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

For all tests in this list use a significance level of $\alpha = 0.02$.

Exercise 1

The data set `PL6-25_q1.csv` has information on 11 socioeconomic variables obtained in a survey of 534 subjects in 1985.

(a) Check whether there are any values missing in the data set. Create a new data frame called `df1` that has no missing values and includes only the variables `gender`, `occupation`, and `married`.

Solution:

Read the data and check whether there are missing data

```r
data1 <- read.csv('PL6-25_q1.csv')
str(data1)
```

```
## 'data.frame':    534 obs. of  12 variables:
##  $ X         : int  1 1100 2 3 4 5 6 7 8 9 ...
##  $ wage      : num  5.1 4.95 6.67 4 7.5 ...
##  $ education : int  8 9 12 12 12 13 10 12 16 12 ...
##  $ experience: int  21 42 1 4 17 9 27 9 11 9 ...
##  $ age       : int  35 57 19 22 35 28 43 27 33 27 ...
##  $ ethnicity : chr  "hispanic" "cauc" "cauc" "cauc" ...
##  $ region    : chr  "other" "other" "other" "other" ...
##  $ gender    : chr  "female" "female" "male" "male" ...
##  $ occupation: chr  "worker" "worker" "worker" "worker" ...
##  $ sector    : chr  "manufacturing" "manufacturing" "manufacturing" "other" ...
##  $ union     : chr  "no" "no" "no" "no" ...
##  $ married   : chr  "yes" "yes" "no" "no" ...
```

```r
sum(is.na(data1))
```

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

There are no missing data. We create the new data frame

```r
df1 <- subset(data1, select = c(gender, occupation, married))
str(df1)
```

```
## 'data.frame':    534 obs. of  3 variables:
##  $ gender    : chr  "female" "female" "male" "male" ...
##  $ occupation: chr  "worker" "worker" "worker" "worker" ...
##  $ married   : chr  "yes" "yes" "no" "no" ...
```

```r
attach(df1)
```

and attach it to facilitate our work.

(b) Build a contingency table for `occupation` and `gender`. The rows of the table should correspond to `gender`. Add the totals by row and column. Represent this table as a mosaic plot and discuss the result.

Solution:

We create the table and store it in `tab1`:

```r
(tab1 <- table(gender, occupation))
```

```
##         occupation
## gender   management office sales services technical worker
##   female         21     76    17       49        52     30
##   male           34     21    21       34        53    126
```

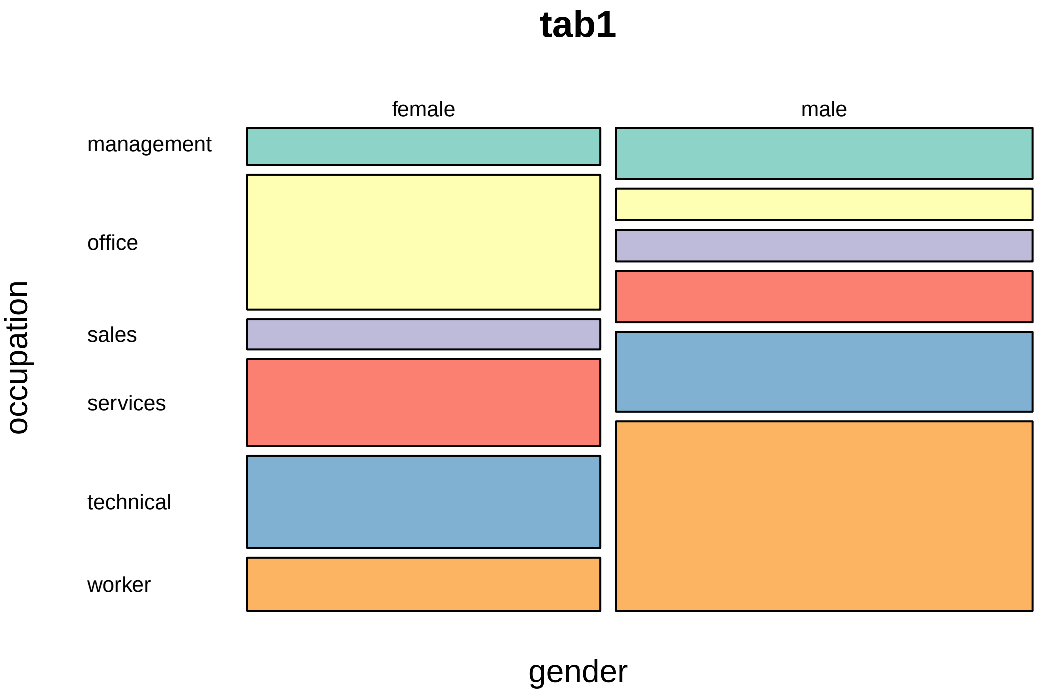
Add the totals:

```r
addmargins(tab1)
```

```
##         occupation
## gender   management office sales services technical worker Sum
##   female         21     76    17       49        52     30 245
##   male           34     21    21       34        53    126 289
##   Sum            55     97    38       83       105    156 534
```

