GENEVA GRADUATE INSTITUTE

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT

GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Statistics for International Relations Research I

Midterm Exam

Rémi Viné

The Graduate Institute — Geneva

Due: Wednesday the 15th of November 2023, at noon (Geneva time)

Instructions

- You have **one week (minus 2 hours) to complete the test**.
  - The total number of points assigned is 30, one fourth of the full class.
- 3 points are assigned to the replicability of your code.
- You are expected to upload your work on Moodle by Wednesday the 15th, at noon.
- Your .pdf document and .R code should address questions in the order in which they appear on the exam.
  - Your .pdf file (or any other well-suited format you prefer) should be a standalone document providing all required answers. Provide details of your computations, process you followed to reach a conclusion. You will not receive full credit if you only write the answer. Make sure your answer is contextualized, using appropriate levels of measurements.
  - Your .R file should be directly replicable, from the data shared with the exam.
- For questions not using R, keep three decimal digits.
- You are expected to implement the materials covered in classes and lab sessions, and possibly use other R commands that can be handy.
- This is an individual exercise, do not work with other students, any suspicion of work copying each other (or being so close to each other that this is not individual work but, say, using extensively an AI) can lead to a zero.

Study survey on studies

Wikipedia defines a famous survey on educational attainment as such: "The Programme for International Student Assessment (PISA) is a worldwide study by the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations intended to evaluate educational systems by measuring 15-year-old school pupils' scholastic performance on mathematics, science, and reading. It was first performed in 2000 and then repeated every three years. Its aim is to provide comparable data with a view to enabling countries to improve their education policies and outcomes. It measures problem solving and cognition."

- You are given a description of the sampling design in the document:
  $MIDTERM\_PISA2018\_TecReport - Ch - 04 - Sample - Design.pdf$.
- You are given a summary report written by OECD:
  $MIDTERM\_PISA\_2018\_Insights.pdf$
- You are given two fake data sets:
  $PISA\_fake\_student\_level.csv$
  $PISA\_fake\_school\_level.csv$

First, use the sampling design document.

1. What is the population? *(0.25pt)*
2. Does the survey design include all people in the population with a non-zero probability to be sampled? *(0.25pt)*
3. Within a selected school and among the population target, is the sampling design a random probability sampling? *(0.5pt)*
4. In a country, was type of sampling design was elaborated to select schools (systematic, stratified, clustered, a mixture of them)? *(0.25pt)*
5. In a selected school of 278 students in the target population and where the paper-based assessment is implemented, indicate (if possible) the probability of a student belonging to the target population to be selected in the sample? *(0.75pt)*
6. Apart from special cases (exclusion of entire school), are all schools equally likely to be selected? Explain the possible layers of differences, if any. *(0.75pt)*
7. Assume, because 7th grade students of a given school where is a spring outing at the moment of the survey and this was not anticipated by the school authorities. Therefore, the field researcher could only gather responses from 30% of the students in this school and, for logistic reasons, they cannot come back after the outing. Would these students be included in the survey? Explain. *(0.25pt)*

Now, look at PISA results as shown in the report shared.

8. What type of variable is PISA score? *(0.25pt)*

9. What is the level of measurement of the PISA score? *(0.25pt)*

10. Using figure 4 (of the *.pdf*), do we know the score of the best student per country? *(0.25pt)*

11. Using figure 4 (of the *.pdf*), do we know the median scores of countries? *(0.25pt)*

12. From figure 4 (of the *.pdf*), comparing Chile and Slovak Republic, what can you say about the low performing students and the top performing students? *(1pt)*

13. Discuss the difference between the middle decile score of Singapore students in Figure 4 (about 535) and the average score of Singapore students in Figure 1 of the report (score = 549). If you had to choose what could be the distribution of students' scores in Singapore, what figure below would it be? *(0.75pt)*

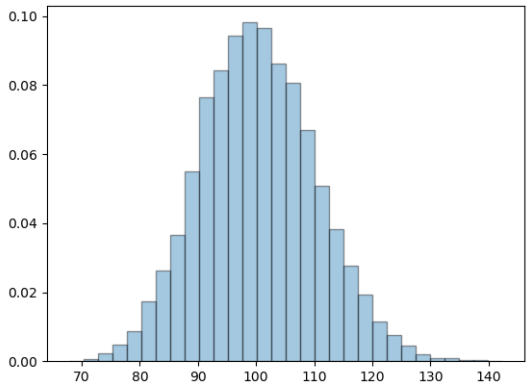

Figure 1

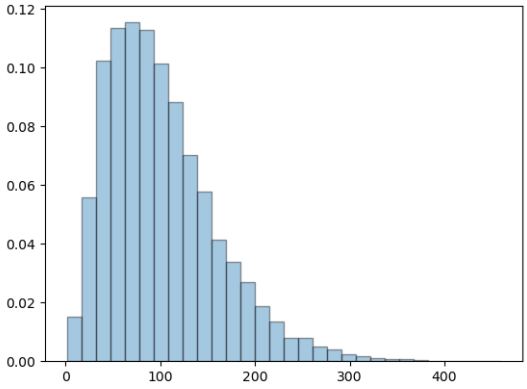

Figure 2

14. Using Figure 2 of the report, and focusing on Switzerland, can you indicate in what interval there is **at least** 3/4 of all observations. *(0.75pt)*

