Statistical Literacy — MINT

Lecture 1: Introduction

Rémi Viné

The Graduate Institute | Geneva

September 25th, 2023

Outline

- Course information
  - Course objectives
- Introductions
  - Logistics
  - Consultations, assessment, and feedback
  - Course material and schedule
- Background of statistics
  - Quick mathematical notation refresher
  - Percentages and growth
  - Datum, Data
  - Variables
  - Data in hands
  - Descriptive and inferential statistics
  - Sampling and biases
  - Biases

Course objectives

- This course aims to help you develop
  - a thorough **understanding** of basic statistical concepts and methods used in social science,
  - the ability **to follow** and **critically assess** empirical literature using statistical analyses, and
  - the ability **to use** statistical software (Stata) to manage and analyze data.

Why learn Stats?

- You can use statistics to:
  - **describe** and **examine** patterns and trends in political phenomena (*ex.*, the proliferation of bilateral investment treaties);
  - to help **explain the variation** in certain outcomes of interest (*ex.*, differences in the design of institutions, variation in patterns of conflict and cooperation);
  - to **estimate the effect** of certain events or interventions (*ex.*, how a new legislation affects policy outcomes).
- Even if you do not use statistics, having a basic understanding of how the most common techniques work will be useful to understand and engage with the broader literature.

Statistical use in practice

- Whether a statistical approach is appropriate for you will depend on:
  - your **research question** and the focus on measurement or on causation,
  - your **familiarity & comfort level with statistical methods**,
  - and **data availability and other constraints** (access to data, time, *etc.*).

Introductions

- **Instructor**: Rémi Viné
- **Teaching assistants:**
  - Antoine Cornevin
  - Simeon Lauterbach
  - Guilherme Suedekum
  - Xiaoyu Yan

Logistics

- Introductory course in statistical literacy.
  - Next semester you will have some more advanced classes
- The course meets:
  - On **Mondays, 2:15-4 pm** (lectures, in **A1A**)
- The course Moodle page contains the syllabus, lecture slides, assignments, additional resources, as well as a Forum for class-related discussions.
- You must have a basic calculator (for the exam).

Logistics of review sessions

- There will be weekly review sessions, starting from this week
  - By now, you should have filled up the doodle form and been allocated a review session (you cannot change review sessions)
- Review sessions will be performed by the teaching assistants
  - Antoine Cornevin
  - Simeon Lauterbach
  - Guilherme Suedekum
  - Xiaoyu Yan

→ Participation is mandatory

Weekly schedule and feedback

- We will upload the **lecture slides** right before each lecture.
- **Problem sets** will become available for download on Moodle after lectures (*i.e.*, Monday, 4 pm). They will be due before the subsequent week's lecture (*i.e.*, Monday, 2:15 pm), or before your review session if before the lecture.
    - Send your problem sets to your TA
    - Problem sets are mandatory and account for the overall grade
    - Some questions are labelled "Difficult". Although you are expected to try to solve them seriously, these contents will **not** be part of the final exam material
- We will post the **solutions to each problem set** on Moodle once the deadline for submission is past.

Consultations

- Our weekly **office hours** are as follows:

Rémi	Monday	4:15pm-6pm	P1-503
Antoine	Friday	10:15am-12pm	P1-655
Simeon	Monday	4:15pm-6pm	TBD
Guilherme	Wednesday	10:15am-12pm	P1-655
Xiaoyu	Tuesday	10:15am-12pm	P1-655

- **Ask questions on the Forum** instead of emailing them to us. Often a colleague is able to work through a problem with you and the resulting exchange benefits the whole class.

Course schedule

- Here is our schedule:
  - Descriptive statistics (W1-W4)
    - Fill up a survey (W1)
    - Assignment 1 (W4)
  - Probability, random variables (W4-W7)
  - Inference tools (W8-W10)
    - Assignment 2 (W10-11)
  - Linear regression (W11-W12)
  - *Final Exam* (W13)

Course material and resources

- Textbook: Diez, David, Mine Cetinkaya-Rundel, and Christopher Barr. 2019. *OpenIntro Statistics*. 4th Edition
  - Available for download for free.
  - All the compulsory readings are from this book. Optional supplementary readings and resources may be provided as needed.
- Another useful reference is Neil A. Weiss. 2015. *Introductory Statistics.* Pearson, 10th edition (some are available at the library)
- Additional resources (especially for the labs):
  - Long, James. 2019. *R Cookbook*. 2nd Edition.
  - Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* 1st Edition.
  - Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. Sage.

