

PLANTER ESTADÍSTICA 2018

TO WHAT EXTENT HAVING PREVIOUS INFORMATION CHANGES YOUR ANSWER IN GENERAL KNOWLEDGE QUESTIONS



Laura Adell
Miriam Lorenzo
Astrid Mayné
Catalina Palacios
Ariadna Sorribas

TO WHAT EXTENT HAVING PREVIOUS INFORMATION CHANGES YOUR ANSWER IN GENERAL KNOWLEDGE QUESTIONS

May 4, 2018

Laura Adell, Míriam Lorenzo, Astrid Mayné, Catalina Palacios, Ariadna Sorribas

Abstract

The purpose of this study is to show that, when we are asked a question, our answer is conditioned by all the other things we already know. To prove this, we asked to two groups of people of the same population two questions, but each question was given in different ways depending on the group: to one group we just stated the question and to the other one we gave them some extra information.

With this statistical study we want to show that we are really conditioned by all the information that we have, so our answers will never be objective. To demonstrate it we just have to compare the results of both groups and see if the answers are similar or if we can observe a tendency to the same answer in the group that has been conditioned.

Contents

1	Introduction	3
2	How did we obtain the data?	3
2.1	Selection of the questions	3
2.2	Selection of the sample	4
2.3	Description of the surveys	4
3	Analysis of the data	5
3.1	Representation of the data	5
3.2	Statistic χ^2 test	7
3.3	Gender, age and region aspects	13
3.3.1	Gender	13
3.3.2	Age	15
3.3.3	Region	16
4	Conclusion	18
5	Annex	19

1 Introduction

The idea of this paper came up to us one day in which we were searching information in several media to do an academic work. We saw that the same event was explained in different ways on different sources. We often accuse mass media of having manipulated the real facts to condition the opinion of people but we do not really know to what extent we are safe to say that. In other words, we wanted to know whether it was possible to influence other people's opinion by giving them different previous information. Therefore, our motivation for this project was that we wanted to know how much someone's opinion can change if you give them some extra information, so we can see if they are sure of what they thought to be true before or if they can be conditioned.

What we expect to encounter in the results is a tendency in the questions where extra information was given to choose the "more logical" answer based in what we said to them, while in the other question the answers should be more distributed between the four options because, as most of them do not know the answer, it is a random election and the four options have the same probability to be chosen.

2 How did we obtain the data?

2.1 Selection of the questions

Our main objective was to select simple questions involving themes of general culture. We selected the questions using data from Idescat [1], so we could define concrete answers depending on the results we obtained. For the first question we looked for the most popular names for girls in the past years. We noticed that Martina, that has been the most popular name this year, has maintained its popularity since 2013. The four previous years was Julia. We decided to ask for 2013 since it was the year of the last change in Catalan newborns. For the second question, we looked at statistics about how many people acquired the Spanish nationality each year and we noticed that there was a crescent tendency, but it was broken in 2014. So we decided to give information that would mislead the surveyed into thinking that the crescent tendency continued in 2014 to see if the people would follow this conditioning or not.

2.2 Selection of the sample

To obtain the information we created two surveys in Google Forms (see in the Annex 18 and 19) that we sent to people from our school. We noticed that this way we would only obtain results from people of the same age (between 11 and 20 years old) so we decided to try to send it to our families and friends to have a more varied sample.

To obtain the most realistic result we could, we did both groups as equal as possible. To do that, we chose to separate the people we surveyed by gender (male, female, other), region (which province they are living) and by age (11-20, 21-30, 31-40, 41-50, 51-60, +60). We also did these separations to see if there were some other things we could study, such as if people is more impressionable at some age or if depending on your gender your tendency to be conditioned is bigger.

2.3 Description of the surveys

Our work consists of two surveys with the same two questions and 4 answers. We first ask the age, the gender, and the country or region. The difference between both surveys is that in one we give previous information and in the other we don't.

