



Census at School Ben-Nyc

Planter_070-Maig 2022



*Centre: IES Salvador Espriu
Tutora: Lucia Bayo
Autors: Noa Vilaboa, Hexi Wei,
Yujie Bai, Kai Blacha, Laura Miró*

TABLE OF CONTENTS

ABSTRACT	2
INTRODUCTION	2
OBJECTIVES	3
GENERAL METHODOLOGY	3
(a) Data collection	3
(b) Statistical work	4
Formulation of hypotheses	4
Cleaning of data	4
Study of samples	4
Study of populations	4
Conclusions	4
HYPOTHESES	4
STATISTICAL WORK	5
1. Physical characteristics:	5
1.1. Right foot length	5
1.2. Right foot length and arm span	6
2. Cultural characteristics	9
2.1. Famous, rich,...	9
2.2. Looking up to someone	11
2.3. Importance of having access to the Internet	13
2.4. Importance of having a computer	15
2.5. Importance of having a computer vs having access to the Internet	16
2.6. Possible correlation between importance of having a computer and importance of having access to the Internet	17
2.7. Highest level of education to be achieved	19
2.8. Memory test time	21
2.9. Possible correlation between memory test time and time for homework	23
2.10. Time for homework for students who plan to have a graduate degree	25
GENERAL CONCLUSIONS	26
WEBLIOGRAPHY	26

ABSTRACT

In this statistical project we have tested specific hypotheses on answers given to some of the questions of a questionnaire, *Census at School*, hosted by an American website maintained by The American Statistical Association and Population Association of America : <https://ww2.amstat.org/censusatschool/students.cfm>. Our two samples contained a total of responses from more than 280 students aged 15 - 17 from New York and Barcelona.

Our general objective was to know if the populations of students in NY and BCN could be considered as the same population respect to the physical and cultural characteristics studied.

Among other things that we have discovered throughout this work, we have learnt that both populations intend to reach the same level of education, that the length of the right foot of BCN students is larger than that of NY, or also that the students from NY and BCN want to be equally famous.

The conclusion to our main objective is that these two populations are different, although they are not completely different, as they coincide in some aspects.

INTRODUCTION

This project comes from a questionnaire called Census at School (<https://ww2.amstat.org/censusatschool/students.cfm>), which allows students in grades 4-12 from all around the world to respond to the same questions. Then the students analyze their class census results, and may compare their class with random samples of students in the United States and other countries. The questionnaire is formed by 40 brief questions, for example, the length of your foot or your height.... Also, there are questions which include doing a little test to find out your memory or your reaction time.

Census at School began in the United Kingdom in 2000 to promote statistical literacy in schoolchildren by using their own real data. The program is operative in the UK, New Zealand, Australia, Canada, South Africa, Ireland, Japan, and the United States. The U.S. component of Census at School is hosted by the American Statistical Association's Education Outreach Program and cosponsored by partner Population Association of America.

This project wants to see if Barcelona's and New York's populations aged 15-17 are similar and if they could be mistaken for one another. The project will only use data collected from the Census at School questionnaire students responses from both cities. Students from Barcelona answered a copy of the questionnaire, but our data could not be uploaded or shared.

This project is the sum of three of the best projects written by all the Math students of 4th ESO from Institut Salvador Espriu, Barcelona. In pairs, students were assigned several questions of the questionnaire in order to practice the concepts learned along the unit of Statistics.

OBJECTIVES

Our objectives for this project are:

1. Practice all we have learned during the 4th ESO Math statistics unit.
2. Expand our knowledge
3. Improve our English level
4. Prepare us for TR (Trebball de Recerca de Batxillerat)
5. Find out if the American sample (New York) and the Salvador Espriu sample (Barcelona) could be considered similar or not, that's it, if the two of them could belong to the same population.

GENERAL METHODOLOGY

4th ESO Math classes in Salvador Espriu worked with almost all the questions in the questionnaire, as a project for the subject. We students, in groups, chose some of the questions to analyze the answers in depth and compare our results with New York students, which we thought could have some similarities with us. **The guessing was whether we could consider both samples to belong to the same population for some of the characteristics.**

In order to participate in the "Planter" our teacher asked some of the students with the best projects to put our projects together and work a bit more. Here you have the common memory we have written, **which does not cover all the questions in the questionnaire, but it still looks for possible similarities between the populations from which both samples come.**

All of us have worked in the same way:

(a) Data collection

All the data collected come from the answers to the **questionnaire** on the website "[Census at School](#)"

We have used **two samples**:

- on the one hand students of 1r de Batxillerat (Abril 2021) and students of 4th of ESO (December 2021) from Institut Salvador Espriu, Barcelona
Sample size: 77
- on the other hand, all the students from NY from grades 10 and 11 who answered the questionnaire in 2021 and in 2020. (We do not know the schools they belong to)
Sample size: 122

Sampling: convenience, because we have taken all the available data on the website and in our school, without having the possibility of taking random data from **both populations (students aged 15-17 from BCN and NY)**

(b) Statistical work

Formulation of hypotheses

Taking questions as our referents, first of all we wrote hypotheses to work with.

Each group worked with at least three hypotheses:

- (a) one with a qualitative variable involved, to work with proportions,
- (b) one with a quantitative variable involved, to work with means and deviations,
- (c) a third one with two quantitative variables involved, to work with possible correlations.

In this report we have several hypotheses of each kind.

Cleaning of data

Before making graphs and calculations, we started cleaning the data by searching and eliminating outliers outside the domain. For example, a length of 5 cm for a right foot or a preference of -300 in a range from 0 to 1000.

Study of samples

Next, we studied the data obtained from the samples, making box plots, scatter plots, other graphs, and calculations as means, standard deviations, etc. We have used mainly GeoGebra and Google spreadsheets.

Study of populations

Except for the hypotheses where we looked for possible correlations, we used online statistical calculators to obtain confidence intervals in order to compare the results and to make decisions. Depending on the variable we were working with, qualitative or quantitative, we used either a statistical calculator for proportions or a statistical calculator for means.

Conclusions

Lastly, except for the study of possible correlations, we arrived at conclusions by using the confidence intervals calculated. **In case the confidence intervals of both samples overlapped we considered them to be samples that could belong to the same population, that is, BCN and NY students could be confused.**

Our teacher explained to us that this was not exactly what we should have done, but it could be taken as a rough introduction to inferential statistical work.

HYPOTHESES

Our **general hypothesis** for this project is that we can consider the population of BCN and the population of NYC (students aged 15-17) to be the same for the characteristics studied, or, in other words, both samples could be considered a part of the same population.

The **hypotheses of each particular case** are written at the beginning of each hypothesis section and subsequently contrasted with all the relevant calculations.

STATISTICAL WORK

1. Physical characteristics:

1.1. Right foot length

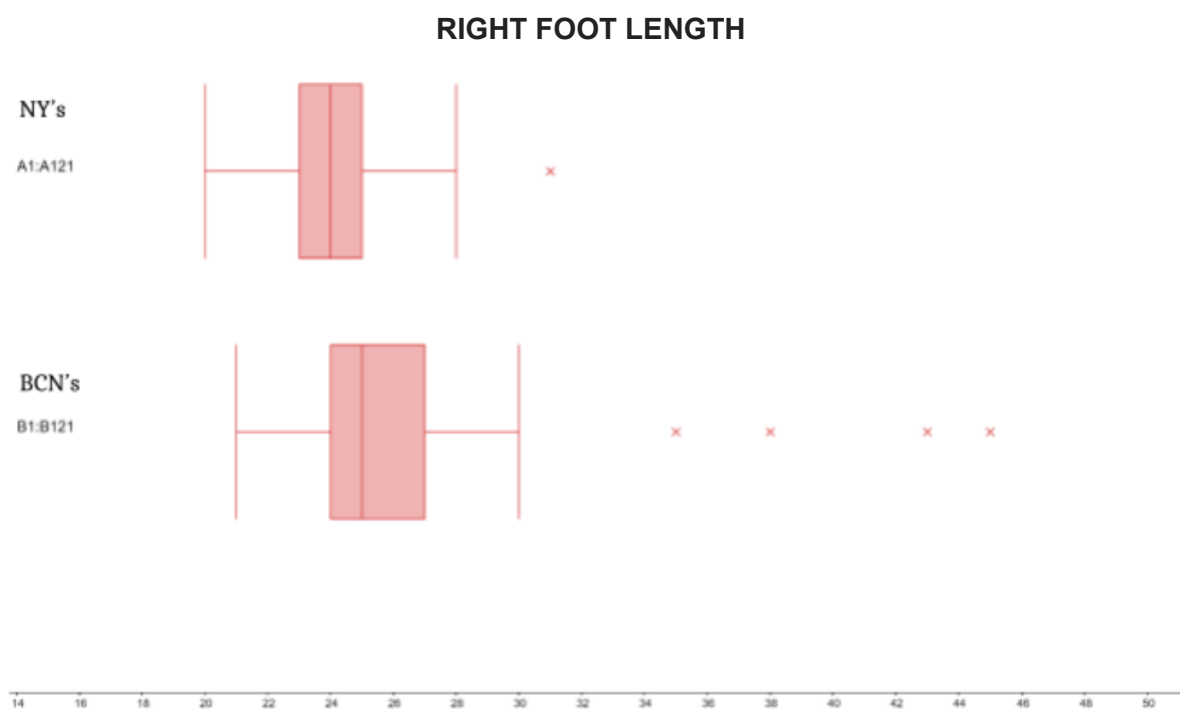
Question: 5. What is the length of your right foot (without your shoe on)? Answer to the nearest centimeter. - Right foot length (cm)

