Statistical Literacy — MINT

Lecture 12: Some extra content, wrap-up and exam information

Rémi Viné

The Graduate Institute | Geneva

December 11th, 2023

Outline

Housekeeping

Housekeeping

- Last lecture of the course; next week is the exam
- No problem set this week, as it is meant for your preparation of the exam
  - Hence, you were provided a Moodle-based mock exam
  - But review sessions this week to correct PS11 (not about revisions - this is your own work)

To go further

Overall

Takeaways

- The class aimed to give you the foundations of statistics and its basic vocabulary
- This is essential to have a critical mind when reading quantitative studies
  - To better pinpoint misinterpretations (a causal interpretation based on correlations...)
  - To better know what visualization to use and how to produce a correct graph
  - To possibly understand if better choices could have been made
- Ability to conduct some simple analysis in your ARP or any other projects

Go further

Know what you know, and what you don't know

- Inferential statistics is a very diverse field
  - We covered some **parametric models**
    - Non-parameteric, semi-parametric also exist
  - We covered some **classical (or frequentist) inference**
    - Bayesian (mostly for dynamic analyses), Predictive (for ML cases) types of inference also exist
  - We used some **supervised learning** methods
    - There exist many unsupervised learning methods (*e.g.*, clustering, Principal Component Analysis)

On the assignment

Assignment on descriptive statistics

- Data is complex
  - Never underestimate the time to
    - Get access to the data
    - Clean data sets
    - Harmonize different data sets to then merge them
- Visualize data
  - Choose the right graph, sort the variable correctly, label axes (with the units)
- Interpret data
  - You must be specific: "Out of the 179 students in the sample, the mean age is about 25 years while the median is lower by 1.4 years, which implies to suspect some right skewness in the distribution of age."

▸ Illustrations

Assignment on inference

- Confidence intervals
  - Proceed step by step, use Stata only for calculations
  - Validity conditions matter ($npq >> 1$ for proportion, $n > 30$ for averages) - and show the calculations!
  - Be clear with the correct point estimate: $\hat{p}$ for proportions, $\bar{x}$ for averages
- Hypothesis testing
  - Compare correctly $|test - statistic| \;\; \& \;\; |criticalvalue|$

Exam - Logistic

The exam

Generalities

- The exam is worth 45% of total grade, it covers all content of the class, 90 min (fixed on Moodle)
- **You must be in the auditorium (A1A, as usual) at 2pm**
  - → **You must not be late**, as you won't be authorized to do the exam
- This is an individual exam, any means of communication is strictly prohibited (digital or not)
- Your laptop must have only one internet browser open, with only one tab: the Moodle tab for the exam itself, you can have any PDF documents open
  - → If we see any other tab or browser, that will justify a zero grade

The exam

Generalities - cont'd

- It will be conducted on Moodle, a multiple-choice exam
- It is an in-class exam, no negative points for wrong answers (unless multiple solutions - the answered with the "squares" - same as the mock exam)
- You are not expected to use Stata for the exam, but to understand statistical outputs (regardless of the software used)

The exam

Material

- We will provide you the draft paper; you must sit where draft will be placed
- Bring working pen(s)
- You can use any paper-based documents, along with PDF documents (PS corrections, slides - even with annotations)
- Use the Z table and the T table (in PDF) that you would have downloaded from Moodle
- You need a basic (working) calculator
- Bring a working laptop and its charger (if you do not have any, reach out as soon as possible and contact the IT)
- You **must** have your student's card (if lost, contact us - much - before, and come with an ID with a picture)

Exam - content and a few examples

Content: all the class material

Non-exhaustive

- Descriptive statistics
  - Understand a data set (Sampling, biases, types and units of variables)
  - Centrality, dispersion, skewness (Indicators, Visualization)
- Inferential statistics
  - Probability 1.0.1. (Axioms, Key Properties)
  - Random variables (Discrete, Continuous)
  - Distributions, in particular the Normal and its standardization
  - Confidence intervals (For Proportions, For Averages, Confidence Level, Critical Value)
  - Hypothesis testings (For Averages, Types of errors, Significance Level, Test-Statistic)
  - Correlation, linear regressions (Interpret - coefficients and model, Select, Predict, Check Inference, Underlying Assumptions)

Exam

- You are strongly advised to complete the mock exam that is available on Moodle
  - It helps make sure you are clear with the different points covered in class
  - It helps grasp the format of the exam: Moodle-based, multiple-choice questions
- Beware that the exam might be slightly different: more (or less) questions, different topics covered in the class than those in the mock exam

Example 1

- Assume the waiting time in the population follows a Normal distribution. A representative sample of $n = 12$ is taken, what can be concluded regarding the sampling distribution?