Initial questions in both surveys:

- Gender
- Age
- Region

Survey 1:

- Question 1:
Knowing that the most common name among the newborns of 2012 was Júlia, what do you think was the most common girl's name among the newly born in 2013 in Catalonia??
 - Maria
 - Paula
 - Júlia
 - Martina

- Question 2:
How many people do you think obtained the Spanish nationality in 2014?
 - 93 700
 - 150 300
 - 312 100
 - 480 200

Survey 2:

- Question 1:
What do you think was the most common name of girl among the newly born in 2013 in Catalonia?
 - Maria
 - Paula
 - Júlia
 - Martina
- Question 2:
Knowing that in 2012 there were 115500 people who obtained the Spanish nationality and in 2013 there were 261300, how many do you think were in 2014?
 - 93 700
 - 150 300
 - 312 100
 - 480 200

3 Analysis of the data

3.1 Representation of the data

After having done the survey to 450 people (one half in the first survey and the other half in the second), we did this pie charts to show the data we obtained:

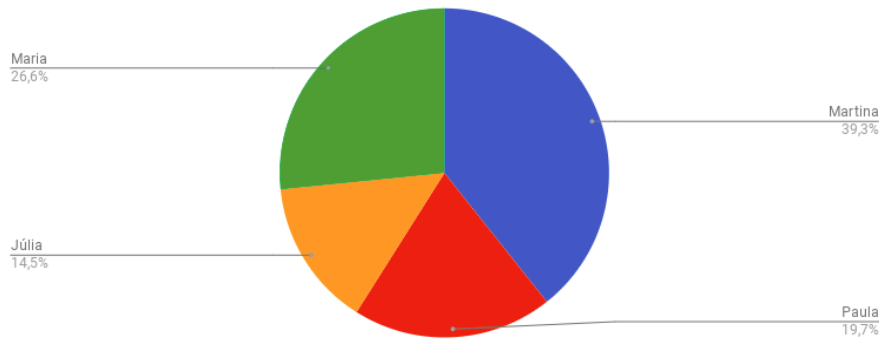


Figure 1: Pie chart for the question 1 in the first survey: *Knowing that the most common name among the newly born of 2012 was Júlia, what do you think was the most common girl's name among the newly born in 2013 in Catalonia?*

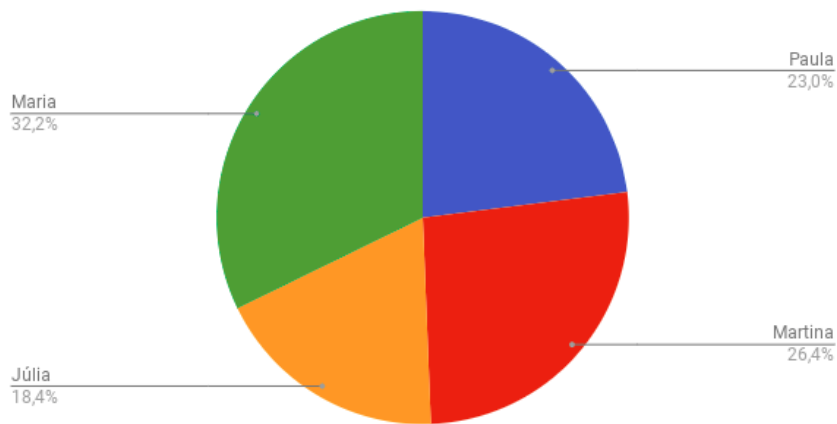


Figure 2: Pie chart for the question 1 in the second survey: *what do you think was the most common girl's name among the newly born in 2013 in Catalonia?*

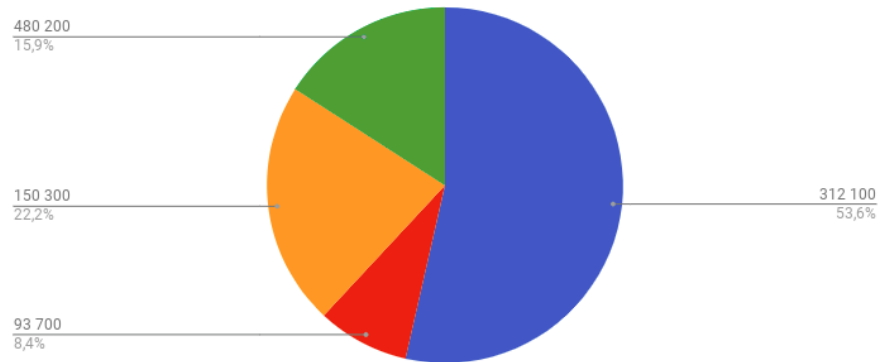


Figure 3: Pie chart for the question 2 in the second survey: *Knowing that in 2012 there were 11,55500 people who obtained the Spanish nationality and in 2013 there were 261300, how many do you think were in 2014?*

