STAT 210
Applied Statistics and Data Analysis:
Problem List 5
(Due on week 6)

Exercise 1

The data set `Auto` in the `ISLR` package has information on nine variables for 392 vehicles. For this question, we will be only interested in two of them, `mpg` and `origin`. The first variable, `mpg`, corresponds to fuel consumption in miles per gallon for each vehicle, while `origin` is coded as 1 (American), 2 (European), and 3 (Japanese).

(a) Create a data frame named `q2.df` with the two columns corresponding to these variables. Do a boxplot of `mpg` as a function of `origin`. Comment on what you observe.

(b) Using the information in `mpg`, add a factor `fmpg` to `q2.df` created according to the following rule: if `mpg` is below 20, the value for the factor is `low`; if `mpg` is between 20 and 35, the value is `med`, and if `mpg` is above 35, the value is `high`. One way to do this is using the function `cut`. Also, change the labels in the `origin` factor to `Am`, `Eu`, and `Jap`.

(c) Produce a table of `origin` and `fmpg` and do a mosaic plot. The table should have `origin` as rows and `fmpg` as columns. Comment on what you observe. Produce a second table with proportions calculated relative to the different levels of `origin`. Again, comment on what you observe.

(d) We want to determine whether the fuel consumption categories that we created are homogeneously distributed for the different origins of the vehicles.

- Which test or tests do you know that can be used for this?
- What are the underlying assumptions?
- Are they satisfied in this case?
- Carry out all the tests you mentioned and discuss the results.
- What are your conclusions?

Solution

```
library(ISLR)
str(Auto)
```

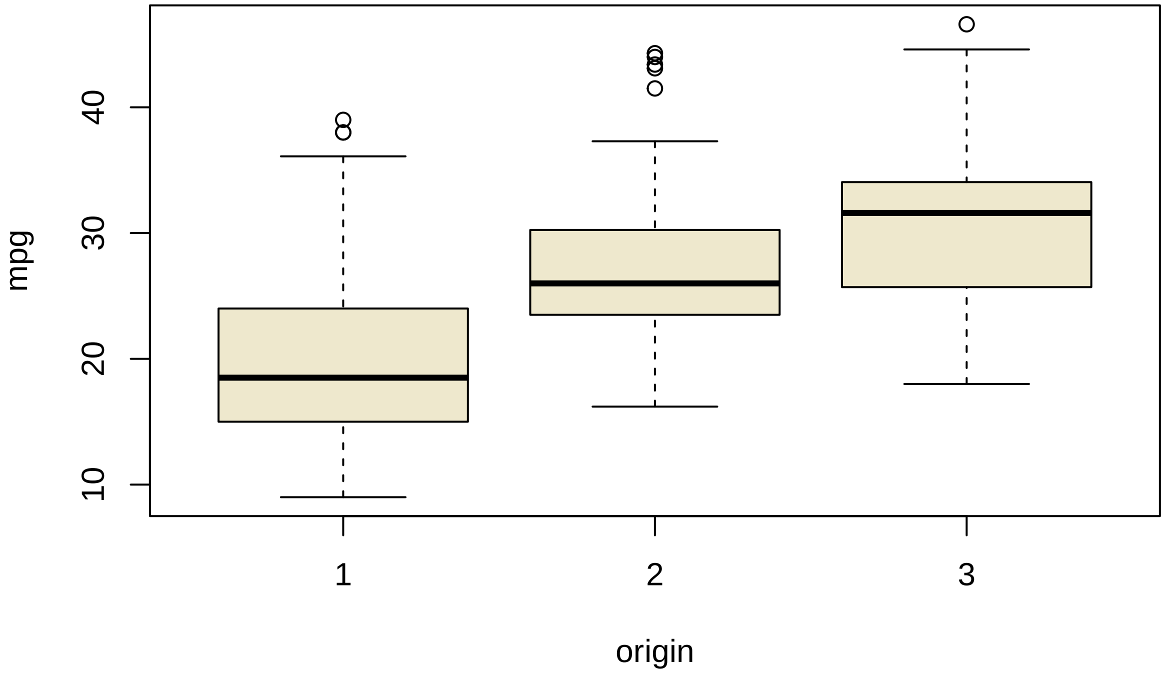
```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2
```

(a) We use `subset` to create this data frame and then we graph the boxplot

```
q2.df <- subset(Auto, select = c(mpg, origin))
str(q2.df)
```

```
## 'data.frame':    392 obs. of  2 variables:
##  $ mpg   : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ origin: num  1 1 1 1 1 1 1 1 1 1 ...
```

```
boxplot(mpg ~ origin, data = q2.df, col = 'cornsilk2')
```



We can see that American cars have a much smaller average `mpg` than Europeans or Japanese cars. European and Japanese cars are closer, but with higher values for the Japanese cars.