Statistical software: Stata

- We will be using **Stata** as our statistical software. There are many reasons for this:
  - It is widely used in social sciences (and in most institutions).
  - It is a powerful statistical software, and quite flexible.
  - A lot of resources for learning, doing things in Stata and troubleshooting; a **learner community** is always ready to help.
  - You will use it in advanced Stats class
  - Python (Jupyter Notebook) can integrate Stata
- How to install it?
  - Go to Student Toolbox → IT Service Desk
  - Or open directly this link and fill up the form
  - Please install Stata as soon as possible (if not done yet)

Assessments

All assessments but the final are take-home

Problem sets (15%)	To hand over, they will be corrected during review sessions
Review session attendance (10%)	Attend but also participate in review sessions
Assignments (30%)	Two assignments, each is worth 15%
Final exam (45%)	In-class exam

Collaboration policy

- You **may collaborate** in answering the weekly problem sets **if and only if**
  - you make a serious effort on your own first,
  - you write up your own answers (no copied sections from others),
  - you note down on your answer sheet the names of who you collaborated with.
- You **must not collaborate** *for* the two assignments.
- Your new best collaborator, **Chat GPT**? Forbidden for assignements (we will check), authorized - but not advised - for problem sets (you cannot use it for the final exam)

When (and to whom) to submit

- Deadlines for problem sets have been detailed, deadlines for assignments are just before the week's lecture (resp. Week 5 and Week 11)
- Submit your work to your assigned TA (you will receive emails from them)
- Unless motives for being late have been shared early enough to the academic coordinators and duly justified, your work will not be graded if you submit any problem sets and assignment late.

Participation

- We will strive to give all students **ample opportunity to participate** during review sessions. Teaching assistants are here to help you
- You are encouraged to use the Forum on Moodle to share interesting examples and participate in discussions on the use of statistics in social science

Background of statistics

- Quick mathematical refresher
- Percentage and changes
- What are data about?
- Variables: Types and Measurement
- Data in hand: population, sample
- Descriptive and inferential statistics
- Sampling and biases

Some vocabulary

- $<$ smaller than (strictly); $>$ larger than (strictly)
- $\leq$ smaller or equal to; $\geq$ larger or equal to
- $=$ equal to; $\neq$ not equal to (or different than);$\approx$ approximately equal to (useful when rounding)
- $+, -, \times, /$ are the classic operation symbols
- x^2 the square of x (*e.g.*, $3^2 = 9$), $\sqrt{x}$ square root ($\sqrt{9} = 3$)
- $\sum_i x_i$ the sum of all elements over i
- $P(A)$ the probability of an event (here "A")
- $\cup$ union of two events ($A \cup B$ is equivalent to "$A\ OR\ B$")
- $\cap$ intersection of two events ($A \cap B$ is equivalent to "$A\ AND\ B$")
- $\sim$: follows in distribution. For example, $X \sim N(\mu, \sigma^2)$ means "the random variable X follows a Normal distribution of parameters μ and σ^2"

What is a percentage?

- A number can be expressed as a fraction of 100 (*per cent*)
  - 28% of students are more than 1.8 meters high → 28 students out of a hundred are taller than 1.8 meters
  - $28\% = \frac{28}{100}$
- A percentage can be understood as a *share*
  - The share of students beyond 1.8 meters high is 28%

What is a growth rate?