Variable: **Quantitative**

Statistical value: **Mean**

HYPOTHESIS: **Mostly BCN's people have a bigger right foot than NY's people.**

Firstly, we organized the data we received from the questionnaire with the variable that we are interested in (foot length), and discarded the outliers (those answers that did not make sense, e.g. the foot length is 3000 cm). Then we took the data we had left, and we put them in a geogebra table, and we placed it in a box plot. Geogebra shows us some outliers, but we decided not to remove them, since they are coherent and inside the domain.



In this box plot we can observe that most of the BCN box is located more to the right (bigger) than the NY box.

We calculated the mean for each sample, and then with the function `stdev`, we calculated the standard deviation.

As the box plot and the standard deviation show, the BCN's data have bigger dispersion.

Next, we calculated the confidence interval for each city. We expect the population value to be around the sample value, inside the confidence interval, with a confidence level of 95%.

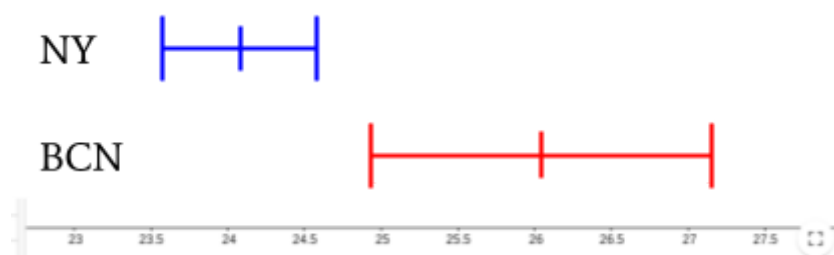
NY's confidence interval:

What is your sample mean?	<input type="text" value="24,079"/>	
What is your sample standard deviation?	<input type="text" value="2,1349"/>	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="71"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		(23.58 , 24.58)

BCN's confidence interval:

What is your sample mean?	<input type="text" value="26,049"/>	
What is your sample standard deviation?	<input type="text" value="4,6584"/>	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="67"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		(24.93 , 27.16)

And we need to compare both confidence intervals to see if they overlap or not.

**Conclusion:**

The two sample confidence intervals are separated and do not overlap, which means that both populations would be different, and our hypothesis would be right (mostly BCN's people have a bigger right foot than NY's people)

So, in Barcelona the foot length mean is expected to be between 24.93 - 27.16 cm and NY's foot length mean is expected to be between 23.58 - 24.58 cm. The mean for Barcelona's foot length is bigger than NY's.

1.2. Right foot length and arm span

Questions:

5. What is the length of your right foot (without your shoe on)? Answer to the nearest centimeter. - Right foot length (cm)

6. What is your arm span? (Open arms wide and measure distance across your back from tip of right hand middle finger to tip of left hand middle finger.) Answer to the nearest centimeter. - Arm span (cm)

Variables: **Quantitative**

Statistical value: **r and R² for different models**

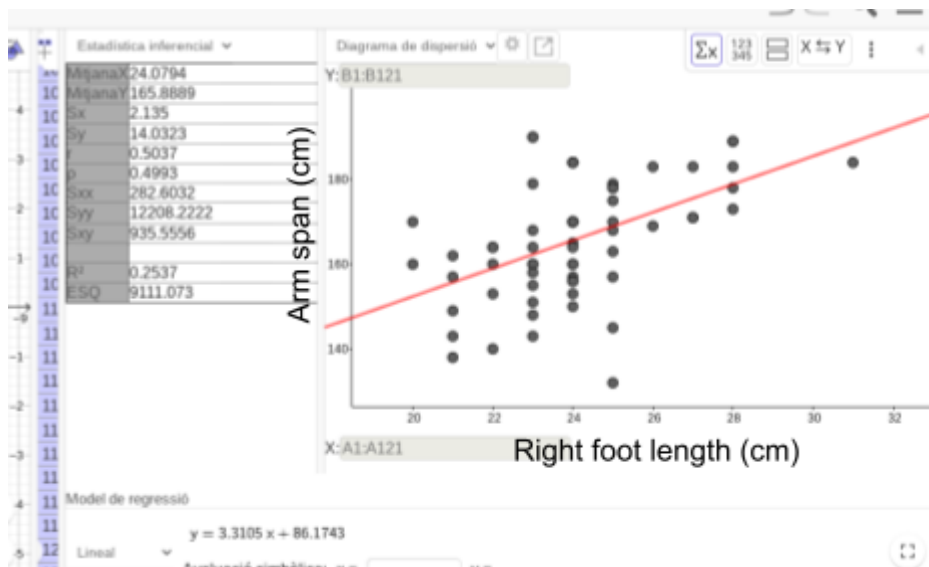
HYPOTHESIS: There exists correlation between the arm span and the length of the right foot, without differences between BCN and NY.

We took the cleaned data for right foot length (section 1.1.) of the New Yorkers and the BCN students, and added the response of the same students on the second quantitative variable (arm span). With Geogebra we made a scatter plot for each city to look for possible correlations between their arm span and the length of their right foot. We tried with different models for the possible relationship between variables.

KEY CONCEPTS:

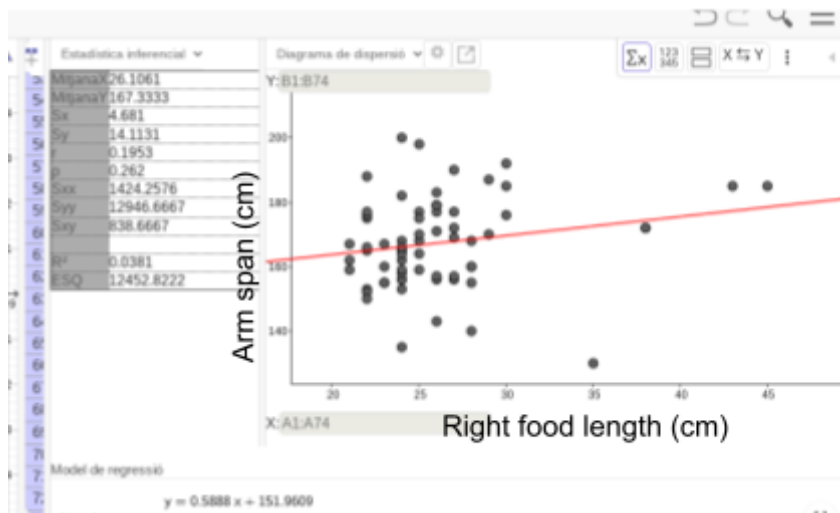
- r (lower case) = linear correlation coefficient of Pearson, only accountable for a possible straight line model, the line of best fit. The nearer to 1 or to -1, the better the prediction.
- R^2 = variation coefficient, the square of r if we try a linear model. R^2 changes with each possible model, linear or curved. It would be the amount of y that is explained with x : $R^2=0.7$ would mean that 70% of y could be explained by x .

NY linear model



$r = 0.5037$
 (we only use r , the coefficient of Pearson, for linear models)
 $R^2 = 0.2537$

BCN linear model



$r = 0.1953$
 $R^2 = 0.0381$

The following table shows the results for the models we tried:

	Barcelona	New York
r, Pearson coefficient	0.1953	0.5037
R² (linear model)	0.0381	0.2537
Linear model	$y = 0,5888x + 151,9609$	$y = 3.3105x + 86.1743$
R² (log model)	0.037	0.2508
Log model	$y = 110.8795 + 17.3755 \ln(x)$	$y = 90.4343 + 80.6661 \ln(x)$
R² (polynomial model)	0.039	0.2541
Polynomial model	$y = 0.012x^2 - 0.1588x + 163.0654$	$y = 0.039x^2 + 1.3805x + 109.8801$

If we compare the results of the two cities by looking at them quantitatively, we can find several similarities. For example, **for the linear model**, they both have a positive slope, they also have similar behaviour, there is a lot of dispersion, and they have a similar r.