(b) Using the function `cut`, we can create the new factor and then we change the labels for `origin`.

```
q2.df$fmpg <- cut(q2.df$mpg,c(0,20,35,50),labels = c('low','med','high'),
                  right=F)
q2.df$origin <- factor(q2.df$origin, labels = c('Am','Eu','Jap'))
str(q2.df)
```

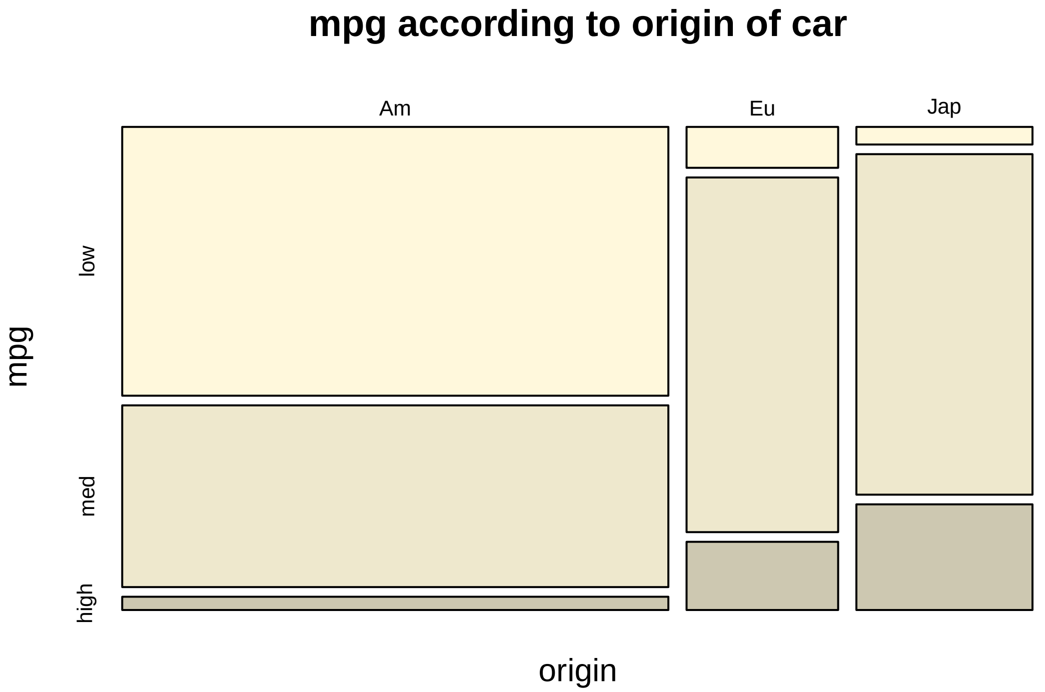
```
## 'data.frame':    392 obs. of  3 variables:
##  $ mpg   : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ origin: Factor w/ 3 levels "Am","Eu","Jap": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fmpg  : Factor w/ 3 levels "low","med","high": 1 1 1 1 1 1 1 1 1 1 ...
```

(c) Table and mosaic plot:

```
(q2.tbl <- with(q2.df, table(origin, fmpg)))
```

```
##       fmpg
## origin low med high
##    Am  142  96    7
##    Eu    6  52   10
##    Jap   3  58   18
```

```
mosaicplot(q2.tbl, ylab = 'mpg', color = c('cornsilk1','cornsilk2','cornsilk3'),
           main = 'mpg according to origin of car')
```



We see in the plot that the proportions of the different mpg categories for European and Japanese cars are similar, although Japanese cars are more frequent on the `high` category and less frequent on the `low`. The proportions for the American cars contrast sharply with the other two. Most cars (more the half) fall in the `low` category and there are very few in the `high` category. This points to American cars being different from the rest, and European and Japanese having similar distributions. The plot also shows that more than half of the cars in the sample come from America.

