Tutorial 6: Tabular Analysis & the χ^2 test

Carlo Knotz

Table of contents

1 Introduction

As (social) data analysts, we are very often interested in *relationships between variables* or *causal effects*: Does one variable (e.g., gender) have an impact on another variable (e.g., political attitudes)?

We can look for evidence of relationships in two ways:

1. A *descriptive* analysis will show first evidence of whether there are any interesting patterns or differences in our dataset;

2. A *formal statistical test* will then tell us if whatever patterns we found earlier are *statistically significant*: If we can generalize the finding from our sample dataset to the general population.

In this tutorial, you learn how you can study relationships between *two categorical* variables with a descriptive tabular analysis and a formal statistical test, the χ^2 ("chi-squared") test. More specifically, we will study if men and women in Norway differ in their patterns of political participation when it comes to being a member of a trade union.

Unlike in the previous tutorials, we will start right away with a real research dataset (data from the 2014 round of the *European Social Survey*).

Tip

Hvis du ønsker å lese en norsk tekst ***i tillegg****: "Lær deg R", Kapittel 5.3.*

2 Hypotheses

Some previous studies have shown that there are some (small) differences in the rates at which men and women participate politically: Specifically, that women are a bit *less* likely to vote or to be active in political movements.[1] But other studies have found that the role of gender has been drastically changing over the last decades, so it is an open question if the relationship between gender and political participation is still true when we look at more recent data.[2]

But, because we are cautious researchers, our guiding hypothesis for this analysis is: *Women participate less politically than men.* As mentioned above, we will look at one specific form of political participation: Trade union membership.

The corresponding *null hypothesis* is that there is *no difference* between men and women when it comes to trade union membership.

[1]See Gallego (2007), "Unequal Participation in Europe", *International Journal of Sociology* 37(4):10-25.
[2]See e.g., Iversen and Rosenbluth (2010), *Women, Work, and Politics*, Yale University Press.

3 Setup

As always, you need to load all the packages you need and then import and clean your dataset. First, make sure that you are working in your project for this course (upper-right corner of the *RStudio* screen; make sure the correct project is active).

3.1 Loading packages

In this tutorial, you will need two packages: `tidyverse` for the data management and `bst290` for a special function to make cross tables. As you know, you load these with the `library()` function:

```r
library(tidyverse)
library(bst290)
```

Then you use the `read_dta()` function from the `haven` package to import the dataset — make sure that you store ("assign") the dataset with the assignment operator (`<-`):

```r
ess7 <- haven::read_dta("ess7.dta") # change the file name if necessary
```

3.2 Data cleaning

Now that the dataset is loaded, you can clean and trim it.

First, you need to convert how the dataset is stored within `R` back to the familiar format with `labelled::unlabelled()`:

```r
ess7 <- labelled::unlabelled(ess7)
```

Then you trim the data down to the variables and observations you really need. We will only work with the data for Norway, and only the following variables:

- `essround`, `idno`, and `cntry`;
- `gndr`: The respondent's gender;
- `mbtru`: Whether the respondent is (or was) a member of a trade union;

You can do both "trimming" steps in one go with the pipe operator (`%>%`):

```r
ess7 %>%
  filter(cntry=="NO") %>%
  select(essround,cntry,idno,gndr,mbtru) -> ess7
```

Always make sure that you store the result, the new "trimmed" dataset. At the very end of the code chunk, we use the "reversed" assignment operator to store the trimmed dataset under the same name (`-> ess7`) in the *Environment*!

If you like, you can also create a "dictionary" of your dataset with `labelled::generate_dictionary()`:

```r
dict <- labelled::generate_dictionary(ess7)
```

3.3 Initial exploratory data analysis (EDA)

Now that you have your dataset in order, you can take a closer look at your main variables. As mentioned in the Introduction, we will look at the relationship between gender and trade union membership in the main part of this tutorial. This means the two variables we need to work with are gender (`gndr`) and trade union membership (`mbtru`).

You can use `attributes()` to get a first overview over the two variables and how they are stored.

```r
attributes(ess7$gndr)
## $levels
## [1] "Male"   "Female"
##
## $label
## [1] "Gender"
##
## $class
## [1] "factor"
attributes(ess7$mbtru)
## $levels
## [1] "Yes, currently"  "Yes, previously" "No"
##
## $label
## [1] "Member of trade union or similar organisation"
##
## $class
## [1] "factor"
```

Both variables are *categorical* variables and should therefore be stored as *factors* — and that is indeed the case.

In a next step, you should check how many observations you have per category of each variable, and if there are some empty categories. You do this with the `table()` function.

In the case of the `gndr`-variable, this looks as follows:

```
table(ess7$gndr)
##
##   Male Female
##    764    672
```

It turns out that there a few more men than women in the dataset, and there are no "empty" categories.