- A growth rate measures a change of a quantitative variable over time
  - It is usually expressed in percentage change
- $Growth\ Rate = \frac{NewValue - OldValue}{OldValue} * 100$
  - Example (source: World Bank): Swiss Gross Domestic Product (GDP) in 2020 = USD 729.4bn and in 2021 = USD 760.2bn
  - GDP growth? $\frac{760.2-729.4}{729.4} * 100 = 4.2$
  - GDP growth in Switzerland from 2020 to 2021 is 4.2%

Refresher: rounding

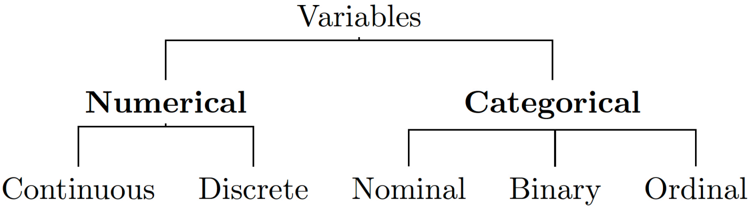
- To round value, round down when next digit is 4 or lower
  - Exception: if you need the minimum natural number, such as the lower number of people to include in a group, then round upward (*e.g.*, $21.1 \to 22$)
- No general rule for the correct decimal digit to use
- Round at third decimal digit:
  - 12.3785 (meters, height of a three floors house)
  - 760150000000 (CHF, Swiss GDP)
  - $0.000000000000000000000019942$ (grams, weight of a carbon atom)
    - Notation (Stata and most other statistical software): $1.9942e-23$

What is data about?

- A collection of facts (or assumed as such) that can be measured
- Useful for:
  - reasoning
  - calculation
  - finding patterns
  - association between different collections

Variables: definition and types

- A variable is a characteristic that can be measured
  - It can take several values or modalities
- There are different **types** of variables:



Variables: levels of measurement

- Different levels of measurement exist
  - Different levels of measurement = different possible calculations

Level of Measurement	Explanation	Example
Categorical	Collection of unordered modalities	
Ordinal	Collection of ordered modalities	
Interval	The zero is conventional, interpret differences	
Ratio	There is a *real* zero, interpret differences and growth	

Variables: levels of measurement

- Different levels of measurement exist
  - Different levels of measurement = different possible calculations

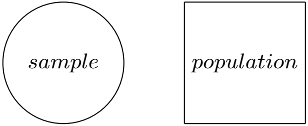
Level of Measurement	Explanation	Example
Categorical	Collection of unordered modalities	Colors, ingredients, sex
Ordinal	Collection of ordered modalities	Likert-scale variables
Interval	The zero is conventional, interpret differences	Temperature, GMAT score
Ratio	There is a *real* zero, interpret differences and growth	Weights, Heights

Data in hands

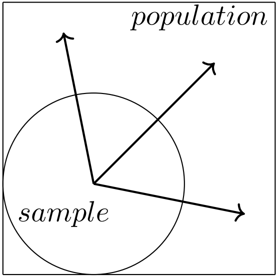
- Population
  - All units in consideration
  - A summary measure derived from population is called **parameter**
- Sample
  - A subset of the units in consideration
  - A summary measure derived from a sample is called **statistic**
  - A census is a subset including all units → population

Two main universes of statistics

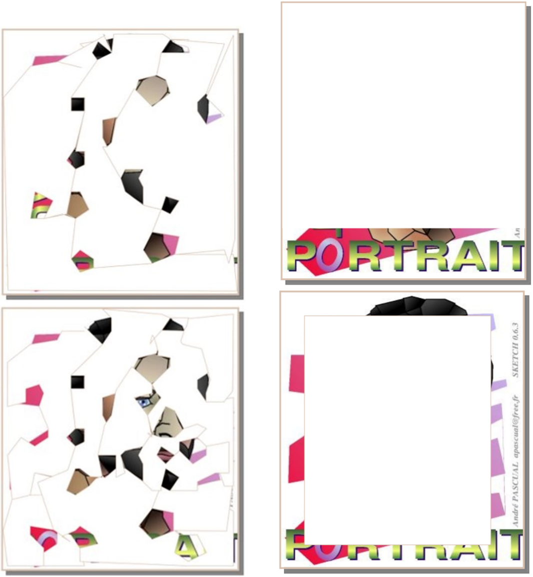
1. Descriptive statistics
   - Describe what is in the data available
     - Indicators (summary measures)
     - Tables
     - Charts



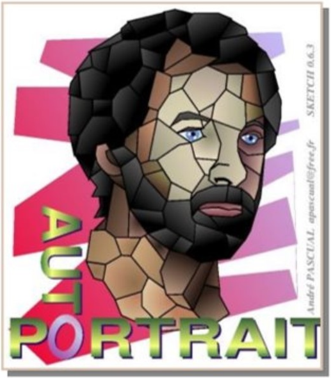
2. Inferential statistics
   - Go beyond and deliver general conclusion on the reality of the population
     - Predict
     - Generalize



Why does sampling (choice of the sample to select within the population) matter?