We now produce the second table using `prop.table`

```
prop.table(q2.tbl,1)
```

```
##       fmpg
## origin        low        med       high
##    Am  0.57959184 0.39183673 0.02857143
##    Eu  0.08823529 0.76470588 0.14705882
##    Jap 0.03797468 0.73417722 0.22784810
```

We see that nearly 60% of American cars are in the `low` category, in contrast with only around 2% in the `high` category. For Japanese and European cars the proportions are similar, with over 73% in the `med` category and more cars in `high` than in `low`.

(d) We have the Chi-square test and Fisher's exact test. Both compare observed and expected values for the contingency table. The first uses a Chi-square approximation for the sampling distribution of the test statistic and requires that the expected value for each cell in the table be at least 5. This is may not be true due to the small number of cars in certain categories.

To check this, we calculate the table of expected values.

```
colSums(prop.table(q2.tbl))%*%t(rowSums(prop.table(q2.tbl))) *392
```

```
##          Am        Eu       Jap
## [1,]  94.375 26.193878 30.431122
## [2,] 128.750 35.734694 41.515306
## [3,]  21.875  6.071429  7.053571
```

All values are above five, so the conditions for the test are satisfied. Another way of obtaining this table is shown below, after doing the chi-square test. The test is executed with the command

```
chisq.test(q2.tbl)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  q2.tbl
## X-squared = 116.25, df = 4, p-value < 2.2e-16
```

The test gives a p-value which is practically zero, so there is strong evidence to reject the null hypothesis of homogeneous distributions.

We can verify that the test is valid by extracting the table of expected values from the output of the test:

```
chisq.test(q2.tbl)$expected
```

```
##       fmpg
## origin      low       med      high
##    Am  94.37500 128.75000 21.875000
##    Eu  26.19388  35.73469  6.071429
##    Jap 30.43112  41.51531  7.053571
```

We now do Fisher's test:

```
fisher.test(q2.tbl)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  q2.tbl
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Again, the p-value is practically zero and we reach the same conclusion.

Exercise 2

The Southern State of the country of Nirvania, malaria is an endemic disease. Studies carried out show that 22% of the population in the state are infected. A new study in the city of Utopia, capital of the state, showed that among 1100 people tested, 198 were infected.

(a) The health authorities want to know if this result is in agreement with previous studies. Which would be your hypotheses in a statistical test? What tests do you know which apply in this situation? Explain why they are adequate and describe their underlying assumptions. Select a test or tests, apply them, and discuss in detail the results.

(b) Utopia is divided by a river into two regions, east and west Utopia. Of the sample used for the survey, 527 subjects were from east Utopia and of those, 89 were infected. Using this information, test whether there is a difference between east and west Utopia. Again, describe clearly the hypotheses you are testing, the reasons for choosing a particular test, the underlying assumptions, and discuss the results.

Solution

(a) If p is the proportion of infected people, we test

$$H_0 : p = 0.22 \quad \text{vs.} \quad p \neq 0.22$$

We considered two tests for this, the proportions test, which uses a normal approximation and the binomial test, which uses the binomial distribution. Since the sample is reasonably large and we are looking at proportions, the normal approximation seems reasonable. We carry out both

```r
prop.test(198,1100,0.22)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  198 out of 1100, null probability 0.22
## X-squared = 10.025, df = 1, p-value = 0.001545
## alternative hypothesis: true p is not equal to 0.22
## 95 percent confidence interval:
##  0.1579915 0.2042802
## sample estimates:
##    p
## 0.18
```

```r
binom.test(198,1100,0.22)
```

```
##
##  Exact binomial test
##
## data:  198 and 1100
## number of successes = 198, number of trials = 1100, p-value = 0.001188
## alternative hypothesis: true probability of success is not equal to 0.22
## 95 percent confidence interval:
##  0.1577217 0.2040042
## sample estimates:
## probability of success
##                   0.18
```

The p-value is smaller with the binomial test but both tests lead to the same conclusion: the result is not consistent with the value of 0.22 for the proportion of infected people.

(b) Let p_e and p_w denote the proportions of malaria infected people in east and west Utopia, respectively. We want to test

$$H_0 : p_2 = p_w \quad \text{vs.} \quad H_1 : p_e \neq p_w$$

We use the proportions test which uses a normal approximation that requires large sample size, which is the case.