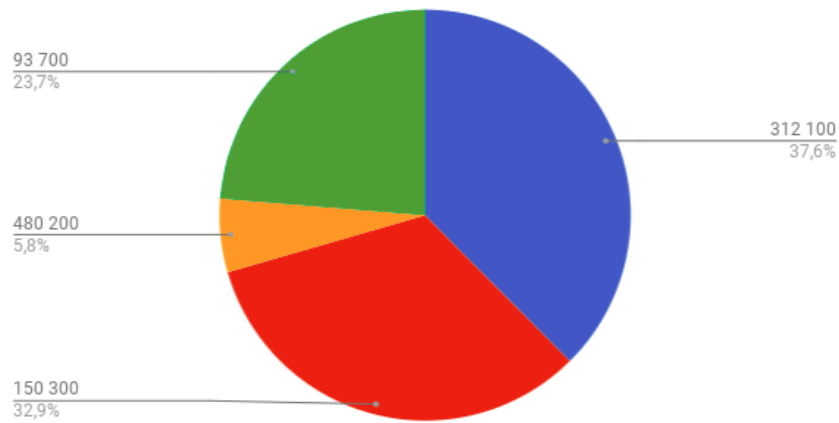


Figure 4: Pie chart for the question 2 in the first survey: *How many people do you think obtained Spanish nationality in 2014?*

3.2 Statistic χ^2 test

We are working with qualitative data so we searched for a statistic tool appropriated to what we are working on. Since the majority of statistical parameters or tests are better suited for quantitative data we decided to apply the χ^2 test.

χ^2 test: The χ^2 test is an statistical hypothesis test used to determine whether there is a significant difference between the expected frequencies and the observed frequencies.

Our main hypothesis consists on the idea that the more information we have, the more likely we are to choose an option, even if we have no idea whether it is the right answer or not, that seems congruent with the information given to us beforehand. Therefore, we can say that our main hypothesis H_1 states that: "the information given beforehand has a direct effect on the answers the sample chose" and, in order to perform the χ^2 test we need a null hypothesis, in this case being H_0 : "The information given beforehand, the independent variable, has no direct relationship with the option chosen by the sampled population".

H_0 : there is no dependence between answers and information.

H_1 : there are evidence of some relation between if there is or not previous information given and the change of answer depending on it.

For having a simplest view we have represented this information in two graphics 5 and 6:



Figure 5: Answers in question 1: *what do you think was the most common girl's name among the newly born in 2013 in Catalonia?*

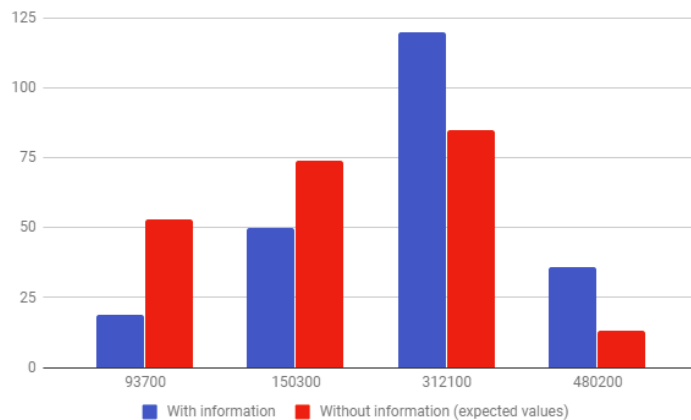


Figure 6: Answers in question 2: *How many people do you think obtained the Spanish nationality in 2014?*

In order to know if the results have any relationship with the questions we will be applying the χ^2 test. First of all, we will calculate the statistic parameter and if the result is above 0,05 we will immediately discard the independence hypothesis, therefore the data will depend on the information given in the question. We will use the results obtained from the question asked without previous information as neutral values in order to determine whether the information added influences the following response.

First, we put in a table the answers we have received and then compute the expected values for each answer:

Table of contingency for question 1:

	Martina	Paula	Júlia	Maria	TOTAL
With information	88	44	33	60	225
Without information	59	52	41	73	225
Total	147	96	74	133	450

Table 1: Contingency table for question 1

The expected values we will use to calculate the χ^2 are the probabilities of choosing each name supposing that there is the same chances of choosing an answer with or without information.

