Hypothesis Testings

Angela Lee

12/4/2018

0.1 summary

To perform the following hypothesis testings to determine the relationship between the lung capacity and smoking. 1. Two Sample t-test 2. Wilconxon (Wilcox) Rank-Sum test 3. Chi-Square Independence test

Before we conduct any test, we assume that: i. let x1, x2,…xn = a random sample from population 1. ii. let x21, x22,…xm = a random sample from population 2. iii. Two populations are independent. iv. both X1 and X2 are normal.

Here is the structure for the data.

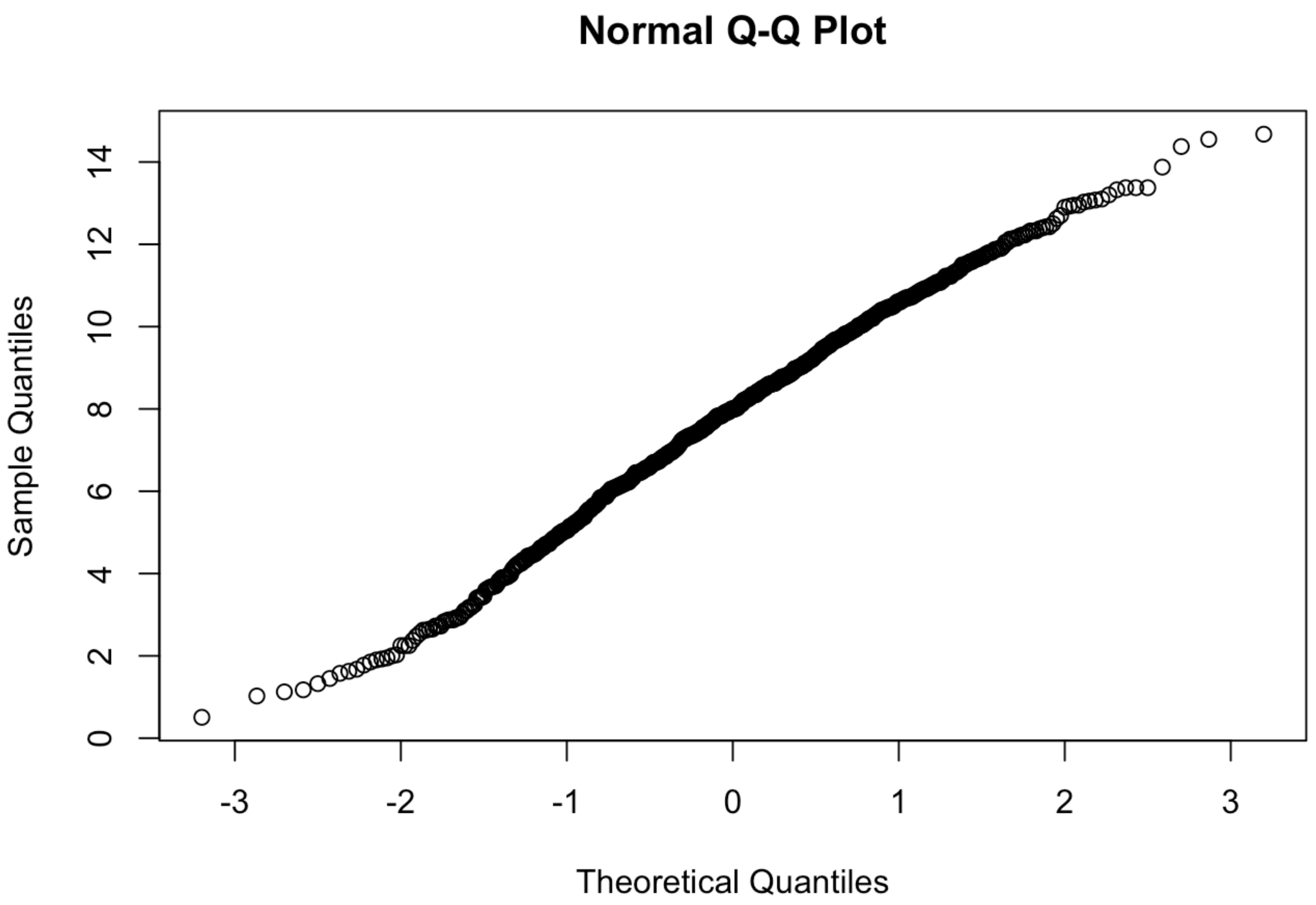
```
str(lungD)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    725 obs. of  6 variables:
##  $ LungCap  : num  6.47 10.12 9.55 11.12 4.8 ...
##  $ Age      : int  6 18 16 14 5 11 8 11 15 11 ...
##  $ Height   : num  62.1 74.7 69.7 71 56.9 58.7 63.3 70.4 70.5 59.2 ...
##  $ Smoke    : chr  "no" "yes" "no" "no" ...
##  $ Gender   : chr  "male" "female" "female" "male" ...
##  $ Caesarean: chr  "no" "no" "yes" "no" ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 6
##   .. ..$ LungCap  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ Age      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Height   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ Smoke    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Gender   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Caesarean: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

0.2 normality

We need to check the normality for continuous variables. For example, the lung capacity is a continuous variable. By looking at the following qqplot for the lung capacity, the points seem to fall about a straight line. The lung capacity variable is normal.