```r
prop.test(c(89,198-89), c(527,1100-527))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(89, 198 - 89) out of c(527, 1100 - 527)
## X-squared = 0.70904, df = 1, p-value = 0.3998
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06850906  0.02581622
## sample estimates:
##    prop 1    prop 2
## 0.1688805 0.1902269
```

According to the test there is no evidence of a difference between east and west Utopia.

Exercise 3

For this question use the data set `Titanic`. Open the help and get familiar with the data included in the set. Observe that this is not a data frame but a table, a type of structure we had not met before. We will focus on two variables, `Class` and `Survived`.

(a) Build a contingency table of `Survived` and `Class`. `Survived` should correspond to the rows of your table. (One way to do this table is to use the function `apply`).

(b) Do a mosaic plot for the table in (a). Differentiate the classes by shades of grey or different colors. Comment on what you observe on this graph.

(c) Add a margin row and column to the table with the corresponding totals.

(d) Build a table with the proportions with respect to the total number of persons in the table. Comment on the results.

(e) Build a table with the proportions with respect to the total number of passengers in each class. Comment on the results.

(f) We want to test whether the distribution of surviving passengers in the different classes is the same. What test would you use for this and why? What conditions need to be satisfied? Discuss whether they are in this example. Carry out this test and comment on your results.

Solution

(a) We explore the data set with `str`. Then we create a table and store it in `Titanic.tbl`:

```
str(Titanic)
```

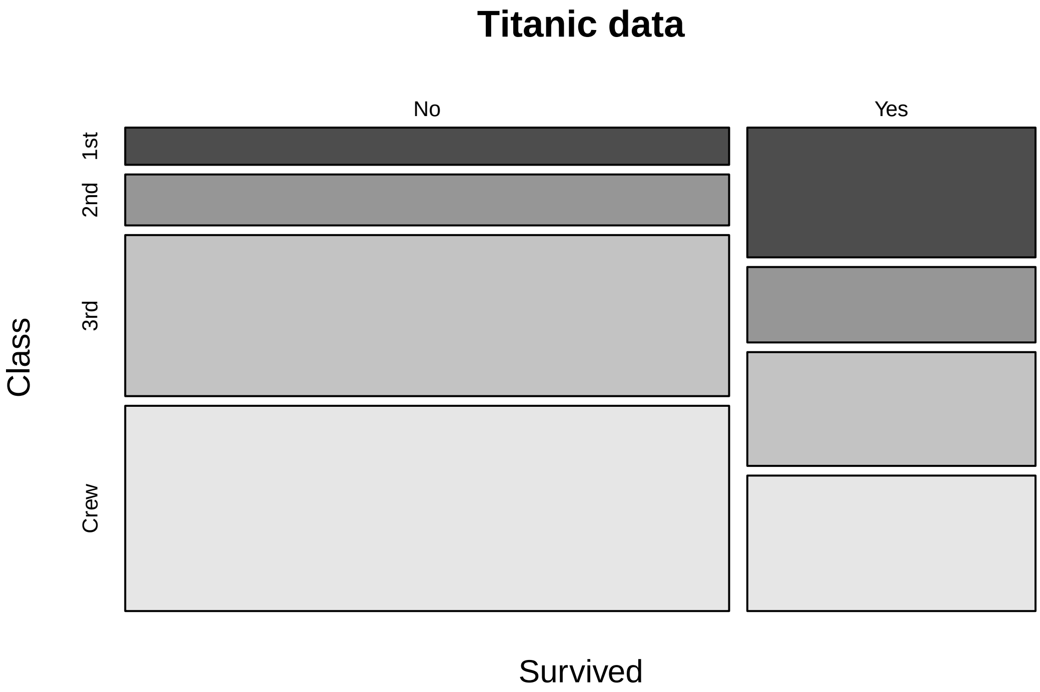
```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

```
(Titanic.tbl <- apply(Titanic,c(4,1),sum))
```

```
##         Class
## Survived 1st 2nd 3rd Crew
##      No  122 167 528  673
##      Yes 203 118 178  212
```