  1. The sampling distribution has 12 observations
  2. The sampling distribution is Normally distributed
  3. The sampling distribution has at most 89% of all observations located in the interval $\mu_{\bar{x}} \pm 2\sigma_{\bar{x}}$

Example 2

- With a population data of $N = 140$, the average of a given variable, say the daily spending in supermarkets, is exactly CHF 90.
  Is it possible for the median to be identical?

  1. Yes
  2. No
  3. More information would be needed

  Is it possible that the Mode is larger than both the average and the median?

  1. Yes
  2. No
  3. More information would be needed

Example 3

- A researcher takes a representative sample of 51 people and investigates their wake up time (in minutes starting at midnight). They find an average of 440 minutes and a standard deviation of 70 minutes. Looking at these results, a colleague concludes that people in the population wake up on average at 8 am. A $\alpha = 1\%$ you want to verify such a statement.

- What distribution is required to find the critical value?

  1. Z distribution
  2. T distribution

- What is the critical value? (write it rounding at the third decimal digit)

$\rightarrow$

What next?

Back to square one

- Statistical methods are powerful
  - They will allow you to be assertive on relationships between variables of interest (but not here)
    - → They might even give you a *causal* link
  - They will provide you with a magnitude of relationships
- But be extra cautious!!
  - Make sure to only use what you understand
    - → This is particularly important in case of statistics using software with commands, packages that are already giving you everything
  - Occam's razor: make sure to only use what is needed
    - → Popper's falsifiability principle: the easier your way to address a problem the more "scientific" it is, as refutability is more easily tested
    - → Often the cure is worse than the disease: an intricate technique to overcome one issue you may have with the data might lead to extra assumptions, lack of interpretability of your results, *etc.*

Next?

- You can pursue the statistical journey by taking the more advanced class and thus learn specific tools to implement quantitative analysis
  - $\Rightarrow$ Learn about **causality**
  - $\Rightarrow$ Look at academic papers, understand, replicate, and assess the core analysis in them

$\rightarrow$ Be rigorous and prove B. Disraeli wrong:

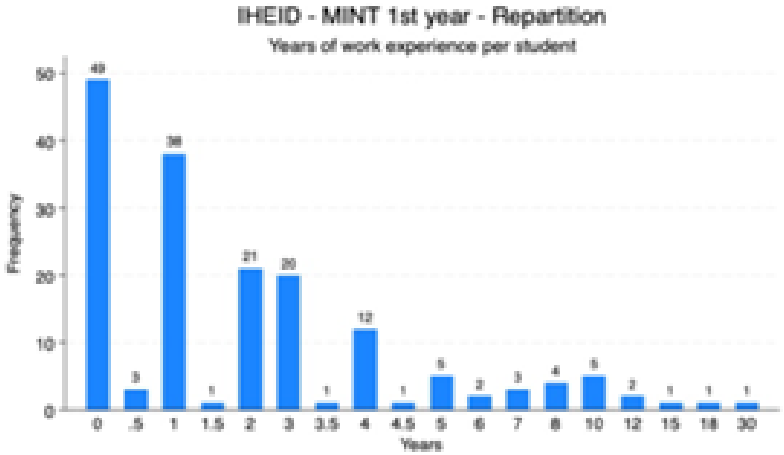
"There are three types of lies – lies, damn lies, and statistics."