```
qqnorm(lungD$LungCap)
```



0.3 set up the Hypothesis

Here is the hypothesis for the difference in means, assuming that variances are unknown. The null hypothesis is that there is no difference in the mean lung capacity for smokers and the non-smokers.

$$H_0 : \mu 1 - \mu 2 = 0$$

The alternative hypothesis is that the mean lung capaicity of smokers is different than the non-smokers.

$$H_A : \mu 1 - \mu 2 \neq 0$$

0.4 boxplot

By examining the boxplot for the mean lung capacity smokers and non-smokers, we can see that the mean lung capaity for smokers is different than the non-smokers. However, we need to check the statistical significance for the mean difference by performing t-test and checking the p-value.

```
library(ggplot2)

ggplot(lungD, aes(x = Smoke, y = LungCap, fill = Smoke)) +
  geom_boxplot() +
  labs(title = "fig 1. Average Lung Capacity for Smokers and Non-Smokers", x = "Smoke
", y = "Lung Capacity")
```

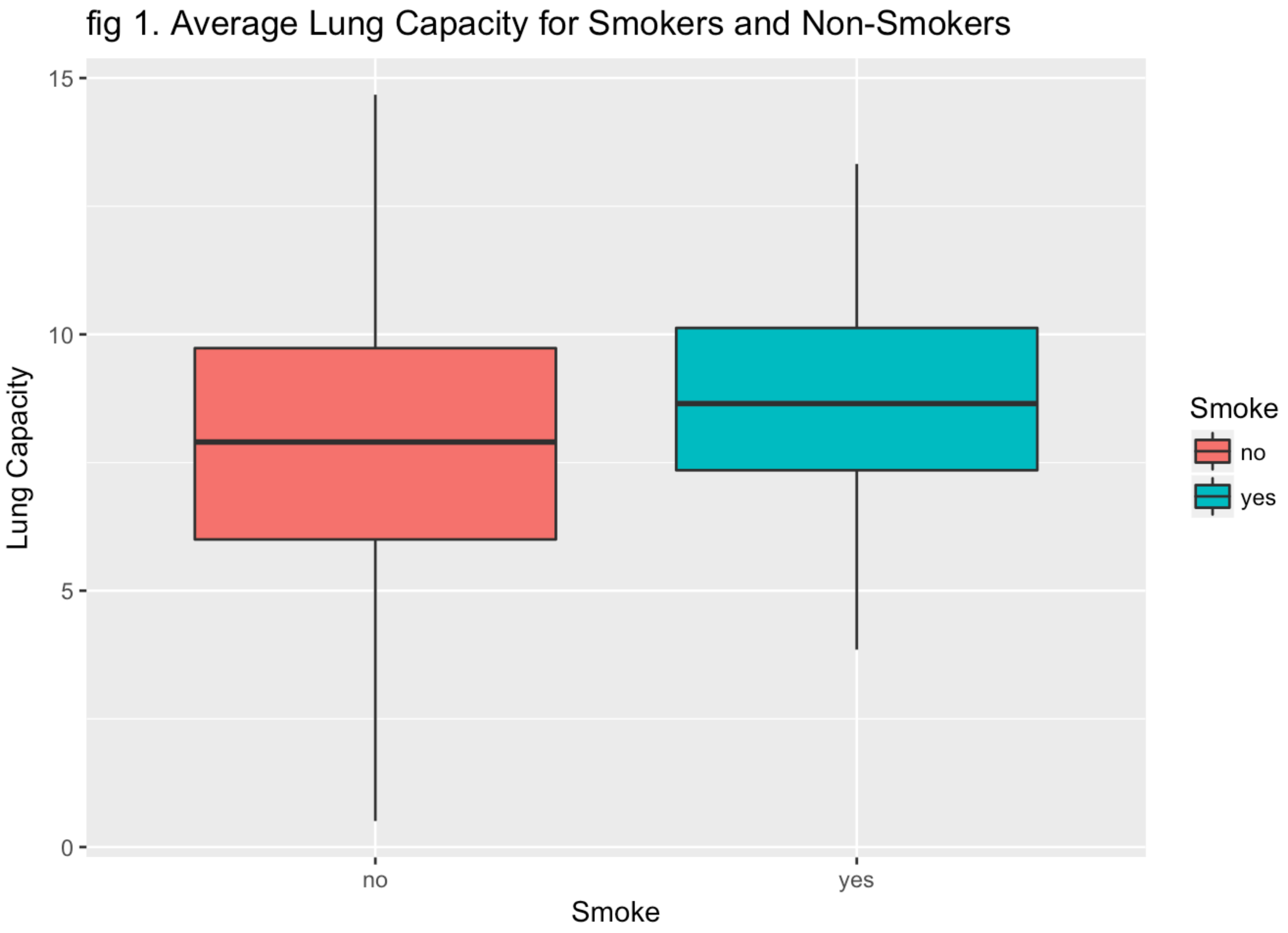


fig 1. Average Lung Capacity for Smokers and Non-Smokers

0.5 check variance

It is important to perform levene test to examine the statistical significance for the population variance before we perform the t-test. This is because we need to pass an argument for equal variance (var.eq = TRUE/FALSE) in the t.test. If the variances are the same (var.eq= TRUE), we get a tighter bound. If the variances are not the same, we get a loser bound. We can actually determine the sample variance for the lung capacity between the smokers and the non-smokers here. The variance of lung capaicty for the non-smokers is 7.43 and the variance of lung capacity for the smokers is 3.55.

```
var(lungD$LungCap[lungD$Smoke == 'no'])
```

```
## [1] 7.431694
```

```
var(lungD$LungCap[lungD$Smoke == 'yes'])
```

```
## [1] 3.545292
```

0.6 Levene test

Levene test is used to test the statistical significance for the population variance.

We create hypotheses and perform the levene test to see how statistically significant is the variance difference.

$$H_0 : population\ variance\ of\ the\ lung\ capacity\ for\ smokers\ and\ non - smokers\ are\ the\ same.$$

$$H_A : population\ variance\ of\ the\ lung\ capacity\ for\ smokers\ and\ non - smokers\ are\ different.$$

We reject the null hypothesis because the p-value in the following Levene test is smaller than 0.05 (p-value = 0.0003408). Result: at 5% statistical significance, we conclude that the population variance lung capacity between the two groups are different.

0.7 check p-value for statistically significance.

Since the boxplot indicates the average lung capacity for smokers and non-smokers are different, let’s assume the variance is also different. In the following t-test, we will set the var.eq = False. By looking at the p-value in the t-test, we can see the p-value (p-value = 0.0003927) is smaller than 0.05 (p> 0.05). We can reject the null hypothesis that there is no difference between mean lung capacity between smokers and non-smokers.

