Data Analysis Report Question 3: Correlation between Partials and Exam

Elias Reutelsterz

May 2025

Contents

1 Context

The present research question refers to the mathematics placement test of the Universidad Nacional del Sur, named ATI, from the year 2025. Successful participation in this test is required for admission to study in a selection of the careers offered at UNS. In total, this encompasses 47 careers. We have data available for the first attempt of this test. There is also a second attempt that can be taken if one fails the first, but we do not have data for its results.
UNS offers a preparatory course for the ATI test that includes practice assignments. The four practice exams are called Seguimientos, and the university offers the opportunity to earn a type of bonus in the ATI test if one passes at least two Seguimientos (more than 6 out of 10 possible points) with at least three participations out of four possible. The bonus consists of having the test counted as passed if one achieves more than 50 percent of the points, instead of the usual 60 percent.
The preparatory courses are divided into different groups, called COMISIONES. Furthermore, all COMISIONES in group **M** receive a different exam than the COMISIONES in group **B**.

2 Problem setting

2.1 Research Question

Is there a correlation between completing/passing partials and the exam beyond the advantage of the grading system itself?

2.2 Analysis Guide

2.2.1 Verification of Available Data

- Do we have individual data for each partial?
- Do we have final exam scores for all students?
- Is there information about the grading system and its rules?

2.2.2 Missing Data to Request (if necessary)

- Exact details of the grading system and weightings
- Dates of activities vs. exam
- Specific grading criteria per activity

2.2.3 Analysis Pipeline

1. **Data Cleaning:** Separate students according to their grading regime (with/without advantage)
2. **Descriptive Analysis:** Calculate simple correlation between activities and exam
3. **Control by System:** Analyze correlation within each grading group
4. **Residual Analysis:** Study the correlation after removing the system's effect
5. **Controlled Regression:** Use models that adjust for the grading mechanism
6. **Interpretation:** Distinguish between true correlation vs. system advantage

3 Data Analysis

3.1 Data Overview

First, we consider the data on test results stored in Ati_anonimo.xlsx. This includes the following columns with their respective meanings:

Column	Meaning
COMISION	commission including group M or B
ACTA	commission index
NRO. INSCR	unique student number
CARRERA	career index
Carrera-Nom	career name
DNI	Unique national ID number
NOTA	Passed/Failed/Not taken

Table 1: Overview ATI Test Data

Furthermore, we have tables for the results of the Seguimientos for each of the 26 commissions, with the following content (only relevant information displayed):

Column	Meaning
DNI	Unique national ID number
Comenzado el	Time of test started
Finalizado	Time of test finished
Calificación	Number of total points earned in this Seguimiento
P. i/j	Points earned in exercise i of the total j points

Table 2: Overview Seguimiento Data

3.2 Data Preprocessing

Since we only need the columns 'DNI', 'COMISION', 'NOTA' from the Ati file, all duplicates with the same entries in these three columns were deleted after loading. This affected 345 out of a total of 4326 rows. These now unique individuals form the basis for a DataFrame to which we add the Seguimiento results. It was noticed that one commission from another university location was included, which was removed from the analysis. Furthermore, there was one student with a DNI that was not purely numerical, which is actually impossible. This student was also excluded from the analysis.

3.3 Results

3.3.1 Descriptive Statistics

The resulting dataset shows the following student distributions.

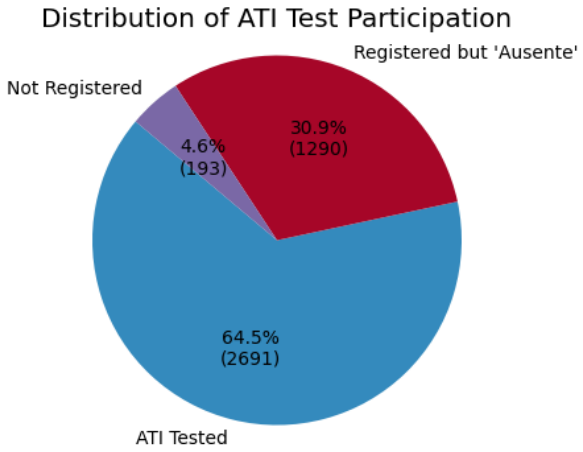

Figure 1: Distribution of students in dataset

'Not Registered' means that there was data for Seguimientos, but no data for an ATI Test.

Now, looking at the number of Seguimientos completed by students, we get the following distribution.

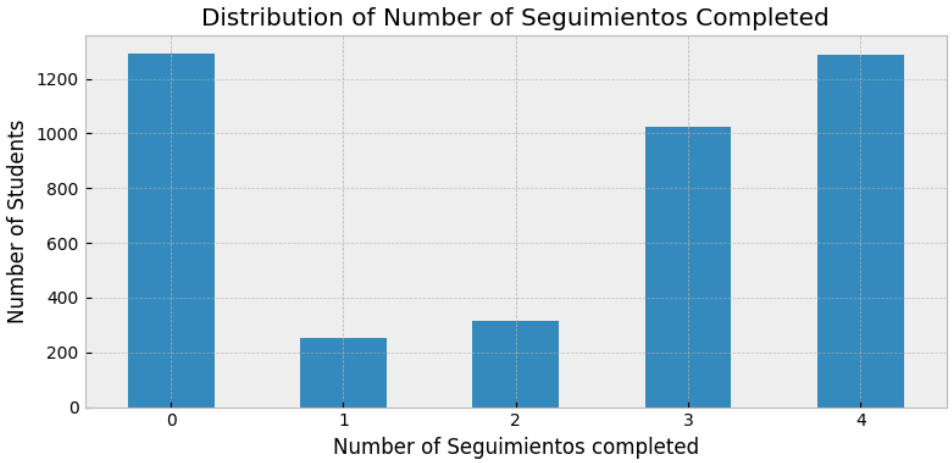

Figure 2: Distribution of seguimientos completed

3.3.2 Influence of completing Seguimientos on the Test Outcome

If we now look at the individual results of the Seguimientos, we can calculate whether the respective student received an Advantage or not (explained in 1). The conditional distributions of the ATI Test grades look as follows.