Conclusion:

As all the models (whether linear, polynomial....) give us a R² smaller than 0.5, and neither BCN's nor NY's r (coefficient of Pearson) is bigger than 0.7, there is no way that these two variables are correlated, no matter in BCN or in New York.

Even so, the R²s and the rs for BCN and for NY are very different. The NY's mean of the R² for all the models is more or less 0.25, meanwhile BCN's one is only 0.037, and the equations are very different from each other. So we can not consider them the same population.

2. Cultural characteristics

2.1. Famous, rich,...

Question: [38. Which would you prefer to be? Select one. \(rich, famous, happy, healthy...\)](#)

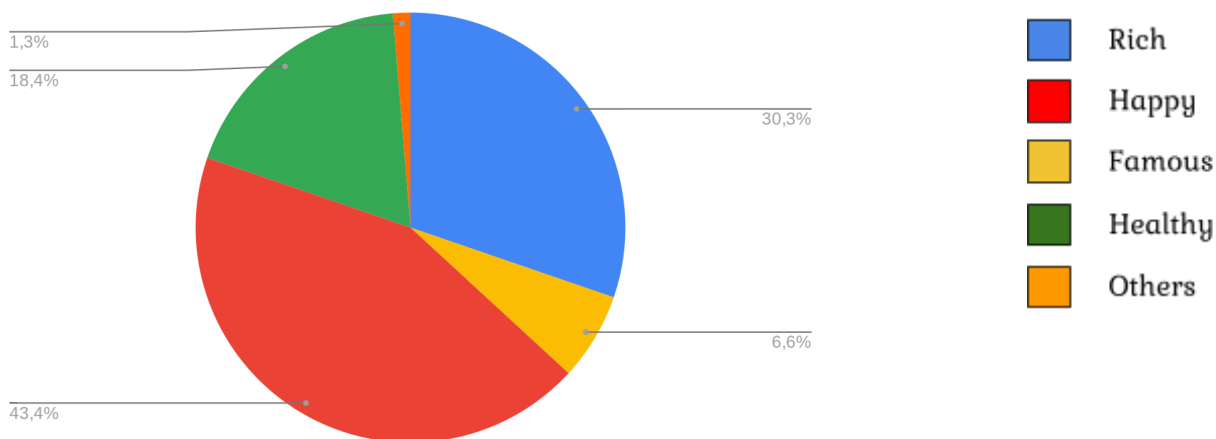
Variable: **Qualitative**

Statistical value: **Proportion**

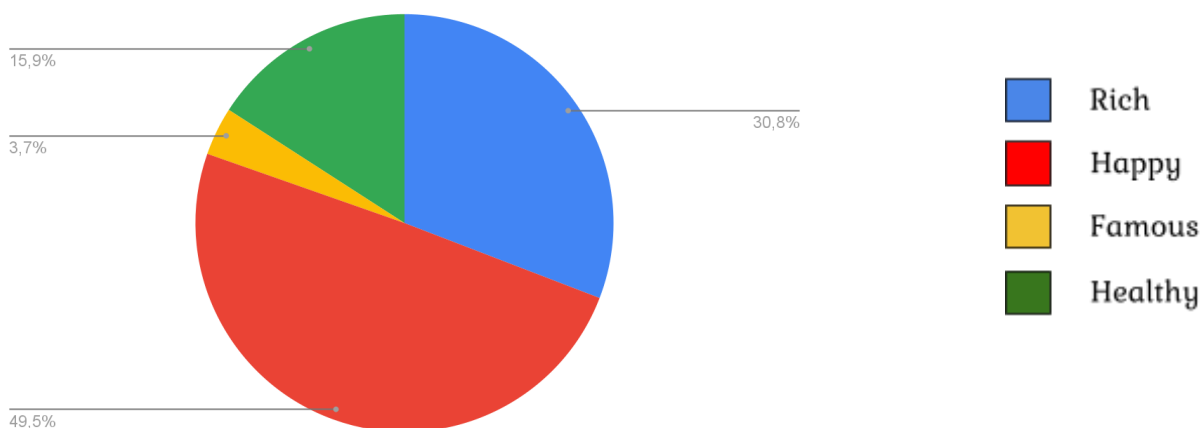
HYPOTHESIS: NY's students want to be famous, more than BCN's students

First, with the spreadsheet, we used the answers that people gave us to calculate the absolute frequency for each option, the number of times the students' choice was repeated (rich, happy...), After having counted the number of times these options were chosen by the students, the answers of each city were organized in a pie chart, so that we could clearly see the proportion of preferences of each city sample with a very visual comparison.

BCN Students








NY Students








As we can see in the two pie charts of the samples, the percentage of people who want to be rich in both is not the highest. But, in fact, our hypothesis deals only with the category "Famous".

Following our hypothesis, we were interested in the proportion of people who wanted to be famous in BCN and in NY (populations). We calculated the confidence interval of each city sample to see between what numbers we expect our population proportion to be. **We expect the population value to be around the sample value. We use samples to know about the populations**

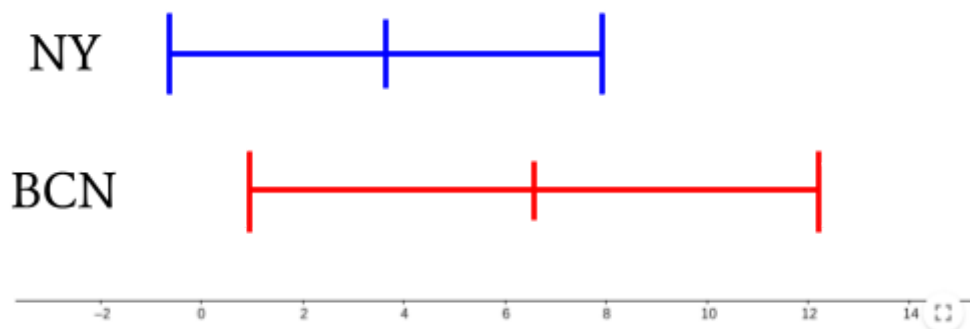
NY's confidence interval:

What is your sample proportion?	<input type="text" value="3,7"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="74"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		
		(-0.6 , 8)

BCN's confidence interval:

What is your sample proportion?	<input type="text" value="6,58"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="74"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		
		(0.93 , 12.23)

Later on, we need to compare both confidence intervals to see if the two populations are similar or not.



Conclusions:

The difference between the two confidence intervals isn't very big, they overlap each other in the most part of them, so we could consider that both populations could be the same one.

Our hypothesis (NY's students want to be famous, more than BCN's students) was wrong, and the opposite is not true either. So we concluded that they both have the same desire to be famous.

2.2. Looking up to someone

Question: 39. Think about someone you most look up to (Politician, Teacher, Business person...). This could be someone you know personally or have read about or seen on TV. From the following list, choose the category that best describes this person.

Variable: **Qualitative**

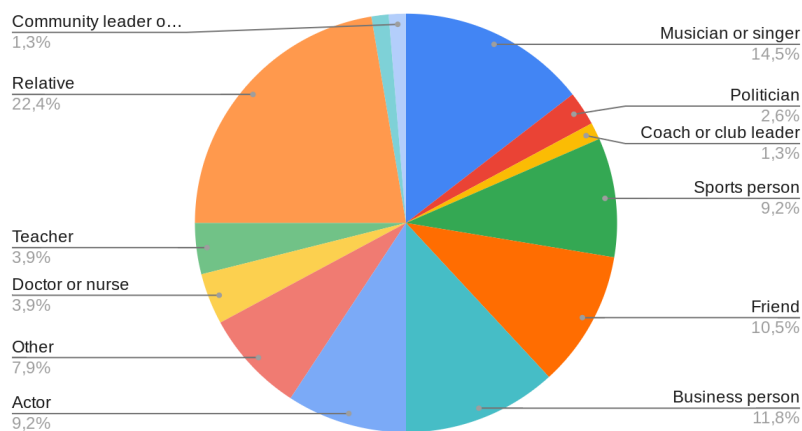
Statistical value: **Proportion**

HYPOTHESIS: More than 40% of people look up to a business person. In New York this percentage is higher.