Do the mosaic plot. We use `RColorBrewer` to generate the color pallete.

```r
library(RColorBrewer)
mosaicplot(tab1, col=brewer.pal(6,"Set3"), las = 1)
```



We see that the proportions are very different. For instance, for the category `worker`, placed at the bottom of the plot, the proportion of male workers is three or four times bigger than for women, but for `office`, the opposite situation is true. It is very llikely that the distribution of `occupation` depends on the gender.

(c) Using the table you built in (b), create a table for the same variables representing the proportions by `occupation`. Discuss the results.

Solution:

We round to two decimals the values in the table

```
round(prop.table(tab1, 2),2)
```

```
##         occupation
## gender   management office sales services technical worker
##   female       0.38   0.78  0.45     0.59      0.50   0.19
##   male         0.62   0.22  0.55     0.41      0.50   0.81
```

We confirm the observation we made about the mosaic plot: the proportions of the two genders are very different for the categories of `occupation`. For `worker`, males are approximate four times more frequent than females, but for `office` the reverse is valid. For `management` and `sales` males have a higher proportion, for `services` females are more frequent, and for `technical` they are evenly split.

(d) You have to test whether the distribution of occupations across the two genders 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:

The adequate test in this case is the Chi-square test, which compares the expected frequencies assuming equal distributions for the genders with the observed values. The test requires that the expected values for all entries in the table be greater than or equal to five. We will verify this after running the test.

```
(ch1 <- chisq.test(tab1))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 93.486, df = 5, p-value < 2.2e-16
```

The *p*-value is practically zero, indicating strong evidence against the null hypothesis of equal distributions. We check the condition for the expected values by retrieving the matrix of expected values from the output of the test, which is stored in `ch1`:

```
round(ch1$expected,2)
```

```
##         occupation
## gender   management office sales services technical worker
##   female      25.23   44.5 17.43    38.08     48.17  71.57
##   male        29.77   52.5 20.57    44.92     56.83  84.43
```

We see that all entries are bigger than 5, and the assumption for the test is valid.

(e) Build a contingency table for `occupation` and `married`. The rows of the table should correspond to `married`. Add the totals by row and column. Represent this table as a mosaic plot and discuss the result.

Solution:

We create the table and store it in `tab2`:

```
(tab2 <- table(married, occupation))
```

```
##        occupation
## married management office sales services technical worker
##     no          17     38     9       35        31     54
##     yes         38     59    29       48        74    102
```