```
t.test(lungD$LungCap ~ lungD$Smoke, mu = 0 , alternative = 'two.sided', conf= 0.95, v
ar.eq = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  lungD$LungCap by lungD$Smoke
## t = -3.6498, df = 117.72, p-value = 0.0003927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3501778 -0.4003548
## sample estimates:
##  mean in group no mean in group yes 
##          7.770188          8.645455
```

```
#OR!
t.test(lungD$LungCap[lungD$Smoke == 'no'], lungD$LungCap[lungD$Smoke == 'yes'])
```

```
##
##  Welch Two Sample t-test
##
## data:  lungD$LungCap[lungD$Smoke == "no"] and lungD$LungCap[lungD$Smoke == "yes"]
## t = -3.6498, df = 117.72, p-value = 0.0003927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3501778 -0.4003548
## sample estimates:
## mean of x mean of y
##  7.770188  8.645455
```

```
#because the value is less than 0.05 we reject the null. reject that there is no diff
between mean lung capacity between smokers and nonsmokers.
```

0.8 Wilcox Test

If the median lung capacity of the smokers and non-smokers are not normal in the qq-plot, we use Wilcox Rank Sum Test. Here are the hypotheses for the median lung capapcity of smokers and non-smokers. Wilcoxon Rank Test is used to check the statistical significance for the population median.

$$H_0 : median\ lung\ capacity\ for\ smokers\ and\ non-smokers\ are\ the\ same.$$

$$H_A : median\ lung\ capacity\ for\ smokers\ and\ non-smokers\ are\ different.$$

By looking the following Wilcox test result, we reject the null hypothesis that the population median of the lung capacity for the smokers and non-smokers are the same because the p-value is smaller than 0.05 (p-value = 0.005538). Result: at 5% statistical significance, we conclude that the median lung capacity between the two groups are different.

0.9 contingency table

A contingency table is a type of table that displays the multi-variate frequency distribution of the variables. Below is the contingency table for gender and smoking.