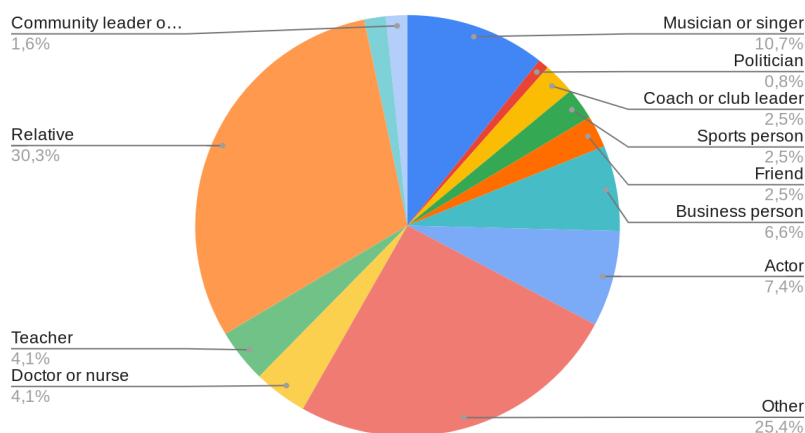
In this hypothesis, we had to calculate the percentages of the different people who students look up to. Especially the percentage of students who look up to a business person, which we thought would be more than 40% in BCN and in New York it would be higher. To test it, we made these calculations for both groups of students, the sample of Espriu students and the sample of New York students:

- Percentages: first, we calculated the percentages of people who look up to all the different people in the group. There were some people who wrote their own answers, and we classified them as *Other*. With this calculation, we did a graph for each sample.

Graph Salvador Espriu



Graph New York



- Confidence interval: these percentages are only for the samples. For having a better approximation for the populations we did the confidence interval of the percentage of students in the samples who look up to business persons. To do the intervals we used an online calculator. The results with a 95% confidence level were:

For **Espriu Students**: 4.55% - 19.05%

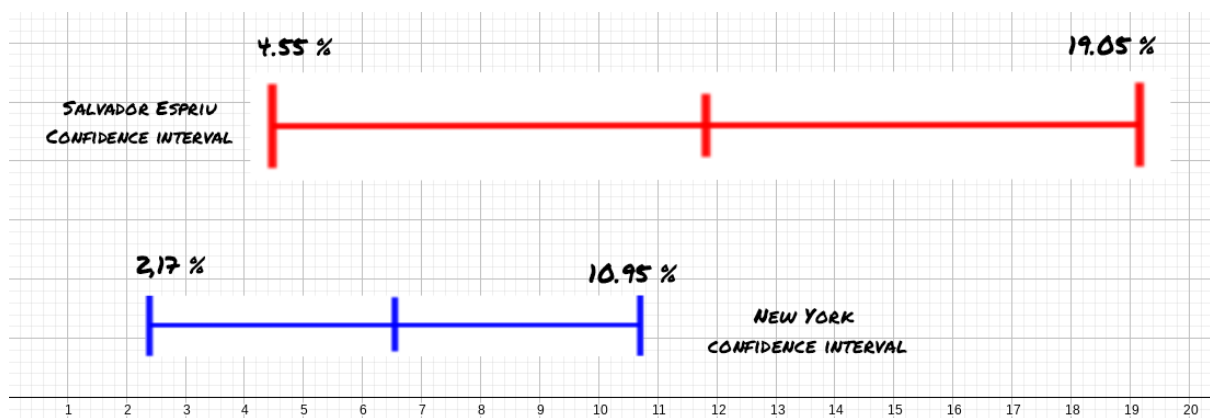
For **New York students**: 2.17% - 10.95%

Calculator

What is your sample proportion?	<input type="text" value="11,8"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="76"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		
		(4.55 , 19.05)

Calculator

What is your sample proportion?	<input type="text" value="6,56"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="122"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is		
		(2.17 , 10.95)



In **conclusion**, our hypothesis was wrong.

On the one hand, the percentage of people who look up to business persons is less than 40% in both groups.

On the other hand, it seems that the percentage is bigger for Espriu Students, between 4.55% and 19.05%, than for New York inhabitants, between 2.17% and 10.95%. Nevertheless, the samples are supported, since the confidence intervals do overlap, the two samples may belong to the same population.

(Both samples give great importance to relatives, more than to business persons.)

2.3. Importance of having access to the Internet

Question: 13 (f). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Access to the internet

Variable: **Quantitative**

Statistical value: **Mean**

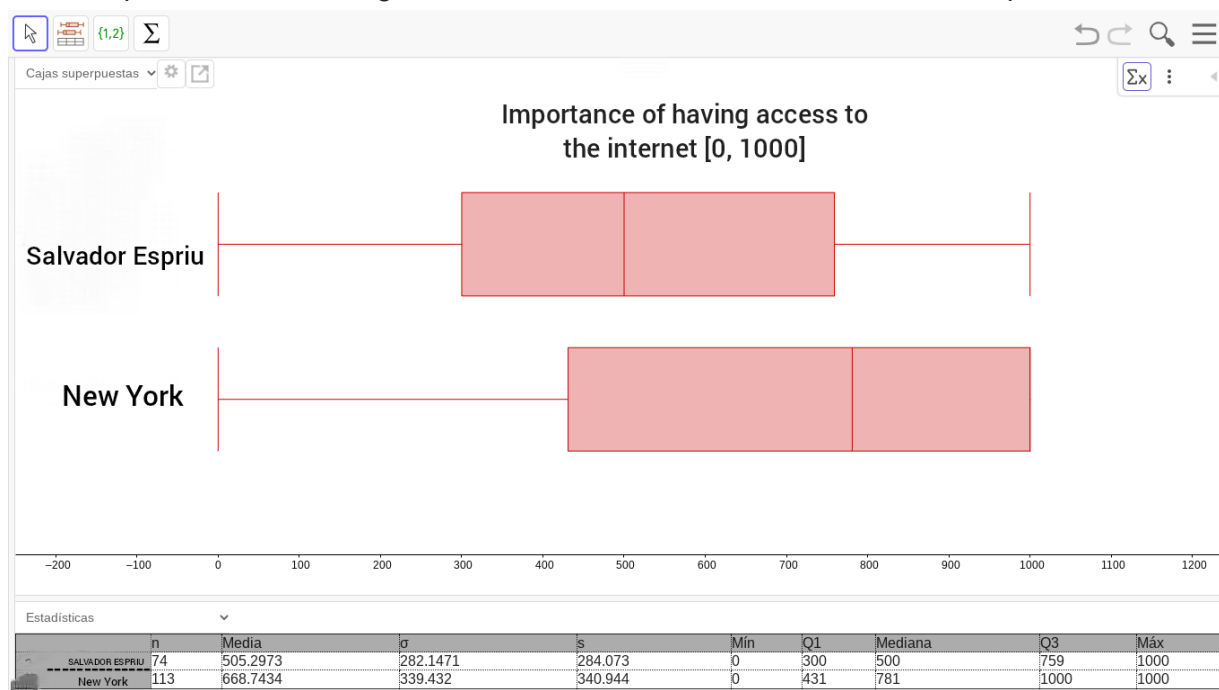
HYPOTHESIS: Students consider access to the internet important, in 800/1000. In New York too.

We had different calculations to do, but first of all we had to revise our samples in case there were values outside the domain. In this case, the domain is between 0 and 1000, because this is the interval they gave in the question. We found some values with an extra 0 as 5000; we replaced these values with the same value but without the extra 0. In the previous example, we replaced 5000 by 500.

Once we revised the data, we started with the calculations and graphs for the samples..

