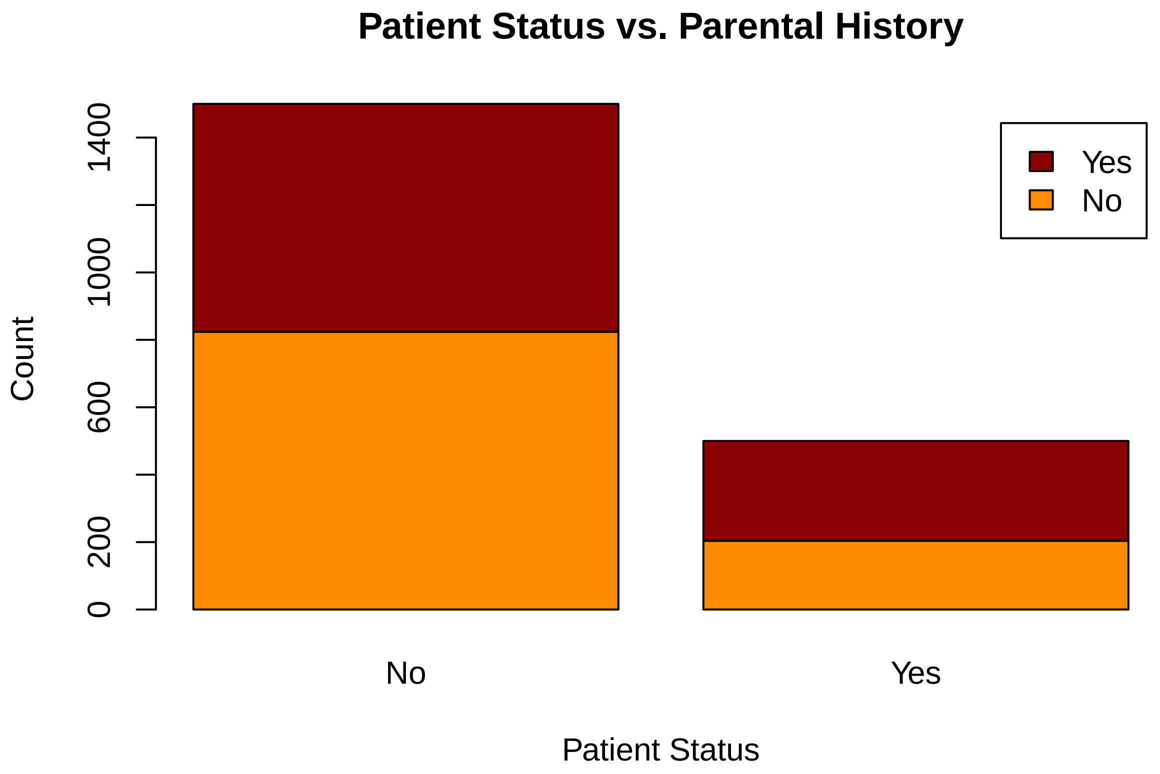
(2) We need to investigate the relationship between patient status and parental history. We will first create a contingency table by leveraging the table() function and using the two variables of interest as the arguments.

```
print(table(patients_dataset$Parental_History, patients_dataset$Status))
```

```
##
##        No Yes
##   No  824 204
##   Yes 676 296
```

We can visualize this contingency table by using the barplot() function.

```
barplot(table(patients_dataset$Parental_History, patients_dataset$Status),
        xlab = "Patient Status", ylab = "Count",
        main = "Patient Status vs. Parental History",
        col = c("darkorange", "darkred"),
        legend.text = c("No", "Yes"))
```



To determine the relationship between patient status and parental history, we ask ourselves the following question: do the proportion of having status as “Yes” (or “No”) differ significantly between having parental history as “Yes” and “No”? Based on the contingency table and the barplot, it appears as though there is a significant relationship between patient status and parental history because it appears as though patients with parental history as “Yes” appear to have a higher proportion of having patient status “Yes” than patients with parental history as “No” (and patients with parental history as “No” appear to have a higher proportion of having patient status “No” than patients with parental history as “Yes”). To determine if this relationship is significant, we can conduct a Chi-squared distribution test.

The null hypothesis is that there is no significant association between patient status and parental status, and the alternate hypothesis is that there is indeed a significant association between patient status and parental status. Like mentioned earlier, we are using a significance level of 0.02.

