THE GRADUATE INSTITUTE GENEVA

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

Due: Tuesday the 15th of November 2022, at 4pm

The Graduate Institute — Geneva

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 Tuesday the 15th, a 4pm.
- 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 four 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.

I Statistical Tools

10 pt

1. The Center for Systemic Peace gathers data on countries regimes over time, the so-called *polity* data. You can find codebooks and other details here.

Democracy score	0	1	2	3	4	5	6	7	8	9	10
Frequency	30	10	4	3	5	6	12	16	21	23	33

Table 1: Polity V, democracy score and country frequency, values as of 2018, one observation per country

   (a) Compute the expected value of the democracy score *0.5 pt*

   (b) What is the median and what is the mode of the democracy score? Can you comment on a broad understanding on the shape of the distribution from the computation of these three measures of centrality? *1 pt*

   (c) Compute the standard deviation of the democracy score *0.5 pt*

   (d) Assume now that a think tank uses only random samples of the countries for the sake of computation simplicity. Only 25 countries are selected. Can you give the standard error of the sampling distribution and compare with the population standard deviation? What happens if the think tank increases the number of countries randomly included into the sample? *1 pt*

2. ILO reports modern slavery around the world. It is constructed around a taxonomy that splits modern slavery into two categories: (i) forced labor, (ii) forced marriage. Table 2 displays some frequencies taken from a 2022 ILO report. The worldwide share of individuals under the yoke of modern slavery in 6.4‰. This accounts for 49'570'000 individuals. Answer each of the following questions, if possible, if not, simply indicate that there is not enough information to find the answer. In your computation, stick using fractions, using decimal format only for results.

   (a) Indicate what is the probability to be a male (worldwide) and to suffer from forced labor (in ‰, and assuming events "to be a male" and "to suffer from modern slavery" are independent events). *0.5 pt*

   (b) What is the probability for an individual suffering from modern slavery to be female and in a forced marriage situation? *0.5 pt*

   (c) Among individuals facing modern slavery, what is the probability that the individual is a child? *0.5 pt*

(frequencies in thousands of people)	Forced labor (FL)	Forced Marriage (FM)
Male (M)	15779	·
Female (F)	11798	·
Adults (A)	·	·
Children (C)	3314	8973
High income (Hi)	5384	1865
Upper middle income (Umi)	8965	·
Lower middle income (Lmi)	8916	·
Low income (Li)	4312	2260

Table 2: Global estimate of modern slavery - Forced labor and forced marriage, 2022

(d) Among individuals facing modern slavery, what is the probability that the individual is a child doing forced labor? *0.5 pt*

(e) Given that the individual in modern slavery is in forced marriage, what is the probability that the person is in a high income country? *0.5 pt*

(f) Given that the person is suffering from forced marriage, what is the probability that the person is in a middle income country (upper or lower)? *1.5 pts*

(g) Are "age group" (A, C) and "type of modern slavery" (FL, FM) independent? *1 pt*

(h) Can you statistically test the (lack of) association of the two variables of the previous question? *(for these computations, you can use rounded numbers iteratively found. Assume that observations are independent. You are advised to use the chisq.test R command to verify that your computation is correct.)* *2 pts*

II Understanding the Data Used

8 pt

A researcher wants to study digital integration in Indonesia. In order to do so, the researcher wants to use the Digital Economy Household Survey 2020 from the World Bank. You are expected to look at the sampling, questionnaires and all relevant materials. **You do not need to use the dataset itself**.

1. Is the study an experiment? Justify. *0.5 pt*

2. Are all selected individuals necessarily responding to the investigator? *0.5 pt*

3. Among the variables in the Module "DEHS 2020 Module 2 Sec IV, Social media platforms", can you find (note that you should not pick *id* variables, you are asked to pick only one per category mentioned, that you should indicate the code of the variable and (very) briefly justify):

   (a) A continuous or a discrete variable *0.25 pt*

   (b) A ordered variable *0.25 pt*

   (c) A binary variable *0.25 pt*

   (d) A categorical variable *0.25 pt*

   (e) The level of measurement of the variable coded "m2_iv_22_4" *0.5 pt*

4. Module "dehs2020_individual_v2" has the following question (coded "m2_vi_43"): "Have you been victim of cyberbullying?" What should you keep in mind interpreting statistics built from such questions? Does the questionnaire try to address these issues? *0.75 pt*

5. Using the same module ("DEHS 2020 Module 2 Sec IV, Social media platforms"), can you indicate what is the expected number of individuals (not households) surveyed? *0.25 pt*

6. If the researcher wanted to implement the *merge*() operation in R so that some information on households specific information ("dehs2020_household_v2" module) and internet activities of individuals can be used together, what variable(s) would you use as key(s)? Explain. *0.25 pt*