A good sample has to be **representative**



Sampling methods

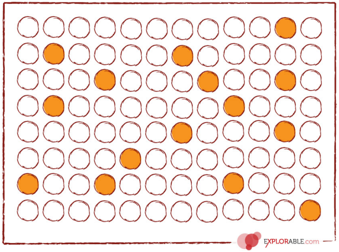
- A *probability sampling design* assigns non-zero probability to be selected to each unit in the population
- A *random sampling design* assigns *identical* probability to be selected to each unit in the population

- When representativity is verified (convincingly), inference can be used

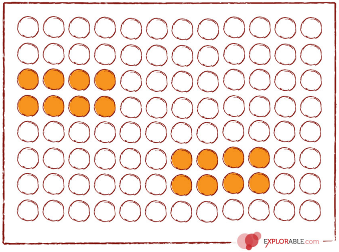
- We assume here that sampling is done *without replacement*
  - Once a unit is selected, its probability to be selected is now zero

Different sampling methods

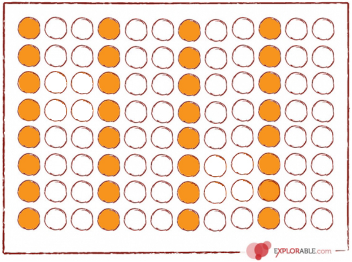
Random Sampling



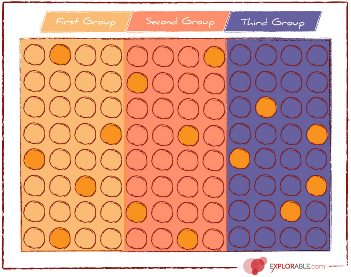
Clustered Sampling



Systematic Random Sampling



Stratified Sampling

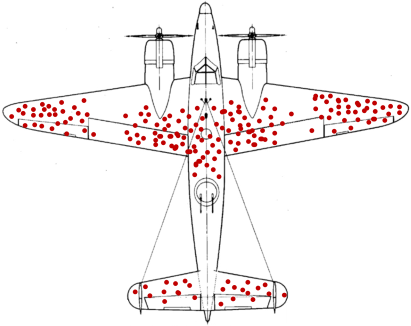


Biases in statistics

- **Biases** will lead to incorrect inference
  - Parameters are incorrectly estimated
- Different types of biases
  - Selection biases
    - Incorrect sampling design
  - Response biases
    - Knowledge issues, cultural issues, social confirmation
  - Non-response biases
    - The selected individuals refuses to participate

Biases in statistics - examples

- Asking:
  - How often do you take illegal drugs?
  - What is the total weekly income of your extended family in USD?
  - Using damaged aircraft coming back, study the weak spots of aircrafts

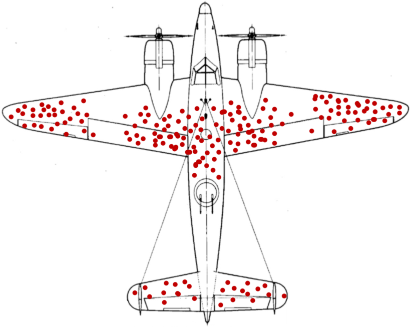


Hypothetical damage observed on a WW II aircraft.

By McGeddon, CC BY-SA 4.0.

Biases in statistics - examples

- Asking:
  - How often do you take illegal drugs? *[Non-response bias]*
  - What is the total weekly income of your extended family in USD? *[Response bias]*
  - Using damaged aircraft coming back, study the weak spots of aircrafts *[Selection bias]*



Hypothetical damage observed on a WW II aircraft.

By McGeddon, CC BY-SA 4.0.

An Important Warning: how to use the material learnt?

- 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 session

- Next session is on organizing data, calculating most usual centrality indicators

- **Do not forget to fill up the survey for assignments!**