Expected values for our answers:

$$E_{i,j} = \frac{T_i T_j}{T}$$

where T is the big total, T_i is the total for the row i , which corresponds to with and without information, and T_j is the total for the column j , which correspond to the names.

Table of expected values for question 1:

	Martina	Paula	Júlia	Maria	TOTAL
With information	73.5	48	37	66.5	225
Without information	73.5	48	37	66.5	225
Total	147	96	74	133	450

Table 2: Table of expected values for question 1

Contingency table for question 2:

	93700	150300	312100	480200	TOTAL
With information	19	50	120	36	225
Without information	53	74	85	13	225
Total	72	124	205	49	450

Table 3: Contingency table for question 2

The expected values we will use to calculate the χ^2 are the probabilities of choosing each value supposing that there is the same chances of choosing an answer with or without information.

Expected values for our answers:

$$E_{i,j} = \frac{T_i T_j}{T}$$

where T is the big total, T_i is the total for the row i , which corresponds to with and without information, and T_j is the total for the column j , which correspond to the answers.

Table of expected values for question 2:

	93700	150300	312100	480200	TOTAL
With information	36	62	102.5	24.5	225
Without information	36	62	102.5	24.5	225
Total	72	124	205	49	450

Table 4: Table of expected values for question 2

The conditions needed to apply the χ^2 test are:

- The data is multinomial data in a contingency table or two way cross-classification table.
- All expected values are at least 5 .
- The cells have counts or frequencies, not percentages or relative frequencies.
- The data comes from one population.

Since our data fulfill all the conditions we can apply the test.

To calculate the statistic parameter χ^2 we will use the formula:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

being:

m, n the dimensions of the table, in this case $m=2$ and $n=1$

$O_{i,j}$ = the frequency observed

$E_{i,j}$ = the theoretical frequency calculated according to the distribution of the null hypothesis.

Then, applying it to the 2 contingency tables we get:

Value χ^2 for question 1 (table 1): 8.523296659

Value χ^2 for question 2 (table 3): 37.47224497

We can compute the degree of freedom by doing the following operation:

$$df = (m - 1)(n - 1) = 3$$

We will look for the value for 0.05 because we consider that if our value is bigger than the value for the 5% we can discard the null hypothesis, which leads to confirm that our

hypothesis of dependency is true. Now, looking at figure 7 the critical values of the χ^2 distribution we can find the value for 0.05:

Table 5-2
Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Figure 7: Critical values of the χ^2 distribution, table extracted from *Introducción a la Estadística* [2]

The value is 7.815. Hence, the hypothesis of independence is true for the tables with values lower than 7.815, while we know that in the ones that have greater value there is a relation between the variables.

We can see that the values in both tables have a parameter bigger than 7.815. This means that there is a relation between the information that was given in the question and what did the people respond. Looking our data in figure 5 we can see that Martina is much more chosen with information than without, just the opposite to Maria. Also, in figure 6 the most popular answers without information are 150300 and 312100, while with information the two highest answers increase considerably their popularity, leading to the lower values to be much less chosen.

