UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Satisfaction surveys improvement study. Analysis of the variables that could bias its reliability

REPORT

Author: Cunibert González Vila
Director: Montserrat Sánchez Romero
Degree: Mechanical Engineering
Release date: June 2020

Abstract

The Office of Planning, Evaluation, and Quality (GPAQ) manages a survey system to assess the performance of the teachers and the subjects. This survey system consists of two different surveys, one for the subjects and another for the teachers, that are conducted biannually and are answered by the students of bachelor's and master's degrees of the Polytechnic University of Catalonia (UPC).

The results obtained from these surveys are used to assess the quality of the subjects and the teaching quality of the staff. The obtained results can impact on the career of the teachers, as they are used as an input in promotions, and must also lead to improvements in the overall quality of the teaching.

If the results obtained from the surveys are biased in any manner, the conclusions derived from their results will not help to improve and could negatively affect the teachers or the subjects.

In the present study, different variables that could impact the teacher's surveys will be analyzed to clarify if there is any bias that could impact the reliability of the obtained results. Surveys conducted on ESEIAAT during the courses 2016/17, 2017/18, and 2018/19 will be analyzed.

Contents

List of tables

List of figures

1. INTRODUCTION

In the UPC (promoted from *Gabinet de Planificació, Avaluació i Qualitat,* GPAQ), surveys are carried out on the student satisfaction on the quality of the subjects and teachers. There is a model focused on teaching and another on subject quality. The survey focused on teaching quality consists of two questions related to the student view on the performance of the teacher, and the survey focused on subject quality consists of three questions related to the student view on the quality of the subject.

<table>
<tr><th>Survey</th><th>Questions</th><th>Rating scale</th></tr>
<tr><td rowspan="2">Teachers</td><td>1. The teacher is accessible to answer questions about the subject matter</td><td rowspan="5">1 - Strongly disagree to
5 - Strongly agree</td></tr>
<tr><td>2. I think he is a good teacher</td></tr>
<tr><td rowspan="3">Subjects</td><td>1. The subject matter was interesting to me</td></tr>
<tr><td>2. The evaluation corresponds with the objectives and level of the subject</td></tr>
<tr><td>3. Overall, I'm satisfied with this subject</td></tr>
</table>

Table 1 Satisfaction surveys model

The key question which has been analyzed in the previous studies is the one related to the overall evaluation of the teacher or the subject.

Survey	Key question
Teachers	I think he is a good teacher
Subjects	Overall, I'm satisfied with this subject

Table 2 Satisfaction surveys key questions

These surveys are used to assess whether the level and focus of the subjects are correct, so that teachers can consult their average rating and they can know if they are adequately communicating to students and if the way of transmitting their knowledge is the most accurate or needs some kind of change, also this rating is used to evaluate the teachers and serves as an input in evaluating promotions and salary complements.

They also serve to maintain the quality control of the university, since it is necessary to know if the students are acquiring the knowledge and abilities to develop their future professional careers and if they are acquiring all the concepts and objectives that their school curriculum establishes.

1.1 Objectives

From the GPAQ there has been an interest to study the satisfaction surveys because there was some concern about what factors could affect the overall rating of the subjects and teachers. For this reason, some projects have been sponsored in this line.

Last year (2019), a study about the subject satisfaction surveys was made in the ESEIAAT, also as a Mechanical Engineering Bachelor's thesis, and the present project will be the continuation of this previous one, studying the teacher satisfaction surveys and completing the study conducted in the ESEIAAT on this topic.

The objective of this project is to analyze the variables that could bias the rating on the key question of the teacher satisfaction surveys to improve the reliability of the results obtained from them.

1.2 Scope

This project will be focused on teacher satisfaction surveys.

The statistical analysis on the variables that could affect the reliability of the teacher satisfaction surveys will be conducted on data obtained in the last 3 years (2016, 2017, 2018) from the teacher satisfaction surveys of all the degrees (bachelor's and master's) offered in the ESEIAAT.

1.3 Justification

It has been noted, by the different studies conducted and by the people involved in the surveys, that some variables could bias its results. For example, it could be that faculty members that teach difficult subjects could obtain a different result that members that teach easier ones. This is the reason why this project is conducted, to assess the potential impact of the different variables involved and to propose improvements on the system to overcome the handicaps that the influence of these variables may pose on the survey system.

2. STATE OF THE ART

The analysis of the state of the art will help in identifying the conclusions and proposals made in the previous studies to establish a background in which the present study can be optimally conducted, benefiting and complementing all the previous work.

2.1 Previous studies

In total, there have been four studies conducted on the topic. They can be divided into two main groups:

- Studies focused on how the data obtained in the surveys must be analyzed statistically.
- Studies focused on the analysis of the variables that could affect the reliability of the obtained results.

2.1.1 Studies focused on how the data obtained in the surveys must be analyzed statistically

This group is integrated by one study, the first conducted in the UPC about the satisfaction surveys. The importance of this study draws in that it sets a methodology (established in the survey system according to *Acord CG/2019/04/15, de 4 de juliol de 2019, del Consell de Govern, pel qual s'aprova el model d'enquestes de l'estudiantat*) to determine which data obtained from the surveys are reliable and will be used and which must be discarded. This is explained in more detail in Section 4.1.

Study nº1. Evaluation of statistical tests on finite populations with a small sample size. Application on the analysis of teaching performance surveys.

- Author: Laura Campeny Carrasco
- Release date: October 2014
- Abstract: A reliable survey model is needed to assess teacher performance. The objective is to find criteria that guarantee sensitivity and specificity to the resolution of the contrasted hypotheses. Various statistical models are evaluated to find the most suitable for this kind of analysis.

2.1.2 Studies focused on the analysis of the variables that could affect the reliability of the obtained results

The most relevant previous work for the present study is this second group. In these studies, some variables have been analyzed, conclusions have been drawn and proposals have been made, so the present study will benefit and complement all this previous work.

This studies can also be divided into two groups:

- Studies conducted by students of the Bachelor's degree in Statistics on satisfaction surveys' data of the FIB (*Facultat d'Informàtica de Barcelona*).
- Studies conducted by students of the Bachelor's degree in Mechanical Engineering on satisfaction surveys' data of the ESEIAAT (*Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa*).

Studies conducted by students of the Bachelor's degree in Statistics on satisfaction surveys' data of the FIB:

From the GPAQ, it was asked to analyze the satisfaction surveys that are carried out in the UPC, since they had never been analyzed in-depth, so there was confusion and general doubt about what factors could affect the overall assessment of the subjects and teachers and which not.

That is why two studies were made. The first analyzing the subject satisfaction surveys and the second analyzing the teacher satisfaction surveys and closing the complete analysis on the topic.

The main objective of these studies was to find the factors that affect the key question of the two satisfaction surveys.

Study nº2. Analysis of surveys to students of the UPC about subjects and teachers (2016).

- Author: Jordi Fuster Arion
- Release date: September 2016
- Abstract: The analysis of the student satisfaction surveys affects aspects of the teaching quality evaluation that can be used to make improvements and

adequate the subject matter of the courses, also it can affect to the teachers' salary supplements. On the other hand, teachers, on several occasions, have demonstrated some concerns about the surveys, both from a methodological and participation point of view. The objective of the project is to address some of these concerns and, if necessary, improve the surveys or provide some compensation criteria.

Study nº3. Analysis of surveys to students of the UPC about subjects and teachers (2017).

- Author: Paloma Menéndez Landa
- Release date: June 2017
- Abstract: In this project, the teacher satisfaction surveys have been analyzed, and it has been exposed in summarized form the analysis that was carried out in the previous study (Jordi Fuster) about the subjects' surveys. This project intends to analyze in-depth the surveys and give some answers to the concerns and doubts that abound about certain aspects related to the subjects and the teaching provided.

Studies conducted by students of the Bachelor's degree in Mechanical Engineering on satisfaction surveys' data of the ESEIAAT:

In 2019, a similar analysis that the one conducted on the FIB was initiated in the ESEIAAT. A first study was conducted analyzing what variables could influence the subject satisfaction surveys on data from the Mechanical Engineering bachelor's degree of the ESEIAAT. The present project will complement it by analyzing the teacher satisfaction surveys and closing the complete analysis of the satisfaction surveys in the ESEIAAT.

Study nº4. Improvement study of the satisfaction surveys model. Reduction of the influence of the process variables in the reliability of the result.

- Author: Pol Agell Vendrell
- Release date: May 2019
- Abstract: In this project, a study of the variables likely to influence the results of the subject satisfaction surveys is carried out and proposals are made to improve the reliability of the results of the surveys.

2.2 Relevant conclusions and proposals of the previous studies

In this section, the relevant conclusions and proposals of the previous studies are listed. It must be noted that the conclusions on the influence of the different variables studied will be analyzed in Section 4.1 so they are out of the scope of this Section.

Study	Conclusions and proposals
Study nº1	The Dirichlet distribution is the one that fits better with the studied problem. A methodology to determine the reliability of the obtained results is established.
Study nº2	Redesign of the survey so that it is sent in a random and stratified way to enough number of students and can be followed to ensure the participation is high.
	It is proposed the use of the median instead of the mean in the results of the subject and teacher satisfaction surveys.
Study nº3	The same survey redesign as the study nº2 is proposed.
	It is concluded that it is not necessary to use the median instead of the mean because the same conclusions of the analysis were drawn using the two statistics.
Study nº4	Quantify the influence of the variables that affect the results and determine correction factors.
	Extend the study to more variables.
	Extend the study to more degrees.
	Complement the study by analyzing teacher satisfaction surveys.

Table 3 Conclusions and proposals from the previous studies

From the conclusions drawn from the previous studies, the following aspects will be taken into account in the present study:

- It will be used the mean in the statistical analysis. The Study nº3 demonstrated that the mean is a convenient statistic and produces equivalent results than the median in testing the influence of the different variables. In Section 3 the statistical tests used are described.
- The study will be extended to more degrees than the last one conducted in ESEIAAT. Instead of analyzing only the data of one degree, the data of all 21 degrees (including bachelor's and master's) taught at ESEIAAT will be included in the analysis.
- The analysis will focus on the teacher satisfaction surveys, thus closing the study conducted in ESEIAAT started by Study nº4.

3. METHODOLOGY

In the analysis of the teacher satisfaction surveys, inferential statistical techniques have been used, more specifically, the Student's t-test. This test has been used to assess if there is enough evidence to claim that a factor influences the teachers' average rating or not.

One of the problems when carrying out the analysis, which was already detected in the previous studies conducted on this topic, is that to apply inferential statistical techniques correctly, the sample used should be representative and random. The data collected by the surveys, because it is sent to all the students and is answered voluntarily, it is not random and could not be 100% representative (for example, only the satisfied students could answer the survey or vice versa). However, it is considered that the inferential analysis of this data can provide useful results because samples are big and data obtained from surveys with low participation (that could be more biased) is not used in the analysis. Also, descriptive statistics will be used and the conclusions drawn will not rely solely on results obtained from statistical inference.

Another problem that arises is that the data will not probably follow a normal distribution, which is one of the requirements of the Student's t-test. This type of data is usually skewed to the right, the mean is not centered and there are more extreme observations in the higher ratings than in the lower ones (big tail in the right), as will be seen in the histograms when analyzing the data. However, it is considered that the Student's t-test will still provide useful results, because is a test that, although assumes that the populations are normally distributed, due to the central limit theorem the test may still be useful when this assumption is not true if the sample sizes are big enough (which holds for this data). In the case of non-normality, the Wilcoxon-Mann-Whitney test, which is a test that does not assume a normal distribution, will also be used, and the results of the two tests will be compared to conclude if a significant difference in the means can be claimed.

Microsoft Excel 2016 has been used for data processing and *Analyse-it v.4.95.1* for the statistical analysis. *Analyse-it* offers, besides all the statistical tests needed to conduct this study, the advantage that it works as an *Excel* add-in, in a way that it is easier to bridge from data processing to data analyzing.

3.1 Student's t-test

To assess which variables affect the rating of a teacher, different Student's t-test will be executed, one for every studied variable.

This methodology will allow assessing if the difference in the mean of two samples is statistically significant so to claim that the theoretical means are different.

In this test, the null hypothesis states that the means of the populations are equal, and the alternative hypothesis states that they are not equal.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The test will be executed with a 5% significance level, which corresponds to a 95% confidence level. If the obtained p-value is lower than the significance level, then there is sufficient statistical evidence to reject the null hypothesis and claim that the means of the two analyzed samples are not equal.

For the equality of means hypothesis testing to generate reliable results, the three assumptions of the method must be satisfied:

- The populations are normally distributed.
- The two samples are independent of each other.
- The population variances are equal.

These assumptions will be checked in every analysis using the tests described in the following sections.

Assumption	Check	Test
Populations are normally distributed	Normality hypothesis test	Anderson-Darling
The two samples are independent	Will not be checked. It is ensured by the survey design	-
Population variances are equal	Homogeneity of variance hypothesis test	Brown-Forsythe

Table 4 Student's t-test assumptions checks

The independence of the two samples will not be checked statistically as the design of the survey ensures that the variables are independent.

As stated at the start of this Section, the data will not probably follow a normal distribution. In that case, the other assumptions will be checked and the Wilcoxon-Mann-Whitney test will also be executed.

If the results obtained with the two tests are equivalent, the statistical result derived from the inferences will be claimed.

3.2 Wilcoxon-Mann-Whitney test

In the case of non-normality in the data, in addition to the Student's t-test, a Wilcoxon-Mann-Whitney test will also be executed. This is a non-parametric test that does not assume normality and allows to assess if there is a shift in location between two samples.

In this test, the null hypothesis states that the Hodges–Lehmann estimator of the two populations are equal, and the alternative hypothesis states that they are not equal.

$$H_0: \Delta_1 = \Delta_2$$
$$H_1: \Delta_1 \neq \Delta_2$$

When the population distributions are similarly shaped, except for a possible shift in central location, these hypotheses can be stated in terms of a difference between means. This will allow us to compare the results of the Student's t-test with the Wilcoxon-Mann-Whitney test before claiming any difference in the means.

The test will be executed with a 5% significance level, which corresponds to a 95% confidence level. If the obtained p-value is lower than the significance level, then there is sufficient statistical evidence to reject the null hypothesis and claim that there is a significant shift in the location of the two populations.

This test assumes the independence of the variables, which is ensured by the design of the surveys.

3.3 Normality hypothesis test

To test if the analyzed data is normally distributed, normality hypothesis tests will be executed, one for every sample of data analyzed.

In this test, the null hypothesis states that the population is normally distributed, and the alternative hypothesis states that it is not normally distributed.

$$H_0: F(x) = N(\mu, \sigma)$$

$$H_1: F(x) \neq N(\mu, \sigma)$$

The Anderson-Darling test will be executed with a 5% significance level, which corresponds to a 95% confidence level. If the obtained p-value is higher than the significance level, then there is sufficient statistical evidence to claim that the data is normally distributed.

3.4 Homogeneity of variance hypothesis test

To test if the two analyzed samples have equal variances, homogeneity hypothesis tests will be executed, one for every pair of samples analyzed.

Many statistical hypothesis tests assume that the variances of the populations are equal. This assumption allows the variances of each group to be pooled together to provide a better estimate of the population variance. A better estimate of the variance increases the statistical power of the test, meaning that smaller sample size can be used to detect the same difference, or that the same sample size can be used to detect smaller differences and make sharper inferences.

In this test, the null hypothesis states that the variances of the populations are equal and the alternative hypothesis states that they are not equal.

$$H_o: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

The Brown-Forsythe test (which is robust against many types of non-normality, unlike other tests like Fisher's F) will be executed with a 5% significance level, which corresponds to a 95% confidence level. If the obtained p-value is higher than the significance level then there is sufficient statistical evidence to claim that the two samples do not differ in variance.

4. SELECTION OF THE VARIABLES TO STUDY

For the selection of the variables to analyze, several sources of information have been taken into account.

All the variables analyzed in the previous studies conducted on this topic have been collected and evaluated, including the relevant conclusions. This will help in avoiding the analysis of variables found to be non-relevant and to continue the previous work on the relevant variables, taking into account previous conclusions, continuing and improving the past work, and clarifying possible contradictions between the conclusions of the different studies. This evaluation will also help in establishing a clear picture of all the variables studied before and the relevant conclusions, to serve as a background for this study to start.

One of the reasons for conducting these studies was to clarify some confusion and general doubt about what factors could affect the overall assessment of the subjects and teachers and which not. In this line, proposals from the teaching staff are included, and the variables proposed by the faculty members will also be evaluated.

After evaluating all the variables studied in the previous work and the variables proposed by the teaching staff, some variables will also be proposed by the author of this study.

4.1 Variables analyzed in the previous studies

Many variables have been analyzed in previous studies. In this Section, all of them will be evaluated and the conclusions drawn from every study will be taken into account in the selection.

Before evaluating all of the variables studied before, they are compiled in a table format so it is easier to see them at a glance.

Study	Survey	Variable	Relevant	Selected
Study nº2	Subject	Semester (1st / 2nd)	NO	NO
		Subject type (compulsory / optional)	YES	YES
		Degree level (bachelor's/master's)	YES	YES
		Number of subject repeaters	Non-conclusive	YES
		Subject difficulty	YES	NO
		Timetable (morning / afternoon)	NO	NO
		Number of students	NO	YES
		Survey participation	NO	NO[1]
Study nº3	Teacher	Semester (1st / 2nd)	NO	NO
		Subject type (compulsory / optional)	YES	YES
		Degree level (bachelor's/master's)	YES	YES
		Number of subject repeaters	NO	YES
		Subject difficulty	YES	NO
		Phase (initial / non-initial)	NO	YES
Study nº4	Subject	Subject type (compulsory / optional)	YES	YES
		Number of students	YES	YES
		Subject difficulty	Non-conclusive	NO

Table 5 Variables analyzed in the previous studies

[1] Although it will not be analyzed as a variable, its influence will be taken into account according to the methodology established in Study nº1 and explained in more detail in this section.

Semester (1st / 2nd):

It was thought that the semester in which the survey is carried out could impact on the obtained results, due to the proximity of holidays, the general mood on the different seasons or other environmental reasons.

This variable was studied in Studies number 2 and 3 and was found to be non-significant to the ratings of the subjects or teachers.

As this variable has been found to be non-significant the two times it was analyzed it is not selected for the present study.

Subject type (compulsory / optional):

Compulsory subjects have a higher number of students in the classrooms than optional subjects. The kind of teaching that can be done in a classroom with a lower number of students is not the same that can be done in a higher populated one. Also, optional subjects are selected by the students according to their likes, so the subject matter can be more of their interest and they can rate the subject or the teacher in a more predisposed way.

This variable has been found to be relevant in all the previous studies. It is an important variable and will also be analyzed in the present study.

Degree level (bachelor's/master's):

In a similar way as with the subject type, master's degrees use to have a lower number of students in their classrooms and the subject matter is more specific and more likely to be of the interest of the student than bachelor's degrees. Also, the students attending master's degrees can have a different opinion than bachelor's students, they are usually older and more mature people with a broader view of college teaching.

This variable was found relevant in the two studies conducted in the FIB and was not analyzed in the study conducted on ESEIAAT, so it is found to be relevant to also study it in ESEIAAT and it will be part of the present study.

Number of subject repeaters:

The number of subject repeaters in a classroom can be indicative of many things. It could be indicative that the subject is difficult or that the teaching is deficient and also could influence the survey's obtained results. Subject repeaters could not be very predisposed to answering the survey objectively and could be impartial in rating the subject or the teacher.

There are no conclusions on the influence of this variable in Study nº2, and Study nº3 found it to be not relevant in the overall rating of the teachers. However, this variable, as it will be explained in the next Section, has been a concern of the teaching staff during this year, so it will be analyzed in the present study.

Subject difficulty:

Subject difficulty could negatively affect the obtained results and could even affect the level of participation because students busy in difficult subjects could not have the time or predisposition to answer the surveys.

This variable was found to be relevant in the two studies conducted in FIB and no conclusions were drawn in the study conducted in ESEIAAT. Although an important variable, it is not selected to be analyzed in the present study because the number of subject repeaters, which is also indicative of the subject difficulty, will be analyzed.

Timetable (morning / afternoon):

The timetable in which the subject is taught was analyzed in study nº2 and it was found to be non-relevant.

In ESEIAAT, the student does not always voluntarily select the timetable. Every degree has its timetable (for example, Mechanical Engineering has a compulsory afternoon timetable) so this variable can be very interdependent with other variables like the degree and could be difficult to draw conclusions from its analysis. As it was found non-relevant in another study and it seems to be a variable not particularly relevant in ESEIAAT, it will not be selected for the present study.

Number of students:

The number of students enrolled in the subject can be an influential factor. The kind of teaching that can be done in a classroom with a lower number of students is not the same that can be done in a higher populated one.

This variable was analyzed in two studies, drawing contradictory conclusions. Study nº2 found it to be non-relevant and Study nº4 found it relevant. As it is a variable that has not been analyzed in the teacher surveys and had been found relevant in the previous study conducted in ESEIAAT, it will be analyzed in the present study, to try to clarify the contradictory conclusions of the previous studies and broaden the information on the influence of this particular variable on the teacher surveys in ESEIAAT.

Survey participation:

Survey participation is probably the most important variable in the reliability of the obtained results of a satisfaction survey because low participation can pose an important statistical bias thus affecting the reliability of the obtained results.

For this reason, a methodology, derived from the proposal of Study nº1 was established in the UPC[2]. Every survey is classified in one of three categories:

- Low / very low reliability
- Medium reliability
- High / very high reliability

This methodology establishes that:

- Low /very low reliable surveys will not be taken into account. The information is available for the teacher to consult but it will not be used for evaluation purposes.
- Medium reliable surveys can be canceled if the teacher involved asks it and justifies it properly.
- In high /very high reliable surveys, allegations involving a lack of representation will not be taken into account.

[2] *Acord CG/2019/04/15, de 4 de juliol de 2019, del Consell de Govern, pel qual s'aprova el model d'enquestes de l'estudiantat*

This variable has been studied in-depth and a methodology has been established to avoid its influence, so it will not be analyzed in the present study. However, its influence will be taken into account, so data from low / very low reliable surveys will not be used in the analysis.

<u>Phase (initial / non-initial)</u>:

Initial phase subjects are usually the most broader and non-specific of all the subjects of a degree. They are usually subjects which could not be of the interest of the students. On the contrary, the subjects which are more oriented towards the specific details of the degree and can be of the interest of the student are the non-initial subjects. This can affect the view of the student on the subject or the teacher and influence the results of the survey.

This variable was analyzed in Study nº3 and was found to be non-relevant for the results of the teacher surveys. However, it will be analyzed in the present study to determine if this holds true in ESEIAAT.

4.2 Variables proposed by the teaching staff

The faculty members' opinions on the survey system are a very important source of information to the studies conducted on the topic. The survey system impacts importantly on the teachers' professional life, so all his opinions must be taken into account.

This year, the GPAQ received a proposal of a variable to analyze from faculty members. This variable is the number of repeaters.

Variable	Previously analyzed	Previously found relevant
Number of repeaters	YES	NO

Table 6 Variables proposed by the teaching staff

Number of repeaters:

As stated in the previous Section, this variable was studied before and was found to be non-relevant. However, as it has been a concern of the teaching staff it will be analyzed in the present study.

4.3 Variables proposed by the author

After compiling all the variables studied before and the ones proposed by the teaching staff, a relevant number of variables have been selected. However, there is a variable that the author of this study finds relevant and that has not been analyzed before.

It is believed by the author to be of high interest to analyze if there is any kind of difference in the overall rating of the teachers depending on their gender.

Variable	Previously analyzed
Teacher's gender (men/women)	NO

Table 7 Variables proposed by the author

4.4 Selected variables

After the evaluation conducted in Section 4, the following variables are selected to be analyzed in the present study.

Variable	Previously analyzed
Subject type (compulsory/optional)	YES
Degree level (bachelor's/master's)	YES
Number of subject repeaters	YES
Number of students	YES
Phase (initial/non-initial)	YES
Teacher's gender (men/women)	NO

Table 8 Selected variables

5. ANALYSIS OF THE INFLUENCE OF THE VARIABLES

The analysis will start with descriptive statistics (histograms, frequency tables, etc), and after examining the results obtained, the different hypothesis tests will be executed and conclusions will be drawn.

First, the whole dataset will be analyzed and, if an influence of the variable is observed, it will be tested in all the semesters to study if the data always behave similarly and the influence can be confirmed.

5.1 Subject type (compulsory/optional)

5.1.1 Analysis of all the dataset

The data of the different teacher satisfaction surveys conducted in 2016, 2017, and 2018 has been sorted according to if the teacher was evaluated in his work in a compulsory or an optional subject.

The final data is composed of 4778 datapoints, 3971 corresponding to compulsory subjects, and 808 corresponding to optional subjects.

Descriptive statistics analysis

Before starting the hypothesis testing, a histogram of every one of the two populations is plotted. This allows us to conduct a graphical analysis of the data before starting the inferences.