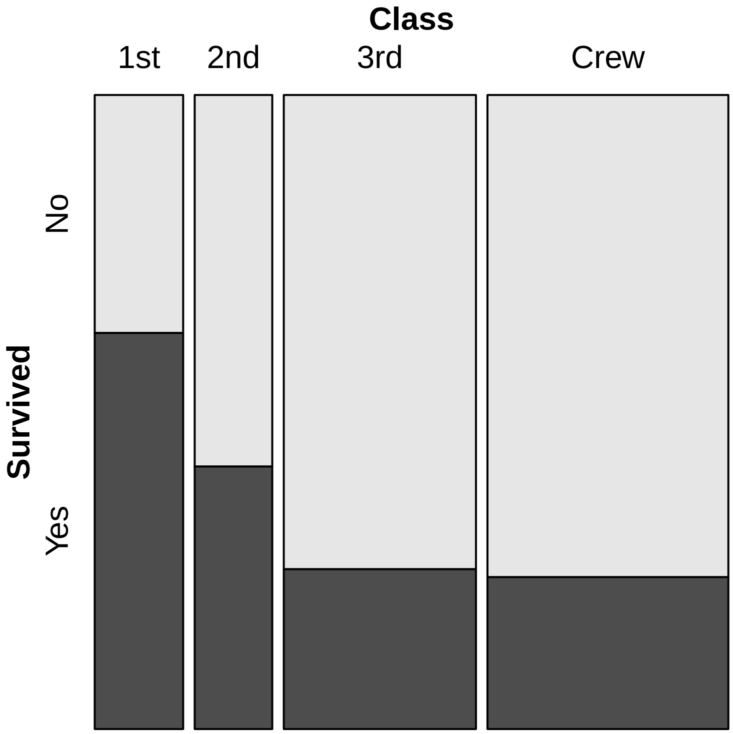
(b)

```
mosaicplot(Titanic.tbl, col = T, main = 'Titanic data')
```



Another way to do this is using the function `mosaic` in the package `vcd`.

```
library(vcd)
mosaic(Titanic.tbl, highlighting = T)
```



Observe, however, that the axes are reversed.

In the graphs we observe that the proportion of passengers that did not survive increases as class decreases, with the largest proportion for the crew.

(c) We can do this with the function addmargins:

```
addmargins(Titanic.tbl)
```

```
##          Class
## Survived 1st 2nd 3rd Crew  Sum
##      No  122 167 528  673 1490
##      Yes 203 118 178  212  711
##      Sum 325 285 706  885 2201
```

(d) We show two ways to do this. First, the total number of passengers is obtained from the previous table as 2201. Dividing the table by this number we get a table of proportions.

```
Titanic.tbl/2201
```

```
##          Class
## Survived        1st        2nd        3rd       Crew
##      No  0.05542935 0.07587460 0.23989096 0.30577010
##      Yes 0.09223080 0.05361199 0.08087233 0.09631985
```

Another way to produce this table is to use the function prop.table:

```
prop.table(Titanic.tbl)
```

```
##          Class
## Survived        1st        2nd        3rd       Crew
##      No  0.05542935 0.07587460 0.23989096 0.30577010
##      Yes 0.09223080 0.05361199 0.08087233 0.09631985
```

These tables show that the highest proportion of non-surviving passengers come from third class and the crew, but since this two classes also have a higher number of passengers, this does not say much. What we really need to make comparisons is a table showing the proportions of the two levels of Survived per class, which is the point of the next part of the question.

(e) Again, we give two procesures for producing this table. First, using sweep and colSums:

```
sweep(Titanic.tbl,2,colSums(Titanic.tbl),'/')
```

```
##          Class
## Survived       1st       2nd       3rd     Crew
##      No  0.3753846 0.5859649 0.7478754 0.760452
##      Yes 0.6246154 0.4140351 0.2521246 0.239548
```

Second, using prop.tables:

```
prop.table(Titanic.tbl,2)
```

```
##          Class
## Survived       1st       2nd       3rd     Crew
##      No  0.3753846 0.5859649 0.7478754 0.760452
##      Yes 0.6246154 0.4140351 0.2521246 0.239548
```

From these tables we see that about 2/3 of the first class passengers survived, but more than 50% of those in second class did not. For third class and crew, this percentage is around 75%.

