R Assignment 7

You are going to investigate the association between nicotine gum use at the 1st annual visit (`AV1GUM`) and sex (`AGENDER`).

The following would be their research question: *Is there a relationship between nicotine gum use and sex?*

```
# read data
lhs<-read.csv(file =file.choose(), header = TRUE)
```

QUESTIONS

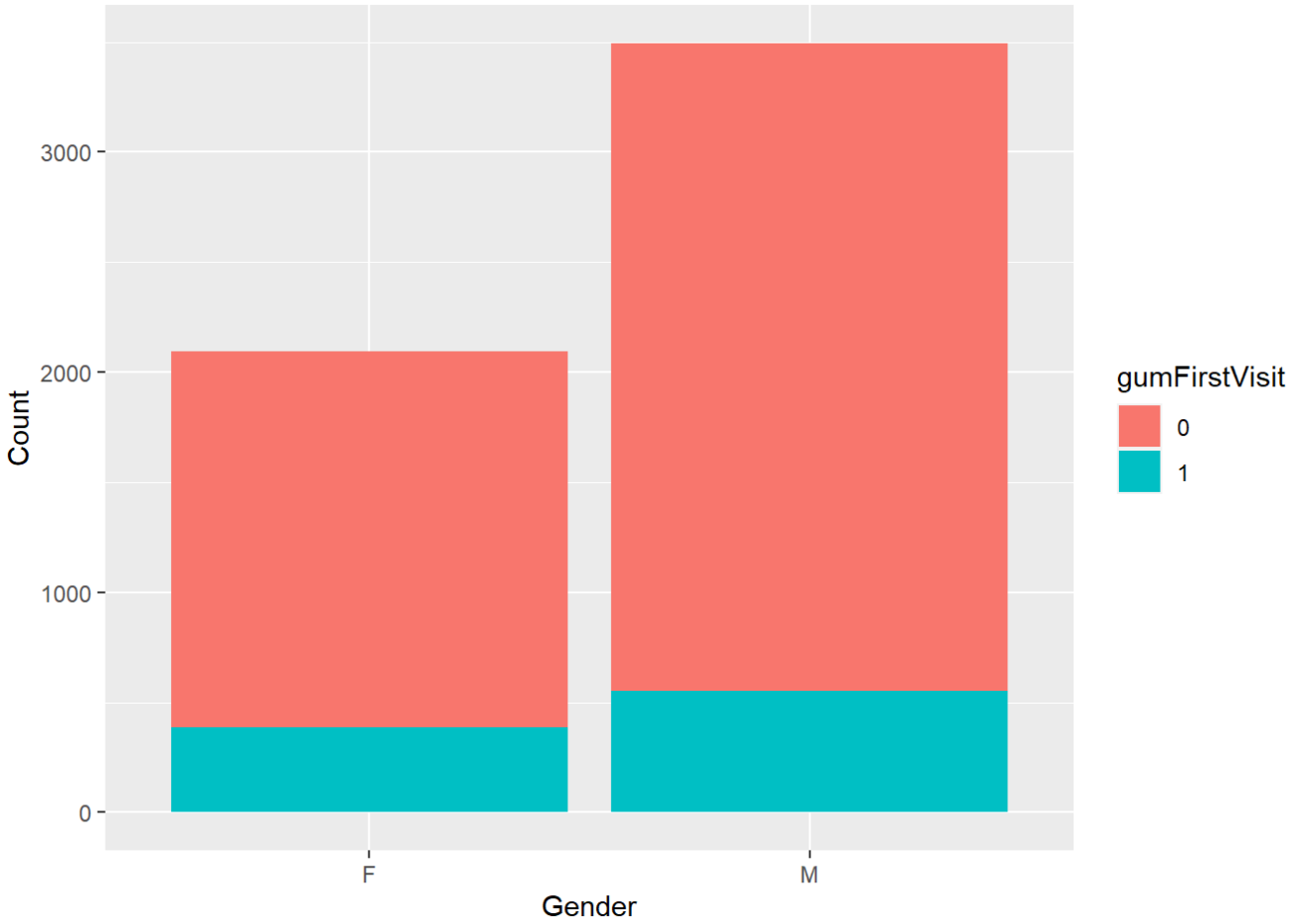
1. Provide a table AND a visualization for this data. What is the observed count for females who were using nicotine gum at the 1st annual visit?

```
tableSexGum = table(lhs$AGENDER,lhs$AV1GUM) # create table for sex and nicotine gum
tableSexGum    # view table
```

```
##
##        0    1
##   F 1708  387
##   M 2942  551
```

```
library(ggplot2)

df = as.data.frame(tableSexGum)
names(df)[1] = "Gender"
names(df)[2] = "gumFirstVisit"
names(df)[3] = "Count"
ggplot(data=df, aes(fill = gumFirstVisit, x = Gender, y = Count)) +
  geom_bar(position = "stack", stat = "identity") +
  ylab("Count")
```



As seen in the table `tableSexGum`, the observed number of females using nicotine gum at the 1st annual visit is 387.