We will need to calculate the expected cell count for each cell.

```r
# Calculate row and column totals in the contingency table
sum_No_Row_Total <- 824 + 204 # total of row "No"
sum_Yes_Row_Total <- 676 + 296 # total of row "Yes"
sum_No_Column_Total <- 824 + 676 # total of column "No"
sum_Yes_Column_Total <- 204 + 296 # total of column "Yes"

# Calculate grand total.
grand_total <- 824 + 204 + 676 + 296

# Calculate expected cell count.
e_No_No <- (sum_No_Row_Total * sum_No_Column_Total) / grand_total # No X No
e_Yes_No <- (sum_Yes_Row_Total * sum_No_Column_Total) / grand_total # Yes X No
e_No_Yes <- (sum_No_Row_Total * sum_Yes_Column_Total) / grand_total # No X Yes
e_Yes_Yes <- (sum_Yes_Row_Total * sum_Yes_Column_Total) / grand_total # Yes X Yes
```

We calculate the test statistic of the Chi-Squared Test of Independence.

```r
# Calculate test statistic by using the formula.
test_statistic_status_parental_history <- (824 - e_No_No) ** 2 / e_No_No +
  (676 - e_Yes_No) ** 2 / e_Yes_No + (204 - e_No_Yes) ** 2 / e_No_Yes +
  (296 - e_Yes_Yes) ** 2 / e_Yes_Yes

cat("The test statistic is: ", test_statistic_status_parental_history)
```

```
## The test statistic is:  29.98618
```

Now that we have the test statistic, we will now calculate the p-value by using the pschisq function. If the p-value is strictly less than the significance level (in this case, 0.02), then we reject the null hypothesis and there is indeed a significant association between these two variables.

Note that there is only one degree of freedom because there are two rows and two columns are there would be $(2-1)*(2-1)=1$ degree of freedom.

```r
cat("The p-value is: ", pchisq(test_statistic_status_parental_history, df = 1,
                               lower.tail = FALSE))
```

```
## The p-value is:  4.351375e-08
```

As the p-value is less than the significance level of 0.02, we reject the null hypothesis and claim that there is a a significant relationship between patient status and parental history. The observed pattern that patients with parental history as “Yes” appear to have a higher proportion of having patient status “Yes” than patients with parental history as “No” (and patients with parental history as “No” appear to have a higher proportion of having patient status “No” than patients with parental history as “Yes”) is indeed significant according to our Chi-squared distribution test.

We can use scientific facts to understand why Status vs. Parental History has a strong association. Genetics might play a strong role here. According to Dr. Sumit Kapadia, research suggests that one parent suffering from varicose veins increases the likelihood of the child developing varicose veins to 40% and both parents suffering from varicose veins increases the likelihood of the child developing varicose veins by 90%. More information can be found on: https://www.drsumitkapadia.com/blog/role-of-genetics-in-varicose-veins/#:~:text=Varicose%20veins%2C%20often%20considered%20a,developing%20them%20rises%20to%2040%25..