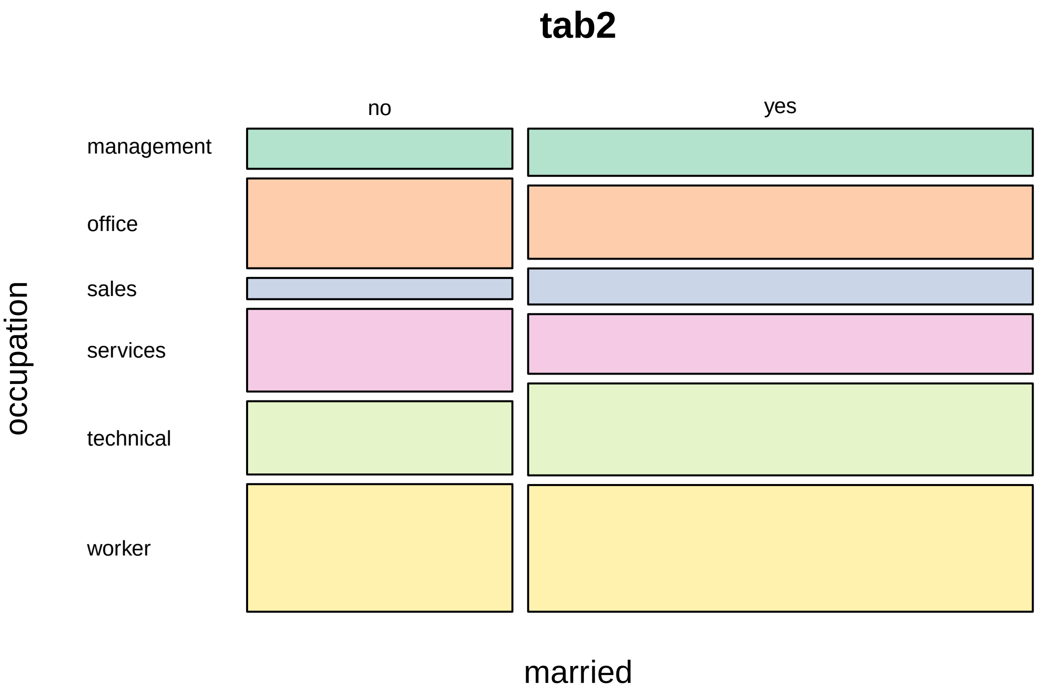
Add the totals:

```
addmargins(tab2)
```

```
##        occupation
## married management office sales services technical worker Sum
##     no          17     38     9       35        31     54 184
##     yes         38     59    29       48        74    102 350
##     Sum         55     97    38       83       105    156 534
```

Do the mosaic plot. we use again `RColorB rewer` with a different pallete.

```
mosaicplot(tab2, col=brewer.pal(6,"Pastel2"), las = 1)
```



We see now that the proportions are more similar than in the previous plot. For instance, for the category `worker`, placed at the bottom of the plot, the proportions of `no` and `yes` are practically the same. The largest differences seem to be for `service` and `sales`.

(f) Using the table you built in (e), create a table for the same variables representing the proportions by `occupation`. Discuss the results.

Solution:

We round to two decimals the values in the table

```
round(prop.table(tab2, 2),2)
```

```
##        occupation
## married management office sales services technical worker
##     no          0.31   0.39  0.24      0.42      0.30   0.35
##     yes         0.69   0.61  0.76      0.58      0.70   0.65
```

We confirm the observation we made about the mosaic plot: the proportions of the two genders are similar for the different categories of `occupation`. Married employees are always more frequent than non-married, and the proportion varies between 58% for `services` to 76% for `sales`.

(g) You have to test whether the distribution of occupations across the two values for married 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:

The adequate test is again the Chi-square test, which compares the expected frequencies assuming equal distributions for `married` with the observed values. The test requires that the expected values for all entries in the table be greater than or equal to five. We will verify this after running the test.

```r
(ch2 <- chisq.test(tab2))
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tab2
## X-squared = 6.5342, df = 5, p-value = 0.2576
```

The p-value now is large, and we cannot reject the null hypothesis of equal distributions. We check the condition for the expected values by retrieving the matrix of expected values from the output of the test, which is stored in `ch2`:

```r
round(ch2$expected,2)
```

```
##         occupation
## married management office sales services technical worker
##     no       18.95  33.42 13.09     28.6     36.18  53.75
##     yes      36.05  63.58 24.91     54.4     68.82 102.25
```

We see that all entries are bigger than 5, and the assumption for the test is valid.

```r
detach(df1)
```

Exercise 2

For all tests in this question, state explicitly the hypotheses, describe the assumptions behind the test, and explain why they are justified.

