Comparison of two qualitative variables
(also known as categorical variables, or factorial variables)

Emmanuelle Becker
May 15, 2018

1 Introduction

Statistics studies populations, whose elements are called individuals. With a one-dimensionnal statistic series, one variable is measured for each element of the population. With a two-dimensionnal statistic series, two variables are measured for each individual. Very often (and this is the classical rule in biology), we do not have access to the entire population and we are thus studying a sample.

A variable can be:

- **Qualitative** (synonyms: categorical, factorial) if the information is non-numeric. The variable can take several modalities.

- **Quantitative** if it can be measured and ordered. The variable can take several values. Depending on the variable that can be taken, the variable can be quantitative and continuous (if all values are possible) or quantitative and discrete (if only some values are expected, for example only natural numbers).

In this practical session, we consider we have two-dimensionnal series with two qualitative variables.

2 Parameters and graphical representations

A parameter is a value that describes a variable\(^1\). With qualitative variables, it is difficult to identify good parameters. The only parameter that can be used is the mode, i.e., the most frequent modality.

The best way to describe a qualitative variable is to use a contingency table. A contingency table means a table that counts the number of individuals for each possible modality.

Some examples that you can try with R:

```r
myTable = read.table("data-individuals.txt", header=T, sep="\t")
str(myTable)
table(myTable$Glasses)
table(myTable$Gender)
```

The graphics that can be used with a one-dimensionnal qualitative variable are most often (i) `barplot` and (ii) `pies`. Some examples you can try with R:

\(^1\) for example, the **mean** and the **median** are parameters for quantitative values
The best way to describe two qualitative variables simultaneously is also to use a contingency table, but this time with two dimensions. A little example that you can try with R:

```r
table(myTable$Glasses, myTable$Gender)
```

### 3 Methodology for statistical test (the $\chi^2$)

With two qualitative variables, the questions that can be addressed are: (i) is one variable inducing a bias for the other variable? (ii) are the two variables independent or not? (iii) is the distribution of variable 1 different for each modality of variable 2?

All these questions rely on the same statistical test, the $\chi^2$ test.

The $\chi^2$ test calculates the expected values for the contingency table if the variables are independent, and then estimates if the observed values are far from these expected values or not.

To run a $\chi^2$ test with R, you need a contingency table. The command is the following:

```r
chisq.test(table(myTable$Gender, myTable$Glasses))
# hint : to see the expected or the observed values :
chisq.test(table(myTable$Gender, myTable$Glasses))$expected
chisq.test(table(myTable$Gender, myTable$Glasses))$observed
```

**Pre-requisites for using a $\chi^2$ test**. To use the $\chi^2$ test, one usually wants the expected values to be all superior to 5. If it is not the case, see the `exact.test`.

**H0** The null hypothesis is that the two variables are independent (no bias, homogeneity).

**H1** The alternative hypothesis is that the two variables are not independent (there is a bias, no homogeneity).

**Interpreting the p-value** The p-value is the probability to observe such a difference between observed and expected values while H0 is true (i.e., just by chance). If the p-value is small, it is unlikely that H0 is true, so we reject H0. If the p-value is not small enough, we do not reject H0 (so we accept it...).

**The appropriate sentence** With a risk $\alpha = 0.05$, we [can/cannot] reject H0, and so we conclude that the variables [are/aren’t] independent.

**Biological interpretation** Always interpret your results biologically, you are a biologist!

### 4 A toy example

In this toy example, we provide you a dataset and we ask precise questions.

**Exercise 1**: Download the dataset `data-individuals` in your environment. You can use
the same commands than in the previous lesson.

**Exercise 2:** Generate a contingency table to see how many students choose each Major.

**Exercise 3:** We want to know if there is a gender bias in the choice of the major:

1. generate a contingency table to see how the major are distributed relatively to the gender of the student;
2. with a $\chi^2$ test command (`chisq.test`), generate the observed table (it should be the same as the table you generated before);
3. with a $\chi^2$ test command (`chisq.test`), generate the expected table;
4. with a $\chi^2$ test command (`chisq.test`), find the value of the statistic, and the p-value: what can you conclude?

**Exercise 4:** We want to know if the majors are chosen with equiprobability.

1. look at the help page of the `chisq.test` function with the command `?chisq.test`;
2. in this help page, find the argument that can be added in the command so that the contingency table is compared to probabilities;
3. perform your $\chi^2$ test command (`chisq.test`);
4. find the value of the statistic, and the p-value: what can you conclude?

**5 An exploratory analysis**

In this section, we provide you a dataset and we would like you to (i) imagine your own questions, and (ii) propose the right method with R to answer your question. **Asking a precise question that can be answered with a dataset is not an easy task! Don’t hesitate to ask us if we can help!**

The dataset is about the use of electronic devices, for both parents and their child. The studied electronic devices for the parents are: smartphone, computer and tablet, and for the children: smartphone and television.