15. Figure 3 of the report shows that the average scores in science. Assume Malaysian students and Emirati students are symmetrically distributed. Where students, in share, are more likely to score in science more than 592? *(1pt)*

16. Using Figure 1 of the report on reading scores, what is the 98% confidence interval of the reading score of Brazilian students? (Show the steps, conclude accordingly, assume $n = 6300$.) *(1pt)*

17. Assume that you only have one representative school where students passed the Computed-Based Assessment (42 students took part of the survey). In the student questionnaire (page 64), question ST222Q09HA asks "I participate in activities in favour of environmental protection." with two possible answers, "Yes" and "No". In the classroom, 25% of all students answered "Yes". Construct a 90% confidence interval of the proportion of students taking part in environmental protection activities. Would you conclude that less than a third of students in the population is taking part in these actions? (Show the steps, conclude accordingly) *(1pt)*

The table II 2-[1/2] here shows the gap of top performers depending on some characteristics.

18. "Education is a top avenue of social reproduction. Auto-selection is one driving force of this phenomenon: disadvantaged and advantaged students internalize their environment and make their expectations accordingly. Therefore, disadvantaged students expect not to complete tertiary education at top level **more** than advantaged students. This is true even is the most equal country in terms of income (based on Gini coefficient), which is the Slovak Republic."

    Conduct a hypothesis testing in order to bring a conclusion on this statement. You are testing a difference between two proportions, assume that the pooled estimate of the proportion is 17.698%. Assume that the sample standard deviation for advantaged students is 5% and 30% for disadvantaged students. Advantaged students account for 10% of the sample, and so do disadvantaged students (other students, the 80% remaining percent, are neither advantaged nor disadvantaged). You are expected to conduct the test step by step and explicit them. *Hint: use the level of significance you deem appropriate, assume 6300 students in total were surveyed in the country (but you workable sample is only composed of the disadvantaged and advantaged students). You may not use the difference directly, as you cannot guarantee to match pairs. (1.5pt)*

Now you can use the two data sets shared in .csv format. One is composed of students directly surveyed in one school assumed to be representative ($PISA\_fake\_student\_level.csv$). The other data set is composed of schools in a given country ($PISA\_fake\_school\_level.csv$). Different scores are given, along with the total income of the parents. For the first data set, these variables are at the student level, for the second data set, these variables are school averages.

19. Write an analysis of the learning situation in the school for which students' data is available (Sex is coded as 0 for girls, and 1 for boys). What is the topic where students score best, where variance is the larger, where there is asymmetry, *etc.*? You are expected to compute indicators you deem relevant, produce appropriate data visualizations (distribution plots, scatter plots, *etc.* - and label them accordingly!). (400 words max) *(4pts)*

20. Write an analysis of the learning situation in the country using the school-level data. You are expected to compute indicators you deem relevant, produce appropriate data visualizations. Specifically, discuss at length the shape of the distributions of the three scores. You must produce a simply summary statistic table (for numerical variables) with relevant indicators. You are expected to produce some Boxplots of the scores by income (create a binary variable for income based on a criterion you deem appropriate and use these two groups for the boxplot), class size (small is $< 20 students$, large is $> 30 students$ and medium is in between). You are also expected to discuss the difference in the spread for scores at the school level (this data set) and at the student level in a school (previous question's data set). (500 words max) *(6pts)*

21. Assume that the schools in the second data set are representative of all schools in the world. OECD states (fictitious) "We regret to notice that the score in science of schools in the world is, with a level of significance of 1% below 500, that is worrying for the future of education". Can you test this statement? *(1pt)*

22. Using the school-level data, create a contingency table (and display frequencies) with class size in rows and income (below/above) median income as columns. Compare the grand total with the total number of observations. *(1pt)*

23. From the contingency table, calculate $P(\text{“}Mediumsizedclass\text{”})$. *(0.25pt)*

24. From the contingency table, calculate $P(\text{“}Below\ median\text{”})$. *(0.25pt)*

25. From the contingency table, calculate $P(\text{“}Below\ median \cap \text{“}Small\text{”})$. *(0.25pt)*

26. From the contingency table, calculate $P(\text{“}Belowmedian \cup \text{“}Large\text{”})$. *(0.25pt)*

27. From the contingency table, calculate $P(\text{“}Abovemedian|\text{“}Large\text{”})$. *(0.25pt)*

28. Are "*Above median*" and "*Large*" independent in probability? *(0.5pt)*

29. Are the two variables statistically independent at $\alpha = 1\%$? *(2pts)*

30. What is the probability to have a sample where class size and binary income are even more independent (that is, where the test-statistic of the previous test is lower than the one found in the previous question)? *(0.5pt)*