3.3 Gender, age and region aspects

3.3.1 Gender

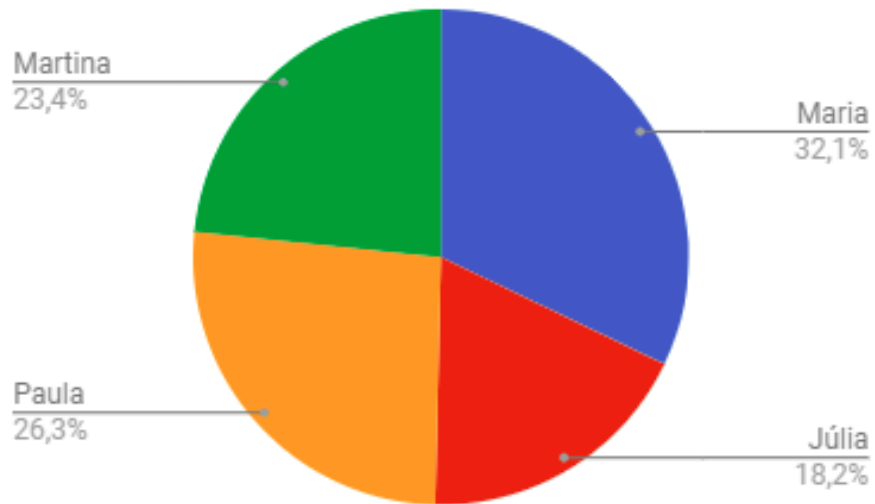


Figure 8: Answers to question 1 by men

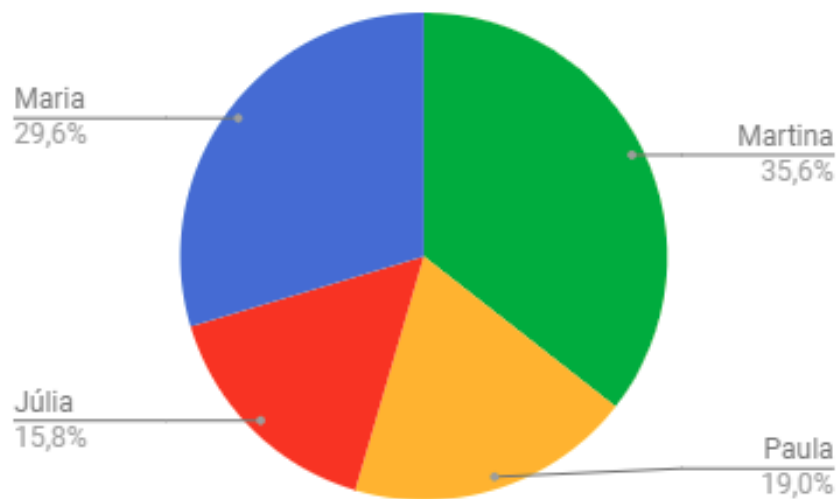


Figure 9: Answers in question 1 by woman

This shows that clearly many more women (36.6%) than men (23.4%) choose *Martina*, while the most popular answer by men is *Maria*. Since *Martina* is the correct answer, this

difference can be given by the fact that it is asking for the most popular girl's name, which can mean the female part of the asked population having more information. However, this does not affect our comparison between the answers to the questions with and without information since both of them have been answered by around 65% of women and 35% of men.