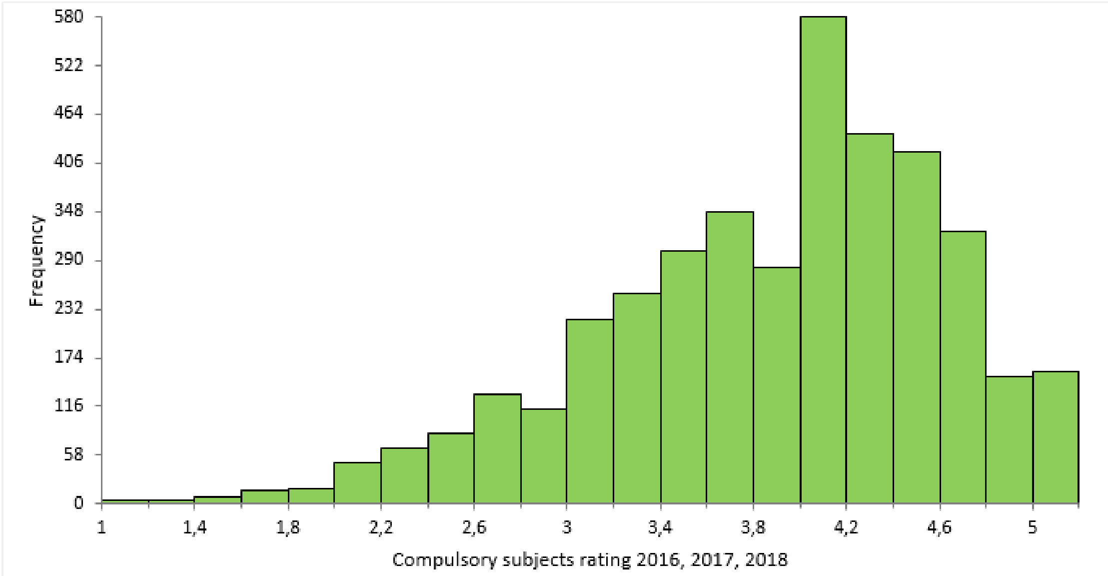


Figure 1 2016, 2017, 2018 Compulsory subjects rating histogram

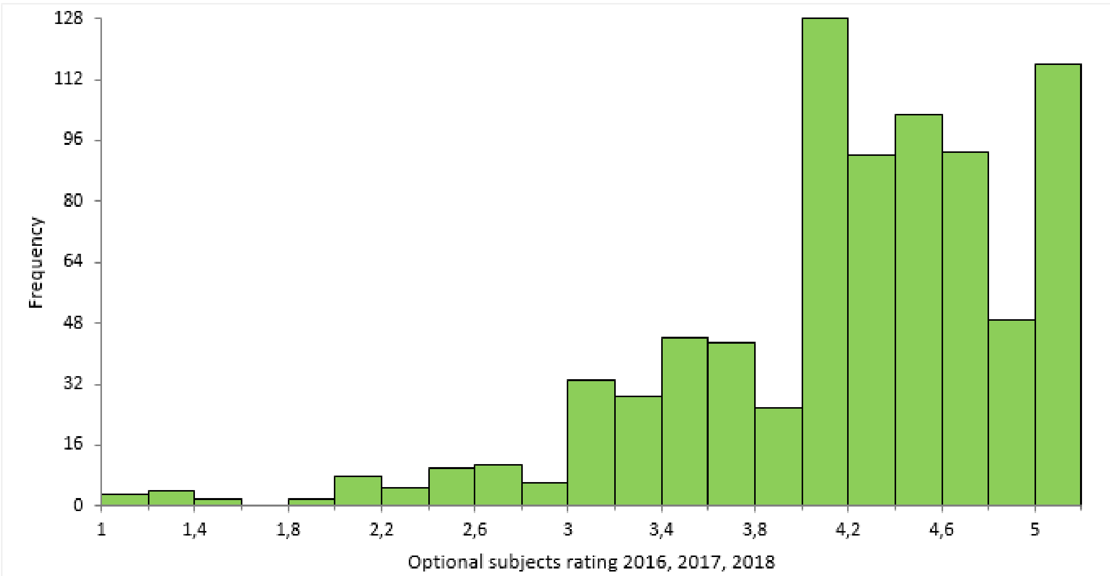


Figure 2 2016, 2017, 2018 Optional subjects rating histogram

From these histograms, some preliminary conclusions can be drawn:

- As was foretold in Section 3, **the data will most probably be non-normal**, as there is a high skew to the right.
- In general, **teachers are well-rated** (around the 4 mark) **no matter the type of subject they teach**. In the two distributions, the most frequent ratings appear to be around the 4 mark. There is a higher frequency of high ratings (higher than 4)

than lower ratings (lower than 2). There appears not to be a significantly[3] higher density in the lower ratings in the compulsory subjects than the optional ones and the distribution is similar (although not equal) for both types of populations.

- Notwithstanding the previous conclusion, **in the optional subjects, there is a higher density of very high ratings** (higher than 4,5 and close to 5) **than in the compulsory ones**. For analyzing this aspect more quantitatively, the distribution tables of the two populations are composed.

Class	Frequency	Relative frequency	Density	Cumulative frequency	Cumulative relative frequency
≥1 to <1,2	5	0,001	0,0063	5	0,001
≥1,2 to <1,4	4	0,001	0,0050	9	0,002
≥1,4 to <1,6	8	0,002	0,0101	17	0,004
≥1,6 to <1,8	17	0,004	0,0214	34	0,009
≥1,8 to <2	18	0,005	0,0227	52	0,013
≥2 to <2,2	50	0,013	0,0630	102	0,026
≥2,2 to <2,4	66	0,017	0,0831	168	0,042
≥2,4 to <2,6	84	0,021	0,1058	252	0,063
≥2,6 to <2,8	131	0,033	0,1649	383	0,096
≥2,8 to <3	113	0,028	0,1423	496	0,125
≥3 to <3,2	219	0,055	0,2757	715	0,180
≥3,2 to <3,4	250	0,063	0,3148	965	0,243
≥3,4 to <3,6	302	0,076	0,3803	1267	0,319
≥3,6 to <3,8	347	0,087	0,4369	1614	0,406
≥3,8 to <4	282	0,071	0,3551	1896	0,477
≥4 to <4,2	580	0,146	0,7303	2476	0,624
≥4,2 to <4,4	441	0,111	0,5553	2917	0,735
≥4,4 to <4,6	420	0,106	0,5288	3337	0,840
≥4,6 to <4,8	324	0,082	0,4080	3661	0,922
≥4,8 to <5	152	0,038	0,1914	3813	0,960
≥5 to <5,2	158	0,040	0,1989	3971	1,000

Table 9 2016, 2017, 2018 Compulsory subjects rating distribution table

From this table, it can be calculated that the relative frequency of rating below 2,2 is 2,6% and the relative frequency of rating above 4 is 52,3%.

[3] Not to be understood in terms of statistical significance, it is just a graphical observation.

Class	Frequency	Relative frequency	Density	Cumulative frequency	Cumulative relative frequency
≥1 to <1,2	3	0,004	0,0186	3	0,004
≥1,2 to <1,4	4	0,005	0,0248	7	0,009
≥1,4 to <1,6	2	0,002	0,0124	9	0,011
≥1,6 to <1,8	0	0,000	0,0000	9	0,011
≥1,8 to <2	2	0,002	0,0124	11	0,014
≥2 to <2,2	8	0,010	0,0496	19	0,024
≥2,2 to <2,4	5	0,006	0,0310	24	0,030
≥2,4 to <2,6	10	0,012	0,0620	34	0,042
≥2,6 to <2,8	11	0,014	0,0682	45	0,056
≥2,8 to <3	6	0,007	0,0372	51	0,063
≥3 to <3,2	33	0,041	0,2045	84	0,104
≥3,2 to <3,4	29	0,036	0,1797	113	0,140
≥3,4 to <3,6	44	0,055	0,2726	157	0,195
≥3,6 to <3,8	43	0,053	0,2664	200	0,248
≥3,8 to <4	26	0,032	0,1611	226	0,280
≥4 to <4,2	128	0,159	0,7931	354	0,439
≥4,2 to <4,4	92	0,114	0,5700	446	0,553
≥4,4 to <4,6	103	0,128	0,6382	549	0,680
≥4,6 to <4,8	93	0,115	0,5762	642	0,796
≥4,8 to <5	49	0,061	0,3036	691	0,856
≥5 to <5,2	116	0,144	0,7187	807	1,000

Table 10 2016, 2017, 2018 Optional subjects rating distribution table

From this table, it can be calculated that the relative frequency of rating below 2,2 is 2,3% and the relative frequency of rating above 4 is 72,10%.

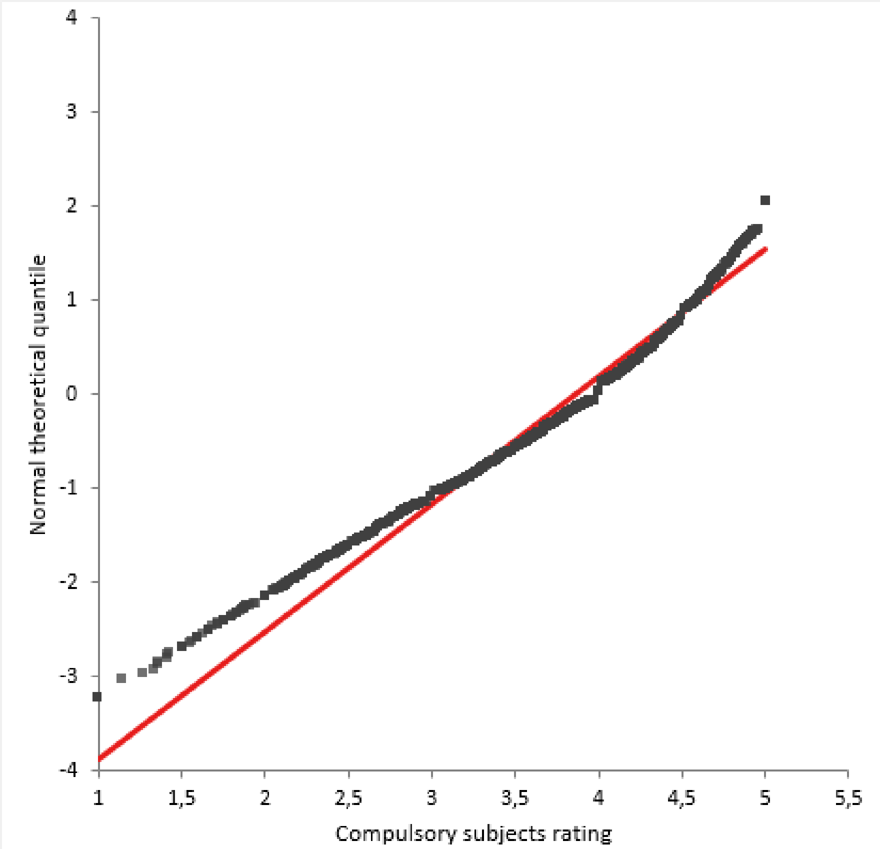
From the distribution tables some preliminary conclusions drawn in examining the histograms can be further understood:

- **Teachers of compulsory subjects do not appear to have a significantly higher density of lower ratings than optional subjects' ones.**
- **In optional subjects, there is a higher density of very high ratings than in compulsory ones**. In the distribution tables, it can be seen than in compulsory subjects the 52,3% of the ratings are above 4, and in optional subjects, this relative frequency grows up to 72,10%. This could cause a difference in the mean of the two populations.

Hypothesis testing

Before starting any hypothesis testing, the two populations are tested for normality.

Normality



Anderson-Darling test

A^2 statistic	38,18
p-value	<0,0001 [1]

H0: F(Y) = N(μ, σ)

The distribution of the population is normal with unspecified mean and standard deviation.

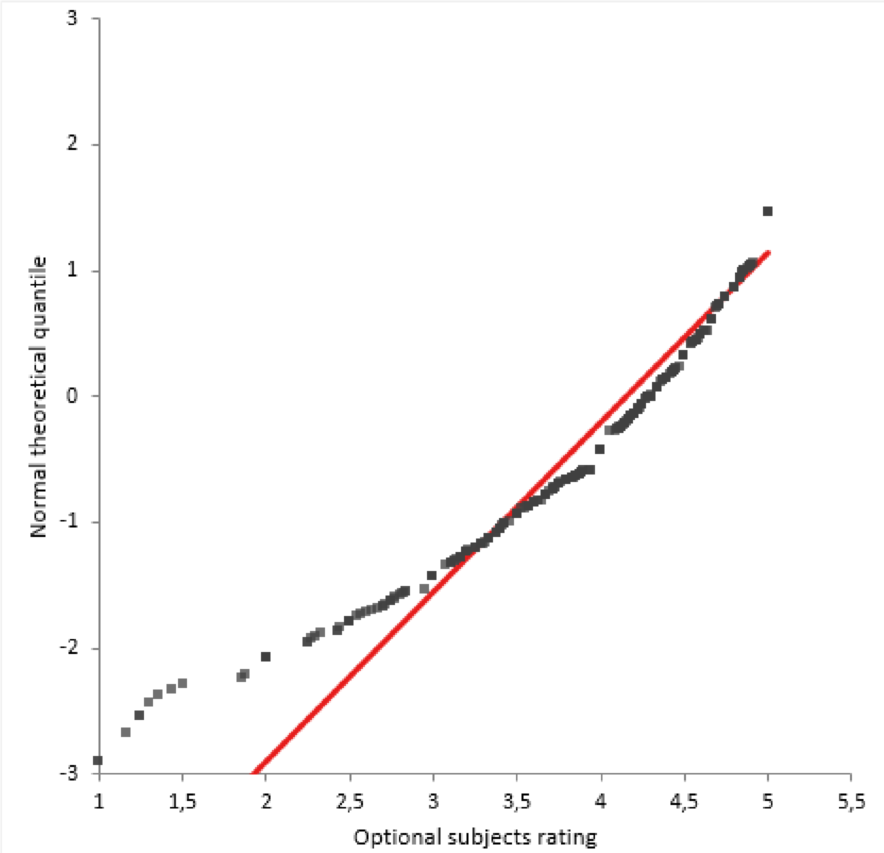
H1: F(Y) ≠ N(μ, σ)

The distribution of the population is not normal.

[1] Reject the null hypothesis in favour of the alternative hypothesis at the 5% significance level.

Figure 3 2016, 2017, 2018 Compulsory subjects Anderson-Darling test of normality

Normality



Anderson-Darling test

A^2 statistic	18,69
p-value	<0,0001 [1]

H0: F(Y) = N(μ, σ)
The distribution of the population is normal with unspecified mean and standard deviation.
H1: F(Y) ≠ N(μ, σ)
The distribution of the population is not normal.
[1] Reject the null hypothesis in favour of the alternative hypothesis at the 5% significance level.

Figure 4 2016, 2017, 2018 Optional subjects Anderson-Darling test of normality

As was expected from examining the histograms, none of the two populations follow a normal distribution. However, they will be tested for homogeneity of variances and taking into account the big number of datapoints and the similarity between their distributions (as seen in the histograms), the Students' t and the Wilcoxon-Mann-Whitney test will be executed to assess if there is any difference in their means.

Brown-Forsythe test:

Descriptives

N	4778

Response by Factor	N	Mean	Mean SE	SD
OB	3971	3,856070568	0,011703941	0,737534039
OP	807	4,148446481	0,026108506	0,741683788

Dispersion

Brown-Forsythe test

F statistic	1,95
Numerator DF	1
Denominator DF	4776
p-value	0,1627 [1]

H0: $\sigma^2_1 = \sigma^2_2 = \sigma^2 ...$

The variance of the populations are all equal.

H1: $\sigma^2_i \neq \sigma^2_j$ for at least one i,j

The variance of the populations are not all equal.

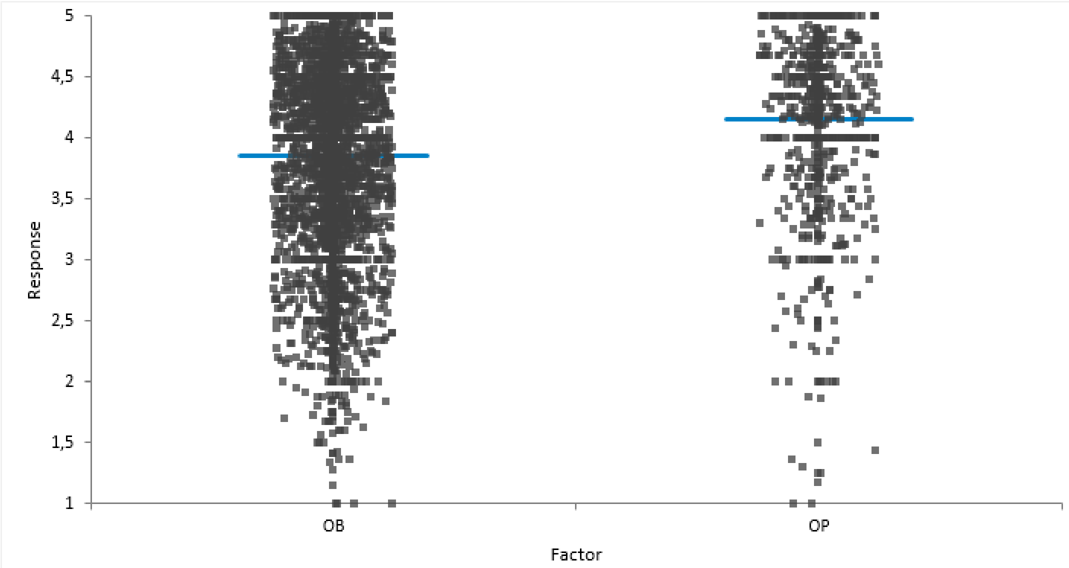
[1] Do not reject the null hypothesis at the 5% significance level.

Figure 5 2016, 2017, 2018 Type of subject homogeneity of variance test

According to the results of the Brown Forsythe-test, the variances of the two populations are equal.

Student's t-test:

Descriptives



N	4778

Response by Factor	N	Mean	Mean SE	SD
OB	3971	3,856070568	0,011703941	0,737534039
OP	807	4,148446481	0,026108506	0,741683788

Location

Mean difference	0,292375912
95% CI	0,236491622 to 0,348260203
SE	0,028505691

$\mu_\Delta = \mu_{OP} - \mu_{OB}$

Student's t test

Hypothesized difference	0

t statistic	10,26
DF	4776
p-value	<0,0001 [1]

H0: $\mu_\Delta = 0$
The difference between the means of the populations is equal to 0.
H1: $\mu_\Delta \neq 0$
The difference between the means of the populations is not equal to 0.
[1] Reject the null hypothesis in favour of the alternative hypothesis at the 5% significance level.

Figure 6 2016, 2017, 2018 Type of subject Student's t-test

Wilcoxon-Mann-Whitney test:

Descriptives

N	4778

Response by Factor	Minimum	1st Quartile	Median	3rd Quartile	Maximum
OB	1,00000000	3,400000000	4,000000000	4,411319073	5,00000000
OP	1,00000000	3,800000000	4,300000000	4,666666667	5,00000000

Location

Hodges-Lehmann shift	0,309523810
95% CI	0,250000000 to 0,352380952

$F(_{OP})=F(_{OB}+\Delta)$

Wilcoxon-Mann-Whitney test

Hypothesized difference	0

Response by Factor	N	Rank sum	Mean rank
OP	807	2339309,0	2898,77
OB	3971	9077722,0	2286,00

W statistic	2339309,00
Z approximation	11,51
p-value	<0,0001 [1]

H0: $\Delta = 0$

The shift in location between the distributions of the populations is equal to 0.

H1: $\Delta \neq 0$

The shift in location between the distributions of the populations is not equal to 0.

[1] Reject the null hypothesis in favour of the alternative hypothesis at the 5% significance level.

Figure 7 2016, 2017, 2018 Type of subject Wilcoxon-Mann-Whitney test

The results of the two hypothesis tests are summarized in the following table:

<table>
<tr><th>Test</th><th>Compulsory subjects mean</th><th>Optional subjects mean</th><th>p-value</th><th>Mean or location difference</th></tr>
<tr><td>Student's t</td><td rowspan="2">3,85</td><td rowspan="2">4,15</td><td><0,0001</td><td>0,292</td></tr>
<tr><td>Wilcoxon-Mann-Whitney</td><td><0,0001</td><td>0,309</td></tr>
</table>

Table 11 2016, 2017, 2018 Type of subject hypothesis tests results

These results show that **there is enough statistical evidence to conclude that the means of the two populations are different, being the mean of the optional subjects higher,** by an amount of approximately 0,3 points[4]. As the two tests provide similar results, the results are considered acceptable.

5.1.2 Semester 1 2016

Data of Semester 1 2016 is composed of 799 datapoints, 646 corresponding to compulsory subjects, and 153 corresponding to optional subjects.

Descriptive statistics analysis

The histograms of Semester 1 2016 two populations will be compared to the histograms of all the dataset, to study if their histograms are similar and the conclusions drawn from analyzing the whole dataset remain true in this subgroup of data.

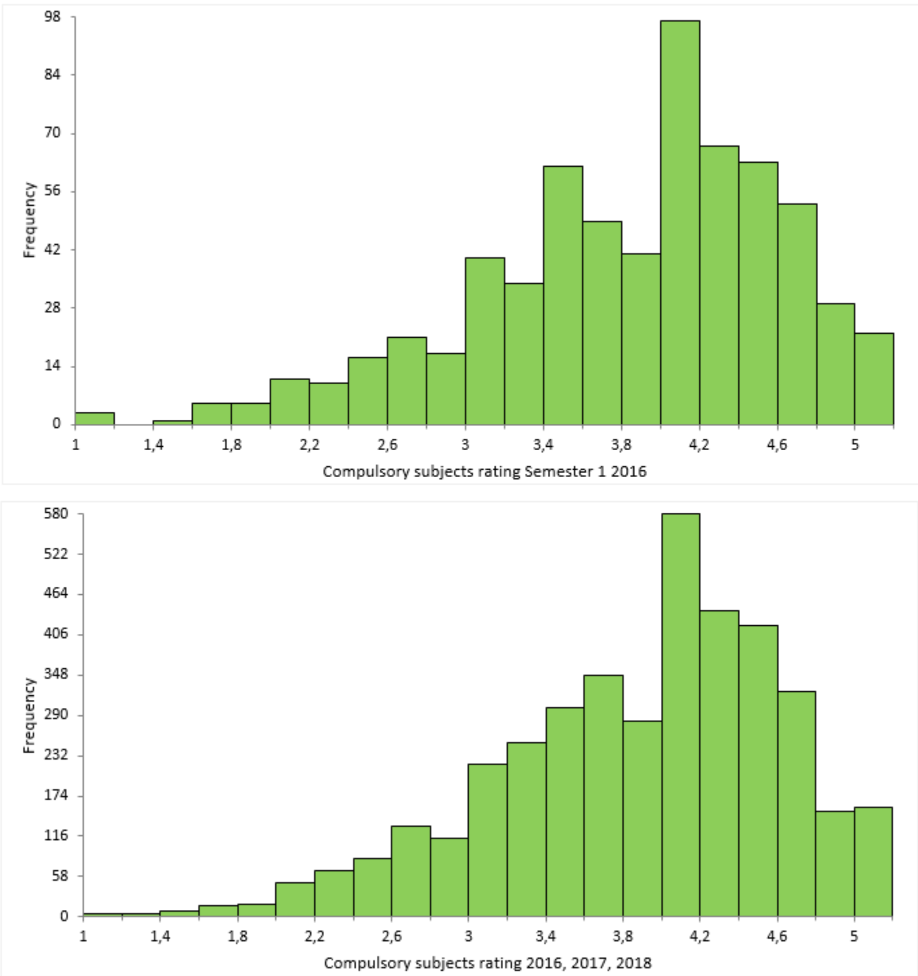
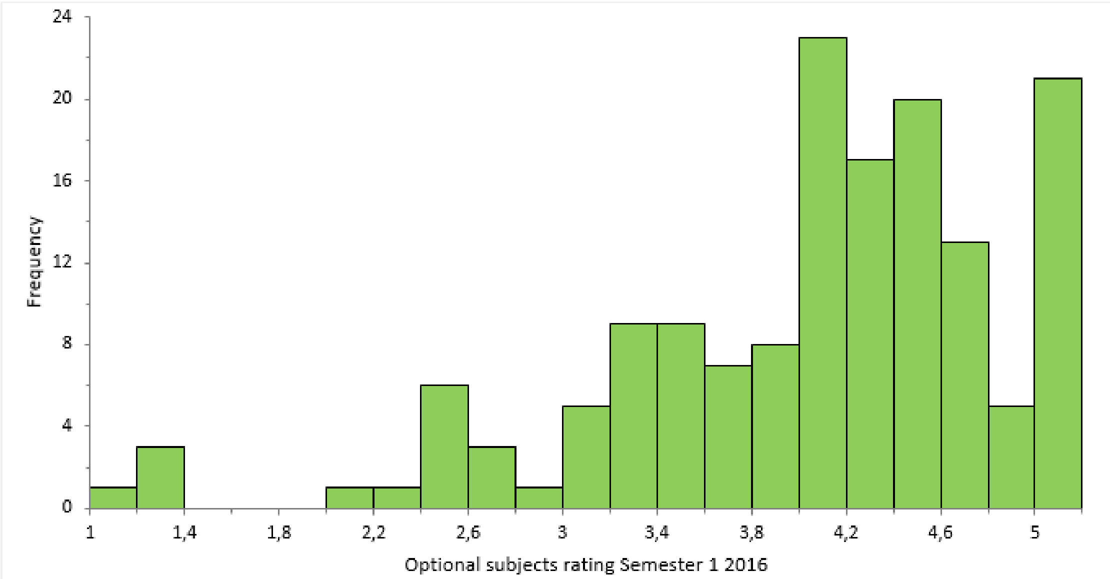


Figure 8 S1 2016 vs 2016, 2017, 2018 Compulsory subjects histogram comparison

[4] Student's t 95% confidence interval: 0,24 - 0,35



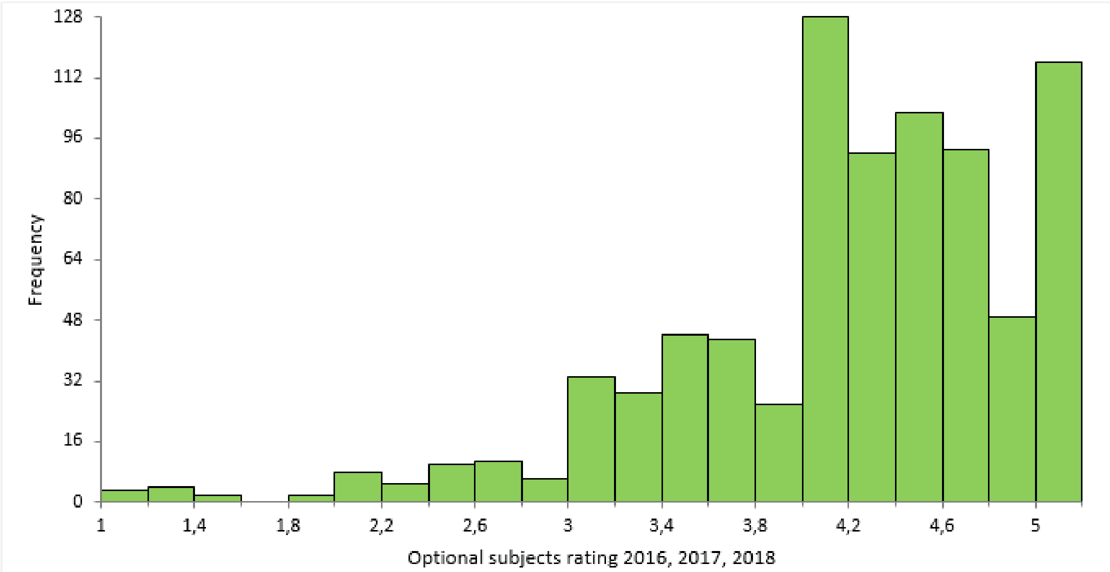


Figure 9 S1 2016 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of Semester 1 2016 are almost identical to the histograms of the whole dataset.

Hypothesis testing

Before starting the hypothesis testing on Semester 1 2016 data, the comparison of the means between S1 2016 and 2016, 2017, 2018 is shown in the following table.

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 1 2016	3,82	4,01	0,19
2016, 2017, 2018	3,85	4,15	0,3

Table 12 S1 2016 vs 2016, 2017, 2018 means comparison

As could have been supposed in seeing the similarity between their histograms, in Semester 1 2016 the optional subjects mean is also higher than the compulsory subjects mean.

However, there is an important difference in the optional subjects' mean values. This difference, which will also appear in analyzing some of the other periods, will be analyzed in Section 5.1.8.

From now on, to improve the readability of the study, the test results will be summarized in tables. All the raw data of the tests are available in the Annexes.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,7894	Variances are equal

Table 13 S1 2016 normality and homogeneity of variances test results

The data, as it was supposed, is non-normal but the variances are equal, so the Student's t and the Wilcoxon-Mann-Whitney tests will be executed.