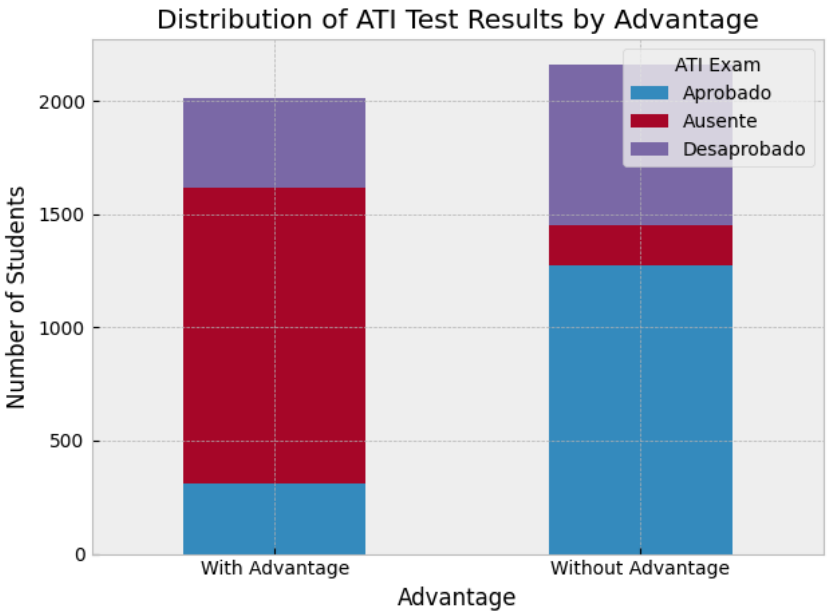

Figure 3: Distributions dependent on advantage

The resulting chance of passing the ATI test depending on having an Advantage or no Advantage is therefore as follows.

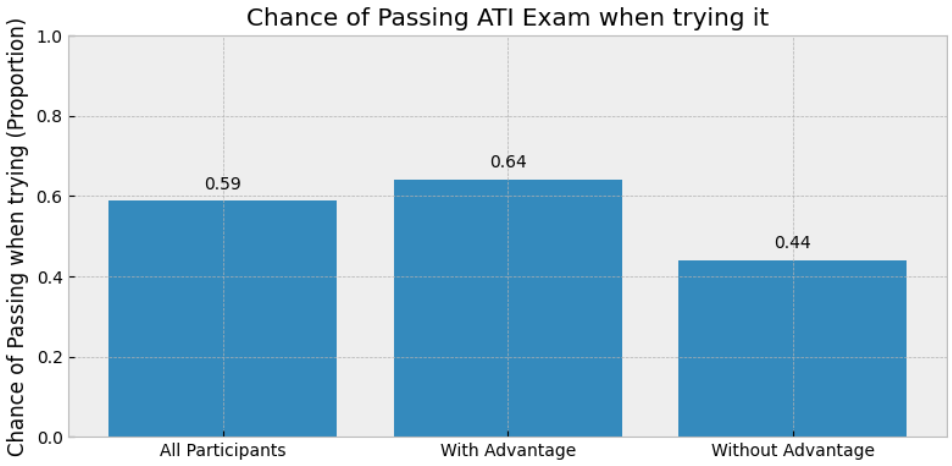

Figure 4: Chance of passing ATI test

The correlation between Advantage from the Seguimientos and passing the ATI test is 0.18. If we look at the sum of the following Seguimiento(s) i, we even achieve the following correlation with passing the ATI Test:

Seguimiento/s i	Correlation to Passing ATI test
1	0.32
2	0.23
3	0.23
4	0.22
$1 + 2$	0.33
$1 + 2 + 3 + 4$	0.39

Table 3: Correlation of Seguimiento points to ATI test

3.3.3 Influence of Commission Group on Test Outcome

If we divide the dataset into Commission Group **B** and **M**, we obtain the following statistics:

Quantity of interest	Group **B**	Group **M**
Number of participants	1558	2616
Chance of passing ATI test	0.63	0.57
Correlation between Advantage and Grade in ATI test	0.13	0.26

Table 4: Influence of Commission Group

3.3.4 Regression Analysis

Performing a regression analysis with the independent variables group_M (boolean indicating if the student is in Commission group M) and Advantage (1 if they earned an Advantage, 0 if not) and the dependent variable 'Grade in ATI Test', yields the following results.

Variable	Value
const	0.489
group_M	-0.1136
Advantage	0.2341

Table 5: Regression Results

Accordingly, students who earned an Advantage performed on average 23.4% better in the test (based on the Passed/Failed outcome). Membership in Commission Group **M** was thus, on average, a disadvantage of 11.4%. Both coefficients are statistically significant.