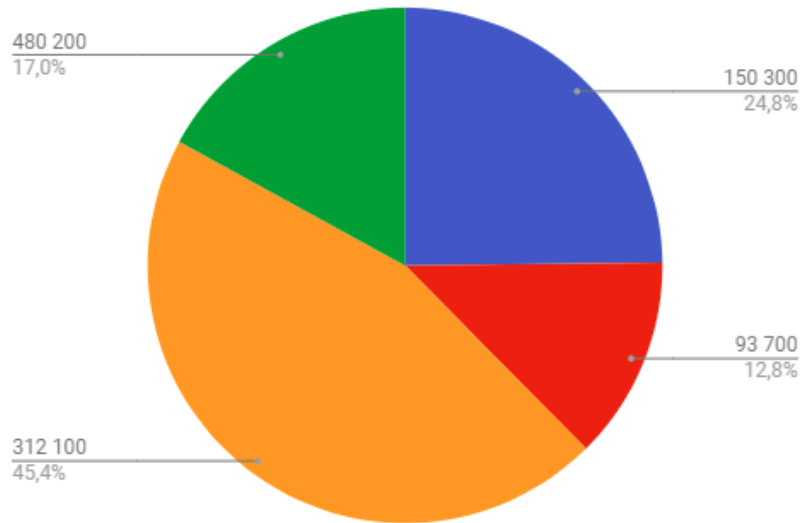


Figure 10: Answers to question 2 by men

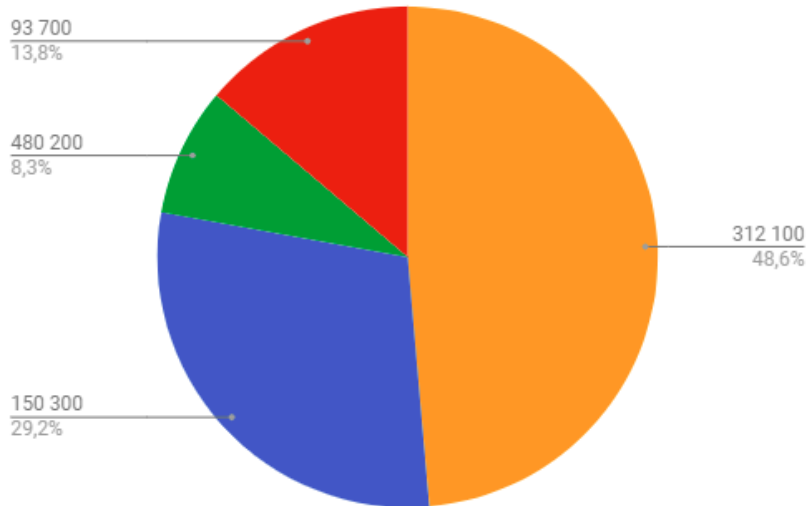


Figure 11: Answers in question 2 by woman

The main difference between the answers of women and men is that only 8.3% of women answered 480200, while 17.0% of the male sample choose it. That could mean that men are more optimistic in front of the nationality achieving situation.

3.3.2 Age

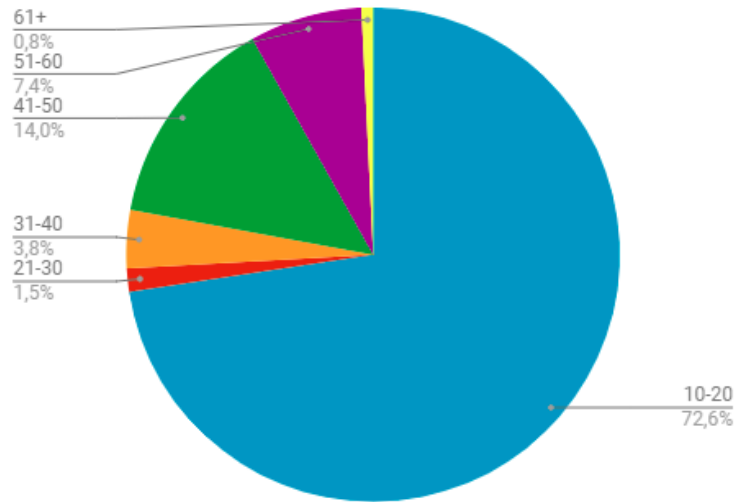


Figure 12: Pie chart for the question: *How old are you?*

Looking at this pie-chart we can see that most of the people that have answered our poll are between 10 and 20 years old. For that reason we have decided not to compare between ages, since it would not have been representative.

3.3.3 Region

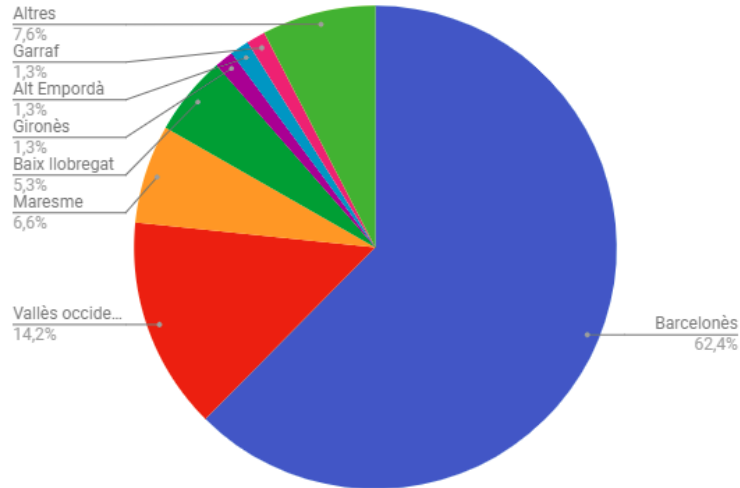


Figure 13: Pie chart for the question: *What region of Catalonia are you from?*

We have seen that most of the people we have asked are from *El Barcelonès* 15 and *El Vallès Occidental* 14. For that reason we have decided to only compare those two regions.

