Inference for categorical data - exercises solutions

E. Pastucha

October 2020

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts --------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1. Study activity and progress

183 of 230 students that started on four of the engineering study programmes at SDU participated in the first math lecture. Of the students that participated in the first math lecture, 98 have completed their study while 85 have dropped out of the study. Similarly for the students that did not participate in the first math lecture, 14 completed their studies while 33 have dropped out of the study.

Conduct a χ^2-test to see if there is a relation between participating in the first math lecture and passing the first year test. Use a significance level of 1%. Sketch your calculations.

The question lists details about the relation between participation in a certain math lecture and the study status of engineering students. The raw observations are written in a 2 by 2 table here:

```
(study_acticity_and_progress <- tribble(
  ~StudyStatus, ~InFirstMathClass, ~NotInFirstMathClass,
  "Completed", 98, 14,
  "Dropped out", 85, 33) )
```

```
## # A tibble: 2 x 3
##   StudyStatus InFirstMathClass NotInFirstMathClass
##   <chr>                  <dbl>               <dbl>
## 1 Completed                 98                  14
## 2 Dropped out               85                  33
```

AD.1

Testing for independence - we want to see whether there is a significant interaction between the behaviour (participation in the math class) and the study status.

AD.2

H0: The behaviour and study status are independent of each other.

HA: The behaviour and study status are *not* independent of each other.

AD.3

- We assume students didn't cooperate on the exam and didn't influence each others work.
- There are more than 5 cases in each category (expected values) - look in the next point.

AD.4

The calculations consists of calculating the difference between the actual number of observations for each combination of the two variables (behaviour and study status) and the estimated number of observations given the assumption from the null hypothesis that the variables are independent.

```r
total <- 98 + 14 + 85 + 33

InFirstMathClass <- 98 + 85
NotInFirstMathClass <- 14 + 33

Completed <- 98 + 14
DroppedOut <- 85 + 33

(number_of_observations <- c(98, 14, 85, 33))
```

```
## [1] 98 14 85 33
```

```r
(expected_number_of_observations <- c(InFirstMathClass * Completed / total,
                                      NotInFirstMathClass * Completed / total,
                                      InFirstMathClass * DroppedOut / total,
                                      NotInFirstMathClass * DroppedOut / total))
```

```
## [1] 89.11304 22.88696 93.88696 24.11304
```

From the calculated differences, a χ^2 value is computed using the equation:

$$\chi^2 = \sum_{i}^{n} \frac{(e_i - o_i)^2}{e_i}$$

```r
(differences <- (number_of_observations - expected_number_of_observations))
```

```
## [1]  8.886957 -8.886957 -8.886957  8.886957
```

```
(chi_squared_value <- sum(differences^2 / expected_number_of_observations))
```

