TP4 : Comparison of two qualitative variables
(also known as categorical variables, or factorial variables)

Emmanuelle Becker

February 26, 2019

Contents
1 Introduction 1
2 Parameters and graphical representations 2
3 Methodology for statistical test (the $\chi^2$) 2
4 A toy example 3
5 An exploratory analysis 4
6 Small project about type I and type II errors 5
   6.1 dpqr functions .......................... 5
   6.3 Loading the dataset .......................... 5
   6.4 First statistical tests .......................... 6
   6.5 Simulating H0 .......................... 6
   6.6 Simulating H1 : Type II error .......................... 7
   6.7 Conclusions .......................... 8

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 :

```r
barplot(table(myTable$Glasses), main="Wearing glasses?")
barplot(table(myTable$Gender), main="Genders", col=c("magenta","cyan"))
pie(table(myTable$Glasses), main="Wearing glasses?")
pie(table(myTable$Gender), main="Genders")
```

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
```

\(^1\)for example, the **mean** and the **median** are parameters for quantitative values
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 Fisher test.

$H_0$ The null hypothesis is that the two variables are independant (no bias, homogeneity).

$H_1$ The alternative hypothesis is that the two variable are not independant (the 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 $H_0$ is true (ie. just by chance). If the p-value is small, it is unlikely that $H_0$ is true, so we reject $H_0$. If the p-value is not small enough, we do not reject $H_0$ (so we accept it...).

The appropriate sentence With a risk $\alpha = 0.05$, we [can/cannot] reject $H_0$, and so we conclude that the variables [are/aren’t] independant.

Biological interpretation Always interpret your result 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. perfom 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

**Exercise 5:** 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.

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.
6 Small project about type I and type II errors

6.1 dpqr functions

For each usual probability law, there are 4 functions in R that allow:

1. to get the probability associated to a given value: d
2. get the probability of being under a given value, or calculate the distribution function (French : fonction de répartition): p
3. calculate the quantiles: q
4. generate a random sample: r

Thus, for the Normal distribution, we find: \texttt{dnorm, pnorm, qnorm} and \texttt{rnorm}.

\begin{Verbatim}
Examples :
> ?dnorm
> dnorm(0)
> dnorm(2, mean=5, sd=2)
> pnorm(-1.5)
> qnorm(0.067)
> qnorm(pnorm(-1.5))
> qnorm(0.975)
> rnorm(10)
\end{Verbatim}


The Deep Blue matches against Kasparov are two matches of six chess games each between Russian world champion Garry Kasparov and an American supercomputer called Deep Blue.

- The first match took place in Philadelphia in 1996 and was won by Garry Kasparov (4-2).
- The second confrontation took place in New York in 1997 and was won by Deep Blue (3.5-2.5), but out of the conditions required at the world championships. Deep Blue had evolved slightly and was then called Deeper Blue.

The \texttt{Echecs.txt} file summarizes the main information about these parts.

\textbf{Is Kasparov better than Deep Blue ? Is Deep Blue better than Kasparov ?}

6.3 Loading the dataset

\begin{Verbatim}
# Change your working directory
setwd(<complete>)

# Load the dataset
tab = read.table("Echecs.txt", header=T, sep="\t")
head(tab)
\end{Verbatim}
**Exercise 6:** First, we are interested in the results obtained.

- Load the dataset as described above.
- Briefly describe the results obtained, and contained in tab. You can use tables, charts, or any other representation that you think is relevant.

**6.4 First statistical tests**

We expect a perfect writing quality (especially for the tests, with a highlighting of the observed numbers, expected ones, of the p-value of the test, as well as a conclusion and its interpretation).

**Exercise 7:** Find the appropriate test.

- At the error risk $\alpha = 5\%$, can we say that we win more easily with one color than with another?
- At the error risk $\alpha = 5\%$, can we say that the winner is consistent with a uniform distribution (1/3 wins for Kasparov, 1/3 wins for DeepBlue, 1/3 of draws)?
- At the error risk $\alpha = 5\%$, are the results of the two confrontations (1996 and 1997) homogeneous? In other words, do we see a significant effect of the transition from Deep Blue to Deeper Blue?
- Finally, what can we conclude from this double confrontation?

**6.5 Simulating H0**

In this part, we will only consider games that did not lead to a draw (there were $n=7$ games won by one participant). Our null hypothesis $H_0$ is that the probability of success of Kasparov is 0.5.

We will therefore consider that the number of games gained by Kasparov follows a binomial law of parameters $Bi(n=7, p=0.5)$

Observe the following code (you can also execute it, unless you won’t see the graph):