```
##
##           no yes Sum
##   female 314  44 358
##   male   334  33 367
##   Sum    648  77 725
```

We can visualize the conditional and marginal distributions of the two variables in the population from the segmented barplot. The segmented portions in figure 2 represent the conditional distributions of smoking population affiliation with gender. Figure 3 shows the marginal distritions for the gender.If the gender and smoking were not associated, the conditional and marginal distributions would be identical. However, since data for the entire population is mostly unavailable, we must perform the inference methods to determine the association for two variables in the sample. One common procedure is the Chi Square Independence Test.

fig 2. Total Smokers Segmented by Gender

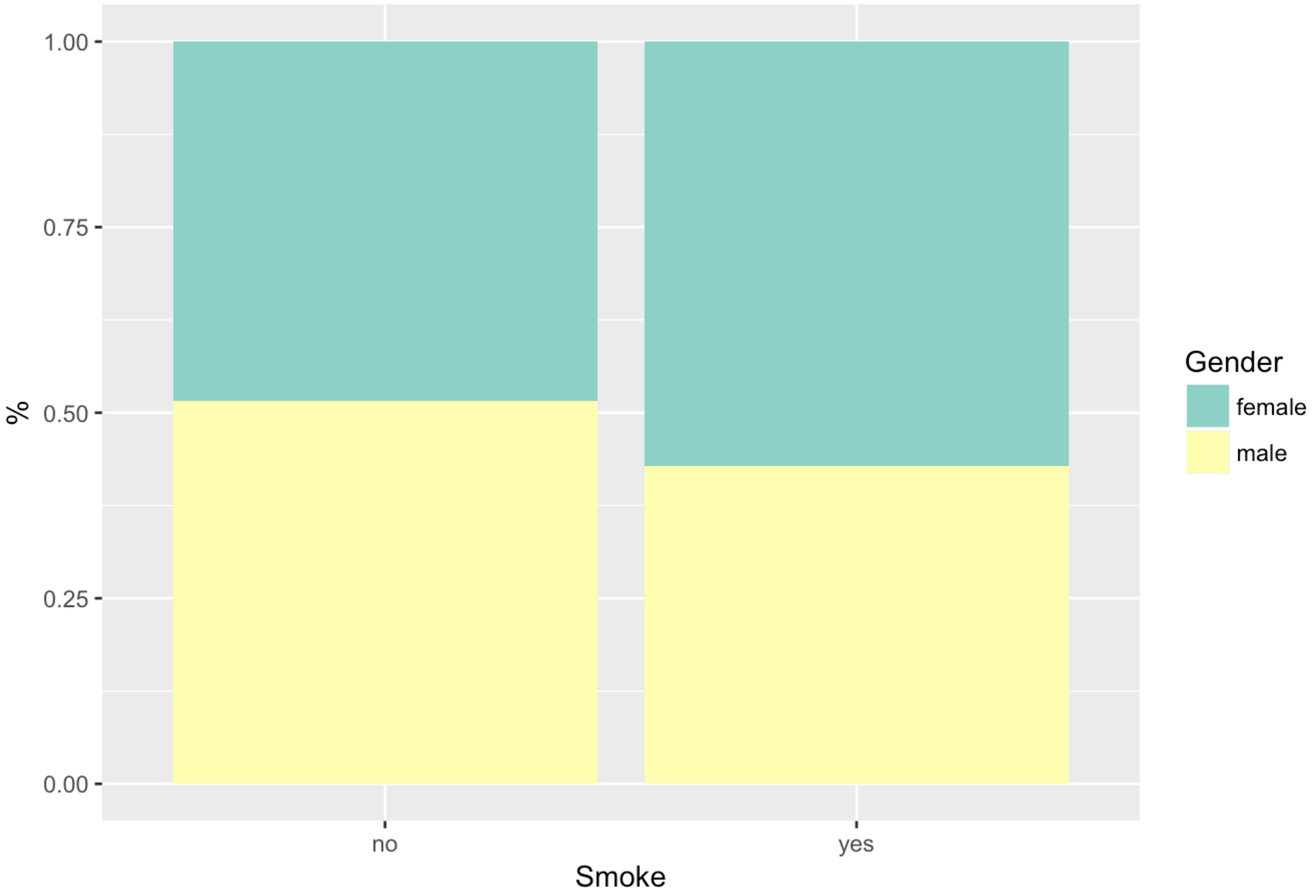
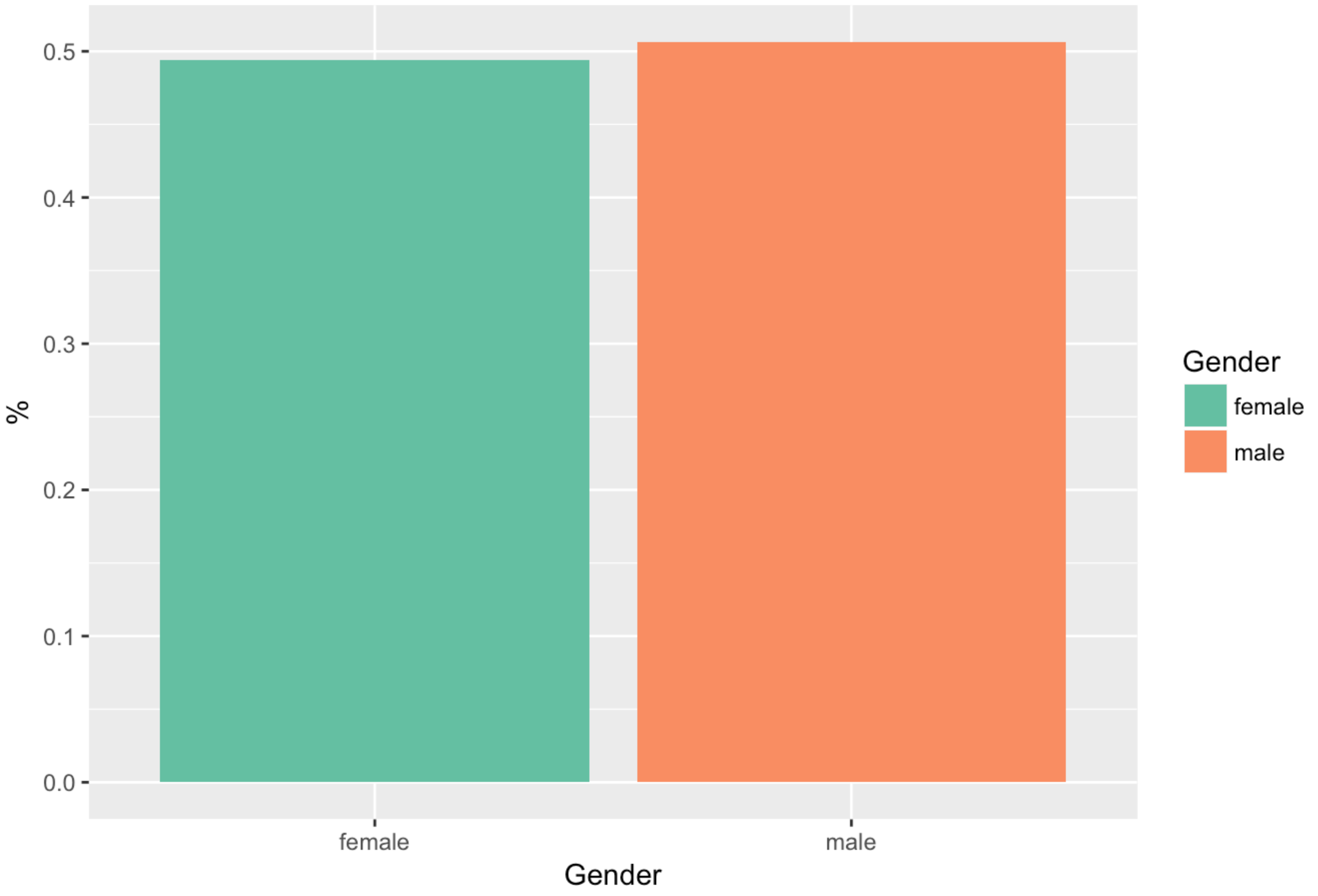


fig 3. Marginal Distributions for Gender



0.10 chi square test

Pearson’s chi square test is used to test the (categorical) variables and to evaluate that how likely any observed difference between the sets arose by chance.

Let’s say if we’d like to make an inference for the association between the gender and smoking, we use the chi square independence test to evaluate the statistical significance for the inference. Here are the hypotheses:

$$H_0 : gender\ and\ smoking\ are\ not\ associated.$$

$$H_A : gender\ and\ smoking\ are\ associated.$$

In order to determine the statistical signficance for the association between gender and smoking. We use chi square independence test and set the alpha to be 5%.

0.11 chi square Test for gender and smoking

Since the p-value is greater than 0.5 (p-value = 0.1866), we don’t reject the null hypothesis. We conclude that gender and smoking are not associated at alpha = 5% significance level.

```
t = chisq.test(tab1); t
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 2.0773, df = 4, p-value = 0.7215
```

0.12 expected frequencies

The idea behind the chi square independence test is to compare the observed frequencies with the frequencies we would expect if the null hypothesis of non-association is true. Here is the test statistic formula, where O = observed frequency. E = expected frequency.

$$x^2 = \sum(O - E)^2/E$$

here are the expected frequencies for gender and smoking.

```
t$expected
```

```
##
##                no      yes Sum
##   female 319.9779 38.02207 358
##   male   328.0221 38.97793 367
##   Sum    648.0000 77.00000 725
```