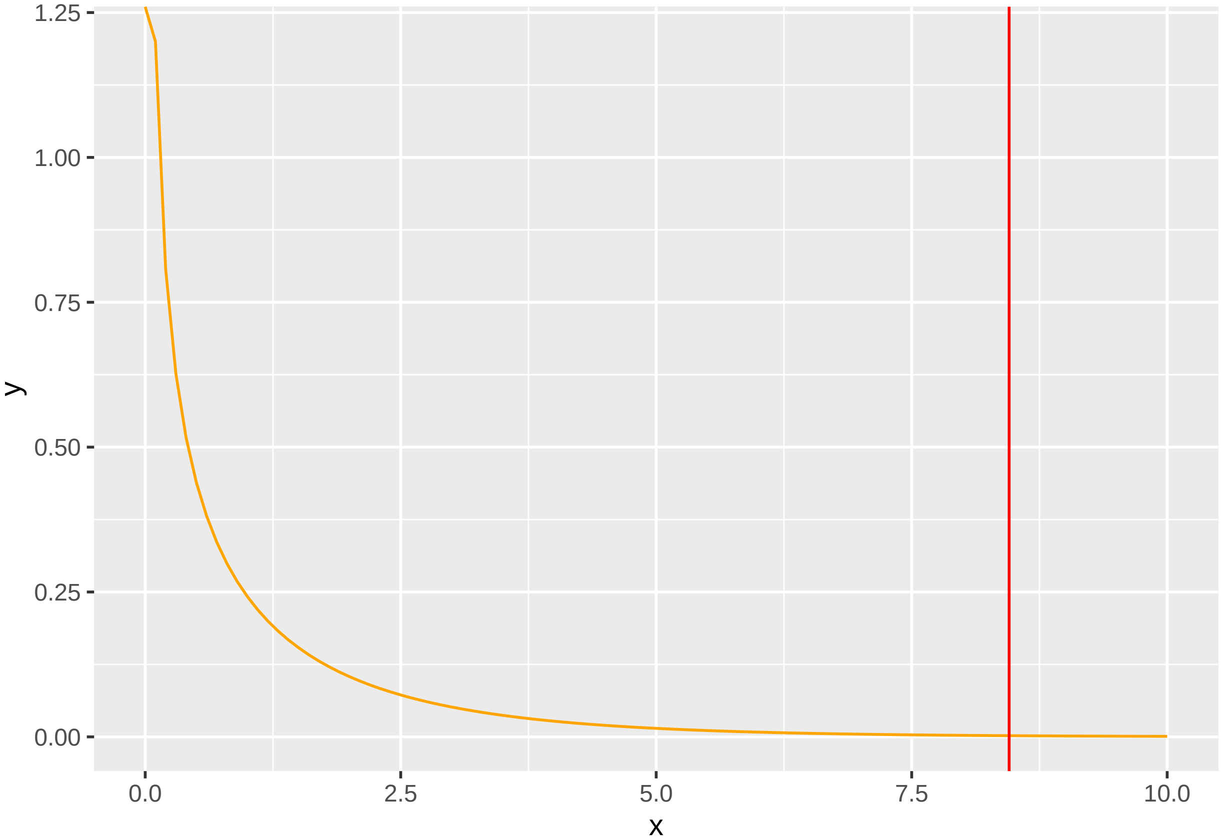
```
## [1] 8.45358
```

Finally a p value is calculated from the determined χ^2 value and the relevant number of degres of fredom (df = 1).

```
p_value <- 1 - pchisq(chi_squared_value, df = 1)
p_value
```

```
## [1] 0.003643257
```

```
ggplot(data.frame(x = seq(0, 10, length = 100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = 1), color = 'orange') +
  geom_vline(xintercept = chi_squared_value, color = 'red')
```



AD.5

The calculated p value 0.003643257 is much lower than the specified significance level of 1% = 0.01. Therefore the null hypothesis can be *rejected*, which means that the study progress and the behaviour depends on each other.

The calculations above can be verified by the *chisq.test* function in R, which gives the output below, that matches with the described calculations.

```r
study_acticity_and_progress %>%
  select(-StudyStatus) %>%
  chisq.test(correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 8.4536, df = 1, p-value = 0.003643
```

2. Olive oil and heart attacks

An experimental study looked at the effect of changing the diet of people to see whether it was related to the number of heart attacks in the different groups. The three groups were the following:

- EVOO: A small potion of Extra Virgin Olive Oil was added to the dayly diet
- Nuts: A small potion of nuts was added to the daily diet
- Control: No changes in the daily diet.

The outcome of the study is shown below, with the number of participants and events (heart attacks) in each group.

Condition	Participants	Events
EVOO	2543	96
Nuts	2454	83
Control	2450	109

AD.1

Testing for goodness of fit - we want to see if there is a difference in occurence of heart attacks within 3 groups.

AD.2 Set up hypothesis.

H0: The probability of events are the same in the three experimental conditions.

HA: The probability of events are not the same in the three experimental conditions.