2. Write what the null and alternative hypotheses are in the context of the question.

In the context of the above mentioned research question, the null hypothesis is, that there is no relationship between nicotine gum use and sex. The alternative hypothesis is, that there is a relationship between nicotine gum use and sex.

3. Choose a significance level and justify why you chose this significance level. What is the test statistic and the degrees of freedom from the Chi-squared test of independence? What is the resulting p-value from that test? State your conclusion in the context of this question. If an association was found,consider whether you can make a causal statement about the association and state your conclusions accordingly.

The chosen significance level for the following test is $\alpha = 0.05$. It leaves a small chance, that the null hypothesis gets rejected, even though it is true, but since there are no drastic consequences for wrongly rejecting the null hypothesis, it is justifiable to choose this significance level instead of a smaller one, as this also lessens the chance of wrongly failing to reject the null hypothesis.

```
chiSexGum = chisq.test(tableSexGum, correct = FALSE)
chiSexGum$statistic
```

```
## X-squared
##  6.825189
```

```
chiSexGum$parameter
```

```
## df
##  1
```

```
chiSexGum$p.value
```

```
## [1] 0.008988105
```

Since the conditions for the Chi-squared test are met, as explained in question 5, the `chisq.test` -function is used. The test statistic is 6.825, there is one degree of freedom. The calculated p-value is 0.009.

The probability of observing our χ^2 statistic or one more extreme if our null hypothesis that there is no relationship between nicotine gum use and sex is true, is below our significance level of 0.05. Thus, we have sufficient evidence to reject the null hypothesis and conclude that there is a relationship between use of nicotine gum and sex.

Since the data is from a random sample, but not from a random assignment as to who uses nicotine gum, we cannot make any causal statements about the association.

4. What is the expected count for females who were using nicotine gum at the 1st annual visit? Why are we interested in the expected counts (think about how this step relates to the null hypothesis and the process of testing theories)?

```
sum1 = colSums(tableSexGum)[2]
propG = sum1 / sum(tableSexGum)
sumF = rowSums(tableSexGum)["F"]
expF = sumF * propG
```

The expected count for females who were using nicotine gum at the 1st annual visit under the null hypothesis is 351.67. Since our null hypothesis states, that there is no relationship between sex and nicotine gum use, we are interested in the expected count, to compare it with the observed count.The expected count is needed in the process of testing the hypothesis, as that includes summing up the squares of all differences between observed and expected counts.

5. Does your data meet the conditions to use the Chi-square test? Explain why or why not. What is p-value from Fisher's exact test?

There are to conditions to use the chi-squared test. 1. The subjects must be independent, which is given

here, because we have a random sample and 2. the expected count in each cell must be greater or equal to 5, which is also true, as can be seen below.

```
sum0 = colSums(tableSexGum)[1]
propNG = sum0 / sum(tableSexGum)
sumM = rowSums(tableSexGum)["M"]

sumF * propG >= 5 &&
  sumF * propNG >= 5 &&
  sumM * propG >= 5 &&
  sumM * propNG >= 5
```

```
## [1] TRUE
```

```
fisherSexGum = fisher.test(tableSexGum)
fisherSexGum$p.value
```

```
## [1] 0.009643719
```

The p-value from fishers exact test is 0.096.

CONCEPTUAL QUESTIONS

> 6.What does the sampling distribution show us (the spread of our data or the spread of possible sample statistics)?

The sampling distribution shows the spread of possible sample statistics.

> 7. Can we observe the true sampling distribution? Why or why not?

We cannot observe the true sampling distribution because we only have one sample.

> 8. What sampling distribution are we interested in when we conduct a hypothesis test? Why is this?

We are interested in the sampling distribution under p_0 as we want to test, how likely it is to get the observed data, if the null hypothesis was true.

> 9. If the central limit theorem conditions met, are we saying that our data is normal or that the sampling distribution is normal?

If the CLT conditions are met, we are saying that the sampling distribution is normal.

10. Why do we check the CLT conditions and compute the standard error by plugging in $\hat{p}$ when constructing confidence intervals, but by plugging in p_0 (the null value) when doing hypothesis testing?

check the CLT conditions and compute the standard error by plugging in $\hat{p}$ when constructing confidence intervals, because we want to make statements about the present sample and state the confidence for the calculated point estimate. We use p_0 when doing hypothesis testing, because we want to test the present sample against p_0. $\hat{p}$ is known and does not need to be tested.