Student's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Student's t	3,82	4,01	0,0065	0,19
Wilcoxon-Mann-Whitney			0,0007	0,22

Table 14 S1 2016 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different**, being the mean of the optional subjects higher, by an amount of approximately 0,2 points. As the two tests provide similar results, the results are considered acceptable.

Semester 1 2016 data behaves similarly to the whole dataset.

5.1.3 Semester 2 2016

Data of Semester 2 2016 is composed of 718 datapoints, 620 corresponding to compulsory subjects, and 98 corresponding to optional subjects.

<u>Descriptive statistics analysis</u>

Histogram comparison:

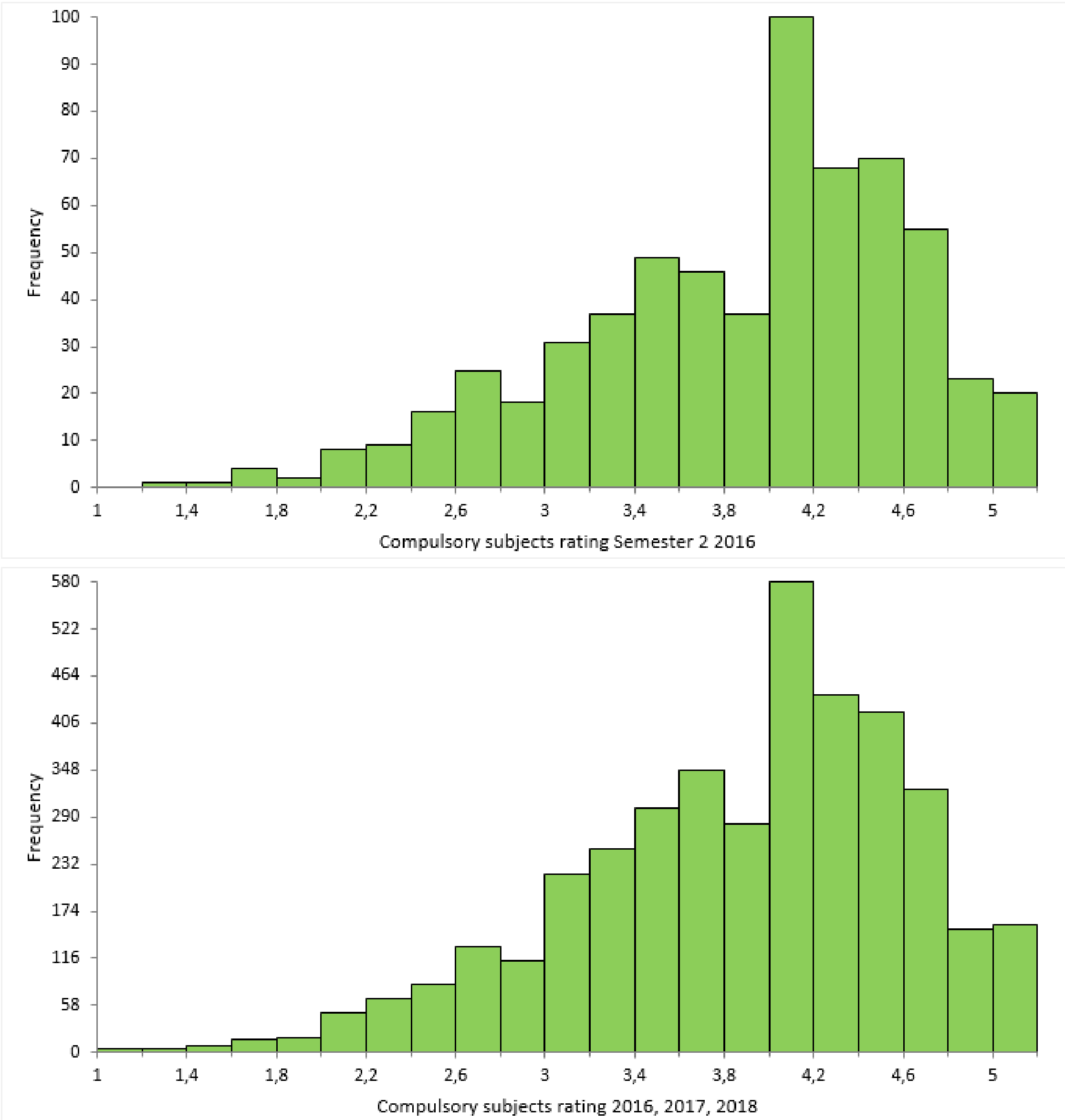
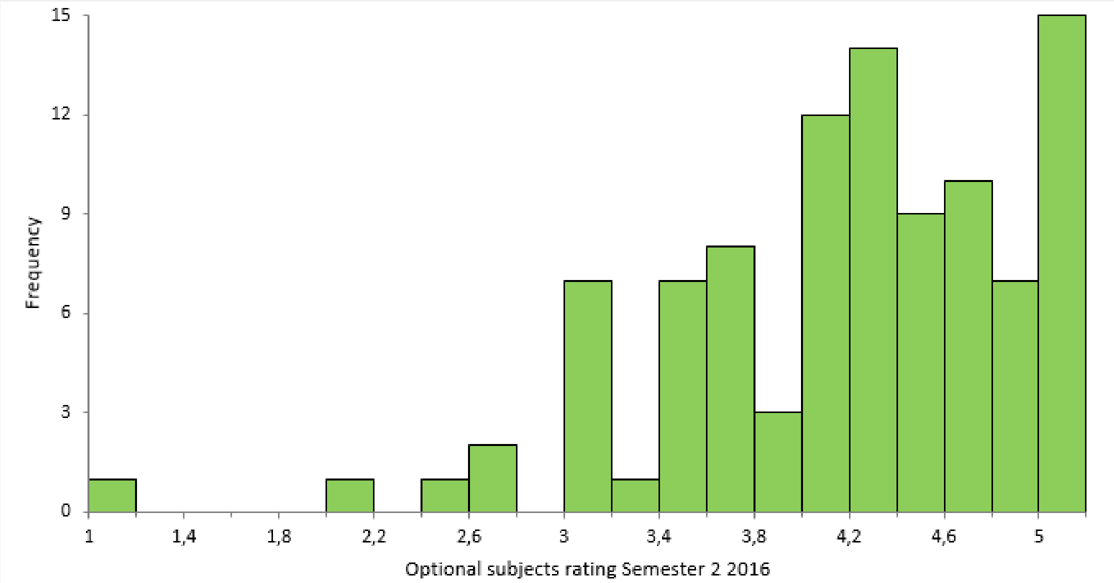


Figure 10 S2 2016 vs 2016, 2017, 2018 Compulsory subjects histogram comparison



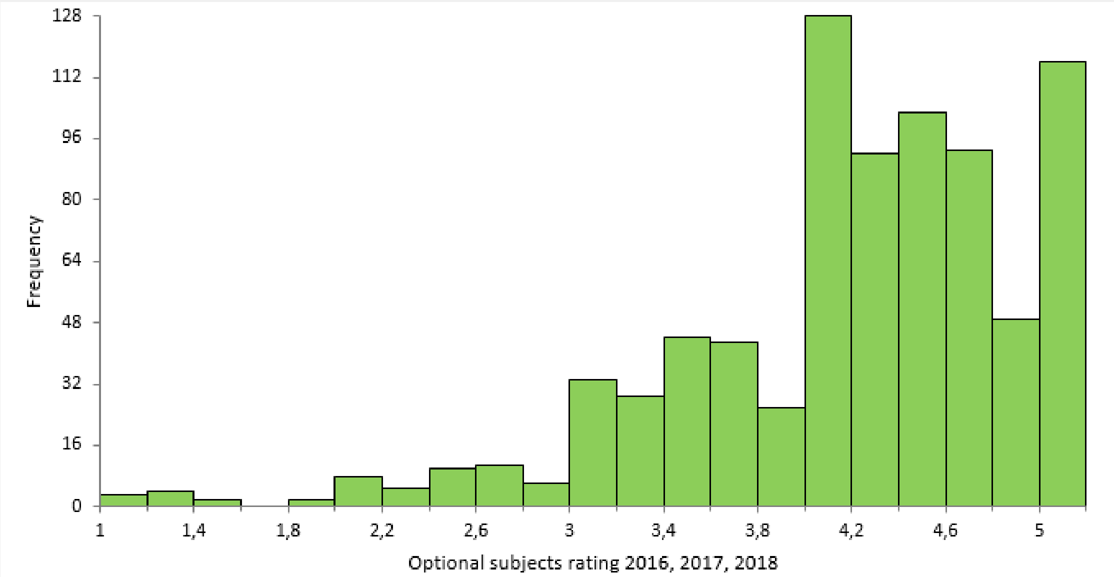


Figure 11 S2 2016 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of the Semester 2 2016 are almost identical to the histograms of the whole dataset.

Hypothesis testing

Comparison of means between Semester 2 2016 and 2016, 2017, 2018:

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 2 2016	3,85	4,14	0,29
2016, 2017, 2018	3,85	4,15	0,3

Table 15 S2 2016 vs 2016, 2017, 2018 means comparison

The means of Semester 2 2016 are almost identical to the general means of 2016, 2017, 2018.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,7387	Variances are equal

Table 16 S2 2016 normality and homogeneity of variances test results

As is it already known, the distributions are non-normal, but their variances are equal.

Student's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Student's t	3,85	4,14	0,0004	0,29
Wilcoxon-Mann-Whitney			<0,0001	0,30

Table 17 S2 2016 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different, being the mean of the optional subjects higher**, by an amount of approximately 0,3 points. As the two tests provide similar results, the results are considered acceptable.

Semester 2 2016 data behaves similarly to the whole dataset.

5.1.4 Semester 1 2017

Data of Semester 1 2017 is composed of 884 datapoints, 732 corresponding to compulsory subjects, and 152 corresponding to optional subjects.

Descriptive statistics analysis

Histogram comparison:

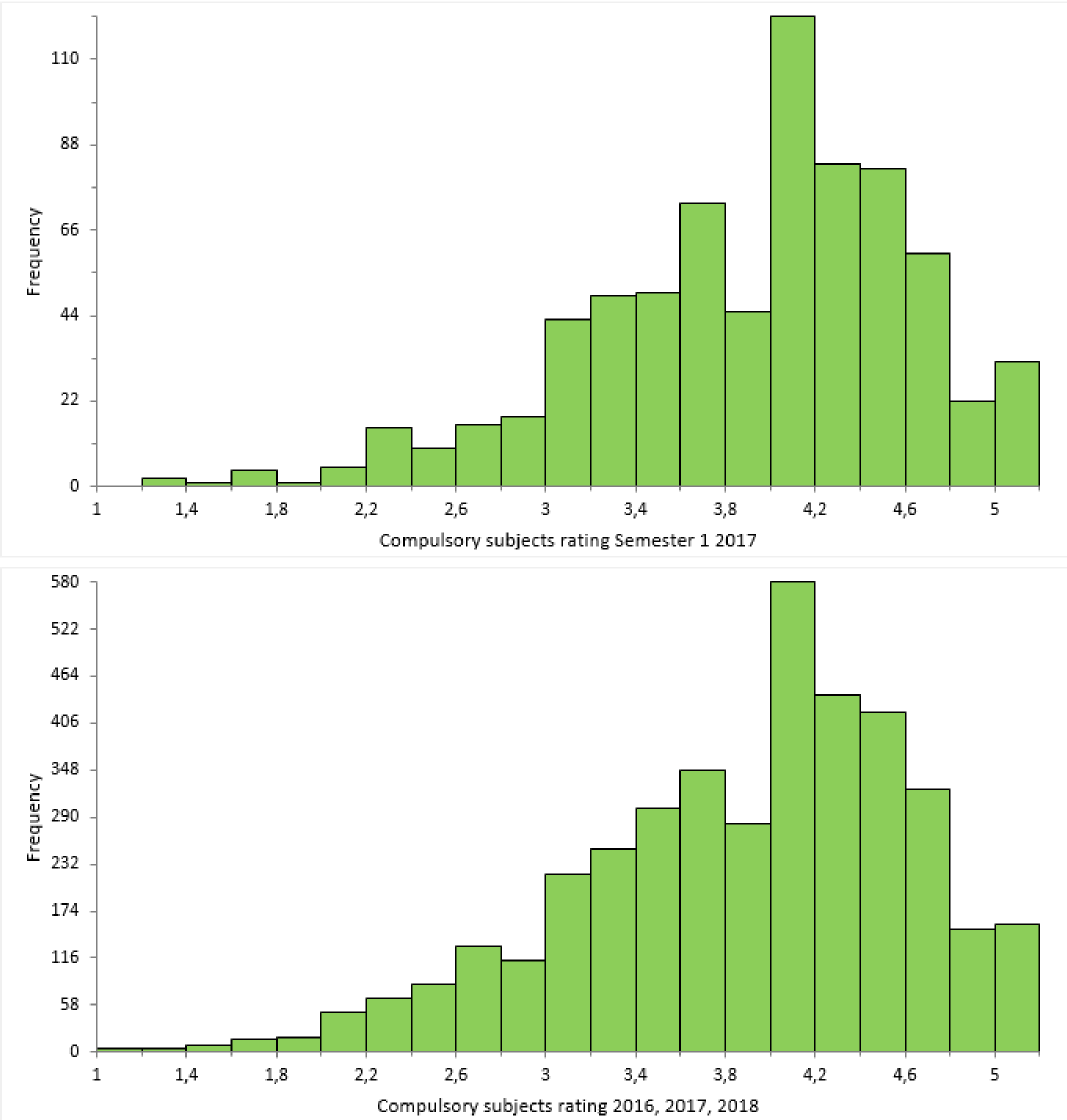


Figure 12 S1 2017 vs 2016, 2017, 2018 Compulsory subjects histogram comparison

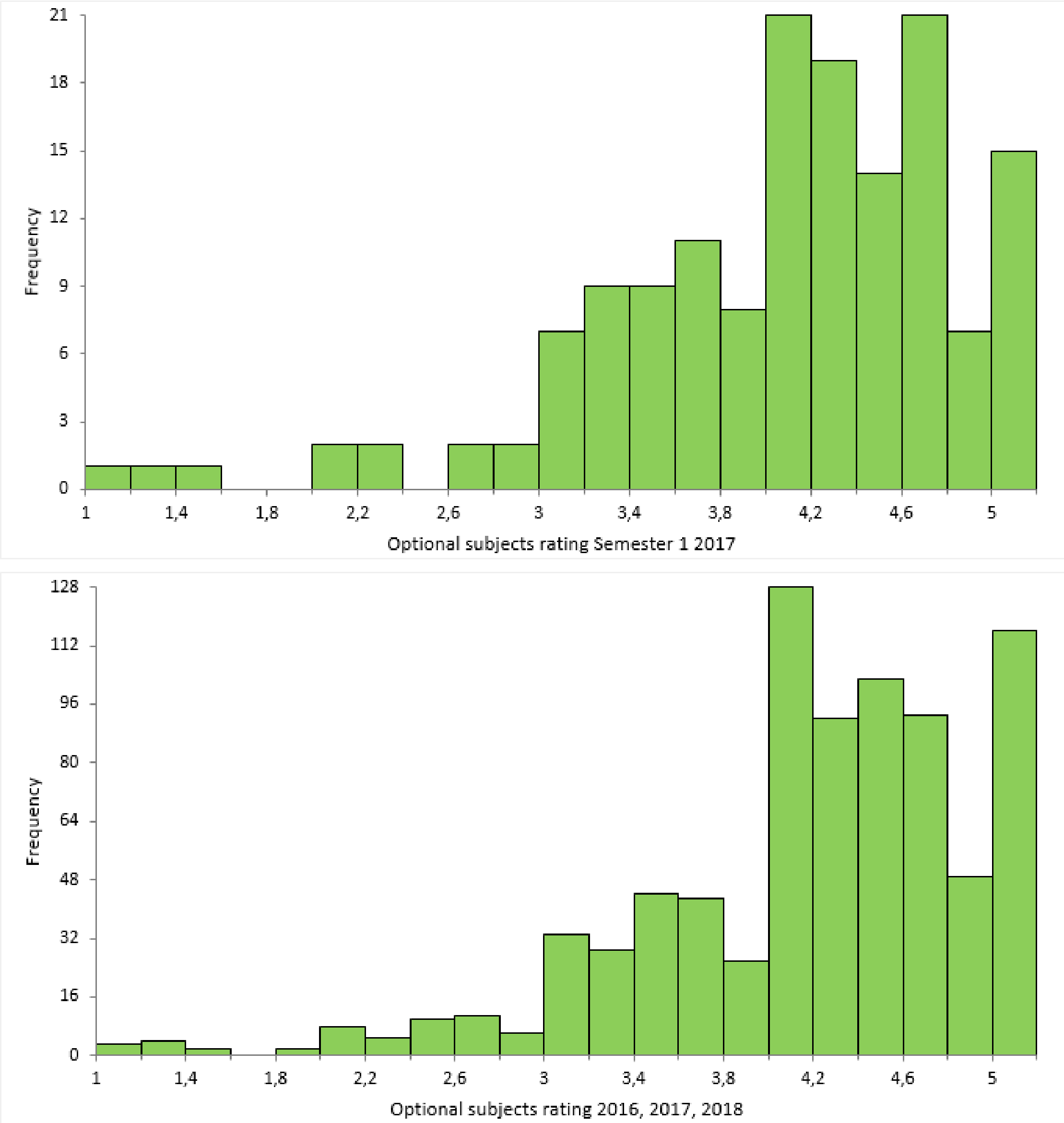


Figure 13 S1 2017 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of Semester 1 2017 are almost identical to the histograms of the whole dataset.

<u>Hypothesis testing</u>

Comparison of means between S1 2017 and 2016, 2017, 2018:

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 1 2017	3,89	4,04	0,15
2016, 2017, 2018	3,85	4,15	0,3

Table 18 S1 2017 vs 2016, 2017, 2018 means comparison

In Semester 1 2017 the optional subjects mean is also higher than the compulsory subjects mean.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,2871	Variances are equal

Table 19 S1 2017 normality and homogeneity of variances test results

As is it already known, the distributions are non-normal, but their variances are equal.

Student's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Student's t	3,89	4,04	0,0220	0,15
Wilcoxon-Mann-Whitney			0,0039	0,17

Table 20 S1 2017 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different, being the mean of the optional subjects higher,** by an amount of approximately 0,15 points. As the two tests provide similar results, the results are considered acceptable.

Semester 1 2017 data behaves similarly to the whole dataset.

5.1.5 Semester 2 2017

Data of Semester 2 2017 is composed of 787 datapoints, 652 corresponding to compulsory subjects, and 135 corresponding to optional subjects.

<u>Descriptive statistics analysis</u>

Histogram comparison:

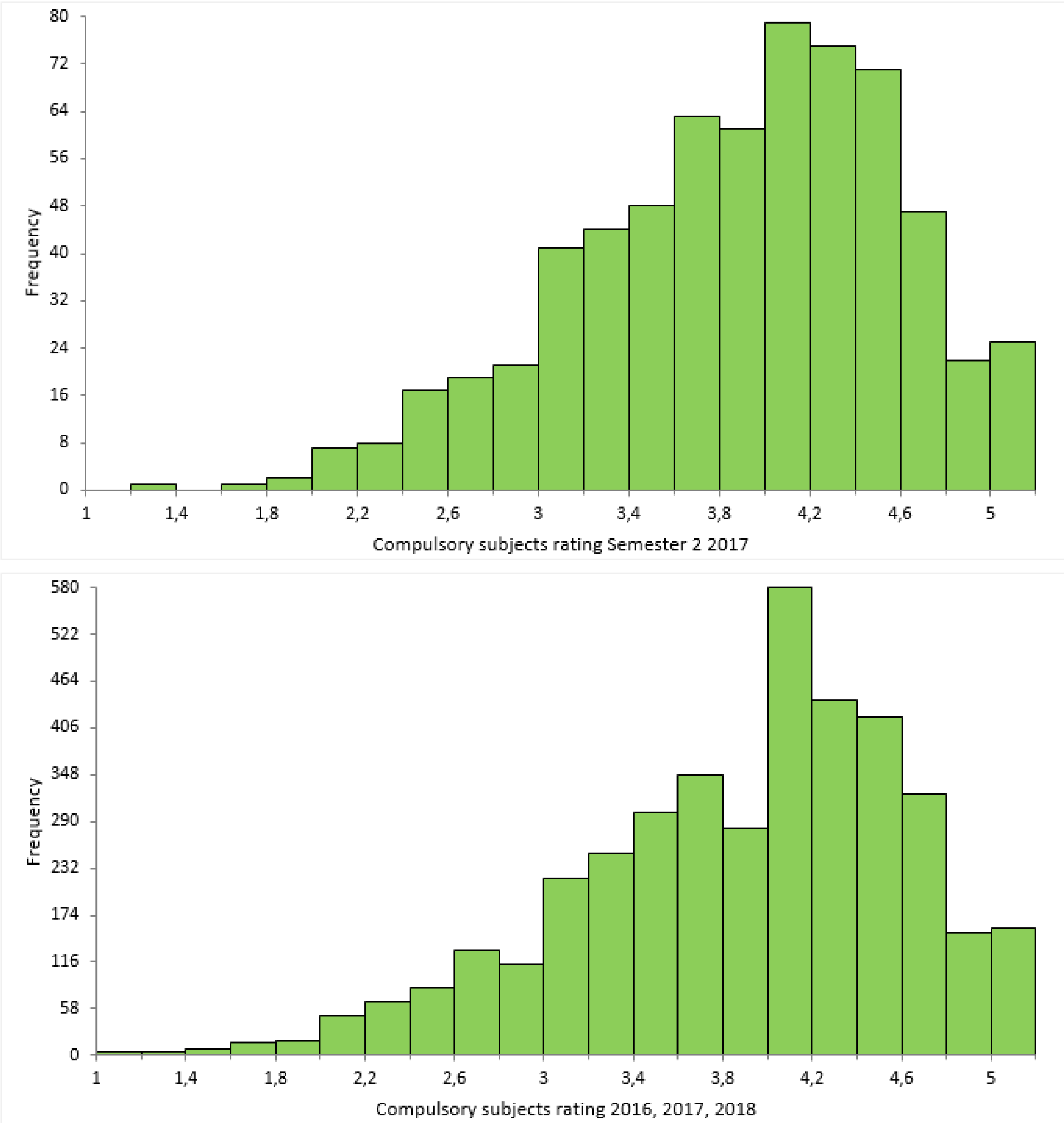


Figure 14 S2 2017 vs 2016, 2017, 2018 Compulsory subjects histogram comparison

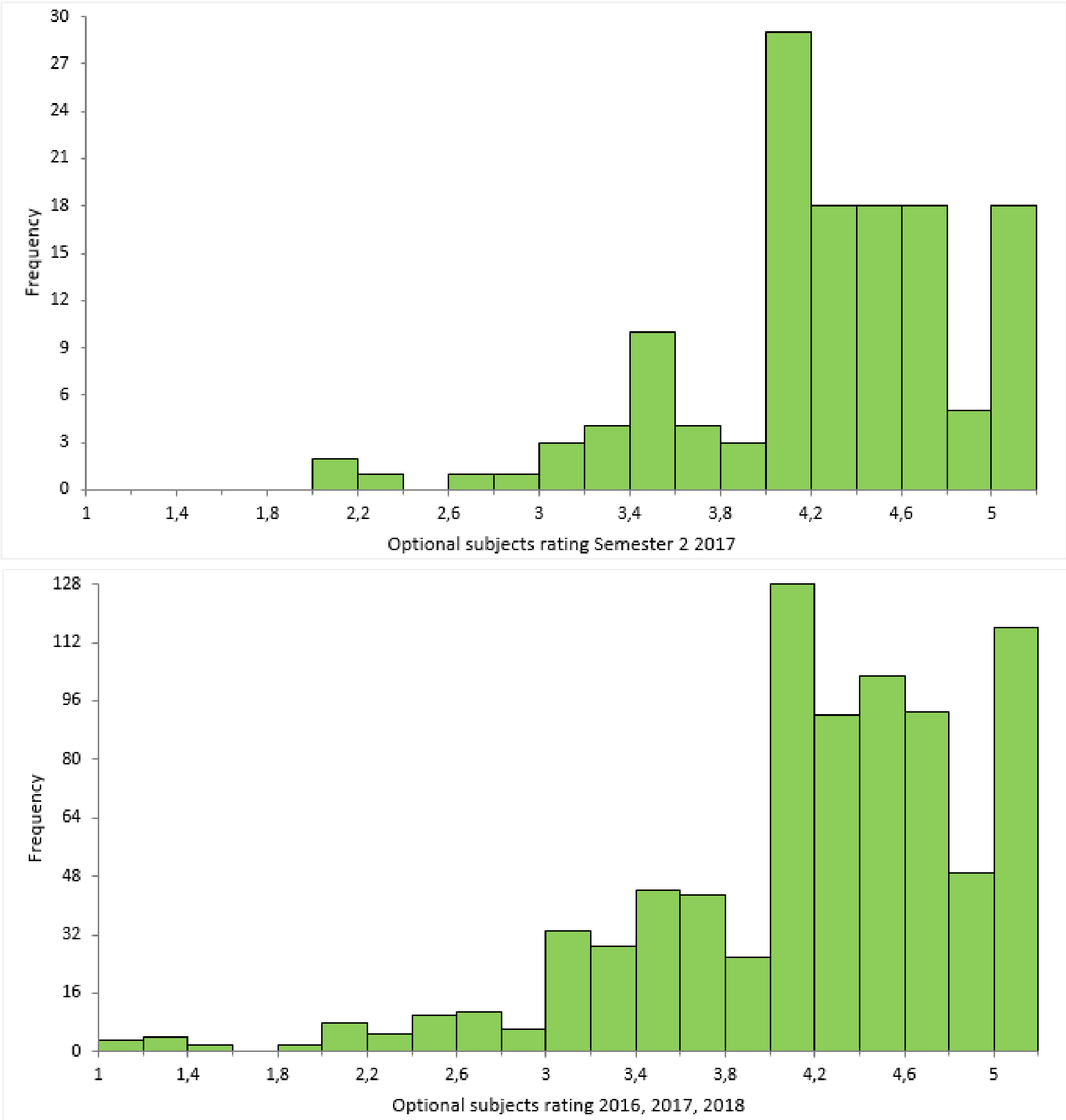


Figure 15 S2 2017 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of the Semester 2 2017 are almost identical to the histograms of the whole dataset.

<u>**Hypothesis testing**</u>

Comparison of means between S2 2017 and 2016, 2017, 2018:

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 2 2017	3,85	4,21	0,36
2016, 2017, 2018	3,85	4,15	0,3

Table 21 S2 2017 vs 2016, 2017, 2018 means comparison

The means of Semester 2 2017 are similar to the general means of 2016, 2017, 2018. The mean of optional subjects is higher than compulsory ones.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,0137	Variances are not equal

Table 22 S2 2017 normality and homogeneity of variances test results

As is it already known, the distributions are non-normal, and in this case, the variances are also not equal. In this case, instead of the Student's t, the Welch's t-test, which is a test designed for unequal variances (but still has the assumption of a normal distribution, this is why the Wilcoxon-Mann-Whitney test will also be executed), will be used.