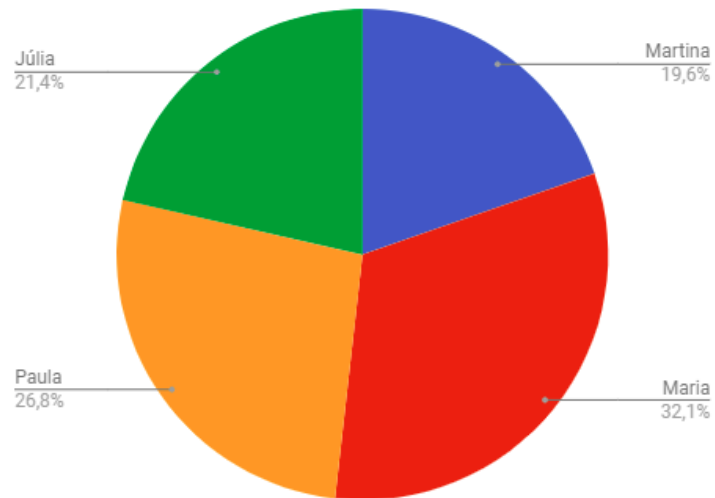


Figure 14: Answers to question 1 by the people from *El Vallès Occidental*

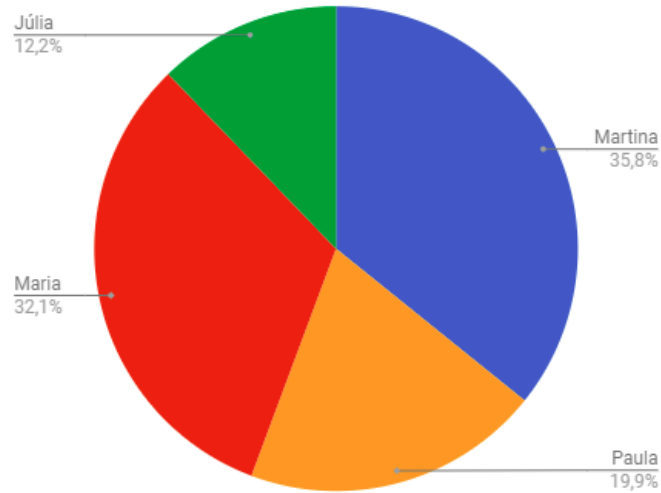


Figure 15: Answers to question 1 by the people from *El Barcelonès*

We can see that *Martina* is the least chosen answer in 14 (19.6%), while it is the most popular answer (35.8%) in 15. Anyway, if we look at the proportion of regions depending of the model of questions we can see that this should not affect the results because have nearly the same percentages.

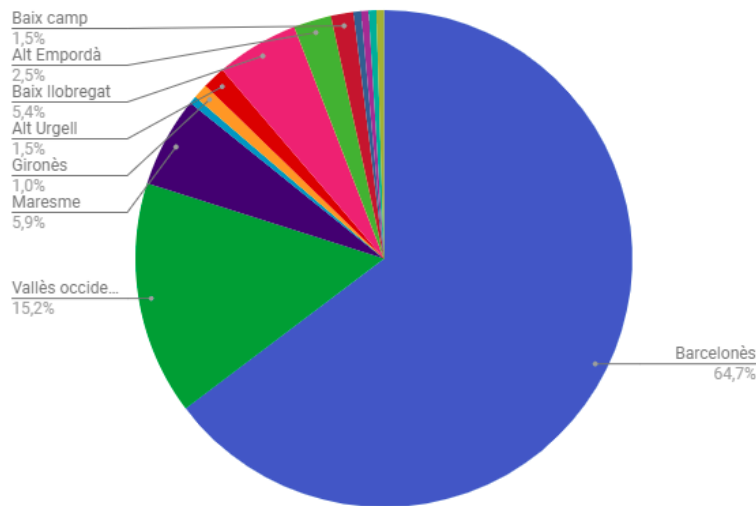


Figure 16: Answer to the question: *What region of Catalonia are you from?* in the first survey