The situation is similar in the case of the `mbtru` variable:

```
table(ess7$mbtru)
##
##  Yes, currently Yes, previously              No
##             663             277             493
```

4 Bivariate analysis

4.1 Descriptive analysis

As explained in Kellstedt & Whitten (2018, Chapter 8), the standard way to starting looking for relationships between categorical variables is to cross-tabulate them. This table would be organized like this:

- The *independent* variable would appear in the *columns*;
- The *dependent* variable would appear in the *rows*;

In this analysis, we expect that gender (`gndr`) has an effect on trade union membership (`mbtru`). This means,

- `gndr` is the *independent* variable;
- `mbtru` is the *dependent* variable;

4.1.1 A cross-table showing frequencies

To cross-tabulate the two variables, you can use the `table()` function, which you already know from before. The variable that should be shown in the rows is named first, then the variable that should be shown in the columns:

```
table(ess7$mbtru,ess7$gndr)
##
##                    Male Female
##   Yes, currently   328    335
##   Yes, previously  164    113
##   No               271    222
```

This table shows you the "raw" numbers of respondents that fall into each category — the *frequencies*. As it is now, this table is a bit difficult to read because it is not easy to see if there is an over- or underrepresentation of women or men among trade union members.

4.1.2 A cross-table showing percentages

A better way to present the data is in the form of a cross-table that shows *column percentages* (see again Kellstedt and Whitten 2018, Chapter 8). Creating such a table is possible in `R` (see Hermansen 2019, *Lær deg R* Chapter 5.3.1), but that is a multi-step process with lots of potential for error messages and frustration.

To make this easier and quicker, you can use a special function from the `bst290` package called `krysstabell()`. When using this function, you just need to specify a) the dataset; b) the variable that should be shown in the rows; and c) the variable to be shown in the columns:

```
krysstabell(dataset = ess7,
            rowvar = "mbtru",
            colvar = "gndr")
##                    gndr
## mbtru                 Male Female    Sum
##   Yes, currently     42.99  50.00  46.27
##   Yes, previously    21.49  16.87  19.33
##   No                 35.52  33.13  34.40
##   Sum               100.00 100.00 100.00
```

This table is a bit easier to read: You see that around 43% of men in the sample are members in a trade union, while the share of trade union members among women is 50%. Conversely, fewer women than men used to be members previously or are not members. This indicates that there might be a relationship between gender and trade union membership in Norway — but here, women are *more, not less* likely to be members than men.

The `krysstabell()` function has an export-option (`export=TRUE`) that you can use to get output that you can copy/paste into your Word document and style there (see also Tutorial 2).

4.2 Running a χ^2 test

The previous descriptive analysis suggests that there *might* be a relationship between gender and trade union membership. *But:* It is possible that this reflects only *sampling variation*: After all, we are working with a *sample* of the entire Norwegian population. Maybe we were really unlucky and ended up by mere chance with a survey sample in which women are more likely to be trade union members, even though this is not true for the wider Norwegian population?

To find out if our relationship is *systematic* or *statistically significant*, we can use the χ^2 test. As you know, this is a formal statistical test for relationships between two categorical variables.

Running this test in `R` is easy because there is a built-in function for it: `chisq.test()`. Usually, you run this function directly on the cross table.[3]

This means: You first cross-tabulate your two variables and save the result:

```r
crosstab <- table(ess7$mbtru,ess7$gndr)
```

Then you use the table that you just saved in the `chisq.test()` function:

```r
chisq.test(crosstab)
##
##  Pearson's Chi-squared test
##
## data:  crosstab
## X-squared = 8.3335, df = 2, p-value = 0.0155
```

[3]You can technically also enter two factor-variables like this: `chisq.test(ess7$mbtru,ess7$gndr)`. However, creating the table first helps to remind you that you always run the actual test on the table, not on the underlying data.

4.3 Interpretation

```
chisq.test(crosstab)
##
##  Pearson's Chi-squared test
##
## data:  crosstab
## X-squared = 8.3335, df = 2, p-value = 0.0155
```

The output should be possible to understand if you read Kellstedt and Whitten (2018, see Chapter 8):

- `df` refers to the degrees of freedom. You calculate these as $df = (r-1)(c-1)$ where r is the number of rows of the cross table, and c is the number of columns. In this case, this is $(3-1)(2-1) = 2 \times 1 = 2$.
- `X-squared` is the χ^2 statistic. This is the result of a longer calculation that is explained in Kellstedt and Whitten (and also the interactive dashboard included in the `bst290` package; see `bst290::practiceStatistics()`);
- `p-value` is, obviously, the associated p-value.

Can you make sense of the result?

5 Conclusion

You now know how to use `R` to test for relationships between two categorical variables and to see if you can generalize any patterns you find in your sample data to the underlying general population.