AD.3

- we assume independence, as it was a scientific experiment (let's assume scientific integrity),

- There are more than 5 cases in each category (expected values) - look in the next point.

AD.4

The chi squared value is defined as the sum of squared deviations from the expected values divided by the expected value.

$$\chi^2 = \sum_{k=1}^{3} \frac{(x_k - e_k)^2}{e_k}$$

where x_k is the observed number of events for case k and e_k is the expected number of events for same case. Finally the calculated χ^2 is found in a table to get the p value.

Initially the values given in the exercise are stored in two variables.

```r
participants_in_each_group <- c(2543, 2454, 2450)
events_in_each_group <- c(96, 83, 109)
```

From the two variables, the probability of an event is calculated, by summing the observations from the three groups. The probability is then used to calculate the expected values.

```r
(total_e <- sum(events_in_each_group))
```

```
## [1] 288
```

```r
(total_n <- sum(participants_in_each_group))
```

```
## [1] 7447
```

```r
(probability <- total_e / total_n)
```

```
## [1] 0.03867329
```

```r
(expected_number_of_events <- probability * participants_in_each_group)
```

```
## [1] 98.34618 94.90426 94.74956
```

Then the χ^2 value is calculated and the method *pchisq* is used to find the proper value in the chi squared table with two degrees of freedom (as there were three observations in the dataset).

```r
expected_number_of_events - events_in_each_group
```

```
## [1]   2.34618  11.90426 -14.25044
```

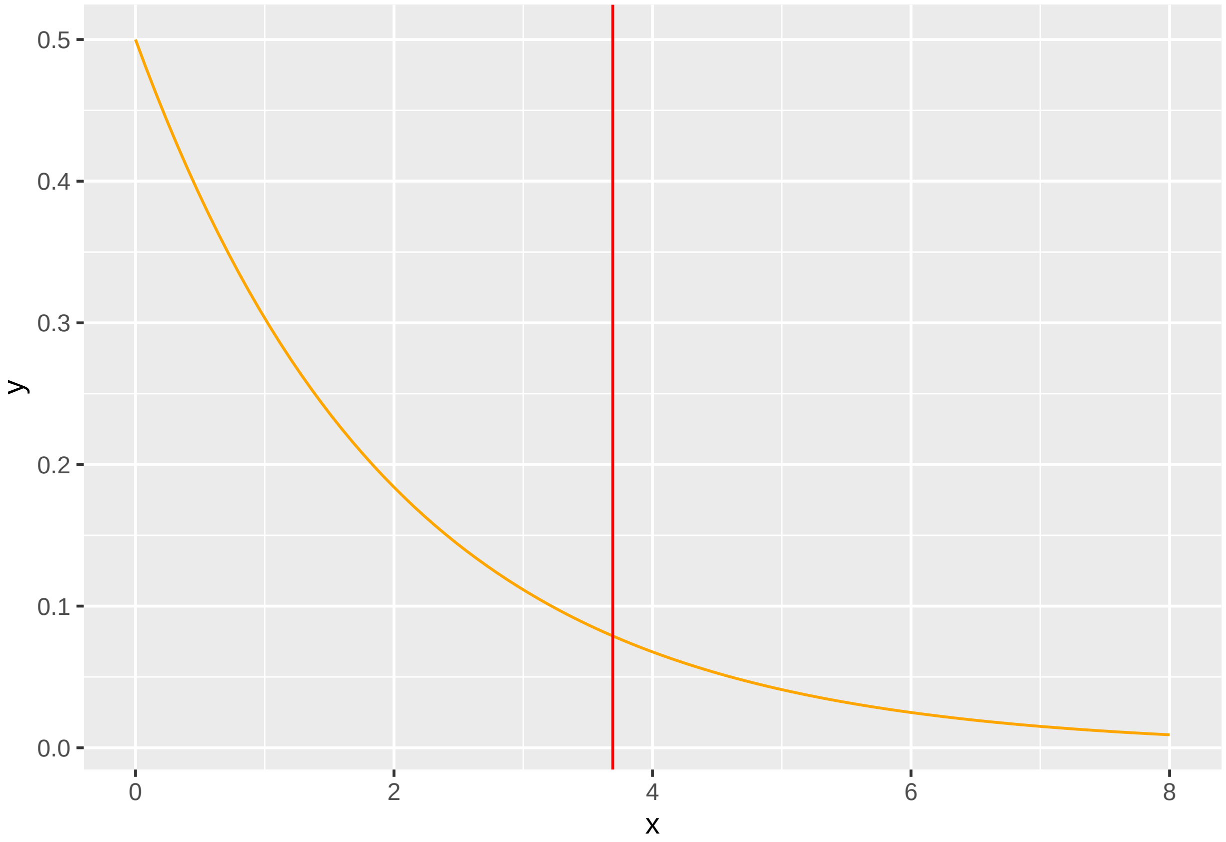
```r
(chi_sq <- sum((expected_number_of_events - events_in_each_group)^2 /
                expected_number_of_events))
```

```
## [1] 3.692455
```

```r
(pval <- 1-pchisq(chi_sq, df=2))
```

```
## [1] 0.1578314
```

```r
ggplot(data.frame(x = seq(0, 8, length = 100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = 2), color = 'orange') +
  geom_vline(xintercept = chi_sq, color = 'red')
```