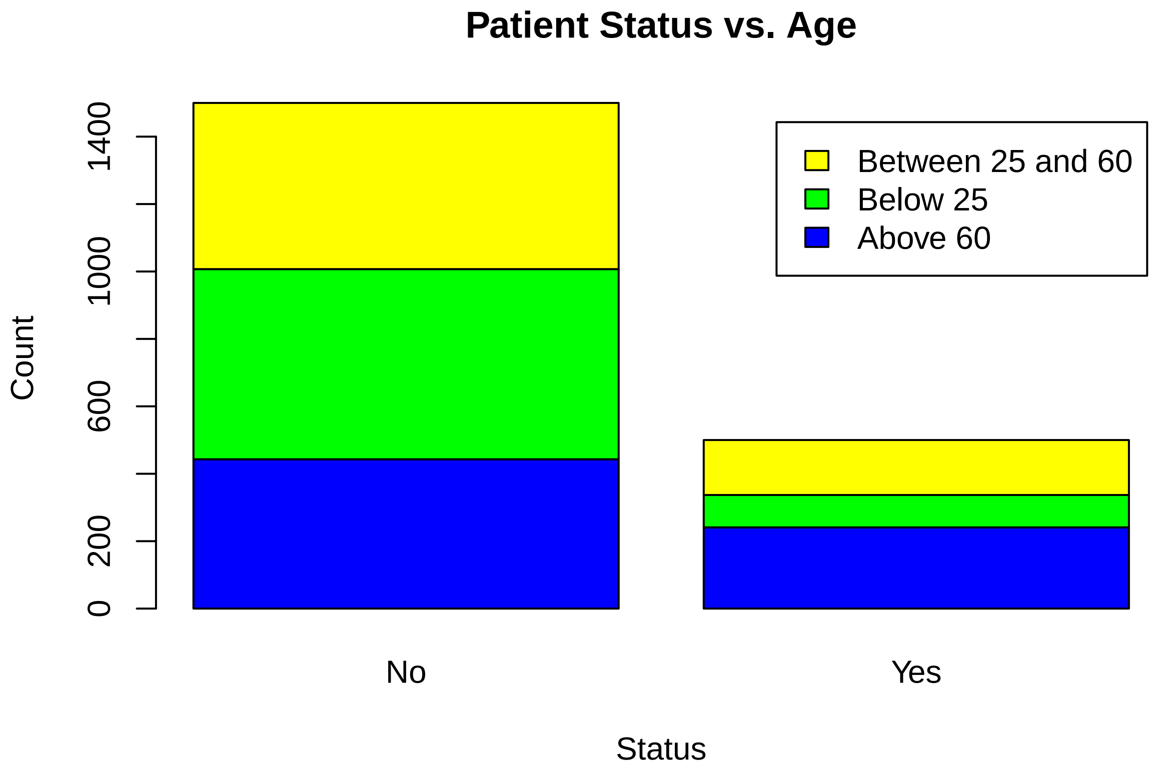
(3) We will need to assess the connection between patient status and age. We will first create a contingency table by leveraging the table() function and using the two variables of interest as the arguments.

```
print(table(patients_dataset$Age_Category, patients_dataset$Status))
```

```
## 
##                        No Yes
##   Above 60            443 241
##   Below 25            564  96
##   Between 25 and 60 493 163
```

We can visualize this contingency table by using the barplot() function.

```
barplot(table(patients_dataset$Age_Category, patients_dataset$Status),
        xlab = "Status", ylab = "Count", main = "Patient Status vs. Age",
        col = c("blue", "green", "yellow"),
        legend.text = c("Above 60", "Below 25", "Between 25 and 60"))
```



Based on the contingency table and the barplot, we see that there is a lower proportion of patients below 25 that answered “Yes” than patients who are between 25 and 60 (likewise, there is a lower proportion of patients between 25 and 60 that answered “No” than patients who are below 25). We also see that there is a lower proportion of patients between 25 and 60 that answered “Yes” than patients who are above 60 (likewise, there is a lower proportion of patients who are above 60 that answered “No” than patients who are between 25 and 60). Consequently, we observe that there is a direct relationship between age and likelihood of answering “Yes”. To determine if this relationship is significant, we will need to use a Chi-squared test.

The null hypothesis is that there is no significant relationship between patient status and age, and the alternative hypothesis is that there is a significant relationship between patient status and age. As mentioned earlier, we are using a significance level of 0.02.

We need to calculate the expected cell count for each cell.

```r
# Calculate row and column totals in the contingency table.
sum_a60 <- 443 + 241 # sum of row "Above 60"
sum_b25 <- 564 + 96 # sum of row "Below 25"
sum_25t60 <- 493 + 163 # sum of row "Between 25 and 60"
sum_No <- 443 + 564 + 493 # sum of column "No"
sum_Yes <- 241 + 96 + 163 # sum of column "Yes

# Calculate the grand total
grand_total <- 443 + 564 + 493 + 241 + 96 + 163

# Calculate the expected cell count.
e_a60_No <- (sum_a60 * sum_No) / grand_total # "Above 60" X "No"
e_b25_No <- (sum_b25 * sum_No) / grand_total # "Below 25" X "No"
```

```r
e_25t60_No <- (sum_25t60 * sum_No) / grand_total # "Between 25 and 60" X "No"
e_a60_Yes <- (sum_a60 * sum_Yes) / grand_total # "Above 60" X "Yes"
e_b25_Yes <- (sum_b25 * sum_Yes) / grand_total # "Below 25" X "Yes"
e_25t60_Yes <- (sum_25t60 * sum_Yes) / grand_total # "Between 25 and 60" X "Yes"
```

We calculate the test statistic of the Chi-Squared Test of Independence.

```r
test_statistic_status_age <- (443 - e_a60_No) ** 2 / e_a60_No +
  (564 - e_b25_No) ** 2 / e_b25_No + (493 - e_25t60_No) ** 2 / e_25t60_No +
  (241 - e_a60_Yes) ** 2 / e_a60_Yes + (96 - e_b25_Yes) ** 2 / e_b25_Yes +
  (163 - e_25t60_Yes) ** 2 / e_25t60_Yes

cat("The test statistic is: ", test_statistic_status_age)
```

```
## The test statistic is:  76.68749
```

Now that we have the test statistic, we will now calculate the p-value by using the pschisq function. If the p-value is strictly less than the significance level (in this case, 0.02), then we reject the null hypothesis and there is indeed a significant association between these two variables.