Welch's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Welch's t	3,85	4,21	<0,0001	0,36
Wilcoxon-Mann-Whitney			<0,0001	0,35

Table 23 S2 2017 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different**, being the mean of the optional subjects higher, by an amount of approximately 0,35 points. As the two tests provide similar results, the results are considered acceptable.

Semester 2 2017 data behaves similarly to the whole dataset.

5.1.6 Semester 1 2018

Data of Semester 1 2018 is composed of 888 datapoints, 742 corresponding to compulsory subjects, and 146 corresponding to optional subjects.

Descriptive statistics analysis

Histogram comparison:

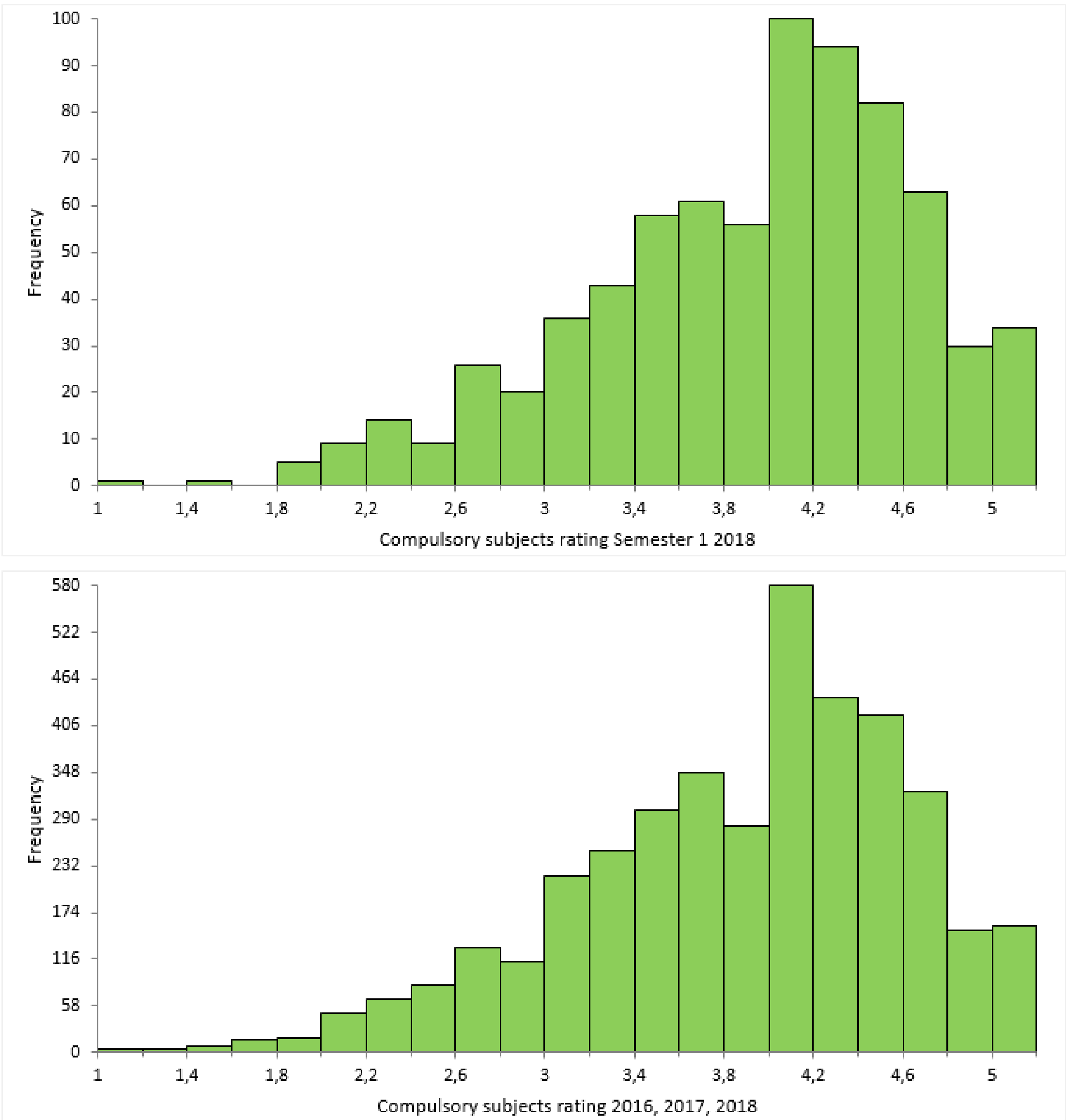
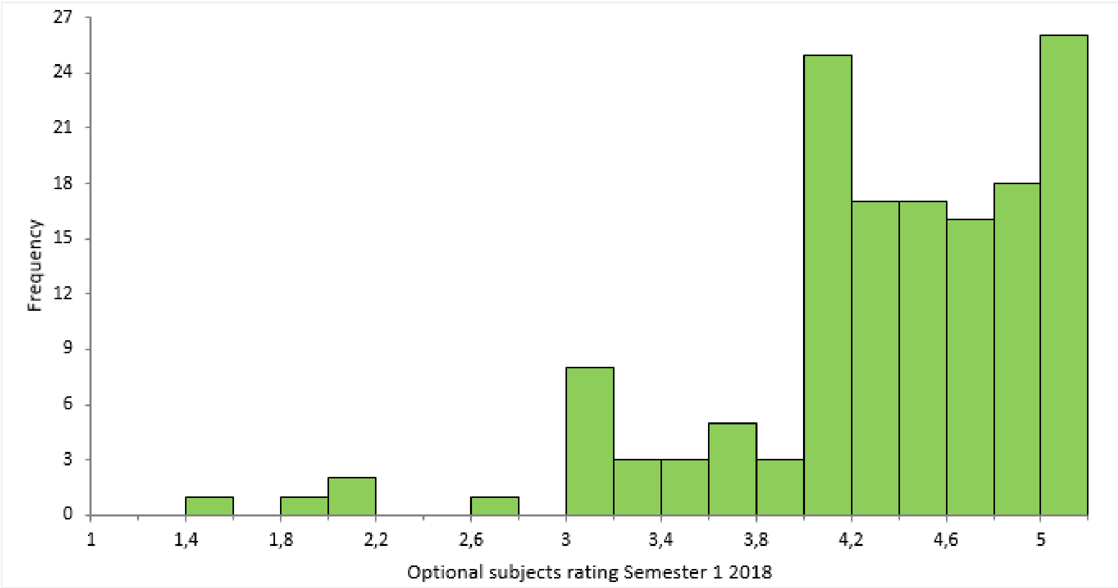


Figure 16 S1 2018 vs 2016, 2017, 2018 Compulsory subjects histogram comparison



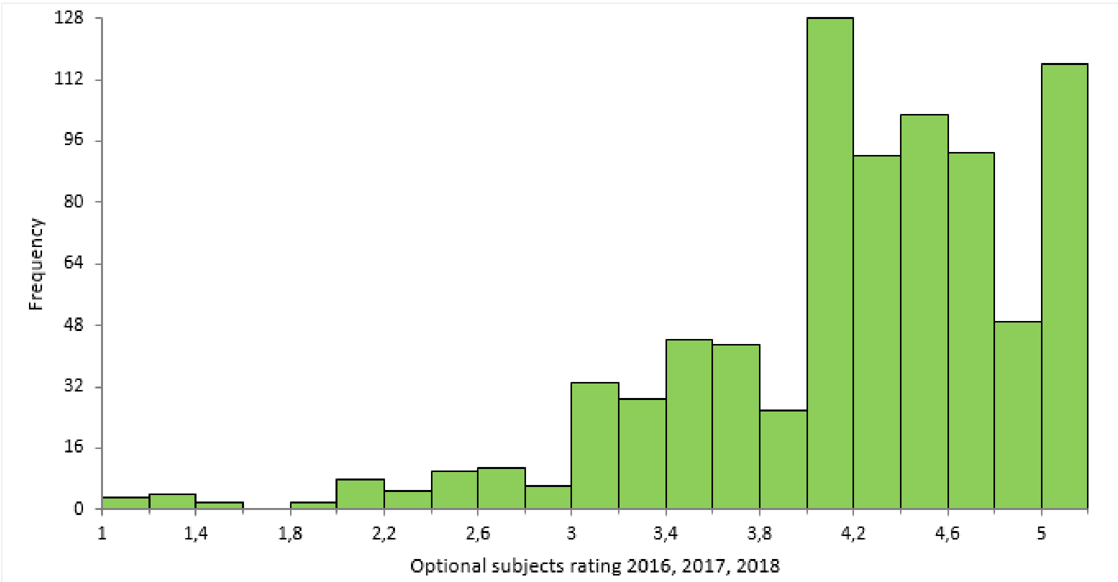


Figure 17 S1 2018 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of Semester 1 2018 are almost identical to the histograms of the whole dataset.

Hypothesis testing

Comparison of means between S1 2018 and 2016, 2017, 2018:

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 1 2018	3,90	4,29	0,39
2016, 2017, 2018	3,85	4,15	0,3

Table 24 S1 2018 vs 2016, 2017, 2018 means comparison

The means of Semester 1 2018 are similar to the general means of 2016, 2017, 2018. The mean of the optional subjects is higher than the compulsory ones.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,1020	Variances are equal

Table 25 S1 2018 normality and homogeneity of variances test results

As is it already known, the distributions are non-normal, but their variances are equal.

Student's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Student's t	3,90	4,29	<0,0001	0,39
Wilcoxon-Mann-Whitney			<0,0001	0,39

Table 26 S1 2018 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different**, being the mean of the optional subjects higher, by an amount of approximately 0,39 points. As the two tests provide similar results, the results are considered acceptable.

Semester 1 2018 data behaves similarly to the whole dataset.

5.1.7 Semester 2 2018

Data of Semester 2 2018 is composed of 702 datapoints, 579 corresponding to compulsory subjects, and 123 corresponding to optional subjects.

Descriptive statistics analysis

Histogram comparison:

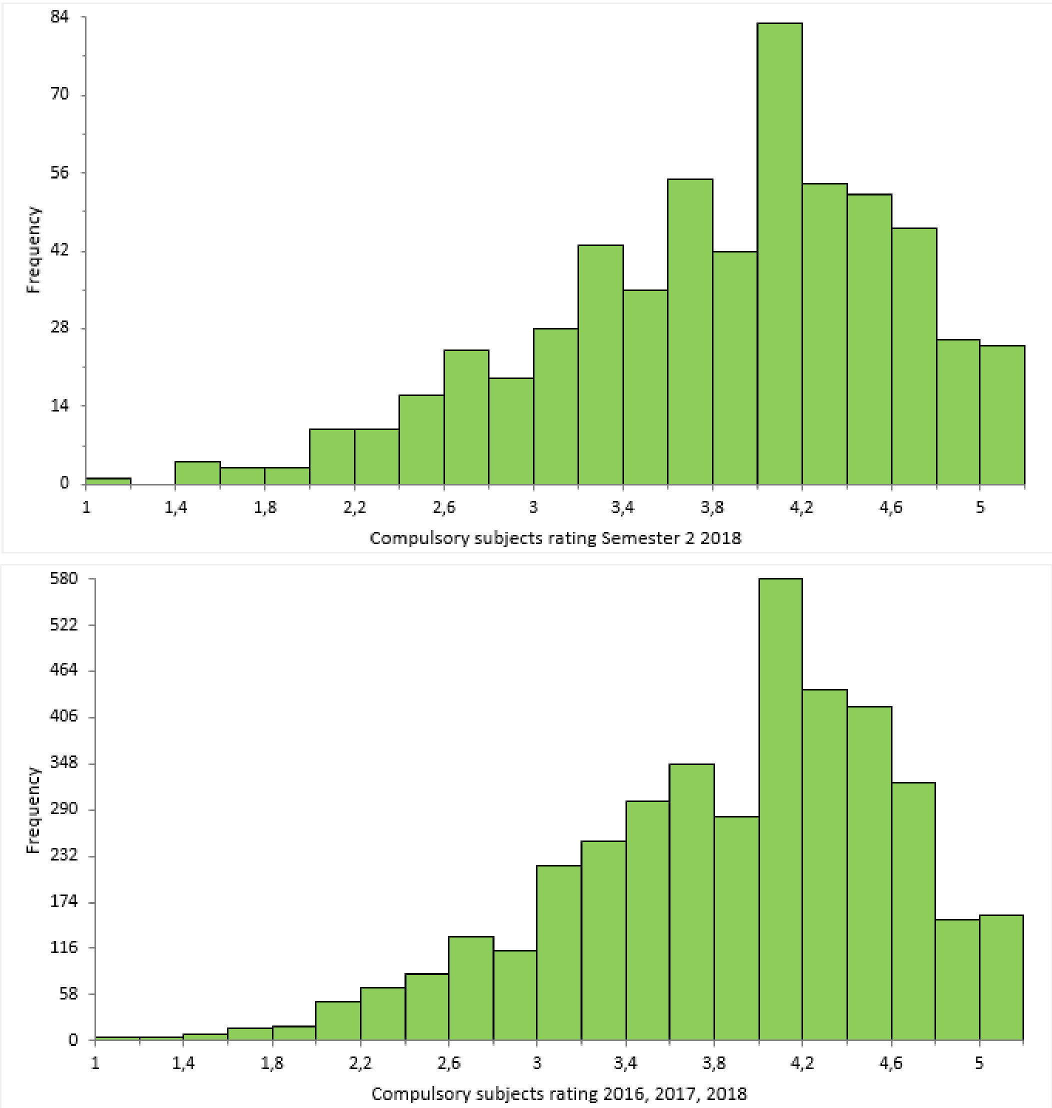


Figure 18 S2 2018 vs 2016, 2017, 2018 Compulsory subjects histogram comparison

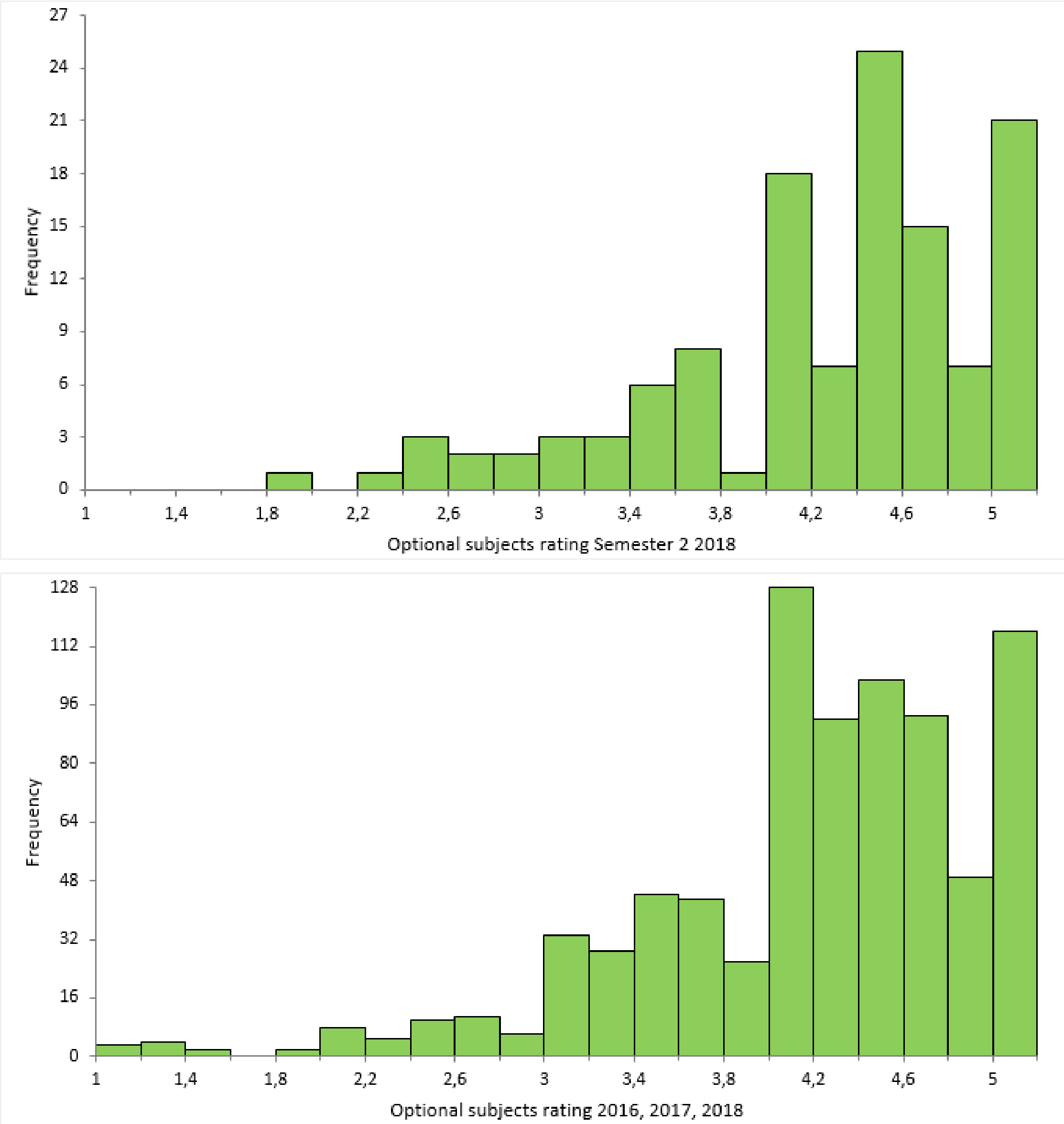


Figure 19 S2 2018 vs 2016, 2017, 2018 Optional subjects histogram comparison

The histograms of Semester 2 2018 are almost identical to the histograms of the whole dataset.

Hypothesis testing

Comparison of means between S2 2018 and 2016, 2017, 2018:

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
Semester 2 2018	3,80	4,22	0,42
2016, 2017, 2018	3,85	4,15	0,3

Table 27 S2 2018 vs 2016, 2017, 2018 means comparison

The means of Semester 2 2018 are similar to the general means of 2016, 2017, 2018. The mean of the optional subjects is higher than the compulsory ones.

Normality and homogeneity of variances tests:

Test	Data	p-value	Result
Anderson-Darling	Compulsory subjects	<0,0001	Non-normal
Anderson-Darling	Optional subjects	<0,0001	Non-normal
Brown-Forsythe	Type of subject	0,0611	Variances are equal

Table 28 S2 2018 normality and homogeneity of variances test results

As is it already known, the distributions are non-normal, but their variances are equal.

Student's t and Wilcoxon-Mann-Whitney tests:

Test	Compulsory subjects mean	Optional subjects mean	p-value	Mean or location difference
Student's t	3,80	4,22	<0,0001	0,42
Wilcoxon-Mann-Whitney			<0,0001	0,42

Table 29 S2 2018 Type of subject hypothesis tests results

The results of the two hypothesis tests show that **there is enough statistical evidence to conclude that the mean of the two populations is different**, being the mean of the optional subjects higher, by an amount of approximately 0,42 points. As the two tests provide similar results, the results are considered acceptable.

Semester 2 2018 data behaves similarly to the whole dataset.

5.1.8 Comparison of means between the different periods

As was noted during the analysis conducted in Semester 1 2016, although all of the studied periods have similar behavior (all of them have similar histograms and in all of them the mean of the optional subjects is statistically significantly higher than the compulsory subjects mean), there is some difference in some of the values compared to the whole dataset. In this Section, these differences will be analyzed.

The different means obtained in all the periods studied are presented in the following table.

Period	Compulsory subjects mean	Optional subjects mean	Difference of means
S1 2016	3,82	4,01	0,19
S2 2016	3,85	4,14	0,29
S1 2017	3,89	4,04	0,15
S2 2017	3,85	4,21	0,36
S1 2018	3,9	4,29	0,39
S2 2018	3,8	4,22	0,42

Table 30 Subject type means in the different studied periods

As can be seen, the means of some of the periods seem significantly[5] different from the rest, especially in the optional subjects' means. For this reason, these differences will be analyzed statistically with hypothesis testing. The maximum and minimum of the means of compulsory and optional subjects will be tested to determine if the differences are statistically significant.

[5] Not in statistical terms. The statistical significance of this differences will be studied in this Section.

	Compulsory subjects	Optional subjects
Period of minimum value	S2 2018	S1 2016
Minimum value	3,8	4,01
Period of maximum value	S1 2018	S1 2018
Maximum value	3,9	4,29
Brown-Forsythe p-value	0,560	0,0601
Student's p-value	0,0139	0,0024
Wilcoxon-Mann-Whitney p-value	0,0250	0,0025
Difference of means (Student)	0,1	0,27
Difference of locations (WMW)	0,09	0,23

Table 31 Type of subject minimum and maximum means hypothesis test results

The test results conclude that there is enough statistical evidence (with a 5% significance) to conclude that the maximum and minimum means of both compulsory and optional subjects are different. However, it must be noted that the obtained p-values are much bigger than p-values obtained in the different analysis before (most of them were lower than 0,0001). This holds especially true for the compulsory subjects, for which the obtained p-values with a significance of 1% would conclude that there is no difference in the means.

To further understand this difference, the means of the different periods are plotted in a line graph.

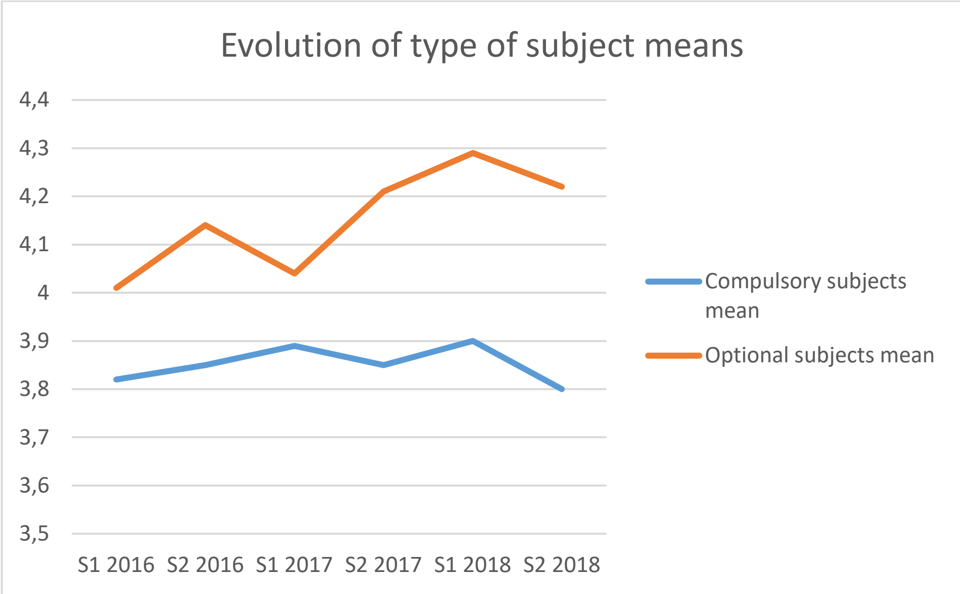


Figure 20 Evolution of type of subject means line graph

The line graph shows that the mean of the compulsory subjects remains almost constant (and that is why the detected difference was lower and the p-values greater) and the optional subjects mean seems to increase in a trend throughout the analyzed period.

If the difference of the means is plotted, a similar trend, explained by the compulsory subjects remaining constant and the optional subjects increasing, is also observed.

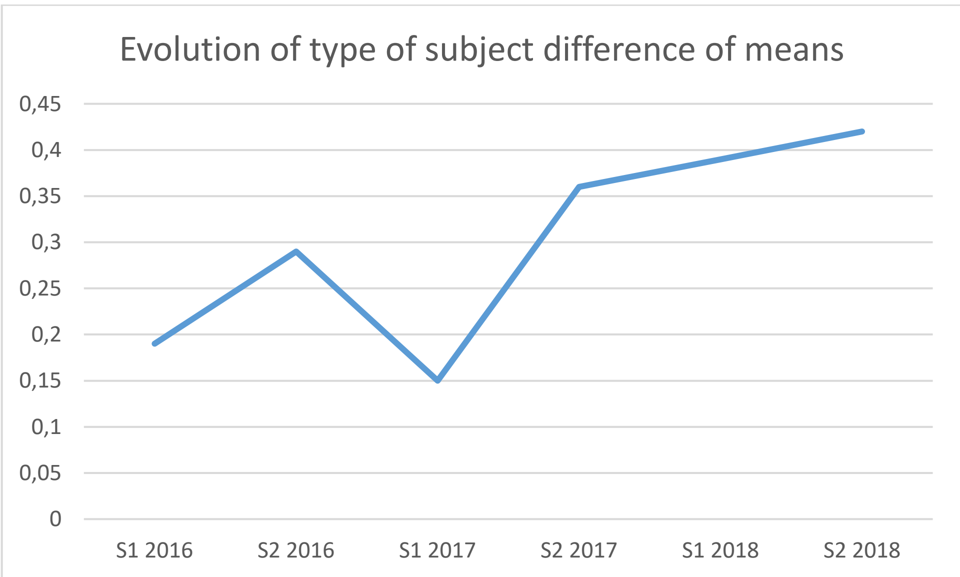


Figure 21 Evolution of type of subject difference of means

To analyze if there is a positive correlation in the two observed trends, linear regression tests are used.