(a) Health authorities have determined that the proportion of people who smoke tobacco in the country of Arcadina is 23%. The capital city Arcadia carries out a six-month campaign against smoking, and a survey at the end of the campaign shows that out of 1200 persons interviewed, 242 smoke. Is there evidence that the campaign was effective? How would you test whether this? State clearly the hypothesis you are testing and describe the assumptions for the test or tests you propose. Why do you think that the assumptions are satisfied in this case? Carry out the test or tests and comment on the result(s).

Solution:

If p denotes the proportion of smokers in the capital, we are testing

$$H_0 : p = 0.23 \qquad vs. \qquad H_A : p < 0.23$$

We can either use the proportions test, which uses a normal approximation, or the exact binomial test. The first test assumes that the sample is large enough for the approximation to be valid. In this case, it is reasonable to assume that the approximation is valid, but since we can also carry out the exact test, we can compare both results.

```
prop.test(242, 1200, 0.23, alternative = 'less')
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  242 out of 1200, null probability 0.23
## X-squared = 5.2807, df = 1, p-value = 0.01078
## alternative hypothesis: true p is less than 0.23
## 95 percent confidence interval:
##  0.0000000 0.2218108
## sample estimates:
##         p
## 0.2016667
```

The p-value is 0.011 and we reject the null hypothesis of equal proportions at the 2% significance level. For the binomial test we get

```
binom.test(242, 1200, 0.23, alternative = 'less')
```

```
##
##  Exact binomial test
##
## data:  242 and 1200
## number of successes = 242, number of trials = 1200, p-value = 0.01
## alternative hypothesis: true probability of success is less than 0.23
## 95 percent confidence interval:
##  0.0000000 0.2216701
## sample estimates:
## probability of success
##              0.2016667
```

The result is practically the same.

(b) In the sample, 698 of those interviewed were males and 165 of them said that they are smokers. Is there evidence of a difference in the proportion of smokers between males and females?

Solution:

In this case we carry out a two-sample test for proportions, which uses the normal approximation. The null hypothesis is that the proportions are equal against the alternative that they are different. The test is:

```
prop.test(c(165,242-165),c(698, 1200-698) )
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(165, 242 - 165) out of c(698, 1200 - 698)
## X-squared = 11.985, df = 1, p-value = 0.0005363
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03671331 0.12929315
## sample estimates:
##    prop 1    prop 2
## 0.2363897 0.1533865
```

The *p*-value is small and we reject the null hypothesis of equal proportions for males and females.

(c) A similar survey was carried out simultaneously in the city of Avalon, where there was no anti-tobacco campaign, and out of 1150 persons interviewed, 283 were smokers. What test would you use to compare the proportion of smokers in the two cities? Carry out this test and discuss your results.

Solution:

Again, we have to do a two-sample proportions test.

```
prop.test(c(242, 283), c(1200,1150))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(242, 283) out of c(1200, 1150)
## X-squared = 6.4251, df = 1, p-value = 0.01125
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.078963354 -0.009877225
## sample estimates:
##    prop 1    prop 2
## 0.2016667 0.2460870
```

The *p*-value is below the 0.02 level set for the test, and we have evidence to reject the null hypothesis of equal proportions of smokers in the two cities.

Exercise 3

For this question, we use the data set `data_q3.csv`. Read the data and store it in a data frame named `q3df`. We will only use two variables in this data set, `blood_type` and `Sugar_in_blood`. The first is an integer-valued variable with values between 1 and 4. The second is a numerical variable measuring the sugar level in blood in milligrams per deciliter (mg/dL). The blood type is coded as follows:

Code	Blood type
1	O
2	A
3	B
4	AB

(a) Create a new factor called `blood_factor` in `q3df` using the information in `blood_type` but using the letter code in the table.