Note that there is only one degree of freedom because there are three rows and two columns are there would be $(3 - 1) * (2 - 1) = 2$ degrees of freedom.

```r
cat("The p-value is: ", pchisq(test_statistic_status_age, df = 2,
                               lower.tail = FALSE))
```

```
## The p-value is:  2.225995e-17
```

Since the p-value is less than significance level of 0.02, we reject the null hypothesis and claim that there is indeed a significant relationship between patient status and age. In other words, there is a direct relationship between these two variables. There is a higher proportion of patients between 25 and 60 that have answered “Yes” for status than patients below 25. There is a higher proportion of patients above 60 that have answered “Yes” for status than patients between 25 and 60. We can also say that there is a higher proportion of patients below 25 that have answered “No” for status than patients between 25 and 60, and there is a higher proportion of patients between 25 and 60 that have answered “No” than patients above 60. This observation is indeed significant according to our Chi-squared test.

We can use scientific facts to understand why Status vs. Age has a strong association. According to Mayo Clinic (2024), aging can cause deterioration in the valves that can assist with controlling the flow of blood. Eventually, this would cause the valves to let the blood flow back into the veins and accumulate there. More information can be found on: https://www.mayoclinic.org/diseases-conditions/varicose-veins/symptoms-causes/syc-20350643.

(4) Recall that the Chi-squared score is the Chi-squared test statistic. Recall the Chi-squared scores for each association (Status vs. Gender, Status vs. Parental History, and Status vs. Age).

```r
cat(" Chi-squared Score for Status vs. Gender: ", test_statistic_status_gender,
    "\n", "Chi-squared score for Status vs. Parental History: ",
    test_statistic_status_parental_history, "\n",
    "Chi-squared score for Status vs. Age: ", test_statistic_status_age)
```

```
##  Chi-squared Score for Status vs. Gender:  14.21345
##  Chi-squared score for Status vs. Parental History:  29.98618
##  Chi-squared score for Status vs. Age:  76.68749
```

If all three associations had the same number of degrees of freedom, then Status vs. Age would have had the strongest association as it had the highest Chi-squared value. However, the number of degrees of freedom for the Status vs. Age association is 2 while the other two have only 1 degree of freedom. One way to overcome this barrier is to calculate the Chi-squared score divided by the number of degrees of freedom for each association. Note that Status vs. Gender and Status vs. Parental History will have the Chi-squared score equal to the Chi-squared score divided by the number of degrees of freedom since both associations have only 1 degree of freedom. Calculate the Chi-squared score divided by the number of degrees of freedom or Status vs. Age.

```
cat("The Chi-squared score per degree of freedom for Status vs. Age: ",
    test_statistic_status_age / 2)
```

```
## The Chi-squared score per degree of freedom for Status vs. Age:  38.34374
```

The Chi-squared score per degree of freedom is still higher for Status vs. Age than for Status vs. Gender and Status vs. Parental History. Consequently, we have determined that age has the most significant impact on varicose vein status as opposed to parental history and gender.

We can also use p-values associated with each Chi-squared score to determine which association is the strongest (one with the lowest p-value).

```
cat("The p-value associated with the: ", "\n", "Chi-squared score for Status vs. Gender: ",
    pchisq(test_statistic_status_gender, df = 1, lower.tail = FALSE), "\n",
    "Chi-squared score for Status vs. Parental History: ",
    pchisq(test_statistic_status_parental_history, df = 1, lower.tail = FALSE),
    "\n", "Chi-squared score for Status vs. Age: ",
    pchisq(test_statistic_status_age, df = 1, lower.tail = FALSE))
```

```
## The p-value associated with the:
##  Chi-squared score for Status vs. Gender:  0.0001631996
##  Chi-squared score for Status vs. Parental History:  4.351375e-08
##  Chi-squared score for Status vs. Age:  2.002685e-18
```

The p-value associated with the Chi-squared score for Status vs. Age is the lowest out of the three associations, so the association between Status vs. Age is the strongest, and age has the most significant impact on varicose vein status.