Compulsory subjects mean linear regression:

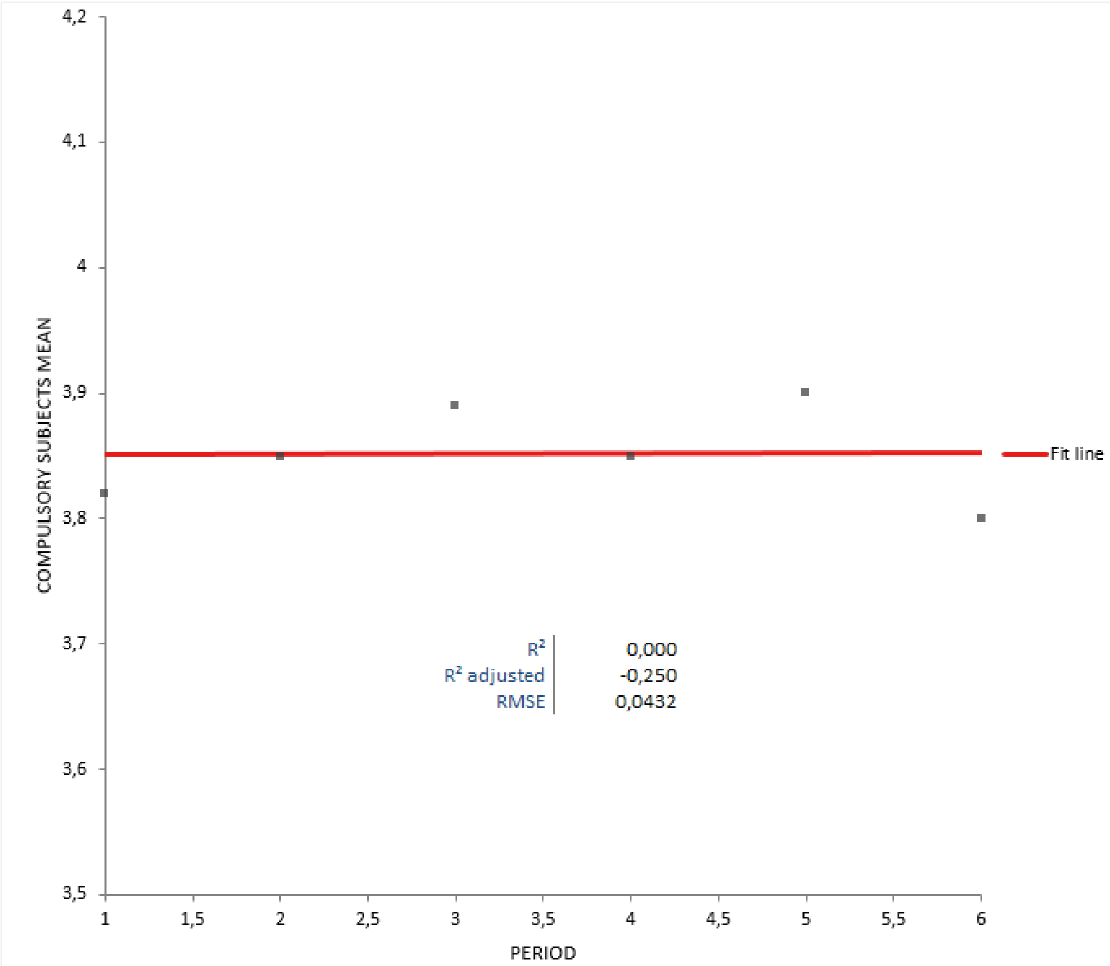


Figure 22 Compulsory subjects mean linear regression

The R^2 value equals zero, so there is no positive correlation and the values follow a random path around a constant value, so the mean of the compulsory subjects does not have an increasing trend in the studied period.

Optional subjects mean linear regression:

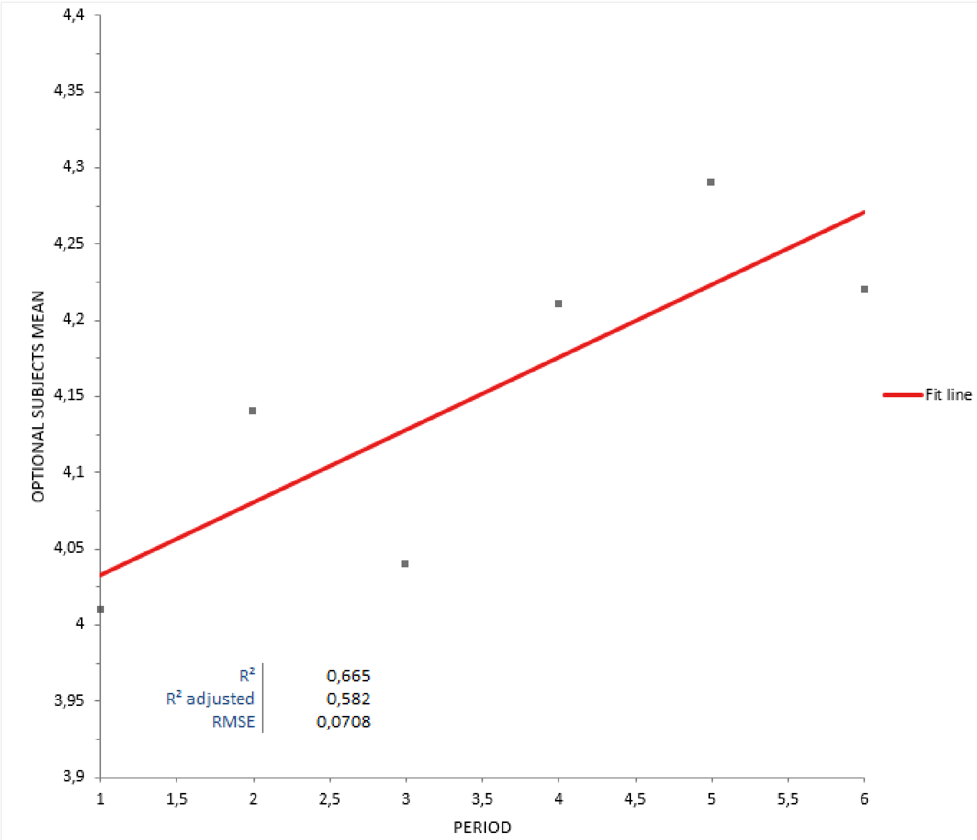


Figure 23 Optional subjects mean linear regression

With an R^2 of 0,665, there is a considerable correlation, so the mean of the compulsory subjects increased in a kind of linear fashion during the studied period, that does not necessarily mean that it will continue to increase, but it explains why there is a trend in the difference of the means of the two populations.

In conclusion, the evidence shows that **the difference in the means between compulsory and optional subjects increased in a kind of linear fashion during the studied period** because the compulsory subjects mean remained almost constant and the optional subjects mean increased in a kind of linear way during the period. This means that, in general, **the results of the surveys have increased in a kind of linear fashion during the period 2016 - 2018 in ESEIAAT** and explains the differences detected during the analysis of the different periods' data. However, despite this trend, **the mean of the optional subjects was always statistically significantly higher than the mean of the compulsory subjects in all the studied periods**.

5.1.9 Worst-case comparison

To determine if the statistical differences found in the previous analysis affect the teacher surveys empirically, worst-case comparisons are made.

According to the previous analysis, a worst-case for a teacher would be to only teach compulsory subjects, and a best-case would be to only teach optional subjects. Therefore, if the previous inferences are true, the teachers who only teach optional subjects will have a better rating than the teachers who only teach compulsory subjects.

In ESEIAAT, during the 2016-2018 period, 289 teachers taught only compulsory subjects and 41 teachers taught only optional subjects.

Type of teacher	Number of teachers	Mean rating
Teaches only compulsory subjects	289	3,81
Teaches only optional subjects	41	4,11

Table 32 Type of subject worst-case comparison

It appears to be a difference between the means of the two populations. As the previous statistical inferences pointed, teachers who teach only optional subjects appear to be better rated than teachers who teach only compulsory subjects.

To prove this assumption, hypothesis testing is executed on the data.

Data	Test	p-value	Mean or location difference	Conclusion
Only compulsory	Anderson-Darling	<0,0001[6]	-	Non-normal
Only optional	Anderson-Darling	0,0070[6]	-	Non-normal
Both	Brown-Forsythe	0,5802	-	Equal variances
Both	Student's t	0,0078	0,29	Significantly different
Both	Wilcoxon-Mann-Whitney	0,0036	0,30	Significantly different

Table 33 Type of subject worst-case comparison hypothesis testing

[6] As was noted in the previous analysis, this kind of data is non-normal.

The results of the two hypothesis tests show that there is enough statistical evidence to conclude that the means of the two populations are different, being the mean of the optional subjects higher, by an amount of approximately[7] 0,3 points.

This allows us to conclude that **teachers who only teach optional subjects are generally better rated than teachers who only teach compulsory subjects**.

The distribution of teachers in ESEIAAT according to their percentage of optional subject surveys respect their total surveys[8] is as follows (the frequency of 0% is cropped to allow the correct visualization of the rest of the distribution):

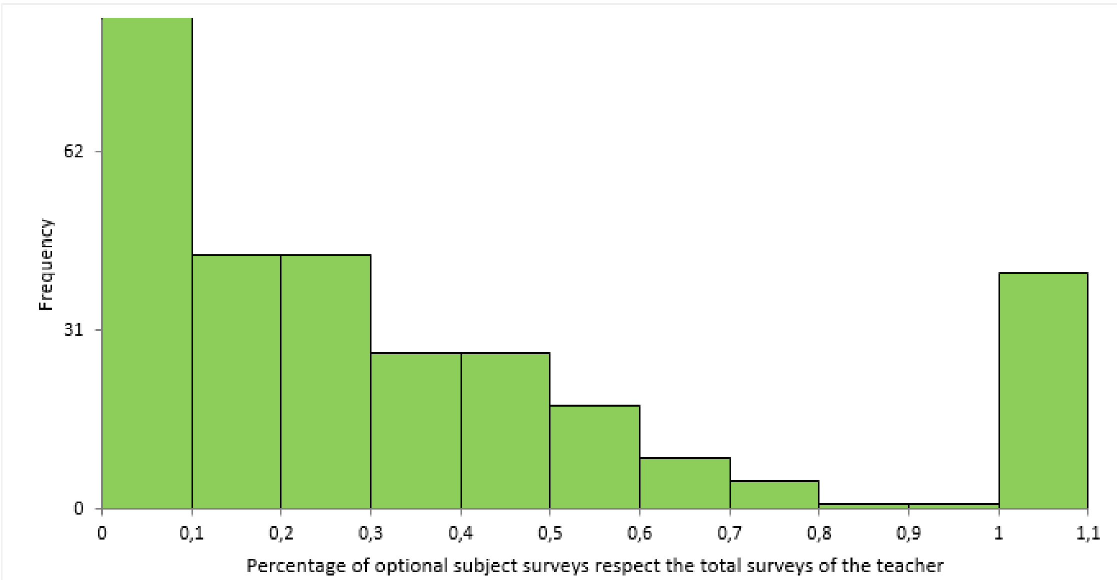


Figure 24 Percentage of optional subject surveys respect the total surveys of the teacher histogram

In this worst-case testing, the two extreme conditions have been tested, but what happens between these two extremes has not been analyzed yet.

For that, the difference between teachers who only teach compulsory subjects and teachers who also teach some optional subjects will be tested. Teachers whose optional percentage is higher than 0% and less than 70% will establish this second population, in this way we set aside the optional subjects extreme condition (100%). The cutoff is established at the 70% mark because between 70%- 90% the sample of data is small and conclusions drawn from the analysis will not be representative of that population.

---

[7] Student's t test 95% confidence interval: 0,08 - 0,5

[8] Optional percentage = $\frac{Optional\ subjects\ surveys\ of\ the\ teacher}{Total\ surveys\ of\ the\ teacher} \cdot 100$

In ESEIAAT during 2016-2018, 289 teachers only taught compulsory subjects and 189 teachers who also taught some optional subjects[9].

Type of teacher	Number of teachers	Mean rating
Teaches only compulsory subjects	289	3,81
Teaches some optional subjects	189	3,92

Table 34 Only compulsory subjects vs some optional subjects teachers comparison

As it appears in the table above, the difference between the mean of the two populations is not as big as it was comparing the extremes.

Data	Test	p-value	Mean or location difference	Conclusion
Only compulsory	Anderson-Darling	<0,0001	-	Non-normal
Some optional	Anderson-Darling	0,0005	-	Non-normal
Both	Brown-Forsythe	0,0005	-	Different variances, Welch's t-test will be used.
Both	Welch's t	0,0484	0,11	Statistically significant
Both	Wilcoxon-Mann-Whitney	0,1713	0	Not statistically significant

Table 35 Only compulsory subjects vs some optional subjects teachers hypothesis testing

The obtained results are not conclusive. Welch's t-test results classify the difference as statistically significant and Wilcoxon-Mann-Whitney classifies it as not statistically significant. However, as the Welch's t p-value is almost at the significance level (0,0484 vs 0,0500) and the obtained results with the Wilcoxon-Mann-Whitney (which is robust against non-normality, unlike the Welch's t) conclude that the difference is not statistically significant, we will conclude that **the means of teachers who only taught compulsory subjects and teachers who also taught some optional subjects are not significantly different**.

As a conclusion of this Section, there is enough evidence to conclude that **teachers who only teach optional subjects are generally better rated than the rest of the teachers**.

[9] According to the previous definition of this population.

5.1.10 Difference of means between subject types in the same teacher

All the analysis carried until now concluded that teachers who only teach optional subjects are generally better rated than the rest of the teachers, but it did not demonstrate that this is due to the variable biasing the results of the survey. It could perfectly be, for example, that the best teachers are more involved in the optional subjects.

In this Section, we will assume that the same teacher should be evaluated similarly no matter the type of subject they teach, and if we can demonstrate that the same teacher is better rated in optional subjects than compulsory ones, this will demonstrate that the variable is indeed affecting the results of the survey.

For this analysis, the population of teachers who have been evaluated in the same number of compulsory and optional subjects surveys will be studied. In this way, the difference in the rating of the same teacher in the different types of subjects can be evaluated.

In ESEIAAT during the 2016-2018 period, there have been 15 teachers who have been evaluated the same number of times in the two types of subjects.

Teacher	Number of total surveys	Mean on compulsory subjects	Mean on optional subjects	Difference of means
PDI_108	12	3,40	4,25	0,85
PDI_127	6	4,68	4,72	0,04
PDI_129	10	4,04	4,40	0,36
PDI_165	6	4,68	4,18	-0,50
PDI_167	4	2,75	3,29	0,54
PDI_175	8	3,94	4,89	0,95
PDI_272	10	4,27	4,57	0,30
PDI_323	2	4,65	4,70	0,05
PDI_334	4	3,07	2,35	-0,72
PDI_413	4	2,58	4,83	2,25
PDI_415	4	3,54	4,10	0,56
PDI_428	8	4,40	4,58	0,18
PDI_441	10	2,64	4,09	1,45
PDI_471	4	4,10	4,51	0,41
PDI_65	4	4,39	4,00	-0,39

Table 36 Teachers who have been evaluated the same number of times in the two types of subjects

From this table, it can be seen that 12 out of the 15 teachers, which corresponds to 80% of this population, were better rated in the optional subjects than in the compulsory subjects they taught.

This same comparison will be extended to all teachers that were surveyed in the two types of subjects (no matter in what percentage).

Result of the comparison	Number of teachers	Percentage
Better rated in compulsory subjects	54	28%
Better rated in optional subjects	142	72%

Table 37 Comparison in the rating of the different type of subjects of teachers who taught compulsory and optional subjects

As can be seen in the summary table above, **72% of the teachers of ESEIAAT were better evaluated in their work in optional subjects**.

5.1.11 Conclusions

After all the analysis performed, there is enough evidence to conclude that:

- **The variable "type of subject" influences the results** of the teachers' survey.
- **Teachers who only teach optional subjects are generally better rated than the rest of the teachers.**
- However, this is not the case for teachers who teach some[10] optional subjects. **Teachers who teach some optional subjects are generally rated in the same way as teachers who only teach compulsory subjects**.
- Most usually[11], **the same teacher is better rated in his work in optional subjects than in compulsory ones**.

[10] In this study, this means less than 70% of their total subjects taught.
[11] In ESEIAAT during the period 2016-2018, this means 72% of the times.

5.2 Degree level (bachelor's/master's)

5.2.1 Analysis of all the dataset

The data of the different teacher satisfaction surveys conducted in 2016, 2017, and 2018 has been sorted according to if the teacher was evaluated in his work in a bachelor's or master's degree subject.

The final data is composed of 4778 datapoints, 2644 corresponding to compulsory subjects, and 2134 corresponding to optional subjects.

This variable is more evenly distributed than the type of subject, so there are almost the same surveys conducted in bachelor's and master's degrees subjects. This similarity in size between the two populations will help in its analysis.

Descriptive statistics analysis

Histogram comparison:

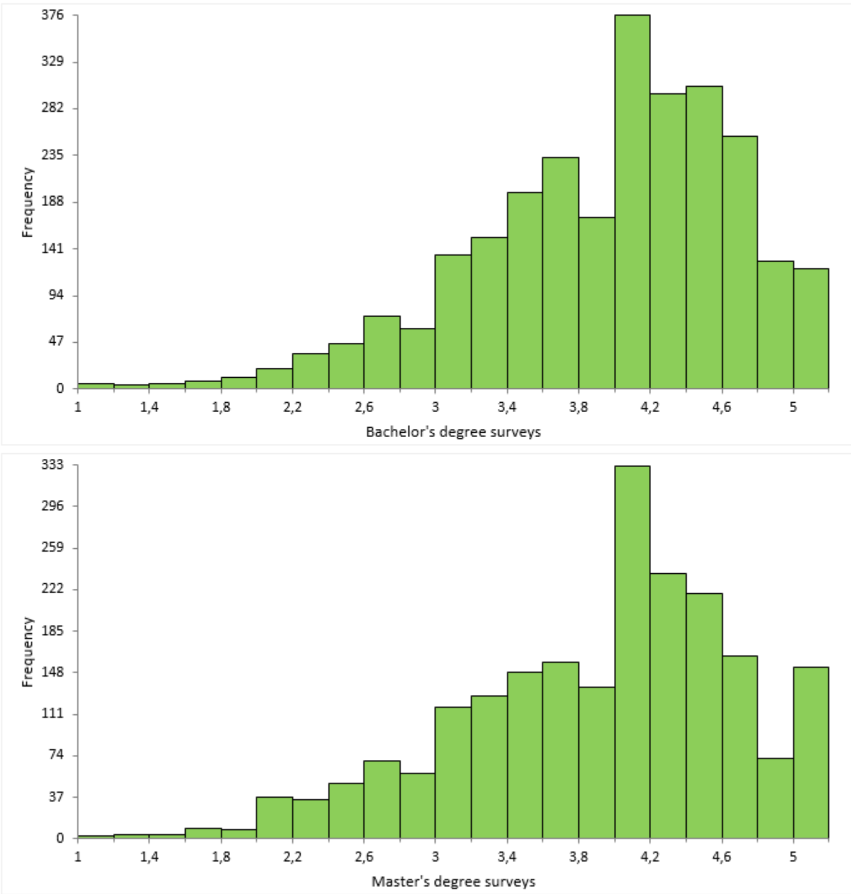


Figure 25 Bachelor's and master's surveys histogram comparison

The two histograms appear to be very similarly shaped, so, at least graphically, **there appears not to be a significant difference between the two surveys**.

Hypothesis testing

Comparison of means:

Period	Bachelor's degree mean	Master's degree mean	Difference of means
2016-2018	3,93	3,88	0,05

Table 38 Bachelor's vs Master's degree surveys means comparison

As was foretold in the graphical analysis, there seems to be no significant difference between the means of the two populations. In fact, the mean on bachelor's degrees surveys is higher (although only by 0,05 points) than the master's degrees mean, contradicting the hypothesis that in master's degrees teachers could be more positively evaluated. This difference could be random, so hypothesis testing is conducted.

Normality and homogeneity of variances tests:

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Bachelor's	<0,0001	-	Non-normal
Anderson-Darling	Master's	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,0215	-	Variances are not equal. Welch's t-test will be conducted.
Welch's t	Both	0,0306	0,05	Statistically significant
Wilcoxon-Mann-Whitney	Both	0,0646	0,03	Not statistically significant

Table 39 2016-2018 degree level hypothesis testing

The results of the two hypothesis tests contradicted, but the Welch's t p-value was near signification (0,0306 vs 0,0500) and the Wilcoxon-Mann-Whitney test p-value, which is robust against non-normality, is over the 5% so we will conclude that **the difference between the two means is not statistically significant**.

That poses a contradiction to the results previously shown in the other studies conducted on this topic. Studies nº2 and nº3 concluded that the degree level influenced the results of the satisfaction surveys. This contradiction could be caused by some of the following reasons:

- The studies nº2 and nº3 did not test the normality of the data nor the homogeneity of variances. They assumed that this kind of data is always normal, which is not the case, and did not test the homogeneity of variances. As they used the Student's t-test, the non-normality and possible heterogeneity of variances could have biased the results. However, the perceived differences were very different, in FIB there were perceived differences between bachelor's and master's surveys of 0,2 points, approximately, and in ESEIAAT the perceived difference is 0,05 points. This makes us believe that this reason could not be the only one responsible for the discrepancy.
- In the FIB there are not as many master's degrees taught than in ESEIAAT, the ratio of master's over bachelor's surveys in ESEIAAT is around 45% and in the FIB studied data it was around 20%. The smaller sample of master's surveys could have biased the results.
- It could just be that the influence of this variable is different in different schools or has changed over time (the other studies analyzed the 2010-2014 period).

This shows that **the influence of the different variables can be different depending on the School and period studied**. It can change over time and over the different Schools of the UPC. **The conclusions drawn from the present study cannot be extrapolated to other schools outside ESEIAAT or other periods outside 2016-2018.**

5.2.2 Analysis of the different semesters

Histogram comparison:

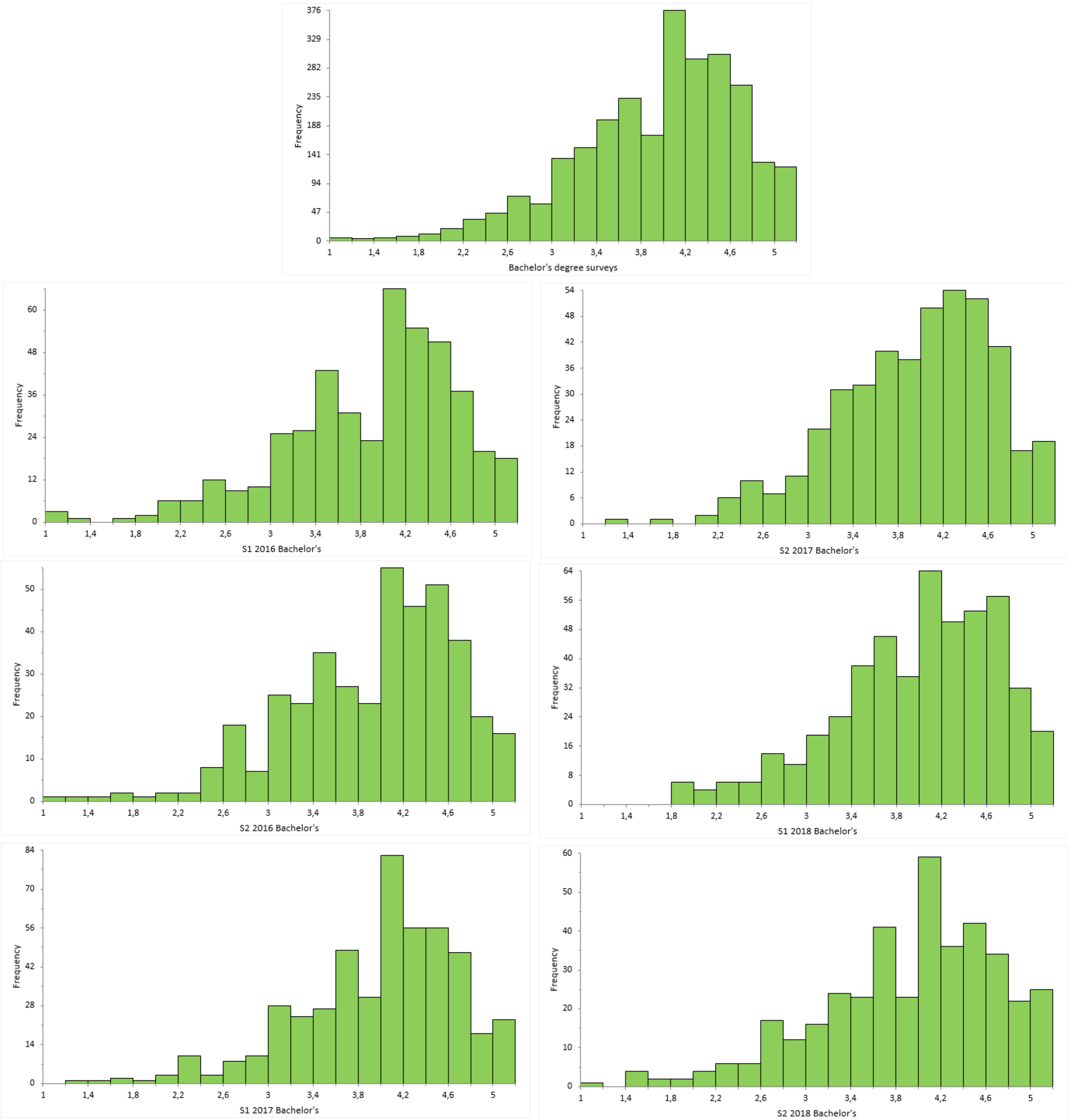


Figure 26 Different semesters bachelor's degree surveys histogram comparison

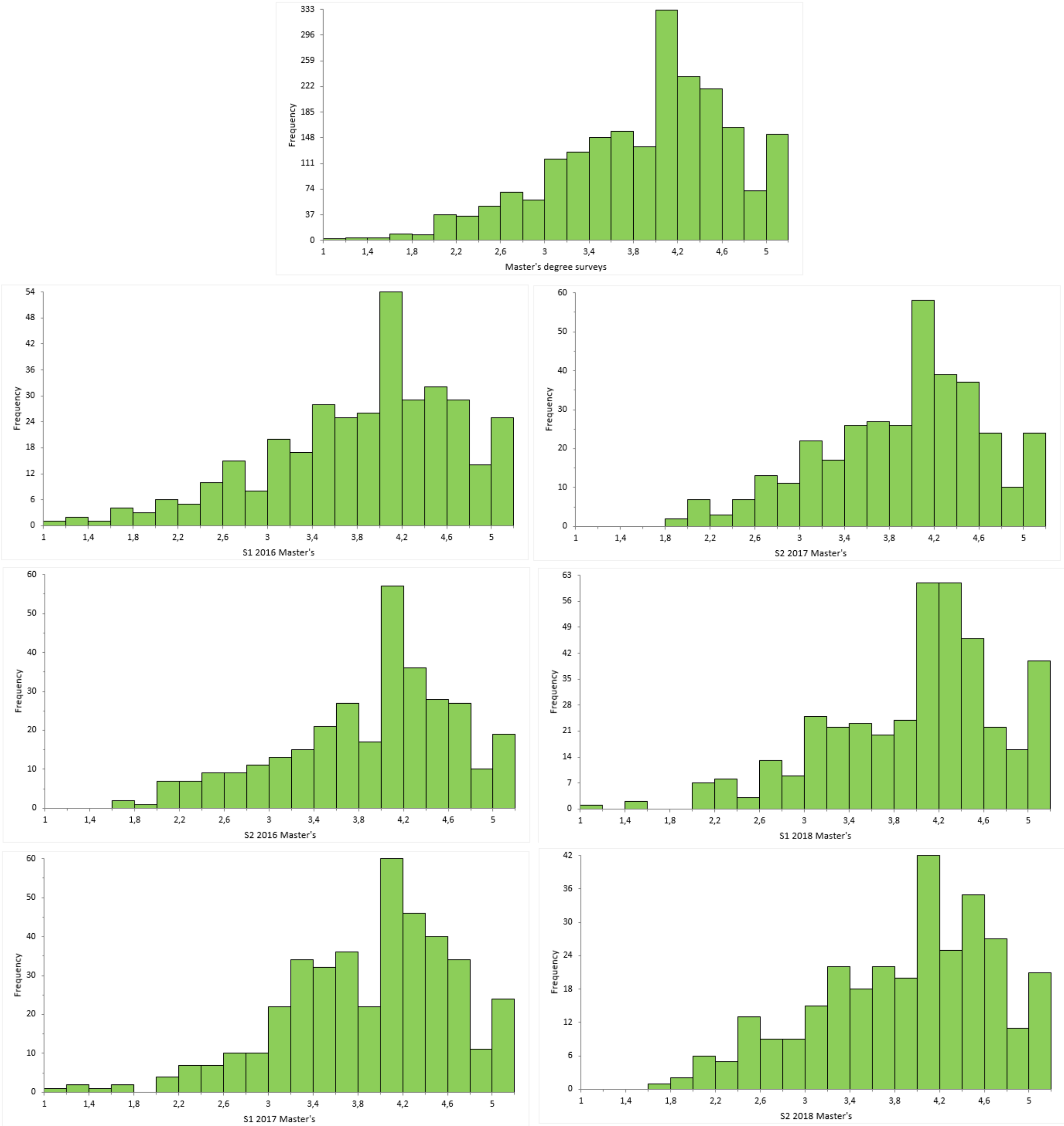


Figure 27 Different semesters master's degree surveys histogram comparison

Comparison of means:

Period	Bachelor's mean	Master's mean	Difference of means	WMW p-value	Result
Semester 1 2016	3,88	3,83	0,05	0,4882	Difference not significant
Semester 2 2016	3,91	3,86	0,05	0,4209	Difference not significant
Semester 1 2017	3,96	3,87	0,09	0,0650	Difference not significant
Semester 2 2017	3,94	3,88	0,06	0,3856	Difference not significant
Semester 1 2018	3,97	3,96	0,01	0,8151	Difference not significant
Semester 2 2018	3,89	3,86	0,03	0,5059	Difference not significant

Table 40 Degree level different semesters comparison of means

The difference in the means of all the semesters have been found not significant, so the different semesters behave similarly as the whole dataset.