Solution:

We start by reading. the data

```
q3df <- read.csv('data_q3.csv')
str(q3df)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender        : chr  "M" "F" "M" "F" ...
##  $ age           : int  22 33 46 24 37 31 38 38 21 31 ...
##  $ blood_type    : int  4 4 4 3 2 3 4 2 1 3 ...
##  $ Sugar_in_blood: num  95.2 83.5 92.7 95.8 114.1 ...
```

We create the new factor using the `factor` command.

```
q3df$blood_factor <- factor(q3df$blood_type, labels=c("O", "A", "B", "AB"))
str(q3df)
```

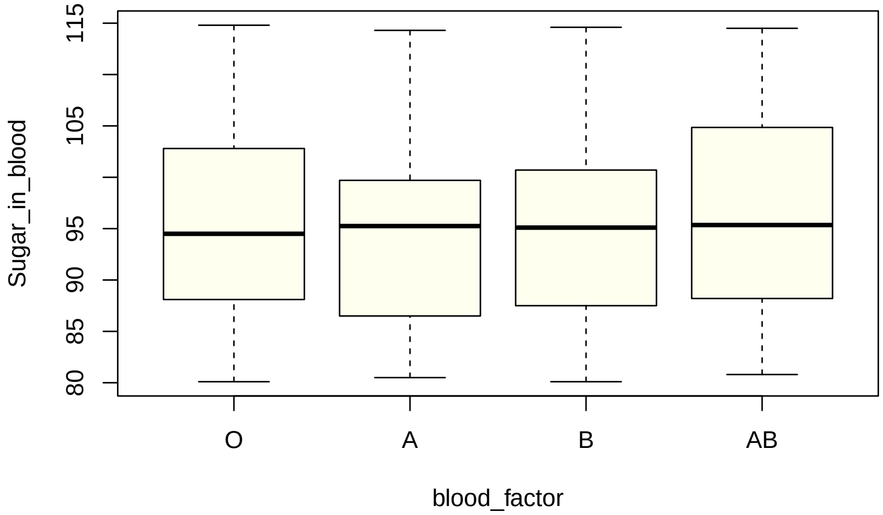
```
## 'data.frame':    500 obs. of  6 variables:
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender        : chr  "M" "F" "M" "F" ...
##  $ age           : int  22 33 46 24 37 31 38 38 21 31 ...
##  $ blood_type    : int  4 4 4 3 2 3 4 2 1 3 ...
##  $ Sugar_in_blood: num  95.2 83.5 92.7 95.8 114.1 ...
##  $ blood_factor  : Factor w/ 4 levels "O","A","B","AB": 4 4 4 3 2 3 4 2 1 3 ...
```

(b) Do boxplots of `Sugar_in_blood` as a function of `blood_factor` and comment on the graph.

Solution:

Boxplots:

```
plot(Sugar_in_blood ~ blood_factor, data = q3df, col = 'ivory')
```



The distributions are similar. They have approximately the same minimum, first quartile, median, and maximum. The main difference is the third quartile, which increases in the order `A`, `B`, `O`, and `AB`. the range of values for the four blood types is approximately the same. The median for type `O` seems lower than the rest.

(c) We want to divide the data in `Sugar_in_blood` into three groups. Up to 87 mg/dL is `low`, above 87 and up to 100 is `normal`, while above 100 is `high`. Create a new ordered factor in `q3df` called `sugar_level` with this information and levels `low < normal < high`.

Solution:

We do this with the function `cut`

```
q3df$sugar_level <- cut(q3df$Sugar_in_blood, c(0,87,100,Inf),
                 labels = c('low', 'normal','high'), ordered_result = TRUE)
str(q3df)
```

```
## 'data.frame':    500 obs. of  7 variables:
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender        : chr  "M" "F" "M" "F" ...
##  $ age           : int  22 33 46 24 37 31 38 38 21 31 ...
##  $ blood_type    : int  4 4 4 3 2 3 4 2 1 3 ...
##  $ Sugar_in_blood: num  95.2 83.5 92.7 95.8 114.1 ...
##  $ blood_factor  : Factor w/ 4 levels "O","A","B","AB": 4 4 4 3 2 3 4 2 1 3 ...
##  $ sugar_level   : Ord.factor w/ 3 levels "low"<"normal"<..: 2 1 2 2 3 2 1 3 1 3 ...
```

(d) Create a contingency table for `blood_factor` and `sugar_level`. Plot this information in a mosaic plot and discuss the results.