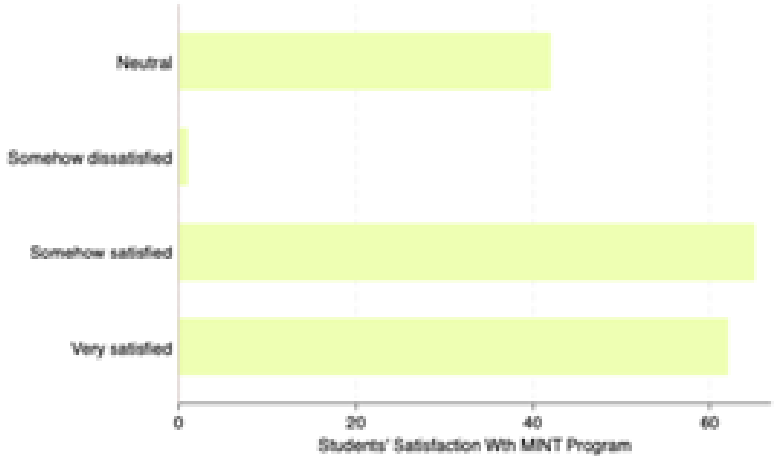
Last words

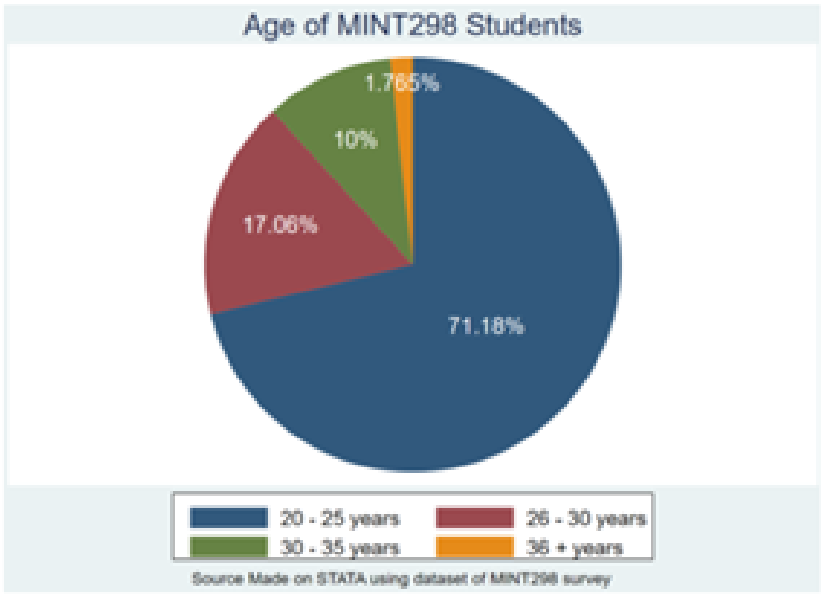
- Thank you to all Teaching Assistants, Antoine, Guilherme, Simeon, and Xiaoyu for their outstanding work
- Good luck with what comes next to you!

Assignment on inference

Choice of graphs



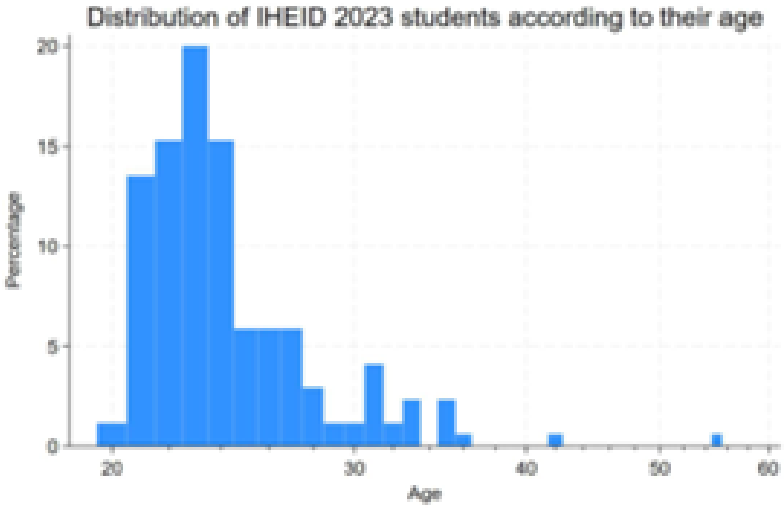


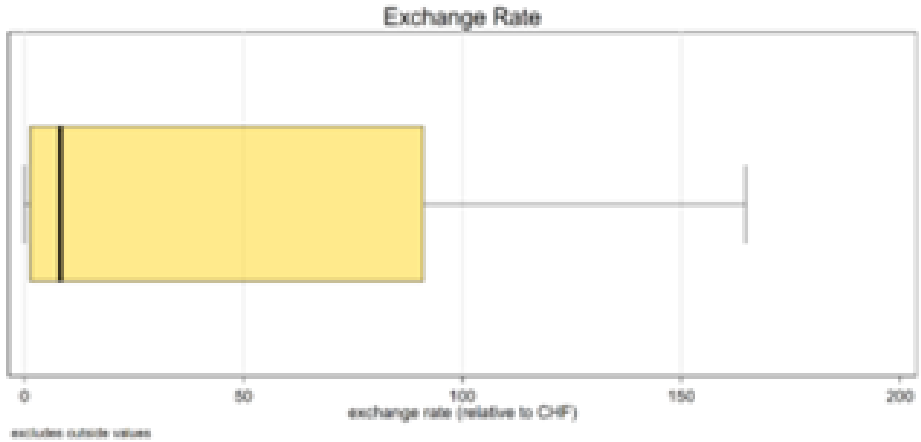


Back

Assignment on inference

Labels, dropping observations





Back