5.2.3 Conclusions

After all the analysis performed, there is enough evidence to conclude that:

- **The variable "degree level" does not influence the results** of the teacher survey, at least in ESEIAAT during 2016-2018.
- **Teachers are rated similarly no matter in what level of studies they teach**.
- The contradiction found with the other conducted studies tells us that the influence of the different variables can be different depending on the School and period studied. It can change over time and over the different Schools of the UPC. **The conclusions drawn from the present study should not be extrapolated to other schools outside ESEIAAT, or other periods outside 2016-2018,** without taking into account this fact.

5.3 Number of subject repeaters

For the analysis of this variable, the data has been sorted according to the percentage of repeaters enrolled in the subject.

It is logical to assume that optional subjects could have fewer repeaters than compulsory ones. If a student fails an optional subject, in most of the cases, he can select another optional subject instead of repeating it, unlike compulsory subjects, which must be passed to obtain the degree. The histograms of the two populations are plotted to study this hypothesis.

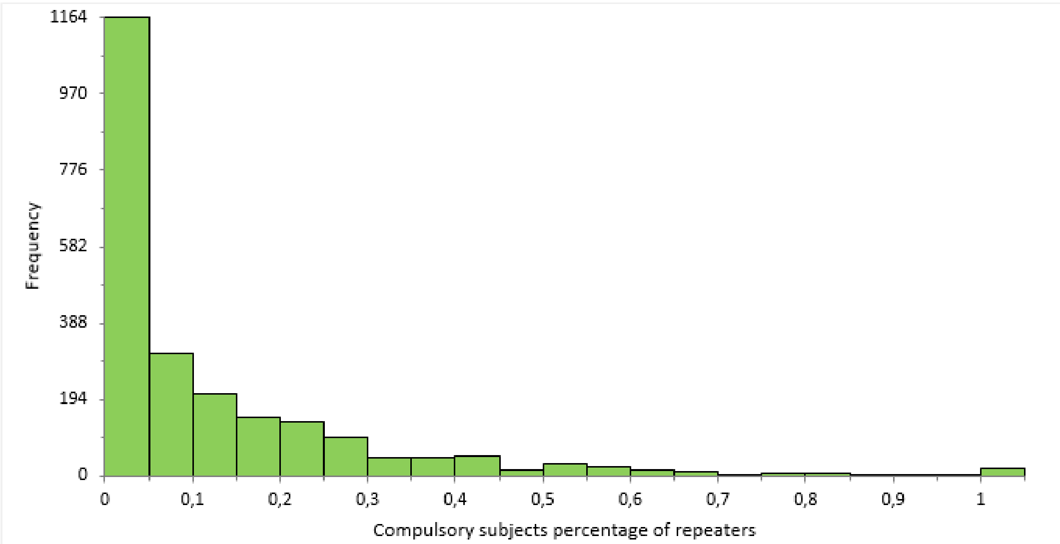


Figure 28 Compulsory subjects percentage of repeaters histogram

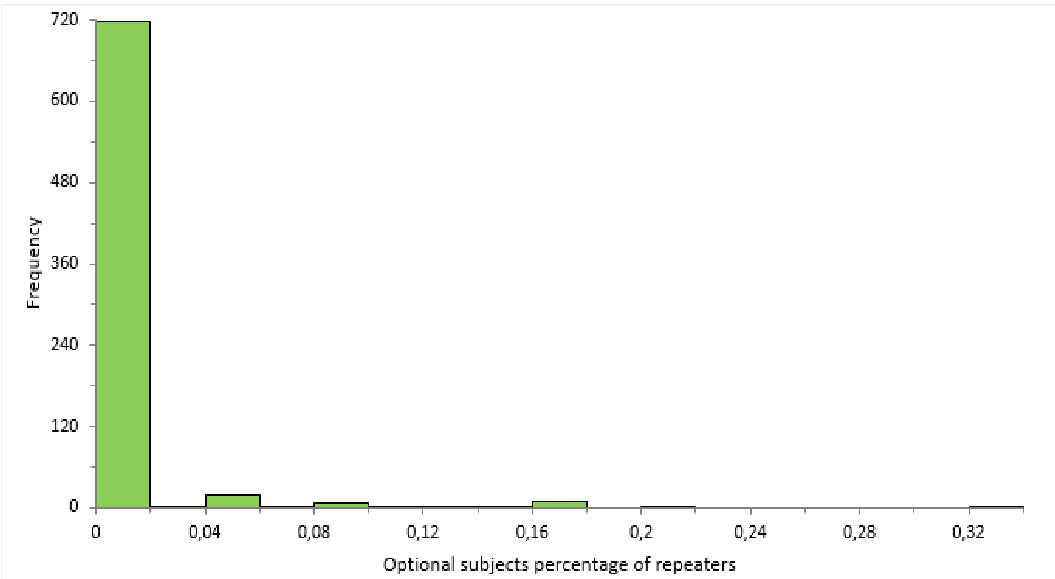


Figure 29 Optional subjects percentage of repeaters histogram

As it can be seen in the histograms, the number of subject repeaters is lower in optional subjects than in compulsory ones. As the type of subject influences the results of the survey (as concluded in Section 5.1) and the distribution of subject repeaters in optional subjects is not very representative, analyzing the pool of the two populations could bias the obtained results. For this reason, this variable will only be analyzed in the compulsory subjects dataset, which has a more representative sample of subject repeaters.

The data has been divided into four groups:

- Group A: Surveys with 0% of repeaters.
- Group B: Surveys with more than 0% and less than 25% of repeaters.
- Group C: Surveys with more or equal to 25% and less than 50% of repeaters.
- Group D: Surveys with more or equal to 50% of repeaters.

5.3.1 Analysis of all the dataset

The final data is composed of 2360 data points distributed in the following manner:

Group	Number of datapoints	Survey rating mean
A: 0% of repeaters	1058	3,88
B: >0% to <25% of repeaters	912	3,87
C: ≥25% to <50% of repeaters	260	3,94
D: ≥50% of repeaters	133	4,00

Table 41 Subject repeaters group distribution

There appears to be an increase in the teacher rating when the number of subject repeaters increases, which would mean that the higher the number of repeaters, the better the teacher will be evaluated. However, this could be caused by the sample being smaller, and thus more biased, as the number of repeaters increases. In fact, if we compare the mean rating of group A vs groups B, C, and D, this difference seems to fade:

Group	Survey rating mean
A	3,88
B, C, D	3,89

Table 42 Comparison of means between A and B, C, D repeaters groups

To further understand this difference, the histograms of the four populations are plotted.

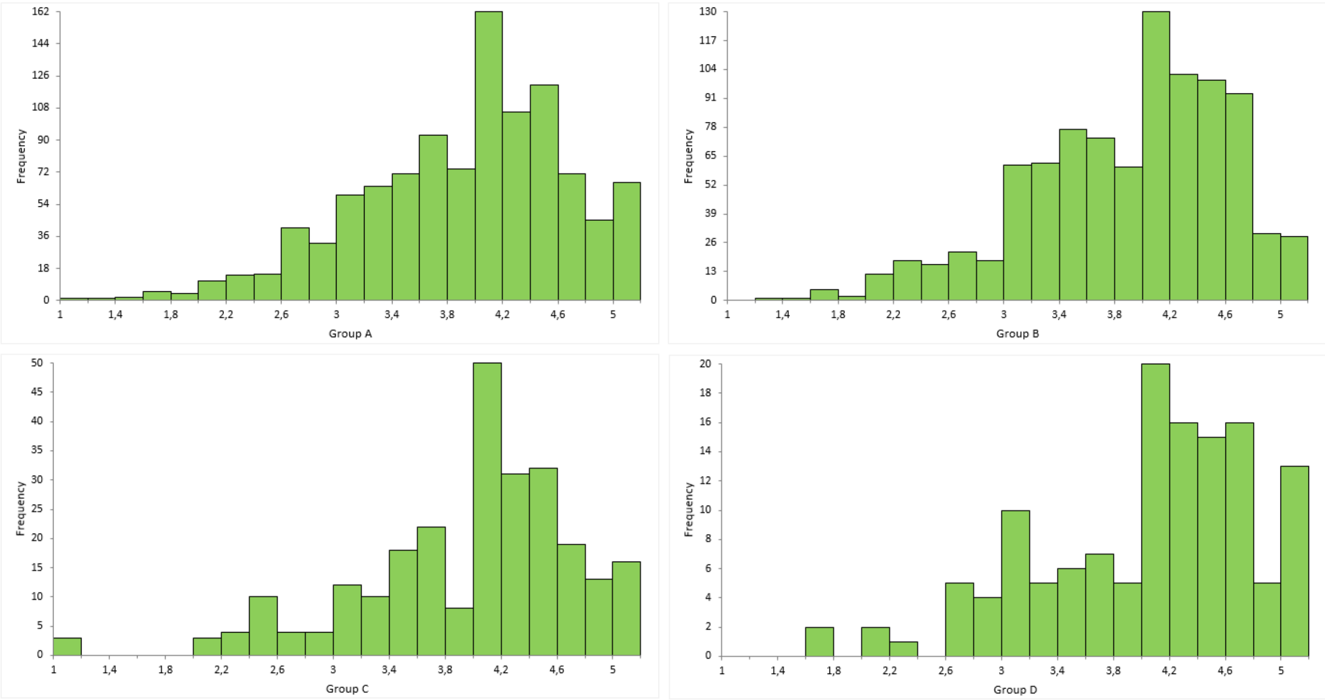


Table 43 A, B, C, D repeaters group histograms

The histogram of group D is different from the rest, being a histogram with more density of high ratings, which could be indicative of a higher mean.

Comparison of group D with group A:

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Group A	<0,0001	-	Non-normal
Anderson-Darling	Group D	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,5290	-	Equal variances
Student's t	Both	0,1004	0,11	Not statistically significant
Wilcoxon-Mann-Whitney	Both	0,0386	0,03	Statistically significant

Table 44 Subject repeaters group D vs A hypothesis testing

Hypothesis testing shows contradictory results. The student's t-test indicates that there is no statistically significant difference in the mean of the two groups, whereas Wilcoxon-Mann-Whitney indicates that there is, indeed, a statistically difference between the two groups. However, the p-value of the Wilcoxon-Mann-Whitney test is near 5%, which indicates that this difference is in the limit of significance.

Now we are going to test the rest of the groups, C vs A and B vs A.

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Group A	<0,0001	-	Non-normal
Anderson-Darling	Group B	<0,0001	-	Non-normal
Anderson-Darling	Group C	<0,0001	-	Non-normal
Brown-Forsythe	A, B	0,6251	-	Equal variances
Brown-Forsythe	A, C	0,7149	-	Equal variances
Student's t	A, B	0,6103	0,02	Not significant
Wilcoxon-Mann-Whitney	A, B	0,5821	0,00	Not significant
Student's t	A, C	0,3064	0,05	Not significant
Wilcoxon-Mann-Whitney	A, C	0,1602	0,07	Not significant

Table 45 B vs A and C vs A repeater groups hypothesis testing

The hypothesis testing results show that there is not a statistically significant difference in the means of groups B and C in comparison with group A. This shows that **the approximate cutoff value in which the number of repeaters starts to influence the results of the survey is around 50% of repeaters**. However, it must be taken into account that this influence is near significance. We will test the behavior of group D in all the semesters to see if this difference occurs in all of them.

5.3.2 Group D vs A semester comparison

Histogram comparison:

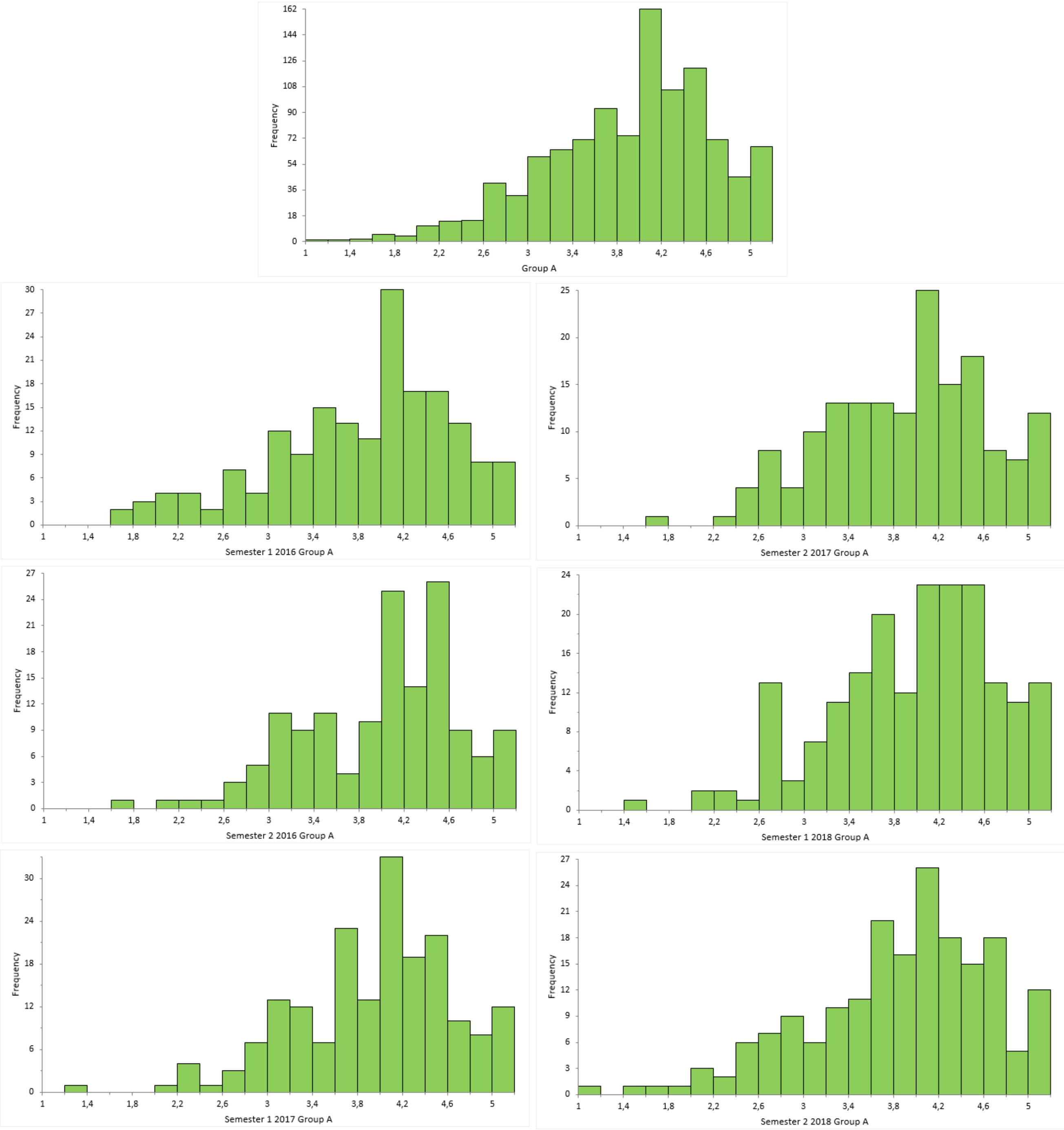


Figure 30 Group A different semesters histograms

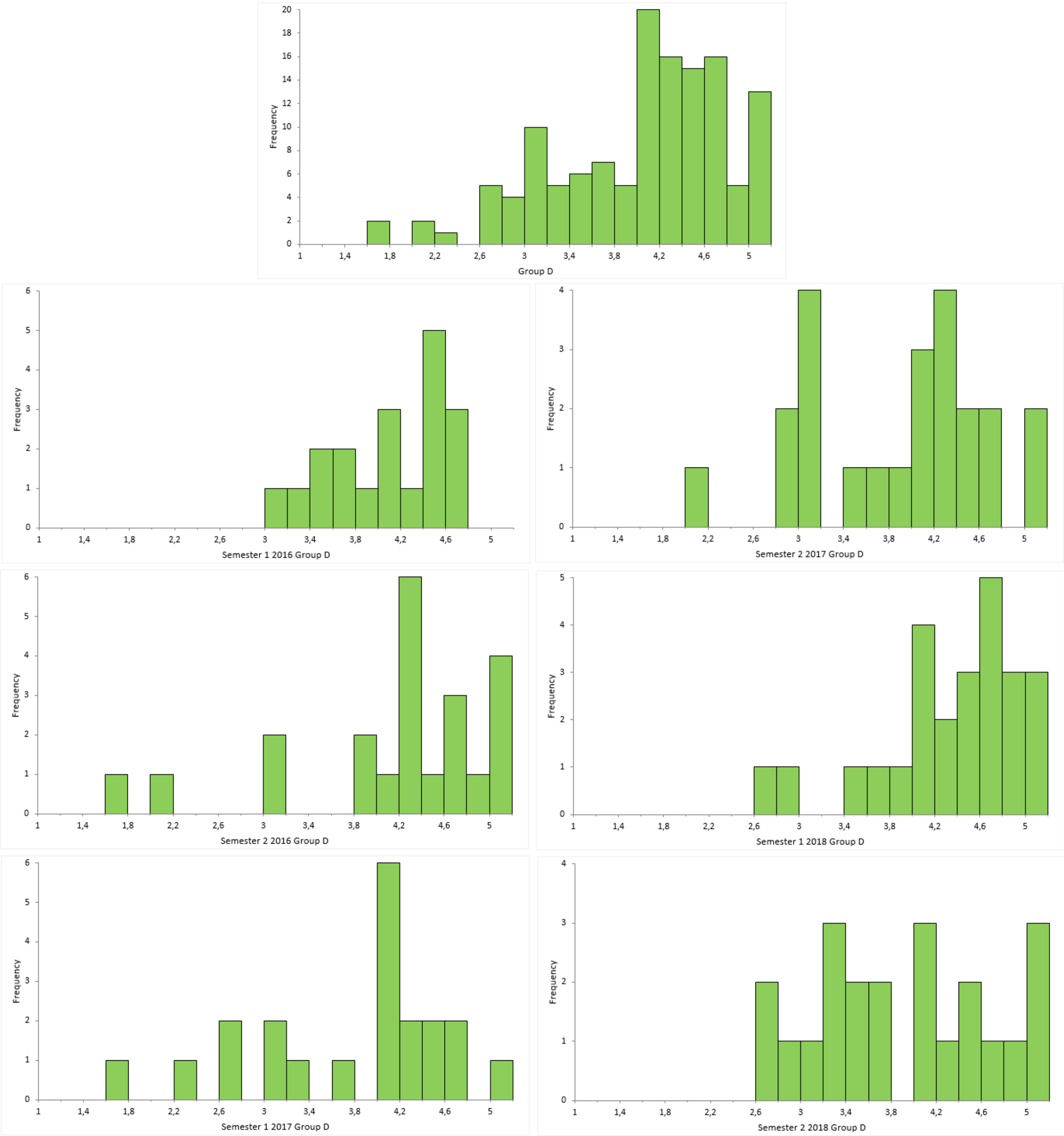


Figure 31 Group D different semesters histograms

The histograms of Group D in the different semesters are very heterogeneous. They vary in shape in the different semesters, so a different behavior of this group in the different semesters can be anticipated. Moreover, the sample sizes of this group are small (around 20 data points), so the value of the mean can be more affected by nonrepresentative extreme values. For this reason, in addition to the means, the medians will also be compared.

Semester	Group A mean	Group D mean	Difference of means
S1 2016	3,80	4,06	0,26
S2 2016	3,97	4,10	0,13
S1 2017	3,92	3,75	-0,17
S2 2017	3,88	3,85	-0,03
S1 2018	3,94	4,29	0,35
S2 2018	3,81	3,88	0,07

Table 46 Different semesters Group A vs D means comparison

Semester	Group A median	Group D median	Difference of medians
S1 2016	4,00	4,11	0,11
S2 2016	4,00	4,27	0,27
S1 2017	4,00	4,00	0,00
S2 2017	4,00	4,00	0,00
S1 2018	4,00	4,50	0,50
S2 2018	3,97	3,83	-0,14

Table 47 Different semesters Group A vs D medians comparison

The tables above show what it was seen in the histograms, group D data behaves differently in the various semesters. In some semesters it follows what was seen analyzing the whole dataset (its mean is higher than group A mean) and in other semesters it behaves contrary and its mean is lower than group A's. Moreover, because of the small samples, the comparison of medians gives different results in S1 2017, S2 2017, and S2 2018, this confirms that nonrepresentative extreme values could be biasing the results.

All the mean and median comparisons above have not been tested for its statistical significance. As the histograms appear to be heterogeneous, the sample sizes are small and appear to be biased and the comparisons show different behaviors in the different semesters and contradictory results, hypothesis testing will not be executed on this data. Group D data behaves differently in every semester and the results obtained analyzing the whole dataset do not explain every one of the periods. For this reason, it is considered that **the statistical conclusions obtained in Section 5.3.1 on Group D data do not completely explain its behavior**.

5.3.3 Conclusions

When analyzing the whole dataset, of all four groups composing the data of this variable, only one (Group D, which comprises surveys with more than 50% of repeaters) was found to have a mean statistically different from the rest. However, the p-value was near significance and the results of the two hypothesis tests were contradictory, which shows that the results of this hypothesis testing must be taken with care.

Moreover, after analyzing group D in the different semesters, it was shown that it behaves differently in every semester, not always confirming the results obtained in the analysis of the whole dataset. Also, its sample sizes are small and prone to be biased by nonrepresentative extreme values, as demonstrated by the contradictory results of the mean and median comparisons.

In the other four groups, that is all the surveys with 0% to 50% of repeaters, it was not found a statistically significant difference in their means.

After all these justifications, it is concluded that:

- **The variable "number of repeaters" does not influence the results** of the teacher survey.
- **Teachers are rated similarly no matter the number of repeaters in the class**.
- Notwithstanding the previous conclusion, it has been found that in ESEIAAT during 2016-2018, surveys of **subjects with more than 50% of repeaters could be better rated than the rest of the subjects**. This behavior does not happen in all the analyzed semesters and its statistical significance is in doubt, so this evidence must be taken with care.

5.4 Number of students

For this analysis, every survey has been sorted according to the number of students enrolled in the subject.

Before starting the analysis, the histogram of the distribution of the surveys according to the number of students enrolled is plotted.

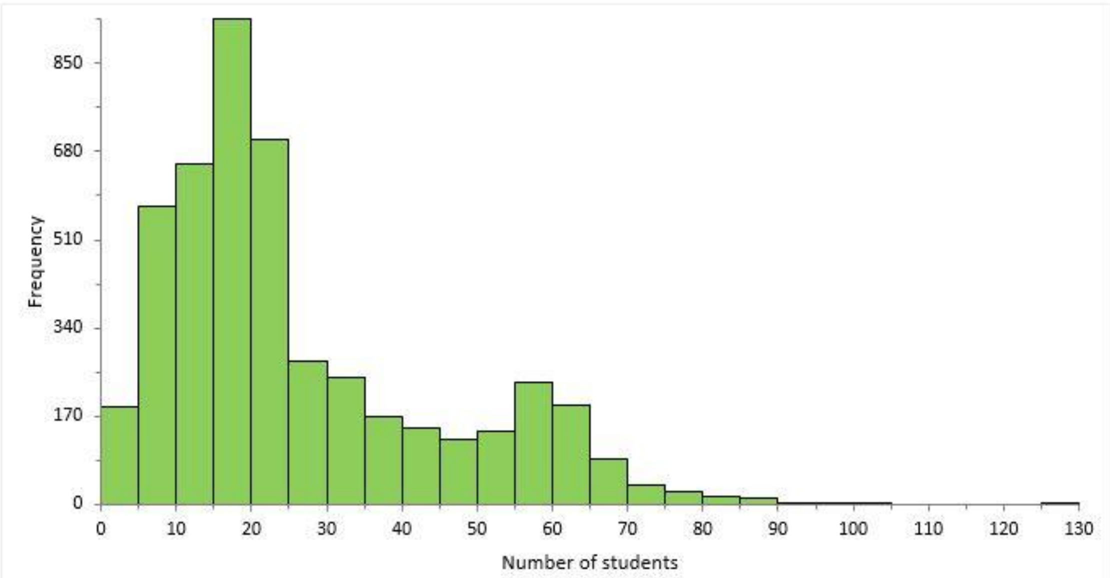


Figure 32 Distribution of the surveys according to the number of students enrolled

This distribution has two peaks, one around 20 students and another around 60 students. We will divide the distribution into two groups, each one containing one of the peaks:

- Group A: Less than 20 students enrolled in the subject.
- Group B: More than 20 students enrolled in the subject.

The two groups have been selected to have a similar size sample, this way the hypothesis testing will work better, none of the samples will be more prone to bias than the other, and the comparative will be more robust.

Group	Sample size
A: <20 students	2355
B: ≥20 students	2423

Figure 33 Number of students sample sizes

5.4.1 Analysis of all the dataset

The histograms of the rating distributions in the two groups are plotted to see if there are any differences.