```r
n = 7
proba_victory = 0.5

## Diagram
x = seq(0, n, le = n+1)
Fx = pbinom(x, size=n, prob=proba_victory)
# Fx if the distribution function of Bi(n=7, p=0.5)
# in French : Fx est la fonction de répartition de la loi Bi(n=7, p=0.5)
plot(x, Fx, xlab="x", ylab="F(x)", type="p",
     main = "Distribution function ie. P(X <= x)")
abline(h = 0.95, col="red", lty="dashed")
# The dashed line represent the values whose probability to appear
# under H0 are less than 5%
```

**Exercise 8:** When you see this graph, what is the minimal value of games that Garry Kasparov has to win so that $H_0$ is rejected?
We can also compute this value directly with R:

```r
## To reject H0
# (remind : q is the reciprocal function of the distribution function)
# (in French : q est la fonction réciproque de la fonction de répartition)
min_number_of_victories_to_reject_H0 = qbinom(0.95, size=n, prob=proba_victory)
min_number_of_victories_to_reject_H0
```

So, if you got it, between 0 and 5 games won by Kasparov, the designers of Deep Blue will say "Well, at risk of error $\alpha = 5\%$, we can not reject H0 and so Deep Blue is as strong as Kasparov". And this is actually true, because we designed the graph so that Kasparov wins with a probability of $p = 0.5$ (it was our hypothesis).

**Exercise 9:** What would have been the non-rejection interval of H0 for $n = 10, 20, 50, 100$?

- Draw 4 graphs with the 4 distribution functions for $n = 10, 20, 50, 100$.
- You will be careful to add a line indicating the delineation between the H0 acceptance zone and the H0 rejection zone at $\alpha = 5\%$ risk.

### 6.6 Simulating H1: Type II error

In the previous sections, we saw that between 0 and 5 parts out of 7 won by Kasparov, designers Deep Blue can say "At risk of error $\alpha = 5\%$, we can not reject H0 and so Deep Blue is as strong as Kasparov". But they may be wrong in saying this, and we are going to work on that type II error rate.

Suppose Kasparov is actually better than DeepBlue and beats him with a probability of 0.9. The number of victories Kasparov, if there are 7 parts, will now follow a binomial law $Bi(n = 7, p = 0.9)$

```r
n = 7
barplot(dbinom(0:n,size=n,prob=0.9), col=c(rep("blue", 5+1), rep("red",n-5)))
```

**Exercise 10:** Interpretation of this graph.

- What does the graph represent? What do the colors on the chart represent? Do not hesitate to complete it to add titles etc ...
- Under the assumption that in reality the number of victories of Kasparov follows a law $Bi(n = 7, p = 0.9)$, what is the probability for DeepBlue designers to accept H0 wrongly?

This error of accepting H0 while H0 is wrong is named second-order error, or type II error, or error $\beta$. This error rate depends on:

1. the risk of error $\alpha$ used to reject or not H0, since it is this risk of error that determines the values leading to the acceptance or rejection of H0;
2. how Kasparov actually is better than DeepBlue (ie. the $p$ in the binomial law);
3. how many games are played (ie. the $n$ value of the binomial law).
Not easy to understand? Let’s produce one more graph...

```r
mystery_function <- function(n, trait, color) {
  x <- seq(0.5, 1, le = 100)
  y <- 1 - pbinom(qbinom(0.95, n, 0.5) - 1, n, x)
  lines(x, y, lty=trait, col=color)
}
x = seq(0.5, 1, le = 100)
y = seq(0, 1, le = 100)
plot(x,y, xlab="x ???", ylab="y ???", type="n")
abline(h = 0.95, col="red", lty="dashed")
mystery_function(7,1,1)
mystery_function(10,2,1)
mystery_function(20,2,1)
mystery_function(30,2,1)
mystery_function(100,2,1)
legend(0.85,0.5,lty=c(1,2,2,2,2),legend=c("n = 7","n = 10","n = 20",
                                  "n = 30","n = 100"), col=c(1,1,1,1,1))
```

**Exercise 11:** Explain what the mystery function does and comment on this graph. You will help yourself by changing the color of the curves to find you there, and you will think to add titles of graphic and adequate axes.

6.7 Conclusions

**Exercise 12:** Try to think about the following questions:

- What feedback can you draw from this statistical analysis?
- Can we really say that we "accept" a H0 hypothesis?