AD.5 Conclusions

The determined p value 0.1578314 is larger than the usual used significance value of 5%, therefore the null hypothesis cannot be rejected.

The same result can be obtained by using the *chi.sq* method.

```
chisq.test(x = c(96, 83, 109),
           p = c(2543, 2454, 2450),
           rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(96, 83, 109)
## X-squared = 3.6925, df = 2, p-value = 0.1578
```

3. SDU Admissions in 2019

```r
admissions <- suppressMessages(readr::read_delim('admission_sdu.csv', delim = '\t'))
head(admissions, 5)
```

```
## # A tibble: 5 x 4
##   Education                   Type      `number of applicants` `number of place~
##   <chr>                       <chr>                      <dbl>             <dbl>
## 1 Amerikanske studier         Bachelors                    212                50
## 2 Audiologi                   Bachelors                    116                33
## 3 Audiologopædi               Bachelors                    188                50
## 4 Biokemi og molekylær biolo~ Bachelors                    234                40
## 5 Biologi                     Bachelors                    220                69
```

AD.1

Testing for goodness of fit - we want to see if randomly selected sample represents well population of SDU freshman.

AD.2

H0: The chosen sample fairly represents all students enrolled at SDU in 2019.

HA: The chosen sample is a biased representation of all students enrolled at SDU in 2019.

AD.3

- independence - if the sample was trully taken randomly independence is satisfied.
- number of cases

We calculate proportion within freshmen population of each group of students, and then use that proportion to calculate expected values.

Expected values are bigger than 5 for each category.

```r
all <- sum(admissions$`number of places`)
(table <- admissions %>%
  group_by(Type) %>%
  summarize(available_places = sum(`number of places`)) %>%
  mutate(observed_values = c(72,18,10)) %>%
  mutate(proportion = available_places/all) %>%
  mutate(expected_values = proportion* 100))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 5
##   Type               available_places observed_values proportion expected_values
##   <chr>                         <dbl>           <dbl>      <dbl>           <dbl>
## 1 Bachelors                      3151              72     0.763             76.3
## 2 Diploma Engineeri~              605              18     0.146             14.6
## 3 Engineering                     375              10     0.0908             9.08
```

AD.4

```
(test_statistic <- sum((table$observed_values-table$expected_values)^2/table$expected_values))
```

```
## [1] 1.101923
```