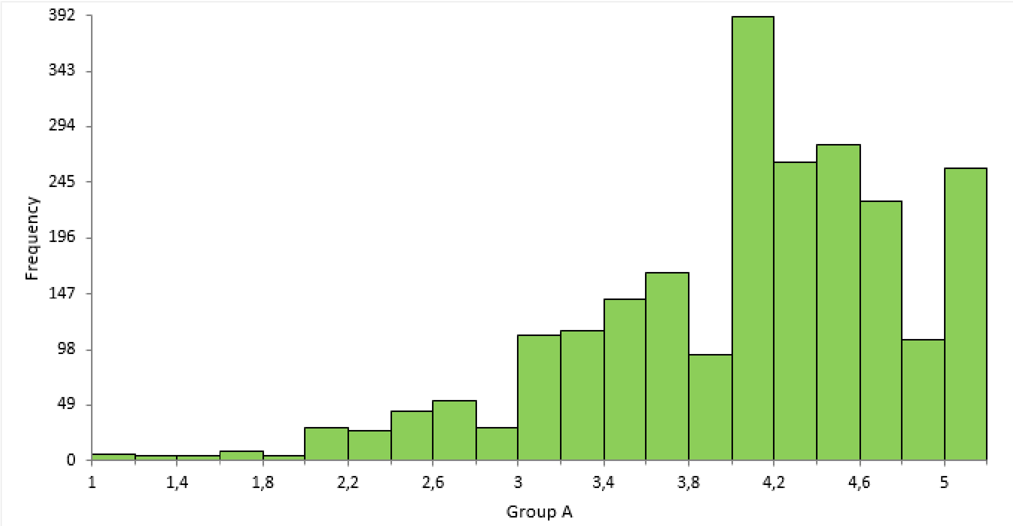


Figure 34 Number of students Group A mean distribution histogram

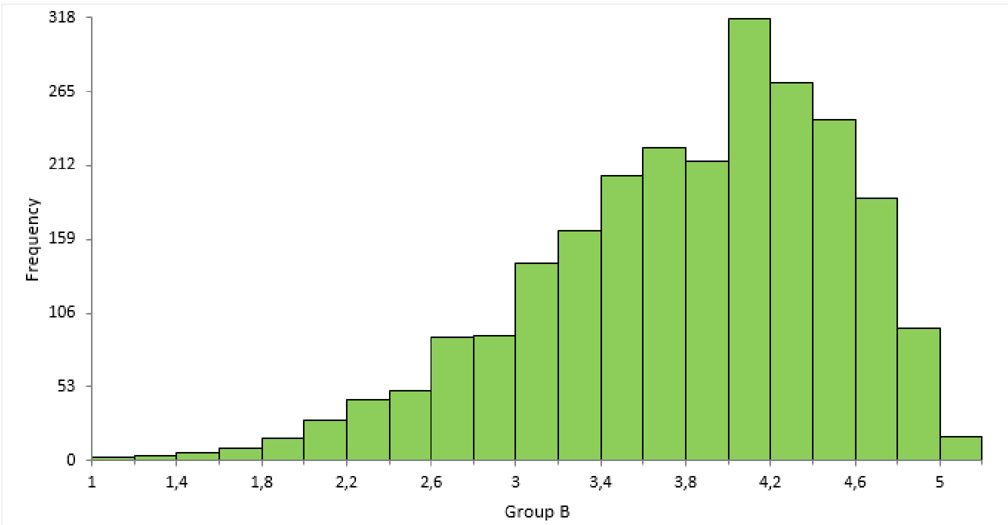


Figure 35 Number of students Group B mean distribution histogram

There is a higher quantity of high ratings in Group A histogram, so a difference in the overall mean can be expected.

Group A mean	Group B mean	Difference of means
4,02	3,79	0,23

Table 48 Number of students Group A and B mean comparison

It seems that there is a difference between the two means, being the average rating of Group A (that is, less populated classrooms) higher than Group B's (more populated classrooms).

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Group A	<0,0001	-	Non-normal
Anderson-Darling	Group B	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,5179	-	Equal variances
Student's t	Both	<0,0001	0,23	Statistically significant
Wilcoxon-Mann-Whitney	Both	<0,0001	0,25	Statistically significant

Table 49 Number of students group A vs B hypothesis testing

There is a highly (p-value under 0,0001) statistically significant difference in the means of the two populations. However, in Section 5.1 it has been found that the type of subject influences the results, so this comparison will be made again only on compulsory subjects. In this way, we can isolate the "number of students" variable and test if it really influences the results.

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Group A	<0,0001	-	Non-normal
Anderson-Darling	Group B	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,7377	-	Equal variances
Student's t	Both	<0,0001	0,18	Statistically significant
Wilcoxon-Mann-Whitney	Both	<0,0001	0,19	Statistically significant

Table 50 Number of students group A vs B hypothesis testing only on compulsory subjects

Analyzing the variable only on compulsory subjects, there is still indeed a highly statistically significant difference in the means of the two populations. **There is enough**

statistical evidence to conclude that the mean of the two populations is different, being the mean of the less populated (less than 20 enrolled students) subjects higher, by an amount of approximately 0,20 points.

5.4.2 Analysis of the different semesters

The analysis will continue to be made only on data of compulsory subjects, to avoid the influence of the type of subject.

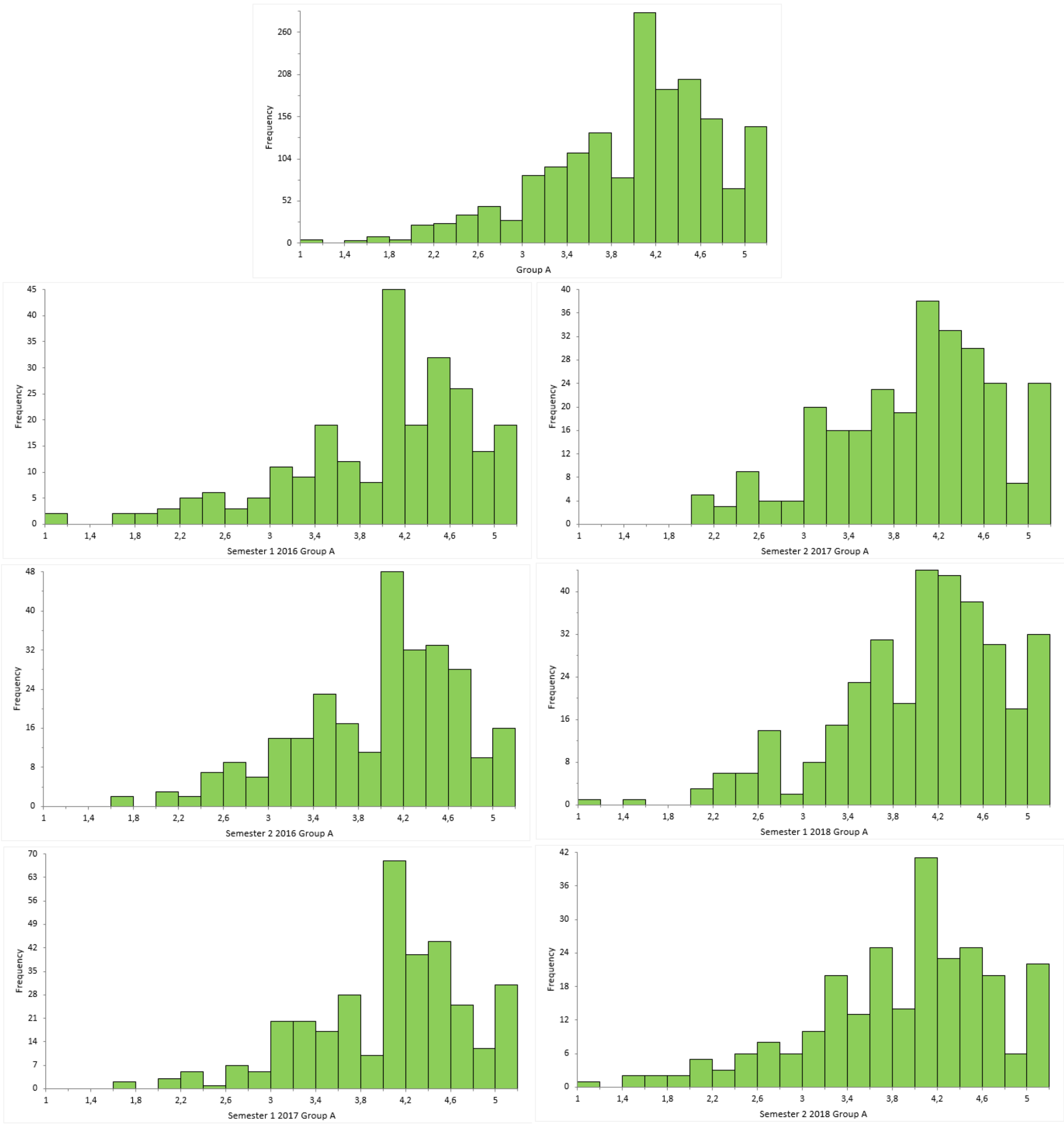


Figure 36 Number of students group A different semesters histograms

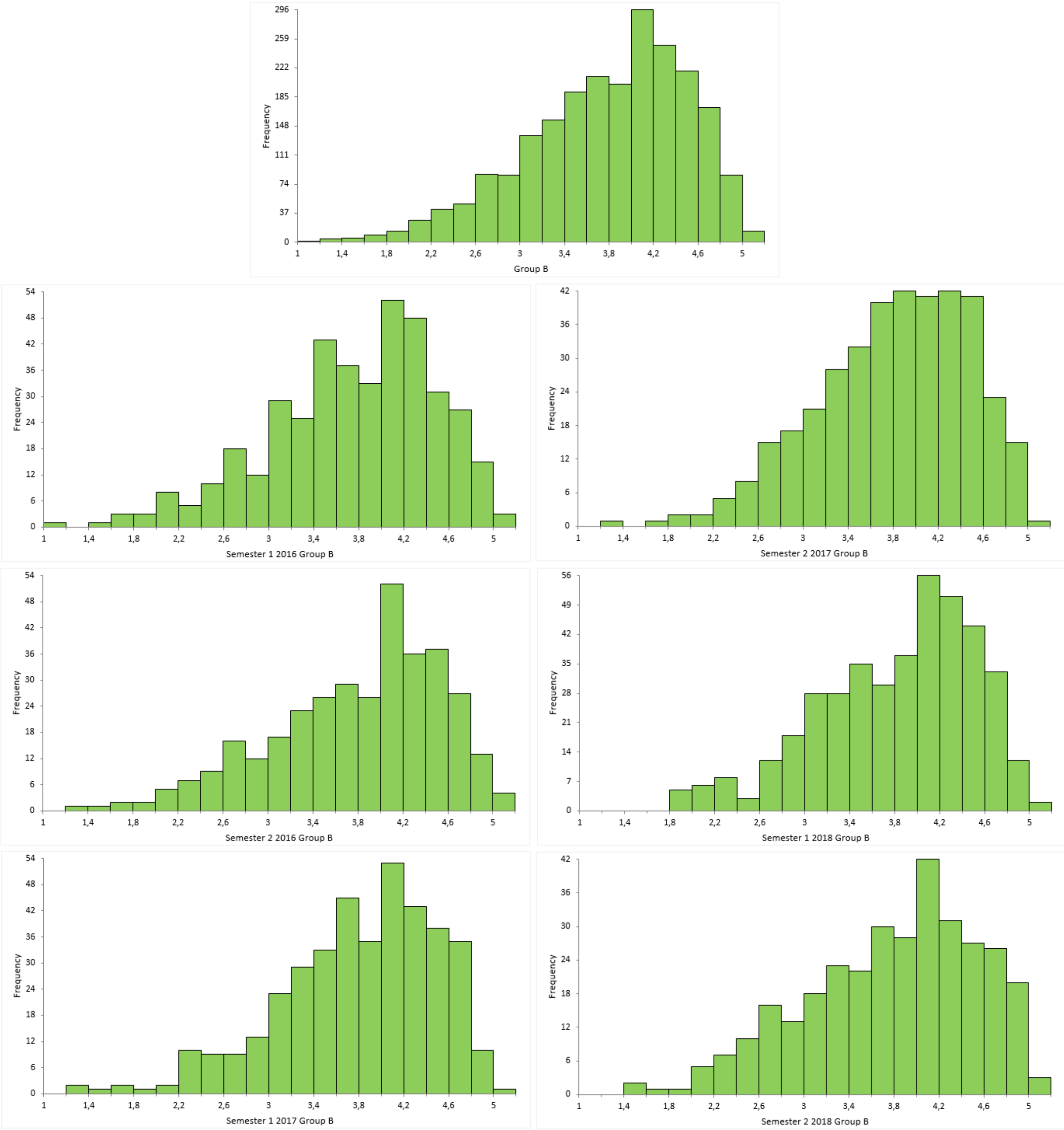


Figure 37 Number of students group B different semesters histograms

Comparison of means:

Period	Group A mean	Group B mean	Difference of means	WMW p-value	Result
Semester 1 2016	3,96	3,74	0,22	<0,0001	Statistically significant difference
Semester 2 2016	3,94	3,78	0,16	0,0078	Statistically significant difference
Semester 1 2017	4,01	3,79	0,22	<0,0001	Statistically significant difference
Semester 2 2017	3,94	3,79	0,15	0,0040	Statistically significant difference
Semester 1 2018	4,01	3,81	0,2	<0,0001	Statistically significant difference
Semester 2 2018	3,84	3,77	0,07	0,1367	**Not statistically significant difference**

Table 51 Degree level different semesters comparison of means

The variable behaved similarly in all the Semesters except Semester 2 2018. In all the rest of the Semesters, there was a statistically significant difference between the means of group A and B, being the average rating of surveys in classrooms populated by less than 20 students bigger, by approximately 0,20 points, than the rating of surveys of more populated classrooms.

In Semester 2 2018, there was not a statistically significant difference between the means of the two populations, however, the observed means behaved similarly like the rest of the periods, being the observed mean of the group A bigger than group B's.

It is concluded that the variable behaves similarly, as analyzed in the whole dataset, in all the periods.

5.4.3 Worst-case comparison

To determine if the statistical differences found in the previous analysis really affect the teacher surveys empirically, worst-case comparisons are made.

According to the previous analysis, a worst-case for a teacher would be to only teach in populated classrooms, and a best-case would be to only teach in less populated classrooms. Therefore, if the previous inferences are true, teachers who only teach in less populated classrooms will have a better rating than teachers who only teach in populated classrooms.

In ESEIAAT, during the 2016-2018 period, 138 teachers taught only on populated classrooms and 63 teachers taught only on less populated classrooms.

Type of teacher	Number of teachers	Mean rating
Teaches only on less than 20 students classrooms	63	3,99
Teaches only on more than 20 students classrooms	138	3,68

Table 52 Number of students worst-case comparison

It appears to be a difference between the means of the two populations. As the previous statistical inferences pointed, teachers who teach only on less populated classrooms appear to be better rated than teachers who teach only on more populated classrooms.

To prove this assumption, hypothesis testing is executed on the data.

Data	Test	p-value	Mean or location difference	Conclusion
Only less than 20 students	Anderson-Darling	<0,0001	-	Non-normal
Only more than 20 students	Anderson-Darling	0,0135	-	Non-normal
Both	Brown-Forsythe	0,6839	-	Equal variances
Both	Student's t	0,0051	0,31	Statistically significant
Both	Wilcoxon-Mann-Whitney	0,0014	0,34	Statistically significant

Table 53 Type of subject worst-case comparison hypothesis testing

The results of the two hypothesis tests show that there is enough statistical evidence to conclude that the means of the two populations are different, being the mean of the teachers who only teach in less populated classrooms higher, by an amount of approximately 0,3 points.

This allows us to conclude that **teachers who only teach in less populated classrooms are generally better rated than teachers who only teach in more populated classrooms**.

The distribution of teachers in ESEIAAT according to their percentage of surveys in teaching in less populated classrooms respect their total surveys is as follows.

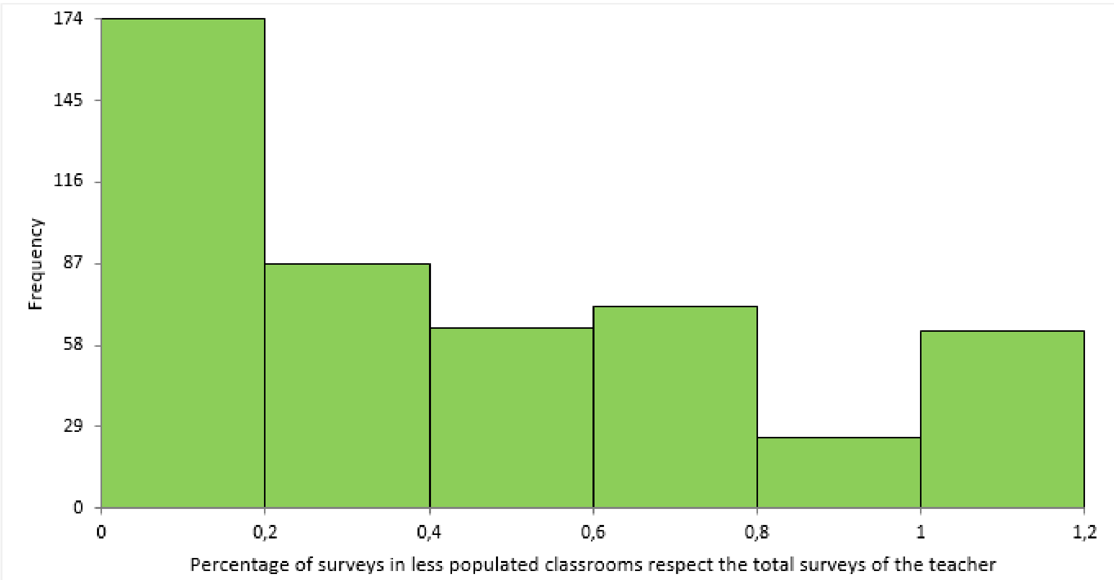


Figure 38 Percentage of surveys in less populated classrooms respect the total surveys of the teacher histogram

In this worst-case testing, the two extreme conditions have been tested, but what happens between these two extremes has not been analyzed yet.

For that, the difference between teachers who only teach in more populated classrooms and teachers who also teach in less populated classrooms will be tested. Teachers whose optional percentage is higher than 0% and less than 80% will establish this second population, in this way we set aside the less populated classrooms extreme condition (100%). The cutoff is established at the 80% mark because between 80%- 90% the sample of data is small and conclusions drawn from the analysis will not be representative of that population.

In ESEIAAT during 2016-2018, 138 teachers only taught in classrooms with more than 20 students and 259 teachers who taught in all kinds of classrooms.

Type of teacher	Number of teachers	Mean rating
Teaches only in classrooms with more than 20 students	138	3,68
Teaches in all kind of classrooms	259	3,84

Table 54 Only populated classrooms vs all kind of classrooms teachers comparison

As it appears in the table above, the difference between the mean of the two populations is not as big as it was comparing the extremes.

Data	Test	p-value	Mean or location difference	Conclusion
Only populated	Anderson-Darling	0,0135	-	Non-normal
All kind of classrooms	Anderson-Darling	<0,0001	-	Non-normal
Both	Brown-Forsythe	0,0002	-	Different variances, Welch's t-test will be used.
Both	Welch's t	0,0204	0,16	Statistically significant
Both	Wilcoxon-Mann-Whitney	0,0508	0	Not statistically significant

Table 55 Only populated classrooms vs all kind of classrooms teachers hypothesis testing

The obtained results are not conclusive. Welch's t-test results classify the difference as statistically significant and Wilcoxon-Mann-Whitney classifies it as not statistically significant. The results are not conclusive and the Wilcoxon-Mann-Whitney p-value is in the limit of significance, which means that the difference is in the limit of being significant.

As a conclusion of these results, it can be stated that there is a difference in the rating of the teachers depending on the number of the students in the classroom, being that **teachers who teach in classrooms of less than 20 students are generally better rated than teachers who teach in classrooms of 20 or more students**.

5.4.4 Conclusions

After all the analysis performed, there is enough evidence to conclude that:

- **The variable "number of students" influences the results** of the teachers' survey.
- **Teachers who teach in less populated classrooms**[12] **are generally better rated than teachers who teach in more populated classrooms**[13], no matter the type of subject**.**

5.5 Phase (initial/non-initial)

For the analysis of this variable, surveys have been sorted according to if the subject was in the initial or the non-initial phase of the degree. Master's degrees do not have initial and non-initial phases, so they are out of the scope of this analysis. Moreover, to avoid the influence of the type of subject (there are many more compulsory subjects in the non-initial than in the initial phase), the analysis has been carried out only on compulsory subjects.

5.5.1 Analysis of all the dataset

The histograms of the rating distributions in the two groups are plotted to see if there are any differences.

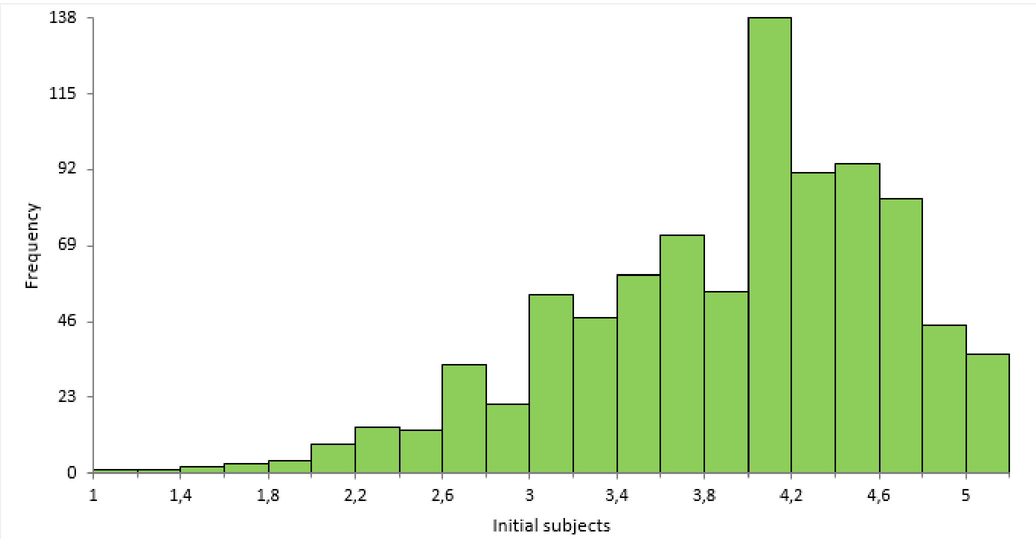


Figure 39 Initial subjects histogram

[12] Defined as classrooms with less than 20 students.
[13] Defined as classrooms with 20 or more students.

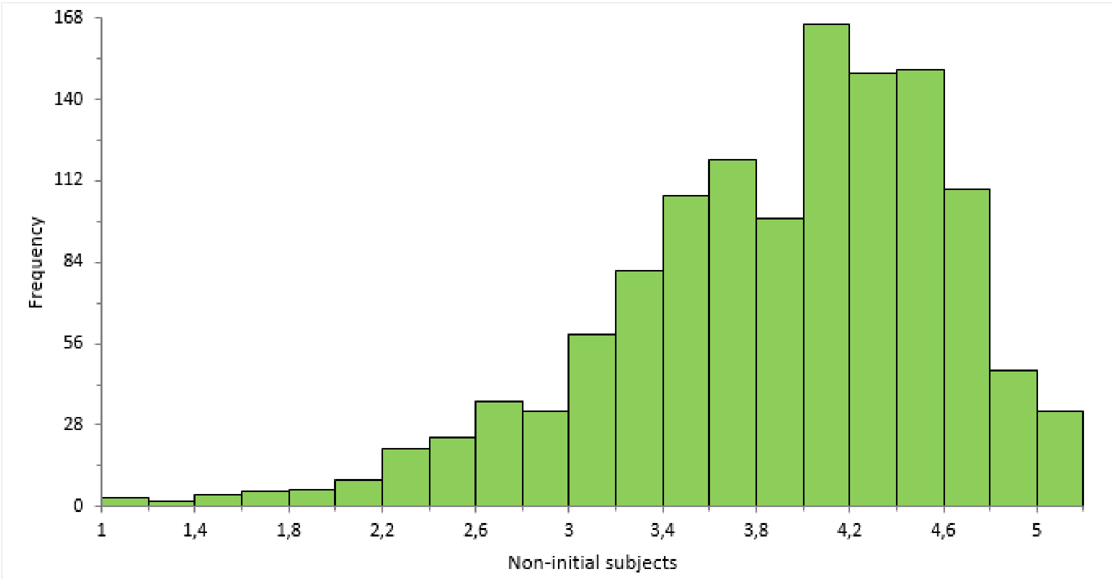


Figure 40 Non-initial subjects histogram

The histograms are similarly shaped, so the distribution of the two populations are similar and no difference in their means is expected.

Initial subjects mean	Non-initial subjects mean	Difference of means
3,90	3,87	0,03

Table 56 Initial and non-initial subjects means

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Initial	<0,0001	-	Non-normal
Anderson-Darling	Non-initial	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,3034	-	Equal variances
Student's t	Both	0,3042	0,03	Not statistically significant
Wilcoxon-Mann-Whitney	Both	0,1832	0,04	Not statistically significant

Table 57 Phase variable hypothesis testing

The results conclude that **there is not a statistically significant difference** in the means of the two populations.