7. What type of sampling design was implemented? *0.25 pt*

8. Can you indicate, if any, what are the issues of the sampling and/or the implementation of the survey? *0.5 pt*

9. What is the population in the study? *0.25 pt*

10. In the module "dehs2020_individual_v", it is found that 8274 out of 8620 respondents do not own a passport. Is the statement "96% of Indonesians do not own a passport" descriptive or inferential? What would be the complement statement (descriptive if you answered inferential and inferential if you answered descriptive)? *0.25 pt*

11. Given that 62.78% of surveyed were married, build a 99.9 confidence interval with respect to the share of the *population* married. Make sure to justify each step (among which why using a given critical value) and to interpret the confidence interval. *1.5 pts*

12. What is the minimum number of surveyed individuals necessary to have a margin of error equal to 0.01, when the level of significance remains $\alpha = 0.001$ and the variable of interest remains the proportion of married individuals? Comment on the relative change of the margin of error as compared to the change of the required sample size. Can you explain this discrepancy in the relative changes? *1.5 pts*

III Data Work

9 pt

In 2001, Acemoglu, Johnson and Robinson published an influential paper that aimed to conclude a long lasting debate about the core roots for development.[1] The main debate was trying to decipher what is the core element justifying development (or not) of countries over the long history. In this paper, the working hypothesis is the following:

$$\begin{array}{c} (Potential)\ Settler\ Mortality \\ \Downarrow \\ Settlements \\ \Downarrow \\ Early\ Institutions \\ \Downarrow \\ Current\ Institutions \\ \Downarrow \\ Current\ Performance \end{array}$$

One of the strengths of the paper is to convince the reader that the link between each step is a one-way arrow, hence *causality* rather than an association between these. They conclude that settler mortality in 1900 played a large role in the construction of early institutions (rule of law, democracy, *etc.*), which affects directly (they try to show that there are no confounding factors) current institutions, which affects current performance of countries.

In the present exercise, you are not aimed to discuss any causality but simply to perform some data preparation, to look at some visualizations of the data, and to compare some variables. You are advised to look at the variables' explanations in the appendix of the paper (page 53 of the .pdf file you are provided). Unless specified otherwise, all observations in the database(s) should be used.

1. For the variable “baseco”, replace NA observations with 0. How does it change the mean? How do you interpret it? *0.5 pt*

[1] Acemoglu, Daron, Simon Johnson, and James A. Robinson. ”The colonial origins of comparative development: An empirical investigation.” American economic review 91, no. 5 (2001): 1369-1401.

2. Construct a table of descriptive statistics including the following variables: “logpgp95”, “euro1900”, “cons1”, “cons90”, “logem4”, “democ00a”, “meantemp”, “lt100km”, “latabs”. (You may need to merge the databases beforehand.) You are to choose the statistics you want to have in the table and you can build the table in your preferred way, you should at least be able to comment on the centrality, the spread, the asymmetry . *0.75 pt*

   (a) Can you comment on the centrality, the spread, and any other relevant information that you deem necessary of the variable “euro1900”? *0.25 pt*

   (b) What variable displays the largest left skewness? What does it tell about the shape of the distribution of the present variable (comment only using the skewness coefficient)? *0.25 pt*

   (c) For all the following variables, what is the minimum share of observations that lies between ±2 standard deviations? *0.5 pt*

3. Construct the same descriptive statistics table but grouping by whether the country was or not colonized by the UK, France, Germany, Spain, Italy, Belgium, the Netherlands, or Portugal.

   (a) Are there any differences in sample sizes and data availability? (You are not asked to derive a formal statistical test but simply to quickly comment on these aspects.) What does it tell you about a potential comparison of, say, “meantemp” across these two groups? *0.75 pt*

   (b) Without making any formal test, interpret the share of European or of European descent in 1900 across both groups. *0.5 pt*

4. Construct the following data visualizations and interpret them.

   (a) We want to confirm the previous discussion on the differential distributions of the share of European or of European descents according to whether the country was colonized in 1900. Draw an overlay histogram and edit the x-axis, y-axis, title, legend title (put “Colony in 1900”) and the legend (put “Yes” and “No”, respectively). Does it confirm the discussion from the (by group) summary statistics table? *1.25 pts*

   (b) Now, we are interested in visualizing the (potential) link between (log) settler mortality and current institutions (the risk of expropriation in 1990). Can you show the joint distribution between these two variables and distinguish by the dummy variable “Africa” that is equal to zero when the country is not in Africa and to one when the country is in Africa. Do you see a difference in the association between (log) settler mortality and the risk of expropriation in 1990 whether the countries belong to the African continent or not? *1.25 pts*

5. The last visualization might push us to statistically test whether (log) Settler Mortality in 1900 differs according to whether the countries belong to Africa or not. Run a mean comparison test, and compare your conclusion from the correctly articulated hypothesis testing (with $\alpha = 0.01$) with the corresponding 99% confidence interval. Make sure beforehand that you have visually checked for Normality of the distributions and checked for equal variance. Comment on these two aspects (and assume that non-normality should not be a problem to build the hypothesis test). *3 pts*