(f) We want to test the homogeneity of the distribution of the variable Survived for the different classes, which compares the proportion of surviving passengers for the four levels of Class. The test for this is the Chi-square test that compares the observed values with the expected values under the assumption that all classes have the same proportion of surviving passengers. The test requires that all entries in the expected value matrix be greater than or equal to five. We verify this with the following command

```
chisq.test(Titanic.tbl)$expected
```

```
##         Class
## Survived      1st       2nd      3rd    Crew
##      No  220.0136 192.93503 477.9373 599.114
##      Yes 104.9864  92.06497 228.0627 285.886
```

We see that all entries are above the required level. We tun the test with the following command

```
chisq.test(Titanic.tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Titanic.tbl
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

The p-value is practically zero, so we have strong evidence to reject the null hypothesis that the proportion of surviving passengers is the same for all classes.

Exercise 4

In a certain country, the proportion of adults between 18 and 50 years that smoke 'frequently' is 22.5%. The authorities carry out a pilot one-year anti-tobacco campaign in a city and want to evaluate its effect on the proportion of 'frequent' smokers, to decide whether to extend the campaign country-wide.

(i) In a randomly chosen sample of 120 persons in the 18-50 years range, 17 are found to be 'frequent' smokers. Is there evidence of a decrease in the proportion of 'frequent' smokers in the city? Which would be your hypotheses in a statistical test? What tests do you know which apply in this situation? Explain why they are adequate and describe their underlying assumptions. Select a test or tests, apply them and discuss in detail the results.

(ii) In the survey, 65 persons were females and 9 of them were 'frequent' smokers. Look at the proportions for frequent smokers among the male and female populations and test whether there is evidence of a difference between the two. Again, describe clearly the hypotheses you are testing, the reasons for choosing a particular test, the underlying assumptions, and discuss the results.

Solution

(i) We test

$$H_0 : p = 0.225 \qquad vs \qquad H_1 : p < 0.225$$

because we are testing for a decrease in the proportion of frequent smokers. We reviewed two tests, the proportions test, which uses a normal approximation and the binomial test, which uses the binomial distribution. Since the sample is reasonably large and we are looking at proportions, the normal approximation seems reasonable. We carry out both.

```
prop.test(17,120,0.225,'less')
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  17 out of 120, null probability 0.225
## X-squared = 4.313, df = 1, p-value = 0.01891
## alternative hypothesis: true p is less than 0.225
## 95 percent confidence interval:
##  0.0000000 0.2066287
## sample estimates:
```

```
##         p
## 0.1416667
```

```r
binom.test(17,120,0.225,'less')
```

```
##
##  Exact binomial test
##
## data:  17 and 120
## number of successes = 17, number of trials = 120, p-value = 0.01542
## alternative hypothesis: true probability of success is less than 0.225
## 95 percent confidence interval:
##  0.0000000 0.2048901
## sample estimates:
## probability of success
##              0.1416667
```

We see that the p-value is smaller with the binomial test, but both are smaller than 0.05 and bigger than 0.01. This means that, at the usual levels, the decision would be the same: At the 5% level we conclude that there has been a decrease in the frequent smoker population but at the 0.01 level we would not.

(ii) Let p_f and p_m denote the proportions of 'frequent' smokers in the female and male populations, respectively. We want to test

$$H_0 : p_f = p_m \qquad vs \qquad H_1 : p_f \neq p_m$$

The proportions test (`prop.test`) can be used in this situation. The normal approximation underlying the test is justified by the size of the sample.

The number of males and frequent male smokers in the sample are

$$120 - 65 = 55, \qquad 17 - 9 = 8$$

```r
prop.test(c(8,9),c(55,65))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(8, 9) out of c(55, 65)
## X-squared = 7.4809e-31, df = 1, p-value = 1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1254252  0.1394112
## sample estimates:
##    prop 1    prop 2
## 0.1454545 0.1384615
```

With this p-value we would not reject the null hypothesis of equal distributions for females and males at the usual levels.