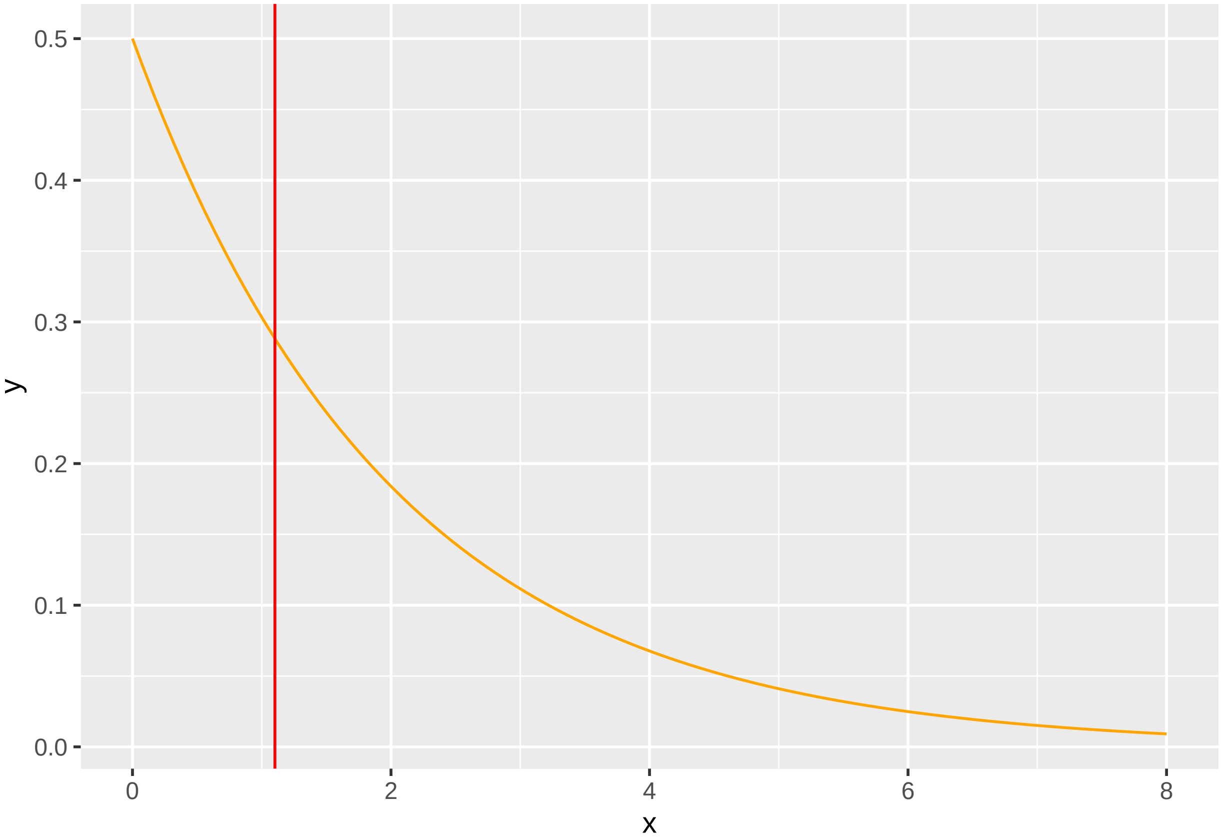
```
(p_value <- 1 - pchisq(test_statistic, df = 2))
```

```
## [1] 0.5763954
```

```
p_value > 0.05
```

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

```
ggplot(data.frame(x = seq(0, 8, length = 100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = 2), color = 'orange') +
  geom_vline(xintercept = test_statistic, color = 'red')
```



AD.5 Conclusions

The determined p value 0.5763954 is larger than the usual used significance value of 5%, therefore the null hypothesis cannot be rejected. This is a fair sample.

The same result can be obtained by using the *chi.sq* method.

```r
chisq.test(x = c(72, 18, 10),
           p = c(76.276931, 14.645364, 9.077705),
           rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(72, 18, 10)
## X-squared = 1.1019, df = 2, p-value = 0.5764
```