- Means: with google spreadsheets, we did easily the mean for each group. These were the results:
 - For Espriu students: 505/1000
 - For New York inhabitants: 669/1000
- Standard deviations: to later do the confidence intervals, we needed the standard deviation, which shows us how far are usually the samples from the mean. We did easily with the google spreadsheet, and these were the results:
 - For Espriu Students: 284
 - For New York inhabitants: 341

Box plot: with Geogebra we made a double box plot, which represents Espriu students and New York students in the same graph, to compare them better. We can see that the New York box plot is placed more to the right and that means that New Yorkers sample gives more importance to having access to the Internet than Salvador Espriu's students



- Confidence intervals: to obtain information about the populations, we obtained the confidence intervals of both samples. The results with a 95% confidence level were:

For BCN students (SE): (440.32 – 569.68)

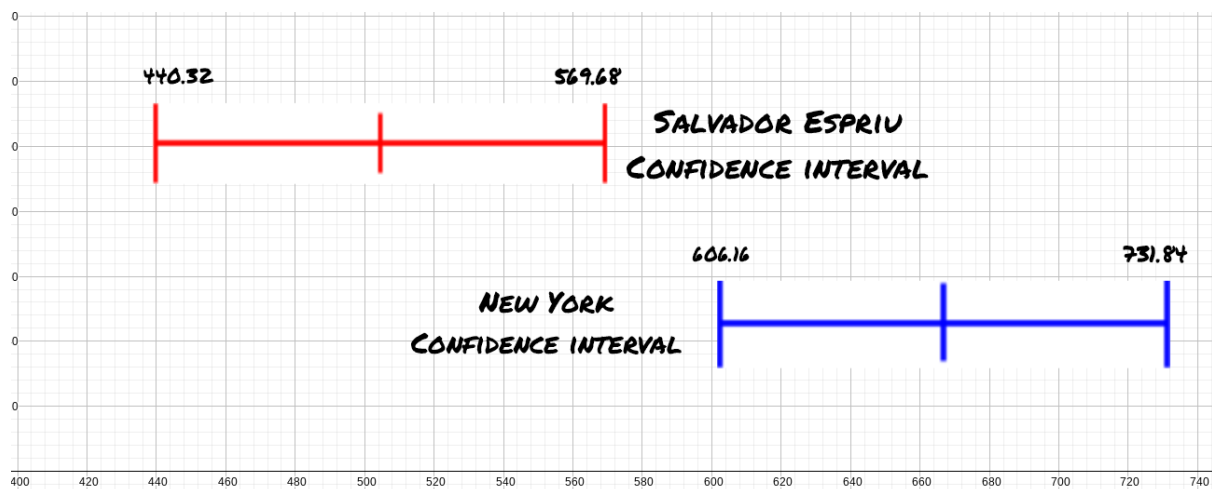
For NY students: (606.16 – 731.84)

Calculator

What is your sample mean?	<input type="text" value="505"/>	
What is your sample standard deviation?	<input type="text" value="284"/>	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="74"/>	
How big is the population?	<input type="text" value="10000"/>	
Your confidence interval is		
		(440.32 , 569.68)

Calculator

What is your sample mean?	<input type="text" value="669"/>	
What is your sample standard deviation?	<input type="text" value="341"/>	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="113"/>	
How big is the population?	<input type="text" value="10000"/>	
Your confidence interval is		
		(606.16 , 731.84)



In **conclusion**, our hypothesis was wrong. In New York, students consider it more important to have access to the internet, than Salvador Espriu students. New York students consider access to the internet important between (606.16 , 731.84), while Barcelona students do it between (440.32 , 569.68).

As both confidence intervals do not overlap, we conclude that both samples could not belong to the same population.

2.4. Importance of having a computer

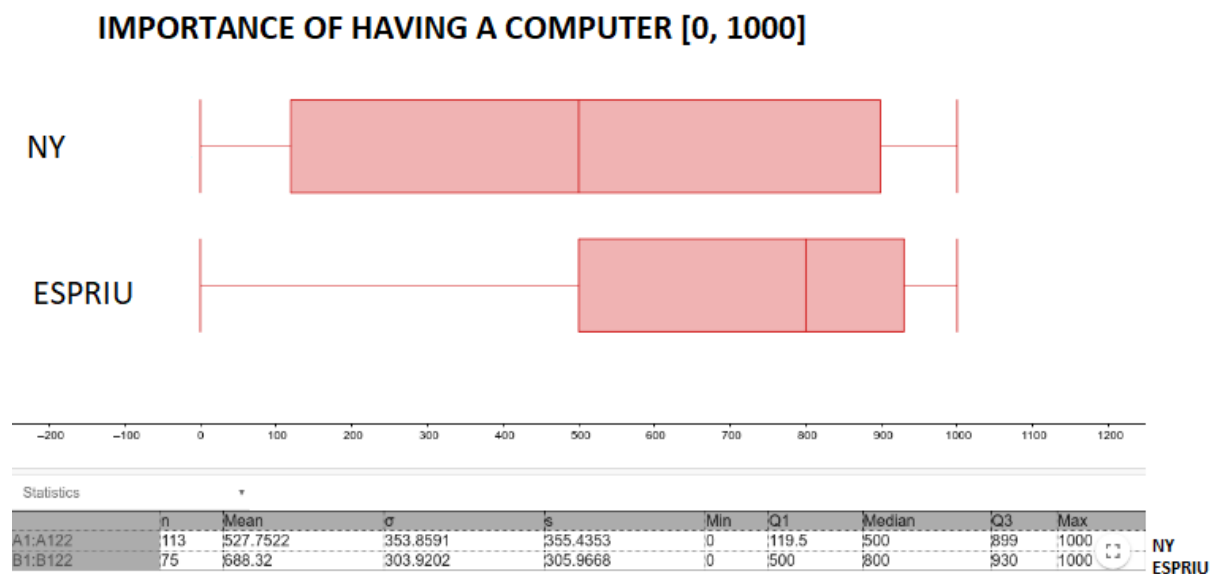
Question: 13 (e). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Owning a computer

Variable: **Quantitative**

Statistical value: **Mean**

HYPOTHESIS: Espriu students consider owning a computer in 700/1000. In New York too.

First, we revised our data to be sure all the data were in the domain. In this case, the domain is between 0 and 1000 as it was asked in the question. And once cleared the data, we put the data in Geogebra to make the calculations and the boxplot.



Down the boxplot there's relevant information as the mean or the standard deviation (indicated with the s) that we use to calculate the confidence intervals with a 95% confidence level.

**New York confidence interval:
(461, 592)**

**Espriu confidence interval:
(619, 765)**

Conclusion:

Espriu's students consider owning a computer more important than New York's students do. There is not overlapping between the intervals, so we can confirm that for this characteristic New York students and Espriu students do not form part of the same population.

Our hypothesis was right only for NY.

Calculator

What is your sample mean? ⓘ

What is your sample standard deviation? ⓘ

What confidence level do you need?
Typical choices are 90%, 95%, or 99% % ⓘ

How big is your sample? ⓘ

How big is the population? ⓘ

Your confidence interval is (461.95, 592.05) ⓘ

Calculator

What is your sample mean? ⓘ

What is your sample standard deviation? ⓘ

What confidence level do you need?
Typical choices are 90%, 95%, or 99% % ⓘ

How big is your sample? ⓘ

How big is the population? ⓘ

Your confidence interval is (619.45, 756.55) ⓘ

2.5. Importance of having a computer vs having access to the Internet

Questions: 13 (e). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Owning a computer

13 (f). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Access to internet

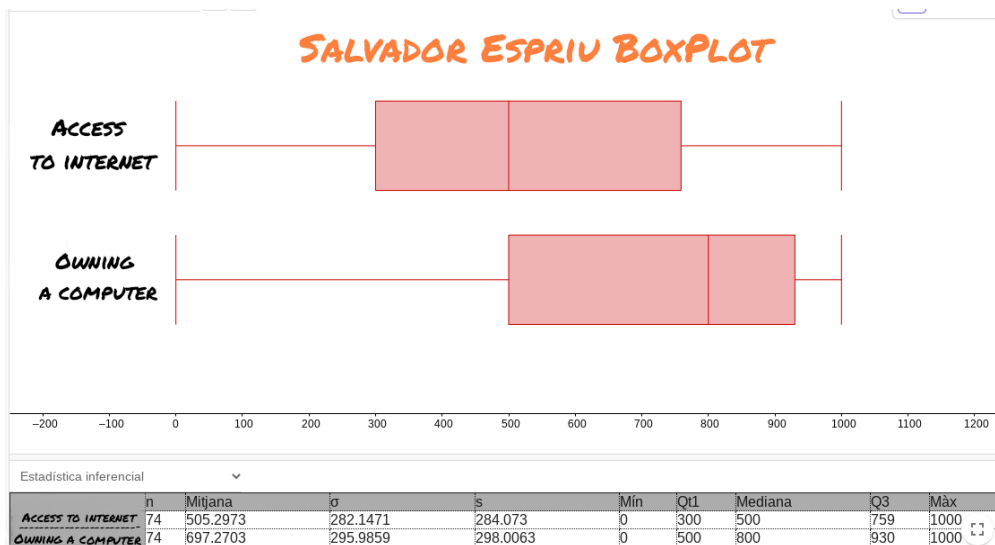
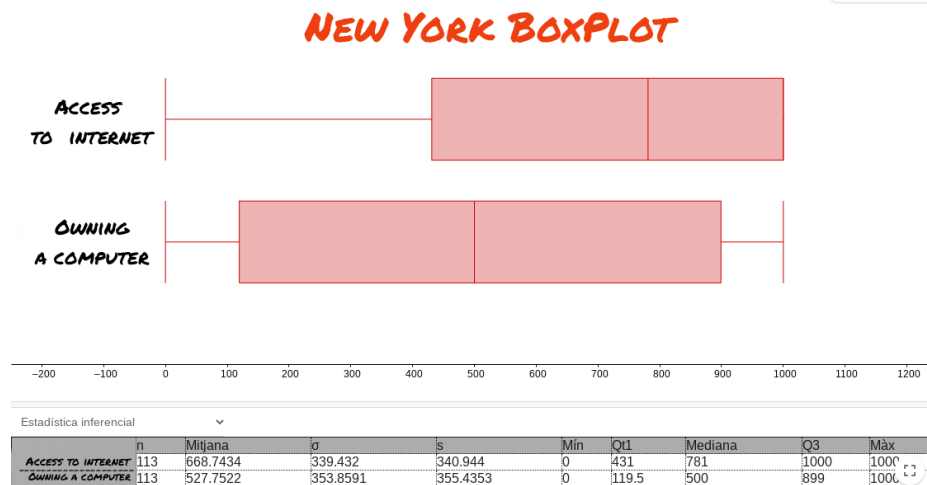
Variables: **Quantitative**

Statistical values: **Box plots (only samples)**

HYPOTHESIS: People of BCN's sample consider it more essential having access to the internet than having a computer. In New York too.