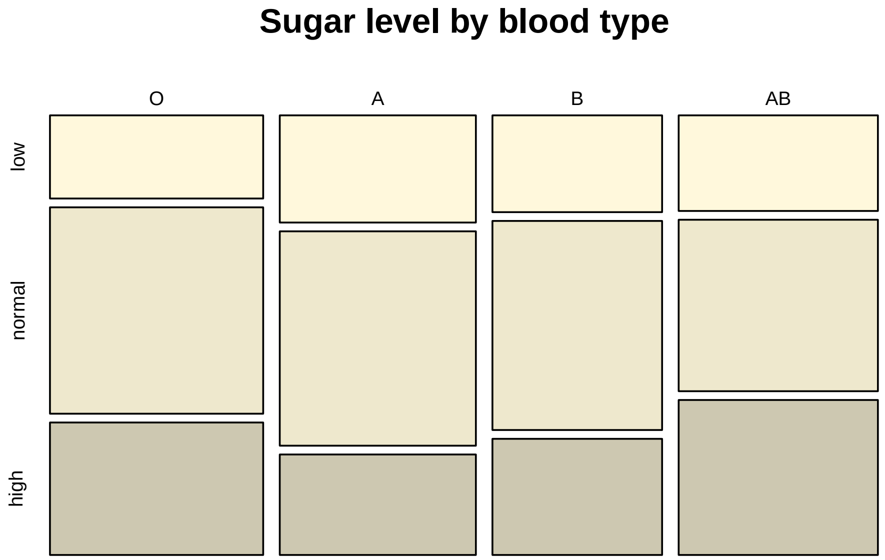
Solution:

Contingency table:

```
(tt1 <- table(q3df$blood_factor,q3df$sugar_level))
```

```
##
##      low normal high
##   O   27     67   43
##   A   32     64   30
##   B   25     54   30
##   AB  29     52   47
```

```
mosaicplot(tt1,color = c('cornsilk1','cornsilk2','cornsilk3'),
           main = 'Sugar level by blood type')
```



We see that there are some differences between the sugar levels for the different blood types, but they do not seem to be too important, and they are consistent with what we observed in the boxplots. In all blood types, a normal sugar level is more frequent than the other two levels. For blood types `O`, `B`, and `AB`, `high` is more frequent than `low`, while for type `A` it is the other way round. The proportion of `normal` sugar levels decreases along the sequence `A`, `B`, `O`, and `AB`.

(e) We want to test if the `sugar_level` is the same for all values of `blood_factor`. What test would be adequate for this? State clearly which assumptions are needed and verify that they are satisfied. Carry out this test and discuss the results.

Solution:

We can use the chi-square test. It requires that in the matrix of expected values, all entries be at least 5. We check this after the test.

```
(chtest <- chisq.test(tt1))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tt1
## X-squared = 6.5116, df = 6, p-value = 0.3684
```

The p value is large and we do not reject the null hypothesis of homogeneous distributions across the blood groups. The matrix of expected values is

```
chtest$expected
```

```
##
##         low normal high
##   O  30.962 64.938 41.1
##   A  28.476 59.724 37.8
##   B  24.634 51.666 32.7
##   AB 28.928 60.672 38.4
```

All values are above 5 so the chi-square approximation is valid.