5.5.2 Analysis of the different semesters

Histogram comparison:

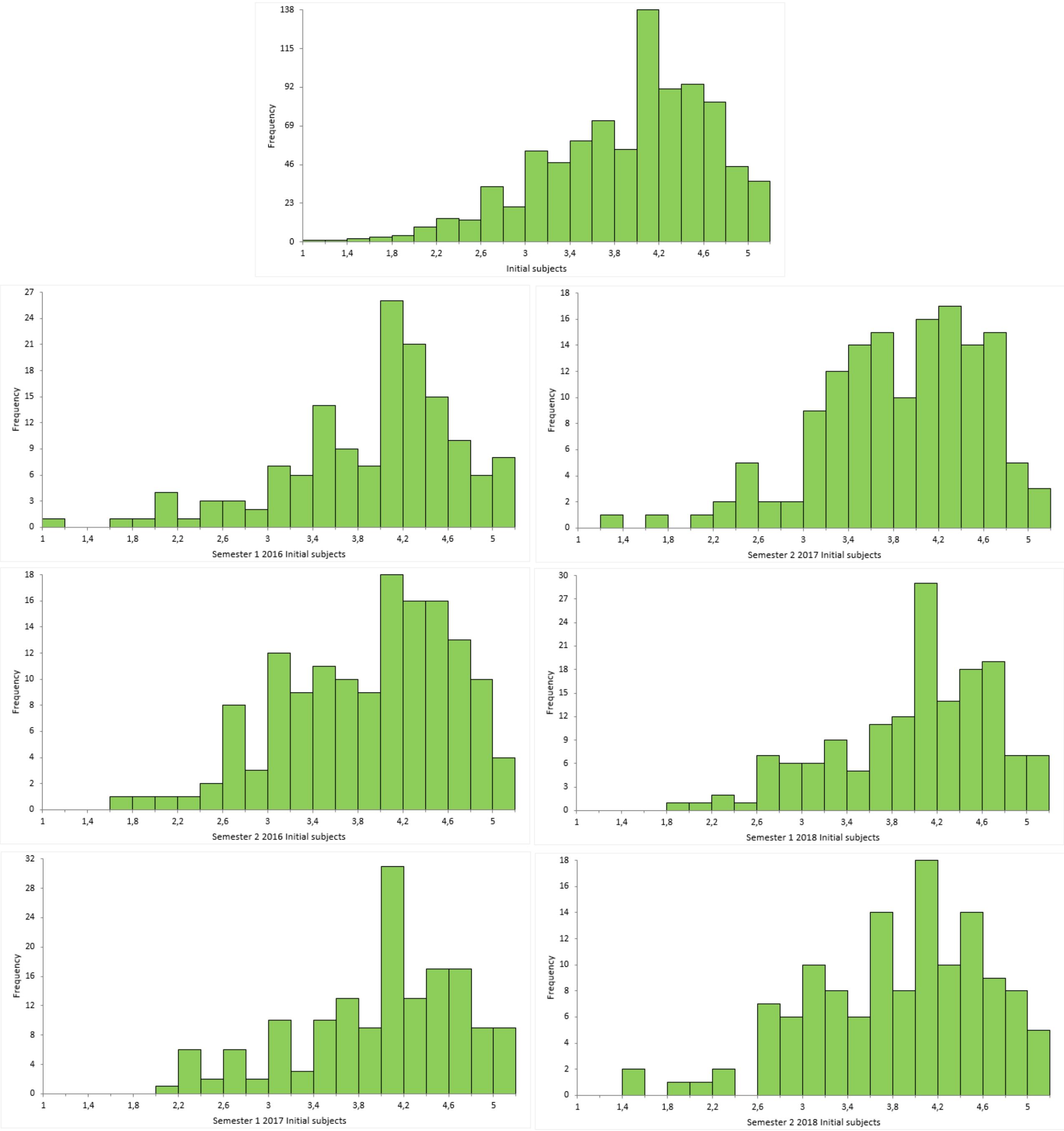


Figure 41 Initial subjects different semesters histograms

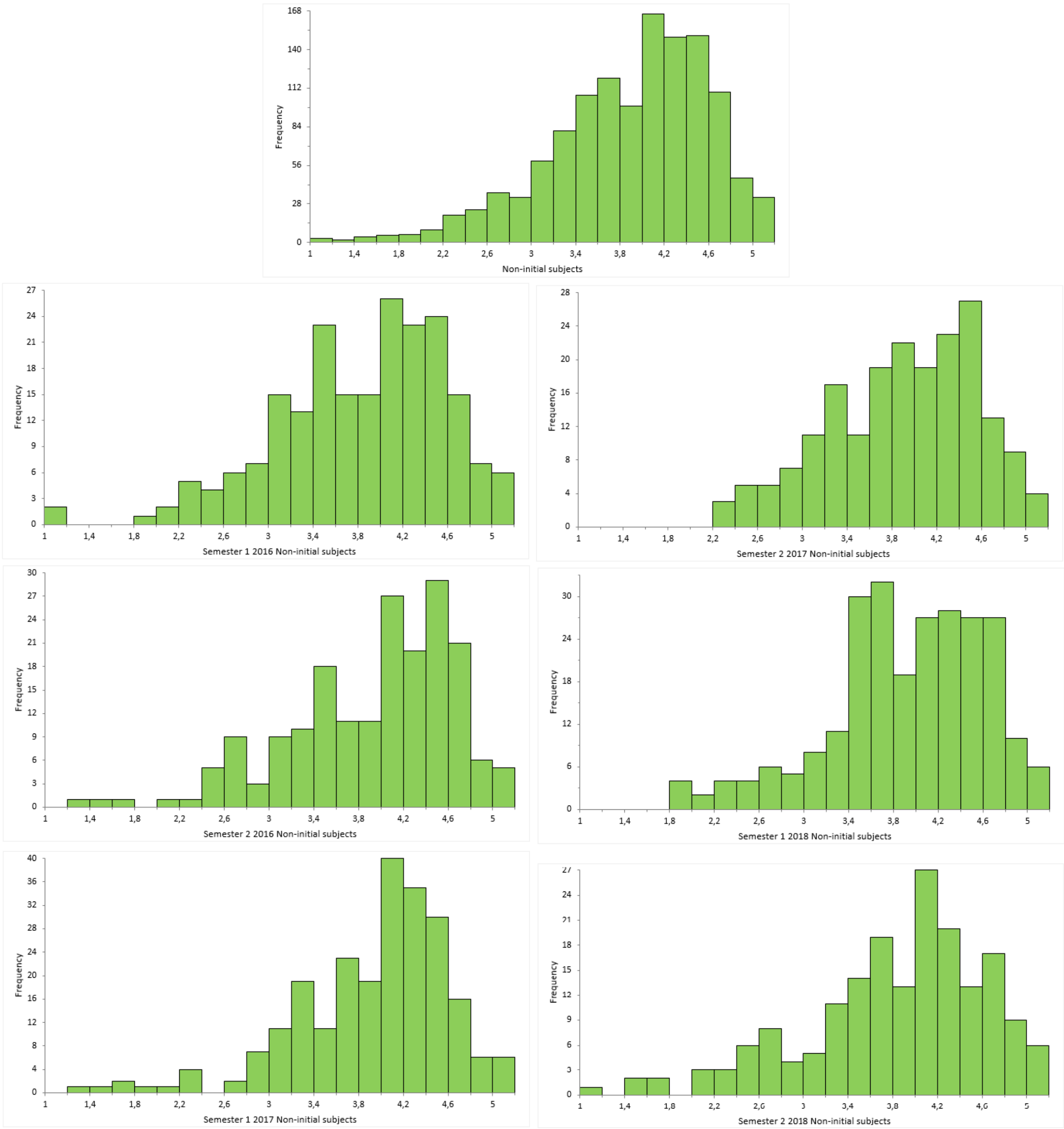


Figure 42 Non-initial subjects different semesters histograms

Comparison of means:

Period	Group A mean	Group B mean	Difference of means	WMW p-value	Result
Semester 1 2016	3,90	3,80	0,10	0,1567	Not statistically significant
Semester 2 2016	3,88	3,90	-0,02	0,6634	Not statistically significant
Semester 1 2017	3,97	3,91	0,06	0,2235	Not statistically significant
Semester 2 2017	3,84	3,88	-0,04	0,7991	Not statistically significant
Semester 1 2018	3,97	3,89	0,08	0,1639	Not statistically significant
Semester 2 2018	3,82	3,80	0,02	0,9609	Not statistically significant

Table 58 Degree level different semesters comparison of means

All the differences observed in the different semesters are not statistically significant, so **there are no differences in the mean of initial and non-initial subjects surveys in any of the semesters**.

5.5.3. Conclusions

After all the analysis performed, there is enough evidence to conclude that:

- **The variable "phase" does not influence the results** of the teachers' survey.
- **Teachers are rated similarly no matter the phase of the subject they teach.**

5.6 Teacher's gender (men/women)

For the analysis of this variable, surveys have been sorted according to the gender of the teacher evaluated.

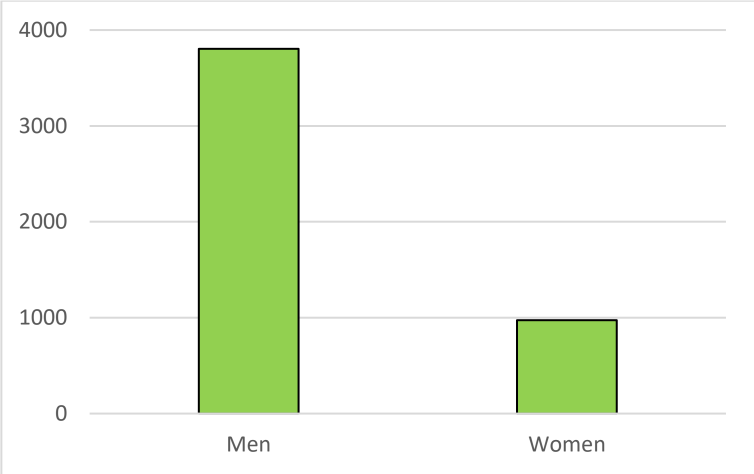


Figure 43 Surveys sorted according to the teacher's gender

There are 3 times more surveys conducted on men than women. That could cause some bias in the comparison between the two populations, however, as the size of the women population is big (around 1000 data points) it is not probable that it could be biased by nonrepresentative extreme values.

5.6.1 Analysis of all the dataset

The histograms of the rating distributions in the two populations are plotted to see if there are any differences.

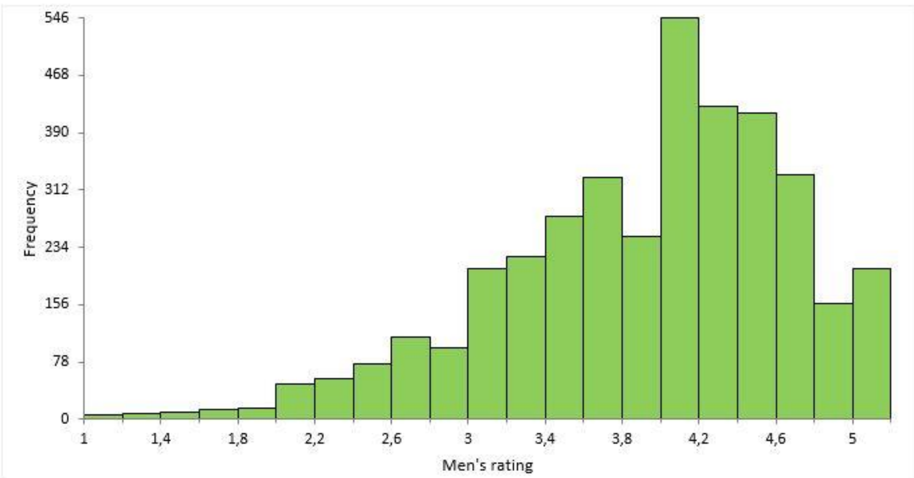


Figure 44 Men's rating histogram

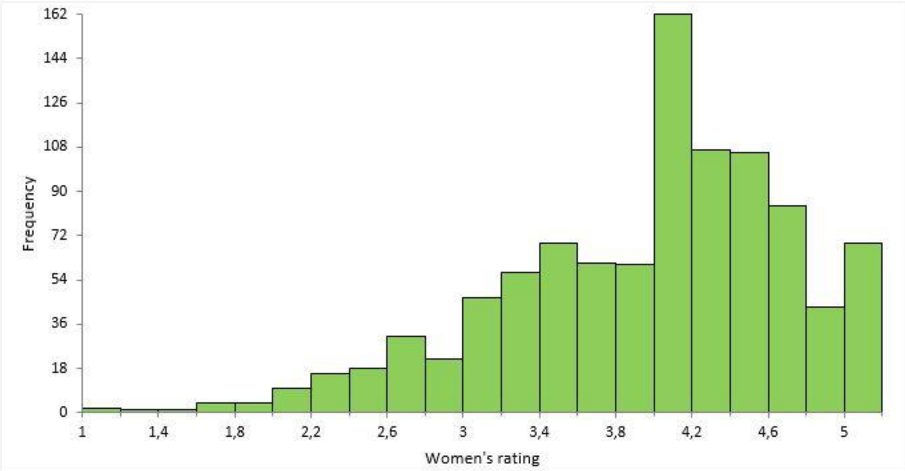


Figure 45 Women's rating histogram

The histograms of the two populations are similarly shaped so the distribution of the ratings is similar and no difference in their means is expected.

Men's mean	Women's mean	Difference of means
3,90	3,93	0,03

Table 59 Men and women means

Test	Population	p-value	Mean or location difference	Result
Anderson-Darling	Initial	<0,0001	-	Non-normal
Anderson-Darling	Non-initial	<0,0001	-	Non-normal
Brown-Forsythe	Both	0,8215	-	Equal variances
Student's t	Both	0,1618	0,03	Not statistically significant
Wilcoxon-Mann-Whitney	Both	0,1244	0,03	Not statistically significant

Table 60 Gender hypothesis testing

The results conclude that **there is not a statistically significant difference** in the means of the two populations.

5.6.2 Analysis of the different semesters

Histogram comparison:

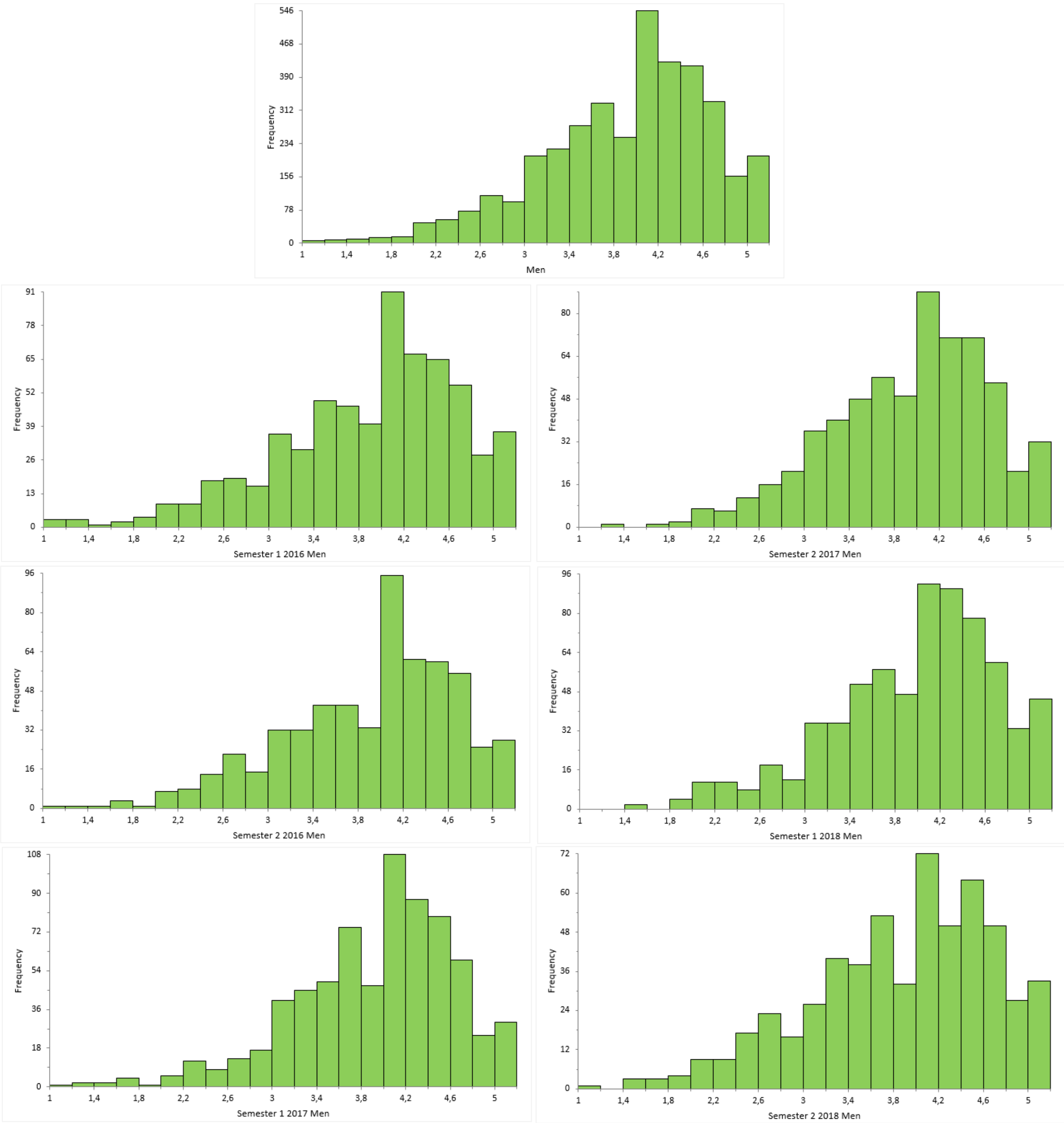


Figure 46 Different semesters men's rating histograms

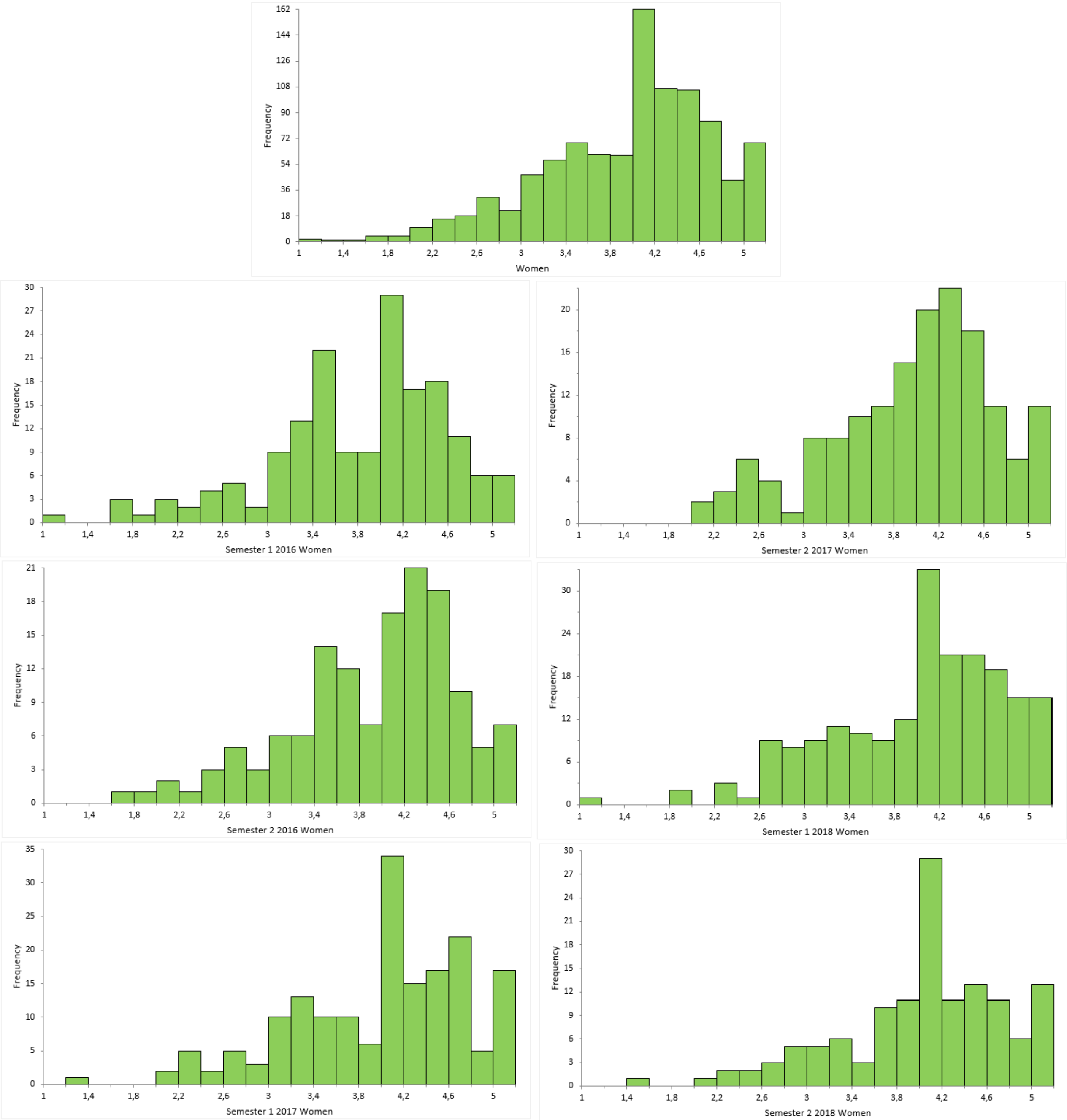


Figure 47 Different semesters women's rating histograms

Comparison of means:

Period	Men mean	Women mean	Difference of means	WMW p-value	Result
Semester 1 2016	3,87	3,8	0,07	0,1961	Not statistically significant
Semester 2 2016	3,88	3,92	-0,04	0,6607	Not statistically significant
Semester 1 2017	3,9	3,97	-0,07	0,1586	Not statistically significant
Semester 2 2017	3,91	3,94	-0,03	0,5122	Not statistically significant
Semester 1 2018	3,96	3,99	-0,03	0,4355	Not statistically significant
Semester 2 2018	3,85	4	-0,15	0,0683	Not statistically significant

Table 61 Degree level different semesters comparison of means

All the differences observed in the different semesters are not statistically significant, so **there are no differences in the mean of men and women surveys in any of the semesters**.

5.6.3. Conclusions

After all the analysis performed, there is enough evidence to conclude that:

- **The variable "gender" does not influence the results** of the teachers' survey.
- Teachers are rated similarly no matter their gender. **Men and women are evaluated equally in the satisfaction surveys**.

6. CONCLUSIONS

6.1 Conclusions obtained from the study

Out of the 6 variables analyzed in the present study, **only 2 variables were found to influence the results of the survey**. These variables are the type of the subject (compulsory or optional) and the number of students in the classroom (less than 20 students or 20 or more students).

Variable	Influences the results of the survey
Subject type (compulsory/optional)	**YES**
Degree level (bachelor's/master's)	NO
Number of subject repeaters	NO
Number of students	**YES**
Phase (initial/non-initial)	NO
Teacher's gender (men/women)	NO

Table 62 Influence of the different variables analyzed

Of the 2 variables found to influence the results of the survey, although having a similar influence (of about 0,3 points on average), **the most influential one is the number of students in the classroom**.

Notwithstanding the previous conclusions, the evidence shows that **the current satisfaction survey system is robust**. The most statistically significant difference found is about 0,3 out of 5 points (a 6% difference). **There have not been noted any big discrepancies or differences in the ratings of the different teachers as an influence of biasing variables**. The only differences found are that **teachers who only teach compulsory subjects are generally better rated (with 0,3 points in average) than the rest of the teachers** and that **teachers who teach in classrooms with less than 20 students are generally better rated than teachers who teach in classrooms with 20 or more students**.

As a final conclusion, it has been noted that **the influence of the different variables in the survey system can vary in time and can be different in the different Schools**. For this reason, the conclusions of the present study only apply in ESEIAAT during 2016-2018 and must not be extrapolated to other UPC Schools or periods without taking much care.

List of conclusions obtained in the different analyses:

- **Teachers who only teach optional subjects are generally better rated than the rest of the teachers.**
- **Teachers who teach some optional subjects are generally rated in the same way as teachers who only teach compulsory subjects**.
- Most usually, **the same teacher is better rated in his work in optional subjects than in compulsory ones**.
- **Teachers are rated similarly no matter in what level of studies they teach**.
- **Teachers are rated similarly no matter the number of repeaters in the class**.
- It has been found that in ESEIAAT during 2016-2018, surveys of **subjects with more than 50% of repeaters could be better rated than the rest of the subjects**.
- **Teachers who teach in classrooms of less than 20 students are generally better rated than teachers who teach in classrooms of 20 or more students**, no matter the type of subject**.**
- **Teachers are rated similarly no matter the phase (initial or non-initial) of the subject they teach.**
- Teachers are rated similarly no matter their gender. **Men and women are evaluated equally in the satisfaction surveys**.

6.2 Suggestions for improvement

The present study found that, in general, the satisfaction survey system of the UPC, applied in ESEIAAT, is robust and does not have many systematic problems. As it has been found that the influence of the different variables can vary in time and on the different Schools, **it is not advisable to try to apply any correction factors**, as they will not probably work well as time passes and will not probably work for every School on the UPC.

The data obtained from the satisfaction survey system must be used to continually improve the quality of the University, and its robustness must be continually assessed. For this reason, **it is recommended to conduct similar studies like the present one**, to the satisfaction surveys data of the different Schools, to assess the robustness of the survey system across the University.

From the analysis carried out in the present study, it has been found that teachers who only teach optional subjects are generally better rated than teachers who only teach compulsory subjects. Teaching compulsory or optional subjects can be a very different experience for a teacher. It has been found that approximately 50% of ESEIAAT teachers only teach compulsory subjects, so it is recommended that, whenever possible, **optional subjects should be distributed more evenly between the members of the faculty**. In this way, the teaching experience will be more even between the staff and the whole educational system will benefit.

It has also been found that the number of students in the classroom is the variable that influences the most in the obtained results. For this reason, it is recommended that, if it is not possible to **reduce the number of students** of the more populated classrooms, at least **the most populated subjects should be distributed more evenly** between the members of the faculty, in a similar way as proposed before for optional subjects. In fact, as optional subjects tend to be less populated, it is possible that when distributing more evenly the optional subjects between teachers, the two issues could be solved.

6.3 Future lines of research

In all the previous studies on the topic, it has been assessed the satisfaction survey system, suggesting improvements in the survey system itself, but it has not been suggested improvements in the educational system. For this reason, the following future lines of research are suggested:

- Study the viability of distributing optional subjects more evenly between the faculty staff.
- Study how the number of students in the classroom affects the quality of teaching and which should be the recommended maximum number of students in a classroom.
- Study the economic aspects and the viability of reducing the number of students in the classrooms.

Moreover, the robustness of the satisfaction survey system must be assessed continuously and, for this reason, the following lines of research are suggested:

- Continue the study in different Schools, to assess the robustness of the system across the whole University and obtain more knowledge on its performance and more suggestions for improvement.
- Study the influence of the variables in data from all the UPC Schools, to try to assess the influence of the different variables in a global way.

7. BIBLIOGRAPHY

Berger, R. and Casella, G. *Statistical Inference.* Second Edition. Duxbury, 2002.

Hayes, B. E. *Measuring Customer Satisfaction and Loyalty: Survey Design, Use, and Statistical Analysis Methods*. Third Edition. ASQ Quality Press, 2008.

Ellis, R., and Hogard, E. *Handbook of Quality Assurance for University Teaching*. First Edition. Routledge, 2018.

Fuster Arion, Jordi. *Anàlisi de les enquestes a l'alumnat de la UPC sobre assignatures i professorat (fase 1)*, 2016.

Menéndez Landa, Paloma. *Análisis de las encuestas al alumnado de la UPC sobre asignaturas y profesorado (fase 2)*, 2017.

Agell Vendrell, Pol. *Estudio de mejora del modelo de encuestas de satisfacción. Reducción de la influencia de las variables del proceso en la fiabilidad del resultado*, 2019.

Douglas, J., Douglas, A., and Barnes B. *Measuring student satisfaction at a UK university.* Quality Assurance in Education Vol. 14 No.3, pp. 251-267, 2006.

Gruber, T., Fuß, S., Voss, R., and Gläser-Zikuda, M. *Examining student satisfaction with higher education services: Using a new measurement tool*. International Journal of Public Sector Management, Vol. 23 No. 2, pp. 105-123, 2010.

Kärnä, S., and Julin, P. *A framework for measuring student and staff satisfaction with university campus facilities*. Quality Assurance in Education, Vol. 23 No. 1, pp. 47-66, 2015.

8. ANNEXES

Annex to this report are the Excel spreadsheets provided by the GPAQ and the Excel spreadsheets containing all the statistical tests conducted. Every spreadsheet containing the statistical tests is named with the Section it contains information of.

All the Annexes are contained in a compressed file.