We hadn't to revise the data, because the samples we used in this hypothesis had been revised previously for other hypotheses, 2.3 and 2.4.

With Geogebra we made two double boxplots.



In **conclusion**, Espriu students sample consider it more essential to own a computer than having access to the internet, just the opposite of New York sample. We were partially wrong.

2.6. Possible correlation between importance of having a computer and importance of having access to the Internet

Questions: 13 (e). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Owning a computer

13 (f). How important is the following issue to you? Rate it on the scale from 0 (not important) to 1000 (very important). Access to the internet

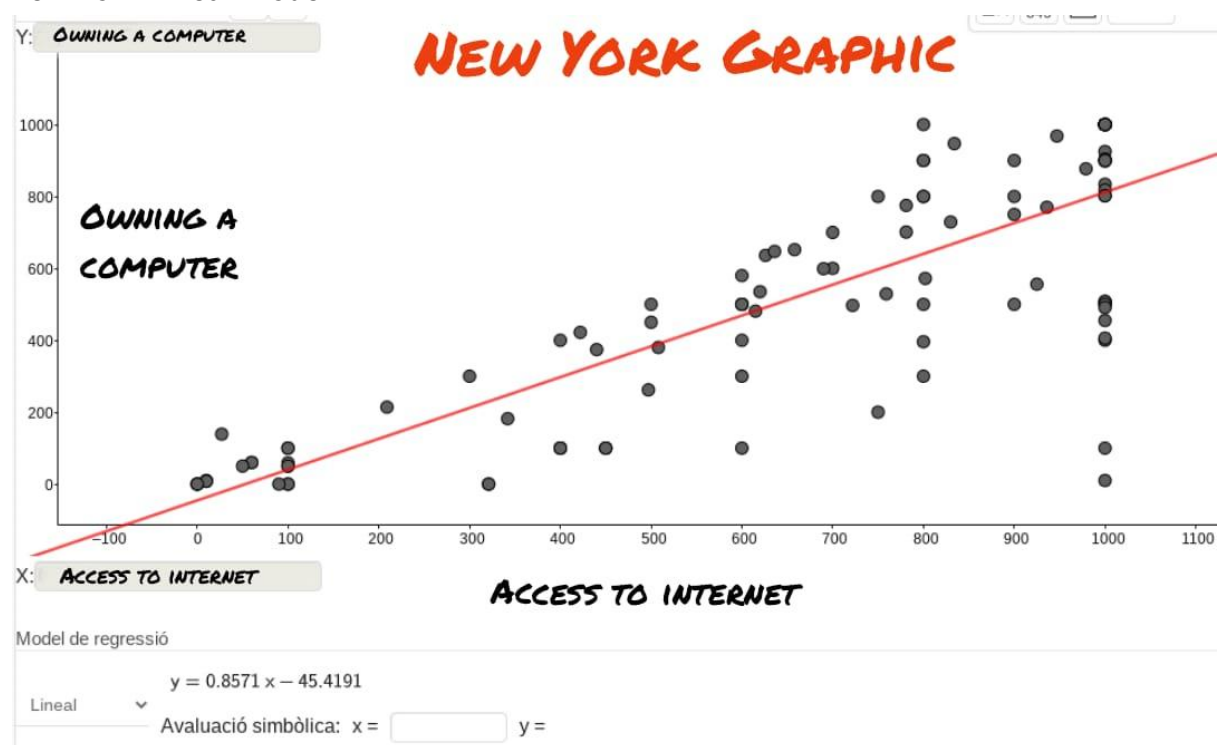
Variables: **Quantitative**

Statistical value: **r and R² for different models**

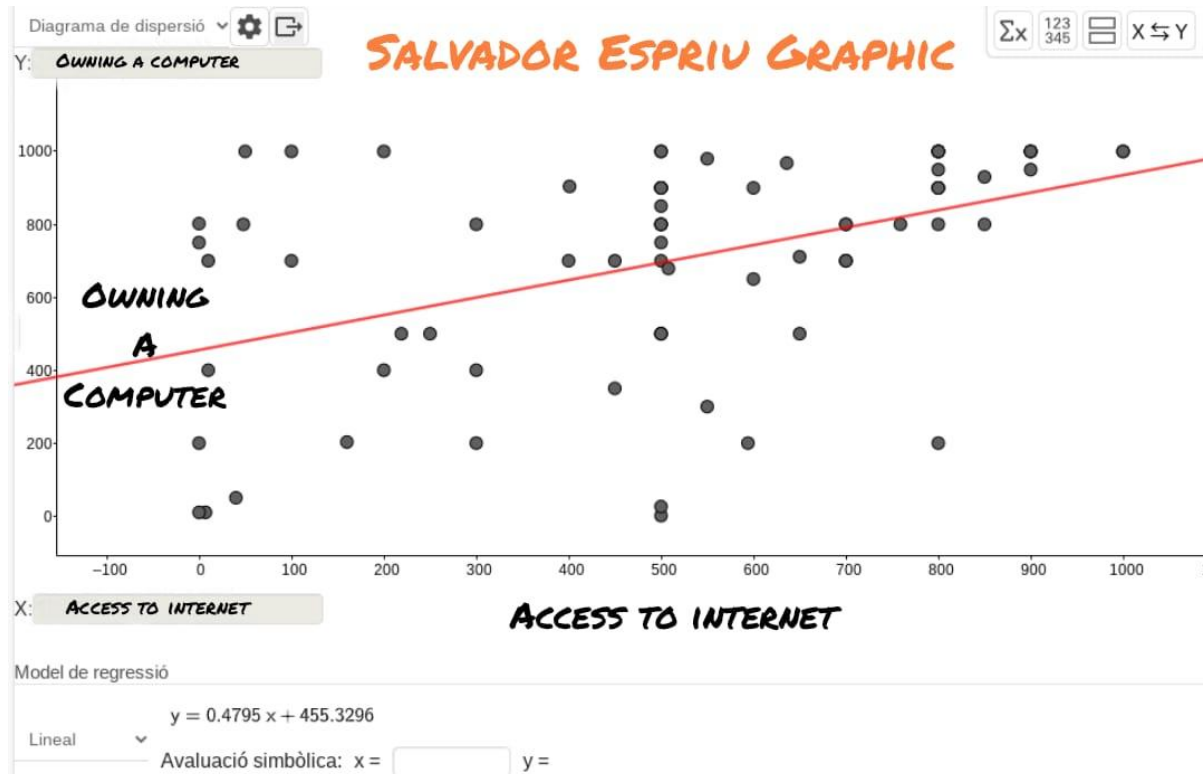
HYPOTHESIS: There is a weak correlation between the importance of having access to the internet and the importance of owning a computer. In New York, the correlation is stronger.

In this hypothesis, we wanted to see if there was a correlation between the importance that people give to having access to the internet and the importance people give to own a computer. We first checked that all the values were possible, but as we had worked with these data in the previous hypothesis, we had all well. To see the possible correlation and check possible models, we put all the data in Geogebra to make a scatter plot.

New York Linear model



Espriu Linear model:



	Espriu	New York
Pearson Coefficient	0.4583	0.8221
R ² Linear model	0.2101	0.6759
Linear model equation	$y = 0.4795x + 455.3296$	$y = 0.8571x - 45.4191$
R ² Polynomial model	0.2203	0.6777
Polynomial model equation	$y = 0.0004x^2 + 0.1435x + 501.6478$	$y = 0.0002x^2 + 0.6826x - 17.9872$

Conclusion:

There is no linear correlation or a very weak correlation for Salvador Espriu students, because the r is only 0,4583. For New York students there exists some linear correlation, because their Pearson coefficient is 0,8221.