In this case Fisher's exac test gives an error (at least in my computer):

```
fisher.test(tt1)
```

```
## Error in fisher.test(tt1): FEXACT error 6.  LDKEY=619 is too small for this problem,
##   (ii := key2[itp=1187] = 2836205, ldstp=18570)
## Try increasing the size of the workspace and possibly 'mult'
```

We can increase the size of the workspace in this case (this was not required and was not done in class)

```
fisher.test(tt1, workspace = 2e6)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tt1
## p-value = 0.3663
## alternative hypothesis: two.sided
```

The p-value we obtain is similar and we reach the same conclusion.

(f) Assume now that you only have two blood types, `A` and `B`, and that there are only two sugar levels, `normal` and `high`. You want to test if the proportion of `normal` and `high` sugar levels is the same for the different blood types. In this context, describe the null and alternative hypotheses and explicitly identify type I and type II errors. Describe the test statistic and the (asymptotic) sampling distribution.

The null hypothesis is that the frequencies of the two levels of `sugar_level` are the same for the two different types of `blood_factor` (these are the two 'populations'). Let us call p_A the proportion of `low` sugar level in the subjects with blood type `A`, and similarly for p_B and blood type `B`. Then the null and alternative hypotheses are

$$H_0 : p_A = p_B \quad \text{vs} \quad H_A : p_A \neq p_B.$$

A type I error occurs if we reject the null hypothesis of equal proportions when it is true, while a type II error occurs if the proportions differ but we do not reject the null hypothesis of equal proportions.

In this situation we can use the chi-square test or the proportions test. I will describe the first but the second is also valid.

In this context, the tables of observed and expected values are 2×2 matrices. Let $o_{ij}, i = 1, 2, j = 1, 2$ be observed values and similarly $e_{ij}, i = 1, 2, j = 1, 2$ be the expected values. The test statistic is

$$\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

Under the assumption that all entries in the table of observed values are greater than or equal to 5, this test statistic has a chi-square distribution with 1 degree of freedom. In the situation that we considered previously in the question, with four blood types and three sugar levels, the distribution has $(4 - 1) \times (3 - 1) = 6$ degrees of freedom.

Exercise 4

(a) A random sample of 200 recent blood donors at a certain blood bank shows that 68 were type A blood. Does this suggest that the actual percentage of type A donations differs from 42%, the percentage of the population having type A blood? What tests do you know that apply in this situation? Explain why they are adequate and describe their underlying assumptions. Carry out a test of the appropriate hypotheses using a significance level of .01. Would your conclusion have been different if a significance level of .05 had been used?

Solution

For this question we can use the proportions test or the binomial test. We do both

```r
prop.test(68, 200, 0.42)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  68 out of 200, null probability 0.42
## X-squared = 4.9312, df = 1, p-value = 0.02638
## alternative hypothesis: true p is not equal to 0.42
## 95 percent confidence interval:
##  0.2755772 0.4106806
## sample estimates:
##    p
## 0.34
```

The p-value is above 0,01, and we do not reject the null hypothesis that the proportion is 0.42. However, if we had tested at the 5% significance level, we would have rejected the null hypothesis.

The second test is the binomial test:

```r
binom.test(68, 200, 0.42)
```

```
##
##  Exact binomial test
##
## data:  68 and 200
## number of successes = 68, number of trials = 200, p-value = 0.02191
## alternative hypothesis: true probability of success is not equal to 0.42
## 95 percent confidence interval:
##  0.2746683 0.4101559
## sample estimates:
## probability of success
##                   0.34
```

The p-value is similar, and we would have reached the same conclusion at both significance levels.

(b) A sample obtained at a different blood bank shows that out of 175 donors, 64 were of type A. Is there evidence to suggest that this proportion differs from that of part (a) of this question? Describe clearly the hypotheses you are testing, the reasons for choosing a particular test, and the underlying assumptions, and discuss the results.

# Solution

We use the proportions test to compare the two proportions

```r
prop.test(c(68,64), c(200,175))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(68, 64) out of c(200, 175)
## X-squared = 0.16957, df = 1, p-value = 0.6805
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.12803574  0.07660716
## sample estimates:
##    prop 1    prop 2
## 0.3400000 0.3657143
```

The p-value is high and we do not reject the null hypothesis of equal proportions.