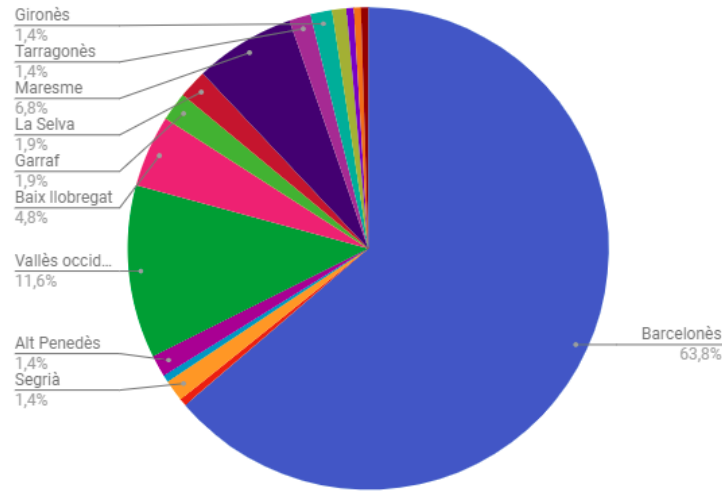


Figure 17: Answer to the question: *What region of Catalonia are you from?* in the second survey

4 Conclusion

After having observed, analyzed and studied our results through a χ^2 Test 3.2, we can conclude there is a direct dependence between the information given in the question with its correspondent answer.

Looking at our survey, we have seen that in the question *What do you think was the most common girl's name among the newly born in 2013 in Catalonia?* the most common answer is *Maria*, whereas when we add the extra information to the question: *Knowing that the most common name among the new born of 2012 was Júlia...* it significantly increases the percentage of *Martina* responses. We believe this is because the immediate response is *Maria*, but the new information given allows us to think more on this first impulse and as a consequence changing the answer to another name.

In the second question: *How many people do you think they got Spanish citizenship in 2014?* We have observed that the two average values, 312100 and 150300, are usually used to respond with a similar proportion when there is no information, but when they take the information: *Knowing that in 2012 there were 115500 people who obtained Spanish nationality and in 2013 they were 261300* the person establishes a logical relation of mistaken growth, responding much more to the two options superiors to 261300, which means an important descent of the popularity of the two answers below that value.

To sum up, we can conclude that people do change their answers depending on the previous information they have. Thanks to the fact that we have done a statistics survey, we can clearly show how without information the percentage of all the answers is similar. Whereas when we provide information to the subjects, the answers given tend to a certain conclusion, just as we thought in the introduction. Moreover, this shows how people can be influenced very easily to make them think in a certain way and how just by giving a specific information you can make someone change their initial belief into your answer of interest. This leads to think that with the correct previous information you can make people think what you want, even without them knowing it and thinking it is their own beliefs. The information we gave was real, but it was just guiding the subjects to the wrong answer. What would happen if the information given was not even true? Would it be also as easier as this time to control people's answers? We think that manipulating people's answers happens every day, we are given a specific information that leads to take a determined answer. Just by showing part of the truth or even by lying we are being controlled to think in certain way, to say certain things and act in a certain manner. However, we have also seen that there are much more things that influence on the answers, which explain why not all the answers have the same popularity on the survey without information, so it is impossible to predict anyone's choice, even knowing the previous information that they are given.

5 Annex

CONEIXEMENT GENERAL

Plantar d'estadística 1

*Necessari

Gènere *

Femení

Masculí

Altres

Edat *

10-20

21-30

31-40

41-50

51-60

61+

A quina comarca viu? *

La vostra resposta

Sabent que el nom més comú entre les recent nascudes del 2012 va ser Júlia, quin creu que va ser el més comú el 2013 a Catalunya? *

Maria

Paula

Júlia

Martina

Quantes persones creu que van obtenir la nacionalitat espanyola el 2014? *

93 700

150 300

312 100

480 200

Figure 18: Questions from first survey

CONEIXEMENT GENERAL

Plantel·ler d'estadística 1

* Necessari

Gènere *

- Femení
- Masculí
- Altres

Edat *

- 10-20
- 21-30
- 31-40
- 41-50
- 51-60
- 61+

A quina comarca viu? *

La vostra resposta

Quin creu que va ser el nom de noia més comú entre les recent nascudes el 2013 a Catalunya? *

- Maria
- Paula
- Júlia
- Martina

Sabent que el 2012 va haver 115500 persones que van obtenir la nacionalitat espanyola i el 2013 van ser 261300, quantes creu que van ser el 2014? *

- 93 700
- 150 300
- 312 100
- 480 200

Figure 19: Questions from second survey

Bibliography

[1] Idescat enquesta nadons, <http://www.idescat.cat/nadons/?lang=es&t=2013>

[2] Introducción a la Estadística, Sheldon M. Ross, 2007, Reverté