The best model for New York data is the polynomial, which is slightly better than the linear one. However, the equation of the polynomial model is too difficult compared to the linear regression model, so it is more advisable to use a linear model.

The hypothesis would be essentially right and both samples would not belong to the same population.

2.7. Highest level of education to be achieved

Question: 35. What is the maximum level of education you plan to attain?

Variable: Qualitative

Statistical value: Proportion

HYPOTHESIS: Students in Barcelona plan to attain a lower level of education than students in New York.

Firstly, I cleaned the data and deleted any blank or randomly submitted data. For example, someone submitted a "-" as a response. In these responses, there were not any other outliers of this kind, so I just cleaned this one.

Graduate degree (màster, doctorat,...)
Graduate degree (màster, doctorat,...)
Graduate degree (màster, doctorat,...)
Some college (FP grau superior)
Undergraduate degree (grau)
-
Graduate degree (màster, doctorat,...)
High school (batxillerat, FP grau mig)
Graduate degree (màster, doctorat,...)
Some college (FP grau superior)

Then I created a table which collected all the times the variables were selected with their percentages and confidence intervals using a confidence level of a 95%:

New York:






Variables	Percentages	Confidence interval
Some college	17,6	(15.25 , 19.95)
Other	13,2	(11.11 , 15.29)
High school	5,5	(4.09 , 6.91)
Graduate degree	68,1	(65.23 , 70.97)
Less than high school	2,2	(1.3 , 3.1)
Undergraduate degree	10,9	(8.98 , 12.82)
Total	100	(100 , 100)

Barcelona:






Variables	Percentages	Confidence interval
Some college (FP grau superior)	5,3	(0.23 , 10.37)
Other	0	0
High school (batxillerat, FP grau mig)	9,3	(2.73 , 15.87)
Graduate degree (màster, doctorat,...)	61,3	(50.28 , 72.32)
Undergraduate degree (grau)	22,6	(13.14 , 32.06)
Less than high school	1,3	(-1.26 , 3.86)
Total	100	(100 , 100)

Both In Barcelona and New York, the most selected option was Graduate Degree, as we can check by looking at their relative frequencies or their confidence intervals, which do overlap (50.28, 72.32) and (58.97, 77.23). To calculate this confidence intervals, I used a Confidence Interval Calculator for a proportion:

Barcelona:**Calculator**

What is your sample proportion?	<input type="text" value="61,3"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="75"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is	(50.28 , 72.32)	

New York:**Calculator**

What is your sample proportion?	<input type="text" value="68,1"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="100"/>	
How big is the population?	<input type="text" value="100000"/>	
Your confidence interval is	(58.97 , 77.23)	

In conclusion, our hypothesis was wrong because confidence intervals overlap and this means that the difference between groups is not statistically significant. So they plan to attain a similar level of education.

2.8. Memory test time

Question: 11. Test your memory, how many pairs can you uncover? (Students had to complete a memory test in which they had to uncover a number of pairs of images and time themselves)

Variable: **Quantitative**

Statistical value: **Mean**

HYPOTHESIS: The mean of seconds spent doing the memory test in New York is higher than the one in Barcelona.

Before calculating the mean for both Barcelona and New York memory test results, I had to clean the outliers that were wrong data. In this case, they were numbers too small to even be possible to belong to the domain. For example, I had to discard the results of 0,50s and 0,58s because they were way outside the domain of (28, 57) in Barcelona and of (27, 118) in New York.

Outliers marked in red:

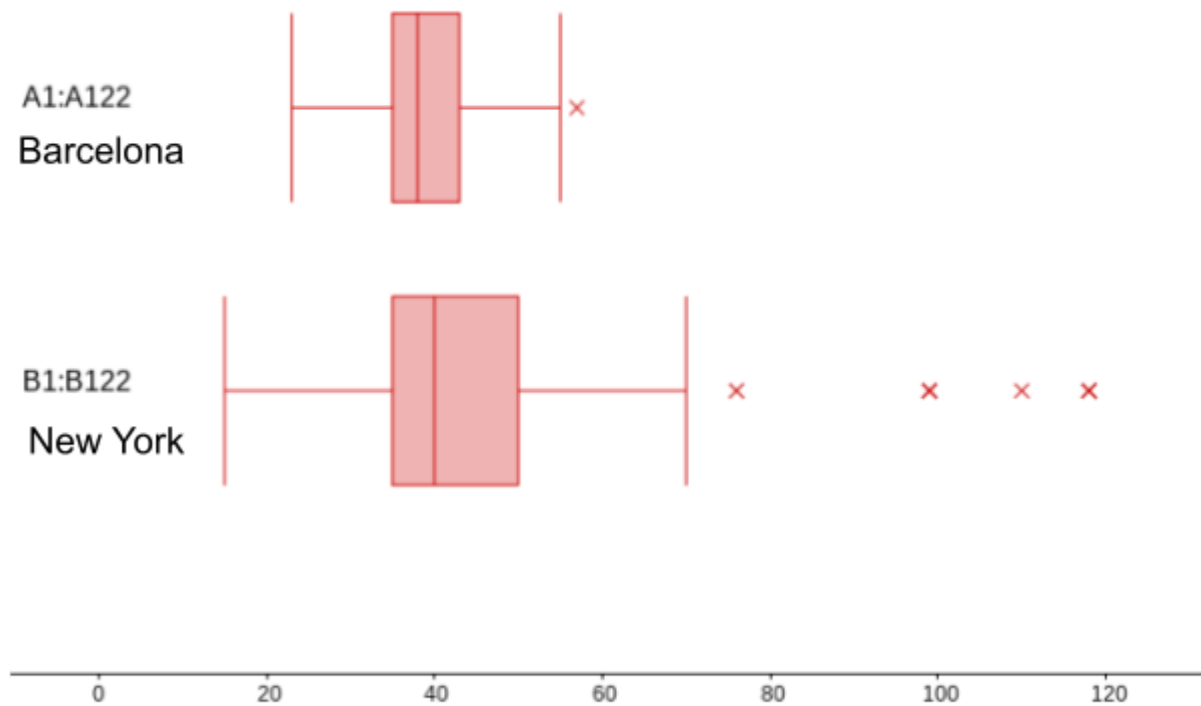
0.58 s	
65 s	
	50
	54
	37
	50
	36
	53
	55
	47
	30
	50
	37
	45
	29
	43
0.50	

Then, using Geogebra and its spreadsheet and box plot functions, I checked the median, mode, and mean in New York and Barcelona:

	Barcelona	New York
Mean	38,9	44,6
Median	38	40
Mode	37	40

There are also a few outliers in both box plots, but they should not be removed just because they are further than the average data on the plot with a domain of (27, 118) because they could be slow response times. From the box plots, we can also see that there's a wider range of responses in New York by looking at the whiskers.

Box plot graph of the memory test times—A:Barcelona, B:New York



I also calculated the **confidence intervals** of both samples which did not overlap:

(36.67, 41.13) for Barcelona and (42.83, 46.37) for New York. This means that the difference between groups is statistically significant.

In conclusion, we can say that my hypothesis was correct given that after the calculations made, Barcelona's mean was lower and both confidence intervals did not overlap. This means that the populations pictured are not similar in this aspect.