There are again some de-bugging exercises (plus quizzes) that you can use to practice tabular analysis and interpreting χ^2 tests ("De-bugging exercises: Tabular analysis"). There, you will study the relationship between gender and voting in elections (using the small practice dataset).

6 (Voluntary) Creating mosaic plots

A cross table showing column percentages is a good way to show relationships between categorical variables — but there is another option that can be even better in some cases: a *mosaic plot* (see also Hermansen 2019, *Lær deg R*, pp. 91-2). Such a plot is, simply put, a graphical representation of a cross table and it is often the case that relationships are easier to see in the form of a graph than when they are presented in table-form.

Creating a mosaic plot is not difficult because there is a dedicated function for it in `R`: `mosaicplot()`. This function works basically the same way as the `chisq.test()` function: You first create a simple cross table of two categorical variables with `table()` and then feed the result of that operation into the `mosaicplot()` function. The only thing to be aware of is that the variables go into the `table()` function in the opposite order than they would otherwise.

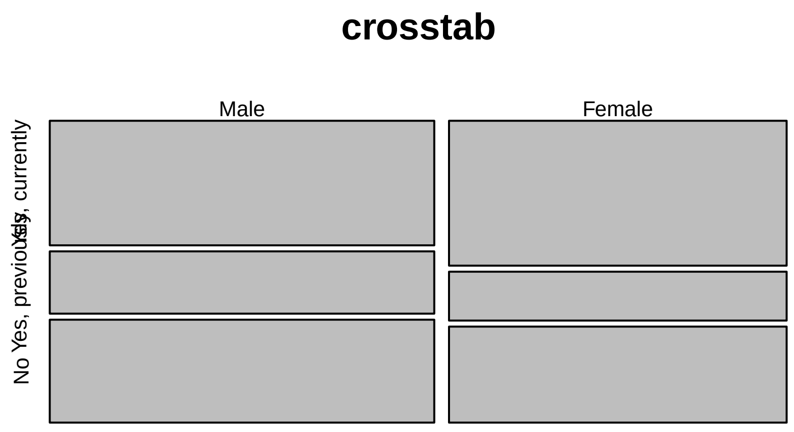
For example, to get a mosaic plot that shows the relationship between gender and trade union membership, you would first create a cross table:

```
crosstab <- table(ess7$gndr,ess7$mbtru)
```

Notice that the variables are in the *opposite order* compared to when we created the frequency table above (compare to section 4.1.1): The independent variable (`gndr`) comes first, and second is the dependent variable (`mbtru`).

Then you use the saved table in the `mosaicplot()` function:

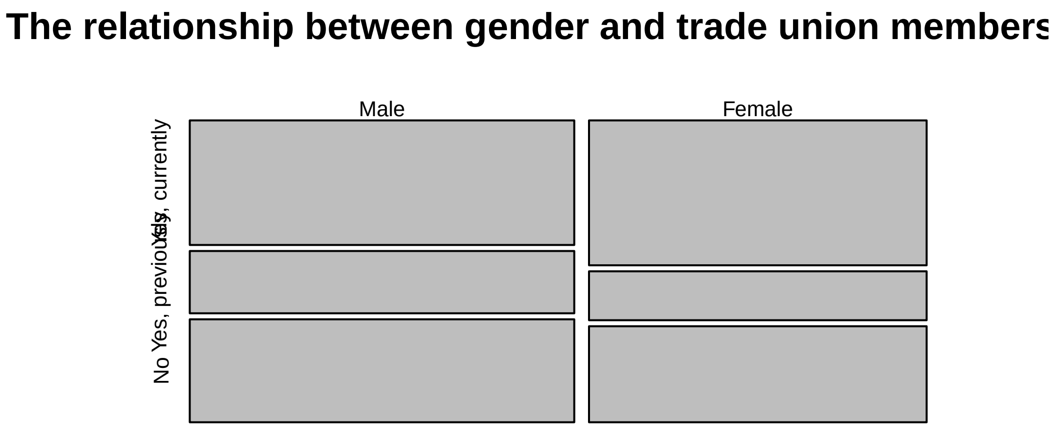
```
mosaicplot(crosstab)
```



The graph shows quite nicely the higher share of women who are currently members of a trade union and, conversely, the higher share of men who are not or who used to be.

You can also give the plot a nicer title by setting the `main`-option:

```r
mosaicplot(crosstab,
           main = "The relationship between gender and trade union membership")
```



When you are happy with the result, you can export the graph with the `pdf()` function (`ggsave()` does not work here, because we are using `base R` graphics, not `ggplot2`):

```r
pdf("mosplot.pdf") # this opens a "graphics device"

mosaicplot(crosstab, # this creates the graph
           main = "The relationship between gender and trade union membership")

dev.off() # you always need to "turn off" the "graphics device" when you have saved your grap
```