2.9. Possible correlation between memory test time and time for homework

Questions: 33 (c) Estimate how many hours a week you spend doing homework

11. Test your memory, how many pairs can you uncover? (Students had to complete a memory test in which they had to uncover a number of pairs and time themselves)

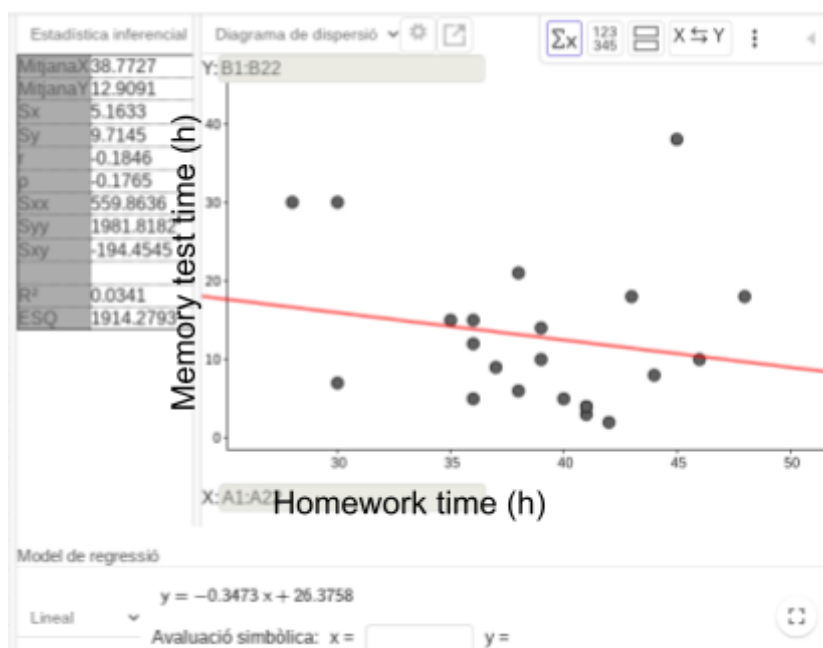
Variables: **Quantitative**

Statistical value: **r and R² for different models**

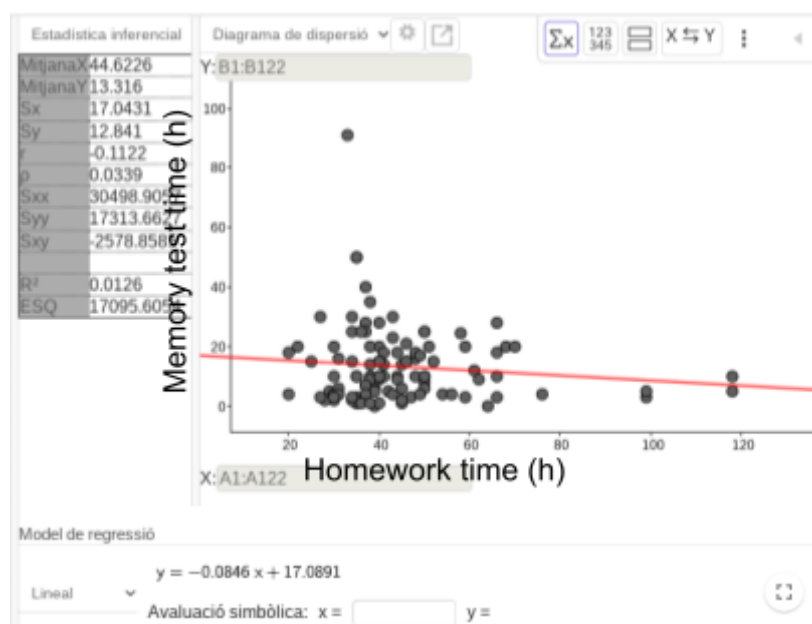
HYPOTHESIS: There is correlation between time taken doing the memory test and time spent doing homework.

Firstly, as always, I had to clean the data from the homework column and proceeded to check if the two data groups (memory test results and time spent doing homework) had any correlation. To do so, I took the data of both questions and put them into columns. Using Geogebra I created a scatter plot of the data:

Scatter plot of the Barcelona data:



After obtaining this scatter plot for the possible Barcelona correlation, I did the same with the New York data after they were cleaned.

Scatter plot of the New York data:

Then I organized the results of this linear model and other ones onto a table:

	Barcelona	New York
R Square (Linear)	0,0341	0,0126
R, Pearson coefficient	-0,1846	-0,1122
Linear model	$y = -0,3473x + 26,3758$	$y = -0,0846x + 17,0891$
R² (Log model)	0,0071	0,009
Log model	$y = -2.6709 + 4.5604 \ln(x)$	$y = 27,4559x - 3,7786 \ln(x)$
R² (Polynomial model)	0,0079	0,014
Polynomial model	$y = -0.0062x^2 + 0.6081x - 0.0001$	$y = -0.0009x^2 + 0.0265x + 14.1412$

In **conclusion**, these data and diagrams confirm that there is a bad correlation between the time students spend doing homework and the memory test time, both in Barcelona and in New York, because all the models give us a R² smaller than 0.5, and neither BCN's nor NY's r (Pearson Coefficient) is bigger than 0.7.

Therefore, the hypothesis was incorrect and both variables have no correlation.

2.10. Time for homework for students who plan to have a graduate degree

Questions: 35. What is the highest level of education you plan to attain?

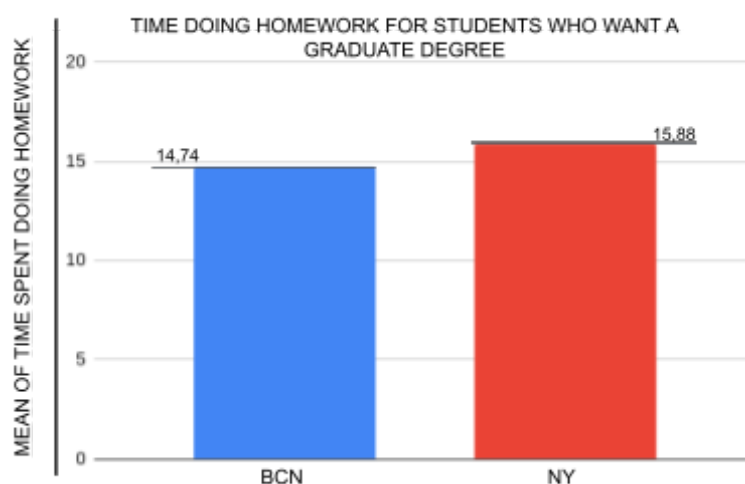
33 (c). Estimate how many hours a week you usually spend doing the following activity: - Doing homework

Variables: **Quantitative**

Statistical value: **Mean**

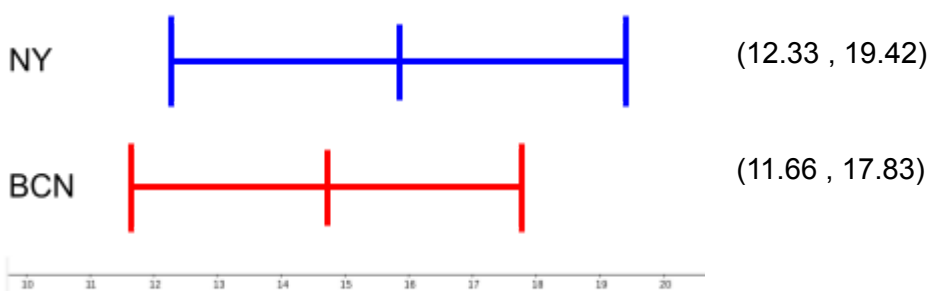
HYPOTHESIS: The BCN students who plan to have a graduate degree spend more time doing homework than NY students who proposed to have a graduate degree

We first calculated the average of the time spent doing homework for the part of the samples that wanted to obtain a graduate degree.



The average for NY's sample is a bit higher than for BCN's

Next, we calculated their confidence intervals



Conclusion:

As both confidence intervals overlapped, we could consider that both samples belong to the same population.

GENERAL CONCLUSIONS

Taking into account all the statistical work previously shown, we can say that the characteristics studied can be classified according to whether the two samples appear to belong to the same population or not.

In the first section, which talked about physical characteristics, the conclusions of the two hypotheses told us that the BCN and NY populations should be considered as different, but in the second section, that of cultural characteristics, there is a greater variety of criteria.

In hypotheses 2.1, 2.2, 2.7 and 2.10, the conclusion is that the population of the two cities could be considered as the same, while in hypotheses 2.3, 2.4, 2.6, 2.8, and 2.9, and 1.1 and 1.2, these two populations should be considered as different. Hypothesis 2.5 dealt only with samples.

Although it is true that most hypotheses we have studied suggest that the populations of NY and BCN are different, it should not be forgotten that they are not completely different as they have some similarities, some of them perhaps surprising, as the kind of people students from both samples look up to.

We expected that especially in the cultural questions (section 2) there would be differences because we thought that we were in front of culturally different countries, which should have discrepancies in lots of issues. Our work has shown us that previous ideas had to be tested to arrive at conclusions without prejudices. Similarities or differences, which must be related to particular characteristics of individuals, not to the individuals themselves, can not be taken for sure without research.

For the characteristics studied, only a few, and with our small samples of convenience, we must conclude that both populations are more different than equal, although they have their similarities and could be considered as the same population in some respects.

Finally, we think that we have met our objectives, listed at the beginning of this report.

WEBLIOGRAPHY

Census at School official website

<https://ww2.amstat.org/censusatschool/>

Online statistics calculators

<https://select-statistics.co.uk/calculators/>

Geogebra Classic online

<https://www.geogebra.org/classic?lang=ca>

[Spreadsheet with BCN data](#)

[Spreadsheet with NY data](#)